Progress in the AUSTRALIAN BIOCOMMONS

A/Prof Andrew Lonie

Director, Australian BioCommons

Australian Research Data Commons

Australian BioCommons

Thanks to rapid advances in sensing technologies...

DNA

mRNA

proteins

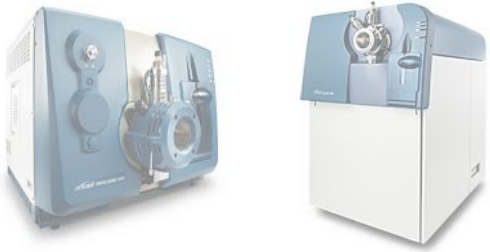
metabolites

...life sciences have become data-intensive

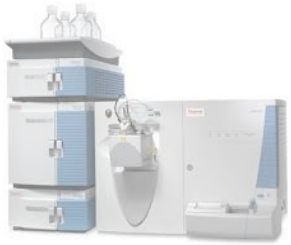
DNA

mRNA

proteins



metabolites



genomics

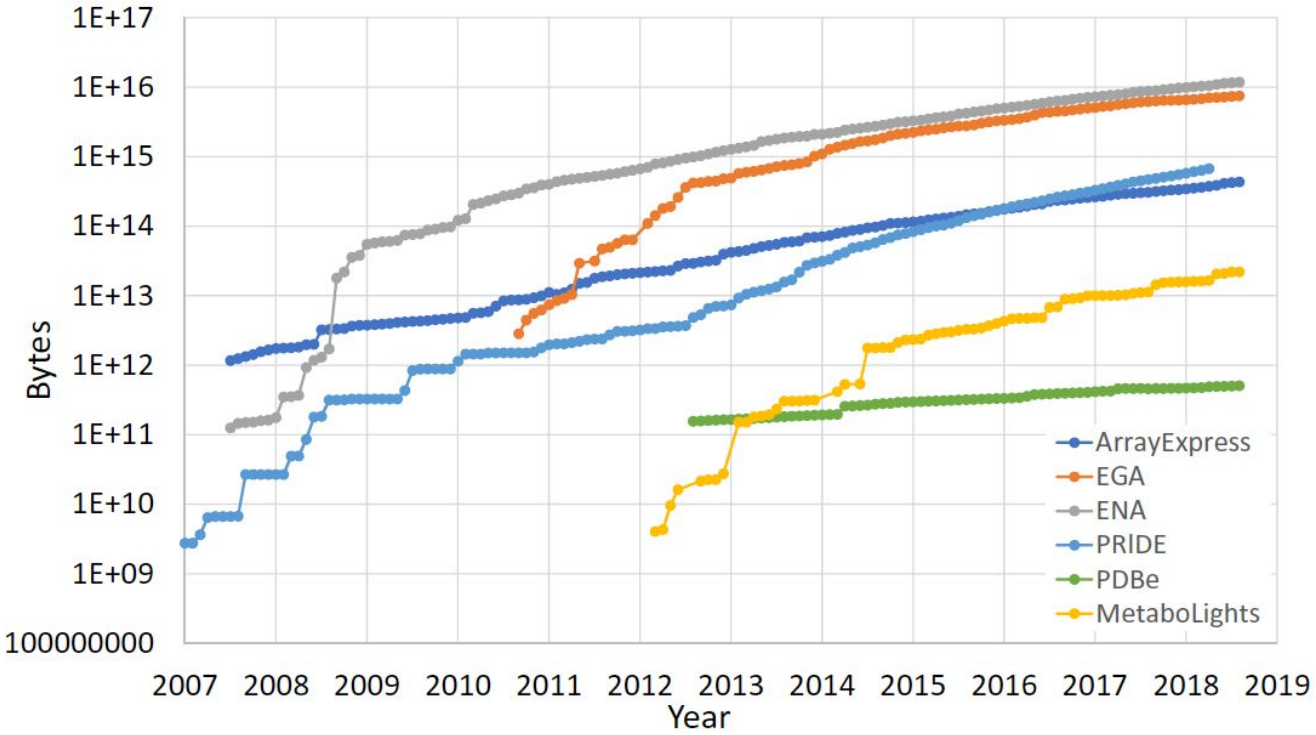
proteomics

transcriptomics

metabolomics

Australian BioCommons

...life sciences have become data-intensive



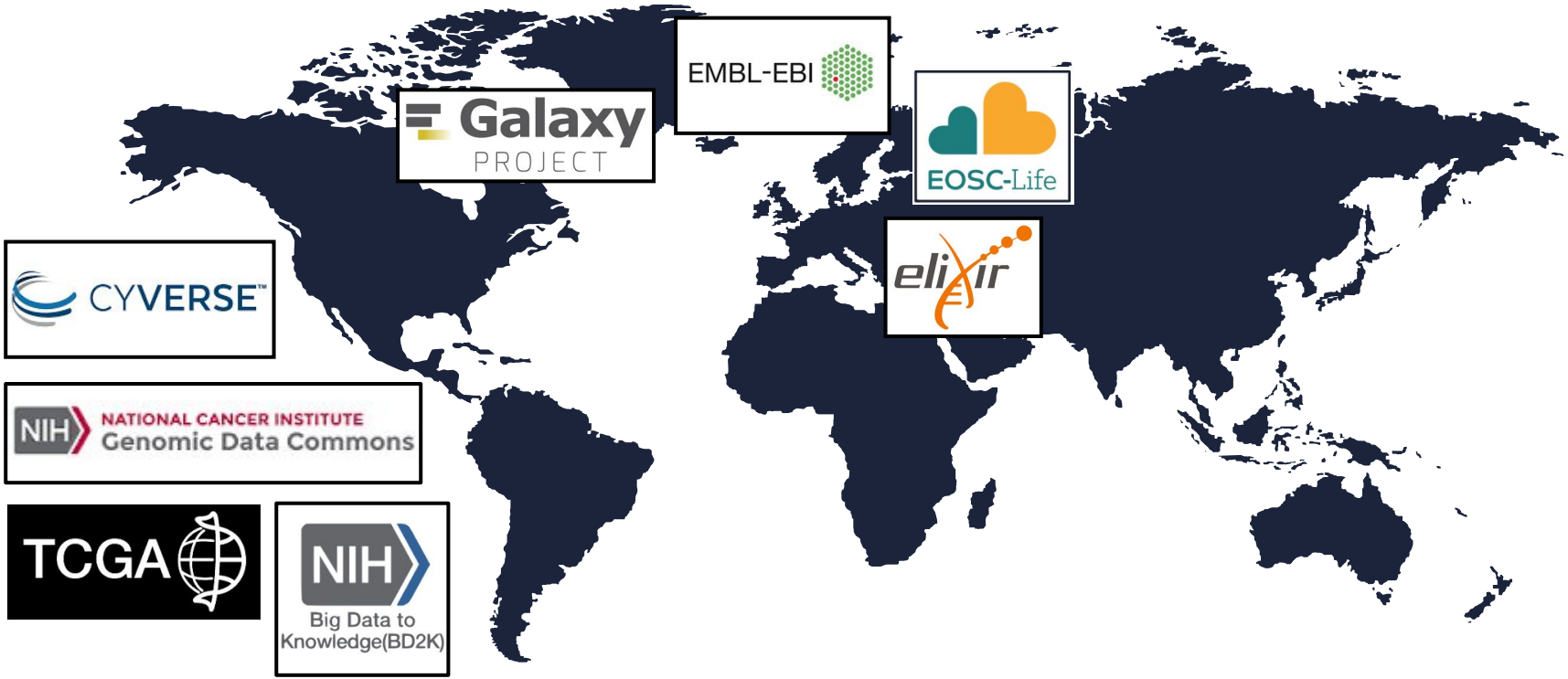
Data growth by EMBL-EBI data resource. Fig 2B. Cook et al, Nucleic Acids Research, 2019, Vol. 47, Database issue D15–D22 doi: 10.1093/nar/gky1124

160PB

10PB in 2012

Australian BioCommons

There is a global wave of investment



Australian BioCommons

We have an opportunity in Australia

NCRIS
National Research Infrastructure for Australia
An Australian Government Initiative

eResearch $911M
Complex Biology $216M
In addition to existing operating

Australian Government

FACILITIES FOR THE FUTURE
UNDERPINNING AUSTRALIA'S RESEARCH AND INNOVATION

Government Response to the 2016 National Research Infrastructure Roadmap
Research Infrastructure Investment Plan

		2017–18 to 2021–22**	Total to 2028–29
Digital data & eResearch platforms	Funding will ensure a regular upgrade cycle for maintaining Australia's Tier 1 HPC capability, across both the Pawsey Supercomputing Centre (Pawsey) and the National Computational Infrastructure. In addition, a scoping study	219.0	911.1
Complex Biology	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.	53.4	216.2

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 aim is 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?

Australian BioCommons

Who is Australia's Life Science Research Community?

100,000

Total pool of Australian publicly funded researchers

Australian BioCommons

Who is Australia's Life Science Research Community?

30,000

We estimate 30% are bioscience related

+200,000 students

Australian BioCommons

Who is Australia's Life Science Research Community?

50%

15,000

30%

9,000

20%

6,000

Australian BioCommons

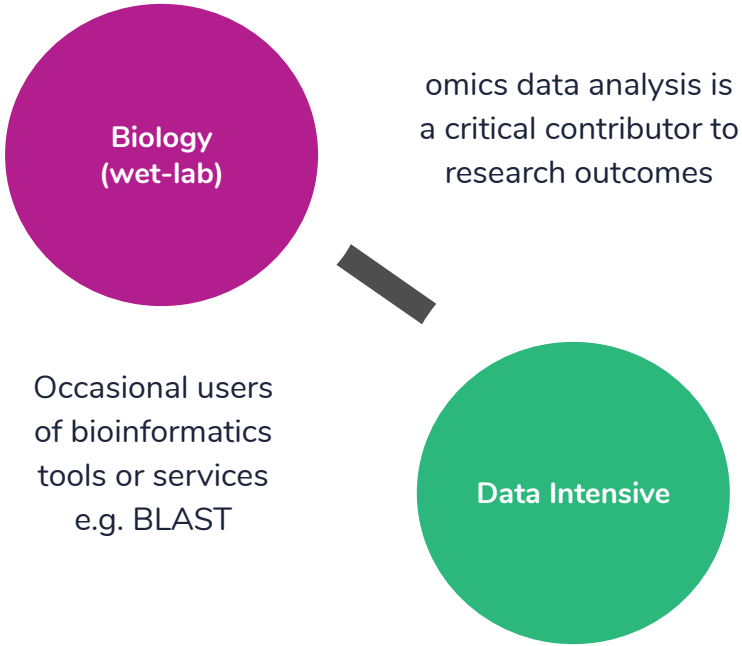
What are their skill level in bioinformatics?

Four broad expertise types

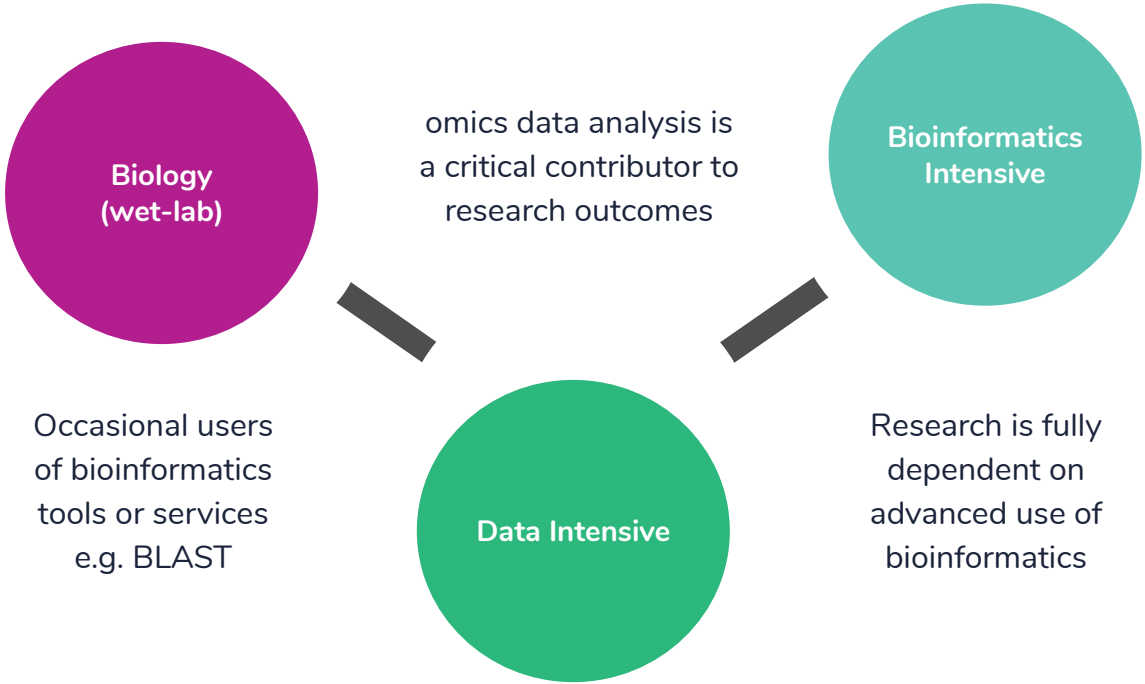
Occasional users of bioinformatics tools or services e.g. BLAST

Australian BioCommons

Four broad expertise types

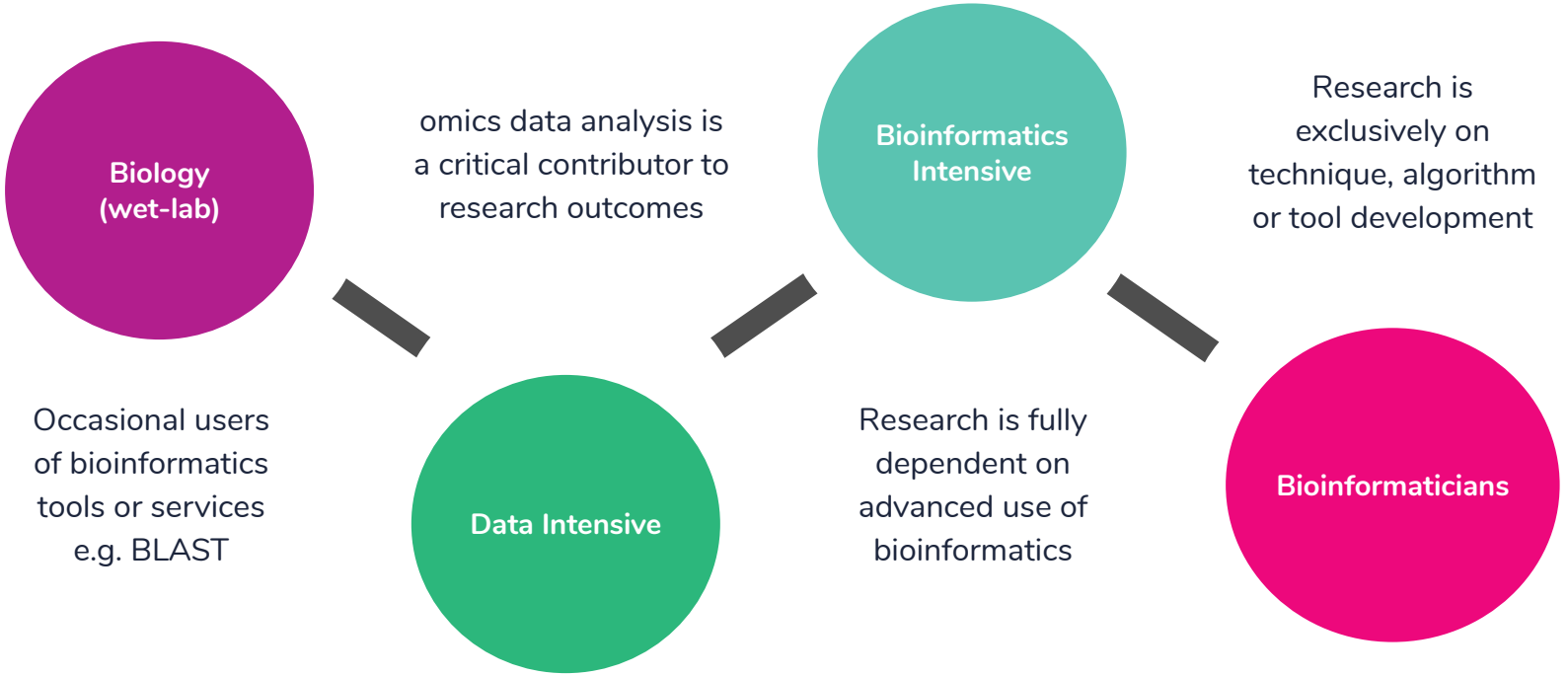


Four broad expertise types



Australian BioCommons

Four broad expertise types

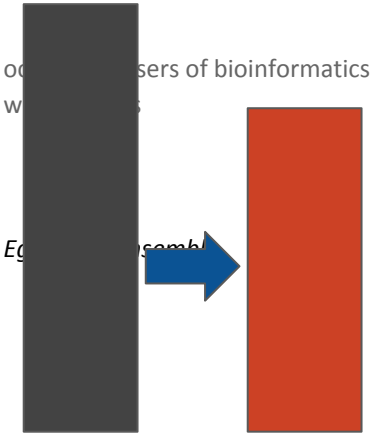
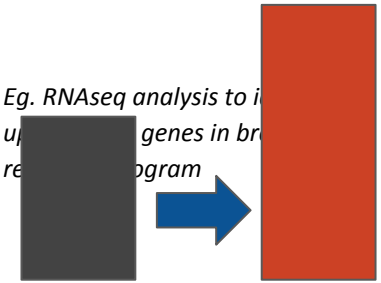
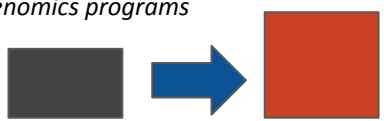
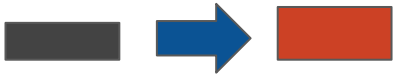


Australian BioCommons

Estimated # Australian biology researchers in 2018: **30,000**			
20,000 *(→ 15,000)* ***biology-focussed bioscience researchers*** occasional users of bioinformatics web services *Eg BLAST, Ensembl*	**7,000** *(→ 12,000)* ***data-intensive bioscience researchers*** 'omics data analysis is a critical contributor to the research outcomes *Eg. RNAseq analysis to identify upregulated genes in broader research program*	**2,000** *(→ 3,000)* ***bioinf-intensive bioscience researchers*** research is fully dependent on advanced use of bioinformatics *Eg. Genomic cancer research, population genomics/agricultural genomics programs*	Estimated #: **1,000** *(In 5 years → 1,500)* ***bioinformaticians*** research into/application of techniques & tool development *Eg. research generating new tool or statistical method; core facilities applying complex analyses*

Australian BioCommons

Estimated # Australian biology researchers in 2018: **30,000**

20,000 *(→ 15,000)* ***biology-focussed bioscience researchers***	**7,000** *(→ 12,000)* ***data-intensive bioscience researchers***	**2,000** *(→ 3,000)* ***bioinf-intensive bioscience researchers***	Estimated #: **1,000** *(In 5 years → 1,500)* ***bioinformaticians***
oc sers of bioinformatics w s	'omics data analysis is a critical contributor to the research outcomes	research is fully dependent on advanced use of bioinformatics	research into/application of techniques & tool development
Eg sembl	*Eg. RNAseq analysis to i u genes in br re ogram*	*Eg. Genomic cancer research, population genomics/agricultural genomics programs*	*Eg. research generating new tool or statistical method; core facilities applying complex analyses*
			

Important Transitions

Australian BioCommons

Defining a BioCommons

NCRIS
National Research Infrastructure for Australia
An Australian Government Initiative

BIOPLATFORMS AUSTRALIA

EMBL Australia Bioinformatics Resource

National Consultation (2017)

Attendance of over 150 research leaders in Life Science

Australian BioCommons

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

Australian BioCommons

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/U WA)
Prof John Mattick (Director, Garvan Institute)
Prof Kathryn North (Director, MCRI)
Prof Nicki Packer (Macquarie U & Inst for Glycomics, Griffith U)
A/Prof Tony Papenfuss (President ABACBS, Computational biology WEHI & Peter Mac)
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)

Australian BioCommons

International Study Tour (2018)

BIOPLATFORMS AUSTRALIA · ARDC · aarnet

1. ELIXIR Germany/deNBI (Alf Pohler)
2. European Commission (Edina Nemeth, Brussels)
3. EU Australian Embassy (Alex Cooke, Brussels)
4. ELIXIR Compute Platform (Steven Newhouse + Busheel 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 "Cartos" team (Anthony Philippakis & Jonathan Scheffli & Danielle Ciofani & Dan Macarthur)
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 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' (Clarix 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)

Australian BioCommons

Study trip: broad conclusions

Global scale compute and data infrastructures are increasingly underpinning global scale research in life sciences

- Cloud First is pervasive across EU and US
- In the US (but not EU) true partnerships arising with cloud providers
- The concept of Data Commons is very strong in the US, focused on data and method sharing
- ELIXIR is doing a very good job of coordinating data infrastructure across Europe
- ELIXIR/EBI compute strategy firmly cloud focussed - federated compute + data across EU
- Federated approaches to data infrastructure are developing, and accessible
- National Bioinformatics Infrastructures can deliver benefits across industry engagement
- Galaxy is extremely well regarded as a community analysis platform in both Europe and the USA

The Australian BioCommons - principles

- A national focus on capabilities and communities
- Partner internationally: participate in and contribute to larger critical mass efforts where possible; reuse and improve rather than build anew
- Build a software and expertise capability that will reduce duplication of infrastructure management in Australia and allow efforts to be re-focussed on methods development and dissemination
- Promote the development of, and build on, high throughput cloud infrastructure that is interoperable with international (initially US and European) equivalents, using established, well supported software platforms
- Streamline the exchange of tools, workflows, data and training and expertise both nationally and internationally

Australian BioCommons

The Australian BioCommons

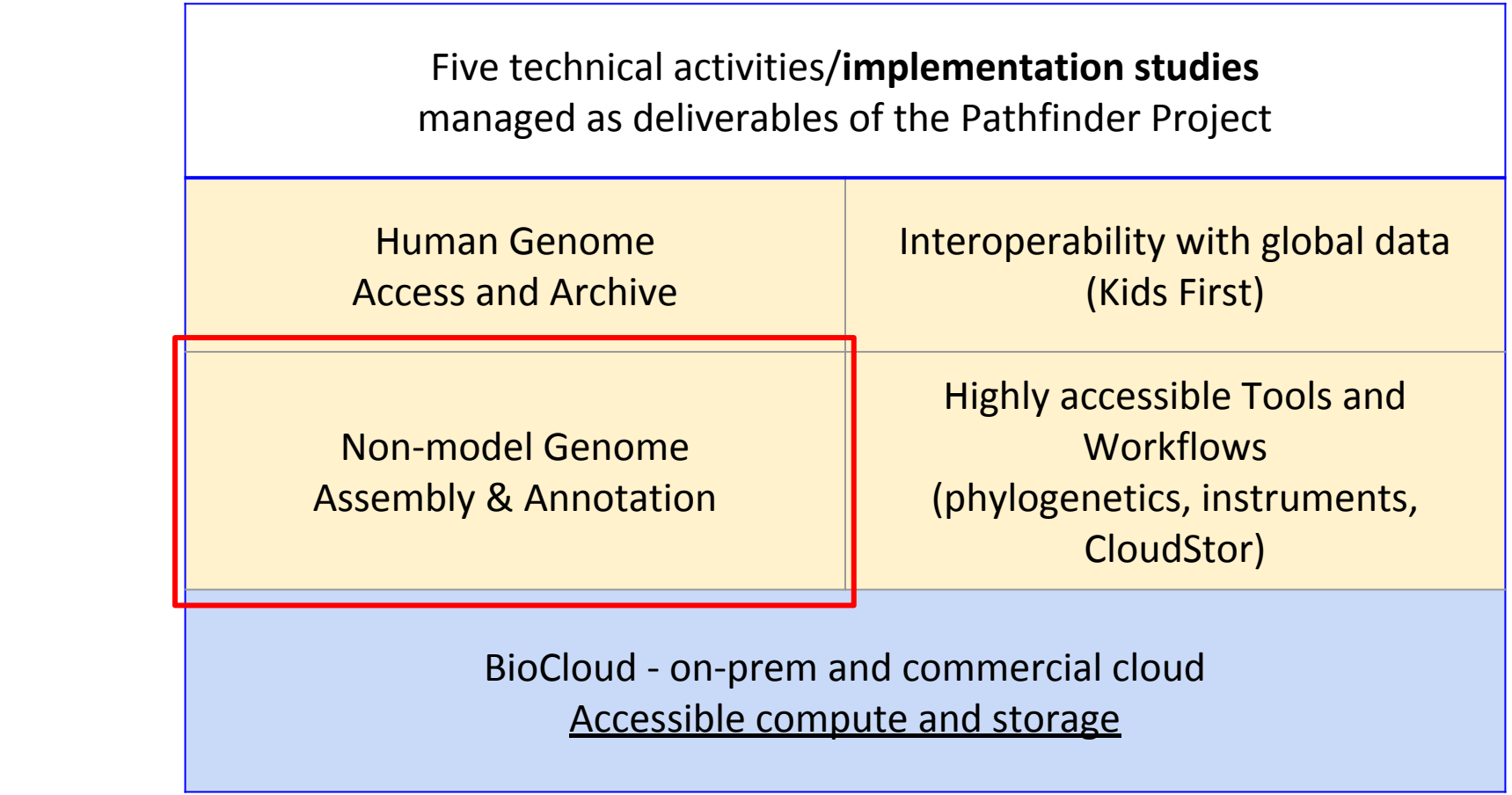
We are not establishing a bioinformatics service:

- We are aiming to reduce duplication of bioinformatics infrastructure management to allow researchers to focus on methods development and use (and away from systems administration tasks)

- We are committed to facilitating workforce upskilling through providing:
  - a national bioinformatics training program (aimed at both life-scientists and bioinformaticians)
  - integrated services to enable exchange of expertise around tools,. techniques etc run across a variety of infrastructures and services

BioCommons 2019

Five technical activities/**implementation studies** managed as deliverables of the Pathfinder Project	
Human Genome Access and Archive	Interoperability with global data (Kids First)
Non-model Genome Assembly & Annotation	Highly accessible Tools and Workflows (phylogenetics, instruments, CloudStor)
BioCloud - on-prem and commercial cloud <u>Accessible compute and storage</u>	

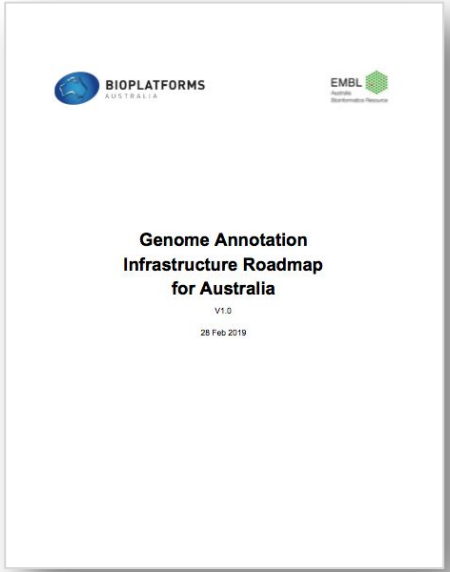


Australian BioCommons

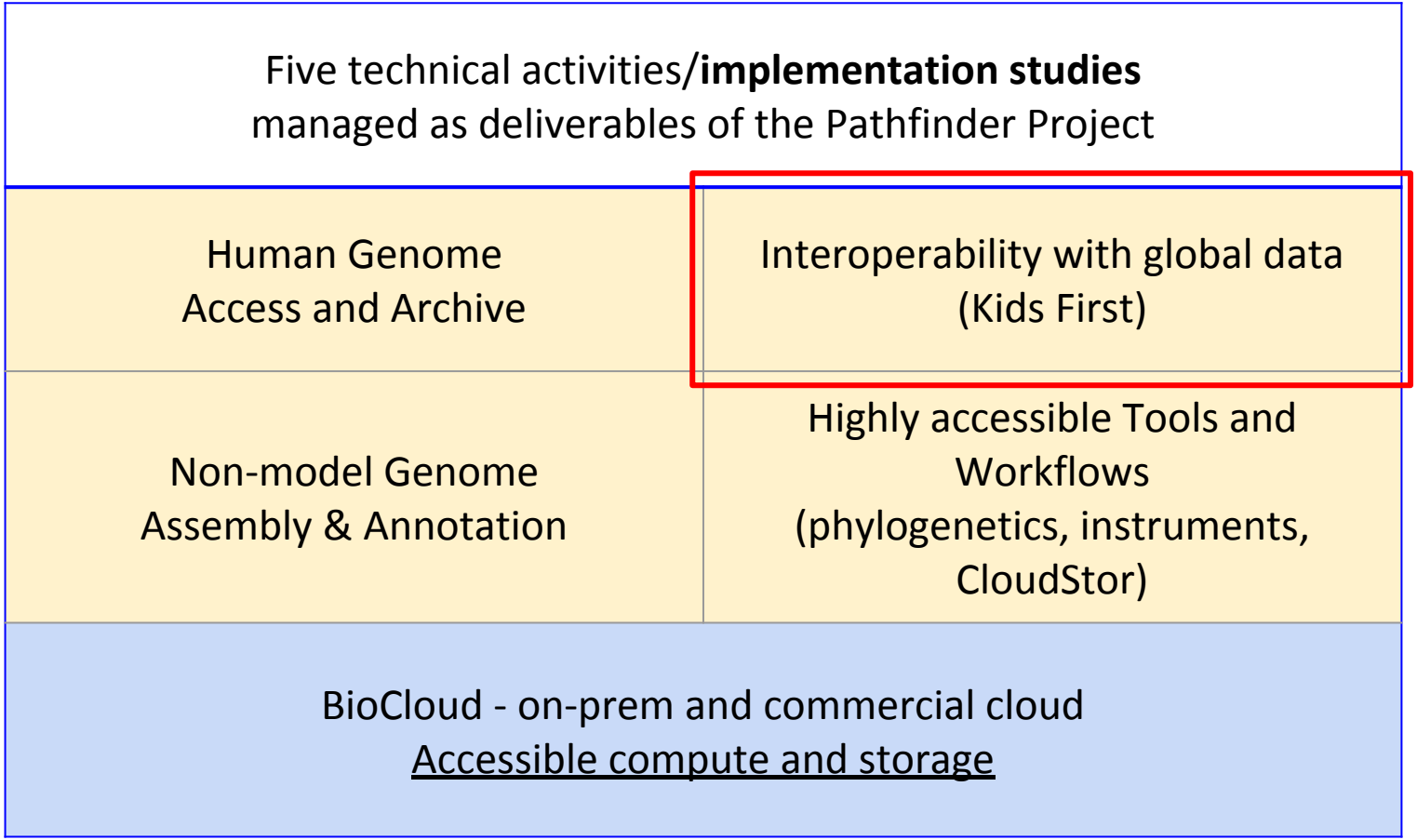
Systems for de novo genome assembly and annotation

- **Communities and infrastructure services identified for common omic-based challenges:** e.g. Genome annotation; Multi-omics integration; Comparative Genomics; etc

- **Development of a Genome Annotation Infrastructure Roadmap for Australia**
  - First iteration released for community comment last week
    - http://bit.ly/aus-genome-annotation – COMMENTS WELCOME!
  - Subsequent iterations following consultation with:
    - International entities operating genome annotation infrastructure elsewhere (e.g. Ensembl, EBI)
    - Australian Infrastructure Providers (e.g. NCI, Pawsey, ARDC)
  - Final community agreed roadmap planned for November 2019, to inform future BioCommons investment



Australian BioCommons



Australian BioCommons

The Australian Bioinformatics Commons Paediatric Cancer Pathfinder Project - Harmonised analysis of geographically separated and jurisdictionally protected data resources

Mark Cowley, Allison Heath, Maely Gauthier, Marie Wong-Erasmus, Jack DiGiovanna, Michele Mattioni, Paul Coddington, Brian Davis, Frankie Stevens, Chris Myers, Steven Manos, Adam Resnick, Andrew Lonie

1. Why?

Large-scale cancer WGS, RNA-Seq and methylome analyses have made a substantial impact on our understanding of many cancers, including their aetiology, identifying disease subtypes, novel pathways and new drug targets. While there are a number of extensive genomic cancer research programs globally, most focus on adult cancer; however, as all high-risk *paediatric* cancer subtypes are rare diseases, statistically significant correlation between subtype and genomic variation is inherently dependent on large sample numbers.

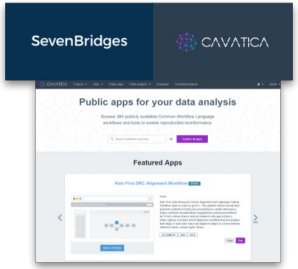
With only 200 new cases of high-risk paediatric cancer in Australia per year, it is imperative that we aggregate data with global data to understand and develop strategies to effectively treat high-risk childhood cancer

AWS **US East**

2a. US Partners

Gabriella Miller Kids First Paediatric Research Program (Kids First)

Kids First is a global-scale National Institutes of Health initiative devoted to exploring and analyzing genetic predisposition and/or somatic association within various childhood cancers and structural birth defects. Data from approximately 8,000 DNA and RNA samples from children affected with cancer or structural birth defects and their families are ready for analysis now, and the resource is expected to grow to more than 30,000 over the next few years. **Kids First** analyses are built on the **CAVATICA** platform - a mature, highly capable and widely used genomics analysis platform currently underpinning the **Kids First Data Resource**, developed and supported by **Seven Bridges Genomics.**



4. Global harmonised analysis

The approach is to extend the existing **CAVATICA** platform to enable efficient harmonised analyses across *geographically separated and jurisdictionally protected* data resources, leveraging commercial cloud standards.

The extended **CAVATICA** orchestration engine will allow **ZERO** and **Kids First** workflows and analysis tools to be used interchangeably and seamlessly across both datasets, effectively aggregating the separate datasets into a single virtual pan-continental dataset from the researcher's perspective, highly accessible through a global best practice analysis platform.

This world-class virtual data resource will allow us to address an initial critical research question focussed on discovery of phenotypically distinct brain tumour subtypes: specifically, we propose to harmonise the brain cancer transcriptome data from **ZERO** (n=160), and the **Kids First** PBTA/CBTTC project (n=921), and identify phenotypically distinct novel brain tumour subtypes.

By computationally pooling ZERO and KIDS FIRST datasets, we will have the power to identify rare brain cancer subtypes.

3. Research & Infrastructure partnership

A collaborative partnership between **ZERO, Kids First, Seven Bridges Genomics**, and **the Australian Bioinformatics Commons** (a $2.5m joint research infrastructure program funded by **Bioplatforms Australia**, the **Australian Research Data Commons**, and **AARNet**) aims to establish internationally federated computational infrastructure that will enable the harmonisation of **ZERO** Australian paediatric cancer data with the extensive genomic datasets from **Kids First**.

BIOPLATFORMS AUSTRALIA

NCI

aarnet Australia's Academic and Research Network

Australian Research Data Commons

AWS **Sydney**

2b. Australian Partners

ZERO Childhood Cancer program (ZERO)

Led by the **Children's Cancer Institute** in partnership with **Kids Cancer Centre at Sydney Children's Hospital**, aims to recruit 400 children on the National Clinical Trial by September 2020, in addition to the 58 children recruited on the pilot study (a total of 260 patients enrolled to date), applying deep whole genome (WGS) and transcriptome (RNA-Seq) sequencing, and methylome profiling to obtain a multi-dimensional molecular portrait of each child's cancer.

Australian BioCommons

Five technical activities/**implementation studies**
managed as deliverables of the Pathfinder Project

Human Genome Access and Archive	Interoperability with global data (Kids First)
Non-model Genome Assembly & Annotation	Highly accessible Tools and Workflows (phylogenetics, instruments, CloudStor)

BioCloud - on-prem and commercial cloud
Accessible compute and storage

Australian BioCommons

BioCommons ARDC Platforms program

Australian Research Data Commons

UoM + USyd + QCIF/UQ + NCI + Pawsey + AAF + AARNet + ARDC +(CSIRO, UNSW, MacquarieU, UoAdelaide)

1. ***Connecting instruments to analysis on the cloud***
2. ***Expanding and improving high accessibility GUI-focussed BYOD platforms (GUI-BYOD)***
3. ***Expanding and improving high flexibility CLI-focussed BYOD platforms (CLI-BYOD)***

BIOPLATFORMS AUSTRALIA

Australian BioCommons

PAWSEY supercomputing centre

NCI AUSTRALIA

qcif

THE UNIVERSITY OF MELBOURNE

aarnet

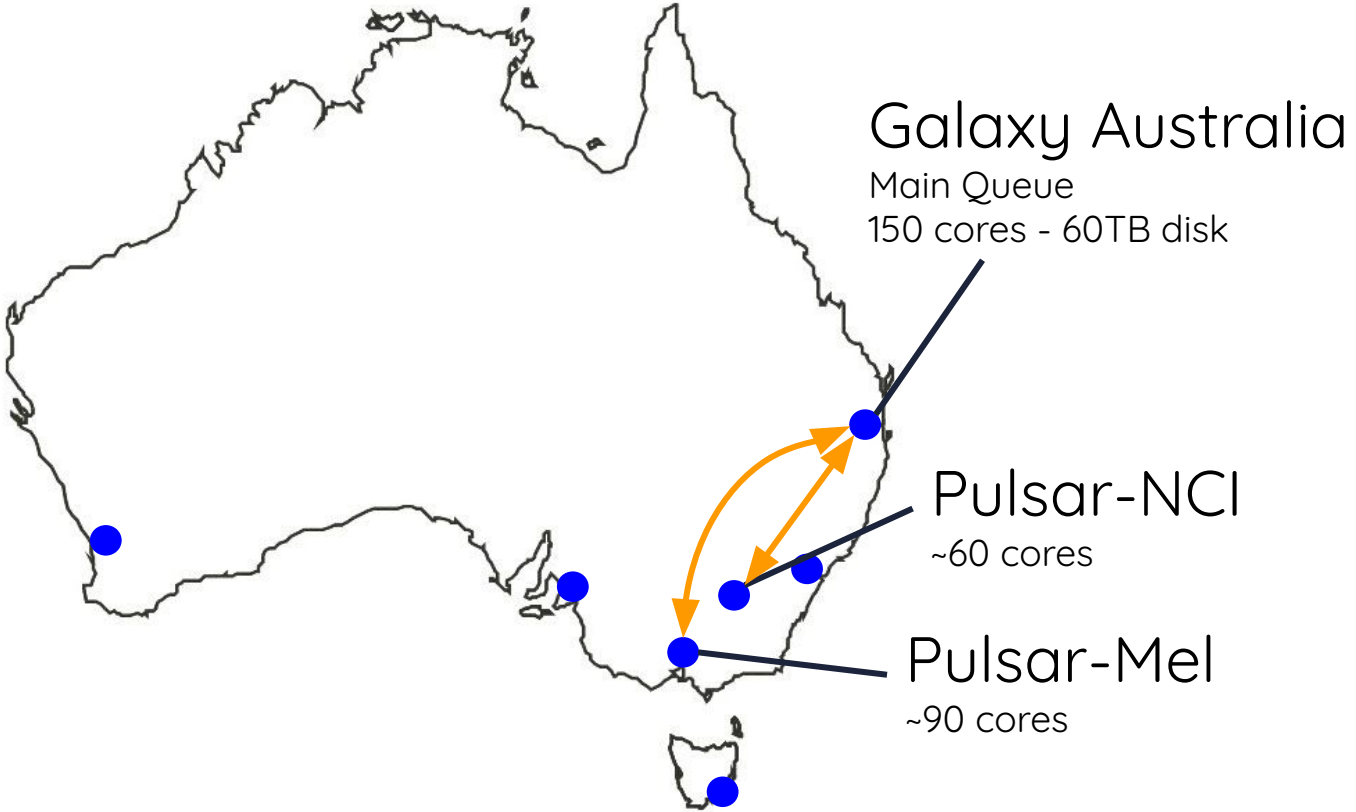
CSIRO

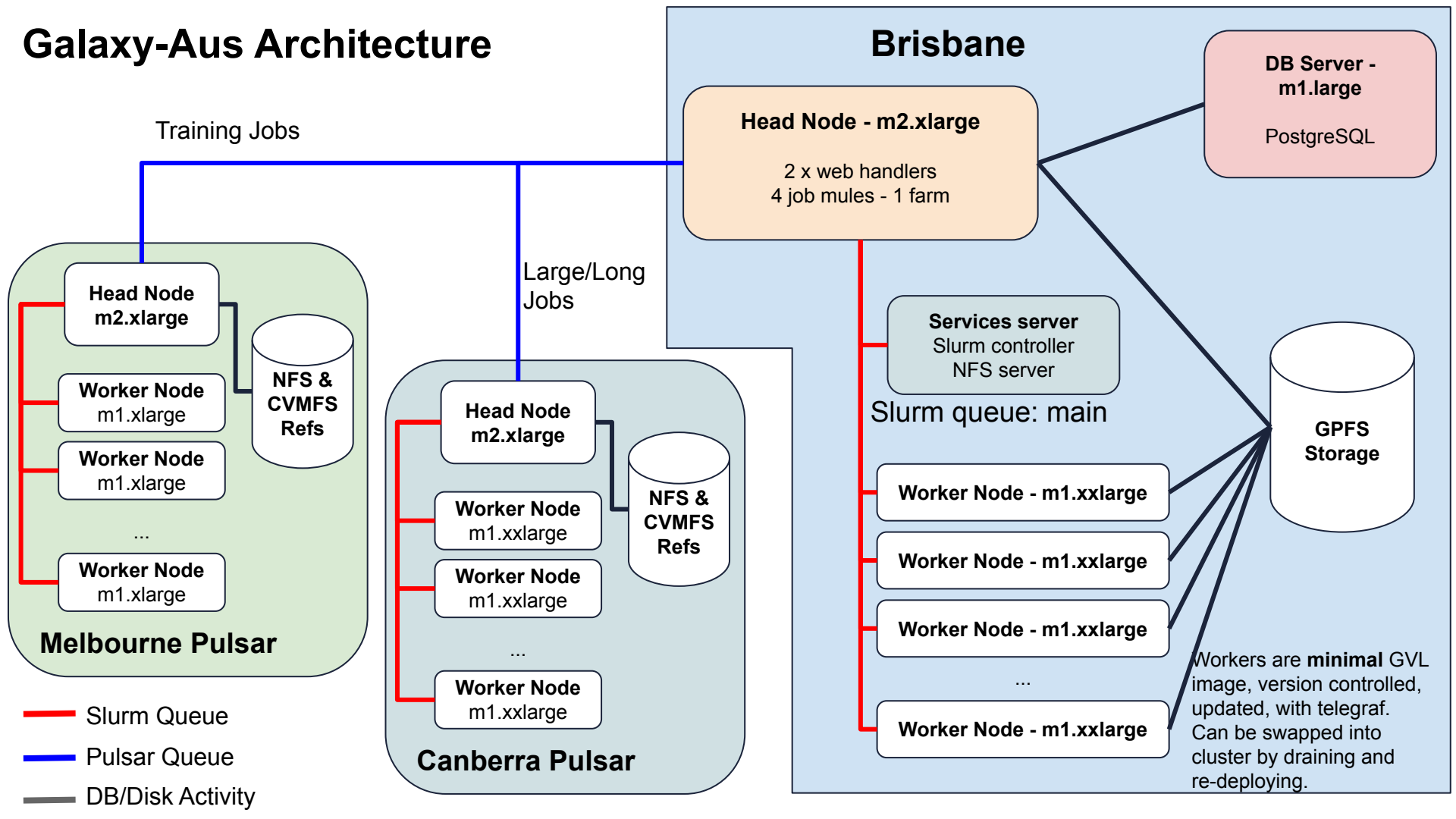
THE UNIVERSITY OF SYDNEY

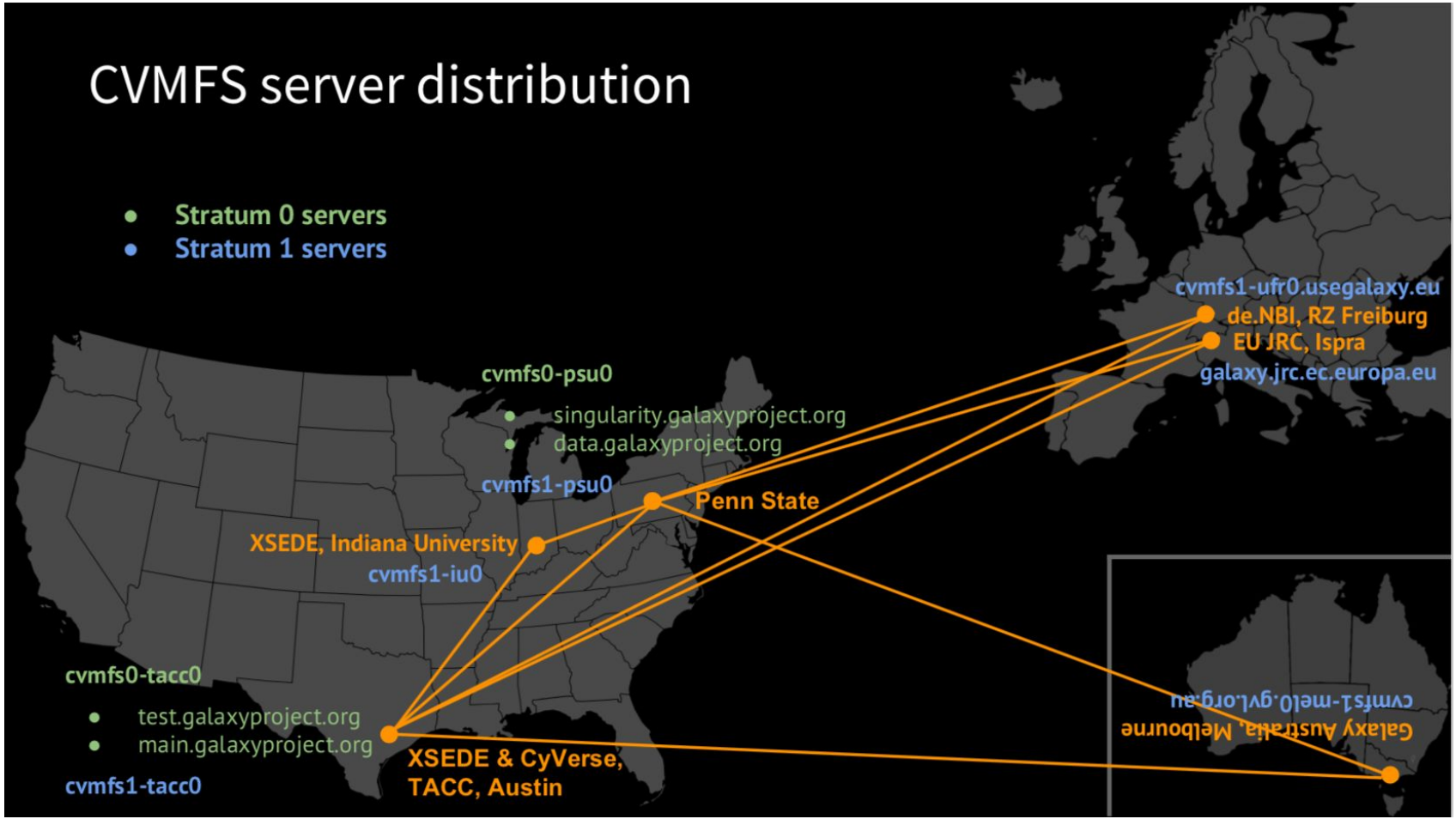
AUSTRALIAN ACCESS FEDERATION

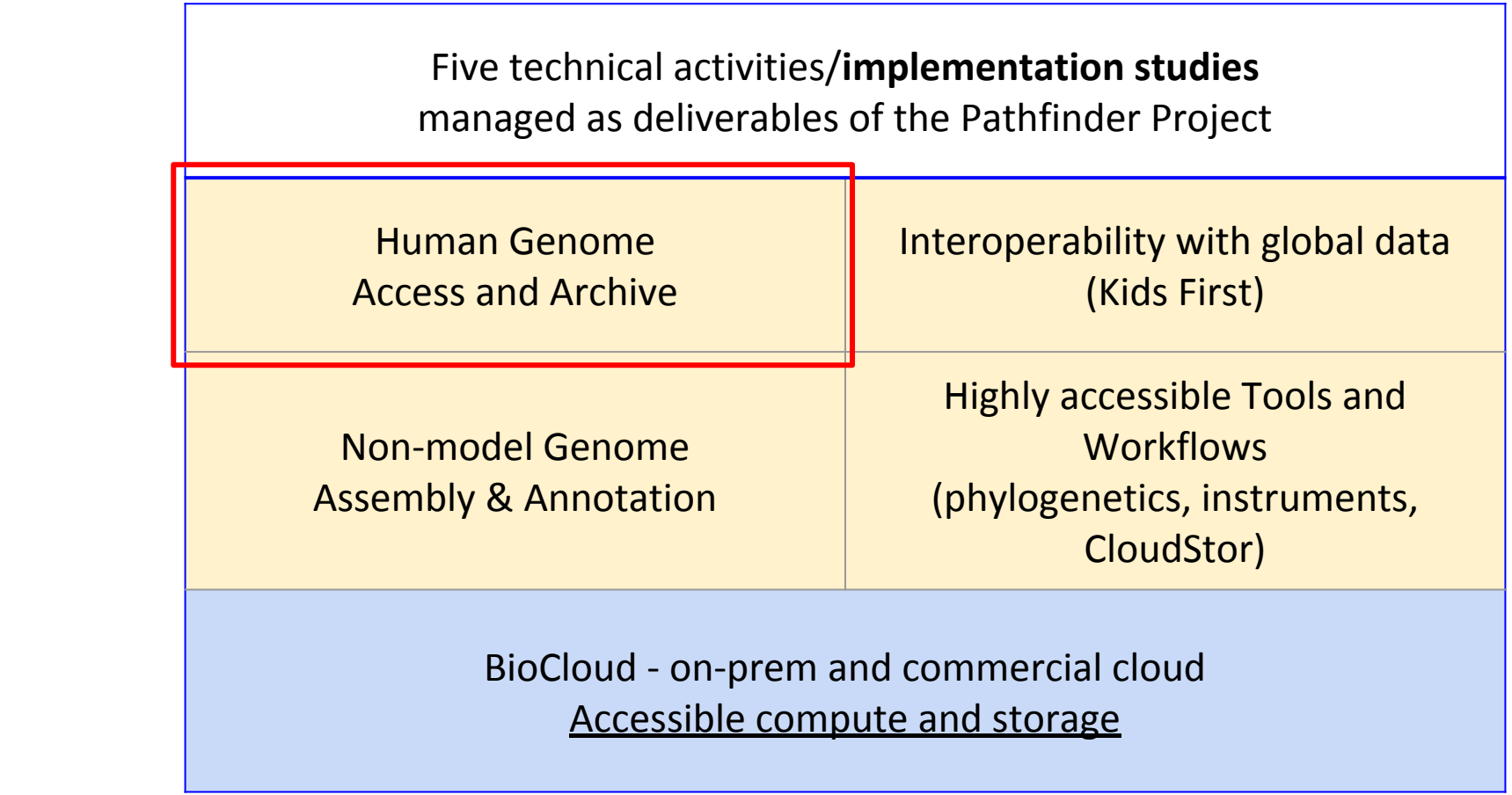
Galaxy Australia Infrastructure Locations

2018-now




Galaxy-Aus Architecture
Brisbane
Training Jobs
Head Node - m2.xlarge
2 x web handlers
4 job mules - 1 farm
DB Server - m1.large
PostgreSQL
Large/Long Jobs
Services server
Slurm controller
NFS server
Slurm queue: main
GPFS Storage
Head Node m2.xlarge
Worker Node m1.xlarge
Worker Node m1.xlarge
...
Worker Node m1.xlarge
NFS & CVMFS Refs
Melbourne Pulsar
Head Node m2.xlarge
Worker Node m1.xxlarge
Worker Node m1.xxlarge
...
Worker Node m1.xxlarge
NFS & CVMFS Refs
Canberra Pulsar
Worker Node - m1.xxlarge
Worker Node - m1.xxlarge
Worker Node - m1.xxlarge
...
Worker Node - m1.xxlarge
Workers are minimal GVL image, version controlled, updated, with telegraf. Can be swapped into cluster by draining and re-deploying.
Slurm Queue
Pulsar Queue
DB/Disk Activity


CVMFS server distribution
Stratum 0 servers
Stratum 1 servers
cvmfs1-ufr0.usegalaxy.eu
de.NBI, RZ Freiburg
EU JRC, Ispra
galaxy.jrc.ec.europa.eu
cvmfs0-psu0
singularity.galaxyproject.org
data.galaxyproject.org
cvmfs1-psu0
Penn State
XSEDE, Indiana University
cvmfs1-iu0
cvmfs0-tacc0
test.galaxyproject.org
main.galaxyproject.org
cvmfs1-tacc0
XSEDE & CyVerse,
TACC, Austin
Galaxy Australia, Melbourne
cvmfs1-melb.gvl.org.au



Australian BioCommons

(Thomas Keane presentation)

Five technical activities/**implementation studies**
managed as deliverables of the Pathfinder Project

Human Genome Access and Archive	Interoperability with global data (Kids First)
Non-model Genome Assembly & Annotation	Highly accessible Tools and Workflows (phylogenetics, instruments, CloudStor)

BioCloud - on-prem and commercial cloud
Accessible compute and storage

Australian BioCommons

We have an opportunity in Australia

NCRIS
National Research Infrastructure for Australia
An Australian Government Initiative

eResearch $911M
Complex Biology $216M
In addition to existing operating

Australian Government

FACILITIES FOR THE FUTURE
UNDERPINNING AUSTRALIA'S RESEARCH AND INNOVATION

Government Response to the 2016 National Research Infrastructure Roadmap
Research Infrastructure Investment Plan

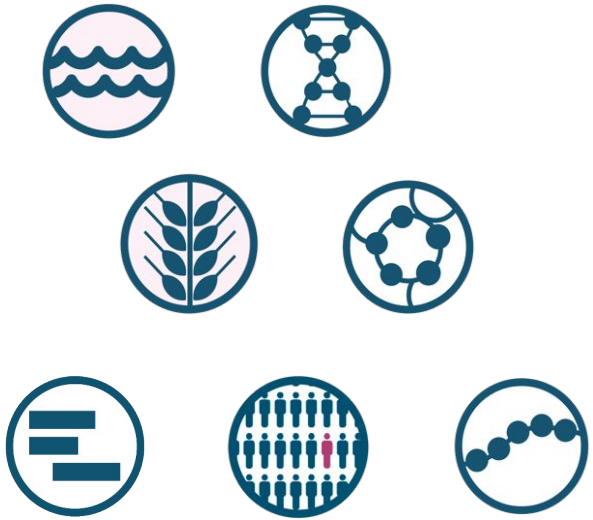
		2017–18 to 2021–22**	Total to 2028–29
Digital data & eResearch platforms	Funding will ensure a regular upgrade cycle for maintaining Australia's Tier 1 HPC capability, across both the Pawsey Supercomputing Centre (Pawsey) and the National Computational Infrastructure. In addition, a scoping study	219.0	911.1
Complex Biology	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.	53.4	216.2

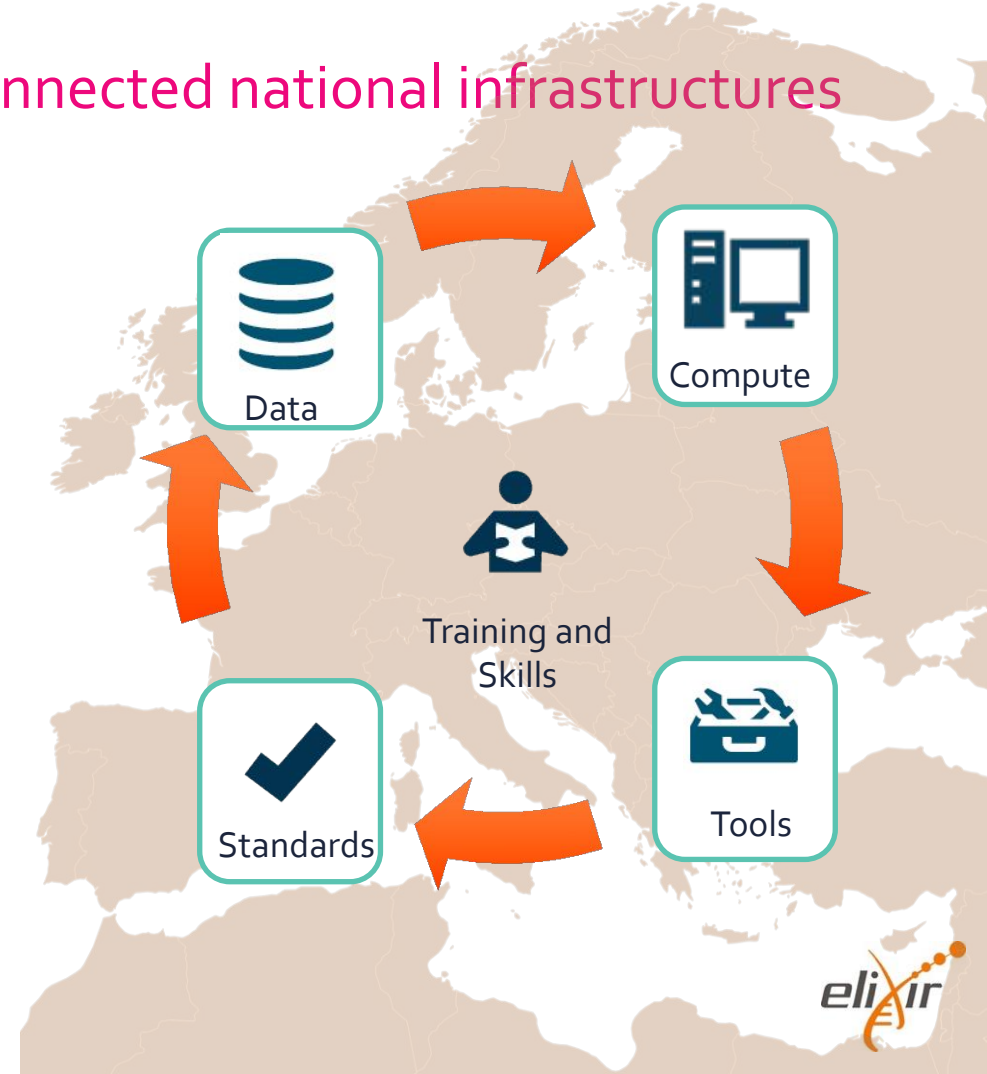
Case Study

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 aim is 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.

Data flowing freely between connected national infrastructures



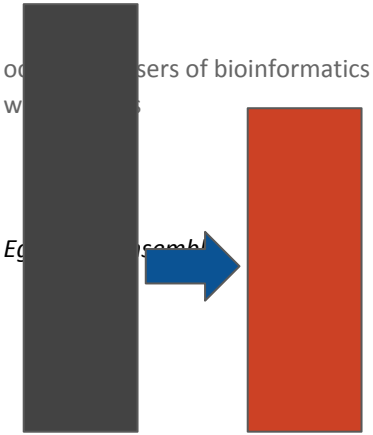
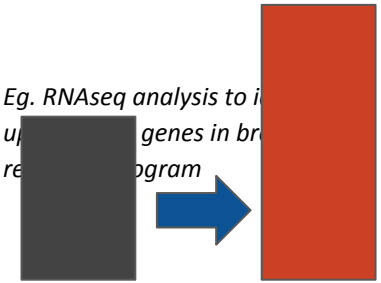
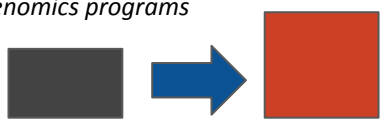
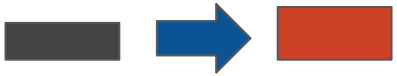


…delivered in partnership with research communities…the ELIXIR Communities

Workforce Transition

Australian BioCommons

Estimated # Australian biology researchers in 2018: **30,000**

20,000 *(→ 15,000)* ***biology-focussed bioscience researchers***	**7,000** *(→ 12,000)* ***data-intensive bioscience researchers***	**2,000** *(→ 3,000)* ***bioinf-intensive bioscience researchers***	Estimated #: **1,000** *(In 5 years → 1,500)* ***bioinformaticians***
oc sers of bioinformatics w s	'omics data analysis is a critical contributor to the research outcomes	research is fully dependent on advanced use of bioinformatics	research into/application of techniques & tool development
Eg sembl	*Eg. RNAseq analysis to i u genes in br re ogram*	*Eg. Genomic cancer research, population genomics/agricultural genomics programs*	*Eg. research generating new tool or statistical method; core facilities applying complex analyses*
			

Important Transitions

Australian BioCommons

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

Australian BioCommons

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

andrew@biocommons.org.au

biocommons.org.au

@AusBiocommons

Australian BioCommons