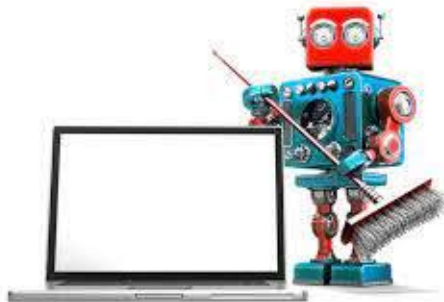


Overcoming challenges of SARS-CoV-2 genomics data sharing for public health surveillance, outbreak investigations and research using the PHA4GE SARS-CoV-2 contextual data specification

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

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

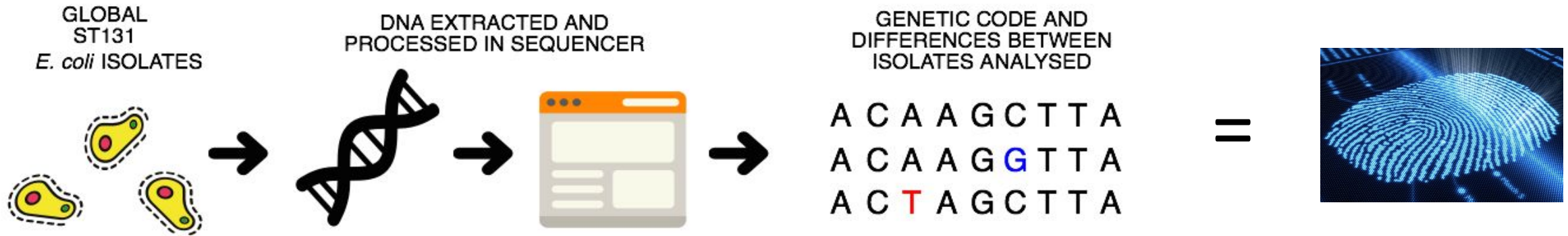
2. Demo of spec: putting standards into practice

- from chaos to control
- improving the quality of open data

3. Implementations of specification

- DataHarmonizer (Canada)
- AusTrakka (Australia)
- Boabab LIMS (South Africa)

Microbial genome sequences can be 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**

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

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

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 Vasani,  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](#)

5.03m Accesses | 706 Citations | 35003 Altmetric | [Metrics](#)

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-2³ (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 deaths⁵.

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 symptoms⁶. 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.

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. Pybus³, Nuno Faria³, Candace Wang^{1,2}, Guixia Yu^{1,2}, Brian Bushnell⁴, Chao-Yang Pan⁵, Hugo Guevara⁵, Alicia Sotomayor-Gonzalez^{1,2}, Kelsey Zorn⁶, Allan Gopez¹, Venice Servellita¹, Elaine Hsu¹, Steve Miller¹, Trevor Bedford^{7,8}, Alexander L. Greninger^{7,9}, Pavitra Roychoudhury^{7,9}, Lea M. Starita^{8,10}, Michael Famulare¹¹, Helen Y. Chu^{8,12}, Jay Shendure^{8,9,13}, Keith R. Jerome^{7,9}, Katie Anderson¹⁴, Karthik Gangavarapu¹⁴, Mark Zeller¹⁴, Emily Spencer¹⁴, Kristian G. Andersen¹⁴, Duncan MacCannell¹⁵, Clinton R. Paden¹⁵, Yan Li¹⁵, Jing Zhang¹⁵, Suxiang Tong¹⁵, Gregory Armstrong¹⁵, Scott Morrow¹⁶, Matthew Willis¹⁷, Bela T. Matyas¹⁸, Sundari Mase¹⁹, Olivia Kasirye²⁰, Maggie Park²¹, Godfred Masinde²², Curtis Chan²², Alexander T. Yu⁵, Shua J. Chai^{5,15}, Elsa Villarino²³, Brandon Bonin²³, Debra A. Wadford⁵, 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
<input type="checkbox"/> Upper respiratory (e.g., Nasopharyngeal or oropharyngeal swab)
<input type="checkbox"/> Lower respiratory (e.g., sputum, tracheal aspirate, BAL, pleural fluid)

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

6 - Specimen Type (check all that apply)

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

<input type="checkbox"/> NPS in UTM	If possible:
<input type="checkbox"/> Throat Swab in UTM	<input type="checkbox"/> BAL
<input type="checkbox"/> Other (Specify):	<input type="checkbox"/> Sputum

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**

Template and terminology

Sample collection and processing													
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

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

Supporting documentation

pha4ge / SARS-CoV-2-Contextual-Data-Specification

<> Code Issues Pull requests Actions Projects Wiki Security Insights Settings

master 1 branch 3 tags Go to file Add file Code

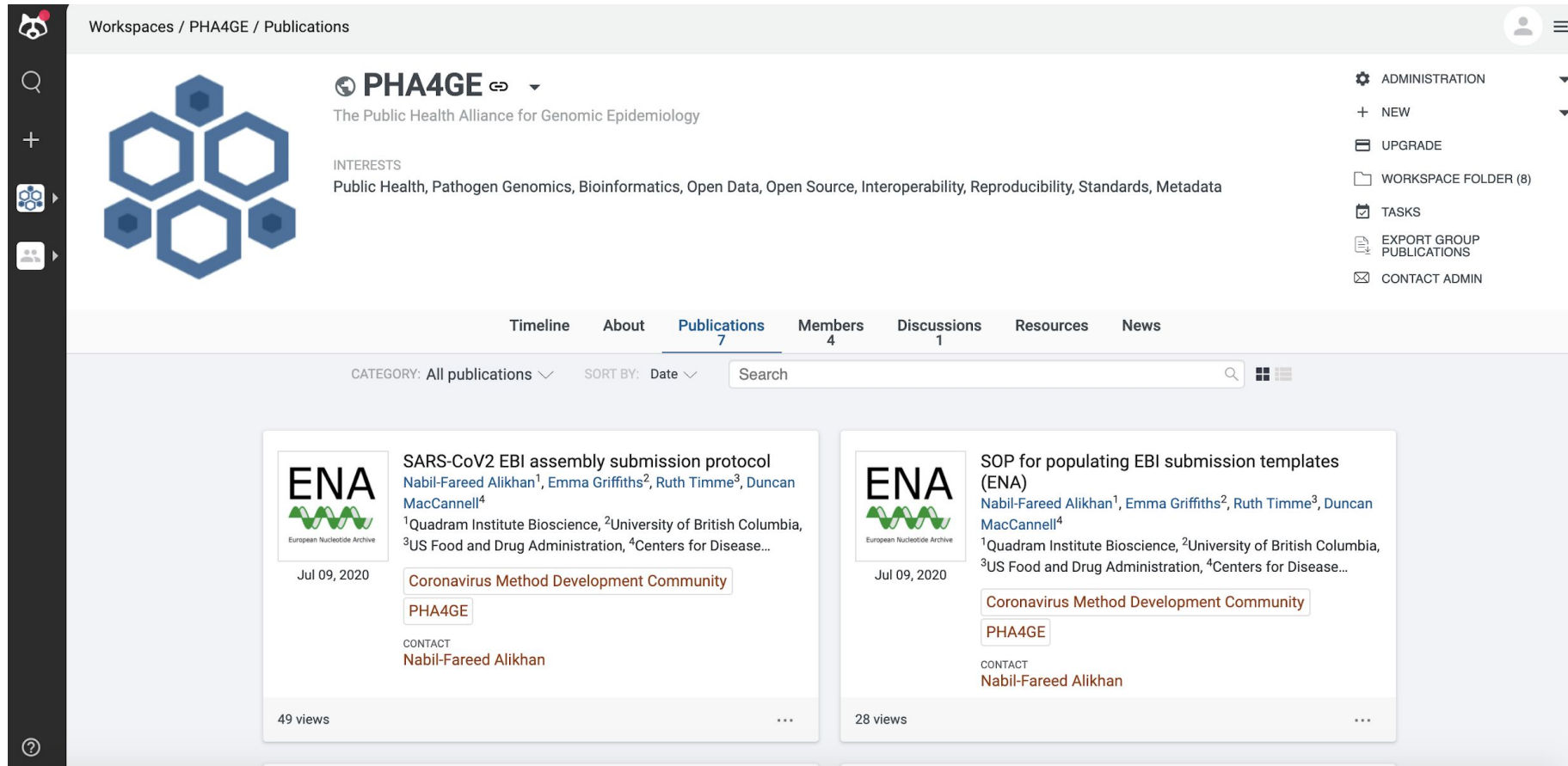
fmaguire Make sure all INSDC are represented 2c69e91 on Jul 15 48 commits

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>

Protocols to mobilize harmonized data



The screenshot displays the PHA4GE workspace on the Protocols.io platform. The workspace is titled "Workspaces / PHA4GE / Publications" and features the PHA4GE logo and tagline: "The Public Health Alliance for Genomic Epidemiology". The interests listed are: "Public Health, Pathogen Genomics, Bioinformatics, Open Data, Open Source, Interoperability, Reproducibility, Standards, Metadata". The navigation menu includes: "Administration", "New", "Upgrade", "Workspace Folder (8)", "Tasks", "Export Group Publications", and "Contact Admin". The main content area shows a list of publications under the "Publications" tab (7 items). Two publications are visible:

- SARS-CoV2 EBI assembly submission protocol**
Nabil-Fareed Alikhan¹, Emma Griffiths², Ruth Timme³, Duncan MacCannell⁴
¹Quadram Institute Bioscience, ²University of British Columbia, ³US Food and Drug Administration, ⁴Centers for Disease...
Jul 09, 2020
Coronavirus Method Development Community
PHA4GE
CONTACT: Nabil-Fareed Alikhan
49 views
- SOP for populating EBI submission templates (ENA)**
Nabil-Fareed Alikhan¹, Emma Griffiths², Ruth Timme³, Duncan MacCannell⁴
¹Quadram Institute Bioscience, ²University of British Columbia, ³US Food and Drug Administration, ⁴Centers for Disease...
Jul 09, 2020
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>

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^{ID}, Andrew J. Page, Nabil-Fareed Alikhan, Dan Fornika, Finlay Maguire, Catarina Inês Mendes, Simon H. Tausch^{ID}, Allison Black, Thomas R. Connor, Gregory H. Tyson, David M. Aanensen, Brian Alcock, Josefina Campos, Alan Christoffels^{ID}, 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

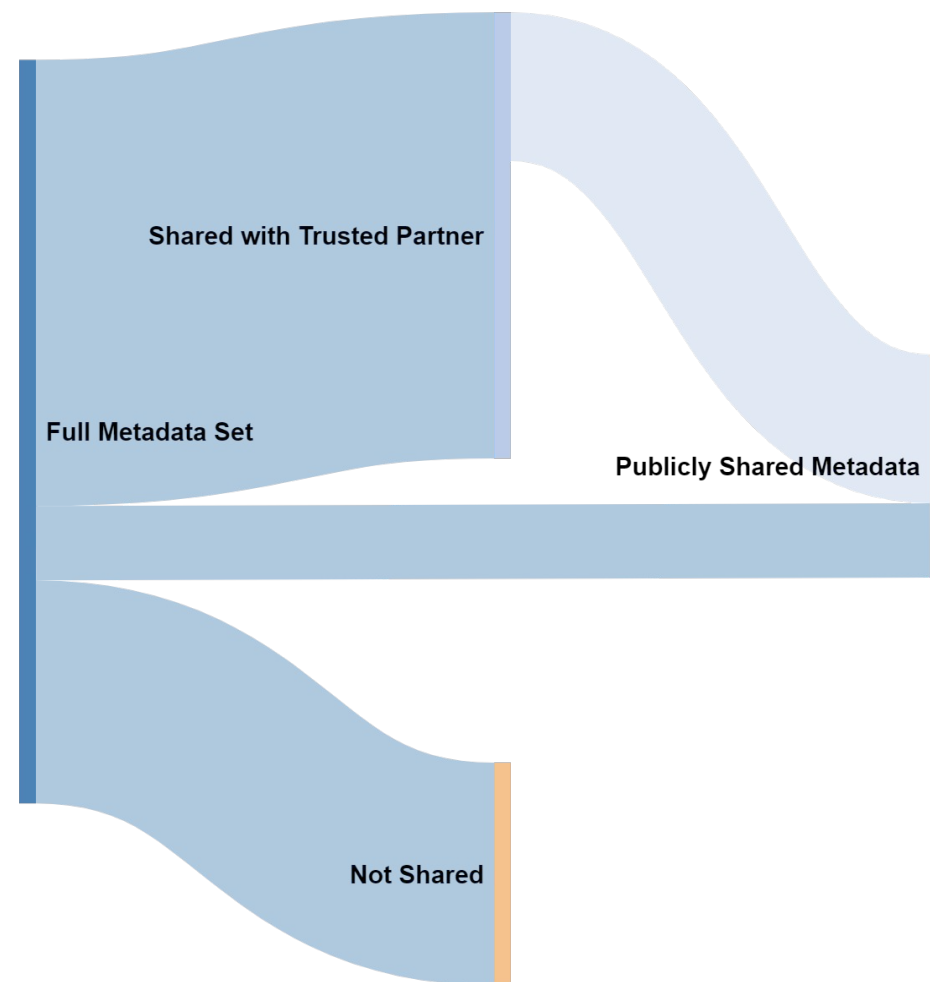
Version 1 : Received: 7 August 2020 / Approved: 9 August 2020 / Online: 9 August 2020 (15:53:58 CEST)

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

The screenshot shows the SoundCloud interface for a podcast episode. At the top, the SoundCloud logo and navigation links (Home, Stream, Library) are visible. A search bar and user options (Sign in, Create account, Upload) are also present. The main content area displays the episode title "26 SARS-CoV-2 contextual data specification for open genomic epidemiology" by "Micro Binfie Podcast", posted "5 days ago" with a "# Science" tag. A waveform player is shown at the bottom left, indicating a duration of 37:46. On the right, there is a promotional image for the "Micro Binfie Podcast" featuring a cartoon purple character with a microphone.

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



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)

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**

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!



Welcome to AusTrakka
From genomics to public health decisions for Australia

Combining Genomics & Epidemiological Data
Human, Environment or Food Sample
Genomic and Epidemiological Data

Promoting Data Sharing Across Public Health Labs

Ensuring Better Health Outcomes for Australians
More Efficient Public Health Outcomes
Healthier Australia

The graphic is a light grey rectangular box with a green shield icon at the top center. It contains three columns of information. The first column has a green header, a green box with icons of a person, a hospital, and a leaf, and a purple box with a DNA helix and a document icon. The second column has a map of Australia with data points. The third column has a blue box with a stethoscope, clipboard, and monitor icon, and an orange box with a group of people icon.

Thank you for listening and participating!

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