THE AUSTRALIAN BIOCOMMONS

an exemplar of international engagement in research infrastructure

Jeff Christiansen, Paul Coddington, Brian Davis, Ian Duncan, Rhys Francis, Christina Hall, Carina Kemp, Steven Manos, Tiffanie Nelson, Sarah Nisbet, Gareth Price, Frankie Stevens, Andrew Lonie
Thanks to rapid advances in sensing technologies...

DNA  mRNA  proteins  metabolites
...life sciences have become data-intensive

160PB
10PB in 2012

What do life-scientists need?

Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators

Barone et al (2017)
PLoS Computational Biology
doi.org/10.1371/journal.pcbi.1005755
There is a global wave of investment
We have an opportunity in Australia

<table>
<thead>
<tr>
<th>Facilities for the Future Underpinning Australia’s Research and Innovation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Digital data &amp; eResearch platforms</strong></td>
</tr>
<tr>
<td>Funding will ensure a regular upgrade cycle for maintaining Australia’s Tier 1 HPC capability, across both the Pawsley Supercomputing Centre (Pawsey) and the National Computational Infrastructure. In addition, a scoping study is under way to explore the integration of the Geoscience Australia core and deep-sea drilling platforms.</td>
</tr>
<tr>
<td><strong>Complex Biology</strong></td>
</tr>
<tr>
<td>Investments enable upgrades to next generation sequencing, mass spectrometry, cryo-storage, phenotyping and virtual laboratories that will drive new medical treatments, reduced future medical costs and facilitate emerging opportunities in biomedicine, medical technology, agribusiness and environmental conservation.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2017-18 to 2021-22**</th>
<th>Total to 2028-29</th>
</tr>
</thead>
<tbody>
<tr>
<td>eResearch</td>
<td>$911M</td>
</tr>
<tr>
<td>Complex Biology</td>
<td>$216M</td>
</tr>
</tbody>
</table>

In addition to existing operating

---

**Case Study**

Bioplatforms Australia (BPA) enables world class research in the life sciences by investing in state-of-the-art infrastructure and associated expertise in the fields of genomics, proteomics, metabolomics and bioinformatics. For example, genomics is the study of genomes—the entirety of our DNA, and was a major focus of the 2030 Strategic Plan. It noted that genomics and precision medicine will play an increasing role in improving health outcomes. Through next generation gene sequencing, BPA is working to integrate genomic medicine into healthcare to create personalised, precision medicine. This aims to shorten diagnosis time and enable early intervention.

Investments in BPA will also include state-of-the-art genomics infrastructure to support the world’s largest coral genomics sequencing project, which will help researchers understand the genetic makeup of corals and how they might respond to climate change.
Who is Australia’s Life Science Research Community?
Who is Australia’s Life Science Research Community?

100,000

Total pool of Australian publicly funded researchers
Who is Australia’s Life Science Research Community?

30,000

We estimate 30% are bioscience related

+200,000 students
Who is Australia’s Life Science Research Community?

- 50% (15,000) - Healthcare
- 30% (9,000) - Agriculture
- 20% (6,000) - Environment
What are their skill level in bioinformatics?
Four broad expertise types

Occasional users of bioinformatics tools or services, e.g. BLAST
Four broad expertise types

Omics data analysis is a critical contributor to research outcomes

- **Biology (wet-lab)**
  - Occasional users of bioinformatics tools or services e.g. BLAST

- **Data Intensive**

Australian BioCommons
Four broad expertise types

- **Biology (wet-lab)**
  - Occasional users of bioinformatics tools or services e.g. BLAST

- **Data Intensive**
  - omics data analysis is a critical contributor to research outcomes

- **Bioinformatics Intensive**
  - Research is fully dependent on advanced use of bioinformatics
Four broad expertise types

- **Biography (wet-lab)**
  - Occasional users of bioinformatics tools or services e.g. BLAST

- **Data Intensive**
  - Omics data analysis is a critical contributor to research outcomes

- **Bioinformatics Intensive**
  - Research is fully dependent on advanced use of bioinformatics

- **Bioinformaticians**
  - Research is exclusively on technique, algorithm or tool development
How many researchers are in each group?
Expected change over the next 5 years
What does the Australian Life Science Research Community require with respect to bioinformatics infrastructure in order to fully participate in, and benefit from an era of data-driven science?
National Consultation (2017)

Attendance of over 150 research leaders in Life Science
National Reference Group

Prof Jacquie Batley (Plant Genetics & Breeding, UWA)
Prof Dave Burt (Director Genomics, UQ)
Prof Peter Cameron (Academic Director, The Alfred Emergency and Trauma Centre/Monash)
Prof Joanne Daly (CSIRO Honorary Fellow)
Prof Frank Gannon (Director, QIMR Berghofer)
Prof Rob Henry (Director, QAAFI, UQ)
Prof Ary Hoffmann (Biosciences, Melbourne U)
Prof Dean Jerry (Dep Director, JCU Centre for Tropical Fisheries and Aquaculture,)
Prof Ryan Lister (Head, Epigenetics and Genomics, Harry Perkins Inst/UWA)
Prof John Mattick (Director, Garvan Institute)
Prof Kathryn North (Director, MCRI)
Prof Nicki Packer (Macquarie U & Inst for Gycomics, Griffith U)
A/ProfTony Papenfuss (President ABACBS, Computational biology WEHI & Petermac)
Dr Maurizio Rossetto (NSW Royal Bot Gardens)
Prof Eric Stone (Director, ANU-CSIRO Centre for Genomics, Metabolomics and Bioinformatics, ANU)
Dr Jen Taylor (Group leader Bioinformatics, CSIRO)
Prof Steve Wesselingh (Director, SAHMRI)
Prof James Whisstock (Monash, EMBL-Australia)
Prof Marc Wilkins (Director, Ramaciotti Centre for Genomics, UNSW)
Prof Tony Bacic (Director, LaTrobe Institute of Agriculture and Food, LTU)
International Advisory Group

Vivien Bonazzi, Senior Advisor Data Science Tech & Innovation, BD2K Initiative (NIH)
Paul Flicek, Lead, Vertebrate Genomics & ENSEMBL, EBI-EMBL
Jaap Heringa, Head, ELIXIR-NL
Jason Williams, Education, Outreach and Training Lead, CyVerse (NSF)
Rebecca Johnson, Director, Australian Museum Research Institute, NSW
Tony Papenfuss, Head, Computational Biology, WEHI, VIC
Mark Walker, Director, Aust Infectious Disease Res Centre, UQ, QLD
Delphine Fleury, Aus Centre for Plant Functional Genomics, SA
Sean Grimmond, Director, Centre for UoM Cancer Research, VIC
International Study Tour (2018)

1. ELIXIR Germany/deNBI (Alf Puhler)
2. European Commission (Szilvia Nemeth, Brussels)
3. EU Australian Embassy (Alex Cooke, Brussels)
4. ELIXIR Compute Platform (Steven Newhouse + Susheel Varma, EBI)
5. ELIXIR Hub (Andrew Smith, Wellcome Genome Campus, Hinxton)
6. European Genome Archive (Thomas Keane, EBI) - also a one day workshop in Local EGA
7. European Bioinformatics Institute:
   a. Ensembl (Paul Flicek)
   b. Mouse Informatics (Terry Meehan)
   c. Microbiome metagenomics (Rob Finn)
8. Amazon Research (Vicki Schneider Cambridge)
9. Cancer Research UK (Geoff Macintyre, Cambridge University)
10. Cyverse-UK @ Earlham Institute (Rob Davey)
11. ELIXIR UK (Gabriella Rustici, Carole Goble)
12. ELIXIR Netherlands (Jaap Heringa)
13. ELIXIR Belgium (Frederik Coppens)
14. Fred Hutch Data Commonwealth (Matthew Trunnell, Fred Hutch Cancer Research Centre)
15. NCI Genomic Data Commons (Bob Grossman, University of Chicago)
16. NIH Data Commons Pilot Phase ‘Carbon’ team (Anthony Philippakis & Jonathan Scheffli & Danielle Ciofani & Dan Macarthur, Broad Institute)
17. US Department of Agriculture (Jack Okamura)
18. Kids First Data Commons (Adam Resnick, Children’s Hospital of Philadelphia)
19. National Science Foundation BIO (Dianne Okamura, NSF-BIO)
20. Galaxy team at Johns Hopkins (Jeremy Goecks and Anton Nekrutenko)
21. National Human Genome Research Initiative Commons aka ANVIL (amongst a number of other components, funding of Galaxy integration with Firecloud and GDC tech over a series of core data resources. Discussions with Galaxy folk @ Johns Hopkins)
22. NIH Data Commons Pilot Phase ‘Helium’ team ‘CommonsShare’ (Claris Castillo, RENCI, University of North Carolina)
23. National Cancer Institute Cancer Research Data Commons (Tony Kerlavage, CIO of National Cancer Institute)
24. NIH Data Commons (Vivien Bonazzi)
25. Cyverse (Nirav Merchant, University of Arizona)
Required elements of a ‘BioCommons’
Required elements of a ‘BioCommons’

Support different expertise levels

For Biologists, Data-intensive and Bioinformatics intensive Researchers and Bioinformaticians
Required elements of a ‘BioCommons’

**Support different expertise levels**
For Biologists, Data-intensive and Bioinformatics intensive Researchers and Bioinformaticians

**Support across all of Life Science**
For Human health, Environmental and Agricultural researchers who apply omics to their research
Required elements of a ‘BioCommons’

Support different expertise levels
For Biologists, Data-intensive and Bioinformatics intensive Researchers and Bioinformaticians

Support across all of Life Science
For Human health, Environmental and Agricultural researchers who apply omics to their research

Simple to use service platforms
Simple to use service platforms to which a researcher can bring their research goals, tools, pipelines and data
Required elements of a ‘BioCommons’

Support different expertise levels
For Biologists, Data-intensive and Bioinformatics intensive Researchers and Bioinformaticians

Simple to use service platforms
Simple to use service platforms to which a researcher can bring their research goals, tools, pipelines and data

Support across all of Life Science
For Human health, Environmental and Agricultural researchers who apply omics to their research

Technical platform for more advanced users
A technical platform on which data integration can more easily occur
Required elements of a ‘BioCommons’

Support different expertise levels
- For Biologists, Data-intensive and Bioinformatics intensive Researchers and Bioinformaticians

Support across all of Life Science
- For Human health, Environmental and Agricultural researchers who apply omics to their research

Technical platform for more advanced users
- A technical platform on which data integration can more easily occur

Simple to use service platforms
- Simple to use service platforms to which a researcher can bring their research goals, tools, pipelines and data

Workforce transition
- Training and skills development for all user types is essential
# Required elements of a ‘BioCommons’

<table>
<thead>
<tr>
<th>Support different expertise levels</th>
<th>Support across all of Life Science</th>
<th>Workforce transition</th>
<th>Global participation</th>
</tr>
</thead>
<tbody>
<tr>
<td>For Biologists, Data-intensive and Bioinformatics intensive Researchers and Bioinformaticians</td>
<td>For Human health, Environmental and Agricultural researchers who apply omics to their research</td>
<td>Training and skills development for all user types is essential</td>
<td>A mechanism through which Australia can represent its interests and participate in global data endeavours in life sciences</td>
</tr>
<tr>
<td>Simple to use service platforms</td>
<td>Technical platform for more advanced users</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple to use service platforms to which a researcher can bring their research goals, tools, pipelines and data</td>
<td>A technical platform on which data integration can more easily occur</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
BioCommons Pathfinder Project (2019)
BioCommons Pathfinder Project (2019)

High level aims are to continue to actively engage the Australian bioscience community to deliver:

- An operating infrastructure providing a core set of bioinformatics services
- A set of research activities and associated communities providing exemplars for others to follow
- A consortium of participants providing guidance and implementation support
- A strategic plan for the Australian BioCommons
- A five year operational plan for the delivery of the BioCommons (through to 2023)
BioCommons Pathfinder Project (2019)
Expanding Galaxy Australia for new communities

A simple to use service platform
Expanding Galaxy Australia for new communities

Hosted web-accessible platform for reproducible, and transparent computational biological research.

Free to use

3566 registered users

800+ bioinformatics tools

200+ reference datasets

600 GB working data storage

30+ publications
Expanding Galaxy Australia for new communities

Metabolomics

Working with Metabolomics
Australia data production facilities.
Installing tools and reusable pipelines for metabolomics data analysis.

Outcome: Democratisation of analysis to users of the facility and increased efficiencies of the facility
Expanding Galaxy Australia for new communities

**Metabolomics**

Working with Metabolomics Australia data production facilities. Installing tools and reusable pipelines for metabolomics data analysis.

**Outcome:** Democratisation of analysis to users of the facility and increased efficiencies of the facility

**Phylogenomics**

Installing command-line driven pipelines for understanding species diversity onto Galaxy Australia.

**Outcome:** Enabling immediate access to analysis by conservation biologists to interrogate DNA sequences to understand species diversity
## Expanding Galaxy Australia for new communities

<table>
<thead>
<tr>
<th>Metabolomics</th>
<th>Phylogenomics</th>
<th>Connect Data Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Working with Metabolomics Australia data production facilities. Installing tools and reusable pipelines for metabolomics data analysis.</td>
<td>Installing command-line driven pipelines for understanding species diversity onto Galaxy Australia.</td>
<td>Working with AARNet to enable direct mounting of personal Cloudstor storage to the Galaxy analysis service.</td>
</tr>
<tr>
<td><strong>Outcome:</strong> Democratisation of analysis to users of the facility and increased efficiencies of the facility</td>
<td><strong>Outcome:</strong> Enabling immediate access to analysis by conservation biologists to interrogate DNA sequences to understand species diversity</td>
<td><strong>Outcome:</strong> seamless movement of data between long term data storage and the analysis environment</td>
</tr>
</tbody>
</table>
A Data Commons for paediatric cancer in Australia

Global participation

Support multiple areas of Life Science - medical genomics
A Data Commons for paediatric cancer in Australia

Kids First Data Commons

An output from the NIH Data Commons Fund

Secure cloud-hosted web-accessible platform for reproducible, and transparent computational biological research.

Gen3 and Cavatica technology

AWS (East Coast US) hosted

Data access control via standard NIH policy and procedures
A Data Commons for paediatric cancer in Australia

Zero Childhood Cancer

A national effort led by Children’s Cancer Institute and The Kids Cancer Centre at Sydney Children’s Hospital.

Brings together all major Australian clinical and research groups working in childhood cancer.

Australia’s first ever personalised medicine program for children with high-risk or relapsed cancer.
A Data Commons for paediatric cancer in Australia

Zero Childhood Cancer
A national effort led by Children’s Cancer Institute and The Kids Cancer Centre at Sydney Children’s Hospital.

Brings together all major Australian clinical and research groups working in childhood cancer.

Australia’s first ever personalised medicine program for children with high-risk or relapsed cancer.

Deploying Kids First DC in Australia
Childhood cancers are thankfully rare, and sensitive human-derived data cannot leave certain jurisdictions

Deploying a Kids First Data Commons in Australia (in AWS-Sydney) allows Australian and US data to be analysed in identical systems and outputs directly compared.

Outcome: Data from small patient numbers in Australia can be safely analysed against much larger global datasets to inform personalised medicine programs.
Systems for de novo genome assembly and annotation

Support multiple areas of Life Science - conservation genomics
Systems for de novo genome assembly and annotation

Genome assembly

- Based on DNA sequence matching
- Computationally building very long virtual DNA sequences from short reads
- ‘Gold standard’ assemblies have chromosome-length DNA assemblies

“assembled” genome
Systems for de novo genome assembly and annotation

Genome annotation

“assembled” genome

Working out what the different bits of the assembled genome are (e.g. genes, control regions etc)

Necessary for inferring biological function from a genome sequence
Systems for de novo genome assembly and annotation

Genome annotation

Working out what the different bits of the assembled genome are (e.g. genes, control regions etc)

Necessary for inferring biological function from a genome sequence

“assembled” genome

“annotated” genome
Lower cost of genome sequencing means more groups are using it as a tool to analyse species which have never been sequenced.

Genome assembly and annotation both are computationally demanding.

Two existing consortia: Bioplatforms Australia Oz Mammals Genomics (OMG) and the Genomics of Australian Plants (GAP).
Genomics in Conservation

Lower cost of genome sequencing means more groups are using it as a tool to analyse species which have never been sequenced.

Genome assembly and annotation both are computationally demanding.

Two existing consortia: Bioplatforms Australia Oz Mammals Genomics (OMG) and the Genomics of Australian Plants (GAP).

Building fit-for-purpose systems

Working with OMG, GAP and other researchers to understand and document their computational challenges.

Identifying appropriate infrastructure (e.g. high memory nodes, cloud compute) and approaches (tool containerisation, extension of wall times etc)

**Outcome:** Conservation researchers can access computational systems that are suitable for their needs.
Phase 1 development of a BioCloud infrastructure

A technical platform for more advanced users
Phase 1 development of a BioCloud infrastructure

Building fit-for-purpose systems

Working with many researchers to understand and document their computational challenges and successes.

Facilitating collaboration between various providers (NCI, Pawsey, Nectar, QRIScloud, UQ, commercial).

Investigating: Development of a high-throughput cloud for bioinformatics

- Software and container orchestration (Lead: Pawsey)
- Data Movement (Lead: AARNet)
- Identity and Access management (Lead: AAF)
- User Experience
- Exploring commercial cloud for research purposes (Lead: USyd)

Outcome: Designing and piloting computational systems that are suitable for biologist’s needs.
Workforce Transition

Training and skills development
Workforce Transition

Upskilling in Bioinformatics

How to gain bioinformatics skills consistently ranks as a primary concern of life science researchers.

How to use tools and infrastructure is required by bioinformaticians.

Bioinformatics training across Australia is highly localised and variable.
Workforce Transition

Upskilling in Bioinformatics

How to gain bioinformatics skills consistently ranks as a primary concern of life science researchers.

How to use tools and infrastructure is required by bioinformaticians.

Bioinformatics training across Australia is highly localised and variable.

BioCommons Training 2019

With national experts, the development of new training material (in 2019: Snakemake and Nextflow, Phylogenetics Trees for Beginners).

Development of a network of highly motivated trainers.

Facilitating national training events at scale ('hybrid model')

Developing an ambitious National Plan for Bioinformatics Training

Outcome: An inclusive and scalable approach to bioinformatics training that complements local efforts.
Global Participation

A mechanism through which Australia can represent its interests and participate in global data endeavours in life sciences
Global Participation

Unites Europe’s leading life science organisations in managing and safeguarding the increasing volume of data being generated by publicly funded research.

Coordinates, integrates and sustains bioinformatics resources across 22 member states (and 1 observer).

Enables users in academia and industry to access services that are vital for their research.
Global Participation

How can Australia engage fully?

Elixir coordinates a number of highly engaged and active communities around several key areas: DATA, COMPUTE, TOOLS, INTEROPERABILITY and TRAINING and communities: HUMAN DATA, METAGENOMICS, METABOLOMICS, PLANT SCIENCES, PROTEOMICS.

The Australian BioCommons and ELIXIR are currently drafting a formal agreement to allow collaboration.

Outcome: Australia can fully engage in ELIXIR activities and Bioinformatics Infrastructure in Australia is much better aligned with peer infrastructures elsewhere.

Unites Europe’s leading life science organisations in managing and safeguarding the increasing volume of data being generated by publicly funded research.

Coordinates, integrates and sustains bioinformatics resources across 22 member states (and 1 observer).

Enables users in academia and industry to access services that are vital for their research.
Where next?
Where next?

- Domain Applications and Services
- Leadership and Governance
- Continuous Pathfinding
- BioCloud
- Resources and Facilities

BPA National & International Consultations
EMBL-ABR community building ANDS, NeCTAR, RDS programs BPA training, GVL, Galaxy services
Enduring National Research Infrastructure


Australian BioCommons
Where next?

- Domain Applications and Services
- Leadership and Governance
- Continuous Pathfinding
- BioCloud
  - Resources and Facilities

2017 - 2024

EMBL-ABR community building ANDS, NeCTAR, RDS programs BPA training, GVL, Galaxy services

BPA National & International Consultations

Enduring National Research Infrastructure

Australian BioCommons
Where next?

Enduring National Research Infrastructure

Domain Applications and Services
Leadership and Governance
Continuous Pathfinding

BioCloud

Resources and Facilities


Pathfinder

BPA National & International Consultations
EMBL-ABR community building ANDS, NeCTAR, RDS programs BPA training, GVL, Galaxy services

BPA National & International Consultations
EMBL-ABR community building ANDS, NeCTAR, RDS programs BPA training, GVL, Galaxy services
Project Partners and Core Team

Patrick Carnuccio, Jeff Christiansen, Thom Cuddihy, Paul Coddington, Marco de la Pierre, Brian Davis, Ian Duncan, Rhys Francis, Simon Gladman, Mark Gray, Dominique Gorse, Johan Gustafsson, Christina Hall, Carina Kemp, Igor Makunin, Steven Manos, Heath Marks, Chris Myers, Tiffanie Nelson, Sarah Nisbet, Gareth Price, Rosemarie Sadsad, Frankie Stevens, Andrew Lonie
Thanks!

jeff@biocommons.org.au

biocommons.org.au

@AusBiocommons
BioCommons Pathfinder Project (2019)

STEERING COMMITTEE

Andrew Gilbert (Bioplatforms Australia)
Ian Duncan (ARDC)
Carina Kemp (AARNet)
Sean Smith (NCI)
Mark Stickells (Pawsey)
John Zic (CSIRO)

A/Prof Oliver Hoffman (Head of Bioinformatics, Melb University Centre for Cancer Research)
Dr Warren Kaplan (Head of Bioinformatics, Garvan Institute)
Prof Dave Burt (Director, UQ Genomics, UQ)
A/Prof Nic Wadell (Head of Medical Genomics, QIMR Berghofer)
Prof Simon Easteal (Director, National Centre Indigenous Genomics, ANU)
Prof Graham King (Director, Southern Cross Plant Science, SCU)
Prof Rebecca Johnson (Director, Australian Museum Research Institute)
A/Prof Tony Papenfuss (Head of Computational Biology - Walter and Eliza Hall Institute and Petermac)