Genome Alberta RP3 – BioNet Alberta
Pillar 3 – Tool Development Competition
Agriculture Stakeholder Consultations
Friday, September 20, 2019

Summary Report

Purpose

BioNet Alberta's mission is to strengthen bioinformatics and computational biology (B/CB) capacity in the province. One mechanism by which this will be accomplished is through a funding competition for the development of B/CB tools and pipelines that address currently unmet needs for the agriculture sector of Alberta. On September 20, 2019, in Lethbridge, Alberta, Genome Alberta held a stakeholder consultation meeting with representatives from Agriculture & Agri-Food Canada (AAFC), Alberta Agriculture & Forestry (AAF), and other provincial stakeholders to share current B/CB challenges in Alberta and inform the development of the program Request for Applications (RFA). Specifically, Genome Alberta sought input on the gaps and challenges in 'big date' analyses in the province, what resources already exist, and how the tool development competition under Pillar 3 of BioNet Alberta may be developed to address these issues. This summary report details the outcomes and highlights from this meeting.

Attendees:

Dr. Steven Morgan Jones, Genome Alberta Board of Directors Chair (Meeting Chair)

Agriculture and Agri-Food Canada (AAFC)

- Rodrigo Ortega-Polo (Lethbridge)
- Arun Kommadath (Lacombe)
- Francois Eudes (Lethbridge)
- Raja Ragupathy (Lethbridge)
- Xianqin Yang (Lacombe)
- Rob Gruniger (Lethbridge)
- John Laurie (Lethbridge)
- Rahat Zaheer (Lethbridge)
- Miles Bushwaldt (Saskatoon)
- Etienne Lord (Quebec)
- Wayne Xu (Manitoba)
- Carolyn Amundsen (Lethbridge)
- Kristin Low (Lethbridge)
- Andre Laroche (Lethbridge)
- Devin Holman (Lacombe)

Alberta Agriculture & Forestry

- Kirill Krivushin
- Mark Hicks
- Saida Essendoubi
- Robin King

Stakeholders

- Reynold Bergen, BCRC
- Dave Moss, CCA
- Tanya McDonald, Lakeland College
- Josie Van Lent, Lakeland College

Observers

- Eric Merzetti, BioNet Alberta Network Manager, University of Lethbridge
- Gijs van Rooijen, Chief Scientific Officer, Genome Alberta
- Niall Kerrigan, Senior Program Officer, Genome Alberta
- Ryan Mercer, Research Program Manager, Genome Alberta
- Lori Querengesser, Economic Development, Trade, and Tourism, Government of Alberta

Identified Issues for Federal (AAFC) and Provincial (AAF) Agriculture Departments

People

Similar challenges exist for both AAFC and AAF when it comes to B/CB. In alignment with previous consultations and discussion with the broader community, **People** were identified as the biggest hurdle to efficient ‘big data’ analyses within the Alberta agriculture research system. Current provincial Bioinformaticians are stretched thin with numerous request for B/CB data analyses from the research community. Many individuals providing B/CB support are either Biologists that have learned data science skills, or Computer Scientists that have some biological understanding. Compounding this strained personnel capacity is that AAFC researchers are not fully aware of provincial resources in other organizations, institutions, departments. This results in only a few ‘known’ provincial Bioinformaticians being tasked to the bulk of analyses within these government departments. If these individuals are unable to assist due to overcommitments on other work, then the analyses may be outsourced to other provinces.

BioNet Alberta is set to tackle some aspects of the shortage in skilled personnel through training workshops and augmenting Alberta’s B/CB profile to aid in recruitment. The Network will also be developing a provincial asset map to inform the research community with resources and expertise are currently available. However, the current demand for B/CB skills is strikingly high and, therefore, alternative strategies for training and staffing need to be explored.

Data Management & Storage

AAFC and AAF maintain their own internal datasets through various data programs, processes, and policies. Another challenge for these agriculture departments is efficient and effective **data storage and management**. Different sequencing platforms and technologies create diverse forms of raw (structured or unstructured) data. A lack of data standardization results in a significant drain on personnel time, and barriers to interoperability between departments and institutions prevents efficient data exchange and analyses. Data management portals that can communicate (e.g. cross-talk) with various platforms would allow for more robust analyses and comparison of different data sets, while also allowing access to researchers in different locations. Curated databases that provide an amalgamation point for genomic and associated metadata are critical to efficient data analyses and deriving true biological signals from large data sets. Incomplete records, references, and linkages will not result in maximal benefit arising from the analysis of generated genomic data. Database development can be a challenging task for current provincial bioinformaticians, and additional support towards the creation of data management tools could alleviate some of these hurdles.

Similarly, when genomic data are generated, the raw form (e.g. sequence reads) does not provide any relevant biological information. Genome assembly and other downstream analyses create additional data files that can be quite large and require external storage (on-site or cloud-based). As the rate of genomic data generation is expected to continue increasing, there will be a critical point in which on-going storage of all data files will be a computational and costly challenge. Perspectives differ on the value of storing data from each step of an analyses, and questions remain around the utility of older data sets generated by outdated technology platforms.

Computing Power & Infrastructure

The last major issue related to B/CB in these government departments involves **computational power and infrastructure**. Both groups acknowledged lags in data transfer rates creates challenges in retrieving data sets from off-site storage (e.g. cloud) or exchanging data with external collaborators. While some bioinformatic analyses can be conducted on regular computers, large, complex, and potentially multi-omics analyses requires significant processing power that would necessitate using a high-performance computing cluster. AAFC-derived data must stay within the federal system where internal computational resources are lacking; there is a single computational cluster for use in Ottawa, but data transfer rates and lengthy access queues are challenges. As government entities, neither AAFC nor AAF can access the national computational resources from Compute Canada. In contrast, collaborative research partnerships with academic researchers allow these government entities to share their data with outside entities and make use of computational power that exists externally.

Feedback from Other Stakeholders

Representatives from the Canadian Cattlemen's Association (CCA) and Beef Cattle Research Council (BCRC) were both supportive of the BioNet Alberta strategy to deliver more robust approaches to handling the masses of data in the age of Smart Agriculture (Smart-Ag). There was encouragement to capitalize on Alberta's capacity and strengths in machine learning and integrate these computational approaches to solving challenges in 'big data'. While these groups focus on production-oriented outcomes, there was also suggestions to keep in mind how biological data could be combined with production-level information as the livestock sector continues the rapid adoption of genomics technologies.

B/CB Tools – More than code

Early in the development of BioNet Alberta's Pillar 3 strategy, it was believed that novel analytical tools (e.g. computational codes) would be the key to maximizing benefit of genomic data sets. Feedback from stakeholders indicated that many (almost too many) B/CB tools are already freely available and could be implemented within new pipelines or adapted and improved for specific applications. It is also evident that '*tools'* may be more appropriately described as '*Enablers'* that allow users to overcome limitations in understanding, analyzing and drawing conclusions from collected datasets. This may include coding scripts, algorithms, data pipelines, user interfaces, databases and platforms, and even strategies (e.g. statistical consultation) for experiment design.

Stakeholder suggestions for reproducible, scalable, and flexible enablers that could be developed under Pillar 3 of BioNet Alberta are listed below:

- 'Plug-n-play' bioinformatics; developing easy-to-use and reproducible pipelines and workflows where the most appropriate existing tools can be integrated depending on desired outcomes and available data.

- Pre-developed pipelines or packages (e.g. R packages) for specific applications

- Re-purposing common B/CB tools from other sectors into applications for livestock and crops

- Comparative analyses of available method accuracy and efficacy to reduce ambiguity in selecting the most appropriate for a given application

- Robust data mining tools and predictive models for detecting biological signals in existing and/or large data sets

- Applications for improving referencing genome mapping, assembly and variant calling on third generation (e.g. long-read) sequencing data, including microbiome metagenomics

- Data integration and statistical validation to reduce uncertainty in the analyses of multi-omic (genomic, transcriptomic, epigenomic) data sets; how to infer biological relevance after layering complex data

- Development of user-friendly data management systems for improving platform cross-talk and data standardization, interfacing, exchange, and analyses.

- Approaches to increase computational efficiency and reduce power needed for complex analyses; reduces the need to access high-performance computing

- Tools for phenomic and meta-data analyses, including quantitative and qualitative imaging data; digitization of phenomic data

- Deep learning and/or machine learning of structured and unstructured data for genomic predictions

Non-computational enablers for Smart-Ag:

- Guidelines for determining appropriate statistical power calculations required for genomic research studies; involvement of biostatisticians to ensure big data analyzed is robust enough to infer meaning

- Addressing policy issue (e.g. data privacy, storage, exchange) barriers that prevent maximizing the utility and application of big data

Pillar 3 Program Suggestions

The stakeholder feedback for how the Pillar 3 program should be structured included suggestions to allow for national collaboration. This would potentially bring new insights from abroad into the Alberta landscape. There were also suggestions to ensure the BioNet Alberta Asset Map was available for researchers to see what resources already exist in the province upon launch of the RFA.