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

Emma Griffiths, PhD

Hsiao Public Health Bioinformatics Lab

Department of Pathology and Laboratory Medicine

University of British Columbia

Vancouver, Canada

PUBLIC HEALTH BIOINFORMATICS

Public Health Alliance for Genomic Epidemiology

Sequencing SARS-CoV-2 genomes has helped track the spread of the virus worldwide.

Sequence data

Contextual data

Sample metadata

Lab results

Clinical/Epi data

Methods toolbox

Methods

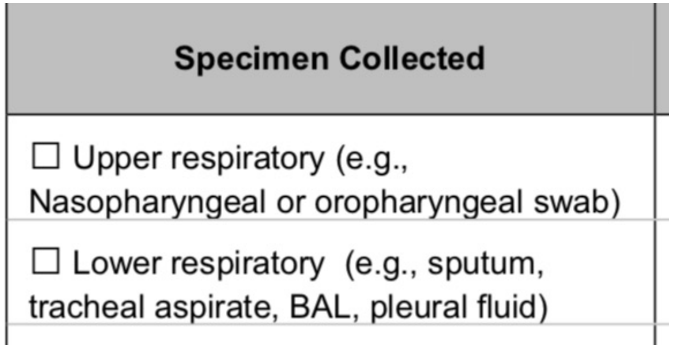
Contextual data (metadata) used for **surveillance** and **outbreak investigations**:

- **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

Challenge: How data is encoded impacts how it can be integrated and used for analyses.

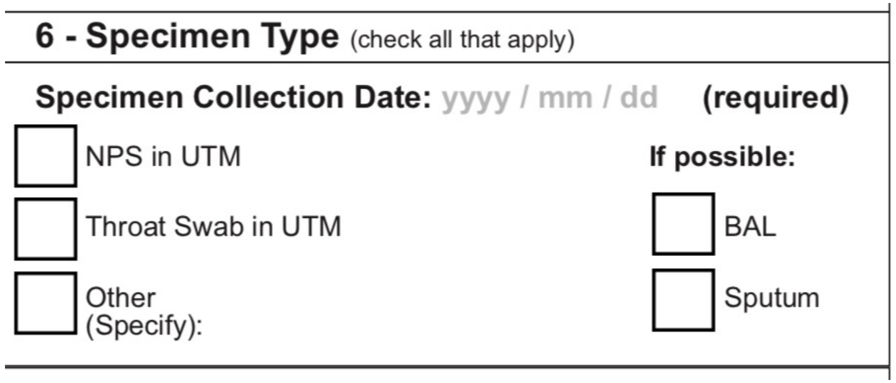
Specimen:



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

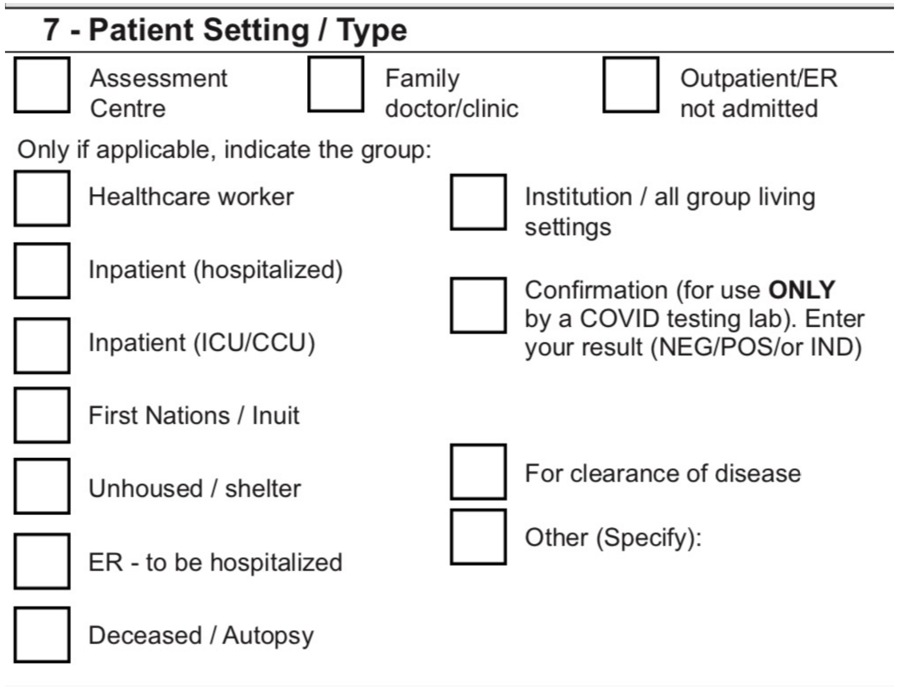
□ Other (Specify):

If possible:

□ BAL

□ Sputum

Patient Setting:



7 - Patient Setting / Type

□ Assessment Centre

□ Family doctor/clinic

□ Outpatient/ER not admitted

Only if applicable, indicate the group:

□ Healthcare worker

□ Inpatient (hospitalized)

□ Inpatient (ICU/CCU)

□ First Nations / Inuit

□ Unhoused / shelter

□ ER - to be hospitalized

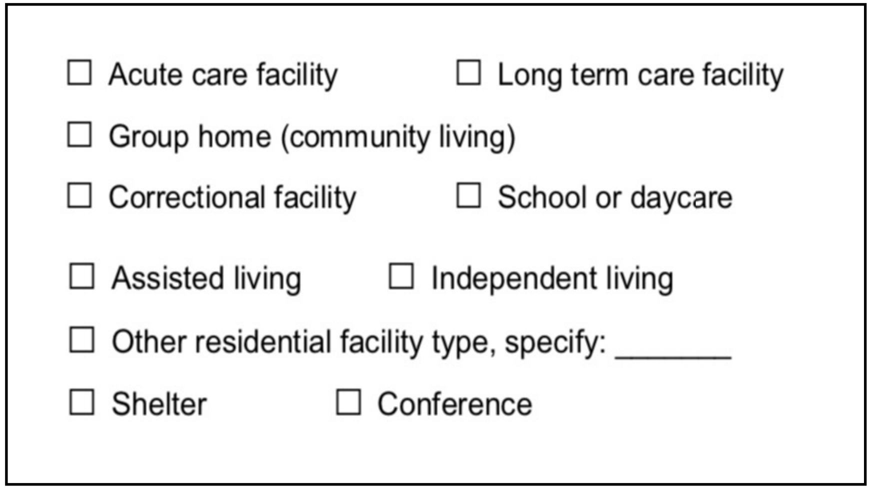
□ Deceased / Autopsy

□ Institution / all group living settings

□ Confirmation (for use **ONLY** by a COVID testing lab). Enter your result (NEG/POS/or IND)

□ For clearance of disease

□ Other (Specify):



□ Acute care facility □ Long term care facility

□ Group home (community living)

□ Correctional facility □ School or daycare

□ Assisted living □ Independent living

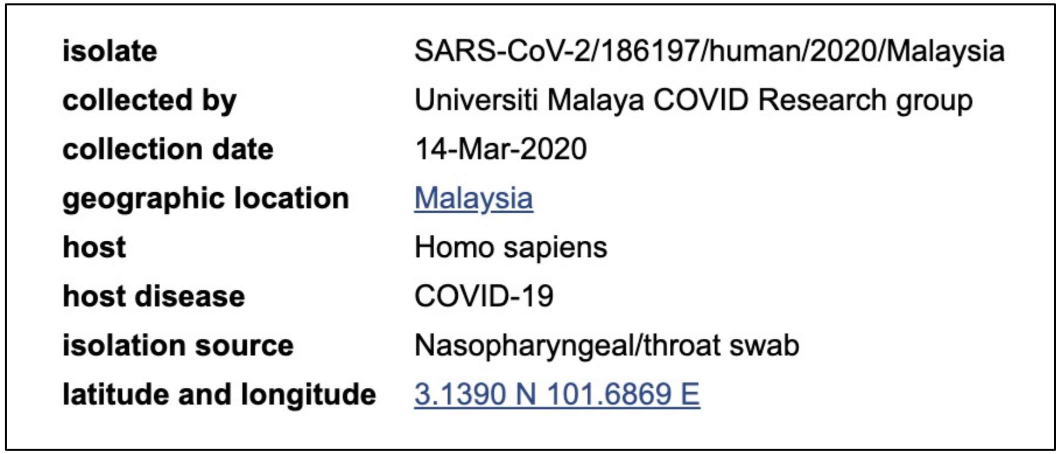
□ Other residential facility type, specify: _______

□ Shelter □ Conference

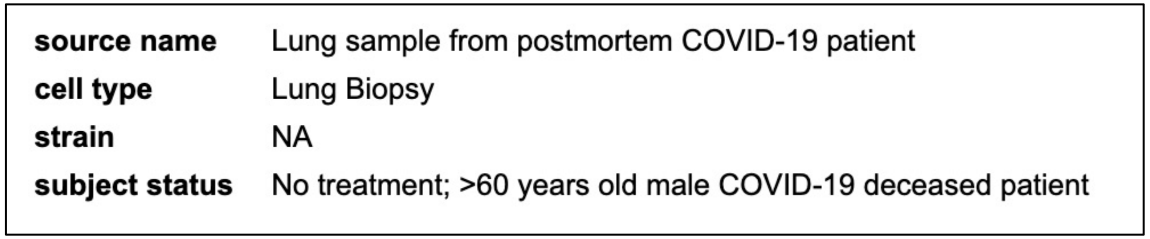
Public Health Alliance for Genomic Epidemiology

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



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

Compare these public datasets:

- different **standards**
- different **granularity**
- **free text**
- different **formats**
- different **interpretation**
- **privacy** concerns
- **methods**

Different data structures make information less interoperable and more difficult to integrate.

That means you need to spend more time and resources to clean/re-structure information before you can use it.

Public Health Alliance for Genomic Epidemiology

Data Structures | Bioinformatic Pipelines and Visualizations | Infrastructure
Public Repositories | Reference, QC and Validation | Workforce Development
Data Sharing and Ethics | Users and Applications

https://www.pha4ge.org | https://www.github.com/pha4ge | 🐦 @pha4ge

Data Structures Working Group

- 21 Members (9 countries, 4 continents)

Goal: develop/promote data standards

- COG-UK, SPHERES, CanCOGeN, Latin American Genomics Network
- identified need for fit-for-purpose contextual data standard for SARS-CoV-2 genomics

African Centre of Excellence for Genomics of Infectious Disease

Laboratório Nacional de Computação Científica

ANLIS MALBRÁN

Quadram Institute
Science Health Food Innovation

U.S. Department of Health and Human Services
Centers for Disease Control and Prevention

U.S. FOOD & DRUG ADMINISTRATION

NARMS
National Antimicrobial Resistance Monitoring System

FRED HUTCH
CURES START HERE®

BfR
Bundesinstitut für Risikobewertung

wellcome sanger institute

BIG DATA INSTITUTE

McMaster University

Comprehensive Antibiotic Resistance Database

DALHOUSIE UNIVERSITY
Inspiring Minds

UBC

BC Centre for Disease Control

University of Basel

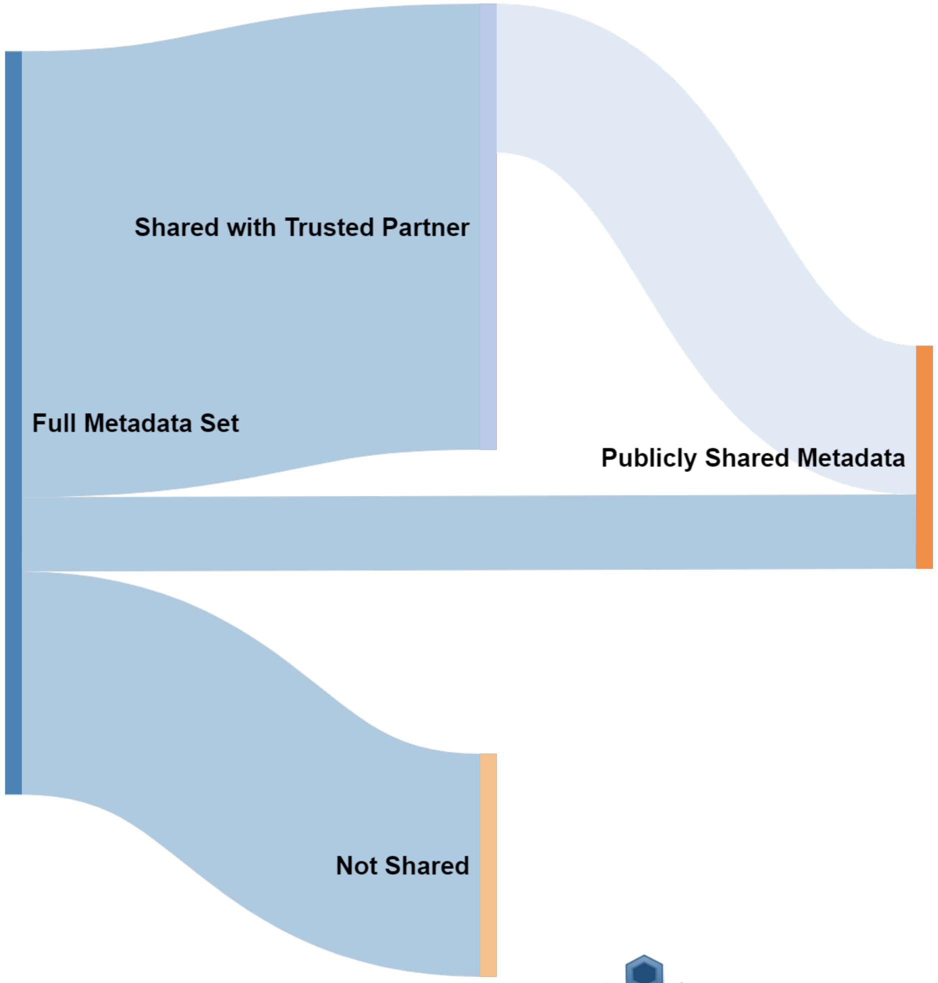
LISBOA
UNIVERSIDADE DE LISBOA

Public Health Alliance for Genomic Epidemiology

The SARS-CoV-2 Contextual Data Specification Package

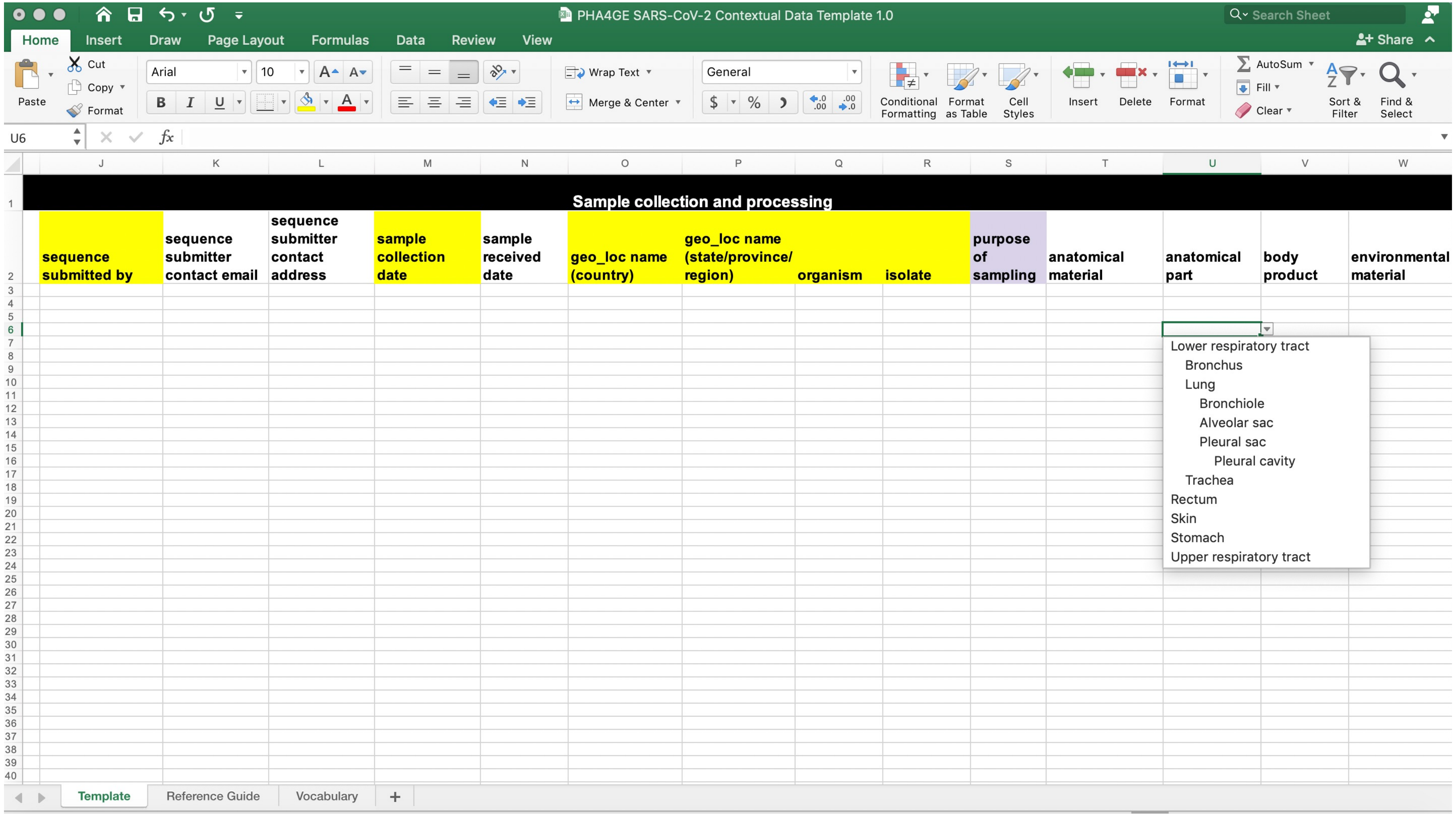
PHA4GE SARS-CoV-2 Full Specification Content

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



Public Health Alliance for Genomic Epidemiology

Template and terminology



PHA4GE SARS-CoV-2 Contextual Data Template 1.0

Sample collection and processing

J	K	L	M	N	O	P	Q	R	S	T	U	V	W
sequence submitted by	sequence submitter contact email	sequence submitter contact address	sample collection date	sample received date	geo_loc name (country)	geo_loc name (state/province/ region)	organism	isolate	purpose of sampling	anatomical material	anatomical part	body product	environmental material

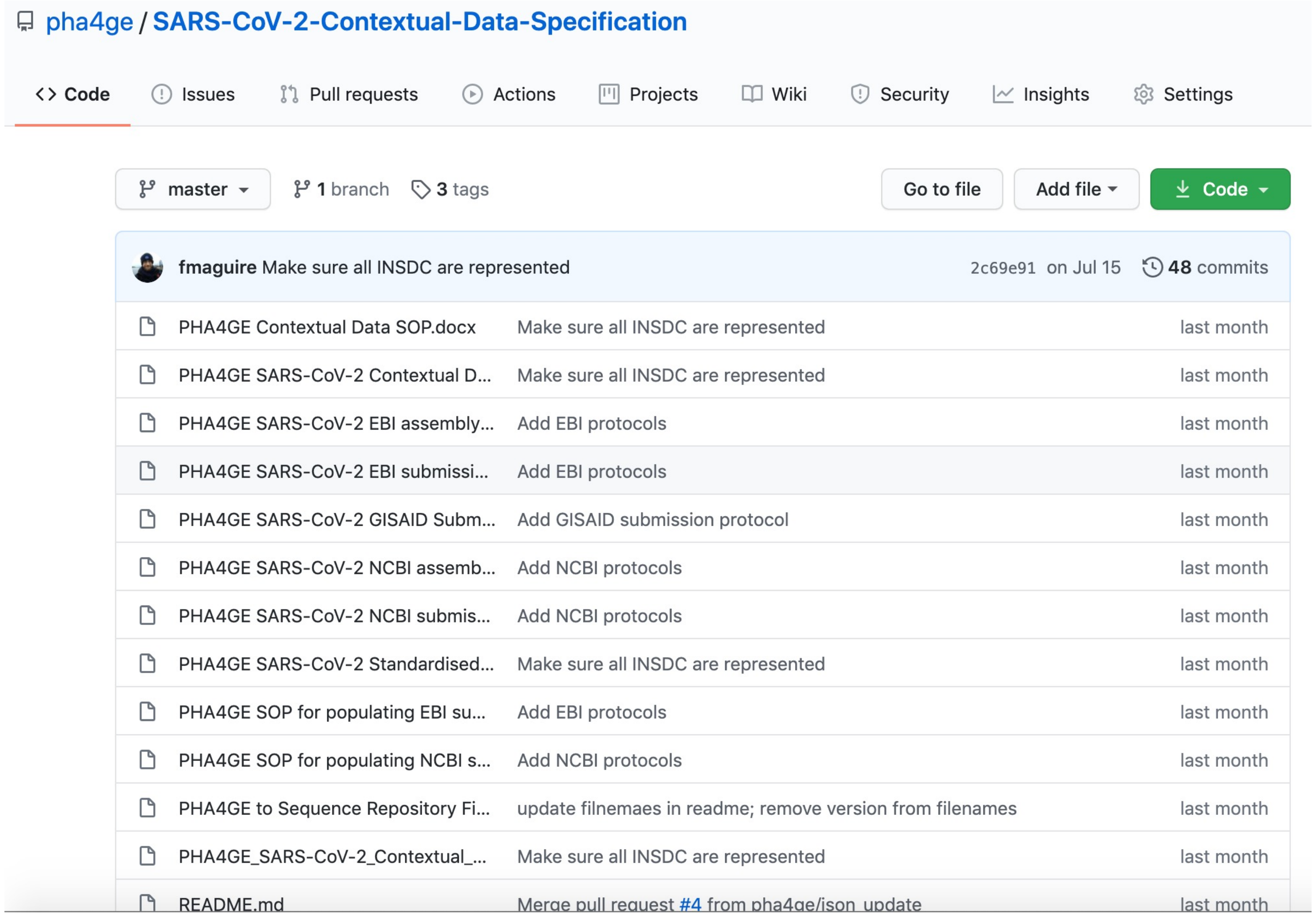
- Lower respiratory tract
  - Bronchus
  - Lung
    - Bronchiole
    - Alveolar sac
    - Pleural sac
      - Pleural cavity
  - Trachea
- Rectum
- Skin
- Stomach
- Upper respiratory tract

- **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



fmaguire Make sure all INSDC are represented 2c69e91 on Jul 15 48 commits

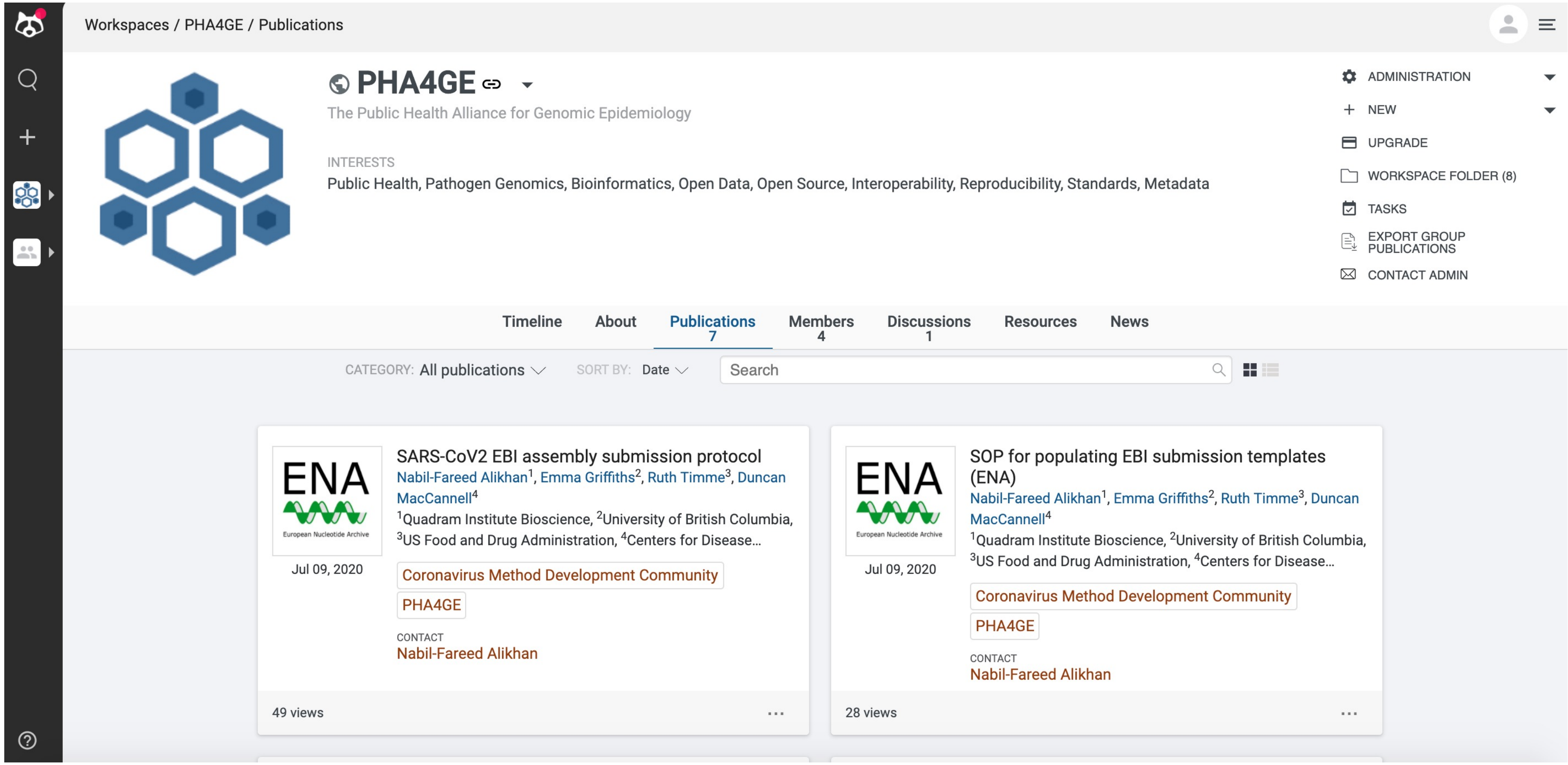
PHA4GE Contextual Data SOP.docx	Make sure all INSDC are represented	last month
PHA4GE SARS-CoV-2 Contextual D...	Make sure all INSDC are represented	last month
PHA4GE SARS-CoV-2 EBI assembly...	Add EBI protocols	last month
PHA4GE SARS-CoV-2 EBI submissi...	Add EBI protocols	last month
PHA4GE SARS-CoV-2 GISAID Subm...	Add GISAID submission protocol	last month
PHA4GE SARS-CoV-2 NCBI assemb...	Add NCBI protocols	last month
PHA4GE SARS-CoV-2 NCBI submis...	Add NCBI protocols	last month
PHA4GE SARS-CoV-2 Standardised...	Make sure all INSDC are represented	last month
PHA4GE SOP for populating EBI su...	Add EBI protocols	last month
PHA4GE SOP for populating NCBI s...	Add NCBI protocols	last month
PHA4GE to Sequence Repository Fi...	update filnemaes in readme; remove version from filenames	last month
PHA4GE_SARS-CoV-2_Contextual_...	Make sure all INSDC are represented	last month
README.md	Merge pull request #4 from pha4ge/json_update	last month

- **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



Workspaces / PHA4GE / Publications

PHA4GE

The Public Health Alliance for Genomic Epidemiology

INTERESTS

Public Health, Pathogen Genomics, Bioinformatics, Open Data, Open Source, Interoperability, Reproducibility, Standards, Metadata

SARS-CoV2 EBI assembly submission protocol

Jul 09, 2020

Nabil-Fareed Alikhan[1], Emma Griffiths[2], Ruth Timme[3], Duncan MacCannell[4]

[1]Quadram Institute Bioscience, [2]University of British Columbia, [3]US Food and Drug Administration, [4]Centers for Disease...

Coronavirus Method Development Community

PHA4GE

CONTACT

Nabil-Fareed Alikhan

49 views

SOP for populating EBI submission templates (ENA)

Jul 09, 2020

Nabil-Fareed Alikhan[1], Emma Griffiths[2], Ruth Timme[3], Duncan MacCannell[4]

[1]Quadram Institute Bioscience, [2]University of British Columbia, [3]US Food and Drug Administration, [4]Centers for Disease...

Coronavirus Method Development Community

PHA4GE

CONTACT

Nabil-Fareed Alikhan

28 views

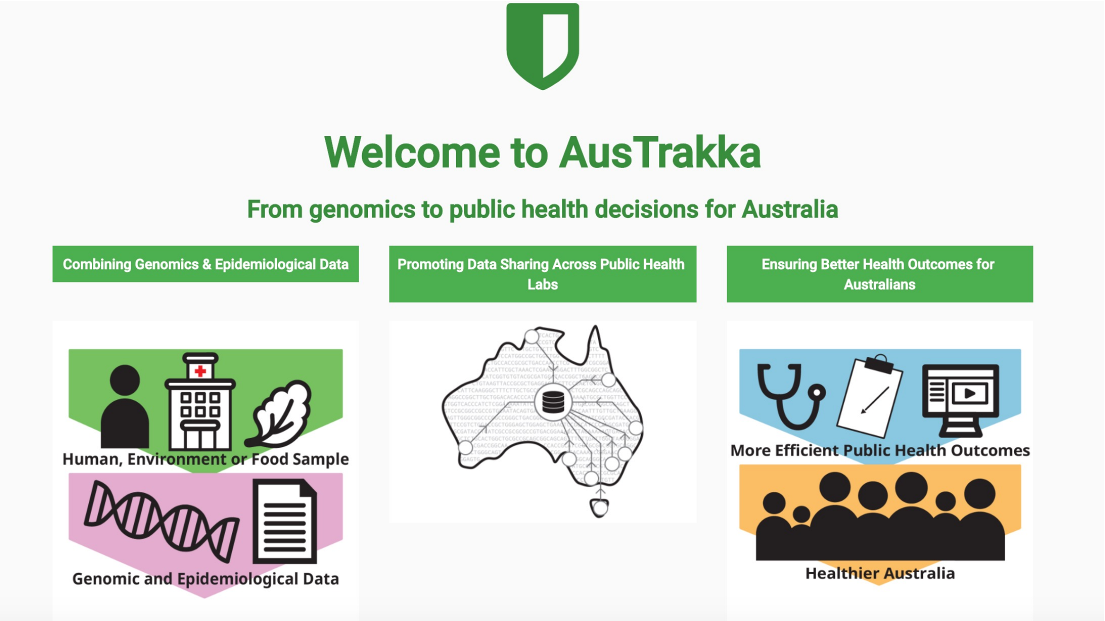
- **7 public repository submission protocols** (GISAID, NCBI, EMBL-EBI) on **Protocols.io**

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

Public Health Alliance for Genomic Epidemiology

Uptake: early adopters

BAOBAB LIMS



Welcome to AusTrakka

From genomics to public health decisions for Australia

Combining Genomics & Epidemiological Data

Promoting Data Sharing Across Public Health Labs

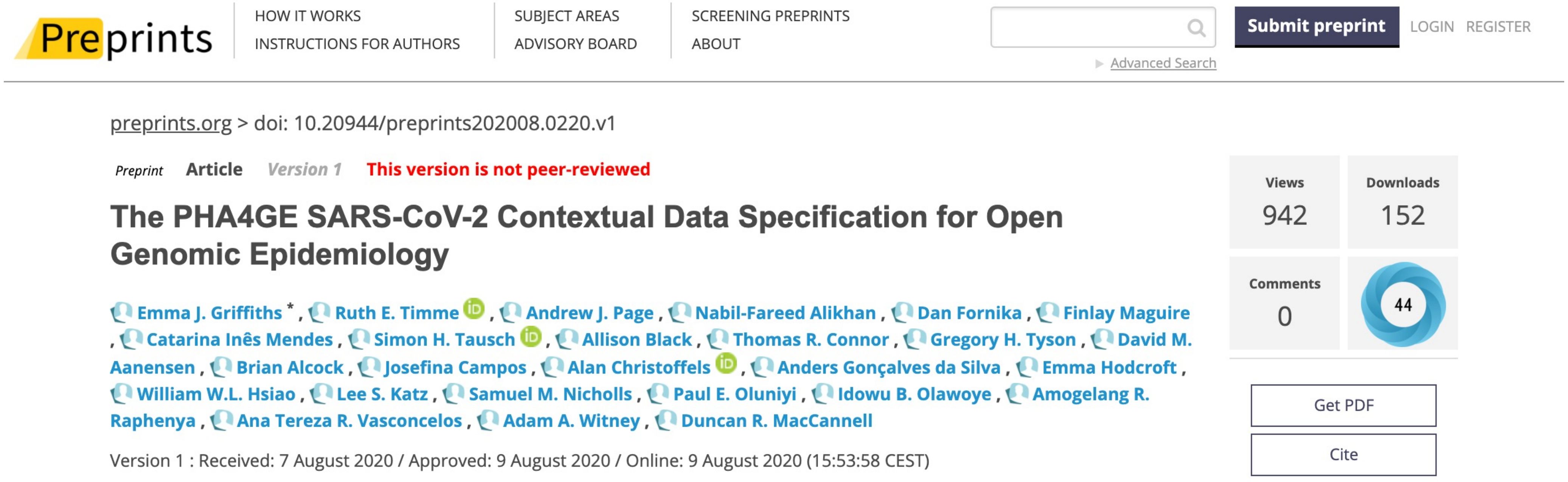
Ensuring Better Health Outcomes for Australians

Human, Environment or Food Sample
Genomic and Epidemiological Data
More Efficient Public Health Outcomes
Healthier Australia

CanCOGeN VirusSeq

SPHERES

Want to know more?



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

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

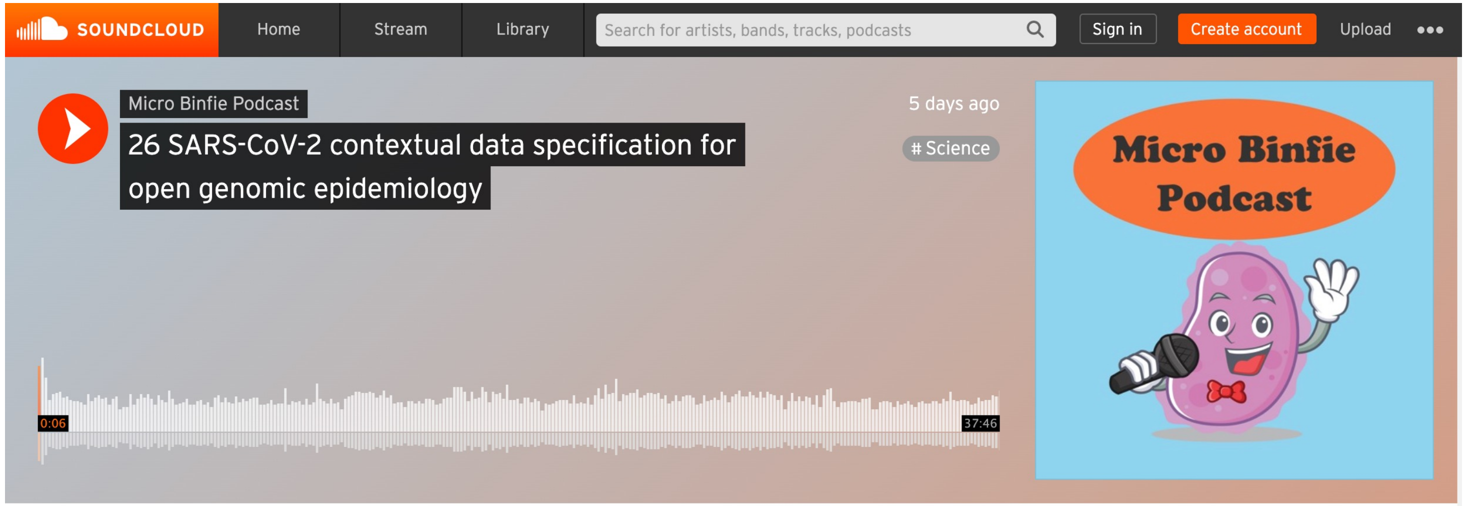
The PHA4GE SARS-CoV-2 Contextual Data Specification for Open Genomic Epidemiology

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)

Views	Downloads
942	152

Comments 0



Micro Binfie Podcast

26 SARS-CoV-2 contextual data specification for open genomic epidemiology

5 days ago

Science

Stay tuned for workshops coming soon!

Public Health Alliance for Genomic Epidemiology

Summary

- contextual data is critical for **interpreting** genomics data/analyses
- specification structures metadata **consistently across labs**
- share with **public** repos, **trusted partners**, use for more **efficient private analyses**
- **future-proof** metadata

- success depends on community uptake

Ongoing challenges: trust, equitable data sharing, better understanding of risk, co-ordination, sharing mechanisms

→ **public health genomics ecosystem**

Public Health Alliance for
Genomic Epidemiology

Thank you to
the Data Structures WG & friends,
PHA4GE consortium,
The Bill and Melinda Gates Foundation,
and to 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