TD110 - Final Report - Data and Services Discovery projects - Transformative Data Collections

Title

Roadmap to integrate Indigenous genome assemblies into a national instance of the international human reference genome resource

Approach

This century will be marked by the transformation of healthcare from being largely treatment-based to being more predictive and preventative through availability of complex multidimensional genomic, epigenomic, transcriptomic and other biomedical data. At the heart of this transformation is the human reference genome.

The human reference genome is essential because: (a) it provides the standardized coordinate system to anchor information about function, clinical significance and population variation; and (b) it is the substrate used to align DNA sequence reads. Its central role as a standard and an essential component for delivering improved health outcome is recognized worldwide.

Despite two decades of research and improvement, it remains imperfect in important ways and is continuously being modified and improved. *Notably and most relevant in the Australian context, it doesn't adequately represent Indigenous Australians and other minority groups with ancestral links to Australia and people from the Pacific region*.

The National Centre for Indigenous Genomics at ANU has genome assemblies that can address this representational bias. In this project we addressed the question: What is needed to enable these genome assemblies to be incorporated into the evolving human genome reference resource and used by researchers and clinicians?

We investigated: the current state of the human reference genome: how it is being used in research and clinical practice; how it is likely to develop over the next few years; the current state of research and development in Australia; the importance of establishment of developments in this area within Australia.

We took three-pronged approach: (a) literature mining, (b) preliminary data analysis, and (c) discussions with the community of practitioners to identify essential requirements.

Following individuals have directly contributed to the generation of this report:

Following individuals have directly contributed to the generation of this report:

1. Australian National University
    a. Prof. Simon Easteal, Director, National Centre for Indigenous Genomics
    b. Dr. Hardip Patel, Bioinformatics Lead, National Centre for Indigenous Genomics
    c. Dr. Yu Lin, Group Leader, Computer Science and Genomics, Graph based representation of genomes
2. University of Adelaide

    a. Dr. Bastien Llamas, ARC Future Fellow, Indigenous reference genome to understand human diversity
    b. Dr. Yassine Souilmi, Post-doc, structural variant properties of human genomes using assembly graphs
3. University of Melboure
    a. A/Prof Stephen Leslie, Group Leader, Expert population geneticist
    b. Dr. Ashley Farlow, Research Fellow, Population genomics and bioinformatics

Face-to-face meetings and discussions were conducted with following individuals and organisations either directly or indirectly for the generation of this report:

1. Dr. Kate Birch (Head of Data and Technology) and Dr. Natalie Thorne (Lead specialist clinical genomics), Melbourne Genomics Health Alliance
2. Dr. Daniel McArthur, Lead for gnomAD and ExAC browser
3. Dr. Dan Andrews, Group Leader, Genome Informatics lead providing clinical genomics analysis services for Canberra Clinical Genomics
4. Dr. Kerstin Howe, Dr. Valerie Schneider, Dr, Tina-Graves Lindsay and Dr. Paul Flicek, Principal Investigators for Genome Reference Consortium
5. Andrew Howard, Cloud Team Manager at the National Computational Infrastructure
6. Dr. Michael Dobbie, CEO of the Australian Phenomics Network
7. Dr. Gareth Baynam, Clinical Geneticist
8. Dr. Andrew Lonie, Dr. Rhys Francis, Australian BioCommons
9. Dr. Marcel Dinger, Head of School, School of Biotechnology and Biomolecular Sciences, UNSW (current board member for NCIG and former CEO of GenomeOne)
10. Dr. Mark Daly and Dr. Taru Tukiainen, Institute for Molecular Medicine Finland
11. Dr. Shivashankar H. Nagraj, Advance Queensland Research Fellow, Queenland University of Technology
12. Dr. Brendan McMorran and Dr. Simon Jiang, Tiwi Island Renal Failure Study investigators with existing Indigenous genomics data, Australian National University
13. Professor Eric Stone, Director, Biological Data Science Institute, Australian National University
14. Dr. Paul Lacaze, Public Health and Preventive Medicine, Monash University

The project identified that the requirements to ensure Indigenous inclusion are far greater than anticipated. There is a critical unmet need for a national genome reference resource. The roadmap resulting from this project identifies how this critical unmet need can be addressed.

FAIR

This project identifies a roadmap for developments that need to occur within Australia to enable inclusion of ancestral diversity within a global human reference genome resource. The current the Human Reference Genome complies with all FAIR prinicples. Data are published, code and metadata are well curated by the international Reference Genome Consortium. The data can be readily accessed and freely downloaded. There are well established global standards for data description and for interoperability. The data are created for the express purpose of reuse. In many ways it is an exemplar of the application of FAIR principles.

The tradition of implementing FAIR principles is well established and will persist in future manifestations of the reference genome. However, this project identified the possible need to accommodate a level of access control, at least in the immediate future, to enable representation of Indigenous and other marginalised populations.

This requirement for access control underlines the need for Australia to be directly involved in international developments in this area, which currently it is not. Without this direct involvement the conditions required for appropriate inclusion of Indigenous populations may not be developed. The reference genome is an international initiative. Addressing Australia's lack of involvement in its development to date is critical to ensuring appropriate application of FAIR principles.

Collaboration and coverage

This project identified the need for an Australian reference genome resource to underpin future developments in important areas of biomedical research and to enable the transformative impact of precision medicine. The project has raised awareness of the need for this resource among contributors, collaborators and consumers, and as part of NCIG's ongoing engagement program, among Aboriginal and Torres Strait Islander communities and organisations.

More generally, the human reference genome is the resource central for all of genomics and much of biomedical research. It is a unifying resource that binds biomedical research to clinical practice. Its role in biomedical discoveries is increasingly important as the direct study of human biology underpinning health and wellbeing replaces much of work for which model systems have historically been required.

A constant feedback loop between research and its clinical application is developing, as evidenced by international projects such as Genomics England, All of Us Research Program, UK Biobank Project, FinnGen Project.

This close coupling will facilitate novel annotations of the human reference genome for functional properties and its roles in diseases and subsequent treatment and predictive strategies. A national facility focussed the human reference genome, which is the essential heart of these developments is essential.

It would enable a cohesive set of standards to emerge that will be applicable to research and health service delivery across Australia, that seamlessly integrate with the existing operations of healthcare delivery

The resource would coordinate efforts of funding and other government agencies (NHMRC, ARC, ARDC, NCI, BioPlatforms Australia) to ensure that their investments in research are appropriately integrated for translational outcomes that benefit of the Australian community.

It would reduce duplication of effort and provide a validated framework for translations of discoveries. Researchers would benefit from the use of a diversity inclusive standardized genome resource that would enable access to interoperable and reusable datasets. It would a unique point of contact for national and international engagements.

Sustainability

NCIG's motivation for initiating this project was to identify how existing long-read genome assemblies representing Aboriginal communities could be incorporated into the human reference genome. NCIG has strong strategic and operational support from ANU that makes its long-term sustainability highly probable.

However, the project revealed requirements that go well beyond NCIG's remit. Indigenous representation first requires the establishment of a national resource within which that representation can occur. The roadmap identifies a need for 12 month's seed funding to develop a business, strategic and operational plans for how such a resource could be established in a sustainable manner.

Appropriate agencies and business models would be developed. Importantly the resource will form an essential part of the fabric of national research infrastructure, a strong recommendation arising from the project is that it should be established with high quality research at its core. Obsolescence is a likely outcome if this is not the case, and is perhaps the greatest risk to sustainability.

Learnings

Three import findings were revealed through this project:

1. These is a serious lack of standardisation, both within Australia and internationally, in the use of the human reference genome. *Most of the genomic data is Findable and Accessible. However, due to the lack of standardization in the use of the human reference genome in biomedical research and clinical applications, large quantities of genomics data lack interoperability or reusability.*
   The proposed national genome reference resource will deliver enduring standards for genomics that will enable all genomic data to be interoperable and reusable. It will proactively work with international organisations such as the Global Alliance for Genomics and Health to ensure correct technological solutions are implemented to improve FAIRness of the data ensuring high-quality controlled access for maximal reuse.

2. Diversity inclusion, the key issue motivating this project, cannot be addressed without the development of the proposed national facility. NCIG has created the genome assemblies required to enable inclusion, but the national capabilities required to enable their use by researchers are substantial and well beyond the Centre's remit.

   Globally, lack of diversity is being addressed by creating haploid assemblies of nearly 300 individuals of diverse ancestry and transitioning from linear representation of the reference genome to a graph-based reference genome representation. Migration to a graph-based reference genome will be challenging. Analytical software, data representation standards, variation reporting standards and annotations of the genome will need to migrate to a graph-based system. These changes in linear to graph based representation will occur incrementally over the next 10 years with increased complexity

of the reference genome graph and sophistication in managing information content around the reference genome.

3. There is currently no Australian involvement in international human reference genome initiatives and there is not existing program aimed at developing capability in Australia, ARDC is well placed to play a key role in addressing gap, which is essential for Australia's future success in biomedical research and its clinical translation.

Impact

The impact of the project is long term. It identified a requirement - a national genome reference resource – that is essential to ensure Indigenous inclusion in the benefits of genomic research and precision medicine.

The need for a national reference genome resource goes much further, however. The fast pace of international developments and the volumes of human genome data for which a reference genome is required make establishment of a local managed resource essential to ensure that Australia remains internationally competitive in many important areas of biomedicine, and in the clinical and population health translation of that research.

The task of establishing such a resource will be a substantial undertaking. This project has identified that there is currently no initiative within Australia focused on achieving this objective. Furthermore, lack of involvement to date in international developments means that Australia has limited expertise in the area.

This project lays the foundation for the next step in the process, in which ARDC is well situated to play a key role: the development of a detailed budgeted plan in conjunction with national and international researchers and in conjunction with appropriate agencies.

Report prepared by: Dr. Hardip Patel and Prof. Simon Easteal, National Centre for Indigenous Genomics, The Australian National University
Date: 10/10/2019