Submission
2016 National Research Infrastructure Roadmap
Capability Issues Paper

Name	**Associate Professor Jonathan Arthur**
Title/role	**Head of Bioinformatics**
Organisation	**Children's Medical Research Institute**

Questions

Question 1: Are there other capability areas that should be considered?

No. However, in defining focus areas, care needs to be taken to ensure the coordination of research capability that spans multiple focus areas. For example, 'omics-related disciplines (including genomics, proteomics, and bioinformatics) would fall across many disciplines within the focus area of Health and Medical Science, Underpinning Research Infrastructure, and Data for Research and Discoverability. It is critical to ensure effective linkages between these capability areas to support these disciplines.

Question 2: Are these governance characteristics appropriate and are there other factors that should be considered for optimal governance for national research infrastructure.

Yes. The comment regarding the presence of pervasive focus areas and the need to create governance systems that ensure coordination across capability areas is particularly important for 'omics-related disciplines.

Question 3: Should national research infrastructure investment assist with access to international facilities?

Yes.

Question 4: What are the conditions or scenarios where access to international facilities should be prioritised over developing national facilities?

Access to international facilities should be prioritised over developing national facilities where Australian researchers will have the same level of access to the facility as they would have if it was developed nationally, (assuming the same, or reduced, level of investment is required to access the international facilities,) i.e., there are no geographic, technical, social, political, or other barriers that would reduce the efficacy of the facility, at the same level of investment, for Australian researchers.

Question 5: Should research workforce skills be considered a research infrastructure issue?

Yes. Career paths need to be developed to enable individuals to pursue a professional career in the provision of research infrastructure. This is needed to attract and retain highly skilled technical staff in the area of research infrastructure, instead of encouraging these individuals to move into research itself (or out of the research workforce entirely to industry positions) in order to obtain a career development pathway. In addition, these highly skilled and experienced staff are imperative to enabling the effective ongoing operation and maintenance of the research infrastructure, thus ensuring the full benefit of the investment.

Question 6: How can national research infrastructure assist in training and skills development?

Question 7: What responsibility should research institutions have in supporting the development of infrastructure ready researchers and technical specialists?

Question 8: What principles should be applied for access to national research infrastructure, and are there situations when these should not apply?

Researchers who receive nationally competitive *government* funding for a project, program, fellowship, or equipment, should have full, free access to the relevant national research infrastructure required to support the project, program, etc., without having to further demonstrate merit (i.e., have the project assessed again for track record, significance, research plan, etc.) or otherwise compete for the required resources.

For example, if an NHMRC biomedical research project involving the whole-genome sequencing of several hundred/thousands of patients is awarded, the research team should not need to compete via a merit allocation scheme for sufficient disk space to store the genomic data or sufficient compute cycles to perform the analysis.

Furthermore, major national research initiatives, or the Australian contribution to international research initiatives, regardless of funding source (government, philanthropic, or other), such as large-scale genome sequencing or proteome capture, should also receive full, free access to the relevant national research infrastructure required to support the initiative.

Researchers with *non-government* nationally competitive research grant funding or without such funding, should be able to compete for an allocation from a portion of infrastructure capability set aside for this purpose on the basis of project merit, with due consideration and weighting in the merit allocation process given to those researchers who have already obtained funding.

It should also be possible to purchase an allocation, presumably at full cost recovery or market rates, using funds derived from an alternative source.

Question 9: What should the criteria and funding arrangements for defunding or decommissioning look like?

Question 10: What financing models should the Government consider to support investment in national research infrastructure?

Question 11: When should capabilities be expected to address standard and accreditation requirements?

Question 12: Are there international or global models that represent best practice for national research infrastructure that could be considered?

Question 13: In considering whole of life investment including decommissioning or defunding for national research infrastructure are there examples domestic or international that should be examined?

Question 14: Are there alternative financing options, including international models that the Government could consider to support investment in national research infrastructure?

Health and Medical Sciences

Question 15: Are the identified emerging directions and research infrastructure capabilities for Health and Medical Sciences right? Are there any missing or additional needed?

The identification of big health data as an emerging direction is correct. However, the full depth of need has not been captured in the Issues Paper. In particular, the focus of the Issues Paper is on the linkage of data sets, with a view to enabling better clinical health outcomes. This is critically important, but the need for research infrastructure capability extends much further upstream into the initial capture, processing, and analysis of the underlying data sets, especially omic datasets, for basic medical research and thence translation.

Section 5.2.3 is entitled “Omics”. However, the text of this section is very sparse and refers almost exclusively to genomics**. It is imperative for national research infrastructure to support all the major omic disciplines, including genomics, transcriptomics, proteomics, and metabolomics**.

Research infrastructure is required to support omic-based analysis from data acquisition through to linkage with clinical data for precision medicine applications in a clinical environment. There are several key elements of this infrastructure (addressed below), only some of which have been captured in the Issues Paper.

Large data storage. The infrastructure needs in this area have only been *partially captured* by the Issues Paper. See further comments at Question 30 and Question 33.

High performance computing. The infrastructure needs in this area have been *largely captured* by the Issues Paper, although some further comments are provided at Question 30.

Establishment of bioinformatics infrastructure. In order to effectively utilize the raw hardware provided by investment in high performance computing and large data storage, omic researchers typically need:

i) access to large reference databases such as GenBank, Ensembl, UniProt, the Cancer Genome Atlas, the Exome Aggregation Consortium database, etc., and
ii) an array of (typically) open source bioinformatics software packages.

At the present time, there is an excessive amount of duplicated effort within the Australian research community associated with the maintenance of these two essential enabling elements.

Large reference databases are regularly updated and researchers need access to the latest version of the database for new research and controlled access to older versions for the purpose of verifying and duplicating older analyses. A component of the national research infrastructure should include highly skilled technical personnel who can download and maintain requisite reference data sources on the national infrastructure for subsequent access by all researchers. This *reduces the duplication* in time, effort, data storage, and data transfer costs associated with each researcher doing this independently.

Similarly, open source bioinformatics software packages are also regularly updated and new packages are regularly introduced as a result of research in bioinformatics algorithm development. Most software packages are developed in a research environment, where funding is available for algorithm development, but not for documentation, user-interface development, quality assurance, or product support. A component of the national research infrastructure should include an investment in highly skilled technical personnel who can "translate" bioinformatics research output (in the form of new algorithms) into wider research usage. This would include facilitating centralized installation and version control of the software on the national infrastructure, development of appropriate documentation and training modules, and product support as required. This *reduces the duplication* of time and effort associated with each researcher performing the tasks independently and *optimizes the use* of the resources by *coordinating* these essential software product support issues.

These approaches are critically important given the current short supply of research bioinformaticians and computational (or quantitative) biologists. Provision of these elements of research infrastructure will enable omics research, critically dependent on bioinformatics, to proceed with smaller teams of bioinformaticians. Bioinformaticians developing new algorithms will be able to focus efforts on research and bioinformaticians collaborating with medical researchers and clinicians will be able to devote more time to the biological discovery and translation, rather than the engineering of underlying systems.

Establishment of standardized, centralised NATA-accredited data workflows. The Issues Paper refers to the need for omics data to be collected and analysed through using well defined clinical and industry standards. This capability is likely to be well beyond the capacity of most individual research groups and many research organizations. Even where capacity exists, in the current environment, there would be excessive duplication as each organisation attempts to build its own accredited workflows that in many cases will be essentially identical.

In order to facilitate access and enable research, the capability is required to construct standardized, centralized NATA-accredited data workflows for each of the omics. These would be built on the underlying national infrastructure identified elsewhere in the Issues Paper, including high performance computing, large data storage, data transfer, and data access/authentication infrastructure. This would enable researchers to quickly and easily achieve standardized processing of basic omic workflows, *reducing duplication* and freeing them to concentrate on specific, custom workflows as required by specific research questions.

As noted above in relation to bioinformatics infrastructure, this is also critically important as a result of the short supply of bioinformaticians, as it allows more optimal usage of this limited (human) resource, through re-directing research effort from building and maintaining data workflows, to utilizing standardized, central workflows to conduct independent and collaborative research.

Question 16: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Question 17: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Health and Medical Sciences capability area?

Environment and Natural Resource Management

Question 18: Are the identified emerging directions and research infrastructure capabilities for Environment and Natural Resource Management right? Are there any missing or additional needed?

Question 19: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Question 20: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Environment and Natural Resource Management capability area?

Advanced Physics, Chemistry, Mathematics and Materials

Question 21: Are the identified emerging directions and research infrastructure capabilities for Advanced Physics, Chemistry, Mathematics and Materials right? Are there any missing or additional needed?

Question 22: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Question 23: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Advanced Physics, Chemistry, Mathematics and Materials capability area?

Understanding Cultures and Communities

Question 24: Are the identified emerging directions and research infrastructure capabilities for Understanding Cultures and Communities right? Are there any missing or additional needed?

Question 25: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Question 26: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Understanding Cultures and Communities capability area?

National Security

Question 27: Are the identified emerging directions and research infrastructure capabilities for National Security right? Are there any missing or additional needed?

Question 28: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Question 29: Is there anything else that needs to be included or considered in the 2016 Roadmap for the National Security capability area?

Underpinning Research Infrastructure

Question 30: Are the identified emerging directions and research infrastructure capabilities for Underpinning Research Infrastructure right? Are there any missing or additional needed?

Data storage should be an additional consideration for Underpinning Research Infrastructure. This is, to some extent, implied in the Issues Paper, especially the Data for Research and Discoverability section. However, the emphasis in the Issues Paper is on the utilization of data, rather than the need for sufficient raw storage provision.

Omics based research is generating increasingly large amounts of data and, as noted in the Data for Research and Discoverability section, this needs to be maintained so that it is reliable and trustworthy. An essential pre-requisite for this is access to sufficient disk space to store the data. At present, competitive grant funding does not typically provide funding for data storage as a direct research cost. At the same time, associated infrastructure funding is insufficient to cover data storage costs as an indirect research cost. This leaves many projects using unreliable (but cheap) data storage solutions, e.g., keeping the data on external hard drives. Alternatively, projects may proceed at a smaller scale, with the potential to be underpowered, or not proceed at all due to lack of data storage space.

A central, large data storage capacity, freely available to all researchers, would eliminate these difficulties for researchers, enable more and higher quality research, and eliminate duplicate costs in setting up and maintaining reliable, trustworthy data storage systems. These could also be appropriate configured and co-located with national compute capability to minimize data transfer times, saving further costs.

Additional high performance computing facilities are also required, as noted in the Issues Paper. Consideration should be given to establishing specialist segments of the national HPC capacity to service particular disciplines, recognising that different disciplines utilize different algorithms that benefit from specific instrument configurations. It may be beneficial to consolidate the management of the Tier 1 HPC facilities and, potentially all associated data centric facilities, to achieve greater efficiencies, eliminate duplication, and ensure effective coordination of the multiple infrastructure aspects.

Access to data should be simplified and duplication eliminated. As noted at Question 8, separate applications for research funding and access to compute resources and large data storage should not be required.

Centralization of data storage and compute is predicated on easy, ultra-fast, data transfer. Thus, investment in better high capacity networks, as noted in the Issues Paper, is essential.

Question 31: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Question 32: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Underpinning Research Infrastructure capability area?

Data for Research and Discoverability

Question 33 Are the identified emerging directions and research infrastructure capabilities for Data for Research and Discoverability right? Are there any missing or additional needed?

Consideration needs to be given to the funding model for access to appropriate configured, reliable, trustworthy, data storage. In essence, approved research projects and major national initiatives should have effectively free access to the large data storage and compute capabilities required to complete the research. This could be managed through the introduction of multiple price points for access for different classes of user (e.g., government funded, non-government funded, unfunded academic, unfunded commercial), or a single price point with the relevant funding supplied at the time of project approval. In the latter case, care would be needed to guard against discrepancies between the funding supplied and what is actually needed to fund the compute and storage costs associated with the research.

Question 34: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Question 35: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Data for Research and Discoverability capability area?

Other comments

If you believe that there are issues not addressed in this Issues Paper or the associated questions, please provide your comments under this heading noting the overall 20 page limit of submissions.