Establishing a national Beacon version 2 network for real-time genomics data discovery

Aggregating data within the realm of rare diseases and cancers is paramount to gaining statistically significant insights into these diseases. To achieve this, the Australian BioCommons has launched the Human Genome Platforms Project (HGPP) to align and build on global standards such as the Global Alliance for Global Health's Beacon specification. This involves the collaboration of numerous institutes across several subprojects focused on Data Access Control (DAC) automation, federated identity management and data archiving.

Challenges

Human genomics data are large, complex and inherently sensitive. Hence, collections tend to remain siloed. The absence of a centralised platform for data browsing, querying and requesting access represents a significant bottleneck particularly to paediatric cancer research. Bringing genomic data from diverse institutions together requires a data sharing solution that has to be simple, federated and secure. Here, we address these issues with Beacon instances to house datasets and a Beacon network to query and aggregate data. Version 1 of Beacon was limited in sharing only genomic variations, whereas the version 2 data model supports cohort, individual and biosample data based queries, among others.

Methods

We were successful in implementing a prototype Beacon v2 network. We first instantiated our own beacon v2 instance using ELIXIR's reference implementation, in addition to ones created at our partner sites (UMCCR, QIMRB, NCI). We then established a Beacon network on our local servers by modifying the aggregator and registry code to support queries across v2 instances. CCI has one instance deployed with genotypes and phenotypes from CCI's Nature Med Publication (26, 1742-1753. 2020), with plans for another instance populated by CINECA 1Kg data in the future. Our collaborators are also exploring the concept of serverless Beacon instances.

Outcomes

We established a Beacon network which enabled us to query across four beacon instances, such as searching for Melanoma patients with a BRAF V600E mutation in patients younger than 20. From this, we were able to issue and resolve queries at both the genotype and phenotype level, demonstrating a major advancement in human genomic data sharing in Australia.

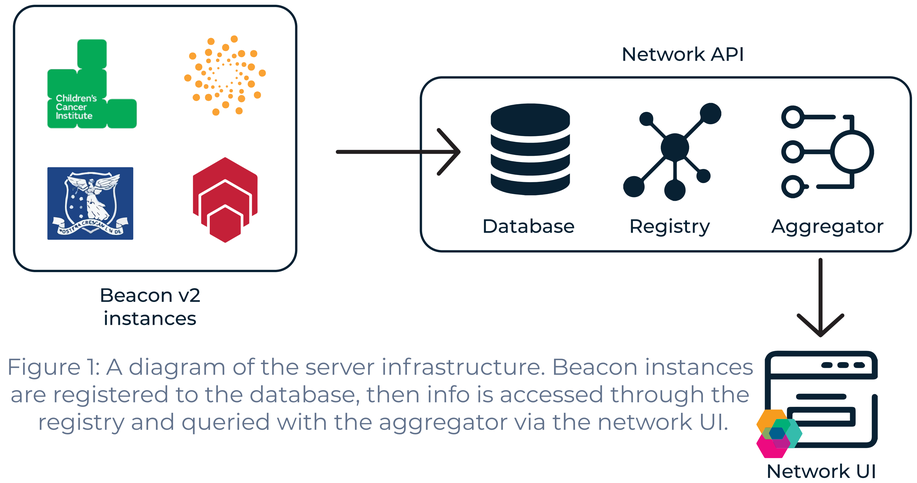

Figure 1: A diagram of the server infrastructure. Beacon instances are registered to the database, then info is accessed through the registry and queried with the aggregator via the network UI.

Conclusion

We propose that this multi-site implementation can be used as a template for genomic data sharing across the country in the spirit of exacting FAIR (Findable, Accessible, Interoperable, Reusable) data sharing practices.

Kamile Taouk (6), Angela Lin (6), Marie Wong-Erasmus (6), Mark Cowley (6), Tiffany Boughtwood (3), Jeff Christiansen (1), Joe Copty (7), Shyamsundar Ravishankar (7), Kylie Davies (2), Matthew Downton (9), Kelsey Druken (9), Ben Evans (9), Clara Gaff (8), Andrew Gilbert (5), Christina Hall (1), Matthew Hobbs (7), Oliver Hofmann (11), Jessica Holliday (1), Warren Kaplan (9), Ross Koufariotis (10), Sarah Kummerfeld (7), Conrad Leonard (10), Andrew Lonie (1), Heath Marks (2), Siobhann McCafferty (4), David Monro (9), Andrew Patterson (11), John Pearson (10), Bernard Pope (1), Shyamsundar Ravishankar (7), Florian Reisinger (11), Andrew Robinson (9), Victor San Kho Lin (11), John Scullen (2), Marion Shadbolt (1), Andrew Treloar (4), Jingbo Wang (9), Scott Wood (10), Mustafa Syed (6).

1. Australian BioCommons 2. Australian Access Federation (AAF) 3. Australian Genomics 4. Australian Research Data Commons (ARDC) 5. Bioplatforms Australia 6. Children's Cancer Institute (CCI) and the Zero Childhood Cancer Program (ZERO) 7. Garvan Institute of Medical Research 8. Melbourne Genomics Health Alliance 9. National Computational Infrastructure (NCI) 10. QIMR Berghofer Medical Research Institute (QIMRB) 11. The University of Melbourne Centre for Cancer Research (UMCCR).