Single Cell and Spatial 'Omics Analysis Infrastructure Roadmap for Australia

V4.1
13 December 2023

Tiffanie M Nelson and Jeffrey H Christiansen

DOI: 10.5281/zenodo.10368976

Contents

1. Executive Summary

Single cell and spatial 'omics is the process of identifying the unique transcriptome from each cell in a population of cells. Experiments identifying cell variation under various states or conditions have implications for understanding fundamental biological processes including disease states. Spatial omics or spatial technologies allows for the correct positioning of cells in complex tissues and environments which has a unique influence over their functioning. The broad benefits of single cell and spatial omics approaches include the identification of new cell states, classification of species, and exploration of cell type specific processes in development, treatment and disease. In Australia, single cell and spatial 'omics is conducted by researchers across a wide variety of life science domains.

This document includes:

- a brief summary of single cell and spatial 'omics 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 single cell and spatial 'omics approaches across Australia), and
- a high-level description of key components of an envisaged shared national single cell and spatial 'omics analysis infrastructure for Australia, which, when implemented, would enable Australian researchers from a wide range of institutions to perform single cell and spatial 'omics work they would otherwise be unable to undertake because of the reported roadblocks, i.e.

  D1. A platform for performing single cell and/or spatial 'omics analyses: to provide all Australian researchers with access to a shared platform with tools and workflows for single cell and/or spatial 'omics analysis, 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 statistical analyses and visualisations of single cell and/or spatial 'omics analysis results: to make it easier for Australian researchers to perform relevant statistical and/or visualisation-based analyses of single cell and/or spatial 'omics data.

  D3. Systems to enable sharing and submission of single cell and/or spatial 'omics data and associated output files from Australia to appropriate global repositories: to make it easier for any Australian researcher to share and publish their single cell and/or spatial 'omics 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 single cell and spatial 'omics analyses. Following engagement with other stakeholder groups (i.e. international entities operating single cell and spatial 'omics 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 Q3 2023.

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 single cell and/or spatial 'omics analysis, in late February 2022, the Australian BioCommons convened a "single cell and spatial 'omics Special Interest Group (SIG)[4]" and invited participation from over 100 researchers across Australia with either experience in, or interest in single cell and/or spatial 'omics[5].

The outcome of the survey, that meeting and further consultations 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 single cell and/or spatial 'omics.

[1] Bioplatforms Australia

[2] National Collaborative Infrastructure Strategy (NCRIS)

[3] Australian BioCommons

[4] When this community was convened it was advertised as a community for practitioners of single cell omics methods. Since the time of the meeting, it has been realised that this community of practitioners should be broadened to include both the single cell and spatial 'omics community.

[5] see Section 3.2 for methodology employed for formation of the group and membership

Following engagement with other stakeholder groups (i.e. international entities operating single cell and/or spatial 'omics analysis 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 Q3 2023.

3. Single Cell and Spatial 'Omics Analysis - Methods and Community

3.1 What are single cell and spatial 'omics, how are they done, and why?

Single cell and spatial 'omics analysis are emerging disciplines that stem from various ‘omics’ technologies, particularly genomics, transcriptomics, and epigenomics, where the sensitivity is improved to a single cell level[6,7]. Applying these technologies at the single cell level is crucial since every cell is specific and unique; nearly every cell population, derived *in vivo* or *in vitro*, is heterogeneous. The emergence of spatial molecular profiling or spatial ‘omics technologies is the process of studying the profiles of cells within their native spatial context within tissues[8]. Single cell and spatial ‘omics technologies are providing unprecedented resolution for the understanding of tissues and systems, and the interaction of single cells on a global scale. Single cell and spatial 'omics’ technologies (inclusive of single cell genomics and transcriptomics and more recently single cell proteomics[9] and metabolomics[10]) have developed and scaled at speed since the first single cell transcriptomic experiment was performed in 2009[11]. The focus of this document is on the technologies of single cell and spatial 'omics and spatial ‘omics technologies.

Early molecular profiling studies had to be undertaken on large biological samples which contained many cells (e.g. hundreds to billions of cells) in order to overcome the small quantities of material present in cells. When undertaken, analyses of this type (e.g. RNA-seq or ribonucleic acid sequencing) result in a molecular profile of the sample that is an average of the contributions from all of the pooled cells within the sample. However, when averaging contributions from many cells, the details of each cellular unit and their individual contents are also pooled and the opportunity to identify different cellular gene expression patterns within/across the sample is lost. By comparison, single cell profiling methods such as single cell RNA-seq (scRNA-seq) for example, permits the comparison of the transcriptomes of individual cells in parallel. Molecular detection methods with single cell resolution along with the accompanying computational methods have scaled steadily and the profiling capability has grown from thousands of individual cells to hundreds of thousands of cells and now also includes the integration of spatial locations within tissue sections[12].

[6] Stein, C. et al. 2021, Journal of Cellular Biochemistry, https://doi.org/10.1002/jcb.30134
[7] Aldridge, S. and Teichmann, S.A., 2020, Nature Communications, https://doi.org/10.1038/s41467-020-18158-5 .
[8] Palla, G. et al. 2022, Nature Biotechnology, https://www.nature.com/articles/s41587-021-01182-1
[9] Kelly, R. 2020, Molecular and Cellular Proteomics, https://www.mcponline.org/article/S1535-9476(20)35153-7/abstract
[10] Lanekoff, I. et al. 2022, Current Opinion in Biotechnology, https://www.mcponline.org/article/S1535-9476(20)35153-7/abstract
[11] Tang, F. et al., 2009, Nature Methods, https://www.nature.com/articles/nmeth.1315
[12] Aldridge, S. and Teichmann, S.A., 2020, Nature Communications, https://doi.org/10.1038/s41467-020-18158-5 .

Early techniques to resolve the differences in single cells used low throughput methods to identify changes in the expression of a single or few genes, including immunohistochemistry coupled with microscopy, single-cell quantitative polymerase chain reaction (qPCR) or single-molecule RNA fluorescence in situ hybridisation (RNA FISH)[13]. Experimental advancements driven by the limitations of these experimental methods, improved the high-throughput generation of cDNA (complementary deoxyribonucleic acid) libraries from messenger RNA (mRNA). For methods that capture the spatial context of cells in tissues, prior to nucleic acid extraction, samples are prepared by creating spatially resolved samples using methods such as laser capture microdissection and tissue image capture to be integrated follow ‘omic data generation. Now, protocols for performing high-throughput single cell and spatial ‘omics are numerous with evolving wet lab methodologies and a list of some of the more common techniques are listed in Table 1.

Table 1. Meaning and Definition of Single Cell and Omics Molecular Techniques

Acronym abbreviation	Acronym meaning and technique description
scRNA-seq, scRNA, RNA-seq	Single cell ribonucleic acid (transcriptome) sequencing.
scATAC-seq	Single cell assay for transposase-accessible chromatin (ATAC) or cis-regulatory elements, also known as a cell’s ‘regulome’.
scDNase-seq	Single cell DNase I hypersensitive site sequencing or cis-regulatory elements, also known as a cell’s ‘regulome’.
scChIP-seq	Single cell chromatin immunoprecipitation (ChIP) identifies the binding sites of DNA-associated proteins which can be used to map global binding sites for a given protein.
Geo-seq	Geographical position sequencing. Combines laser capture microdissection and single cell RNA (transcriptome) sequencing.
tomo-seq	Provides genome-wide expression data with spatial information. Performs RNA-seq on individual cryosections.
CITE-seq	Cellular indexing of transcriptomes and epitopes by sequencing. Oligonucleotide-labelled antibodies are used to integrate cellular protein and transcriptome measurements into an efficient, single-cell readout.
Slide-seq	RNA spatially resolved from tissue sections by transfer onto a surface covered with DNA-barcoded beads. Reveals spatial gene expression patterns.
STRT-seq, STRT	Single-cell tagged reverse transcription sequencing, STRT-seq utilises 5' tag counting of transcripts. performed on microfluidics Fluidigm C1 platform.
MARS-seq	Massively parallel RNA single-cell sequencing. Incorporates unique molecular identifiers into the cDNA.
SMART-seq, SMART-seq2	Switching mechanism at 5′ end of RNA template sequencing. Generates full-length cDNA.

[13] Stegle, O. et al., 2015, Nature Reviews Genetics, https://pubmed.ncbi.nlm.nih.gov/25628217/

CEL-seq, CEL-seq2	Cell expression by linear amplification and sequencing. Incorporates unique molecular identifiers into the cDNA.
Visium Spatial Gene Expression	Next-generation molecular profiling for classifying tissue based on total mRNA. Allows analysis of the transcriptome within the tissue context, which means that the transcript is captured within the tissue, followed by sequencing outside the tissue.
GeoMX Digital Spatial Profiler	Shines UV light on regions of interest from whole, fresh or frozen tissue sections to release photo-cleavable gene barcodes for quantification with either targeted or untargeted next generation sequencing.
Xenium	Fresh or frozen sections can be used to access the RNA for labelling with circularised DNA probes. Probe ligation generates a circular DNA probe which is enzymatically amplified. Slides are placed in the Xenium Analyzer where the sample undergoes successive rounds of fluorescent probe hybridization, imaging, and removal; creating bright, easy to image signals with a high signal-to-noise ratio. An optical signature specific to each gene is generated, enabling target gene identification. Finally, a spatial map of the transcripts is built across the entire tissue section.
CosMx Spatial Molecular Imaging (SMI)	CosMx SMI enables rapid quantification and visualisation of up to 1,000 RNA and 64 validated protein analytes. CasMx SMI provides spatial multi omics (whole transcriptome of RNA expression and protein abundance with spatial context) on formalin-fixed paraffin-embedded and fresh frozen tissue samples at cellular and subcellular resolution.s an integrated system with mature cyclic fluorescent in situ hybridization (FISH) chemistry, high-resolution imaging readout.
merFISH	Multiplexed Error-Robust Fluorescence in situ Hybridization (MERFISH) spatial profiling technology works. MERFISH is a spatially resolved single-cell transcriptome profiling technology, capable of simultaneously measuring the copy number and spatial distribution of hundreds to tens of thousands of RNA species in individual cells.
STOmics	SpaTial Enhanced REsolution Omics-Sequencing (Stereo-seq) technology. Stereo-seq offers researchers a novel tool to explore spatial biology with unprecedented field-of-view and resolution. STOmics-GeneExpression-S1 combines centimetre-field-of-view in situ capture of whole transcriptome information with nanoscale resolution, powering the unrivalled quality of Stereo-seq. STOmics-GeneExpression-S1 Chips are patterned grids of probes containing spatial coordinates. Upon interaction with a tissue section, cDNA is synthesised in situ from mRNA captured by the chip probes. Sequencing the cDNA with their spatial coordinates allows *in silico* reconstitution of the spatial transcriptomic profile of the tissue section, allowing easy visualisation and analysis.

These protocols typically begin by isolating and sorting individual viable cells from the tissue of interest by fluorescence-activated cell sorting (FACS) or by exploiting a microfluidics-based system[7,14]. Emerging techniques are now however evolving without the need to isolate single viable cells, including isolation of single nuclei and multiple rounds of aliquoting and pooling know as split-pooling to introduce unique barcodes to sample material [15,16]. Following isolation of the cells, they are lysed, and as many RNA molecules present in

[14] Haque, A. et al., 2017, Genome Medicine, https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0467-4
[15] Rosenberg, A.B. et al, 2018, Science, https://pubmed.ncbi.nlm.nih.gov/29545511/
[16] Habib, N. et al, 2016, Science, https://pubmed.ncbi.nlm.nih.gov/27471252/

the sample as possible are captured for subsequent steps in the process. Primers are used to preferentially capture mRNA over ribosomal RNAs and mRNA is converted to cDNA by reverse transcription[8]. The cDNA is then amplified with optional steps taken to preserve the cellular origin. The amplified cells are then pooled and sequenced by next generation sequencing using library preparation techniques and nucleic acid sequencing platforms similar to those used for bulk samples[8]. A high-level conceptual workflow showing the general steps that are taken to perform scRNA-seq are shown in Figure 1.

Commercial lab kits and reagents exist for all of the stages of wet-lab scRNA-seq protocols and (other than a multi-channel pipette) removes the need for any extra lab hardware required to perform scRNA-seq, although the reagent costs can be substantial[8]. Droplet-based instruments can encapsulate thousands of single-cells individually, each containing necessary reagents for transcription and molecular tagging thus eliminating the steps required for cell sorting and isolation from tissues. However, these approaches require investment in dedicated hardware and might not be available to a researcher considering scRNA-seq for the first time[8].

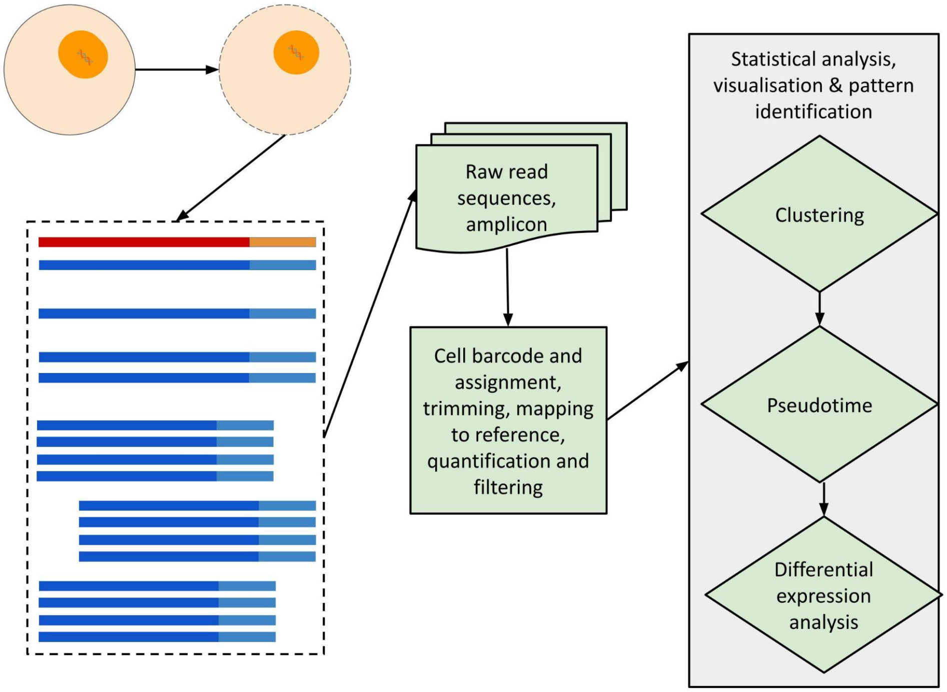

Raw read sequences, amplicon
Cell barcode and assignment, trimming, mapping to reference, quantification and filtering
Statistical analysis, visualisation & pattern identification
Clustering
Pseudotime
Differential expression analysis

Figure 1: A general single-cell RNA-seq workflow

A general workflow of single-cell RNA sequence experiments including : 1/ isolation of single cells, 2/ cell lysis while preserving mRNA, 3/ mRNA capture, reverse transcription of primed RNA into complementary DNA followed by cDNA amplification, 4/ raw sequencing reads are treated for quality control and cell assignment to create an output table. Finally 5/ the output data are visualised and biologically meaningful patterns are identified. Workflow was adapted from and Haque, A., et al., 2021[17] and Andrews, T.S., et al., 2019[18]. A high resolution copy of this diagram is located here.

Following data generation there are a number of steps that are applied to the raw data to complete analysis as shown in Figure 1. Initially following generation, raw data are pre-processed and then if data are from multiple methods or platforms, they are normalised to ensure comparability across samples or experiments. Low quality cells and genes are removed from the analysis using various bioinformatic tools[19] and the data is reduced (a requirement for high dimensional[20] data such as single cell and spatial 'omics data) for easier visualisation and analysis prior to cellular clustering. Identification of cell populations use unsupervised clustering cellular transcriptomic profiles to identify the cell types or states [21]. Additionally, to capture spatial information of the cells when performing spatial 'omics techniques, captured images are analysed via image analysis which involves identifying the cells and then mapping their spatial location onto the tissue[22]. As the data generated from a single cell and spatial 'omic experiment can include expression data for several thousand(s) of genes over millions of cells, differential expression analysis is the primary downstream process used to identify gene markers for cell type detection and also provide inputs to other analyses[23]. Finally, cell trajectory analysis may be performed to allocate cells to lineages and order them based on pseudo times within the lineages of different cell types or states[24].

Initial scRNA-seq studies examined human and mouse primary cells, such as embryos, tumours, neural tissue and lymphocytes[8]. The resolution afforded by scRNA-seq studies have enabled the assessment of transcriptional similarities and differences within populations of cells that were previously assumed to be fairly homogeneous and has revealed higher than expected levels of heterogeneity in samples ranging from immune cells and embryonic cells, for example[8]. Application of scRNA-seq to the study of human disease including cancer and auto-inflammatory disease have uncovered novel cell states when comparing healthy and diseased tissues[25]. The examination of transcriptional differences between individual cells that are rare, such as malignant tumour cells within a tumour mass[26] or identifying the expression for unique cells, such as individual T lymphocytes[27] reveals

[17] Haque, A. et al., 2017, Genome Medicine, https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0467-4

[18] Andrews, T.S., et al., 2021, Nature Protocols, https://www.nature.com/articles/s41596-020-00409-w

[19] Chen, G. et al. 2019, Frontiers in Genetics, https://www.frontiersin.org/articles/10.3389/fgene.2019.00317/full

[20] High dimensional data refers to data that has a greater number of variables than it does observations. For single cell transcriptomic data this refers to a greater number of single cell transcriptomes than tissues sampled. Further information is detailed in Wu, Y. and Khan, K. 2020, Nature Reviews Nephrology, https://www.nature.com/articles/s41581-020-0262-0

[21] Ianevski, A. et al. 2022, Nature Communications, https://www.nature.com/articles/s41467-022-28803-w

[22] Williams, C. G. et al. 2022, Genome Medicine, https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-022-01075-1

[23] Das, S. et al. 2022, Entropy (Basel), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9315519/

[24] Van den Berge, K. et al. 2020, Nature Communications, https://www.nature.com/articles/s41467-020-14766-3

[25] Vieira Braga, F.A. et al. 2019, Nature Methods, https://pubmed.ncbi.nlm.nih.gov/31209336/

[26] Miyamoto, D.T. et al, 2015, Science, https://www.science.org/doi/10.1126/science.aab0917.

[27] Stubbington, M.T.J. et al., 2016, Nature Methods, https://www.nature.com/articles/nmeth.3800.

potential new therapeutic targets. Single-cell transcriptomics has also been used to advance biological understanding of virus transmission during the COVID-19 pandemic and gives a valuable platform for investigating infection, transmission and strategies for prevention[6].

3.2 Who in Australia is performing single cell and/or spatial 'omics analysis, and which species are they tackling?

The increasing accessibility to methods that do not require specialist hardware, along with a reduction in nucleic acid sequencing cost and greater data availability has reduced the barrier for many researchers to carry out single cell molecular profiling projects. During the emergence and early development (from 2009 to 2015) of single cell and spatial 'omics methods, scRNA-seq was conducted by specialist research groups[28], primarily due to the specialised and costly hardware required, such as a FACS machine or microfluidics platform. More recently the technologies and tools needed to conduct scRNA-seq studies have become accessible to more biomedical researchers and clinicians, and this has greatly democratised access to the methodology and progressed the global understanding of the cellular heterogeneity and characteristics of gene expression[28].

Medical science is at the cutting edge of single cell and spatial 'omics technology and data generation[29] for its ability to identify rare cells that play a role in disease[30], understanding the heterogeneity of cancers[31] and characterising immune cells[32]. In the field of agriculture, single cell and spatial ‘omics technologies have many applications in the areas of crop improvement and livestock breeding[33], and specifically in the adoption of single cell and spatial 'omics technologies to provide insight into gene expression changes and marker gene discovery in crops/livestock at specific developmental stages or under varying abiotic conditions[34,35].

Supporting the generation of genomic and other ‘omic 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[36], e.g. Agricultural Science[37] , Nutrition

[28] Stein, C. et al. 2021, Journal of Cellular Biochemistry, https://doi.org/10.1002/jcb.30134
[29] Stein, C. et al, 2021, J Cell Biochemistry, https://doi.org/10.1002/jcb.30134
[30] Jagadeesh, K.A. et al. 2022, Nature Genetics, https://www.nature.com/articles/s41588-022-01187-9
[31] Ren, X. et al. 2018, Genome Biology, https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1593-z
[32] Gusland, N. C. et al. 2020, Frontiers in Immunology, https://www.frontiersin.org/articles/10.3389/fimmu.2020.559555/full
[33] Srivastava, U. and Singh, S. 2022, Genomics of Cereal Crops, https://experiments.springernature.com/articles/10.1007/978-1-0716-2533-0_14
[34] Ryu, K.H., et al., 2019, Plant Physiology, https://doi.org/10.1104/pp.18.01482
[35] Wang, Y. et al., 2021, Journal of Genetics and Genomics, https://doi.org/10.1016/j.jgg.2021.06.001
[36] 10-year strategic plans for science disciplines, developed by the Australian Academy of Science’s National Committees for Science.
[37] The Decadal Plan for Australian Agriculture Sciences, 2017 - 2026, science.org.au/support/analysis/decadal-plans-science/decadal-plan-agricultural-sciences-2017-2026

Science[38] as well as the Australian Government Department of Health's Primary Health Care Plan[39] and Long Term National Health Plan[40].

Undertaking single cell and spatial 'omics 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.

There are many groups and consortia across Australia who are actively working on single cell and spatial 'omics projects. The scientific literature indicates an approximate number of studies using single cell and spatial 'omics, or the technologies, such as scRNAseq or smart-seq produced from Australian-based researchers (see Figure 2).

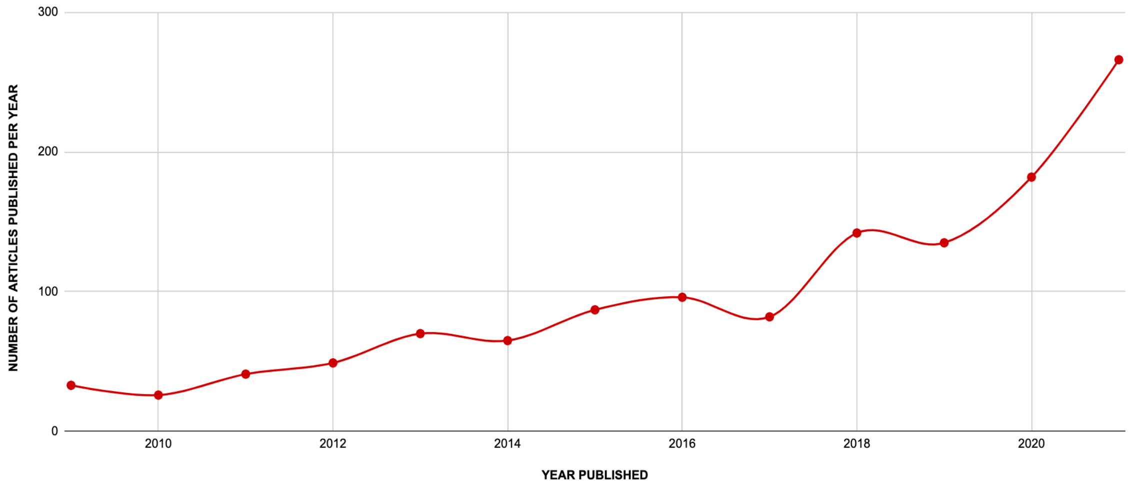


Figure 2: Estimates of the increasing number of single cell and spatial 'omics studies conducted in Australia

To gain an estimate of the number of single cell and spatial 'omics studies that have been conducted historically in Australia, a search was conducted of the Scopus database for articles with the following in the title, abstract, or keyword: single cell/s, single-cell/s, scrna, scatac-seq, geo-seq, tomo-seq, cite-seq, scrna-seq, slide-seq, strt-seq, mars-seq, smart-seq, smart-seq2, cel-seq 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 studies whose focus was on developing or evaluating analysis methods or tools. The complete list of citations including authors, titles and journals can be found here.

[38] Nourishing Australia: A decadal plan for the science of nutrition, 2019 - 2029, https://www.science.org.au/supporting-science/science-policy-and-analysis/decadal-plans-science/nourishing-australia-decadal-plan

[39] Future focused health care: Australia's Primary Health Care Plan, 2022-2032, https://www.health.gov.au/sites/default/files/documents/2022/03/australia-s-primary-health-care-10-year-plan-2022-2032-future-focused-primary-health-care-australia-s-primary-health-care-10-year-plan-2022-2032.pdf

[40] Australia's Long Term National Health Plan https://www.health.gov.au/resources/publications/australias-long-term-national-health-plan

Some of these groups include:

- Labs associated with the Oz Single Cells Consortium[41] which was formed in response to the Human Cell Atlas[42] with a goal to bring together skills and experience to combat disease with single cell technologies;
- Single Cell Omics Network (SCON), a collaborative effort to foster and support single cell research projects[43];
- The Stem Cells Framework Data Initiative which is using single cell technologies to study fundamental questions using stem cells[44];
- The Cellular Genomics Futures Institute whose focus is on the invention of new technologies to understand cellular nucleic acids and related proteins for the purpose of precise diagnosis and treatment of human disease[45];
- As well as researchers from many other institutions who utilise facilities who offer single cell technologies and analyses for single cell research projects at all stages of projects, such as the Baker Heart and Diabetes Institute's Single Cell Omics Platform[46]; Children's Medical Research Centre's CellBank Australia[47]; University of Queensland's Genome Innovation Hub[48]; University of New South Wales' Ramaciotti Centre for Genomics[49]; Monash University's Micromon[50]; Monash Health Translation Precint's Medical Genomics Facility[51]; WEHI's Advanced Genomics Facility[52] and QUT's Central Analytical Research Facility[53].

There are also a number of laboratory groups, consortia and divisions within Australian institutes and universities pursuing fundamental biology and disease with single cell

[41] Oz Single Cell Consortium, singlecells.org.au/
[42] Human Cell Atlas, https://www.humancellatlas.org/
[43] Garvan Institute of Medical Research's Single Cell Omics Network, garvan.org.au/research/garvan-weizmann/scon
[44] Bioplatforms Australia's Stem Cells Framework Data Initiative, bioplatforms.com/projects/stem-cells/
[45] Cellular Genomics Futures Institute, University of New South Wales, research.unsw.edu.au/cellulargenomicsfutures/institute
[46] Single Cell Omics Platform, baker.edu.au/research/research-platforms/single-cell
[47] CellBank Australia, cellbankaustralia.com/single-cell-solutions.html
[48] Genome Innovation Hub, gih.uq.edu.au/research/single-cell-and-short-read-sequencing
[49] Ramaciotti Centre for Genomics, ramaciotti.unsw.edu.au/services/next-generation-sequencing/single-cell-sequencing
[50] Micromon, https://www.monash.edu/researchinfrastructure/micromon/Services/next-gen-sequencing/single-cell-sequencing
[51] Medical Genomics Facility, https://www.mhtpmedicalgenomics.org.au/index.php/services/single-cell-genomics
[52] Walter and Eliza Hall Institute's Advanced Genomics Facility, wehi.edu.au/people/rory-bowden/4536/wehi-advanced-genomics-facility
[53] QUT Genomics Facility, qut.edu.au/research/why-qut/infrastructure/central-analytical-research-facility

technologies, including immunity and cancer[54,55,56,57,58,59,60,61,62,63,64,65] and others who focus on bioinformatics and tool development[66,67,68,69,70,71,72].

A unique focus that has developed alongside single cell and spatial ‘omics technologies are the global efforts to catalogue the discovery of these new cells and their gene activities. Although many of these cell atlases are not a focus of one country and instead are a collaborative effort, the inclusion here is to signify the importance of these catalogues to this emerging field and highlight the role Australian researchers have played in these developments. For instance, the Human Cell Atlas (HCA)[73] is an international collaborative project with a goal to create a comprehensive reference map of all human cells[74]. A number of researchers from a range of Australian research institutes are working on and contributing to the HCA. In particular, scientists from the Garvan Institute of Medical Research in Sydney are leading the development of the HCA Lung Biological Network which has a focus to create a detailed cellular map of the human lung[75]. Since the launch of the HCA, a number of related cell atlases have followed suit with contributions from a number of Australian research institutions, including the Mouse Cell Atlas[76], Plant Cell Atlas[77], Fly Cell Atlas[78], Malaria Cell Atlas[79] and the COVID-19 Cell Atlas[80]. Each resource is working towards understanding the role of gene expression under relevant conditions revealing new insights

[54] Turner Laboratory - Monash Biomedicine Discovery Institute, Monash University, https://www.monash.edu/discovery-institute/turner-lab/research
[55] Walter and Eliza Hall Institute, https://www.wehi.edu.au/
[56] Genome Biology and Genetics Program, Harry Perkins Institute of Medical Research, https://perkins.org.au/research/labs/genome-biology-genetics-program/
[57] Cancer Program, Harry Perkins Institute of Medical Research, https://perkins.org.au/research/labs/cancer-program/
[58] Regulatory Systems Medicine, Victor Chang Cardiac Research Institute, https://www.victorchang.edu.au/heart-research/regulatory-systems
[59]Genomics and Machine Learning Laboratory, Institute for Molecular Biology, University of Queensland, https://imb.uq.edu.au/research-groups/nguyen
[60] South Australian Immunogenomics Cancer Institute, University of Adelaide, https://www.adelaide.edu.au/saigenci/about-saigenci
[61] Garvan Institute of Medical Research, https://www.garvan.org.au/
[62] Wells Laboratory, University of Melbourne, https://biomedicalsciences.unimelb.edu.au/sbs-research-groups/anatomy-and-physiology-research/stem-cell-and-developmental-biology/wells-laboratory-stem-cell-systems
[63] Peter MacCallum Cancer Centre, https://www.petermac.org/
[64] The Florey Institute of Neuroscience and Mental Health, https://florey.edu.au/about
[65] Baker Heart and Diabetes Institute, https://baker.edu.au/
[66] Walter and Eliza Hall Institute, https://www.wehi.edu.au/people/
[67] Lê Cao Laboratory, University of Melbourne, https://lecao-lab.science.unimelb.edu.au/
[68] Computational Trans-Regulatory Biology Group, Charles Perkins Centre, University of Sydney, https://www.sydney.edu.au/charles-perkins-centre/our-research/research-groups/computational-trans-regulatory-biology-group.html
[69] Institute for Molecular Bioscience, University of Queensland, imb.uq.edu.au/about
[70] Garvan Institute of Medical Research, https://www.garvan.org.au/
[71] Diamantina Institute, University of Queensland, di.uq.edu.au/
[72] Oschlack Laboratory, Peter MacCallum Cancer Centre, https://www.petermac.org/research/labs/alicia-oshlack-0
[73] Human Cell Atlas, https://www.humancellatlas.org/
[74] Regev, A. et al., 2017, Elife, https://pubmed.ncbi.nlm.nih.gov/29206104/
[75] Schiller, H.B. et al., 2019, American Journal of Respiratory Cell and Molecular Biology, https://doi.org/10.1165/rcmb.2018-0416TR
[76] Mouse Cell Atlas, http://bis.zju.edu.cn/MCA/
[77] Plant Cell Atlas, https://www.plantcellatlas.org/
[78] Fly Cell Atlas, https://flycellatlas.org/
[79] Malaria Cell Atlas, https://www.malariacellatlas.org/
[80] COVID-19 Cell Atlas, https://www.covid19cellatlas.org/

into fundamental biological functions and specifically the way disease or symptoms manifest.

In February 2022, the Australian BioCommons invited over 100 researchers across Australia to participate in a single cell and spatial 'omics Special Interest Group (SIG). Researchers with experience in, or interest in, single cell and spatial 'omics were identified through a connection with the Oz Single Cell Consortium and elsewhere. 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[81] (number of respondents = 15), and also held an open-invitation video conference attended by an additional seven members of the community to follow-up and gain further information (minutes[82] and a recording[83] of the meeting are available). We additionally engaged multiple other single cell and spatial 'omics researchers in conversation at conferences and in related meetings to identify their needs and challenges[84].

3.3 How are single cell and spatial 'omics analyses being done in Australia?

Respondents to the survey (n = 15) indicated they are applying single cell and spatial 'omics analyses in the areas of human health / medical research (87% of respondents, *n* = 13) and/or fundamental / bioresearch (60% of respondents, *n* = 9). Fewer respondents to our survey (<20% of the respondents, *n* = 4) were using single cell and spatial 'omics technologies for projects classified as 'diagnostics' or 'environmental' fields of science.

3.3.1 Data

Based on information received through the survey (*n* = 15), to generate experimental data, the majority of researchers (80% of respondents, *n* = 12) use 10x Genomics[85] platform followed by 20% of respondents (*n* = 3) using NanoString[86] or Codex[87]. Four of the

[81] Presentation including survey results on single cell and spatial 'omics infrastructure needs and challenges conducted from 15/10/2021 to 30/02/2022 is here.
[82] Meeting minutes from the single cell and spatial 'omics SIG meeting held 3/03/2022 are here.
[83] Recording of single cell and spatial 'omics SIG meeting held 3/03/2022 is here.
[84] The Australian BioCommons engagement team attended the Multi-Omics Conference in December, 2022 where single cell and spatial 'omics researchers and practitioners were engaged. The team also attended and joined in meetings with the Melbourne Academic Centre for Health (MACH) Omics steering group committee to understand needs and challenges in September and October, 2022. The team also engaged a broad group of researchers, infrastructure providers dn informatics providers from Queensland-based universities and institutions on behalf of the Queensland Cyber Infrastructure Foundation to discuss needs and challenges in spatial and single cell and spatial 'omics in December 2022. At the beginning of the consultation period, the team also presented ideas and engaged members of the Oz Single Cells Steering Committee on 16th December 2021.
[85] 10x Genomics, https://www.10xgenomics.com/
[86] NanoString, https://nanostring.com/
[87] Codex DNA, https://codexdna.com/

respondents mentioned using another platform or labelling methodologies, e.g. Oxford Nanopore[88], inDrop[89], Hive[90], or Hyperion Imaging System[91].

The dominant ‘omics analyses being performed (by more than >70% of respondents, *n* = 13) currently and ongoing were transcriptomics and spatial analyses, with 60% of respondents (*n* = 8) performing multi-omics, seven respondents performing genomics and six respondents stating they performed epigenomics currently and plan to continue in the next five years. Three respondents are planning to use proteomics.

Aside from generating their own data, 87% respondents are accessing public datasets (n = 13) or private datasets (53% of respondents, *n* = 8) to aid in their analyses. The top three public databases being accessed to support their analyses are the Gene Expression Omnibus (GEO)[92], Sequence Read Archive (SRA)[93] and the Human Cell Atlas[94], and survey respondents also indicated that they access data held in six other databases, Sfaira[95], Human Protein Atlas[96], Tabula Muris[97], Tabula Sapiens[98] and Array Express[99]. Two respondents did not access existing datasets primarily due to the difficulty in integrating the datasets in their analyses as a result of poor or outdated metadata or due to a lack of suitable tools or pipelines (*n* = 1). One respondent also stated that no data existed or that the existing data could not be accessed.

Only two respondents of the 15 stated that they used a data or code/analysis management framework of any kind, i.e. R Markdown[100] and workflowr[101]. Others responded that they did not (*n* = 4) or left the response unanswered (*n* = 9). When asked about data sharing, respondents noted that data was shared with colleagues, collaborators and group members with sometimes different methods for raw data and processed data. For raw data sharing, respondents were using institutional data management or storage services (*n* = 2) or shared spaces on high performance computer (*n* = 1), CloudStor[102] (*n* = 1) and hard drives (*n* = 1) . Tools used for sharing processed data included GitHub[103] (*n* = 1) , Seurat object[104] (*n* = 1) or as a web page summary (*n* = 1) .

[88] Oxford Nanopore Technologies, https://nanoporetech.com/

[89] inDrop for high-throughput single-cell labelling, Illumina, Inc. https://sapac.illumina.com/science/sequencing-method-explorer/kits-and-arrays/indrop.html

[90] Hive scRNAseq Solution, Honeycomb Biotechnologies, https://perkinelmer-appliedgenomics.com/home/applications/hive-scrnaseq-solution/

[91] Hyperion Imaging System, Fluidigm, https://www.fluidigm.com/products-services/instruments/hyperion

[92] Gene Expression Omnibus (GEO), National Center for Biotechnology Information, https://www.ncbi.nlm.nih.gov/geo/

[93] Sequence Read Archive (SRA), National Center for Biotechnology Information, https://www.ncbi.nlm.nih.gov/sra

[94] Human Cell Atlas (HCA), https://www.humancellatlas.org/

[95] Sfaira, https://theislab.github.io/sfaira-portal/

[96] Human Protein Atlas, https://www.proteinatlas.org/

[97] Tabula Muris, https://tabula-muris.ds.czbiohub.org/

[98] Tabula Sapiens, https://tabula-sapiens-portal.ds.czbiohub.org/about

[99] ArrayExpress Archive of Functional Genomics Data, European Bioinformatics Institute, https://www.ebi.ac.uk/arrayexpress/

[100] R Markdown, https://github.com/rstudio/rmarkdown

[101] workflowr, https://workflowr.github.io/workflowr/

[102] CloudStor, https://www.aarnet.edu.au/cloudstor

[103] GitHub, https://github.com/

[104] Seurat object, https://satijalab.org/seurat/.html

Data sharing challenges faced by respondents included difficulties when sharing with users who are not skilled in statistical or bioinformatic software. Two respondents from twelve stated that they were "*lacking an easy way to share data with non R savvy users in an interactive way*" and "*sharing interactive tools for looking at results is difficult*". One of these respondents stated that a hosted RStudio server[105], hosted Shiny application[106] or iSEE in Galaxy Australia[107] were solutions that could work to solve this challenge. Two other respondents working with human data stated that there were challenges with compliance and regulatory authority when sharing data with colleagues or collaborators and issues with storage space access controls.

All of the respondents (*n* = 10 respondents from 10) indicated they aim to make their datasets publicly available through established databases or other systems including GEO[108] , SRA[109], FigShare[110] and HCA[111].

3.3.2 Tools

Based on the survey, 23 software tools, pipelines or packages were identified as being used by respondents for their single cell and spatial 'omics analysis process. These are listed in Appendix 1 of this document.

There are three dominant single cell analysis environments used by researchers, they are:

1. Bioconductor in R Statistical Software[112] offers a repository of many bioinformatics analysis packages. Single cell packages in Bioconductor make use of the ***singleCellExperiment*** class;
2. Seurat in R Statistical Software[113] provides packages that are a one-stop shop for most common single cell analysis tasks. These use the ***Seurat*** class; and,
3. Scanpy implemented in Python[114], is a toolkit for single cell analysis. Uses the ***anndata*** class and provides a large ecosystem of tools that integrate with Scanpy.

From the respondents, only three tools were used by more than 30% of respondents The top tool used by 60% of respondents (n=9) is the scRNAseq package Seurat[115]. The next most used tool (used by n = 6 respondents) is the pipeline Cell Ranger[116] which is a free software produced by the 10x Genomics company that processes 10x Genomics Chromium single cell data. The third most used tool is Scanpy[117] (used by n = 5 respondents). Two

[105] RStudio server, https://support--rstudio-com.netlify.app/products/rstudio/
[106] Shiny application, https://shiny.rstudio.com/
[107] iSee in Galaxy, https://github.com/neoformit/isee-docker-galaxy
[108] Gene Expression Omnibus, https://www.ncbi.nlm.nih.gov/geo/
[109] Sequence Read Archive, SRA, https://www.ncbi.nlm.nih.gov/sra
[110] FigShare, https://figshare.com/
[111] Human Cell Atlas, HCA, https://www.humancellatlas.org/
[112] R Statistical Software, https://www.r-project.org/
[113] R Statistical Software, https://www.r-project.org/
[114] Python programming language, https://www.python.org/
[115] Seurat, https://satijalab.org/seurat/
[116] Cell Ranger, https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger
[117] Scanpy, https://scanpy.readthedocs.io/en/stable/

other tools were used by two respondents, they are Harmony[118] and Velocyto[119], and the remaining 18 tools or pipelines (Monocle[120], anndata[121], Scran[122], Salmon[123], Giotto[124], STARsolo[125], ShellCheck[126], scvi-tools[127], ArchR[128], GeoMx[129], InferCNV[130], kallisto[131], NATMI [132], Tidyverse[133], PAGA[134], LoomPy[135], zUMIs[136], Optimus[137]) were identified by one respondent only. However, in conversation with researchers, it was communicated that often all or multiple tools in multiple environments are used depending on the questions being asked from the data[138]. Most tools used by the respondents in the survey were written in python language or for use in the R Statistical Software. As one respondent mentioned, a researchers skill set will drive the tool or workflow use because most tools are available in either python, R or can be used via the command line but rarely are the tools available in all three languages or environments[139].

Challenges identified by respondents in the set-up or use of these tools included requirements by some tools to need “a lot of RAM” (*n* = 3 from 7 respondents) and issues regarding software installations on available computational environments. One respondent said that conda[140] helps with these installation issues. Another respondent stated that R versions are difficult because they are released every 6 months but tools or packages are tied to specific R versions so having R version flexibility would be desirable. One respondent mentioned, that the R version challenges and requirement are a particular challenge with new tools that are released[141]. One respondent stated that the combining of python and R functions assists in getting around some of the memory requirement roadblocks.

[118] Harmony, https://portals.broadinstitute.org/harmony/
[119] Velocyto, http://velocyto.org/
[120] Monocle, http://cole-trapnell-lab.github.io/monocle-release/
[121] anndata, https://anndata.readthedocs.io/en/latest/
[122] SCRAN: Single Cell RNA-seq Analysis, https://github.com/elswob/SCRAN
[123] Salmon, https://combine-lab.github.io/salmon/
[124] Giotto, https://rubd.github.io/Giotto_site/
[125] STARsolo, https://github.com/alexdobin/STAR/blob/master/docs/STARsolo.md
[126] ShellCheck, https://github.com/koalaman/shellcheck
[127] scvi-tools, https://scvi-tools.org/
[128] ArchR, https://www.archrproject.com/
[129] GeoMx, https://nanostring.com/products/geomx-digital-spatial-profiler/geomx-data-center/
[130] InferCNV, https://github.com/broadinstitute/infercnv
[131] kallisto, https://pachterlab.github.io/k allisto/about
[132] NATMI: Network Analysis Toolkit for the Multicellular Interactions, https://github.com/forrest-lab/NATMI
[133] Tidyverse, https://www.tidyverse.org/
[134] PAGA: Partition-based Graph Abstraction, https://github.com/theislab/paga
[135] loompy, https://linnarssonlab.org/loompy/
[136] zUMIs, https://github.com/sdparekh/zUMIs
[137] Optimus, https://broadinstitute.github.io/warp/docs/Pipelines/Optimus_Pipeline/README/
[138] From the respondent: “*Often we use all these tools. Cell/space-ranger does the initial mapping. Filtering/cluster/DE can be performed in seurat/monocle/harmony. Depending on the questions we could use all 3 to do specific steps.*”
[139] From the respondent: ”*These tools are split across command line, python and R. A few tools have been dually developed in both python and R but not all. So someone who knows only R is limited to just the R tools unless they learn a new programming language.*”
[140] Conda https://docs.conda.io/en/latest/
[141] From the respondent: “*This is for new packages that are released. Since this field is new the analysis methods are still being developed. Therefore the brand new method/tool which is developed will be tied to the latest R release.*”

3.3.3 Compute infrastructure

3.3.3.1 Types used

Survey respondents (*n* = 15) currently use a variety of computational infrastructure for their analyses. Most respondents used high-performance computing provided by their host institute (*n* = 11), their laptops or personal computers (*n* = 8). Fewer respondents (*n* = 3) are using shared high-performance computers managed by national (e.g. NCI[142], Pawsey[143]) or state (e.g. QCIF/QRIScloud[144]) computing centres. Only one respondent accessed commercial cloud resources, such as Amazon Web Services (AWS)[145], Google Cloud[146] or Microsoft Azure[147].

The majority of respondents (93% (n = 14) stated that they had access to necessary expertise to maintain computational infrastructure, such as installing and updating software, either within their institution or group but also from collaborators and partner institutes. Only one respondent stated that they were unable to access the expertise needed to build and maintain their computational infrastructure.

3.3.3.2. Resourcing

Just over half of the respondents to this question (*n* = 6) said the infrastructure they currently had access to was not sufficient for their current single cell and spatial 'omics work, due to limitations in available memory, instability, lack of an institutional solution, data storage allocations, limited access (due to busy clusters) or inconvenience of moving and keeping track of data and code between different computational environments. One respondent stated that there are too many users on the system they have access to, and another stated competition for compute nodes made access/availability insufficient for their needs. Another respondent stated that the added inconvenience of moving and keeping track of data and code between their local (interactive) environment and a HPC environment means there is limited time to pursue alternative analyses[148]. One respondent stated that there was no suitable institutional solution available to them.

Just over half of the respondents (56%) stated that their current computational infrastructure setup would not be sufficient for single cell omics research in two years time and 73% of respondents thought their current computational infrastructure would not be sufficient to support their research in five years time.

[142] NCI, National Computational Infrastructure, nci.org.au/
[143] Pawsey Supercomputing Centre, pawsey.org.au/
[144] Queensland Cyber Infrastructure Foundation, QCIF
[145] Amazon Web Services
[146] Google Cloud
[147] Microsoft Azure
[148] From the respondent: *“The added inconvenience between moving (and keeping track of!) data and code between local (interactive) and hpc (large mem jobs) means I haven't had time to try out a promising differential expression approach yet (MAST modelling bio replicates and individual cells!).*”

3.4 Challenges being faced

In addition to the specific examples listed in Section 3.3, a variety of broad limitations/roadblocks/challenges/issues with current infrastructure were identified by the survey respondents, SIG and others.

3.4.1 Data sharing and data movement challenges

- Sharing large amounts of raw single cell data between different computational environments is a challenge because the data is often large. For example, an S4 Flow Cell from a NovaSeq[149] run of paired reads (2 x 150 base pair long) is equivalent to 2,190 GB of data. Four respondents stated that this was difficult to transfer from the instrument to analysis or storage environments and used options such as email, CloudStor, shared HPC spaces and internal institutional data storage / management solutions or even relying on posting/couriering hard drives to transfer raw data to colleagues or collaborators. Data movement of large raw datasets was further complicated by the need to move the data around to access appropriate computational environments for analyses and security enforced by some host institutions and the available data movement tools "*did not make data movement easy*";

- Sharing processed data for interpretation with life science research collaborators who have limited bioinformatics/IT skills was a challenge raised by a number of respondents (n = 4). In particular, sharing analysed data as interactive plots and results, including visualisations and cell clustering with collaborators was desirable and "*being able to provide interpretable analysed data between labs is difficult*". However, one respondent suggested a hosted RStudio or RShiny server or GitHub repositories would serve as excellent solutions to these challenges.

- Additionally, some respondents also had special needs with regards to ensuring security and controlled access to data when sharing data derived from human subjects. One member of the SIG identified that they required better sharing mechanisms that align with compliance and regulatory authority requirements when sharing human data with collaborators.

3.4.1 Computational resourcing and environment challenges

- Many researchers in this space rely on software and tools that are written in the R statistical software program and the accompanying R software package manager, Bioconductor[150]. Computational resources that are available (even across a variety of infrastructures) can be insufficient with requirements, particularly in regards to available RAM in use with the statistical software R. Many respondents (n = 13) stated they would like access to R and BioConductor with greater computational memory resources for

[149] NovaSeq instrument, Illumina, Inc. https://sapac.support.illumina.com/bulletins/2018/01/approximate-sizes-of-sequencing-run-output-folders.html

[150] Bioconductor, https://www.bioconductor.org/

analysing large datasets[151]. However, there are challenges with accessing the right environment for R which updates versions every several months and tools often follow or remain usable in a former version of the R software. This can cause challenges when needing to use new tools and maintaining environments with the appropriate version for analysis needs[152]. Several SIG members raised the issue of challenges when using R in an environment with greater memory resources such as HPC, because the interactivity is lost making it less desirable and convenient. Installation of tools becomes a challenge also. In addition R does not scale well for large datasets and this often requires workarounds that involve rewriting code in python language combined with R to run analyses[153].

- Some respondents (n=3) stated that their current compute infrastructure is insufficient as it has "*too many users*", "*[too much] competition for computing nodes*" and "*storage space limitations*". Although respondents stated that access to high memory resources is required for short periods of time or is "*bursty*", moving data around to access different types of resources on different pieces of infrastructure to perform particular steps in a workflow is annoying and time-consuming[154].

3.4.2 Tool and pipeline related challenges

- One respondent in the survey remarked that there are "*literally thousands of tools available*" (see also https://www.scrna-tools.org/). Most respondents stated that their pipelines require flexibility and each analysis is fluid[155], custom[156,157] or bespoke[158].

- Tools are lacking that appropriately facilitate data integrations. Data integration is common practice in single cell and spatial omics analyses whereby different datasets are pulled together or 'integrated' to identify shared cell states that are present across different datasets. Data integration is essential in recognising patterns across different

[151] From the respondent: "*We have issues with compute for large data sets in bioconductor*"

[152] From the respondent: "*R versions are a real struggle- bioconductor releases every 6 months often contain cool new tools, but are tied to R versions. Some methods need a lot of RAM, which is fine on HPC (once you've got your R version setup), but need to transfer data around and know exactly what you want to run in a script.*"

[153] From the respondent: "*We have issues with compute for large data sets in bioconductor. [To overcome] memory requirements, [we] combine python and R functions*".

[154] From the respondent: "*Yes it [the fraction of my time spent computing datasets] is quite bursty, especially the need for high memory, for example, but it tends to be annoying to move on and off and so it [the community cloud institutional resource] gets tied up more than it should be but yes it is quite bursty.*"

[155] From the respondent: "*The analysis that I perform is very fluid and is not suited to a single pipeline. It uses various packages/pipelines (parts).* "

[156] From the respondent: "*Since single-cell analyses are slightly different from standard bulk RNA seq or microarray / proteomics analyses. Available pipelines (web-based) can perform standard analyses (clustering, visualisation, generic cell-types information). However, [in] my experience many hypotheses are developed and then tested for each experiment that makes researchers to process the data locally instead of online servers. Probably a strategy needs to be developed so that researchers could do a drag-drop approach and make their own analysis plan using available tools to keep testing their hypothesis*."

[157] From the respondent: "*We customise often, trialling/benchmarking new tools and incorporating as needed*".

[158] From the respondent: "*Freedom to use whatever tools and bespoke pipelines I want is essential [in considering the use of a shared single cells analysis platform]*".

regions, therefore provenance tracking is a major challenge that impacts the data reuse and interpretation[159].

3.4.3. Other challenges

- Research computing IT support for researchers has diminished at some institutions, putting greater pressure on researchers to manage their own software and tool installations on institutionally supplied shared computational systems[160]. Because single cell and spatial 'omics analysis workflows use different resources and different hardware for each step in the workflow, are custom designed for each analysis and there are significant number of tools available, containers were identified by the SIG as being key to working efficiently[161];

- Having access to standardised datasets, such as a complex mixture of cells with known proportions, to allow benchmarking of new tools and instruments is necessary with the current rate of new tools and instruments[162]. This proves challenging because the generation of standardised reference datasets are not generally fundable activities through Australia’s national life science research granting bodies (such as ARC and NHMRC)[163];

- To allow scalability, R packages can require rewriting in languages that allow for scaling up with large datasets, such as python. There is a strong desire in the community to be supported in learning or using python language to convert current tools to work with large datasets or other methods for scaling up;

[159] From the respondent: “*For us, it is not about sharing analysed data but also how we integrate data to see reproducible patterns across different cell types. The provenance tracking is the major challenge so we can understand the quality of clustering or understand patterns more broadly. It can be hard to understand the quality of the different datasets you are trying to integrate. There are some software like Harmony or Seurat but they don’t carry over the data labels, they just give new ones. For example: when wanting to integrate data from multiple labs or experiments with different parameters of intestinal location or disease state etc. We want to bring together these data to see what variables change the behaviour of the cell clustering or proportions. It is important to be able to understand how the original study found the clusters and then labelled them and then track where they end up. This is a major challenge and impacts the ability to use the existing study in a meaningful way*”.

[160] From the respondent: “There has been a change at [our institution] and maybe in the field, we used to have eResearchers or a team of specialists install software for people but now that is gone and now it is very much up to users because we have tools like Conda now and containerisation. So we use that”.

[161] From the respondent: “*Because single cell workflows use very different resources for different steps, we need to use different hardware. In my lab we containerise everything we do so we can easily move our compute environment between whatever hardware we get access to and containers are really a key in this area and very, very useful.*”

[162] From the respondent: “*Looking at complex mixtures of cells or looking at mixtures of cells whose proportions vary across a spatial slide, these are much harder to benchmark. So the quality of our benchmarking material needs to evolve rapidly too. Of course people aren’t resourced to do that, it is not a fundable activity within NHMRC ideas or ARC discovery project.*”

[163] From the respondent: “*I’ll second that we are doing benchmarking both on the wet and the dry side and we are fortunate enough to have some support from philanthropic funding but would never get this funded via NHMRC or ARC. Some go through to publication but most of it never gets published. It is useful to know whether that tool works or not but unless it is a major advance you are not going to get a publication or bother to try and get to publication. If this is centrally funded this would be very useful.*”

- One respondent stated that they did not have access to expertise to update, build or maintain current computational resources[164] or perform activities such as containerisation, and;
- Training and skills that are targeted to a particular area of study in a research area are hard to find. Mechanisms to share and communicate bioinformatic workflows and methods used by researchers and practitioners in the context of scientific drivers would be valued by the community[165].

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

80% of respondents (n = 12) agreed that if a shared data collaboration/analysis platform for single cell and spatial 'omics 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) quick installation of additional tools/pipelines on request; (2) good documentation on platform use; (3) an ability to transfer data easily to/from storage; (4); access to preferred tools/pipelines; and (5) security of data and analysis; and (6) good documentation on how to use the tools and pipelines.

[164] From the respondent: “*No, we would like to set some up or update our current approach [to build and maintain computational infrastructure] but can't access expertise. Some is supported by [our infrastructure provider], but specific application are via my team or collaborators*”.

[165] From the respondent: “*My general approach is that base level training is useful to people to get moving but after that it explodes and having training in anything niche/targeted is not that beneficial to enough people. Perhaps more useful is a kind of ‘show and tell’ - this is the type of analysis I did. So we have plenty of talks about the science and where it gets to, but not enough on the analysis and what I did and how it worked for me. So a little bit more to help people go onto the methods / bioinformatics. This type of training would be the type of training to go from medium level up to quite advanced*”.

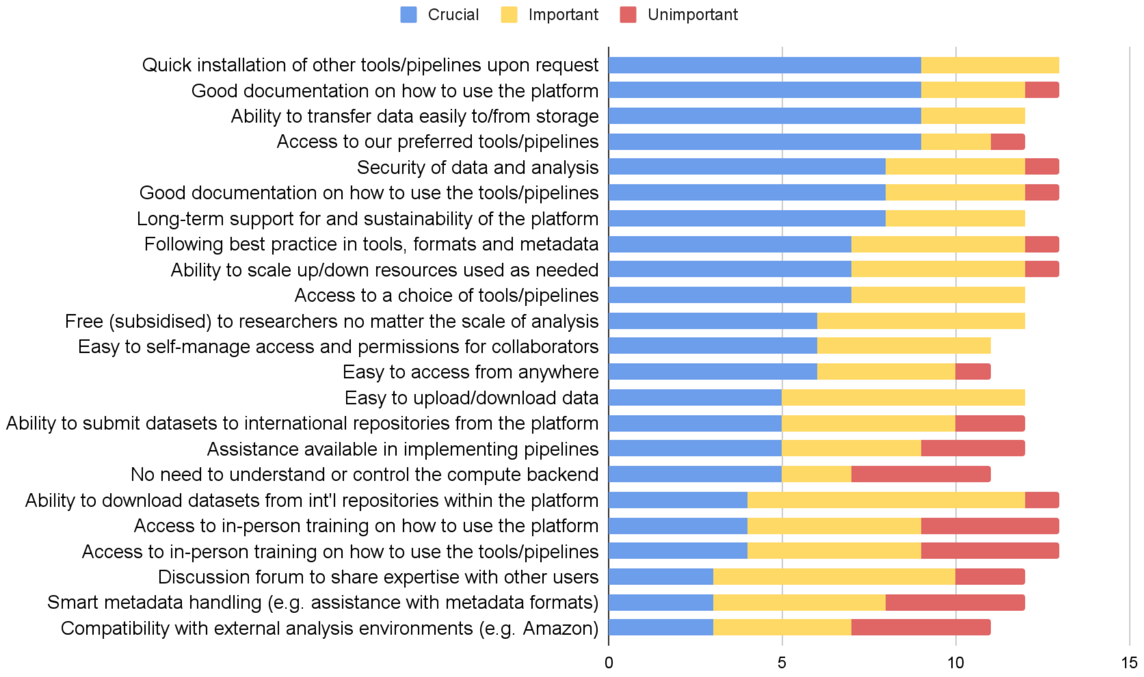


Figure 3: Desired features in a shared single cell and spatial 'omics analysis platform

Survey respondents were asked about which features they considered to be ‘crucial’, ‘important’ or ‘unimportant’ in a shared single cell and spatial 'omics 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 Single Cell and Spatial ‘Omics Analysis Infrastructure

4.1 Goal

The Australian BioCommons aims to develop a ‘single cell and spatial 'omics Infrastructure Roadmap for Australia’ that describes collaborative infrastructure, which, when implemented (from Q3 2023 onwards), will help enable Australian researchers from a wide range of institutions to perform high-quality single cell and spatial 'omics and spatial omics analysis work who would otherwise be unable to do so because of data-, expertise-, software- or compute-related infrastructure roadblocks.

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

- V1 - Content-based on SIG survey results and input from SIG meeting - December 2022.
- V2 - Content modified to incorporate feedback from various national computational infrastructure providers - May 2023
- V3 - Content modified to incorporate feedback from SIG, other researchers undertaking single cell and spatial 'omics analysis, and international groups - June 2023.
- V4 - Content modified to incorporate final feedback from SIG - July 2023.

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 single cell and spatial 'omics analyses to be performed;
    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 single cell and spatial ‘omic data; and,
3. To make it easier to publish high-quality single cell and spatial ‘omic-associated data files in accordance with best-practice open science guidelines.

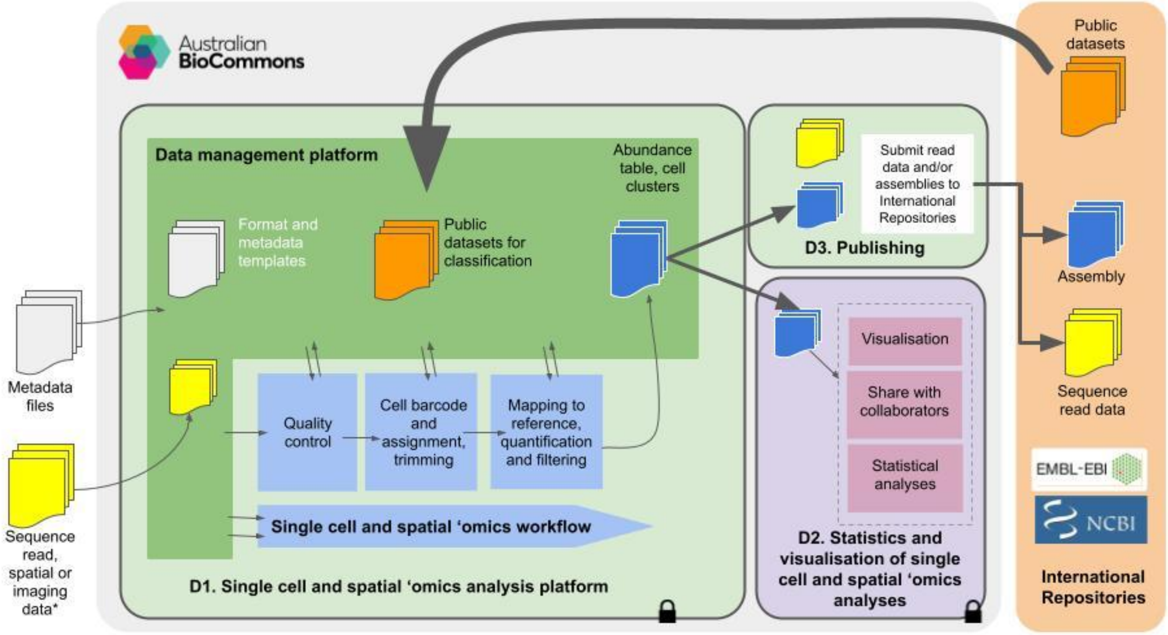
4.3 Outputs

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

D1. A platform for performing single cell and spatial ‘omics analysis

D2. Systems to enable visualisation and statistical analyses of single cell and spatial 'omics data and their data products

D3. Systems to enable sharing and submission of single cell and spatial ‘omics data and associated output files from Australia to appropriate global repositories


Australian BioCommons
Data management platform
Format and metadata templates
Public datasets for classification
Abundance table, cell clusters
Metadata files
Sequence read, spatial or imaging data*
Quality control
Cell barcode and assignment, trimming
Mapping to reference, quantification and filtering
Single cell and spatial 'omics workflow
D1. Single cell and spatial ‘omics analysis platform
Submit read data and/or assemblies to International Repositories
D3. Publishing
Visualisation
Share with collaborators
Statistical analyses
D2. Statistics and visualisation of single cell and spatial ‘omics analyses
Public datasets
Assembly
Sequence read data
EMBL-EBI
NCBI
International Repositories

*User-generated; facility-generated; public data downloaded from repository; shared via a collaborator.

Figure 4. Schematic diagram showing the proposed infrastructure to support single cell and spatial 'omics analyses, and data flow

D1 - Sequence reads, assembled data or other relevant data (either locally sourced, or obtained from various public nucleic sequence repositories) are inputs into the platform for single cell and spatial 'omics Analysis which provides a command-line interface (CLI) or graphical user interface (GUI) access to a community requested selection of tools and workflows for performing single cell and spatial ‘omics analyses (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 datasets and their data products; and **D3** - systems to enable submission/publishing of single cell and spatial ‘omics 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. A high resolution image is here.

D1 - A platform for single cell and spatial 'omics 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 single cell and spatial ‘omics analyses to be performed), it is proposed to implement a platform in Australia [166], that:

A. Includes a set of key tools[167] and/or pipelines for various types of single cell and spatial ‘omics analyses that are requested by the local research community, or chosen by the end-user for the dataset in question (including for example, tools for quality control, cell barcode and assignment, trimming, mapping to reference, quantification and filtering etc):

   a. Installed (plus all other dependencies), with the ability to self-install relevant tools, and optimised on a command line interface (CLI) analysis environment (i.e. across a variety of Tier 1 and 2[168] shared computational infrastructures) underpinned by appropriate computational resources[169]. The platform would include access to some expertise that can provide support for tool and workflow installation and optimisation on relevant computational resources;

   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[170]), 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 as well as various cell atlases), institutional or other data storage, and with the ability to upload/mount user-generated or other datasets[171] that are required as inputs for a single cell and spatial 'omics analysis pipeline.;

D. Has appropriate user authorisation and sharing mechanisms to allow for data sharing, solely at the discretion of a data owner/custodian;

[166] 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.
[167] e.g. a selection of the tools listed in Appendix 1 or described here https://www.scrna-tools.org/
[168] 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.
[169] Including necessary high memory nodes (>1TB RAM). See also biocommons.org.au/pathfinder-biocloud
[170] Galaxy Australia
[171] Including necessary datasets for cell clustering and data integration.

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[172];

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

H. Includes training[173] for all the above.

D2. Systems to enable statistical analyses and visualisations of single cell and spatial 'omics analysis results:

To address objective 2 (*i.e. to make it easier for Australian researchers to perform statistical and visualisation analyses of single cell and spatial 'omics data*), it is proposed to implement:

A. Hosted frameworks to enable researchers to utilise common packages in relevant statistical software, such as R statistical software[174] and python[175] language, for interactive statistical analysis, and exploration of single cell and spatial 'omics datasets. In particular, frameworks will include the ability for single cell and spatial omics datasets to be viewed with an interactive component that includes the required image analysis or segmentation and visualisation;

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 in how to use such systems to support research.

[172] 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

[173] 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.

[174] R statistical software: https://www.r-project.org/ can be used in a number of hosted formats, including via Shiny, a web-based environment to easily access and share: https://www.rstudio.com/products/shiny/

[175] Python: https://www.python.org/

D3 - Systems to enable sharing and submission of single cell and spatial 'omics 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 single cell and spatial 'omics datasets 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 sharing interactive single cell and spatial 'omics' result or dataset (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 from 2023.

As of March 2023, several key activities that are relevant to the proposed infrastructure are already underway:

Component	Planned dates for delivery	Notes
D1-Aa/Ab. Key tools[176] and/or pipelines for data preparation, quality control, tools for quality control, cell barcode and assignment, trimming, mapping to reference, quantification and filtering single cell and spatial 'omic 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, 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[177] shared computational infrastructures)	Ongoing	As of March 2023, four of the tools listed in Appendix 1 (Cell Ranger, R, Salmon, Kallisto) are installed as modules on QRIScloud/UQ-RCC HPC machines (Bunya[178], Wiener[179]); and three of the tools (Cell Ranger, R, Salmon) are installed as modules at NCI. Installation of further tools as modules across NCI, Pawsey, and QRIScloud/UQ-RCC infrastructures to single cell and spatial 'omics and other single cell and spatial 'omics 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).

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

[177] 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.

[178] Bunya high performance computer.

[179] Wiener high performance computer.

D1-Aa. CLI platform appropriately resourced for performing single cell and spatial 'omics analyses	Ongoing	BioCommons partner infrastructures at NCI, Pawsey, and QCIF include computational systems that are capable of performing any part of single 'omics analysis. This includes Wiener at QCIF/UQ with Graphics Processing Units (GPUs) for high throughout imaging data. 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 has been 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[180]), underpinned by appropriate computational resources; and,	Ongoing	As of March 2023, seven of the tools listed in Appendix 1 (CellxG, Seurat[181], Salmon, ScanPy, Kallisto, Salmon-Kallisto, anndata) are installed on Galaxy Australia. At the time of writing, there are over 1,800 tools installed on Galaxy Australia, and more than 7,000 more tools available on the Galaxy Tool Shed. 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 single cell and spatial 'omics analyses	Ongoing	Since 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. It also accesses Microsoft Azure commercial cloud compute to resource certain tools (e.g. AlphaFold) As of July 2022, Galaxy Australia moved its head node and associated servers to Australia's Academic and Research Network (AARNet) to allow for increased on-demand support for users.
D1-Ac. Key tools available as high quality, trusted software	Ongoing	As of March 2023, seven of the tools listed in Appendix 1 (CellxG, Cel-Seq2, R, Seurat, Salmon,

[180] Galaxy Australia

[181] Please note that the current Seurat tool in Galaxy Australia offers a limited number of functions when compared with the complete Seurat package of commands that are available in R, https://satijalab.org/seurat/reference/

containers for self-deployment on institutional or independent computational infrastructures.		Kallisto, ScanPy) are available as containers[182] (either Bioconda, Docker or Singularity).
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 system that is tightly linked to the single cell and spatial 'omics Platform	Ongoing	Considerations for what may be the best technical solution are ongoing. See Requirements of a Data Management Component of the Australian BioCommons
D1-G. Tool and software workflow documentation with community-contributed content.	Ongoing	Tool and workflow documentation and discovery are available via the BioCommons ToolFinder and BioCommons WorkFlowHub, respectively. These Services aim to cover all research communities to provide current access and information to tools and workflows across communities. Support to researchers in supplying access to the workflow manager, Nextflow, through a pilot exploration of a National Netflow Tower Service. Nextflow will allow researchers to launch bioinformatics pipelines on resources of their choice, including on-premises, supercomputing and commercial cloud platforms.
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, BioCommons Training Program and Galaxy Training Program. Community-contributed content is accessible and open to contribution at the Australian BioCommons' How-to-Guides resource.
D1-H. Training for all aspects of the Single Cell and spatial	Ongoing	Introductory level training has occurred for a number of relevant skills, including, software

[182] See BioContainers, https://biocontainers.pro/registry

'Omics Platform, Statistics and Visualisation Platform.		containerisation[183]; getting started with command line[184], Galaxy Australia[185] and R[186] 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. The Galaxy Training Network offers free follow along tutorials with access to the Galaxy platform with 19 tutorials in single cell and spatial 'omics workflows and tools. Further training material for single cell and spatial 'omics can be located through international bioinformatics training repositories such as TeSS, GOBLET or the DReSA (Digital Research Skills Australasia) platform

[183] Containers in Bioinformatics Workshop: Containerising a Pipeline, Building Containers and BYO Pipeline, delivered by Dr Marco de la Pierre

[184] Webinar: Getting Started with Command Line Bioinformatics, delivered by Parice Brandies

[185] Webinar: Galaxy Australia: a Strengthened National Life Science Platform that Engages Globally, delivered by Dr Gareth Price

[186] Webinar: Getting Started with R, delivered by Dr Saskia Freytag

Appendix 1

Table 1. Single cell and spatial 'omics analysis tools for consideration for inclusion in a shared analysis environment.

Note that a single cell and spatial 'omics analysis protocol may also incorporate many other software tools not listed here. In reality, any tool that is used in the bioinformatic and statistical data analysis process has the potential to be installed on available infrastructures[187].

Workflow Step	High-level component	Tool	Brief description	Link to data/software or article
	Analysis Pipeline, Workflow	Cell Ranger	Cell Ranger is a set of analysis pipelines that process Chromium single-cell data to align reads, generate feature-barcode matrices, perform clustering and other secondary analysis, and more.	https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger
	Analysis Pipeline, Workflow	GeoMX	The GeoMx DSP workflow seamlessly integrates with current histology methods to get you robust and reproducible spatial omics data quickly.	https://www.nanostring.com/products/geomx-digital-spatial-profiler/geomx-dsp-overview/
	Analysis Platform	R	R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.	https://www.r-project.org/
	Data Management	Rmarkdown	Turn your analyses into high quality documents, reports, presentations and dashboards with R Markdown.	https://rmarkdown.rstudio.com/
	Database/ data repository	Gene Expression Omnibus (GEO)	GEO is a public functional genomics data repository supporting MIAME-compliant data submissions.	https://www.ncbi.nlm.nih.gov/geo/
	Database/ data repository	Sequence Read Archive (SRA)	Sequence Read Archive (SRA) data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data.	https://www.ncbi.nlm.nih.gov/sra
	Database/ data repository	Human Cell Atlas (HCA)	The Human Cell Atlas is an international collaborative consortium that charts the cell types in the healthy body, across time from development to adulthood, and eventually to old age.	https://www.humancellatlas.org/
	Database/ data repository	European Nucleotide Archive (ENA)	The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation.	https://www.ebi.ac.uk/ena/browser/home
	Database/ data repository	Sfaira	a single-cell data zoo for public data sets paired with a model zoo for executable pre-trained models. a versatile repository that serves pre-trained scRNA-seq models.	https://github.com/theislab/sfaira

[187] For an extensive list of tools related to scRNAseq analyses, see https://www.scrna-tools.org/.

	Database/ data repository	Human Protein Atlas	The Human Protein Atlas is a Swedish-based program initiated in 2003 with the aim to map all the human proteins	https://www.proteinatlas.org/
	Database/ data repository	Tabula Muris	Tabula Muris is a compendium of single cell transcriptome data from the model organism Mus musculus.	https://tabula-muris.ds.czbiohub.org/
	Database/ data repository	Tabula Sapiens	Tabula Sapiens is a benchmark, first-draft human cell atlas of nearly 500,000 cells from 24 organs of 15 normal human subjects.	https://tabula-sapiens-portal.ds.czbiohub.org/about
	Database/ data repository	Array Express	ArrayExpress Archive of Functional Genomics Data	https://www.ebi.ac.uk/arrayexpress/
	Programming Language	Bash	Bash is a Unix shell and command language written by Brian Fox for the GNU Project as a free software replacement for the Bourne shell.	https://www.gnu.org/software/bash/
	Programming Language	Python	Python is a programming language that lets you work more quickly and integrate your systems more effectively.	https://www.python.org/
	Software Manager	Conda	Conda is an open source package management system and environment management system that runs on Windows, macOS and Linux.	https://conda.io/
	Software Manager	Bioconductor	The mission of the Bioconductor project is to develop, support, and disseminate free open source software that facilitates rigorous and reproducible analysis of data from current and emerging biological assays.	https://www.bioconductor.org/
	Tool, Toolkit, Software Package/s	zUMI	zUMIs is a fast and flexible pipeline to process RNA-seq data with (or without) UMIs. The input to this pipeline is simply fastq files.	https://github.com/sdparekh/zUMIs
	Tool, Toolkit, Software Package/s	Seurat	Seurat is an R package designed for QC, analysis, and exploration of single-cell RNA-seq data.	https://satijalab.org/seurat/
	Tool, Toolkit, Software Package/s	Scanpy	Scanpy is a scalable toolkit for analyzing single-cell gene expression data built jointly with anndata. I	https://scanpy.readthedocs.io/en/stable/
	Tool, Toolkit, Software Package/s	Harmony	Harmony is an algorithm for performing integration of single cell genomics datasets.	https://portals.broadinstitute.org/harmony/articles/quickstart.html
	Tool, Toolkit, Software Package/s	Velcyto	velocyto (velox + κύτος, quick cell) is a package for the analysis of expression dynamics in single cell RNA seq data.	http://velocyto.org/
	Tool, Toolkit, Software Package/s	Monocle	An analysis toolkit for single-cell RNA-seq.	http://cole-trapnell-lab.github.io/monocle-release/docs/#installing-monocle

	Tool, Toolkit, Software Package/s	anndata	anndata is a Python package for handling annotated data matrices in memory and on disk, positioned between pandas and xarray.	https://anndata.readthedocs.io/en/latest/
	Tool, Toolkit, Software Package/s	scran	Implements miscellaneous functions for interpretation of single-cell RNA-seq data.	https://bioconductor.org/packages/release/bioc/html/scran.html
	Tool, Toolkit, Software Package/s	Salmon	Salmon: Fast, accurate and bias-aware transcript quantification from RNA-seq data	https://combine-lab.github.io/salmon/getting_started/
	Tool, Toolkit, Software Package/s	giotto	Giotto, a comprehensive and open-source toolbox for spatial data analysis and visualization.	https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02286-2
	Tool, Toolkit, Software Package/s	STARsolo	STARsolo is a turnkey solution for analyzing droplet single cell RNA sequencing data (e.g. 10X Genomics Chromium System) built directly into STAR code.	https://github.com/alexdobin/STAR/blob/master/docs/STARsolo.md
	Tool, Toolkit, Software Package/s	ShellCheck	ShellCheck is a GPLv3 tool that gives warnings and suggestions for bash/sh shell scripts:	https://github.com/koalaman/shellcheck
	Tool, Toolkit, Software Package/s	scvi-tools	scvi-tools accelerates data analysis and model development, powered by PyTorch and AnnData	https://scvi-tools.org/
	Tool, Toolkit, Software Package/s	ArchR	ArchR is a full-featured R package for processing and analyzing single-cell ATAC-seq data.	https://www.archrproject.com/
	Tool, Toolkit, Software Package/s	InferCNV	InferCNV is used to explore tumor single cell RNA-Seq data to identify evidence for large-scale chromosomal copy number variations, such as gains or deletions of entire chromosomes or large segments of chromosomes.	https://github.com/broadinstitute/infercnv
	Tool, Toolkit, Software Package/s	Kallisto	kallisto is a program for quantifying abundances of transcripts from bulk and single-cell RNA-Seq data.	https://pachterlab.github.io/kallisto/about
	Tool, Toolkit, Software Package/s	NATMI	NATMI (Network Analysis Toolkit for Multicellular Interactions).	https://github.com/asrhou/NATMI
	Tool, Toolkit, Software Package/s	Tidyverse	The tidyverse is an opinionated collection of R packages designed for data science.	https://www.tidyverse.org/
	Tool, Toolkit, Software Package/s	PAGA	Single-cell RNA-seq quantifies biological heterogeneity across both discrete cell types and continuous cell transitions.	https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1663-x
	Tool, Toolkit, Software Package/s	Loompy	Loom is an efficient file format for large omics datasets.	http://loompy.org/
	Workflow	Optimus	Optimus is an open-source, cloud-optimized pipeline developed by	https://broadinstitute.github.io/warp/docs/Pipelines/Optimus_Pipeline/READ

			the Data Coordination Platform (DCP) of the Human Cell Atlas (HCA) Project.	ME/
	Workflow Management	Snakemake	The Snakemake workflow management system.	https://snakemake.readthedocs.io/

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

Appendix 2

Survey[188] questions posed to the single cell and spatial 'omics research community

1. How would you describe your level of experience with single cell and spatial 'omics Analysis?

Tick all that apply.

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

2a. What types of single cell omics analyses do you / your group members perform, or envisage performing in the next 5 years? *

Tick all that apply.

- Genomics
- Transcriptomics
- Proteomics
- Metabolomics
- Epigenomics
- Multi-omics
- Chromatin structure/accessibility
- Spatial analyses
- Other:

2b. If you indicated Multi-omics above, what types of omics data will you be combining?

3. Which broad application area(s) do you work in, or will work in, during the next 5 years? *

Tick all that apply.

[188] single cell and spatial 'omics Poll/Survey Results

- Human health / medical
- Fundamental bio-research
- Environmental
- Agriculture
- Diagnostics
- Other:

4. Which data repositories (if any) do you access to support your single cell omics analyses? e.g. ENA, SRA, ArrayExpress, GEO, Human Cell Atlas, Metabolights, PRIDE, etc

5. Which tools / software / pipelines / programs / platforms do you or group members use for single cell omics analysis?

6a. Are there tools / software / pipelines / programs / platforms you'd like to use but that aren't fit for purpose for your study area?

6b. Why aren't the tools / software, etc listed in 6a. not suitable?

7a. 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)?

7b. What are the tools / software etc. listed in 7a and what are the roadblocks you've encountered?

7c. What is your workaround to the roadblocks listed in 7b and why is it inadequate?

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

9. What data generation platform/s or technologies are you currently using to generate data for single cell analysis? e.g. 10x/SmartSeq2/RNAScope etc

10. Do you make use of existing datasets (choose all that apply)?

Tick all that apply.

- Yes, public datasets
- Yes, private datasets (from my previous work or that of collaborators)
- No, because no relevant data exists

- 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 formator 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 single cell omics project(s)? If so, what?

12a. How do you share data within your group and with collaborators?

12b. What difficulties that you have encountered when sharing data within your group or with collaborators?

13a. Do you make your single cell omics datasets publicly available?

13b. Where do you make the datasets mentioned in 13a available?

13c. Have you encountered any difficulties in making datasets available?

14. If you don't make your single cell omics datasets publicly available, why not?

Tick all that apply.

- Commercial confidence issues
- I don't see a benefit in sharing my datasets publicly available
- I don't know how to make my datasets publicly available
- Other:

15. What kind of compute infrastructure setup do you use for single cell omics analysis (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 Australia
- 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 not, why not?

Tick all that apply.

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 single cell omics analysis at a similar scale in 2 years' time
- Yes, as we expect to be doing less single cell omics analysis in 2 years' time

- No, we expect to be doing more single cell omics analysis 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?

- Yes, we expect to be doing single cell omics analysis at a similar scale in 5 years' time
- Yes, as we expect to be doing less single cell omics analysis in 5 years' time
- No, we expect to be doing more single cell omics analysis 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 to perform single cell omics analysis?

Tick all that apply.

- Yes - we can't currently perform single cell omics analysis without such infrastructure
- Yes - our needs are currently met but we'd consider shared infrastructure if it was suitable
- No - we will always prefer to perform single cell omics analysis locally no matter how good a shared platform is

- Other:

21. How important (crucial, important, unimportant) are these general factors to you in a shared single cell omics analysis infrastructure?

- 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 (crucial, important, unimportant) are these data-related factors to you in a shared single cell omics analysis infrastructure?

- 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 (crucial, important, unimportant) are these tool/pipeline-related factors to you in a shared single cell omics analysis infrastructure?

- 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 single cell omics analysis infrastructure?

25. How important (crucial, important, unimportant) are these compute-related factors to you in a shared single cell omics analysis infrastructure?

- 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 (crucial, important, unimportant) are these training-related factors to you in a shared single cell omics analysis infrastructure?

- 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 single cell omics analysis infrastructure? If so, what?

Please let us know here.

Document Control

VERSION	DATE	AUTHOR(S)	DESCRIPTION
V1.0	5 August 2022	Tiffanie Nelson, Jeff Christiansen	A preliminary document detailing the outline of the roadmap draft including the software list obtained from researchers.
V2.0	6 March 2023	Tiffanie Nelson, Jeff Christiansen	Reviewed by Jeffrey Christiansen. Shared with Dominique Gorse.
V3.0	10 August 2023	Tiffanie Nelson, Jeff Christiansen	Reviewed by Sarah Williams and Nicholas Matigan.
V4.0	16 November 2023	Tiffanie Nelson, Luciano Martelotto, Eija Korpelainen, Kevin Dudley	Invitation to review sent to more than 10 national and international single cell and spatial omics researchers. Reviewed in substantial detail by the researchers mentioned.
V4.1	13 December 2023	Tiffanie Nelson	Minor update to list DOI and Creative Commons licensing on the front of the document.