Public Health Ontario | Santé publique Ontario

TECHNICAL NOTES

Phylogenetic Analysis of SARS-CoV-2 in Ontario (Nextstrain interface hosted at Public Health Ontario)

Background

Public Health Ontario (PHO) is working with the Ontario Ministry of Health, Ontario Health, and clinical laboratories and research partners to test for and monitor COVID-19 variants of concern (VOCs) in Ontario.

Nextstrain is an open-source platform to visualize and interact with genomic data.[1] PHO is hosting a Nextstrain build of SARS-CoV-2 samples that have undergone whole genome sequencing (WGS) and bioinformatics analysis at PHO and contributing laboratories. In future, it expected to serve as a platform to share a broader range of SARS-CoV-2 whole genome sequencing results as part of Ontario's COVID-19 Genomics Sequencing Network.

Target Audience

The tool is intended for genomics researchers and experts, epidemiologists, virologists and public health experts.

Purpose of the Tool

The visualization integrates SARS-CoV-2 WGS data with other data types including demographic, spatial and temporal information. This tool can be used to visualize the emergence and growth patterns of different lineages of SARS-CoV-2 positive samples, and enable the tracking of variants of concern and other lineages in Ontario over time, by health region, gender and age groups.

Frequency of Data Updates

All data are updated weekly on Wednesdays and includes sequences released to the GISAID public repository.

Data Source

This tool reflects samples that are SARS-CoV-2 positive and sequenced by PHO Laboratory. At this time, the data is not representative of Ontario overall.

Descriptive Measures

Date. The date a sample was first logged at PHO Laboratory for testing (SARS-CoV-2, VoC PCR, or WGS).

Country. This represents the country in which the SARS-CoV-2 sample was collected and tested, and will be limited to samples in Ontario, Canada.

Pango Lineage. As part of PHO's bioinformatic processing, all genomic sequences are routinely analyzed using the pangolin tool[2] and receive a Pango lineage designation. The dynamic Pango nomenclature system allows for the classification of SARS-CoV-2 genetic diversity. Reassignments of Pango lineages are expected to occur as more SARS-CoV-2 genome sequences become available, thus providing a more complete picture of SAS-CoV-2 evolution and more accurate lineage assignment.

VoC Lineage. The Variant of Concern (VoC) lineage is derived from the Pango lineage designation to visually distinguish in Nextstrain between VoC lineages and non-VoC ineages. VoCs are established nationally and internationally based on the presence of key mutations and evidence of clinical and public health impact, such as increased transmissibility, disease severity, or impact on vaccine and therapeutic efficacy.

Nextstrain Clade. A Year-Letter nomenclature used by Nextstrain to identify major global lineages. See Nextstrain documentation for further details.[3]

GISAID Clade. Clade designation generated by the genome sequence repository known as GISAID. Clade designations are informed by the statistical distribution of genome distances in phylogenetic clusters. See GISAID documentation for further details.[4]

Genotype. Selecting this data field provides the user with an option to select a SARS-CoV-2 gene of interest and specify a gene or genome position to colour the samples based on the amino acid or nucleotide present at that position in each sequence.

Ontario Health Region. Groupings of public health units. Samples are assigned to a public health unit based on the individual's address or, if unknown, the address of the ordering provider. Public health units are assigned to health regions as follows:

- North West - Northwestern Health Unit; Thunder Bay District Health Unit
- North East - Algoma Public Health; North Bay Parry Sound District Health Unit; Porcupine Health Unit; Public Health Sudbury & Districts; Timiskaming Health Unit
- Eastern - Ottawa Public Health; Eastern Ontario Health Unit; Hastings Prince Edward Public Health; Kingston, Frontenac and Lennox & Addington Public Health; Leeds, Grenville & Lanark District Health Unit; Renfrew County and District Health Unit
- Central East - Durham Region Health Department; Haliburton, Kawartha, Pine Ridge District Health Unit; Peel Public Health; Peterborough Public Health; Simcoe Muskoka District Health Unit; York Region Public Health
- Toronto - Toronto Public Health
- South West - Chatham-Kent Public Health; Grey Bruce Health Unit; Huron Perth Public Health; Lambton Public Health; Middlesex-London Health Unit; Southwestern Public Health; Windsor-Essex County Health Unit
- Central West - Brant County Health Unit; City of Hamilton Public Health Services; Haldimand-Norfolk Health Unit; Halton Region Public Health; Niagara Region Public Health; Region of Waterloo Public Health and Emergency Services; Wellington-Dufferin-Guelph Public Health

Age group. Determined using Date (defined above) and date of birth as reported on the test requisition. Age groups include: 19 and under, 20-39, 40-59, 60-79, 80 and over, and unknown where date of birth was not reported.

Sex. This uses sex of the individual at the time of sample collection as reported on the test requisition. Includes male, female, or unknown (did not specify male or female).

Tool Overview

Summary

The web-based interface contains a number of panels that allow the user to display the relationship between the phylogenetic tree, the map, and variability across the genome. Colors remain consistent throughout the application to link different panels. Within the side bar there are controls for the different panel displays, for example any of the panels can be turned off. There are also data controls to change the data to be visualized. A number of controls are available in a sidebar to filter the time period viewed, the layout of the tree, and other views. Nextstrain also has a number of languages included in the tool, for example French or Spanish, which can be selected from the dropdown at the bottom of the left panel. For further information, see Nextstrain official documentation.[5]

Phylogenetic Tree

Nextstrain presents a phylogenetic tree, which is a visual representation of the inferred evolutionary relationship among different viral samples. The relationship is inferred based on their mutation profiles and other sample data such as sample date or health region can be viewed on the phylogenetic tree. The user is able to examine a set of possible relationships among different samples clustered based on time separation or number of mutations, which can help provide context to different SARS-CoV-2 lineages and their possible place and time of origin in Ontario.

Using the various tree options on the left-hand side of the Nextstrain application, the user has the option to view the orientation and content of the phylogenetic tree with many different configurations. This includes visual filtering of samples by both date and divergence, which is defined as the difference in mutation profiles between a samples and the reference sequence, the first COVID-19 genome that was fully sequenced from Wuhan, China in early 2020. Using divergence as the tree orientation, the user can gain an appreciation for the changes in sample mutation profiles that occur over time, and identify how the accumulation of certain mutations can drive the formation of different lineages across geographic regions. Additionally, data filters allow for only samples that share a common characteristic to be highlighted in the tree. At any point, the user can choose to restore the default settings of the tree by resetting the layout to begin a new investigation.

The tree can be zoomed in by clicking on branches, and the RESET LAYOUT option above the tree allows the view to be reset again. Additionally, the tree enables the user to both hover over a particular sample as well as click on it to display additional information. Fields that are displayed in pop-up box include its unique PHO WGS identifier, the year and month of collection, the global lineage assignment, and various map designations such as country or region. Clicking outside of the pop-up box allows the complete map view to be restored.

For more detailed information on how to analyze a phylogenetic tree, including various definitions used to describe phylogenetic trees, please refer to the Nextstrain official documentation.[6]

TREE OPTIONS

Layout. Several layout options are available for tree visualization, including rooted options – rectangular and radial, and an unrooted option. The clock view plots the time from collection of the reference genome (Genbank No. MN908947.3) against divergence (mutations) of each sequence to estimate the number of mutations per year.

Branch. Tree branches can be set to none, clade (Nextstrain Clade) or aa (amino acid). The labels will be placed at the branch node.

Tip. Tip labels can be set to display any of the data fields in the dropdown menu, e.g., Nextstrain Clade. The labels will become visible when a tree is zoomed in.

Branch length – Time. The horizontal distance of each sequenced sample from the reference genome can be calculated and visualized based on time, i.e., date of PHO sample from December 31, 2019 (collection date of the reference genome).

Branch length – Divergence. The horizontal distance of each sequenced sample from the reference genome can be calculated and visualized based on divergence, i.e. number of mutations in the sequenced genome of a PHO sample compared to the reference genome.

Map

The geography panel in the Nextstrain build includes a map that accompanies the information presented in the phylogenetic tree. Depending on the "Map Options" filter(s) that the user has set, the map displays a pie graph for each Ontario Health Region in which the individuals for which the SARS-CoV-2 samples were collected, coloured by the chosen characteristic e.g. Pango lineage.

Circles highlighting different areas on the map represent the relative number of sequences that were identified in each Ontario Health Region. The user can apply different filters on the left-hand side panel to modify the number of samples from a particular region based on set characteristics (e.g. view the number of sequences of lineage B.1.1.7 in the Central West Region). In this manner, the map can provide a visual comparison of the differences in SARS-CoV-2 trends that can be seen across Health Regions. Furthermore, the inside of the circle will act as a pie chart to display the proportion of samples with a filter characteristic (i.e. when coloring the map by lineage, the circles over each Health Region will reflect the proportion of samples from each lineage that correspond to that location). Through a combination of the circle size and its interior pie chart, the user is able to quickly identify trends and patterns among the samples that are specific to an Ontario Health Region.

For more detailed information on how to analyze the Nextstrain map, please refer to the following section in the Nextstrain official documentation.[7]

Diversity and Entropy

The diversity panel, situated just below the phylogenetic tree and map, represents the relative amount of genetic diversity seen across the SARS-CoV-2 genome for the samples that are displayed in the tree. Accordingly, this panel will change as the user zooms in on specific subsets of the dataset. The relative heights of the bars in the panel represent the "entropy" of that particular location, which can be considered as the relative level of change that a particular genomic location demonstrates across the samples. Genomic sites with larger bars correspond to sites where more genetic variation has been observed in the dataset, and therefore are considered to have greater entropy. This panel enables the user to identify regions of the genome that are more prone to mutation, and when combined with a

lineage filter or another type of sample characteristic, may highlight the mutation patterns for different lineages or different time points in the dataset.

Users interested in a specific genomic position can hover over the corresponding bar in the diversity panel and view its amino acid and nucleotide information. Furthermore, clicking on the bar will generate a filter showing the different possible mutations at that site and where they are found both in the tree and on the map. This filter can be reset by selecting a new filter from the "Color By" drop down field on the left-hand side panel.

Below the diversity panel is a visual representation of the different genes in the SARS-CoV-2 genome and their positions relative to each other. Users may find this additional information useful to link a specific genomic position to its corresponding gene. The gene name is often used in media and press releases to describe the location of a mutation of interest, therefore users may find these panels useful in understanding the behaviour of reported mutations and emerging mutations of interest (i.e. the spike protein (S) is often referenced in the media in relation to a number of emerging global variants).

Further Details on Data Sources

- Demographic information is extracted from PHO's Laboratory Information Management System at the time the sample is selected for genome sequencing.

- Whole genome sequences are biological data that have been generated through the genome sequencing of SARS-CoV-2 samples submitted to PHO Laboratory and bioinformatically analyzed.[8,9] Results of which are maintained within PHO's BioComputing cluster and are extracted at **4pm** on the day prior to posting to PHO's Nexstrain build.

- Lineage and clade data are generated through PHO's bioinformatic processing of all SARS-CoV-2 genomic sequences which are routinely analyzed using the pangolin tool, nextclade tool and through deposit to the GISAID repository which provides the GISAID clade information.[2,10] The variant of concern (VoC) lineage is derived from the Pango lineage designation to visually distinguish in Nextstrain between VoC lineages and non-VoC ineages. Data are stored in PHO's genomic database and are extracted at **4pm** on the day prior to posting to PHO's Nexstrain build.

Data Limitations

All information presented through the various panels and visual aids in Nextstrain are intended to represent the possible associations between different sample groups based on observed data and WGS results. Therefore, any conclusions or inferences that may be suggested by the data should not be considered as absolute. Samples that are displayed in Nextstrain meet a minimum data quality threshold as set by PHO to generate a reasonable conclusion and prevent the possibility of serious misinterpretation by users. However, it is important that users appreciate the limitations, and therefore, the overall usability of the data presented through the application. The limitations of the data fields and contents of Nextstrain are further described here.

- Only samples sequenced by PHO Laboratory are included and therefore, results do not represent the province overall. A sub-sample of all sequences are used for display of up to 5,000 sequences in the Nextstrain application as visualization of all samples would be too dense to be useful.

- The criteria used to select samples to sequence has not been consistent over time and as such results should be interpreted with caution. Beginning February 3, 2021 all samples that screened positive for a VoC (N501Y mutation detected) were transferred to the genomics group for sequencing, which has biased the samples included.

- Date reflects the date the sample was first received and logged into PHO's Laboratory Information Management System. There may be a delay between the sample being collected and logged at PHO Laboratory, especially if originally tested for SARS-CoV-2 at another laboratory.

- Health region, age group, and sexare based on information provided on test requisitions. As such, they are subject to potential errors in requisition completion or data entry, which may impact accuracy.

- Health region is assigned using an individual's postal code of residence. If a residential postal code has not been provided, the postal code of the ordering provider is used, which creates the potential for regional misattribution of samples.

- Genome sequences and case data are sample-based. As such, it is possible that more than one sample was sequenced per individual.

- Approximately 10% of samples attempted are unsuccesfully sequenced, these are more likely to represent samples with low viral quantities. Therefore, the sequencing results may be potentially biased towards variants causing higher viral loads and samples collected from people in earlier stages of the disease.

- The nomenclature system for Pango lineages, Nextstrain and GISAID clades are dynamic. As more information is available and global SARS-CoV-2 sequences are added to the public domain designations may change as the virus continues to evolve. With each weekly update of PHO's Nexstrain build any given sample(s) may have an updated lineage and/or clade assignment different from the previous assignment.

- It is very important to note that phylogenetic trees are NOT transmission networks; in fact, they represent a possible model of interconnectivity based on statistical inference that are subject to change as more samples in Ontario are sequenced. Because these models are subject to sampling biases and other technical limitations, the relationships displayed among samples in a tree should not be considered by the user as absolute. Rather, the similarities and differences observed among the samples in a phylogenetic tree should be considered as the most likely representation of similarity based on the data available to PHO at the time of analysis.

- The statistical models used to generate the tree incorporate both genomic information and date to reasonably infer ancestral relationships, similarity groupings, and phylogenetic evolution. While PHO strives to include only data that have been reviewed and examined carefully and extensively, it is possible that sample information can very occasionally be incomplete or inaccurate. In those instances, certain elements of PHO's Nextstrain build may not fully reflect what is occurring in a particular region of Ontario at that given moment. As such, this application is suggested to provide a high-level guide on general trends for COVID-19 in the province and to demonstrate the SARS-CoV-2 WGS efforts that are currently underway at PHO.

References

1. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018;34(23):4121-3. Available from: https://doi.org/10.1093/bioinformatics/bty407

2. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol. 2020;5(11):1403-7. Available from: http://www.nature.com/articles/s41564-020-0770-5

3. Bedford T, Hodcroft EB, Neher RA. Updated Nextstrain SARS-CoV-2 clade naming strategy. 2021 Jan 06 [cited 2021 Mar 24]. In: Nextstrain blog [Internet]. Seattle, WA: Trevor Bedford and Richard Neher; [2019] – . Available from: https://nextstrain.org/blog/2021-01-06-updated-SARS-CoV-2-clade-naming

4. GISAID. Clade and lineage nomenclature, March 2, 2021: clade and lineage nomenclature aids in genomic epidemiology studies of active hCoV-19 viruses [Internet]. Munich: Freunde von GISAID e.V.; 2021 [cited 2021 Mar 24]. Available from: https://www.gisaid.org/references/statements-clarifications/clade-and-lineage-nomenclature-aids-in-genomic-epidemiology-of-active-hcov-19-viruses/

5. Nextstrain; Bedford T, Neher R, Hadfield J, Hodcroft E, Sibley T, Huddleston J, et al. Welcome to Nextstrain's documentation! [Internet]. Seattle, WA: Trevor Bedford and Richard Neher; 2020 [cited 2021 Mar 24]. Available from: https://docs.nextstrain.org/en/latest/index.html

6. Nextstrain; Bedford T, Neher R, Hadfield J, Hodcroft E, Sibley T, Huddleston J, et al. How to interpret the phylogenetic trees: transmission trees vs phylogenetic trees [Internet]. Seattle, WA: Trevor Bedford and Richard Neher; 2020 [cited 2021 Mar 24]. Available from: https://docs.nextstrain.org/en/latest/learn/interpret/how-to-read-a-tree.html

7. Nextstrain; Bedford T, Neher R, Hadfield J, Hodcroft E, Sibley T, Huddleston J, et al. How to interpret what's shown on the map [Internet]. Seattle, WA: Trevor Bedford and Richard Neher; 2020 [cited 2021 Mar 24]. Available from: https://docs.nextstrain.org/en/latest/learn/interpret/map-interpretation.html

8. Genome Sequence Informatics. ncov2019-artic-nf [Internet]. 2020 [cited 2021 Mar 24]. GitHub. Available from: http://github.com/oicr-gsi/ncov2019-artic-nf

9. Simpson J. ncov-tools [Internet]. 2020 [cited 2020 Nov 24]. GitHub. Available from: http://github.com/jts/ncov-tools

10. Nextstrain. nextclade [Internet]. 2020 [cited 2021 Mar 15]. GitHub. Available from: https://github.com/nextstrain/nextclade

Disclaimer

This document was developed by Public Health Ontario (PHO). PHO provides scientific and technical advice to Ontario's government, public health organizations and health care providers. PHO's work is guided by the current best available evidence at the time of publication.

The application and use of this document is the responsibility of the user. PHO assumes no liability resulting from any such application or use.

Citation

Ontario Agency for Health Protection and Promotion (Public Health Ontario). Technical notes: Phylogenetic analysis of SARS-CoV-2 in Ontario (Nextstrain interface hosted at Public Health Ontario). Toronto, ON: Queen's Printer for Ontario; 2021.

For Further Information

For more information, email communications@oahpp.ca.

Public Health Ontario

Public Health Ontario is an agency of the Government of Ontario dedicated to protecting and promoting the health of all Ontarians and reducing inequities in health. Public Health Ontario links public health practitioners, front-line health workers and researchers to the best scientific intelligence and knowledge from around the world.

For more information about PHO, visit publichealthontario.ca.

Ontario