Public Health Alliance for Genomic Epidemiology

PHA4GE Newsletter

October 2020

Editorial

A frequently used phrase has to be "the new normal" besides "you are muted". Yet it does not do justice to the mayhem, chaos, stress and survival-mode that many are experiencing at this time around the world. However, there appears to be an alignment of initiatives (finally) as global partners recognize the value of a combined coordinated effort to navigate through this COVID-19 pandemic. There is hope for those of us living with, and responding to, other epidemics such as HIV, Tuberculosis and Malaria. And the Public Health Alliance for Genomic Epidemiology (**PHA4GE**) has been actively engaged in a range of cross-border initiatives in this space with a view that spans much further in the horizon.

In this issue of the PHA4GE newsletter we hear from **Dr Emma Griffiths** (co-chair of the PHA4GE Data Structures working group) and their exciting recent work on developing **a metadata specification for SARS-COV-2 data collection**. This harmonization tool is well placed for many countries who are developing biospecimen collection protocols for projects directed towards a COVID-19 public health response and more generally for other pathogens.

Dr Nicki Tiffin, co-chair of the recently established Ethics and Data sharing working group describes the Ethics online forum that is under development, as a space for debate around ethical issues in research. The need for heightened ethics awareness during pandemics as outlined by the **African Academy of Sciences** resonates with the PHA4GE Ethics and Data Sharing working group and we encourage participation.

It has been exciting to witness, and be a part of, a pan-African strategy to establish a network of laboratories that are responding to the needs for human capacity and infrastructure resources at the **country-level** and across the **African continent**. These initiatives are well placed and positioned relative to other global COVID-19 genomic networks reported in **May 2020**.

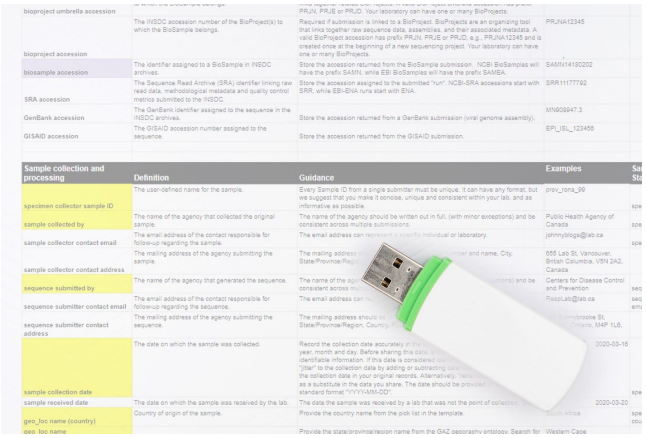
Don't forget to browse the events calendar for up-coming meetings pertinent to data standards and public health.

Getting the right information to the right people: The PHA4GE SARS-CoV-2 contextual data specification

The SARS-CoV-2 pandemic has impacted lives and economies all over the world, with more than 34.8 million cases and over 1 million deaths globally in early October 2020. Sequencing and bioinformatic analyses of viral genomes has already demonstrated many insights into the origin and spread of the disease, due to the ever-increasing amounts of data shared with public repositories like GISAID and the INSDC. Good quality genomics contextual data (sample metadata, lab/epidemiological/clinical data, methods and metrics) are critical for interpreting sequence data and informing decision making based on results, as well as answering biological questions about the virus and the disease. Contextual data elements such as sample collection dates and geographical locations, patient age, gender, health outcomes, pre-existing conditions, symptoms and onset dates, as well as possible and known exposures, are useful for a wide variety of surveillance and other public health activities. These include characterizing lineages and clusters, identifying variants with clinical significance, and correlating genomic trends with outcomes and risk factors.

In order to capitalize on the potential of SARS-CoV-2 sequence data, getting the right information to the right people is

critical, however this process is often hampered by fragmented data collection and management processes. Due to the division of labour across laboratories, departments, agencies and jurisdictions, contextual data is often collected according to local needs and reporting requirements, and structured according to organization-specific data dictionaries, creating data silos and barriers for data sharing. While metadata standards exist, they are broadly scoped to cover as many use cases and pathogens as possible, and include fields that may be subject to privacy concerns, may not be applicable to a pathogen of interest, or exclude fields commonly used in public health surveillance and investigations.



bioproject umbrella accession		PRJNA, PRJEB or PRJD. Your laboratory can have one or many BioProjects.	
bioproject accession	The INSDC accession number of the BioProject(s) to which the BioSample belongs.	Required if submission is linked to a BioProject. BioProjects are an organizing tool that links together raw sequence data, assemblies, and their associated metadata. A valid BioProject accession has prefix PRJN, PRJE or PRJD, e.g., PRJNA12345 and is created once at the beginning of a new sequencing project. Your laboratory can have one or many BioProjects.	PRJNA12345
biosample accession	The identifier assigned to a BioSample in INSDC archives.	Store the accession returned from the BioSample submission. NCBI BioSamples will have the prefix SAMN, while EBI BioSamples will have the prefix SAMEA.	SAMN14180202
SRA accession	The Sequence Read Archive (SRA) identifier linking raw read data, methodological metadata and quality control metrics submitted to the INSDC.	Store the accession assigned to the submitted "run". NCBI-SRA accessions start with SRR, while EBI-ENA runs start with ERA.	SRR11177792
GenBank accession	The GenBank identifier assigned to the sequence in the INSDC archives.	Store the accession returned from a GenBank submission (viral genome assembly).	MN908947.3
GISAID accession	The GISAID accession number assigned to the sequence.	Store the accession returned from the GISAID submission.	EPI_ISL_123456

Sample collection and processing	Definition	Guidance	Examples	Sa
specimen collector sample ID	The user-defined name for the sample.	Every Sample ID from a single submitter must be unique. It can have any format, but we suggest that you make it concise, unique and consistent within your lab, and as informative as possible.	prov_rona_99	spe
sample collected by	The name of the agency that collected the original sample.	The name of the agency should be written out in full, (with minor exceptions) and be consistent across multiple submissions.	Public Health Agency of Canada	spe
sample collector contact email	The email address of the contact responsible for follow-up regarding the sample.	The email address can represent a specific individual or laboratory.	johnnybloga@lab.ca	spe
sample collector contact address	The mailing address of the agency submitting the sample.	The mailing address [illegible] number and name, City, State/Province/Region, [illegible]	655 Lab St, Vancouver, British Columbia, V5N 2A2, Canada	spe
sequence submitted by	The name of the agency that generated the sequence.	The name of the ag[illegible]ions) and be consistent across mu[illegible]	Centers for Disease Control and Prevention	seq
sequence submitter contact email	The email address of the contact responsible for follow-up regarding the sequence.	The email address can [illegible]	RespLab@lab.ca	seq em
sequence submitter contact address	The mailing address of the agency submitting the sequence.	The mailing address should [illegible] State/Province/Region, Country, [illegible]	[illegible]nnybrooke St, [illegible]ronto, Ontario, M4P 1L6,	
sample collection date	The date on which the sample was collected.	Record the collection date accurately in the [illegible] year, month and day. Before sharing this data, [illegible] identifiable information. If this date is considered [illegible] "jitter" to the collection date by adding or subtracting a [illegible] the collection date in your original records. Alternatively, "[illegible] as a substitute in the data you share. The date should be provided [illegible] standard format "YYYY-MM-DD".	2020-03-16	spe
sample received date	The date on which the sample was received by the lab.	The date the sample was received by a lab that was not the point of collection [illegible]	2020-03-20	
geo_loc name (country)	Country of origin of the sample.	Provide the country name from the pick list in the template.	South Africa	spe cou
geo_loc name		Provide the state/province/region name from the GAZ geography ontology. Search for	Western Cape	

Many of the members of the Data Structures working group are part of large sequencing consortia (e.g. COG-UK, SPHERES, CanCOGeN, the Latin American Genomics SARS-CoV-2 Network) that have faced challenges in data harmonization and integration as a result of the barriers described above. In light of these challenges, we have developed a fit-for-purpose SARS-CoV-2 contextual data specification focused on public health needs, designed to accommodate privacy requirements while maximizing information linkage, content and interoperability across datasets and databases (**https://cutt.ly/AgtJyrf**).

The specification was developed by consensus among domain experts, and incorporates existing community standards to describe repository accession numbers and identifiers, sample collection and processing, host information, host exposure information, sequencing methods, bioinformatics and quality control metrics, pathogen diagnostic testing details, as well as provenance and contribution attribution.

The specification package includes:

Components of the PHA4GE SARS-CoV-2 Contextual Data Specification Package
Standardized collection template **Pick lists**: standardized **Reference guide**: field labels, definitions, guidance, expected values, required vs optional fields **SOP**: how to use template, find new terms, highlights practical/ethical/privacy issues **Field mapping to existing standards**: highlight alignment and gap **JSON schema**: machine readable version for incorporation into different applications **7 public repository submission protocols** (GISAID, NCBI, EMBL-EBI) on **protocols.io**

The collection template enables vital information to be collated in a single location, and harmonized across various sources using established principles to improve machine-amenability. Different subsets of the harmonized data can be 1) shared with public repositories e.g. GISAID and INSDC using the PHA4GE protocols, 2) shared with trusted partners e.g. national sequencing consortia, public health partners, and 3) kept private and retained locally with the potential for sharing in the future for particular surveillance or research activities. How, and how much of, the specification is used is ultimately at the discretion of the user. To date, versions of the specification are being implemented in the CanCOGeN (Canada) and SPHERES (USA) SARS-CoV-2 sequencing initiatives, the AusTrakka (Australia) national data sharing platform, by the Global Emerging Pathogens Treatment Consortium (Africa), and in the Baobab LIMS at the South African National Bioinformatics Institute (SANBI).

As countries around the world prepare for new waves of infections throughout the pandemic, a unique opportunity for harmonization in data collection exists. With this specification we have endeavored to create a mechanism for promoting consistent, standardized contextual data collection that can be applied in such a way that community-based data sharing efforts are not excessively burdened. We hope that, given sufficient uptake, this specification will enhance the reusability of collected data, enabling national and international agencies to accelerate the understanding of SARS-CoV-2

epidemiology and biology. Furthermore, the framework for SARS-CoV-2 presented in this work can also be used to build a roadmap for dealing with future public health crises.

To learn more about the specification and how to get started using it, read our **recent preprint** (**https://cutt.ly/YgtJgKT**) or listen to our interview on the **Micro Binfie podcast** (**https://cutt.ly/3gtJk8z**).

To learn more about other activities of the PHA4GE Data Structures workgroup, check out our webpage **https://pha4ge.org/work-group/data-structures**

The Ethics and Data Sharing Working Group is live!

The Ethics and Data Sharing Working Group convened in August 2020; starting off with six members. Whilst the working group's Terms of Reference are still under discussion, some ideas are already shaping their mandate. The group envisions a non-hierarchical structure that promotes inclusive interactions and activities within the ethics domain. Through informal interactions and peer-to-peer learning, substantial ethics skills and knowledge may be honed and could also assist researchers with varied levels of expertise in engaging with ethical issues in research. An opportunity is also created to encourage young researchers and those new to the field of ethics to join the conversation.

An online platform, accessed through a standard web browser, is under development for these interactions. It will necessitate open, honest and frank discussions regarding ethical issues and dilemmas and how research ethics should be implemented. Such conversations, hopefully, will bring the research and ethics communities closer. The envisioned community will drive and guide the forum and this includes

ethicists, researchers, scientists and policy-makers from all fields, disciplines across different countries. Apart from discussions on ethical issues, the platform may extend to information sharing on funding opportunities, training and collaboration opportunities including authors collaboration for new papers, flagging current relevant publications, and a platform to ask for ethics advice and offer guidance to other members of the community. However, the addition of new sections will continually evolve, based on suggestions from the user community.

Please contact Nicki at **nicki.tiffin@uct.ac.za** for more information.

Community

Leveraging resources for COVID-19 genome sequencing and analysis:

Accelerating genomics-based surveillance for COVID-19 response in Africa

https://cutt.ly/HgtHNO2

A genomics network established to respond rapidly to public health threats in South Africa

https://cutt.ly/XgtH6YD

WHO and Africa CDC announce sequencing laboratory network in response to COVID-19

https://cutt.ly/wgtJwBW

WHO chooses SANBI at UWC as national reference lab to join the fight against COVID-19 in Africa

https://cutt.ly/agtJrsv

Secretariat News

We wish to extend a warm welcome to the newest member of the PHA4GE family, Nawaal Nacerodien who has joined the Secretariat Group as an Administrator. We'd also like to thank Anja Bedeker for the administrative support she has provided over the last few months. Anja will continue to work closely with the Data Sharing and Ethics Working Group.

Anja Bedeker and Nawaal Nacerodien

Interested in being featured in our newsletter?

Send an email to communications@pha4ge.org

Web edition: https://cutt.ly/zgtZ5vO

Website: pha4ge.org

Twitter: @pha4ge

Facebook: facebook.com/pha4ge

Events

The GA4GH 8th Plenary Meeting took place recently and there are few upcoming events that may be of interest:

Global Grand Challenges	**2020 Grand Challenges Annual Meeting** 19 - 21 October 2020 Learn More https://cutt.ly/tgtHF7w
ASTMH	**ASTMH Annual Meeting** 15 - 19 November 2020 Learn More https://cutt.ly/egtHKY6
GO FAIR CODATA	**International FAIR Convergence Symposium** 30 November - 4 December 2020 Learn More https://cutt.ly/NgtHL9B
AMERICAN SOCIETY FOR MICROBIOLOGY	**ASM Conference on Rapid Applied Microbial Next-Generation Sequencing and Bioinformatic Pipelines** 7 - 11 December 2020 Learn More https://cutt.ly/lgtHC5H