AUSTRALIAN BIOMOLECULAR DATA CAPABILITY
Phase 3a Report

Contents

PART A

1 Context

The definition of the Australian Biomolecular Data Capability (ABDC) has been explored through a small number of use cases. The use cases were selected for execution convenience, to include a spread of issues, and to cover the domains of human health, agriculture and environment. They did not reflect a decision about a future ABDC and its priorities, or the limits of its potential uses.

The question asked was: how can the proposed ABDC best contribute to opportunities with characteristics such as these? Four use cases were identified.

- Childhood cancer can provide insight into our participation in and use of global data holdings currently under development as well as highlight challenges in the integration of a data commons and advanced informatics in support of a clinical approach.
- The wheat genome is an example of a global endeavour, with national importance, related to traits and breeding, involves geo-spatial and other data types and research and commercial data interests.
- Oz Mammals exemplifies the challenges that will arise in supporting an Australian contribution to the global data commons and also engages a broad range of data types.
- Work already underway to better interconnect NCRIS imaging, microscopy, omics and eresearch infrastructures will inform a Mammalian phenomics case study.

The fourth use case was unable to be progressed due to the pressure of time. However the available expertise supported analyses for additional use cases in Environmental Metagenomics, ABRPI and Indigenous Genomics.

1.1 Methodology

The activity proceeded as follows:

- Consideration was given to the selection of use cases and their characteristics were reviewed in order to ensure a spread across relevant issues was present.
- For each use case:
    - A small group of expert researchers, known from previous activities in the relevant areas, were approached and agreed to help
    - A number of meetings with those researchers produced a summary document for each use case (see attachments)
    - The researchers contributed content, including editing of the summaries themselves
    - The relevance of the summaries to related research interests (such as generalising from wheat to grains) was reviewed with the researchers
    - Analysis of the needs expressed and their potential generalisation were verified in further consultations
- A detailed breakdown of the Oz Mammals case into 27 specific activities to be supported provided the next step forward
- An analysis was undertaken of nine readily available technology platforms against all of the Oz Mammal use case needs. It showed there are very few options for supporting Oz Mammals ‘out of the box’.
    - The technology platforms reviewed were: CYVERSE, CKAN, DNAnexus, GALAXY, GVL, GenomeSpace, Mediaflux and Seven Bridges.
    - A composite “OMICs” platform was added to the set, to achieve an additional example addressing most needs. The composite platform was previously developed for the antibiotic resistant bacteria consortium and is based on Galaxy, plus GenomeSpace and Mediaflux.
    - Some needs were not met by any platform, that is, specific features are missing from all

technology platforms, demonstrating that some development would be necessary. It follows that the total requirement is unable to be met in the short term.

- The technology platforms categorise their offering into products and services. Those categories have similarities across the platforms. The similarity of the categories provides a pathway to identifying key functionality for the ABDC.
- All of that analysis was then reconsidered for the other use cases (i.e in agricultural, medical and other environmental research scenarios).

2 Findings

2.1 Infrastructure Concept

The requirements in the use cases cluster into research performing aspects (on the left below) and data asset construction and use aspects (on the right), which correspond to aspects of Capability I and II. Different forms of data Integration arise in both.

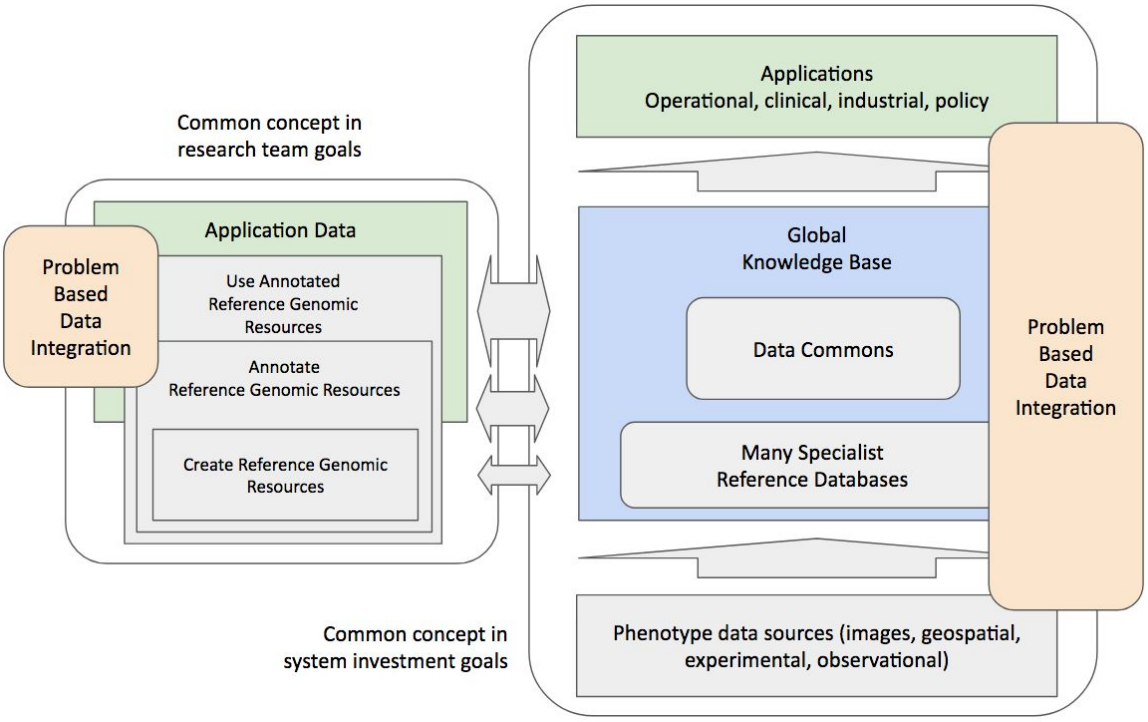

Figure 1

Notes:

- A small set of product and service categories can be identified that fulfill many of the requirements in the use cases considered (for the left component above) and would support related use cases.
- There are very few existing technology platforms that deliver end to end solutions in proven use.
- There are requirements in data integration and analysis pipelines that do not have solutions within the existing technology platforms reviewed.

2.2 Refining the Capabilities

The analysis undertaken in phase 3a is extremely helpful and the methodology appears sound. Further use cases should be identified and explored. Technology platforms that may arise (eg underpinning the NIH data

commons) need to be evaluated.

Capability I has a compelling benefit within the use cases reviewed. However, that benefit includes the realisation of population genomic resources and the identification and use of markers and traits, both of which rely on genomic and phenotype data integration.

The objectives to be set for an investment in Capability II are uncertain due to the number of issues and options included in the topic. However each use case identified steps that could be undertaken. The work suggests that many BPA Framework datasets will have integration opportunities associated with them.

A missing technology component involves the means to connect genomic resources to the data management systems outside research that aggregate and/or integrate other data sources. The variety of data environments to which genomics could add value cannot all be imported into the management of genomic resources. Therefore while Capability I would bring phenotype information into a genomic resource, Capability II needs to integrate a genomic resource into broader data solutions.

In terms of implementation, an ABDC national infrastructure will have four layers addressing: a) systems capability, b) foundational products and services, c) application specific platforms, and d) researcher training and support. The partners and suppliers of each layer should be expected to be different.

The infrastructure should align with international developments to the maximum extent possible. The key design decision affecting research hinges on the approach to b) and c). Very clear signals exist in each use case that a tailored platform is important to each research community. Experience overseas highlights the fact that such platforms can be composed from common services. It is likely that a strategy is needed to ensure onshore development resource integrates globally for b) and where it differentiates, does so by adding value around c).

Because a), digital systems infrastructure, is a key underpinning capability, the scale, configuration, location and access that best supports genomic resources needs to be articulated and implemented.

2.3 Recommendations

The proposal arising from the study is that the two arms of the ABDC should be refined into:

1. The provisioning for a Bioinformatics Commons; and
2. The exploration and achievement of broadly based Bioscience Data Integration.

The Bioinformatics Commons should have two components:

> 1a) support for the widespread use of genomic analyses on a bring your own data basis, and
> 1b) a more intensive resource allocation system able to support the development and use of reference genomes, population genomic resources, markers and traits.

As discussed in the report, the Bioinformatics Commons will involve well understood approaches to some data integration requirements in 1b).

The broader challenge around data integration is outward facing. A further understanding of these important future aspects of Bioscience Data Integration should be developed by trailblazing examples addressing key issues:

> 2a) Multi-modal data integration for progressing research challenges
> 2b) Integration with domain generated data holdings to aid decision making
> 2c) Integration with national and global clinical data holdings for health

2.4 Conclusions

The review of needs, and the investigation into the scale of services meeting similar needs in other jurisdictions, suggests that BPA acting in its own right and with its current participants could:

- Establish and operate a genomic analysis service on a BYO Data basis, and interconnecting the analysis services with global data sets; and
- Support a project based initiation of data integration activities in a similar manner to the existing BPA Framework Data initiative.

These investments would add value to Australian world class bioscientists at all levels of bioinformatics intensity and add momentum to directions the bioinformatics intense bioscientists themselves identify as priorities.

More is possible

The use cases show that significantly more value is available if the analysis service is extended to support long term data collaborative activities, such as the construction and use of population genomic resources. This additional capability would create a National Bioinformatics Commons.

Such a Bioinformatics Commons is crucially dependant on the long haul retention and use of genomic related data. Long term retention and collaboration around genomics data is a relatively new requirement presenting challenges to infrastructure within all leading research jurisdictions.

- The commitment of the Australian Research Data Commons and its future participants will be crucial in achieving the full value possible from a National Bioinformatics Commons.

A project by project data integration programme also leaves open the challenge of the long haul outcomes for the integrated data. A higher ambition would be a data integration facility, an Australian EBI or equivalent.

- A significant commitment from research intensive partners will be needed to achieve this outcome as well as the involvement of national HPC facilities and the ARDC.

The developmental difference in genomic resources available in human health compared to other domains and the different ethical constraints suggests either an approach involving a single facility where human health is a significant priority, or a human health focussed research data integration facility and one or two smaller facilities in the other domains.

Seizing the Advantage

It is important to realise that the twelve year commitment to NCRIS funding (through to 2030), and the significant allocation of that funding to Bioplatforms Australia and the Australian Research Data Commons (over the next four years and ongoing), together with the lift in supercomputing investment in NCI and Pawsey, opens the door to a significant competitive advantage for Australia. An advantage relative to jurisdictions without that funding and timeline, and an advantage that can be secured through these developments.

PART B

3 Use Case Overviews

3.1 Oz Mammals Genomics Consortium

Aims:

- Develop well-assembled genomes for at least 8 representative Australian marsupials
- Generate a comprehensive phylogeny for all Australian terrestrial mammals
- Build reference population genomic resources for threatened species

Key International Resources:

- Genome 10k
- Reference Vertebrate Genome project

Expert group:

- Rebecca Johnson, Director, Aust Museum Research Institute
- Anna MacDonald, Oz Mammals manager
- Dave Burt, Director, UQ Genomics
- Mark Eldridge, Principal Res Scientist, Aust Museum
- Craig Moritz, PI, Oz Mammals
- Andrew Pas, [University of Melbourne
- Janine Deakin, University of Canberra
- Jason Bragg, Royal Botanic Garden, Sydney
- Cameron Jack, Bioinformatics, John Curtin SoM

What would a good outcome look like?

Yr 1	An appropriately resourced and accessible platform with tools/pipelines for mammalian genome assembly, phylogenomics/conservation exon capture and ddRAD analyses, and tools for sharing/interpreting draft phylogenetic data.
Yr 2	An appropriately resourced and accessible system with tools/pipelines for automated and manual mammalian genome annotation, and web accessible tools for sharing draft assembly/ annotation or conservation data.
Yr 3	High quality, well annotated reference genome resources. An initial phylogeny. Analysis methods and outputs available to a broad audience. Web-accessible public resources exposing assembled, annotated reference genome including genotypic and phenotypic data; and experiment-focussed tools for genomic researchers to analyse and interpret WGS/RNAseq/other omic data.
Yr 4-5	Identification of markers of interest in conservation. Tools to support sharing of data in an accessible format in use by conservation organisations to support conservation-focussed decision making supported by genomics. A published and accessible integration and correlation of genomic and other modality phylogenetic information.

This use case is an example of very common research goals and processes and therefore useful for scoping infrastructure. Three general goals can be identified:

- Create a reference genome
    - In this case at least 8 but then growing over time towards 500, each requires
    - ~1000s of core hours on large memory systems, X?TB and 6-9 months bioscience FTE
- Annotate the genome

    - To create a resource that can assist the analysis and interpretation of individuals and which interconnects entries in phenotype datasets
- Identify Markers
    - To relate behavioural and observational data using more easily observable or measurable features that are known to indicate genetic variation.

In the case of Oz Mammals deploying CyVerse with sufficient resource would satisfy the infrastructure requirement. Proper connectivity and interoperation with Genome 10k and the Reference Vertebrate Genome project would significantly improve the overall value.

3.2 Australian Wheat Genome (Crops)

Aims:

- Exploiting genomic variation to allow understanding and directed improvement of wheat varieties in an arid environment/climate, for a range of stakeholders including researchers, breeders and farmers.

Key International Resources:

- The international Wheat Genome Sequencing Consortium reference genome (16Gbp - 5x size of human) hosted in Australia and France.
- ELIXIR and Earlham Institute (UK)

Expert group:

- Graham King (SCU) [applied res]
- Ute Baumann (UA) [bioinf]
- David Edwards (UWA) [applied res, bioinf]
- Rudi Appels (Murdoch) [bioinf and applied research, IWGSC]
- Jen Taylor (CSIRO) [bioinf]
- Delphine Fleury (UA) [applied res]
- Tony Bacic (LTU)
- Matt Hayden (LTU) [bioinf]
- Nathan Watson-Haigh
- GRDC: Pip Wilson

What would a good outcome look like?

Yr 1	Development of a high quality Australian-relevant wheat reference genome resource. Converge and integrate with international genome browser resource (http://wheatis.org).
Yr 2-3	Web-accessible public resources based on Australian reference genome including genotypic and phenotypic data. Experiment-focussed tools for genomic researchers to analyse and interpret WGS/RNAseq/other omic data.
Yr 4-5	Integration into pan-genome context (which is mostly based out of Europe currently). Widespread access to and expertise in genomic techniques for analysis.

This use case is an example where:

- a locally specific reference genome is needed to augment the value of a global reference genome; and
- integration with locally specific industry data is also needed.

Typical Workflow

A researcher or a breeder would start by sending a query - usually a short sequence from a molecular marker or a gene of interest - to the wheat genomics database. It will enable identifying the matching sequence on the

reference genome, and so its location.

A wheat genome database should then enable to retrieve information from the region such as:

- The full sequence of the genes with annotations.
- The Arabidopsis database (TAIR), can list mutant lines for each gene on the Gbrowser.
- Sequence variants to enable designing new molecular markers for fine mapping of a locus, or for molecular assisted selection, a procedure routinely used by wheat breeders. They are several SNP arrays available in wheat (800k SNP axiom, 90k SNP iSelect etc) that should be on the Gbrowser.
- Quantitative trait loci (QTL) are usually located at molecular markers. They often span large regions and the best way to locate them is to use the markers flanking the QTL peak (corresponding to the maximum probability of QTL location). QTL identifies the traits that are linked to the region of interest. QTL are identified in specific set of germplasm (also called genetic populations) that are useful to breeders particularly.
- Genomes sequences: gather information from other genomes such as the 16 wheat varieties shotgun sequenced through BioPlatform Australia (Edwards et al 2012 Journal of Plant Biotechnology).

Improvements

- The Wheat Initiative aims to develop a Wheat genomics database to bring all this together.
    - However, the project is huge, international and unfunded.
- Australian institutions have already developed some tools (eg Uni of Adelaide, La Trobe Uni).
    - Those could be put together to generate a first wheat genome browser to address Australian needs in research and breeding.
- A high-quality Australian genome reference. The current reference, Chinese Spring variety, although extremely useful is unlikely to give all the answers breeders need.
    - Wheat has a very plastic genome that differs greatly intra-species. Chinese Spring is distant from Australian modern varieties and researchers have already observed many divergences at the sequence level, especially for traits related to plant adaptation to climate.
- Wheat breeders use molecular markers more often than genes, mostly because the functions of most genes in wheat are unknown.
    - Breeders want to see their markers located on the genome and the relative position of QTL.

3.3 Childhood Cancer

Aims:

- A biomedical research and clinical application study of a pan-national data commons for genomic and human health data.
- Improved understanding and treatment of pediatric cancer.

Key International Resources:

- The Cancer Genome Atlas (2.5PB)
- NIH Data Commons: Vivien Bonazzi
- Kids First: Adam Resnick

Expert group:

- Paul Ekert (Murdoch Children's Research Institute)
- Adam Resnick (Kids First, CHoP)
- Mark Cowley (Garvan)
- Michelle Haber (Zero Childhood Cancer)
- Kathryn North (Murdoch Children's Research Institute, GA4GH)
- Roger Reddel (Director, CMRI)
- Vanessa Tyrell (Program Manager, ZCC)

- Multiple cancer institutes globally

The combination of NIH Kids First Data Commons, Seven Bridges and TCGA

What would a good outcome look like?

Yr 1	A shared informatics platform for primary analysis, incorporating currently used standard platforms (eg Seven Bridges, DNAnexus, Galaxy - see NIH Data Commons). Likely transactional model - data uploaded, analysed, downloaded back to user's data store. Significant compute resources underpinning. Example best practice workflows for human WGS, RNAseq, WES from international high profile initiatives (PanCancer/ICGC, Broad, Sanger, DKFZ, Genomics England,...).
Yr 2	Highly accessible research analysis platform incorporating access to selected open international core datasets, extending on #1. Initial core datasets selected through combination of merit for existing projects and best practice as implemented by NIH Data Commons, ELIXIR.
Yr 3	Shared data landscape for research, linked to analysis platform, including access to global datasets (not just open), allowing exploration of variants of unknown effect in as broad as context as possible.
Yr 5	Integrated data commons approach, linking major global omic datasets to Australian cohort datasets in a sophisticated analysis environment that is based on shared standards with US and European platforms.

What specific mechanisms and levels of engagement might we consider, where would they add benefit, and for whom? In what timeframe?

The order of development seems to be:

1. Infrastructure for world class per-genome clinical genomics analysis by co-developing and implementing on common standards
2. Improvements in best practice per-genome clinical genomics analysis through access to Kids First tools and methods
3. Accessing high quality global research data landscape as provided through Kids First or equivalent
4. Contributing to that high quality global research data landscape, and
5. Agreement on common annotation and curation platform and process
    - across ZCC (nationally)
    - across ZCC and Kids First
6. Real-time/dynamic clinical annotation based on active, global cohort analysis
7. Evaluate a shared precision-medicine clinical trial platform for patients across disease specific initiatives.

Ultimately, the development of a platform for the shared engagement of patients-as-partners via an integrative data ecosystem supporting researcher, clinicians, and their pediatric patients across the US and Australia.

4 Implications

4.1 Key functionality

This table identifies critical functions for the ABDC required by the use cases.

				Capability		
				I	II	III
Data Management	Sharing and collaboration	Web-accessible data-focussed portal with sharing capabilities	General	✿		
	Metadata and querying	Sophisticated queryable metadata data platform	General			
	Secure storage and sharing	Secure data management and sharing - to meet ethics approval, IP management and/or institutional policies	General			
Analysis	Tools and pipelines	User-managed suites of tools (BYO)	User	✿		
	Common tools and pipelines	Platform-managed suites of tools - structured and managed on national infrastructure	General			
	Sharing and interpretation	Collaboration tools	General			
Community Engagement	Data Commons	Large international datasets available in analysis environment	Application area	✿		
	Accredited analysis platforms	Participation in and adherence to standards for workflows and tools	Application area	✿		
	Data submission	Structured data and tools for interacting with international or national omic repositories	General	✿	✿	
	Data discovery	Structured data and tools for interacting with international or national non-omic focussed repositories	General	✿	✿	
Data Integration	Global data collaboration	Participation in global dataset annotation and building initiatives including standards, interoperability	Application area	✿	✿	
	Domain-specific integration	Generation of individual knowledge interfaces for multiple user groups	Application area		✿	
	Data integration	Integration across multiple modalities and interpretation	General		✿	
Capacity and Capability	Dedicated long term data resources	Strategic commitment + user commitments - accessible by individuals, consortia, facilities, institutions	General	✿	✿	✿
	Dedicated compute and working data resources	Strategic commitment + user dedicated commitments	General			
	Management, operating and support team	Located somewhere rather than dispersed	General and application			
	Training	Delivered regionally	General and application			

There are very few platforms that provide all of this functionality, as is depicted in the following spreadsheet from Attachment 4.

	Scope	Capability	Product category	Functionality	CYVERSE	OMICs Platform	Seven Bridges	DNAnexus	GALAXY	GVL	CKAN	Mediaflux	GenomeSpace
4	General	1a	Data management - capacity	Dedicated data resources (strategic commitment + user commitments) - accessible by individuals, consortia, facilities, institutions	CyVerse Data Store	VicNode storage	Yes	Yes	None	None	If provisioned in association with CKAN	If provisioned in association with Mediaflux instance	If provisioned in association with Mediaflux instance
5	General	1a	Data management - sharing and collaboration	Web-accessible data-focussed portal with sharing capabilities	CyVerse Discovery Environment (DE)	Mediaflux/DaRIS	Yes	Yes	None	None	Yes	Mediaflux itself	Yes
6	General	1a	Data management - metadata and querying	Sophisticated queryable metadata data platform	CyVerse Discovery Environment (DE) / iRODS	Mediaflux/DaRIS	Rudimentary	Probably	None	Rudimentary	Intended for datasets not items. No schema libraries	Mediaflux itself	No
7	General	1a	Data management - secure storage and sharing	Secure data management and sharing - to meet ethics approval, IP management and/or institutional policies	None	ASD compliant encryption at rest for data and metadata	ISO27K, HIPAA compliant	ISO27K, HIPAA compliant	None	None	None	ASD compliant encryption at rest for data and metadata	None
8	General	1a	Analysis - capacity	Dedicated compute and working data resources (strategic commitment + user commitments)	CyVerse Atmosphere (cloud) Powered by CyVerse (HPC)	Working storage: Nectar Allocation Compute: Galaxy-Australia and a GVL powered PathwayTools	Yes	Yes	If provisioned in association with a Galaxy instance	If provisioned to underpin a GVL instance	None	None	None
9	User	1a	Analysis - tools and pipelines	User-managed suites of tools (BYO)	Atmosphere	If user attaches via GenomSpace	Yes, If user wraps them	Yes, If user wraps them	Yes, if user wraps them	No	None	None	None
10	General	1a	Analysis - common tools and pipelines	Platform-managed suites of tools - structured and managed on national infrastructure	Discovery Environment	Yes - many tools in the managed Galaxy component	Yes, but only for human assembly, variant calling	Yes, but only for human assembly, variant calling	Yes - many tools in Galaxy	Yes - Galaxy plus SMRTportal, LOVD etc	None	None	None
11	General	1a	Analysis - sharing and interpretation	Collaboration tools	CoGe CyVerse Phylogenetics analysis apps: TreeView FastTree2	JBrowse, Mauve	Seven Bridges Genome Browser	Unclear	JBrowse / Apollo	JBrowse / Apollo	None	None	None
12	General	1a, 2	Data submission	Structured data and tools for interacting with international or national omic repositories	NCBI SRA Submission - BioProject Creationapp NCBI SRA Submission - BioProject Updateapp NCBI Whole Genome Shotgun (WGS) Data Submission CyVerse Data Commons	Could be developed	No	Unclear	None	None	Could be developed	Could be developed	None
13	General	1a, 2	Data discovery	Structured data and tools for interacting with international or national non-omic focussed repositories	None	None	None	None	None	None	None	None	None
14	Application area	1b	Data Commons	Large international datasets available in analysis environment	Cyverse Data Commons http://datacommons.cyverse.org	None	Unclear	Unclear	Possible	Possible	None	None	Possible
15	Application area	1b	Commercial analysis platforms	Participation in and adherance to standards for workflows and tools	unclear	Yes	Yes	Yes	Yes	None	None	None	None
16	Application area	1b, 2	Global data collaboration	Participation in global standards, interoperability, annotation and dataset building initiatives	BirdCoGe	None	None	Yes	None	None	None	None	None
17	Application area	2	Domain-specific integration	Generation of individual knowledge interfaces for multiple user groups		None	None	None	None	None	None	None	None
18	General	2	Data integration	Integration across multiple modalities and interpretation	BisQue - basic image analysis but probably not suitable for 3D morphometric analyses	None	None	None	None	None	None	None	None
19	General and application area	3	Training	Delivered regionally	CyVerse Wiki CyVerse Learning Center CyVerse Tutorials	OMICS Training Portal	Yes	None	None	None	None	None	Yes

4.2 A Bioinformatics Commons (Capability I)

The exploration of the use cases and the technology platforms that could support them suggests four layers need to be distinguished in the development of a Bioinformatics Commons.

1	Research use of platforms or products and services, and researcher support, in day to day research. ● Skilling researchers and the ready availability of technical support are critical needs.
2	Tailoring and configuration of application specific platforms and their operational support ● Because activities are expected to span many years, the tailoring of products and services into application specific collaboration platforms can be expected to lift productivity.
3	Identification of a small set of foundational products and services together with their international and on-shore development, maintenance and operational support. ● The platforms in 2 can be decomposed into more common products and services, and relevant products and services are under development worldwide. ● Because the product and services will be critical components of the many application platforms, and their maturity is on different trajectories, the ABDC will mature over time.
4	Access to scaled up compute, storage and connectivity facilities and services. ● The resources available will be a direct result of funding applied by NCRIS and by participants. ● A mechanism is needed that allows strategic and tactical funding to flow for general and specific purposes.

Of note:

- Relatively mature products and services exist in analysis and data management. Such tools support researchers and collaborating research teams, to perform research. Products and services to support the construction and stewardship of complex data resources of general utility are less mature.

- The primary sources of hard resource (layer 4) currently available include those at UoM and QCIF/UQ (developed through Nectar), NCI and commercial platforms such as Amazon. The suggestion by AARNet that it could sponsor the establishment of a new national data infrastructure provider is of interest. AARNet and the AAF would be important connectivity providers.

Key considerations:

- The layering suggests very robust service is require at 4 and into 3 with platforms at 2 allowing for more variable quality of service depending on the degree of development they involve. Clear demarcation between the layers in terms of performance delivery is likely to be needed.

- Transparency in the manner in which the capacity created by strategic funding is used will be important. The ability to apply participant funding to capacity used in designated platforms is almost certainly a key feature. Some agreement between a consortium of participants will be needed.

- The workforce available for layers 2 and 3 is difficult to identify and will need to be secured over many years. A group of institutions able to support that workforce need to agree a national framework for doing so and a means to manage resource allocation on a national impact basis.

- The creation of dependencies on infrastructure is a ‘lock in’ that requires durability of the

infrastructure to create the long term ROI for researchers. Long term commitment is essential.

4.3 Bioscience Data Integration (Capability II)

<work in progress>

The exploration of the use cases show that much more work is needed to understand Capability II.

There are many layers to data integration as one moves from left to right across Figure 1. The general flow can be set out as follows.

- For each species of interest, a primary underlying challenge relates to the construction of a reference genome, which can be used to make some progress and which is best treated as a reference data asset - international in scope.
    - It is the correctness of the assembled sequence and the identification of its features **and** the quantity and quality of interpretative data associated with those features that determines the usefulness of the reference genome. The inherent variability, redundancy and transposition in DNA sequences causes considerable complexity
- Significant further development is then required to construct a population based set of variations on the genome for the species, so that the interpretative analysis of a previously unseen genome becomes practical - creating a genomic resource of significant value.
    - The value of that resource increases with the number of members of the population that are sequenced. Generally a population genomic resource is a key reference data asset for much that follows - international in scope for globally present species, otherwise regionally scoped.
- As the quality of the genomic resource improves, the ability to associate phenotype information with genomic variations grows, leading to the identification of markers and other useful barcoding strategies, that in turn yield better identification of species variations in observational settings and hence improved phenotype information.
    - These observable or testable ‘markers’ (such as a molecule or species characteristic) can be associated with omic features and thus with other information associated with those features. Testing an individual for markers can therefore identify its relationship to a broad base of interpretative information that has been associated with that feature thereby connecting field observation or trial data with reference data and laboratory studies.
- The genomic resource can also be annotated with additional observational datasets to improve the understanding of the relationship between genome and phenotype data sets particularly within an experimental setting.
    - These include molecular data such as transcript analysis, proteomics and metabolite analysis and imaging and microscopy data.
- As the overall quality of the linkage from genomics to phenotype improves, new usages become possible including predicting characteristics of individuals within the population and the ability to understand traits and to breed or modify genomes for traits.

In this trajectory, genomic resources related to human health are all significantly further advanced than in other domains. And this gap is more likely to widen than close.

As an illustration, the Oz Mammals use case is centered to the left of Figure 1, the childhood cancer use case is centered strongly to the far right. To the extent that the ‘Problem Based Data Integration’ indicated on the right of Figure 1, may be better depicted as a third column as it is understood more thoroughly.

This would support the idea that the linking of elements in population scale data sets (of any modality), to omics features, can create new associations within those datasets.

- An omic derived cross linking of elements within broad holdings of population scale data can reveal previously hidden patterns - using omics features as a new index
- Such data may also be itself cross linked based on demographic, geographic, observational or other properties
- The new set of omic based relationships in combination with the existing relationships would be expected to yield new insight

The proposition is to refine Capability I and II as follows:

Capability I be named the Bioinformatics Commons and support data integration leading into population genomic resources, and the use of markers and traits. Therefore it is delineated into two parts:

1. A bring your own data facility that allows primitive analyses to be performed by researchers seeking to add genomic characterisation to their research activities and techniques.
2. The provisioning for collaborative data holdings (such as the BPA Framework Datasets), where communities build population scale genomic resources.

Capability II focus on the outward integration of genomic resources with application data sets, and be pursued on a project basis leading to the improved understanding that could define a facility investment. Even so, such projects would need to initiate multi-year or even decadal activities and so be of significant scale (and hence limited in number).