Advanced research computing resources and needs at 4 Canadian Genome Centers

Guillaume Bourque
Director of Bioinformatics
McGill University & Genome Quebec Innovation Center (MUGQIC)
Montréal, Québec
Phone: 514-398-7245
E-mail: guil.bourque@mcgill.ca

Michael Brudno
Director
High Performance Computing for Health (HPC4Health Consortium)
Hospital for Sick Children & University Health Network
Toronto, Ontario
Phone: 416-978-2589
E-mail: brudno@cs.toronto.edu

Steven Jones
Head of Bioinformatics and Associate Director,
Michael Smith Genome Sciences Centre (GSC)
Vancouver, British-Columbia
Phone: 604-877-6083
E-mail: sjones@bcgsc.ca

Lincoln Stein
Program Director, Informatics and Bio-computing,
Ontario Institute for Cancer Research (OICR)
Toronto, Ontario
Phone: 416-673-8514
E-mail: lincoln.stein@gmail.com

March 1st, 2016

Opportunities and challenges in Genomics

The Human Genome Project, completed in 2003, required hundreds of sequencing machines and cost over $1 billion over a 10-15 year period. In 2016, it is possible to sequence an individual's genome in 2-3 days for little more than $1000 (the cost of a day at the hospital). These numbers are not static and some estimates suggest that, by 2020, data will be generated at up to one million times the current rate, which is orders of magnitude faster than the growth of computational power as predicted by Moore's law (i.e., doubling of computing power every two years). The analysis of human genomes, transcriptomes, epigenomes, proteomes, interactomes, metabolomes and microbiomes will provide the basic knowledge necessary to diagnose, understand and cure many diseases leading directly to reductions in health costs for society.

– Canadian Bioinformatics National Strategy

Over the next decade almost every biomedical investigation in basic and clinical research will be enabled through characterization of an accompanying genome sequence. Genomic technologies have become a critical component not only in human health research but also in other fields such as: agriculture, fisheries, forestry and mining. With next-generation sequencing technologies revolutionizing the life sciences, data processing and interpretation, rather than data production, has become the major limiting factor for new discoveries. **In this context, the availability of advanced research computing resources has become a key issue for the genomics community.**

We represent the 4 largest Genome Centers in Canada: the McGill University and Genome Quebec Innovation Centre (MUGQIC) in Montréal, the High Performance Computing for Health (HPC4Health) Consortium in Toronto (that supports The Centre for Applied Genomics (TCAG) and Princess Margaret Genomics Centre), Canada's Michael Smith Genome Science Centre (GSC) in Vancouver and the Ontario Institute for Cancer Research (OICR) also in Toronto. Through support from the Canadian Foundation for Innovation, Genome Canada along with regional partners, the Ontario Government, as well as numerous other funding agencies, our platforms have been at the core of a national genomics initiative (>$1.5 billion invested in the past 15 years), which has helped advance Canada's science and technology agenda.

Over the last decade, we have provided sequencing and informatics services for thousands of regional, national and international research initiatives. These include the International Cancer Genome Consortium (ICGC), the International Rare Disorder Research Consortium (IRDiRC), the International Human Microbiome Consortium (IHMC), the Internal Human Epigenome Consortium (IHEC), modENCODE and the NCI

Genome Data Commons project. In 2015 alone, the >600 staff working in the 4 centers have supported over 2,500 Principal Investigators from all 10 provinces, with cost recovery services exceeding $30 million dollars.

The coming years will mark the long-awaited point of "inflection" in genomics, as costs for accurate genome sequencing move below the $1,000 range. The bioinformatics and computational biology community is currently engaged in preparing its first national strategy document. This strategy will be delivered via a multi-agency government/private-sector consortium, which includes Compute Canada (CC), and will connect the science to national research, training and infrastructure needs. It is indisputable that significant investment in advanced research computing infrastructure will be required to realize the benefits of genetics and genomics research and bring new insights to different fields of the life sciences.

We will now summarize available computing resources and anticipated needs at the 4 Genome Centers.

Current HPC resources

We have been using a number of advanced computing platforms to cater for the needs in compute and storage of our high-throughput genomics platforms. The bulk of the resources utilized are local high-performance computing (HPC) clusters but many CC clusters have also being exploited. These include the guillimin, mammouth, Scinet and Westgrid clusters. Two of the center's local HPC clusters (at HPC4Health and GSC) are also already part of the CC network. In aggregate, the available HPC resources currently available to the Genome Centers are shown in Table 1 and are currently used at near capacity. Additional resources such as Amazon Web Services have also been used but so far that utilization has been limited in scope. It is important to note that given the sensitivity of some of our datasets, our computing resources also have to satisfy distinct privacy, security and redundancy requirements.

	Cores	Online disk (PB)	Tape library (PB)
MUGQIC*	2,604	3	6
HPC4Health	15,952	4.9	4.5
GSC**	11,000	12	14
OICR	8,324	4.5	1
Total	**37,880**	**24.4**	**25.5**

Table 1. Current HPC resources available at the 4 Genome Centers.
*** Includes resources obtained through Compute Canada.**
**** Includes 600 cores of contributed co-located systems**

Anticipated HPC needs (2016-2020)

To support the increase in sequencing throughput at the 4 Genome Centers, compute and storage capabilities have had to grow exponentially over the last few years. To use the MUGQIC as an example, in 2010 an internal cluster with 300 CPU cores was sufficient to cater for all computing needs. In 2015, this internal cluster has been built up to 1,104 cores but, importantly, a resource allocation at CC has also provided access to an additional 1,500 core years. In total this amount to 2,604 cores, which represents a 8.5X increase in 5 years. Similarly, the storage resources went from approximately 1.5 PetaBytes (PB) to 8 PB (split between disk and tape storage), also a 5X increase. Comparable growth in computing and storage has also happened at the other Genome Centers.

We anticipate the computational needs at the 4 Genome Centers to continue to grow exponentially over the next few years. This stems from 3 main factors: 1) the addition of new sequencing instruments (e.g. in 2015, 3 of our centers were awarded a 58M CFI award to purchase 15 new HiSeq X instruments which have boosted our sequencing capacity >5X) and future instrument upgrades, 2) the need to apply more advanced analysis strategies to mine these large datasets to their full potential (e.g. *denovo* assembly of human genomes) and 3) increased demand for access to informatics infrastructure to utilize large international public datasets.

We show in Table 2 a detailed breakdown of our expected computing and storage needs up to 2021. **The main assumptions in this forecast are that our computing resources will have to double every 2 years and that online disk and tape backup resources will have to triple over the same period.** These estimates are based on the fact that sequencing production in our centers currently doubles every 12 months. We note that this is a conservative estimate that assumes aggressive data management and limited computational analyses.

	2014	Current	2018	2020
Cores	24,552	37,880	75,760	151,520
Online disk (PB)	17	24.4	73.2	219.6
Tape storage (PB)	12.5	25.5	76.5	229.5

Table 2. Anticipated HPC needs at the 4 Genome Centers.