Genomics Digital Research Infrastructure to Support Canadian Competitiveness

Genomics is a Data Science

Genomics involves a variety of high-throughput technologies that measures how life works, whether in a single cell or in a complex ecosystem. Throughout this whitepaper genomics broadly refers to the collection of these technologies including specific domains such as nucleic acid sequencing, proteomics, metabolomics and bioinformatics.

Genomics is a big data science with multi-sectoral applications: the genome of a person, lentil, spruce, or salmon contains billions of data points that can be mined for actionable insights that will improve our health, our economy, our food and our environment. Some researchers have suggested that genomics could produce more data by 2025 than ultra-high-resolution astronomy sites or platforms like YouTube and Twitter. The accelerating pace of sequencing data generation is made possible by the rapid co-evolution of enabling technologies and software being applied by a specialized talent pool that interprets and extracts value from that data. It is now possible to sequence a human genome with six billion pieces of information in mere hours at a cost approaching $100, compared to 20 years ago at an estimated $300 million.

Canada is recognized as a global leader in genomics research, the impact of which has led to improvements in many areas, for example: disease diagnosis and treatments, pathogen surveillance, food security, natural resource planning, and climate change. Research, technology development and enabling digital platforms in Canada have been supported through significant public sector investments over the past twenty years. For example, in fiscal year 2019 Genome Canada, the regional Genome Centres, and the Tri-Council together invested approximately $785 million in genomics-related research and technology development, while in the past five years the Canada Foundation for Innovation has invested $117 million in genomics infrastructure and operations.

However, Canada's genomics digital research infrastructure is reaching a tipping point. The genomics research community has expressed the need for substantial investments in digital research infrastructure to secure and sustain its capacity to be a leader in genomics research, and to attract, retain and train personnel – and hereby maintain its competitive edge – in this period of deliberate planning for the post-COVID recovery. Our current systems for collecting, curating, storing, and analyzing data in all life sciences sectors need review. The pandemic has spotlighted barriers and gaps uncovered by initiatives such as the Genome Canada-led Canadian COVID Genomics Network (CanCOGeN), revealing a data landscape where fragmentation and inconsistent standards impede our efforts to create value and realize innovative solutions that address domestic and global challenges.

In order to capture the current state of the digital infrastructure needs of the genomics research community Genome Canada, in partnership with the regional Genome Centres and the Canadian Institutes of Health Research (CIHR), interviewed over 40 academics in a targeted consultation process to inform this white paper. Capturing the current state of data storage and computational needs and challenges unique to the genomics community, we identify common themes and offer recommendations based on our appreciation of the future needs of the Canadian genomics research community as the New Digital Research Infrastructure Organization (NDRIO) continues with its strategic planning process.

The Canadian Genomics Landscape

Genomics, by its very nature, relates to multiple interconnected technologies and economic sectors. Each layer of the genomics landscape emphasizes the scale, complexity and need for dedicated genomic digital research infrastructure.

Multi-Omics. The measurement of DNA, RNA, proteins and metabolites is applicable to all living species. A near infinite number of applications exist such as the study of microbial populations (e.g., metagenomics or microbiomes) that contribute to disease and can be engineered to produce various products and medicines. Genomics can be measured at multiple scales, including across space (e.g., cells and tissues), time (e.g., embryonic development and growth) and perturbations, (e.g., disease and changes to the environment). Each form of genomics has its unique data challenges ranging from sheer quantity to standardization and integration. Increasingly, research projects require complete genomic profiles across different scales of time and space. The necessary computational data infrastructure to support these data structures is not trivial and requires substantial shared resources and highly qualified personnel.

Multi-Sector. Genomics has become essential to multiple economic sectors and research domains—agriculture, fisheries, natural resources, the environment—and Genome Canada is recognized globally for its multi-sector model. Applications include breeding for better food traits, pathogen and pest surveillance, and biomanufacturing of sustainable products and medicines. The recent explosion of the field of precision agriculture is an additional opportunity for Canadian leadership supported by digital initiatives, such as DivSeek (see insert). Genomics is also a critical toolbox in fighting climate change through areas such as Arctic surveillance, bioremediation of tailing ponds and developing heat- and drought-resilient agriculture and forests. The breadth and complexity of the application of genomics requires a purpose-fit foundational digital research infrastructure.

DivSeek International aims to facilitate the generation, integration, and sharing of data and information related to plant genetic resources (PGR). DivSeek provides a forum for members to come together, it acts as a 'hub' to connect and promote interactions between its members, other researchers and organizations to develop and share knowledge and best practice. DivSeek seeks to increase the efficiency of projects undertaken by its members, to limit redundancy and increase the availability of digital data associated with PGR to a wider community.

Context. Significant value is added to genomics research through the metadata that describe it and the contextual information that connects it to the environment. In many cases these contextual or linked data are larger and more complex than the underlying genomics data, for example: deep phenotyping in rare disease, drone imaging in agriculture, geographic information for pathogen surveillance and climate data for management of fisheries. The exponential growth of data collected in the deployment of inexpensive sensor networks exposes both an extreme challenge and opportunity. A national digital research infrastructure strategy that champions genomics provides strong and effective linkages between genomic and contextual data sources to facilitate and accelerate genomic research and application. The data access, sharing, and security challenges addressed by national COVID-19 initiatives like CanCOGeN emphasize this need for building and sustaining linkages in our data system.

Platforms. Genome Canada currently supports ten Genomics Technology Platforms—five in partnership with the Canada Foundation for Innovation's Major Science Initiatives program (CFI MSI). These provide extensive genomics infrastructure dedicated to providing researchers with access to a wide variety of advanced genomic technologies and extensive primary data analysis and bioinformatics support. Additionally, there exists several hundred university and research institute genomic cores generating data. These platforms depend on large internal high-performance computing (HPC) and cloud solutions to generate, analyze, and store genomics data. As keystone infrastructure in the Canadian genomics ecosystem, these platforms must be included in an ongoing needs assessment of Canadian digital research infrastructure.

WHAT WE HEARD FROM THE GENOMICS COMMUNITY

We are in the early days of a Bio Revolution, where advanced biosciences and biotechnology will fundamentally transform our lives. The McKinsey Bio Revolution report estimates that life science applications could have a direct economic impact of up to $4 trillion a year over the next 20 years. This is likely an under-estimate given that 2020, despite the pandemic, has already experienced the biggest biotech IPO market in history. With its abundance of natural resources, Canada is well positioned for leadership in the emerging bioeconomy if we act decisively in our transition to a sustainable and profitable green economic paradigm.

As a cornerstone of the Bio Revolution, genomics needs intentional investments in both the science and the digital infrastructure that supports it. Based on consultations with the genomics community, four themes and associated recommendations were identified. They included the need for:

1. **Streamlined and equitable access** to digital infrastructure.
2. **Dedicated and secure computing resources** for the analysis and storage of datasets.
3. **Canada-wide data sharing** through appropriate infrastructure and data governance.
4. **Specialized tools and talent** to derive value from research.

Streamlined and Equitable Access. The current funding model for digital research infrastructure limits access to computing resources due to regional disparities and burdensome administrative requirements that include complex application processes and redundant technical review. Barriers to accessing a shared digital research infrastructure have resulted in the establishment of independent local computing infrastructures that are not interoperable. Our consultation revealed multiple examples of project data that span lab servers, institutional high-performance-computing clusters, regional academic clusters and multiple private clouds. *Simplifying the process to access computing will increase efficiencies and establish a more equitable landscape sharpening the conversion of investment dollars into real-world impact.*

> *Recommendation #1. The coordination of research funders in support of a funding model for digital research infrastructure with reduced barriers to access.*

Dedicated and Secure Computing Resources. A consistent message from consultees was that existing infrastructure, although incredibly valuable, offers insufficient computing power for the processing and analysis of genomics data. Alternatives such as private sector clouds potentially offer more flexibility and functionality, but the costs are not sustainable for the volumes of data and the required analytics. It was highlighted that genomic technology platforms need dedicated and secure computing to deliver high-quality, short turn-around datasets. Alternatively, individual researchers need discretionary access to intermittent bursts of significant levels of computing to analyze data in short periods of time. Consultees underscored that competing for computing with users in other high-demand areas such as physics and astronomy is not ideal. Moreover, if Canada wants to lead in innovation sectors that provide significant competitive advantages - artificial intelligence (AI), for example – we need to build and provide access to appropriate technology such as graphical processing units.

The high cost of long-term data storage creates a significant barrier to innovation by limiting the sustainability of the data assets and curtailing their useful lifetime. Resources are needed for archival beyond the duration of the initial project to allow for secondary and tertiary mining of the data such that initial investments continue to create returns in terms of new discoveries and solutions. Given the absence of Canadian alternatives genomic data generated with public support must often be deposited in US and European repositories – creating barriers to broad access and preventing us from generating globally-competitive national data assets. *Providing researchers access to dedicated computing and secure storage will allow Canada to house genomic data within its borders and participate in increasingly data-intensive scientific discovery and intellectual property (IP).*

Recommendation #2. The creation of a Canadian genomics digital platform that blends the best of private-sector cloud flexibility with public sector subsidized resources and oversight. For example, a Canadian Genomics Cloud that respects principles of data federation, modularity and interoperability.

Canada-wide Data Sharing. Consultees agreed that there are significant challenges to sharing genomic data within Canada as well as internationally. The challenges can be broken down into three areas: (1) regulations and policy; (2) standards; and (3) technical solutions for secure data repositories.

Regulatory and policy barriers to data sharing are in part due to the institutional and provincial differences in the approach to data stewardship, particularly in healthcare. The legal, privacy and regulatory frameworks in place to protect highly sensitive data have been described as poorly harmonized, which limits the extraction of value from the data. One consultee highlighted an extreme recent example: the need for material transfer agreements to exchange data within their own research institute.

Genomic data standards are an essential requirement for data sharing and to promote data being Findable, Accessible, Interoperable and Reproducible (FAIR). Canada is an international leader in this area largely due to the efforts of the Global Alliance for Genomics and Health (GA4GH), a not-for-profit incorporated in Canada and headquartered in Toronto (see insert). Consultees advocated that a Canadian digital research infrastructure strategy should strongly support genomics standards as a flagship and model for other research domains where they are less mature or fractured.

The need for secure genomic data repositories or safe havens was discussed in multiple contexts. Secure repositories facilitate data sharing or data visiting through controlled access to systems that are trusted and approved by relevant data custodians. For health-related data, CIHR is leading the development of a Canadian genome library (see insert).

The GA4GH convenes the research and healthcare communities to agree on common methods for collecting, storing, transferring, accessing, and analyzing data in order to overcome these barriers, enable broad data sharing that transcends the boundaries of any single institution or country, and fulfill the human right of everyone to benefit from the advances of genomic research, as mandated by the Universal Declaration of Human Rights (Assembly 1948). The GA4GH is a standard setting body that has been incredibly successful in defining how genomics data can be leveraged through implementation within member organizations. Moreover, Canada has multiple leadership roles and driver projects that are a testament to Canadian expertise and international leadership in genomics.

The CIHR-led Canadian **human genome library** is a resource that will facilitate the integration of human genome sequencing efforts. Such a library could maximize the sequencing investments made in the publicly funded Canadian healthcare system. Ultimately, a Canadian human genome library will contribute to the establishment of a 'real time' learning precision healthcare system. The main objectives of this library would be to:

- establish a single-entry point to access human genome analysis, and to set standards for human genome information, accuracy, access, and use;
- establish mechanisms for genome research to inform the clinical context to enable timely and cost-effective precision diagnosis and treatment;
- enable feedback to healthcare practitioners and patients when their genomic information reveals potential health risk; and
- enable more robust, complete, and accurate clinical trials making Canada a country of choice for these types of studies.

In addition to these challenges, it was emphasized that the genomics community is more than universities. Digital research infrastructure should therefore support data sharing among the public and private sectors to convert investments in innovation at universities into economic benefits and impacts for Canada. *A connected and interoperable Canadian digital research infrastructure will facilitate the creation of national digital assets that are purpose-fit to address and adapt to societal challenges and missions.*

> *Recommendation #3. To support the development of secure genomic data repositories and data safe havens with appropriate data governance and standards.*

Specialized Tools and Talent. Consultees across all sectors identified genomic data management and analysis as the most time-consuming and technically challenging aspects of a research project. At present, there is a gap in the availability of specialized tools, such as robust bioinformatic workflows, and skills to overcome this challenge. This gap is forecasted to grow as genomic data exponentially increases in volume and complexity. Additionally, the value of a digital research infrastructure could be increased if expertise were available to help use existing tools and support training on advanced analytics to exploit new methodologies, such as AI. *A renewed digital research infrastructure that incorporates the needs of the genomics research community will facilitate the creation of world-class analytical tools and talent, branding Canada as an attractive and well-supported environment for innovation.*

> *Recommendation #4. To prioritize the inclusion of specialized genomics tools and resources for talent development within a Canadian digital research infrastructure, for example by the inclusion of diverse genomics end users in NDRIO committees.*

CONCLUSION

The Canadian genomics research community, Genome Canada, the regional Genome Centres and CIHR applaud the foresight that NDRIO is demonstrating through this consultation process and look forward to future engagement. NDRIO has an opportunity to build a Canadian digital research infrastructure that will ensure the privacy of the data and the individuals providing it to drive high-impact and internationally-competitive genomics research for the benefit of all Canadians.

The ability of the genomics research community to generate, access, analyze and store their data in an efficient, fair, secure and equitable manner is critical if we are to be successful in translating the knowledge generated through research into innovating technologies and positive social impact. We commit to working with NDRIO and other partners to advance essential data assets and analytic tools and develop and retain a diverse generation of skilled researchers and innovators ready to define the future for Canada and improve our health, environment and economy.

Submitted by Genome Canada, the regional Genome Centres and the Canadian Institutes of Health Research on behalf of the Canadian genomics research community.