Data and Services Discovery projects - Transformative Data Collections

Title

Database for Cardiovascular Functional Genomics

Approach

ARDC funding was requested to run an engagement program and workshop to bring together leaders from cardiovascular research, stem cell science, genomics, computer science, disease modelling, and drug screening to develop a strategy for the establishment, use, ongoing growth and funding support for a 'Database for Cardiovascular Functional Genomics'.

What activities did you undertake?
Engagement was performed in two stages. First, a preliminary questionnaire was circulated to relevant stakeholders in July/August 2019 to identify key challenges and areas of interest or need. The feedback received from this initial survey was used as a framework for planning and focusing a face-to-face workshop which was held on 27th August 2019 at the Victor Chang Cardiac Research Institute in Sydney. The workshop, cosponsored by the Australian Cardiovascular Alliance (AcVA) and Australian eResearch Organisation (AeRO), was arranged in two parts as detailed below:

1) **Information session.** Subject matter experts from cardiovascular science, ethics, data management, and FAIR data presented short talks to set the scene for workshop discussions. The list of speakers and topics covered was a follows:

- Prof Louisa Jorm (Director, Centre for Big Data Research in Health, UNSW) - *Big picture: A population-based linkage spine for administrative and research datasets*
- Warran Kaplan (Chief of Informatics, Garvan Institute of Medical research) - *IT infrastructure and data management: The Genome One Experience*
- Tom Honeyman (ARDC) – *An overview of FAIR data principles*
- Luc Betbeder-Matibet (Director Research Technology Services, UNSW) - *Data management classification approaches*
- Dr Tom Briffa (University of Western Australia) – *Ethics and data privacy*
- Dr David Elliot (Murdoch Children's Research Institute) – *Biobanking: Experience with the cardio-oncology registry*
- Prof Sally Dunwoodie (Australian Functional Genomics Network) - *Australian Functional Genomics Network*
- Prof Angus Lamond (University of Dundee) – *The human induced stem cell initiative (hipsci)*

2) **Roundtable session.** Discussions were held around three themes: Biological samples, the data resource and administration/governance. Each session was structured around topics highlighted in the initial questionnaire feedback. Templates used for each of the discussions are attached in the appendix to this document.

Which participants or collaborators were involved?
Organisations involved included:

- Australian Cardiovascular Alliance (ACvA);
- Australian eResearch Organisation (AeRO);
- Australian Functional Genomics Network;
- NHMRC Accredited Academic Health Research Translation Centres;
- Medical Research Institutes (VCCRI, Centenary Institute, Westmead Institute for Medical Research);
- Local Health Districts (Sydney, Northern Sydney, Western Sydney).

A full list of attendees, their positions and affiliations is attached as an Appendix to this report.

What o*utputs were produced?*
Responses from roundtable discussions were collated and assembled into a consensus agreement/plan that was presented for discussion and is in part summarised in this report. A steering committee of state representatives was selected who will formulate a more detailed strategy for establishment of the proposed collection based on workshop feedback. Specific short term priorities (see also section: Learnings) were identified as: 1) Establishment and publishing of minimum data standard; 2) Appointment of a governance team; 3) Identification of funding opportunities, 4) Engagement with providers (Cloud and/or NCRIS infrastructure) 5) Recruitment of admin/coordinators/developers/

FAIR

Findable
Current: Datasets are siloed in individual lab and/or researchers data servers. Metadata is often limited to information embedded in the title.
Proposed: Datasets will be combined in a domain specific repository (the core of this proposal) with persistent DOI. Basic description of data in the metadata.

Accessible
Current: The data that will form the transformative dataset is fractured and distributed around labs around Australia. The vast majority is not available online.
Proposed: Analysed/curated data would be centralised and made available through a web portal. Data submitted by individual labs would be embargoed until the primary data from the source lab was published.

Interoperable
Current: Data is mostly proprietary, unstructured and without metadata or data element descriptions
Proposed: Move to translate to machine readable (.csv) with machine readable minimum data standards. We would work towards publishing this standard as part of the proposed consortium.

Reusable
Current: Data is not licenced and has limited provenance information. Little consistency between sits or data types
Proposed: Creative commons licence embedded in metadata for data distributed through the proposed resource. Provenance information captured in text format.

A FAIR assessment spreadsheet is attached as an appendix to this report

Collaboration and coverage

A major issue we sought to tackle in this project is that data from basic, mechanistic cardiovascular research, particularly in relation to phenotypic data acquired from induced stem cell derived cardiomyocyte models of physiology and pathophysiology is typically siloed with the individual groups who produce it. Furthermore, few Australian institutions have the capability to acquire, store and manage the diverse data on the scale necessary for integration with genomic data, and this is a significant limitation to the broader research field. As part of our workshop we engaged key peak bodies, research institutions and local area health districts across relevant domains (details of organisation involved in preceding section) who all supported the concept and had representatives attend and contribute. The involvement of these key stakeholders bodies, together represents 100s of cardiovascular researchers throughout Australia who will benefit from the resource. Attendees were nationally diverse, covering New South Wales, Victoria, Queensland, Western Australia and South Australia. We also sought international input from domain experts including Prof Zam Cader (Stembanc.com, Europe), Prof Angus Lamonde (Hipsci, UK).

Sustainability

What agreements are in place to sustain the outcomes of the project?
The project has in principle support through the Australian Cardiovascular Alliance, whose executive director (Prof Gemma Figtree) as well as strategic flagship directors (Prof Louisa Jorm – Big Data and A/Prof James Hudson – Bioengineering) participated in the workshop. This synergy with the peak body representing Australian cardiovascular researchers will ensure that the project is aligned with national research priorities to maintain relevance and sustainability. On a practical level, attendees agreed to contribute to seeding the initial implementation of the database, including 600 datasets that will be acquired from iPSCs derived from patients with cardiovascular disease that will be gathered by the Cardiovascular Functional Genomics network (funded by Medical Research Future Fund).

What are the existing ingredients that enable sustainability?
The Victor Chang Cardiac Research Institute has recently established Australia's first high throughout phenotyping centre, supported by a $25 million investment from NSW government, that will be a primary source of data for this project. This level of investment demonstrates commitment that will support the proposed data resource. We have already leveraged this technology to win competitive grants worth $4M from NHMRC and ARC, demonstrating ongoing feasibility. In relation to the database structure, a linkage spline already exists (under administration of Prof Louisa Jorm – Centre for Big Data Research in Health, UNSW) that will facilitate linkage of phenotypic datasets with patient data and outcomes to maximise the impact of the proposed dataset.

What steps will you take to sustain the data collection and/or outcomes of the project?
A major limiting factor in establishing the data collection was the availability of funding for initiation of the project and this issue was discussed at length in the roundtable sessions. Suitable sources of funding identified were: NSW Office of Health and Medical research

cardiovascular capacity building fund, the Medical Research Future fund Mission for cardiovascular health (specifically in relation to ACvA Big Data and Bioengineering Flagships), the Ramsey Foundation, the National Health and Medical Research Council and the ARDC platforms fund. Priority recruitment areas were identified as: 1) 1-2 administration/dataset coordinators and 2) a Governance committee.

The lack of a minimum data standard was seen as the most significant hurdle to making the data resource FAIR and a publication was proposed to establish data standards for phenotyping platforms. In this regard, a birds of a feather session at the Research Data Alliance Plenary Session in Melbourne (March 2020) was identified as a suitable vehicle for development of such a publication.

Learnings

1. A significant gap exists around standardisation of data formats and metadata in our domain. A minimum information standard for phenotypic data (including eg. Electrophysiology, calcium handling) was proposed similar to MIRIBEL that described in Faria *et. al* (2018) Nature Nanotech. 13(9), 777-785. An opportunity exists here to interact with other international users via the platform manufacturers to establish a data standard.
2. The community requirement is not necessarily for raw data, which might require expertise and/or specialist software to interrogate. Rather there is a preference for analysed datasets and/or concierged access to raw data. The overall consensus was that analysed data relating to electrophysiology, calcium handling, metabolism and multi-omics should be integrated on a central platform with raw data linked and available from source labs on request.
3. For a national dataset to be built around phenotypic data from induced pluripotent stem cell derived cardiomyocytes and tissue, it is necessary to put in place standardised protocols for the generation of the biological specimens that are used to generate data (to ensure consistency and reproducibility of samples). Specifically this might include mandating cell reprogramming methods (Sendai virus), QC for PBMC preparation and storage (ACH^2), pluripotency measures (Pluritest) and differentiation protocols. In the latter case it was accepted that given the advanced status of individual projects using various differentiation protocols, this may be a site specific variable that has to be accepted.
4. The Cardiovascular community (and biological community in general) had little knowledge about what national infrastructure (eg. NCI or other NCRIS funded) might be available for hosting resources such as that proposed here. ARDC education around national infrastructure and/or facilitation of interactions with cloud platforms would provide a useful gateway into this domain.
5. ARDC outreach personnel to advise on establishment of transformative resources, or ARDC supported project personnel to work on initial setup/coordination would be an extremely valuable resource.

Impact

Research impacts of the project?

By creating a centralised, open resource for cardiovascular phenotypic and genotypic datasets, this project will lead to fairer and wider sharing of this data to facilitate discovery across a range of scientific disciplines. Collaborations will be facilitated across the spectrum of cardiovascular disease research and drug discovery in Australia, specifically though leveraging of the umbrella organisations supporting the resource including the Australian Cardiovascular Alliance, the Australian Functional Genomics Network, and Sydney Partnership for Health Education Research and Enterprise (SPHERE) cardiovascular clinical academic group. The researchers involved in the project are international leaders in their fields and as such, the collection will enable tier 1 publications from Australia's leading cardiovascular scientists. The project will bring together disparate fragmented data from labs across the country, will increase the power for discovery of the molecular mechanisms of disease, and allow development of new tools for discovery. Furthermore, the resource will be valuable asset for the pharmaceutical industry in early phase drug discovery, thereby promoting greater involvement of international biotech and large pharma in the research landscape within Australia.

Who or what might benefit from the results of the project (industry, community, government, wider public, etc)? Will you put in place pathways to ensure future impact?

The major community benefit from this project will be derived from advances in research into cardiovascular disease – Australia's biggest cause of mortality. Approximately 30 % of all deaths in Australia occur as a result of cardiovascular disease equating to one person every 12 minutes. At the government level cardiovascular disease is Australia's largest direct healthcare cost, including 11 % of all hospitalisations, at a cost of $8.8 billion dollar per annum. Furthermore, cardiovascular disease disproportionality affects our indigenous community, so is a particular burden in the Australian context. The research enabled by this data resource will lead to the new understanding of disease processes and identification of new cures and therapies to deliver community and government benefits and promote the development of a cardiovascular biotech sector in Australia.

Report prepared by: Adam Hill, Jamie Vandenberg, Matthew Perry
Date: 06/10/19