PHI-Compliant Computing and Storage: a critical need for Canadian biomedical and health research

Submitted in response to NRDIO/NOIRN's Call for White Papers on Canada's Future DRI Ecosystem. December 14, 2020

Richard F. Wintle[1,*], Lisa J. Strug[2], Avery MacLean[3], Rob Naccarato[4], Adam Shlien[5], Stephen W. Scherer[6]

[1]Assistant Director, The Centre for Applied Genomics (TCAG), The Hospital for Sick Children, Toronto, Canada

[2]Associate Director, TCAG; Senior Scientist, Genetics and Genome Biology, The Hospital for Sick Children; Professor of Statistical Sciences and Computer Science and Director of CANSSI Ontario, The Faculty of Arts and Science, the University of Toronto, Toronto, Canada

[3]Director, Research IT, The Hospital for Sick Children

[4]Senior Manager, High Performance computing Facility (HPF), The Hospital for Sick Children

[5]Associate Director, Translational Genetics, Department of Paediatric Laboratory Medicine; Senior Scientist, Genetics and Genome Biology, The Hospital for Sick Children

[6]Director, TCAG; Director, University of Toronto McLaughlin Centre; Senior Scientist, Genetics and Genome Biology, The Hospital for Sick Children

*Designated contact: rwintle@sickkids.ca; 13.9715 PGCRL, 686 Bay St., Toronto, ON, M5G 0A4

1. Summary

Our vision is rooted in the central tenet that every biological investigation will be enhanced by the presence of an accompanying genome sequence and informed by clinical and demographic information. Existing computational and data storage architectures for the analysis of these types of identifiable personal health information (PHI) are currently overburdened. Additionally, data sharing between researchers, either within or between individual institutions, is complicated by the use of separate computational and storage systems, and the need for both inter-institutional data use agreements and forward-looking research ethics approvals that explicitly contemplate broad data sharing. In order for Canadian scientists to take full advantage of new discoveries and assume their place at the forefront of collaborative health research worldwide, a national strategy is needed to address both present and future needs for PHI-compliant computing, storage, and sharing. To stay relevant and competitive internationally, this strategy will further need to fully embrace cloud computing and storage. We envision a role for NDRIO to provide the framework and resources for a federated data storage platform, robust computational resources that can be implemented in easily accessible, cloud-based environments, and seamless sharing between Canadian researchers.

2. Current Status

Researchers at The Hospital for Sick Children can currently use resources provided by *HPC4Health*, a member of Compute Ontario's consortium that is affiliated with Compute Canada. This high-performance computing facility offers fast, secure compute and storage that serves some teaching hospitals affiliated with the University of Toronto, as well as selected other groups, including the Vector Institute and the Institute for Clinical Evaluative Sciences (ICES). *HPC4Health* also employs highly qualified personnel who work closely with researchers to interpret their requirements and help them to use these resources efficiently, and most importantly ensures a PHIPA compliant environment suitable for clinical research, with appropriate governance and regular third-party validation. This latter point is critical not only for clinical, but also basic research using identifiable human data. Data are transferred with the *Canarie* high speed network to connect with both national and international partners. The connection to *Canarie* is enabled through two Provincial Research and Education Networks (*GTAnet-ORION*). Advanced research computing resources operated by the University of Toronto (i.e., *SciNet*) are not PHI-compliant, and thus cannot be used for the majority of biomedical and health research.

Importantly, the environment described above represents physical hardware located within our data centre at The Hospital for Sick Children's *Peter Gilgan Centre for Research and Learning* in downtown Toronto, and requires users to pay annual fees for storage and maintenance. This can be prohibitively costly for many investigators, who thus are unable to download and compute on public datasets, because they do not have access to adequate storage for these data. Although *HPC4Health* includes a private cloud environment, none of these resources take advantage of arbitrarily large, cloud-based storage or computing such as could be provided by commercial cloud providers. This significantly inhibits our research activity to the confines of what is possible within the four walls of our data centre. Most of the barriers to using cloud resources are based in institutional concerns surrounding data privacy and sovereignty, as well as significant costs to access commercial clouds. A national, NDRIO-supported infrastructure would be one possible solution that could provide surety to institutions such as ours that relevant data sets will be easily linked together, while being kept appropriately secure, in a managed, cloud-based environment.

3. Existing Challenges

Ten years ago, in making a case for cloud computing for genomics, Lincoln Stein wrote of "the impending collapse of the genome informatics ecosystem."[1] In 2020, our most pressing challenge is still that of scalability: as costs to generate massive genome-wide sequence data continue to drop, sample sizes in studies will increase, resulting in ever-larger datasets, almost certainly extending beyond current institutional capacities. At present, we are limited to studying clinical data painstakingly extracted from patient charts by individual Clinical Research Associates, based on hypotheses about which data are likely to be important, a well-established paradigm used in genetics research. Our current architectures and regulations are inhibiting our participation in the global paradigm shift that sees the sharing and linking of large-scale clinical and genomics research data. Related to this issue, we are rapidly approaching (and in some cases already at) the point where it is not practical to download or compute on the largest available international data sets, using existing infrastructure; additionally, it is inefficient to use storage space and bandwidth to download data that already exist elsewhere in cloud storage. We need to have the flexibility to bring our software (via containerized resources) to the data, not the reverse. Commercial scale, cloud-based storage and computation would provide a route to alleviate these bottlenecks.

The absence of reciprocal agreements between sites that would address regional disparities in regulatory frameworks for management of clinical data (PHIPA and PIPEDA) are a limiting factor for scientists across our research enterprise. Access to PHIPA-compliant, appropriately governed infrastructure is a key consideration for working with or at other sites, using either on-premises HPC or (potentially) cloud computing. The requirements of international collaborators governed by the General Data Protection Regulation (GDPR) are even more stringent, requiring significant investment in security governance models, monitoring, and validation of the infrastructure. We also discern a need for more affordable, dedicated high speed network connections to move big data from local laboratory instruments to cloud resources for large-scale processing. Commercial services that could extend our local network into the cloud (such as Microsoft ExpressRoute) are prohibitively expensive at this time, with pay models that are not designed for the exploratory nature of scientific research, and come with no assurances of appropriate PHI compliance.

It is important to contemplate not just sharing of data generated now, or that coming in the future, but also re-useability of existing or legacy data sets.[2] These can represent significant prior investments of money, time, and expertise, and should not need to be re-generated. In many cases, the existing data may also represent a specific snapshot in time of a patient's disease case (e.g., for cancer), and thus it may be impossible to re-generate identical data, necessitating re-use. We also recognize that there are many examples of data resources that would provide considerable additional scientific value if they were more easily linked together: clinical, population cohort, healthcare utilization, and demographic data as examples. Our arguments, while rooted in genomics research, are relevant to discovery using any type of PHI data.

As more and more Canadians enrol in genetic/genomic based studies, this wealth of health research data presents a rich resource that can potentially be re-used for studies beyond those that were originally contemplated. This can easily happen, for example, with persons having complex medical conditions with multiple signs and symptoms, who might realistically enrol in two or more different studies. These data must be made accessible, shareable, and in a format that is straightforward for additional researchers to use. Without systems to facilitate this, the data are used only once and additional scientific value that could be derived would be lost. We learn so much from every patient's unique genome that these data may in fact become the justification for future clinical trials, and/or improved diagnostic tests for the next generation.

An extension of this issue is the case in which researchers access existing data sets external to their own institutions, such as those held by the National Centre for Biotechnology Information's (NCBI) *database of Genes and Phenotypes* (dbGaP), the *Genotype-Tissue Expression* (GTEx) project, and others. GTEx, in particular, represents a unique resource of gene expression, histology, and other data from nearly 1,000 individuals, that is impossible to re-create, hence necessitating data sharing for it to achieve utility. In such cases, it is possible that multiple researchers at the same institution may independently request access to the same data (as examples, from a large disease cohort or a set of population controls), resulting in multiple copies of these data being stored and computed on, using the same resources. This is wasteful of local storage space and computation time. As well, the significant administrative time and overheads involved in gaining sharing approval and access to such data sets multiple times should not be underestimated. A related issue is the need to download these data locally to parse and work with them, which is time consuming, compute heavy, and an unnecessary use of resources. In our experience, this has been a major challenge, which could again be solved with a federated data platform approach, allowing multiple users to access and compute on these data.

Another obvious major challenge is that not all researchers have access to privacy-compliant infrastructure, as we do, although even our systems are now strained. Those without access, such as many researchers who are located in traditional university academic departments, have limited options. With the massive public data resources that continue to become available online, including in the humanities and the social sciences, this constitutes a significant disadvantage for Canadian scholars.

4. Future Vision

Genome data can be identifying of individual participants, and thus PHI-compliant architectures for the storage of, and computation on, genomic data are critical to the Canadian biomedical and health research enterprise. Rooted in the open sharing model established by the landmark Human Genome Project, human genomics research has had many successes that similarly rely on a philosophy of sharing of large data sets. Notable recent examples with significant Canadian participation include the now >13,000 genome sequences and associated phenotypes of the Autism *MSSNG* project (www.mss.ng),[3] and the International Cancer Genome Consortium/The Cancer Genome Atlas analysis of over 2,600 tumour genomes.[4] *MSSNG*, for example, now provides cloud-stored genome sequence data accessible by a simple, lightweight data access request process, and has to date shared data with 224 investigators representing 66 institutions and eight private sector research entities, distributed across 19 countries. This was facilitated by forward-thinking research ethics consent language, and the willingness of the cloud storage provider (Google) to establish appropriate data security protocols. These examples provide a guide for what we wish NDRIO to accomplish; what we require is easier and faster processes to enable results such as these, that can be established within Canada, and do not require substantial investments from philanthropic funders or private sector partners, as is the case for *MSSNG*.

Additionally, the execution of the largest population-scale health research projects will in many cases require significant capacity across multiple sites. One recent and timely example, Canada's *COVID-19 Host Genome Sequencing Initiative* ("HostSeq"; http://www.cgen.ca/project-overview), integrates data from three genome centres with that from dozens of clinical sites submitting samples and meta-data, and aims to link to national and provincial data collected as part of the administration of Canada's publicly funded health care system. The challenges of performing laboratory and analytical quality control across multiple sites have recently been examined.[5] Those who wish to be recipients of the resultant shared genome sequence, clinical, demographic, and healthcare data will require the additional resources and strategies contemplated in this document.

Another consideration is the potential for enhanced return on investment. In one example, we have recently estimated that >$20 million of new grant funding was generated via data sharing and insights derived from the SickKids Foundation's $1-3 million investment in the *KiCS* paediatric cancer sequencing program. This is in addition to the creation of the *Precision Oncology for Young People* (PROFYLE) program, and resultant new clinical trials. The scientific and economic values of data become far greater when they are well-managed and shareable.

In short, these goals could be facilitated by a national strategy, led by NDRIO, rooted in the philosophy of building bridges between and among Canadian institutions, and enabling access and contributions to international efforts.

5. Bridging the Gap

We envision NDRIO as facilitating health research in Canada by prioritizing the support and coordination of a secure, PHI-compliant, highly accessible (i.e., on-demand) computing and storage architecture, directly available to Canadian scientists. We emphasize that such a system would not only benefit genomics, but also be broadly enabling to psychological, social, health economic, policy, ethics, and many other domains of health research. Further, we imagine a system in which data are easily flowed from one user to another, without the need for inter-institutional data sharing agreements on a case-by-case basis, within an overall, secure environment. In the most desirable implementation, such an NDRIO-coordinated environment would satisfy the most common institutional requirements for data sharing, allowing for both deposition and retrieval of data by researchers, with minimal roadblocks.

More specifically, we recommend a strategy to address the following three key goals:

1. A national strategy for PHI-compliant, cloud-based data storage and computing, allowing for containerized computation resources accessible by all. The availability of adequate capacity to ensure both the longevity and utility of data generated in Canada would be a key component of this strategy. Such a system would greatly relieve burdens on individual institutions, associated philanthropic foundations, and other funders.
2. Data integrity and security protocols broad enough to satisfy data sharing agreements for data sets currently held and shared by major, international data repositories. This would provide international data holders with confidence that their data will be secure, and Canadian researchers with a simple route to rapidly obtain access to such data without complex, one-at-a-time, institutional approval processes.
3. A national framework for data sharing, inclusive of all jurisdictions across Canada, and the establishment with bioethics experts of appropriate research ethics guidelines and language to allow for seamless, secure data sharing between and among Canadian researchers.

With the implementation of these strategies, NDRIO could ensure that Canadian biomedical and health research science has the resources, expertise, and capability to be in a world-leading position, while greatly alleviating financial and administrative burdens on institutions, researchers, and their funders. This, in turn, would have the effect of freeing up both human and intellectual capital to better address the medical challenges of today and the future.

6. References

1. Stein LD (2010). The case for cloud computing in genome informatics. *Genome Biology* 11(5):207.
2. Wallace SE, Kirby E, Knoppers BM (2020). How can we not waste legacy genomic research data? *Frontiers in Genetics* 11:446
3. Yuen RKC, Merico D, Bookman M, L Howe J, Thiruvahindrapuram B, *et al.* (2017). Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nature Neuroscience* 20(4):602-611.
4. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. *Nature* 578(7793):82-93.
5. Corbett RD, Eveleigh R, Whitney J, Barai N, Bourgey M, *et al.* (2020). A distributed whole genome sequencing benchmark study. *Frontiers in Genetics*, 01 December 2020, https://doi.org/10.3389/fgene.2020.612515