CanCOGeN Pan-Canadian Experience in Data Sharing: Lessons Learned and Recommendations for the Road Ahead

CanCOGeN | RCanGeCO

The authors would like to acknowledge the financial contribution of Genome Canada and its partners to this policy brief in the context of the CanCOGeN project

This Policy Brief can be cited as follow:

Yann Joly, Hanshi Liu and Ma'n Zawati on behalf of the CanCoGeN Data Sharing Group, *CanCOGeN Pan-Canadian Experience in Data Sharing: Lessons Learned and Recommendations for the Road Ahead,* Genome Canada, Ottawa, 2022

Context

The SARS-CoV-2 virus, which causes COVID-19, was first reported in Canada in January 2020, with subsequent community transmission being confirmed in mid-March of the same year. In April 2020, successful requests for emergency funding were made to Innovation, Science, and Economic Development Canada (ISED) by Canada's lead genome sequencing (CGEn) and viral sequencing groups. Genome Canada then launched the Canadian COVID-19 Genomics Network (CanCOGeN), an initiative composed of two ISED-funded projects: VirusSeq, responsible for sequencing viral genomes and sharing them along with accompanying metadata and, HostSeq, a second project led by CGEn with similar objectives that focused on sequencing the human host genome and including both clinical data and metadata. An overarching goal of both projects is to share data and facilitate linkages between patient and viral data for informing public health research and policy.

While capacity building and sequencing progressed at a good pace for both projects, data sharing for research did not materialize with the urgency expected in a pandemic context. This comment is not meant to belittle both the hard work and goodwill shown by the teams of VirusSeq and HostSeq under difficult circumstances. In the case of HostSeq, even though there were significant and complex scientific, regulatory, and logistical challenges associated with sharing human genetic data and accompanying metadata, access to the controlled database was available to users a year after the project start date (https://www.cgen.ca/daco-main). That said, there is no doubt additional measures can be taken to ensure a quicker turnaround time in response to future pandemics.

VirusSeq was slow to share its first viral genomic sequences publicly. The problems for this group seemed to be systemic, resulting in a particularly uncoordinated and limited sharing of viral data along with a minimal amount of metadata for research during the first year of the pandemic (*see figure 1 below*). Furthermore, there is still no centralised process to access more sensitive metadata (vaccination status, ethnicity, specific age, etc.) along with viral or the patients' genomic sequence data from Canadian labs, or to obtain additional personal data to achieve data linkages between host and viral genomic sequences. Other countries at the forefront of COVID genomic research, such as the United Kingdom and the United States, also faced this last challenge. This suggests that there may be room for international collaboration to find a common solution to enable such linkages, which are of tremendous importance for public health research and planning. However, an international initiative of this sort will require substantial investment in time, money, and political will. It would benefit from being undertaken under the auspices of an international organization such as World Health Organization (WHO) or Global Alliance for Genomics and Health (GA4GH).

Currently, the only way to obtain such data is to request it directly from each provincial lab or research project; such requests have their own complex set of procedures, approvals, and outcomes. The outlook improved in the second year of the pandemic when Genome Canada and other agencies clearly established genome data sharing as a priority and established specific governance committees to systematically address any data sharing challenges. Another important contributing factor to this improvement was the substantial financial investment of the Canadian government in the capacity building of public health labs, as well as Canada's national genome sequencing facility (CGEn).

While the current increase in the Canadian genome data-sharing performance is a very positive achievement of CanCOGeN and the Canadian Public Health Laboratory Network (CPHLN), there is consensus that there remains much room for Canadian stakeholders to continue improving genomic data and metadata sharing. There are also concerns that, as we move towards a new phase of the pandemic, some of CanCOGeN's governance framework will be dismantled or passed on to other organisations. In the absence of a clear national mandate for data sharing, this activity will no longer be given the priority it deserves in the public health and research agenda. If this happens, there is a risk that the critical improvement accomplished in data sharing performance will be lost and that Canada will again fall behind other developed economies in this area.

With this challenge in mind, both the CanCOGeN Virus Seq Implementation Committee and the CanCOGeN data sharing committee agreed on the need to develop a Data Sharing Roadmap that would identify the challenges on the path to data sharing and propose concrete strategies and policy avenues to address the current challenge. While the roadmap focuses mainly on the data sharing challenges identified in the VirusSeq project, several of our findings are also applicable to HostSeq.

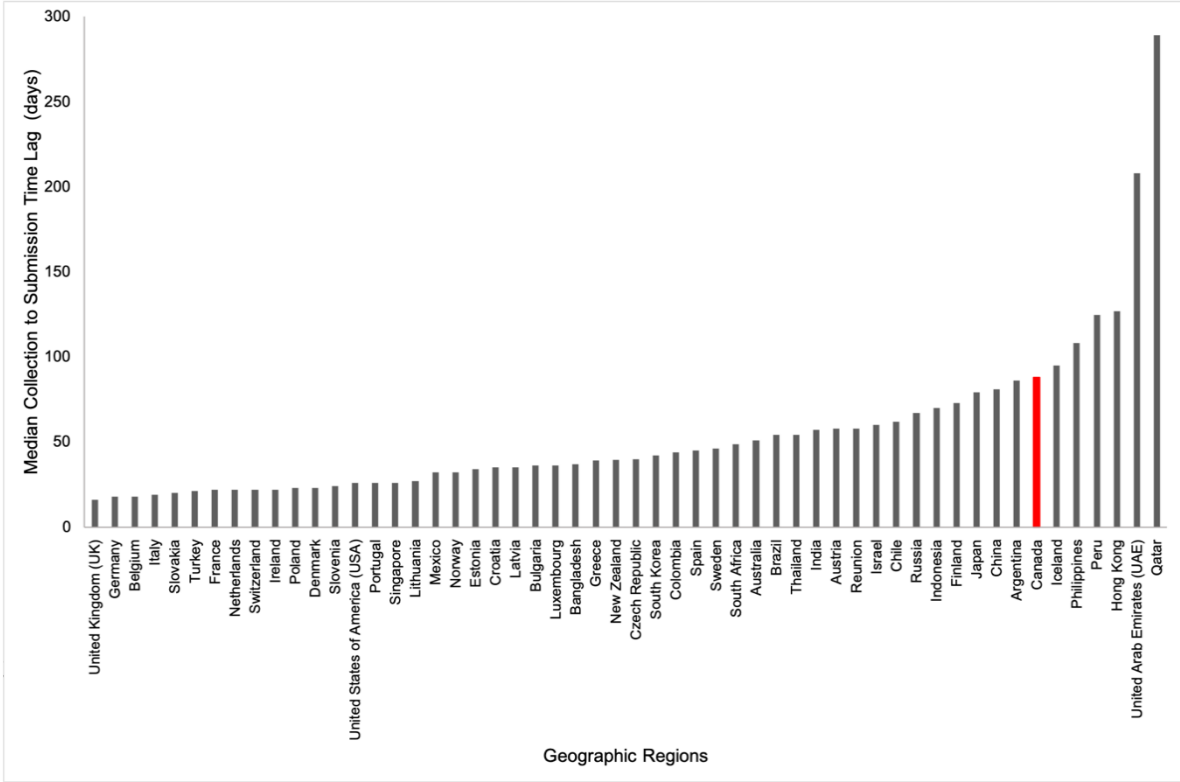

Figure 1

GISAID Median Collection to Submission Time Lag by Geographic Region: Plot illustrating the time lag between sample collection to submission of data in GISAID, in Canada and other geographic regions, as of May 2021. For Canada, the collection to submission time lag had a median of 88 days for over 44,000 viral genomes. The level of completeness of additional metadata provided with sequences varied considerably across provinces. This plot was adapted from Kalia et al. 2021. Please note, only geographic regions submitting more than 1,000 viral genomes are shown.

Key Findings and Recommendations

1. **Viral Sequence and minimal metadata** Nothing in Canadian data protection law and policies prevents the rapid sharing of viral genome sequence data along with a small amount of non-identifying metadata. Assuming minimal precautions suggested in the Privacy Implications of Sharing of Viral Sequence and Minimal Patient Metadata memo are taken, such data do not constitute personal data and can be promptly and openly shared without other formalities. A sustained effort should be made by research funders and privacy offices to engage data stewards (see Recommendation 3 below), public health labs, and organisations responsible for such data to convey this message in a coherent and concise manner.

Since it does not raise particular legal issues, rapid sharing of viral genome data and minimal metadata should be a given and parties that are reluctant to share should be reminded of their ethical duty towards the public good. In the case of a pandemic, possible sanctions (see Recommendation 5 below) should be considered in cases of persistent and willful non-compliance.

2. **Data privacy legislation** Canadian federal and provincial public health laws offer public health agencies broad power to access and use data to fulfill their mission. However, these *laws are insufficiently harmonised at present thereby creating uncertainties. For example, there remains an unnecessary margin of discretion with provincial public health agencies regarding which data can be disclosed, with whom, and under what conditions*. Moreover, in some provinces, it is not clear that 'research' would be an activity that falls under the mission of a public health lab and so benefit from an exemption from the general privacy rule of consent. *Canadian federal, provincial, and territorial data protection legislation should be amended to stipulate, in clear language, that personal information can be collected, used, and disclosed for public health purposes, including*

research, in combination with appropriate de-identification and good information stewardship practices.

In addition, secure, privacy-compliant solutions that will allow for data linkage and deeper analysis, for example, the ability to link data with clinical data sets, socio-demographic factors, etc. should be identified and presented to public health agencies for approval. The experience of the CanCOGeN Coordination Committee shows that such solutions currently exist although the challenge of concretely implementing them may be daunting. It would be helpful if such processes were systematically and uniformly put in place ahead of a pandemic or at its onset. For example, to facilitate data linkage, **we recommend** that: 1) all consent forms include a clause notifying participants of such linkage; 2) common standards for linkage be established; 3) an individual from each of the projects and, where possible, from each of the relevant repositories, be identified and resourced appropriately to ensure efficient coordination and timely delivery of the data; and 4) proper measures to enable central controlled data access, or federated access, be set in place.

3. **Data stewardship** According to the GA4GH, data stewards are entities or senior officials responsible for assuring the quality, integrity, and access arrangements of data and metadata in a manner that is consistent with applicable laws, institutional policies, and individual permissions (see also CIO standard CAN/CIOSC 100-7:20XX (D2) for a complete definition). This includes entities and positions that deal primarily with personal health information, and those that deal with other health information. This could also include information that is not personal (ex. viral genome sequence data) but is also collected for health purposes. **We recommend** that the current system, where provincial agencies are considered data custodians, be replaced by a system of data stewardship, where an individual at each institution is hired to fill the position of a data steward. A lead data steward should also be appointed to harmonize and coordinate the work of stewards at provincial institutions. These individuals will be responsible for promoting good data governance practices, including the transition to a culture more favorable to collaboration and data sharing than the one that currently prevails. Federal and provincial laws should be amended to impose this requirement.

In the meantime, public funding agencies could impose this requirement as a condition for any funding beyond a certain financial threshold and, as an incentive, provide additional money to be used towards the salary of these stewards.

4. Attribution The rules of attribution/acknowledgement for sharing genome data are extremely disparate both nationally and internationally. The rules currently differ depending on the type of genome data shared, the repository, the organisation or

consortium providing data oversight, etc. Such an inconsistent and unpredictable system contributes to a climate of mistrust between users and producers and is a source of delay for data sharing associated with the need to negotiate terms ad-hoc. **We recommend** that all Canadian stakeholders agree on consensus guidelines concerning attribution and acknowledgement for using viral sequence genomic data. Such guidelines could be based on the model currently used for the VirusSeq Canadian Data Portal. The HostSeq publication and acknowledgment policy provides, in turn, a model for databases making their host sequence data available through a controlled-access system.

5. Accountability It is important that transparent and enforceable rules be developed concerning sanctions for improper or negligent acts that have strong negative public health impacts on the sharing of pathogen genomic, and de-identified health data. Such sanctions will lead to greater accountability for all stakeholders involved in the Canadian data ecosystem. **We recommend** that the following sanctions be considered by stakeholders:

Issues	Sanction	Source
Using data shared for health research to intentionally re-identify an individual patient/participant	Penal sanction.	Law
Willful refusal to share data timely that could benefit public health or healthcare	Penal sanctions, the detailed interpretation of what constitutes 'willful' refusal and the implementation mechanism for these sanctions could be explained in a Code of Conduct.	Law
Using data shared for health research in a publication without providing proper attribution to data producer(s)	First, the data provider(s) and data user(s) involved should attempt to negotiate an equitable solution. If negotiations fail, publishers and funders,	Community norms (Canadian publishers and research funders)

	using means at their disposal, should try to promote the adoption of a solution that is acceptable to all parties.	

6. Infrastructure CanCOGeN VirusSeq struggled early in the project to get provincial public health laboratories to share their data in a uniform, coordinated manner with researchers (see figure 1 above). The development of a new infrastructure, *the Canadian VirusSeq Data Portal* was conducive to more coherent data-sharing activities. This was achieved by helping the CPLHN and its members streamline the deposition process and harmonizing their practice by identifying the most relevant metadata to share, locating errors in the process, and suggesting ways to resolve them. The portal increases the analytical capacity of public health. It provides a mechanism for (a) data access requests and (b) collaboration requests, i.e., interested parties can be connected to relevant public health contacts who have richer contextual information compared to what is available publicly in the portal. This collaborative model has been successful in the UK where they have co-developed a suite of bioinformatics tools, diagnostic technologies, and mechanistic research insights. Maintaining the *Canadian VirusSeq Data Portal* ensures that a single repository responsible for sharing viral genome sequences and metadata for research use is available to the community. The portal team is independent of public health stakeholders, embraces the principles of findable, accessible, interoperable, and reuseable (FAIR) open data sharing and provides expert advice to data providers on matters of data annotation, sharing, storage, and governance. *Considering the added value and synergistic potential of the data portal, **we recommend** that it be maintained beyond the CanCOGeN project. Provided additional funding becomes available, the portal's mandate could be expanded to also provide controlled access to more sensitive metadata types.*

Through HostSeq, CGEn, as a data steward, has developed a national database containing genomic as well as personal and health information in line with policies and practices developed by the GA4GH. It has also put in place a data access office that receives and reviews requests for access to data and convened an independent pan-Canadian data access committee to make final decisions. This one-stop-shop model has streamlined the access process and ensured timely approval of requests. It is recommended that these infrastructures be maintained beyond the CanCOGeN project. CGEn also created a GA4GH aligned beacon portal for HostSeq, which allows for the rapid querying of the

dataset for genetic alleles of interest in an open and anonymised fashion, protecting the identity and confidentiality of the participants.

7. CanCOGeN data committees (CanCOGeN Coordination Committee, CanCOGeN Data Sharing Committee) Beyond the data portal, the work played by the data committees of CanCOGeN has contributed to improvements in data sharing through: 1) providing rapid responses to specific questions of the community on data sharing (ex. on the privacy implications of sharing specific metadata, on the suitability of specific data-sharing agreements, on consent forms); 2) performing constant advocacy work and engagement to promote data sharing, across the consortium and in the media; and 3) providing coordinated data curation and specification harmonization activities. Furthermore, these data committees have acquired important know-how on the opportunities and pitfalls associated with data sharing in times of pandemics from their work in CanCOGeN. There is a substantial risk that should such data groups, or a consultant with a similar function, no longer play this role for the CPLHN and its members, data sharing risks becoming once again a low priority. ***We recommend*** *that until Recommendation 3 (data steward) has been implemented in whole or in part, a person or committee be tasked with continuing the work of the CanCOGeN data committees.*

In addition to this, **we recommend** that the role of the CanCOGeN Data Sharing Committee be expanded to include the identification of metrics to measure more rigorously our genomic data sharing performance in times of pandemics. The proposed metrics could assess the timeliness, completeness, and quality of shared data, as well as the impact of such data sharing on research and healthcare. They could also be used as targets in funding agreements. The newly appointed data stewards (see Recommendation 3, above) could be responsible for gathering data on how each public health lab or organisation is performing against the identified metrics.

Adopted by Virus Seq Implementation Committee on: agreement in principle 23/02/22

Adopted by Host Seq Implementation Committee on: agreement in principle 07/04/22

Writing group:

Yann Joly, McGill U., VirusSeq

Ma'n Zawati, McGill U., HostSeq

Hanshi Liu, McGill U., VirusSeq

Advisory Group:

Koko Agborsangaya, Genome Canada

Naveed Aziz, CGEn, HostSeq

Guillaume Bourque, McGill U., VirusSeq & HostSeq

Fiona Brinkman, Simon Fraser University, VirusSeq

Erin Gill, Simon Fraser University; VirusSeq

Emma Griffiths, University of British Columbia, VirusSeq

Will Hsiao, Simon Fraser University; VirusSeq

Steven Jones, BC Cancer Research Institute, HostSeq

Bartha Knoppers, McGill U., HostSeq

Catalina Lopez-Correa, Genome Canada

Kim McGrail, University of British Columbia

Kieran O'Doherty, University of Guelph

Art Poon, Western University, VirusSeq

Megan Smallwood, Genome Canada

Terry Snutch, University of British Columbia; VirusSeq

Eric Sutherland, PHAC

Gijs van Rooijen, Genome Alberta

Daryl Waggott, Genome Canada

Observers:

Matthew Croxen, Alberta Public Health Laboratory, VirusSeq

Sandrine Moreira, INSPQ, VirusSeq

Natalie Prystajecky, BC Centre for Disease Control, VirusSeq

References

The following related activities and documents inform and complete our recommendations

Standards and policies

- Canadian VirusSeq Data Portal, Acknowledgement of Contributions and other governance policies (2021)
- List of minimal metadata set to include with viral sequence data (2020/2021)
- Opportunities for virus genomic data: summary, Caroline Colijin and Art Poon, November 2021
- Barriers to Public-Level Pathogen Genomics Data Sharing, CanCOGeN Virus-Seq Ethics and Governance Group, January 2021
- Privacy Implications of Sharing of Viral Sequence and Minimal Patient Metadata, CanCOGeN Virus-Seq Ethics and Governance Group, Fall 2020
- Host Seq Model Informed Consent Form (2020)
- HostSeq Database Governance Framework (2020)
- HostSeq Databank Publication & Acknowledgment Policy (2021)
- Genome Canada Data Release and Sharing Policies (2017)

Publication

- Song L, Liu H, Brinkman FSL, Gill E, Griffiths EJ, Hsiao WWL, et al. Addressing Privacy Concerns in Sharing Viral Sequences and Minimum Contextual Data in a Public Repository During the COVID-19 Pandemic. Frontiers in Genetics [Internet]. 2022 [cited 2022 Mar 24];12. Available from: https://www.frontiersin.org/article/10.3389/fgene.2021.716541

- Kalia K, Saberwal G, Sharma G. The lag in SARS-CoV-2 genome submissions to GISAID. Nat Biotechnol. 2021 Sep;39(9):1058–60.

- Knoppers BM, Beauvais MJS, Joly Y, Zawati MH, Rousseau S, Chassé M, et al. Modeling consent in the time of COVID-19. J Law Biosci. 2020 Jun;7(1):lsaa020.

- Lin YC, Brooks JD, Bull SB, Gagnon F, Greenwood CMT, Hung RJ, et al. Statistical power in COVID-19 case-control host genomic study design. Genome Med. 2020 Dec 28;12(1):115.

- Tremblay K, Rousseau S, Zawati MH, Auld D, Chassé M, Coderre D, et al. The Biobanque québécoise de la COVID-19 (BQC19)—A cohort to prospectively study the clinical and biological determinants of COVID-19 clinical trajectories. PLOS ONE. 2021 May 19;16(5):e0245031.

Reports

- Alexander Bernier, Hanshi Liu, Robyn McDougall, Yann Joly, Law and Policy of Public Health Information Sharing in Canada (2021)
- Edwin Faraji S. et al., Genomic Data Sharing and Governance Policies: A Selective Comparative Review of Canadian Partners (2022)
- Rhiannon Cameron, Sarah Savić Kallesøe and Emma Griffiths, CanCOGeN VirusSeq Comparison and Analysis of Canadian Public Health SARS-CoV-2 Case Report Forms, https://www.genomecanada.ca/sites/default/files/2020-12-10_crf_report_.pdf

Surveys

- Hanshi Liu, Rim Metina-Belknap, Yann Joly, CanCOGeN QC survey (October 2021)
- Canadians' Opinions Towards COVID-19 Data Sharing: Survey (forthcoming in 2022)

Workshops

- Genome Canada, *COVID Genomic Data Portal Workshop,* 19 April 2021
- Genome Canada, *Public Health Laboratories Workshop,* 6 May 2021