Overcoming challenges of SARS-CoV-2 genomics data sharing for public health surveillance, outbreak investigations and research using the PHA4GE SARS-CoV-2 contextual data specification

Emma Griffiths, Ruth Timme, Finlay Maguire, Ines Mendes, Lee Katz, Damion Dooley, Rhiannon Cameron, Dominique Anderson, Anders Gonçalves da Silva, William Hsiao, Duncan MacCannell

Public Health Alliance for Genomic Epidemiology

Housekeeping

1. Session is being recorded
2. Please keep mics muted until Q&A
3. Please put questions in the chat
4. Please keep cameras off if internet unstable/not presenting
5. Keep phone/apps on silent
6. Slides will be made available after workshop
7. If you'd like to tweet #FAIRConvergence

Who Are We?

Public Health Alliance for Genomic Epidemiology

Workshop Overview

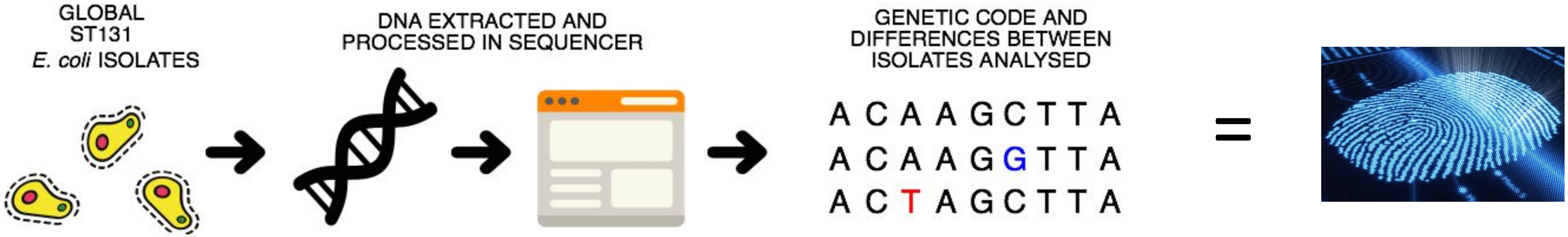
1. **Public health microbial genomics**
    - Importance for COVID-19 response
    - Challenges in data harmonization/integration
    - Overview of PHA4GE SARS-CoV-2 specification package
    - How PHA4GE specification makes genomics contextual data FAIR

2. **Demo of spec**: putting standards into practice
    - from chaos to control
    - improving the quality of open data

3. **Implementations of specification**
    - DataHarmonizer (Canada)
    - AusTrakka (Australia)
    - Boabab LIMS (South Africa)

Public Health Alliance for Genomic Epidemiology

Microbial genome sequences can used as a molecular fingerprint to trace the source of infectious disease.



- Public health agencies exchange information about these fingerprints

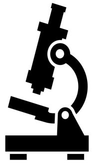
(Dramatic representation from the movie Outbreak)

Contextual data is critical for interpreting the sequence data.

Sequence data

Contextual data

Sample metadata



Lab results

Clinical/Epi data

Methods

Contextual data (metadata) used for **surveillance** and **outbreak investigations**:

- **characterize** lineages and clusters
- identify variants with **clinical significance**
- correlate genomics trends with **outcomes, risk factors**
- **inform decision making** for public health responses and **monitor effects of interventions**

Public Health Alliance for Genomic Epidemiology

Sequencing and sharing of SARS-CoV-2 genomes has had many benefits during the pandemic.

Check for updates

A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants

Bethany Dearlove, Eric Lewitus, Hongjun Bai, Yifan Li, Daniel B. Reeves, M. Gordon Joyce, Paul T. Scott, Mihret F. Amare, Sandhya Vasan, Nelson L. Michael, Kayvon Modjarrad, and Morgane Rolland

PNAS September 22, 2020 117 (38) 23652-23662; first published August 31, 2020;
https://doi.org/10.1073/pnas.2008281117

The proximal origin of SARS-CoV-2

Kristian G. Andersen, Andrew Rambaut, W. Ian Lipkin, Edward C. Holmes & Robert F. Garry

Nature Medicine **26**, 450–452(2020) | Cite this article

To the Editor – Since the first reports of novel pneumonia (COVID-19) in Wuhan, Hubei province, China[1,2], there has been considerable discussion on the origin of the causative virus, SARS-CoV-2[3] (also referred to as HCoV-19)[4]. Infections with SARS-CoV-2 are now widespread, and as of 11 March 2020, 121,564 cases have been confirmed in more than 110 countries, with 4,373 deaths[5].

SARS-CoV-2 is the seventh coronavirus known to infect humans; SARS-CoV, MERS-CoV and SARS-CoV-2 can cause severe disease, whereas HKU1, NL63, OC43 and 229E are associated with mild symptoms[6]. Here we review what can be deduced about the origin of SARS-CoV-2 from comparative analysis of genomic data. We offer a perspective on the notable features of the SARS-CoV-2 genome and discuss scenarios by which they could have arisen. Our analyses clearly show that SARS-CoV-2 is not a laboratory construct or a purposefully manipulated virus.

Cite as: X. Deng *et al.*, *Science*
10.1126/science.abb9263 (2020).

Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California

Xianding Deng[1,2*], Wei Gu[1,2*], Scot Federman[1,2*], Louis du Plessis[3*], Oliver G. Pybus[3], Nuno Faria[3], Candace Wang[1,2], Guixia Yu[1,2], Brian Bushnell[4], Chao-Yang Pan[5], Hugo Guevara[5], Alicia Sotomayor-Gonzalez[1,2], Kelsey Zorn[6], Allan Gopez[1], Venice Servellita[1], Elaine Hsu[1], Steve Miller[1], Trevor Bedford[7,8], Alexander L. Greninger[7,9], Pavitra Roychoudhury[7,9], Lea M. Starita[8,10], Michael Famulare[11], Helen Y. Chu[8,12], Jay Shendure[8,9,13], Keith R. Jerome[7,9], Catie Anderson[14], Karthik Gangavarapu[14], Mark Zeller[14], Emily Spencer[14], Kristian G. Andersen[14], Duncan MacCannell[15], Clinton R. Paden[15], Yan Li[15], Jing Zhang[15], Suxiang Tong[15], Gregory Armstrong[15], Scott Morrow[16], Matthew Willis[17], Bela T. Matyas[18], Sundari Mase[19], Olivia Kasirye[20], Maggie Park[21], Godfred Masinde[22], Curtis Chan[22], Alexander T. Yu[5], Shua J. Chai[5,15], Elsa Villarino[23], Brandon Bonin[23], Debra A. Wadford[5], Charles Y. Chiu[1,2,24†]

Comment on this paper

Large scale sequencing of SARS-CoV-2 genomes from one region allows detailed epidemiology and enables local outbreak management

Andrew J Page, Alison E Mather, Thanh Le Viet, Emma J Meader, Nabil-Fareed J Alikhan, Gemma L Kay, Leonardo de Oliveira Martins, Alp Aydin, David J Baker, Alexander J. Trotter, Steven Rudder, Ana P Tedim, Anastasia Kolyva, Rachael Stanley, Maria Diaz, Will Potter, Claire Stuart, Lizzie Meadows, Andrew Bell, Ana Victoria Gutierrez, Nicholas M Thomson, Evelien M Adriaenssens, Tracey Swingler, Rachel AJ Gilroy, Luke Griffith, Dheeraj K Sethi, Rose K Davidson, Robert A Kingsley, Luke Bedford, Lindsay J Coupland, Ian G Charles, Ngozi Elumogo, John Wain, Reenesh Prakash, Mark A Webber, SJ Louise Smith, Meera Chand, Samir Dervisevic, Justin O'Grady, The COVID-19 Genomics UK (COG-UK) consortium

doi: https://doi.org/10.1101/2020.09.28.20201475

Getting the right information to the right people is critical during health emergencies.

- Need to share data: **within** organization, with **trusted partners**, with **international** agencies/**public** repositories

Private databases:

Specimen Collected
☐ Upper respiratory (e.g., Nasopharyngeal or oropharyngeal swab)
☐ Lower respiratory (e.g., sputum, tracheal aspirate, BAL, pleural fluid)

6 - Specimen Type (check all that apply)

Specimen Collection Date: yyyy / mm / dd **(required)**

☐ NPS in UTM

☐ Throat Swab in UTM

☐ Other (Specify):

If possible:

☐ BAL

☐ Sputum

Public databases:

isolate	SARS-CoV-2/186197/human/2020/Malaysia
collected by	Universiti Malaya COVID Research group
collection date	14-Mar-2020
geographic location	Malaysia
host	Homo sapiens
host disease	COVID-19
isolation source	Nasopharyngeal/throat swab
latitude and longitude	3.1390 N 101.6869 E

source name	Lung sample from postmortem COVID-19 patient
cell type	Lung Biopsy
strain	NA
subject status	No treatment; >60 years old male COVID-19 deceased patient

The SARS-CoV-2 Contextual Data Specification

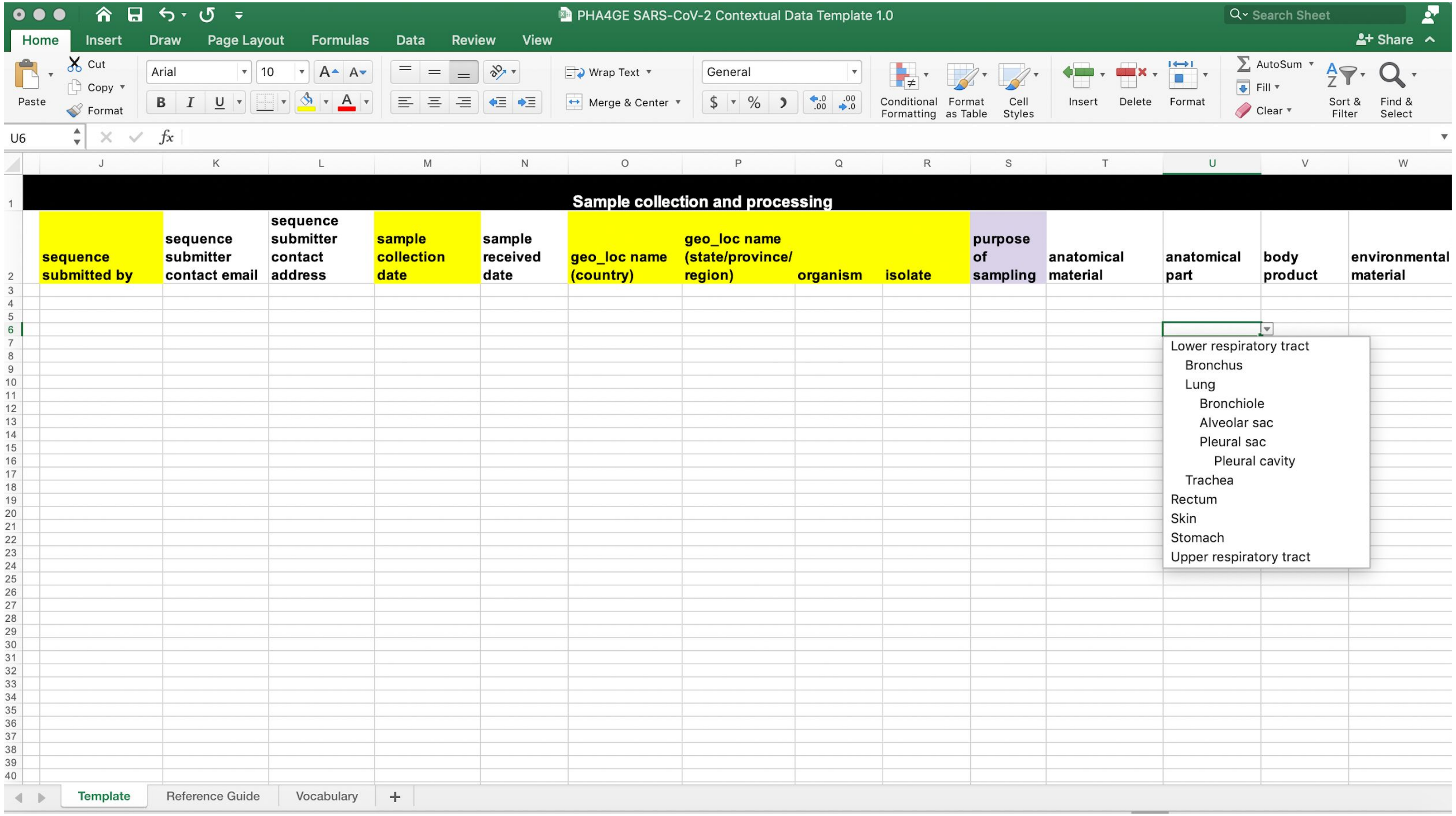
SARS-CoV-2 Specification Content
- Repository accession numbers and identifiers
- Sample collection and processing
- Host information
- Host exposure information
- Sequencing methods
- Bioinformatics and quality control metrics
- Pathogen diagnostic testing details
- Provenance and attribution

Data Sources
- Case report forms
- Public repository requirements
- Existing metadata standards
- Literature

Mapping to Standards
- MIxS 5.0
- MIGS Virus, Host-Associated
- Project/Sample Application Standard
- **OBO Foundry Ontologies**

Public Health Alliance for Genomic Epidemiology

Template and terminology

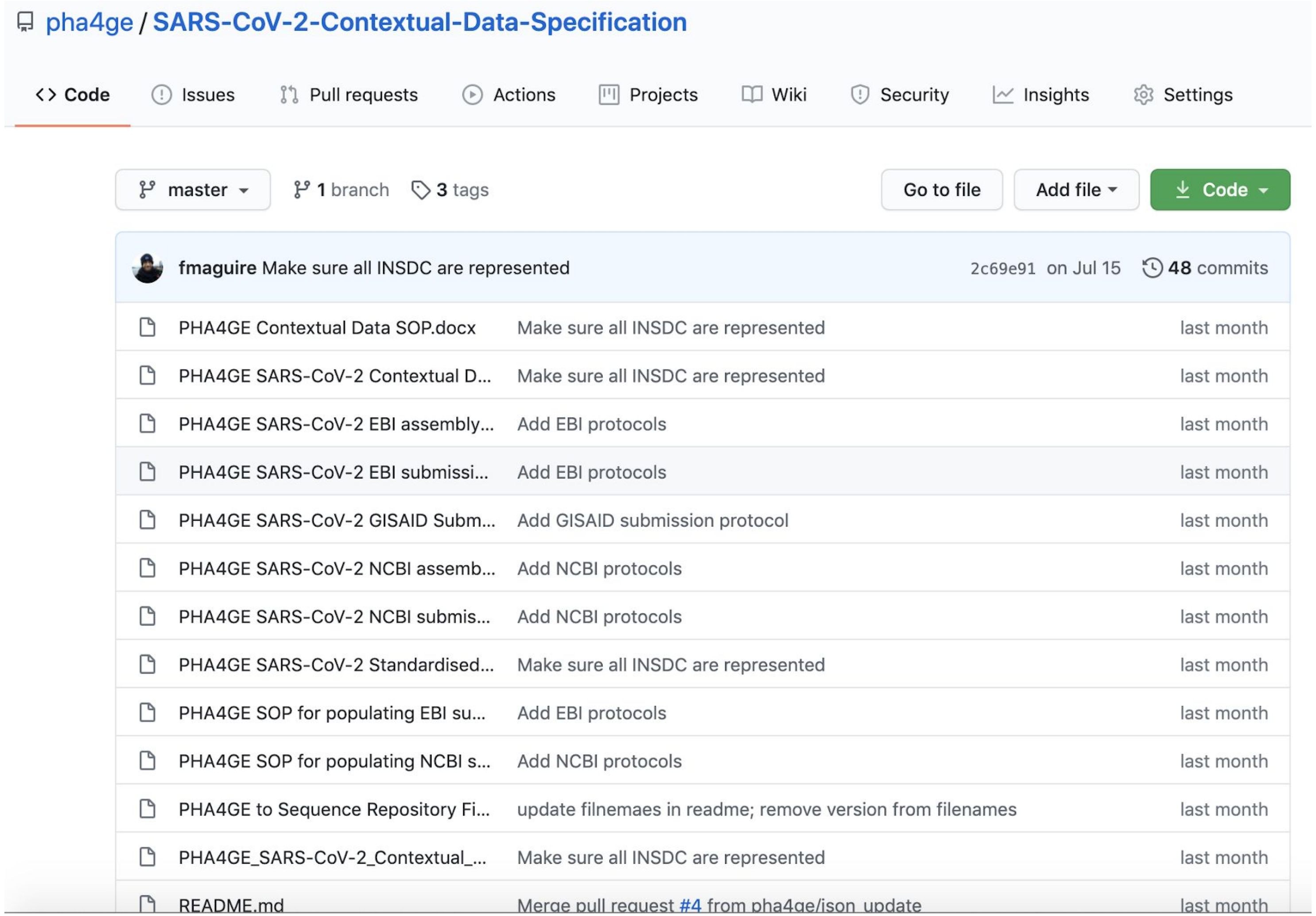


- **Standardized collection template** (colour-coded)
- **Pick lists**: standardized terms
- **Reference guide**: field labels, definitions, guidance, expected values

https://github.com/pha4ge/SARS-CoV-2-Contextual-Data-Specification

Public Health Alliance for Genomic Epidemiology

Supporting documentation

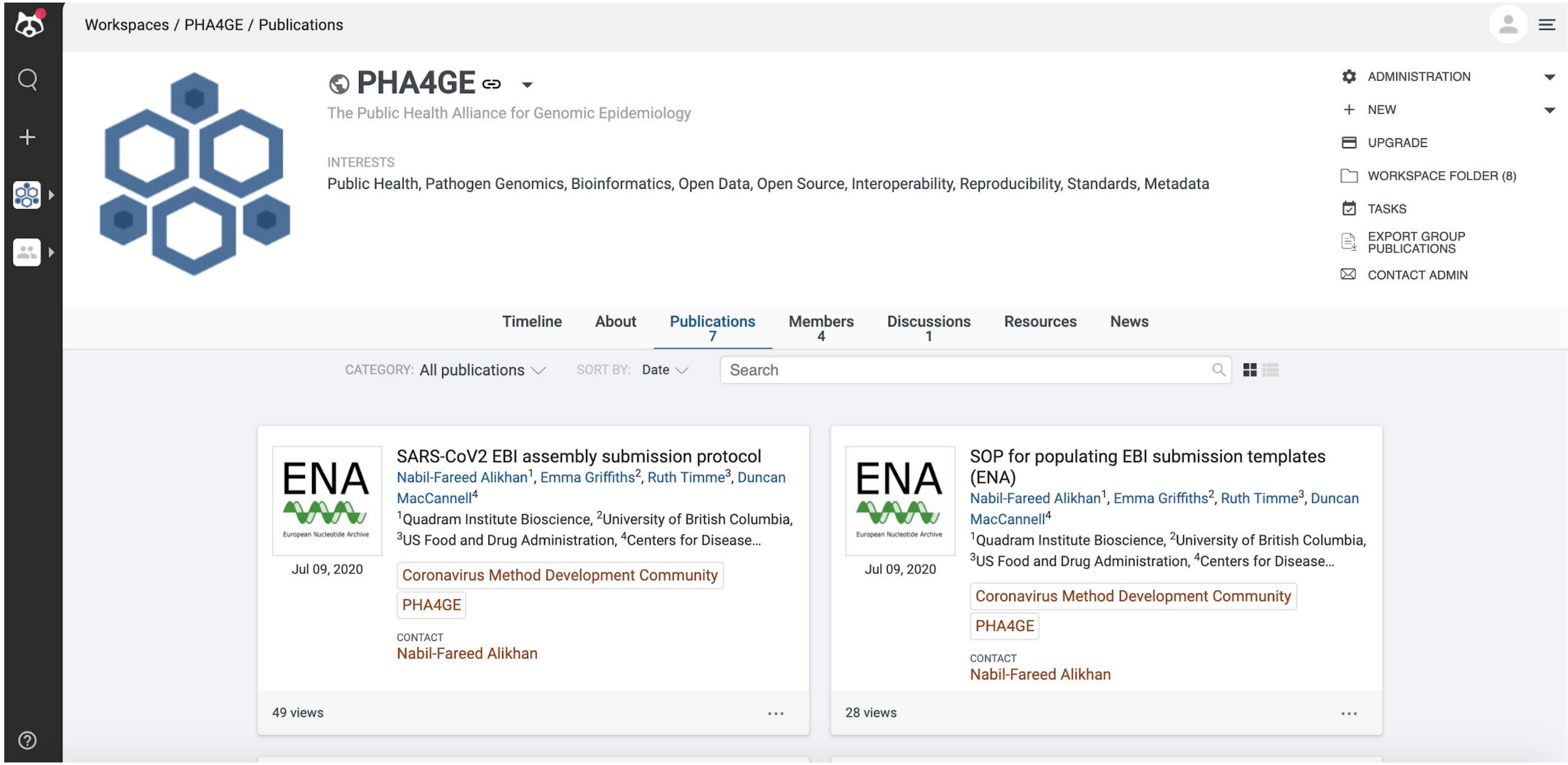


- **SOP**: how to use specification, find new terms, highlight practical/ethical/privacy issues
- **Field mapping to existing standards**: highlight alignment and gaps
- **JSON schema**: machine readable

https://github.com/pha4ge/SARS-CoV-2-Contextual-Data-Specification

Public Health Alliance for Genomic Epidemiology

Protocols to mobilize harmonized data



- **7 public repository submission protocols** (GISAID, NCBI, EMBL-EBI) on **Protocols.io**

https://www.protocols.io/workspaces/pha4ge

Preprints | HOW IT WORKS | INSTRUCTIONS FOR AUTHORS | SUBJECT AREAS | ADVISORY BOARD | SCREENING PREPRINTS | ABOUT

Advanced Searc

preprints.org > doi: 10.20944/preprints202008.0220.v1

https://www.preprints.org/manuscript/202008.0220/v1

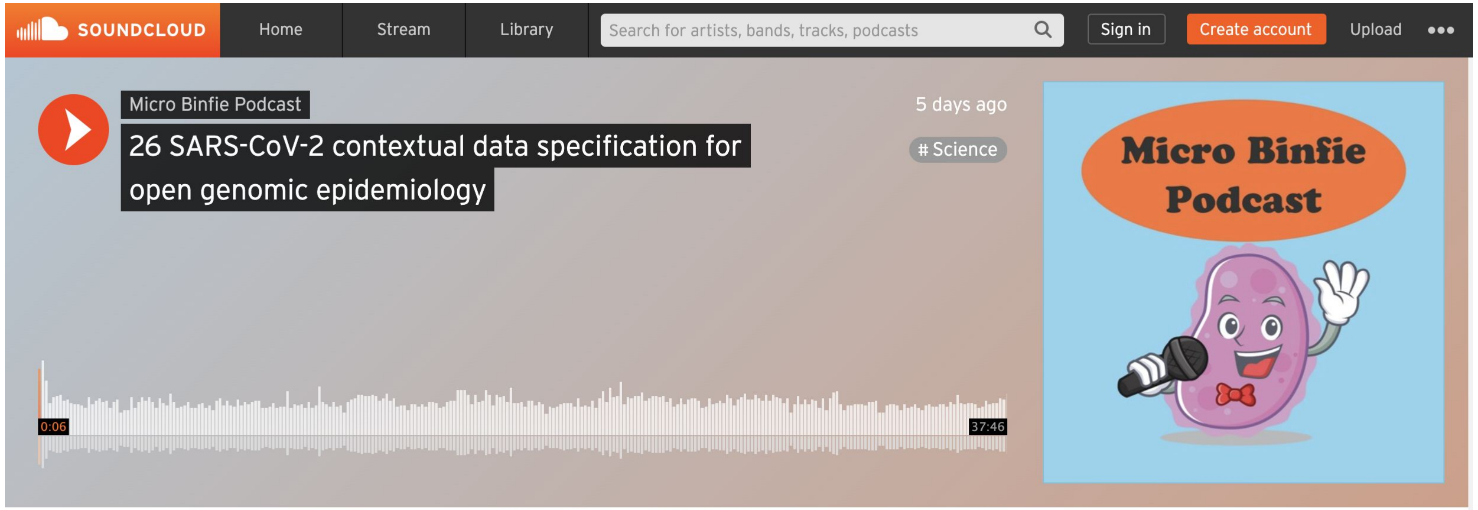
Preprint **Article** *Version 1* **This version is not peer-reviewed**

The PHA4GE SARS-CoV-2 Contextual Data Specification for Open Genomic Epidemiology

Emma J. Griffiths *, Ruth E. Timme, Andrew J. Page, Nabil-Fareed Alikhan, Dan Fornika, Finlay Maguire, Catarina Inês Mendes, Simon H. Tausch, Allison Black, Thomas R. Connor, Gregory H. Tyson, David M. Aanensen, Brian Alcock, Josefina Campos, Alan Christoffels, Anders Gonçalves da Silva, Emma Hodcroft, William W.L. Hsiao, Lee S. Katz, Samuel M. Nicholls, Paul E. Oluniyi, Idowu B. Olawoye, Amogelang R. Raphenya, Ana Tereza R. Vasconcelos, Adam A. Witney, Duncan R. MacCannell

Version 1 : Received: 7 August 2020 / Approved: 9 August 2020 / Online: 9 August 2020 (15:53:58 CEST)

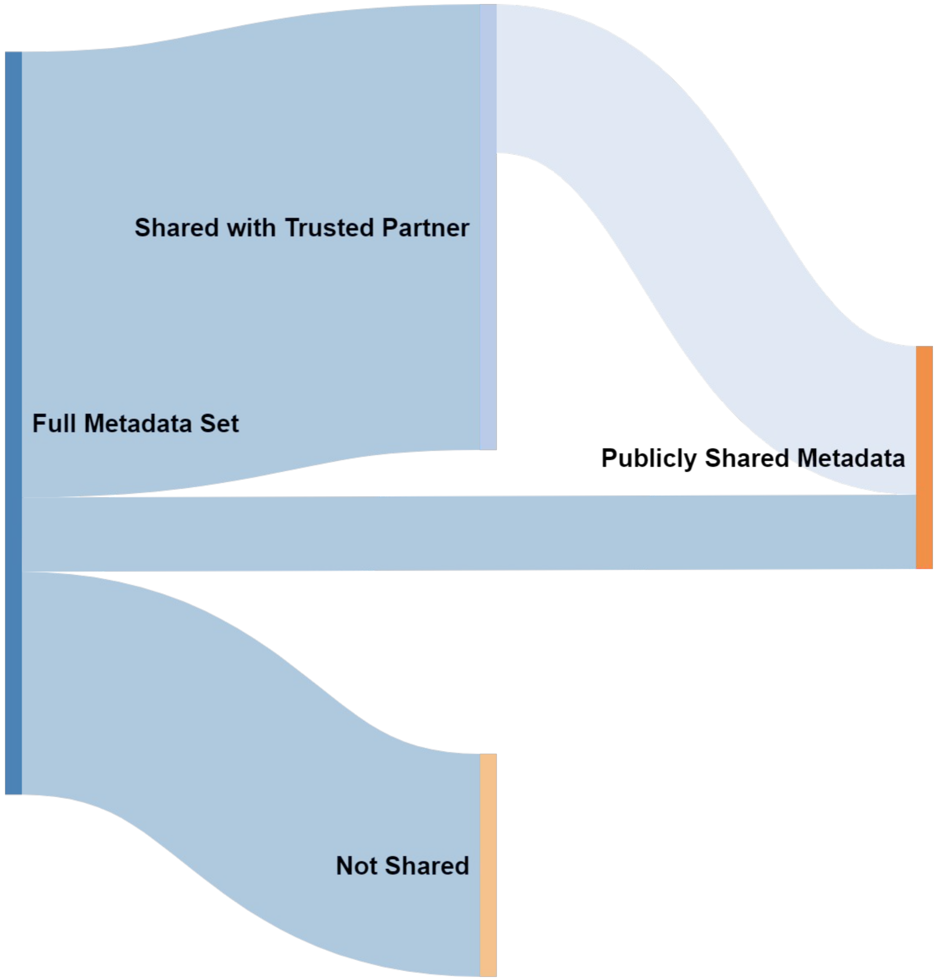
https://soundcloud.com/microbinfie/26-sars-cov-2-metadata#t=0:00



Public Health Alliance for Genomic Epidemiology

How do you use it?

- as much or as little as you want, it's up to you!
- structure metadata **consistently across labs**
- share with **public** repos, **trusted partners**, use for more **efficient private analyses**
- **future-proof** metadata



Public Health Alliance for Genomic Epidemiology

How does the PHA4GE Spec make public health genomics contextual data FAIR?

Findable – *every piece of information has a home, one stop shop*
- data elements standardized, not buried in methods
- ontologies offer URIs (unique, persistent identifiers)

Accessible – *understandable by humans/computers*
- spreadsheet and JSON
- protocols for storage in trusted repositories

Interoperable – *harmonization across users/standards*
- defines data structures for streamlined communication, data integration

Reusable - *enriched datasets*
- genomic information has many uses, enriched contextual data makes data fit for more purposes
- spec usage license (CCBY 4.0)

Public Health Alliance for Genomic Epidemiology

Putting standards into practice:

How to make data FAIR using the PHA4GE spec

Practical examples

- a) Harmonizing variable contextual data
- b) How to submit harmonized data to NCBI

Examples of implementation at organizations

- a) DataHarmonizer (Canada)
- b) Austrakka (Australia)
- c) Baobab LIMS (South Africa)

---Quick Q&A---

- Follow us on twitter
  - @BaobabLIMS
- Online documentation
  - https://media.readthedocs.org/pdf/baobab-lims/latest/baobab-lims.pdf
- *Website*
  - www.baobablims.org
- Get the code (and more)
  - https://github.com/BaobabLims
- Send us an email
  - Training – dominique@sanbi.ac.za
  - Helpdesk – help@baobablims.org

Summary

- spec for SARS-CoV-2 **public health contextual data** for **harmonization** across labs and datasets

- future-proof data

- **FAIR:** providing **consistent structure**, **human/machine-readable**, encourages **data sharing** in responsible way, linking information using **ontologies**

- used by members of **sequencing consortia**

- implemented in **different tools/platforms**

Public Health Alliance for Genomic Epidemiology

Thank you!

Data Structures Team
Ana Ribeiro de Vasconcelos
Josefina Campos
Idowu Olawoye
Paul Oluniyi
Adam Witney
Andrew Page
David Aanensen
Ines Mendes
Emma Hodcroft
Simon Tausch
Allison Black
Ruth Timme
Greg Tyson
Mike Feldgarden
Lee Katz
Brian Alcock
Amos Raphenya
Finlay Maguire
Dan Fornika
Duncan MacCannell

Specification Contributors & Partners
Nabil-Fareed Alikhan
Alan Christoffels
Will Hsiao
Sam Nicholls
Tom Connor
Anders Gonçalves da Silva
Dominique Anderson

Steering Committee & Secretariat
Jamie Southgate
Alecia Naidu
Rangarirai Matima
Nawaal Nacerodien
Peter Van Heusden
Danny Park

This work would not be possible without the contributions and dedication of these wonderful people.

Find us:
https://www.pha4ge.org
https://www.github.com/pha4ge
@pha4ge

Public Health Alliance for Genomic Epidemiology

Thank you!

BAOBAB LIMS

CanCOGeN VirusSeq

SPHERES

Public Health Alliance for Genomic Epidemiology

Thank you for listening and participating!

Public Health Alliance for
Genomic Epidemiology

Get the PHA4GE spec here
https://github.com/pha4ge/SARS-CoV-2-Contextual-Data-Specification

Get the preprint here
https://www.preprints.org/manuscript/202008.0220/v1

Get the DataHarmonizer here
https://github.com/Public-Health-Bioinformatics/DataHarmonizer/releases/

Learn about AusTrakka
https://portal.austrakka.net.au/

Learn about Baobab LIMS
https://github.com/BaobabLims

Contact us
datastructures@pha4ge.org

Public Health Alliance for
Genomic Epidemiology