Article

PHA4GE Quality Control Contextual Data Tags: Standardized Annotations for Sharing Public Health Sequence Datasets with Known Quality Issues to Facilitate Testing and Training

Emma Griffiths [1,*], Catarina Inês Mendes [2], Finlay Maguire [3], Jennifer Guthrie [4], Leonid Chindelevitch [5], Ilene Karsch-Mizrachi [6], Zahra Waheed [7], Rhiannon Cameron [1], Kathryn Holt [8], Lee Katz [9], Robert Petit III [10], Duncan MacCannell [11], Mugdha Dave [12], Paul Oluniyi [13], Muhammad Ibtisam Nasar [14], Amogelang Raphenya [15], William Hsiao [1] and Ruth Timme [16]

[1] Centre for Infectious Disease Genomics and One Health, Faculty of Health Sciences, Simon Fraser University, Burnaby, British Columbia, Canada
[2] Instituto de Microbiologia, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal
[3] Department of Community Health & Epidemiology, Faculty of Medicine, Dalhousie University, Halifax, Nova Scotia, Canada
[4] Department of Microbiology & Immunology, Western University, London, Ontario, Canada
[5] MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College London, London, UK
[6] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA
[7] European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK
[8] Department of Infection Biology, London School of Hygiene and Tropical Medicine, London, UK
[9] Center for Food Safety, University of Georgia, Georgia, USA
[10] Wyoming Public Health Laboratory, Wyoming, USA
[11] National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, Georgia, USA
[12] McMaster University, Hamilton, Ontario, Canada
[13] Chan Zuckerberg Biohub, San Francisco, California, USA
[14] Department of Biology, College of Science, United Arab Emirates University- AL Ain, Abu Dhabi, UAE
[15] Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, Ontario, Canada
[16] Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, College Park, Maryland, USA
* Correspondence: ega12@sfu.ca

Abstract: As public health laboratories expand their genomic sequencing and bioinformatics capacity for the surveillance of different pathogens, labs must carry out robust validation, training, and optimization of wet- and dry-lab procedures. Achieving these goals for algorithms, pipelines and instruments often requires that lower-quality datasets be made available for analysis and comparison alongside those of higher-quality. This range of data quality in reference sets can complicate the sharing of sub-optimal datasets that are vital for the community and for the reproducibility of assays. Sharing of useful, but sub-optimal datasets requires careful annotation and documentation of known issues to enable appropriate interpretation, avoid being mistaken for better quality information, and for these data (and their derivatives) to be easily identifiable in repositories. Unfortunately, there are currently no standardized attributes or mechanisms for tagging poor-quality datasets, or datasets generated for a specific purpose, to maximize their utility, searchability, accessibility and reuse.

The Public Health Alliance for Genomic Epidemiology (PHA4GE) is an international community of scientists from public health, industry and academia focused on improving the reproducibility, interoperability, portability and openness of public health bioinformatic software, skills, tools and data. To address the challenges of sharing lower quality datasets, PHA4GE has developed a set of standardized contextual data tags, namely fields and terms, that can be included in public repository submissions as a means of flagging pathogen sequence data with known quality issues, increasing their discoverability. The contextual data tags were developed through consultations with the community including input from the International Nucleotide Sequence Data Collaboration (INSDC), and have been standardized using ontologies, community-based resources for defining the tag properties and the relationships between them. The standardized tags are agnostic to the organism and the sequencing technique used and thus can be applied to data generated from any pathogen using an array of sequencing techniques. The list of standardized tags is maintained by PHA4GE and can be found at https://github.com/pha4ge/contextual_data_QC_tags. Definitions, ontology IDs, examples of use, as well as a JSON representation, are provided.

The PHA4GE QC tags were tested, and are now implemented, by the FDA's GenomeTrakr laboratory network as part of its routine submission process for SARS-CoV-2 wastewater surveillance. We hope that these simple, standardized tags will help improve communication regarding quality control in public repositories, in addition to making datasets of variable quality more easily identifiable. Suggestions for additional tags can be submitted to PHA4GE via the New Term Request Form in the GitHub repository. By providing a mechanism for feedback and suggestions, we also expect that the tags will evolve with the needs of the community.

Keywords: quality control; reference datasets; public health; metadata; data sharing

1. Introduction

Pathogen genomics surveillance laboratories generate microbial sequence data that can be used in a variety of ways. Examples include the detection and resolution of outbreaks, development of vaccines and diagnostic tests, understanding microbial evolution including antimicrobial resistance and virulence mechanisms, detection of zoonotic events and patterns of transmission, source attribution, and more (Petrillo et al, 2022; Robinson et al, 2013; Munnink et al, 2021; World Health Organization, 2022; Cook, 2021; Hendriksen et al, 2019; Brown et al, 2021). The quality of sequence datasets greatly impacts their utility, the interpretations of analytical results, and the decisions that can be made based on how much confidence one has in the interpretations (Rick et al, 2022; Smits, 2019). The quality of sequence data can vary for many reasons, including, but not limited to, low concentrations of starting materials, expired reagents, deviations from ideal sample handling and storage conditions, errors during library preparation, overloading/underloading of flow cells in the case of long-read sequencing techniques, and contamination within and between sequencing runs (Gargis et al, 2016; Rossen et al, 2018). Quality control metrics of raw reads depend on many factors such as the depth and breadth of coverage of generated reads compared to a reference, the presence of reads from another source (i.e. previous run, host organism, sample contamination) and, the number and density of flow cell clusters in a run (Wagner et al, 2021). Quality control metrics and thresholds often differ across laboratories and surveillance networks. While public health laboratories generate and release a vast number of high-quality sequences, there will often be a proportion of datasets that may fall just short of a set of prescribed baseline quality control metrics. These datasets are then excluded from many types of public health analyses and are often not publicly released. In these cases, the issues associated with these borderline or lower-quality datasets have often been identified, and the datasets can still provide important surveillance insights or information on real-world test performance. Conversely, low-quality datasets can sometimes be included in public repository submissions but are not flagged which creates issues for laboratories using the data.

As pathogen sequencing is increasingly used routinely in public health laboratories for the surveillance of different pathogens, and programs are continually developed and expanded, laboratories must carry out robust optimization of wet- and dry-lab procedures (Carrillo & Blais, 2021). Lower quality datasets are highly useful for optimizing, validating, verifying and benchmarking the performance of algorithms, pipelines and instruments, as well as training new personnel (Figure 1). An example of the utility of high and low quality datasets can be seen in Xiaoli et al (2022) in which SARS-CoV-2 Nanopore/Illumina read datasets generated from public health genomic surveillance were shared as a collection to support benchmarking tools, understanding the genomic epidemiology of different lineages, and identifying variants of concern. The collection also contained a number of SARS-CoV-2 genomes of lower quality due to recognized errors and common sequencing failures (Xiaoli et al, 2022).

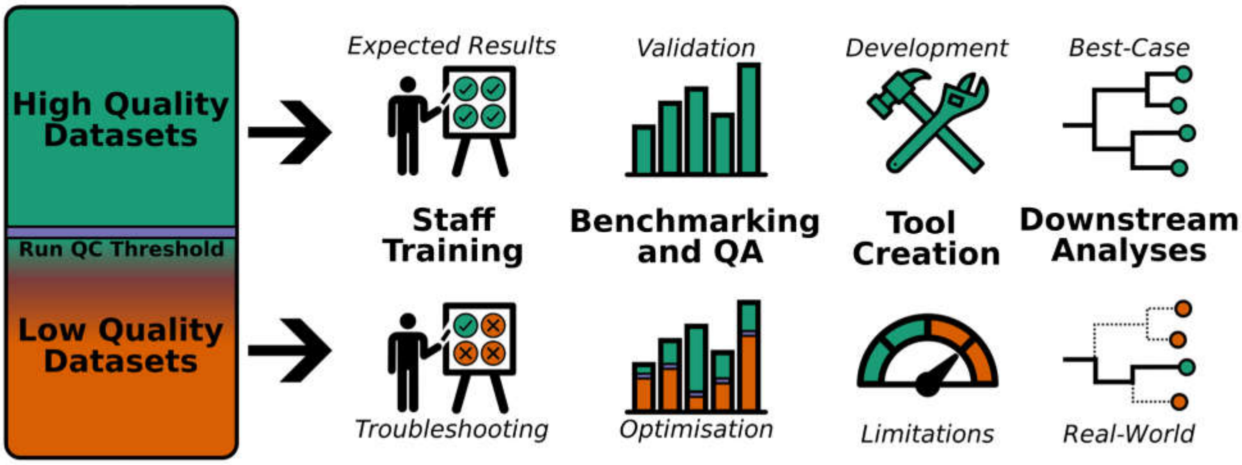

Figure 1. Sequence data quality is assessed using prescribed criteria (i.e. metrics) and thresholds. Datasets of high and lower quality have many uses in public health activities such as staff training and lab procedure/software optimization and validation in ideal and real-world scenarios.

Sharing sub-optimal data can be useful for the broader public health and research community, particularly when the data is carefully annotated with known issues so that it is not mistaken for better quality information, and can be more easily identified in repositories. However, there are currently no standardized attributes for tagging poor-quality datasets, preventing them from being easily searched and made accessible, and all but ensuring that they are excluded from applications that require only high-quality data. Standardized fields and terms have previously proven useful in improving data harmonization and integration as well as communication and data sharing in SARS-CoV-2 surveillance (Griffiths et al, 2022; Lusignan et al, 2020), and the implementation of genomics contextual data (metadata) standards have been encouraged repeatedly in the community (Black et al, 2020; Gozashti & Corbett-Detig, 2021; Pettengill et al, 2021; Schriml et al, 2020; Stevens et al, 2020). As such, a set of standardized attributes to describe the properties, quality and purpose of microbial sequence datasets could make quality control results more explicit in public (and private) repositories.

The Public Health Alliance for Genomic Epidemiology (PHA4GE) is a global coalition of scientists focused on improving the reproducibility, interoperability, portability, and openness of public health bioinformatics software, data and expertise (https://pha4ge.org/). As part of its' mission to improve interoperability and reproducibility, PHA4GE workgroups develop, share and promote consensus data specifications, in an effort to streamline and improve data structures across public health bioinformatics resources (tools, protocols, databases, platforms, and repositories). The overarching goal of this work is the development of an open software philosophy and ecosystem that will empower more stakeholders across global public health to analyze, manage and govern their own data, regardless of resource status. The Data Structures Working Group, tasked with assessing needs and developing these specifications, operates by consensus and comprises diverse perspectives and expertise with members representing many different countries, organizations, microbial sequencing initiatives, and standards development efforts.

To address the challenges of sharing lower-quality datasets, PHA4GE has developed a set of standardized contextual data attributes (fields and terms known as "tags") that can be included in public repository submissions as a means of flagging pathogen sequence data with known quality issues to increase their discoverability, and to facilitate their interpretation and appropriate reuse. The contextual data tags (attributes) were developed through a series of consultations with the public health microbiology research community, including input from the International Nucleotide Sequence Data Collaboration (INSDC), and staff from multiple national, regional and local public health institutions. The development of these tags, standardized using community-based resources known as ontologies, is expected to be an iterative and participatory process with input from users and subject matter experts from across the community. Ontologies are well-defined controlled vocabulary describing a domain, structured in a hierarchy where logical relationships link the terms, and the meanings of terms are disambiguated using persistent identifiers (Smith et al, 2007). As ontologies are developed by community consensus, by applying ontology-based attributes in publicly available data, the pitfalls and variability of institution-specific vocabulary and free text can be avoided. The standardized tags are agnostic to the organism and sequencing technique used, and can be applied to data generated from any pathogen using an array of sequencing techniques. The list of standardized tags for quality control, introduced here, is maintained by PHA4GE and can be found at https://github.com/pha4ge/contextual_data_QC_tags. Recognizing that data needs can change over time, or that different use cases can require additional vocabulary, PHA4GE accepts suggestions for new tags from the community which can be submitted using the New Term Request System on GitHub (see PHA4GE repository linked above).

As testing and implementation are key to the success and uptake of data specifications, PHA4GE partnered with the US Food and Drug Administration Center for Food Safety and Applied Nutrition (FDA CFSAN)'s GenomeTrakr program as a test use case and in the early adoption of the QC tags described in this work. GenomeTrakr is a US Food and Drug Administration (FDA)-led international pathogen surveillance network through which member labs submit sequence data and minimal contextual data in real-time for the purposes of tracking and identifying outbreaks (Timme et al, 2019). GenomeTrakr has focused on surveillance of foodborne pathogens for many years, and with the onset of the COVID-19 pandemic, has also expanded to metagenomic wastewater surveillance of SARS-CoV-2. Wastewater monitoring can provide an early warning of COVID-19 detection in a community or setting (e.g. watershed, institution). An early warning of even a few days can be critical to the success of public health interventions. Metagenomic surveillance of wastewater can be challenging due to the complex nature of samples, therefore, sharing information regarding quality control assessment is of uttermost importance. The GenomeTrakr network has tested and now implements the PHA4GE QC contextual data tags as part of its routine submission process. and provides a worked example of how the tags can be customized by organizations and surveillance initiatives within the public health bioinformatics community.

The sharing of lower-quality datasets and their annotation using the contextual data tags described here will enable public health labs to make use of data that would have otherwise been discarded, and in some cases, side-step the need to generate synthetic data for representing different real-world scenarios. Using these tags will enable the community to more easily establish datasets for training and testing purposes (software and human). The inclusion of ontologized PHA4GE QC tags will also make datasets and quality control results FAIR (Findable, Accessible, Interoperable, and Reusable) (Wilkinson et al, 2016).

2. Methods

The members of PHA4GE are involved in many sequencing initiatives and surveillance networks, and as a result, have a broad collective experience in developing solutions to microbial bioinformatics challenges. Requests for standardized quality control tags were made to PHA4GE from members of the wider public health and research communities via direct communication and social media. The range and types of common quality control issues were identified through a survey via member networks. The proposed list of quality control tags was circulated for feedback within the PHA4GE community and was improved based on feedback. The QC attributes were then mapped to existing ontologies, and ontology terms were created for tags with no existing equivalent and made publicly available in the Genomic Epidemiology Ontology (GenEpiO, https://github.com/GenEpiO/genepio). Definitions were also developed, along with recommendations for their use in INSDC sequence submissions (in collaboration with INSDC representatives). The QC attributes were made publicly available on GitHub in October 2022, and included in a specially designed SRA submission form for pathogen sequence data (available on GitHub).

QC contextual data tags were reviewed and evaluated by GenomeTrakr scientists for tagging known quality control issues in wastewater metagenomics datasets used for SARS-CoV-2 surveillance. The fields were added to the prescribed GenomeTrakr submission requirements, along with additional values for GenomeTrakr-specific pipelines and analyses.

3. Results

Development of Standardized Quality Control Tags

After reviewing lower-quality bacterial and viral datasets to determine the most common reasons for QC failure, the list of reasons and associated information was categorized and structured according to the fields and values in Table 1. After review, fields and terms were ontologized and included in the GenEpiO in order to make the attributes FAIR. GenEpiO QC terms can be searched using different ontology lookup services such as the EMBL-EBI OLS (https://www.ebi.ac.uk/ols/index), and downloaded as part of the GenEpiO web ontology language file available through The Open Biological and Biomedical Ontology (OBO) Foundry (https://obofoundry.org/ontology/genepio.html). The QC tags are also available as a bespoke JSON file through the PHA4GE QC GitHub repository, as well as field and term reference guides providing definitions and further guidance for usage.

Table 1. Standardized fields and values for annotating quality control information in shared pathogen genomics datasets.

Field*	Definition	Ontology ID	Data Type	Values	Example
quality control method name	The name of the method used to assess whether a sequence passed a predetermined quality control threshold.	GENEPIO:0100557	String	No prescribed values	ncov-tools
quality control method version	The version number of the method used to assess whether a sequence passed a predetermined quality control threshold.	GENEPIO:0100558	String	No prescribed values	1.2.3
quality control determination	The determination of a quality control assessment.	GENEPIO:0100559	Enums	no quality control issues identified [GENEPIO:0100562]; sequence passed quality control [GENEPIO:0100563]; sequence failed quality control [GENEPIO:0100564]; minor quality control issues identified [GENEPIO:0100565]; sequence flagged for potential quality control issues [GENEPIO:0100566]; quality control not performed [GENEPIO:0100567]	sequence failed quality control [GENEPIO:0100564]
quality control issues	The reason contributing to, or causing, a low quality determination in a quality control assessment	GENEPIO:0100560	Enums	low quality sequence [GENEPIO:0100568]; sequence contaminated [GENEPIO:0100569]; low average genome coverage [GENEPIO:0100570]; low percent genome captured [GENEPIO:0100571]; read lengths shorter than expected [GENEPIO:0100572]; sequence amplification artifacts [GENEPIO:0100573]; low signal to noise ratio [GENEPIO:0100574]; low coverage of characteristic mutations [GENEPIO:0100575]	low average genome coverage [GENEPIO:0100570]
quality control details	The details surrounding a low quality determination in a quality control assessment.	GENEPIO:0100561	String	No prescribed values	CT value of 39. Low viral load. Low DNA concentration after amplification.

Best Practices for Use

Below are a few simple recommendations for implementing the QC tags (also available in the Field Reference Guide available at GitHub).

1. Providing the name of the method used for quality control is very important for interpreting the rest of the QC information. A method name should always be included (do not include additional QC tags if no method name is provided).
2. Method names can be provided in the form of a name of a pipeline or a link to a GitHub repository. Multiple methods should be listed and separated by a semicolon.
3. Methods updates can make big differences to their outputs. The version of the method used for quality control should be included.
4. The method version can be expressed using whatever convention the developer implements (e.g. date, semantic versioning).
5. If multiple methods were used, record the version numbers in the same order as the method names. Separate the version numbers using a semicolon.
6. If a pick list does not contain a desired value, a new term request should be submitted to PHA4GE via the QC Tag GitHub repository issuetracker New Term Request form (described below under “Community Development and Maintenance”).

Annotation Limitations and Considerations

The QC tags are intended to address issues pertaining to different types of sequencing techniques (single isolate or targeted sequencing, metagenomics). Not all tags may apply to all techniques and so where they are not appropriate then they should not be used. The tags are also intended to describe QC results of sequence data rather than downstream analytical results (e.g. raw reads, consensus sequences or assemblies rather than phylogenies or lineage determinations). Owing to the wide variety of quality control software available, and the differences in criteria and thresholds, the application of these attribute tags may be subjective and dependent on the QC processes performed. To better evaluate and interpret the QC determinations proposed, it is recommended that other information pertaining to QC be included in other contextual data fields not specified in this work (i.e. choice of reference genome), and that the tags be interpreted in light of the other methodological metadata included in the record (i.e. BioSample, Experiment/SRA contextual data). The controlled vocabulary attributes are intended for high-level triage purposes rather than capturing all methods in detail. However, information affecting the selection of one tag over another can also be included in the “quality_control_details” field. It is also important to note that the quality control tags refer to a particular sample obtained at one point in time, and not the comparison of a set of samples across time or from different tissues of the same host.

Implementation of Standardized Quality Control Tags

PHA4GE has collaborated with the INSDC in the design and application of the QC tags. Both organizations recommend the inclusion of the QC tags in submissions, and that the attributes should be included as user-defined fields in SRA (NCBI) or included interactively or programmatically in Experiment metadata via the ENA submission system (note: the Webin command line program does not accept user-defined attributes). To facilitate inclusion in NCBI submissions, PHA4GE has created a modified SRA submission form containing PHA4GE QC fields with drop-down menus of prescribed values which is available at https://github.com/pha4ge/contextual_data_QC_tags.

PHA4GE QC tags were evaluated for use by GenomeTrakr scientists. The generic PHA4GE tags were adapted for GenomeTrakr use through the addition of initiative-specific vocabulary for methods of QC assessment, as well as GenomeTrakr-specific instructions and guidance to data providers. A summary of the way GenomeTrakr is implementing PHA4GE QC tags is provided in Table 2. Example SRA records containing the attributes can be found in Table 3. Multiple values can be provided for each field, as necessary.

To date (January 31 2023), there are over 1603 GenomeTrakr SRA run records containing these QC tags. Many of these records make use of the “no quality control issues identified” tag, as well as the “quality_control_method_name” and “quality_control_method_version” fields, which add value despite there being no issues identified, by providing QC methodology details that would otherwise not have been present in the records.

Table 2. Implementation and updates to the PHA4GE quality control attributes applied by GenomeTrakr Network.

Attribute name	Description	Guidance	Term lists
quality_control_method_name	Name of quality control pipeline, software, or method	Populate using a term from the picklist	GalaxyTrakr SSQuAWK; CFSAN Wastewater Analysis Pipeline (C-WAP)
quality_control_method_version	Version number		
quality_control_determination	User determined assessment of data quality	Populate using a term from the picklist	No quality control issues identified; minor quality control issues identified; sequence flagged for potential quality control issues; not performed
quality_control_issues	Quality control issues relevant for the project or data type	Populate using a term from the picklist	Low quality sequence; sequenced contaminated; low average genome coverage; low % genome captured; read lengths shorter than expected; sequence amplification artifacts; low signal to noise ratio; low coverage of characteristic mutations
quality_control_details	Free text attribute capturing custom entry	Free text entry	None

It should be noted that GenomeTrakr also includes other PHA4GE contextual data fields for describing their methods in their SRA submissions. These attributes include amplicon_PCR_primer_scheme; amplicon_size; dehosting_method; and sequence_submitter_contact_email. These additional prescribed fields are outlined in the PHA4GE SARS-CoV-2 Contextual Data Specification (https://github.com/pha4ge/SARS-CoV-2-Contextual-Data-Specification).

Table 3. Example GenomeTrakr SRA records illustrating PHA4GE QC contextual data tag use.

Data Provider	Surveillance Network	SRA Accession	Run Record
Washington State Department of Health	GenomeTrakr wastewater project	SRR21205381	SRR21205381 Run Record
US FDA, Center for Food Safety and Applied Nutrition	GenomeTrakr wastewater project	SRR19851129	SRR19851129 Run Record
US FDA, Center for Food Safety and Applied Nutrition	GenomeTrakr wastewater project	SRR20046849	SRR20046849 Run Record
New Jersey Department of Agriculture	GenomeTrakr wastewater project	SRR20428498	SRR20428498 Run Record
Texas Department of State Health Services	GenomeTrakr wastewater project	SRR20018633	SRR20018633 Run Record

PHA4GE Contextual Data QC Tags in SRA can be found in the Run records (the hyperlink that starts with SRR in the "Run" chart near the bottom of the webpage). This information can be accessed by clicking the hyperlink and then clicking the "Show additional attributes" link in the Run section.

Community Development and Maintenance

While the initial list of standardized QC fields and values was developed by PHA4GE through community consultation, we recognize it will need to change over time. To ensure the list reflects current QC issues across pathogens and methods (e.g. sequencing techniques, bioinformatics analyses), a mechanism for requesting additional QC tags was created via the PHA4GE QC GitHub repository Issue tracker (New Term Request (NTR) form). A template for submitting new terms is available and community members are welcome to submit suggestions for new terms by following the instructions provided with the template. Suggestions will be evaluated and periodic updates to the list will be performed.

4. Discussion

A major goal of pathogen genomics surveillance programs is to produce high-quality data for use in public health analyses and decision-making. Owing to time, personnel and resource limitations, samples that yield sequences of borderline or slightly poorer quality cannot often be re-sequenced. This sequence data, while perhaps not suitable for surveillance or outbreak analysis, is still useful for the development of tools, the optimization and validation of quality frameworks and sequencing processes, as well as for bioinformatics training purposes. Useful datasets for testing and training purposes can include sequences containing contamination, low yields and/or low average genome coverage, shorter than expected read lengths, sequence amplification artifacts, low signal-to-noise ratio, and low coverage of characteristic mutations.

Due to the lack of standardized attributes in contextual data records, purposefully identifying sub-optimal quality datasets in public repositories is difficult. The PHA4GE Contextual Data QC Tag Specification provides a set of five fields which can be included as user-defined contextual data in public repository raw read sequence submissions. While PHA4GE encourages the use of these fields and terms in any repository, not all public repositories have the mandate or the ability to include user-defined attributes. The PHA4GE tags are implementable in submissions to the INSDC (in SRA (NCBI, DDBJ) and as "Experiment" contextual data in ENA). The tags have been used by the GenomeTrakr pathogen surveillance network to flag general quality control issues (or the lack thereof), as well as to provide additional quality control methods information.

The GenomeTrakr implementation demonstrates how the generic PHA4GE tags can be customized according to initiative-specific needs. GenomeTrakr adds standardized names of QC pipelines used by different data providers in the "quality_control_method_name" field, and has created other "quality_control_issues" tags that were subsequently added to the PHA4GE prescribed list i.e. " low coverage of characteristic mutations". We anticipate that as the tags are implemented for different organisms and initiatives, there may be other useful tags that should be included in the PHA4GE list. PHA4GE encourages feedback and suggestions from the community via the New Term Request form on GitHub. By sharing community needs and requests with PHA4GE in this way, we are able to work with ontology developers and public repository scientists to make new standardized vocabulary available through different channels. Also, it is possible to create different collections of specifications so that tags are honed for particular use cases. PHA4GE also recommends updating records when possible with known quality control issues.

There are many elements to standardizing quality control including specifying types of metrics and their parameters and thresholds, selecting and documenting tools and algorithms, prescribing different checkpoints in wet and dry lab processes, and so on. However, the PHA4GE Contextual Data QC Tag Specification does not delve into these more in-depth aspects, but rather the attributes act as quick, searchable, downstream flags for overall outcomes of QC assessments. Further development in standardized QC language and harmonized QC threshold for such nuanced aspects of QC frameworks is therefore needed. We hope that these simple tags will help improve communication around quality control in public repositories, as well as make datasets of variable quality easier to identify.

5. Conclusion

Availability and Requirements: The software used in this study is available on GitHub.

Project name: PHA4GE QC Contextual Data Tags Specification

Project home page: https://github.com/pha4ge/contextual_data_QC_tags

Operating system: Platform independent.

Programming language: Not applicable.

Other requirements: None.

License: MIT License.

List of abbreviations: CFSAN, Center for Food Safety and Applied Nutrition; DDBJ, DNA Data Bank of Japan; DSWG, Data Structures Working Group; EMBL-EBI, European Molecular Biology Laboratory-European Bioinformatics Institute; ENA, European Nucleotide Archive; FAIR, Findable, Accessible, Interoperable, Reusable; FDA, Food and Drug Administration; GENEPIO, Genomic Epidemiology Ontology; INSDC, International Nucleotide Sequence Database Collaboration; NTR, New Term Request; OBO, Open Biological and Biomedical Ontology; PHA4GE, Public Health Alliance for Genomic Epidemiology; QC, quality control; SRA, Sequence Read Archive.

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Funding: We wish to thank the Bill & Melinda Gates Foundation for supporting the establishment and work of the PHA4GE consortium. EJG and RC were funded by a Genome Canada CanCOGeN grant E09CMA. CIM was supported by the Fundação para a Ciência e Tecnologia (grants SFRH/BD/129483/2017 and COVID/BD/152583/2022). LC acknowledges funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/R015600/1), jointly funded by the UK Medical Research Council (MRC) and the UK Foreign, Commonwealth & Development Office (FCDO), under the MRC/FCDO Concordat agreement and is also part of the EDCTP2 programme supported by the European Union. The work of IKM was supported by the National Center for

Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health.

Acknowledgements: The authors would like to thank the many members of the bioinformatics community who volunteer their time to provide ongoing feedback and support to PHA4GE, without which this work would not be possible.

Competing Interests: The authors declare that they have no competing interests.

References

1. Black A et al (2020). Ten recommendations for supporting open pathogen genomic analysis in public health. Nat Med. 26(6):832-841. doi: 10.1038/s41591-020-0935-z. Epub 2020 Jun 11. PMID: 32528156; PMCID: PMC7363500.
2. Brown B et al (2021). An economic evaluation of the Whole Genome Sequencing source tracking program in the U.S.. PLoS ONE 16(10): e0258262. https://doi.org/10.1371/journal.pone.0258262
3. Carrillo CD & Blais BW (2021). Whole-Genome Sequence Datasets: A Powerful Resource for the Food Microbiology Laboratory Toolbox. Front. Sustain. Food Syst. 5:754988. doi: 10.3389/fsufs.2021.754988
4. Cook S (2021). Genomic surveillance in the roll out of vaccines. PHG Foundation. Accessed Jan 12 2023 https://www.phgfoundation.org/blog/genomic-surveillance-in-the-roll-out-of-vaccines
5. Gargis AS et al (2016). Assuring the Quality of Next-Generation Sequencing in Clinical Microbiology and Public Health Laboratories. J Clin Microbiol. 54(12):2857-2865. doi:10.1128/JCM.00949-16
6. Gozashti L & Corbett-Detig (2021). Shortcomings of SARS-CoV-2 genomic metadata. BMC Res Notes. 14:189.
7. Griffiths E et al (2022). Future-proofing and maximizing the utility of metadata: The PHA4GE SARS-CoV-2 contextual data specification package. GigaScience. 11. https://doi.org/10.1093/gigascience/giac003
8. Hendriksen RS (2019). Using Genomics to Track Global Antimicrobial Resistance. Front. Public Health 7:242. doi: 10.3389/fpubh.2019.00242
9. Lusignan S et al (2020). COVID-19 Surveillance in a Primary Care Sentinel Network: In-Pandemic Development of an Application Ontology. JMIR Public Health Surveill. 6(4): e21434.
10. Munnink BBO et al (2021), Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. Science, 371:6525. doi: 10.1126/science.abe5901.
11. Musen M (2022). Demand standards to sort FAIR data from foul. Nature. 609.
12. Petrillo M et al (2022). A roadmap for the generation of benchmarking resources for antimicrobial resistance detection using next generation sequencing [version 2; peer review: 1 approved, 2 approved with reservations]. F1000Research, 10:80 (https://doi.org/10.12688/f1000research.39214.2)
13. Pettengill JB (2021). Interpretative Labor and the Bane of Nonstandardized Metadata in Public Health Surveillance and Food Safety. Clin Infect Dis. 73(8):1537-1539. doi: 10.1093/cid/ciab615. PMID: 34240118.
14. Rick JA et al (2022). Reference genome choice and filtering thresholds jointly influence phylogenomic analyses. bioRxiv doi: https://doi.org/10.1101/2022.03.10.483737
15. Robinson ER et al (2013). Genomics and outbreak investigation: from sequence to consequence. Genome Med 5:36. https://doi.org/10.1186/gm440
16. Rossen JWA et al (2018). Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology. Clinical Microbiology and Infection. 24. 355-360.
17. Schriml L et al (2020). COVID-19 pandemic reveals the peril of ignoring metadata standards. Scientific Data. 7:188. https://doi.org/10.1038/s41597-020-0524-5
18. Smith B et al (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 25, 1251–1255. https://doi.org/10.1038/nbt1346
19. Smits THM (2019). The importance of genome sequence quality to microbial comparative genomics. BMC Genomics 20:662. https://doi.org/10.1186/s12864-019-6014-5
20. Stevens I et al (2020) Ten simple rules for annotating sequencing experiments. PLoS Comput Biol 16(10): e1008260. https://doi.org/ 10.1371/journal.pcbi.1008260
21. Timme RE et al (2019). Utilizing the Public GenomeTrakr Database for Foodborne Pathogen Traceback. Methods Mol Biol. 1918:201-212. doi: 10.1007/978-1-4939-9000-9_17.
22. Wagner DD et al (2021). Evaluating whole-genome sequencing quality metrics for enteric pathogen outbreaks. PeerJ 9:e12446 http://doi.org/10.7717/peerj.12446
23. Wilkinson M et al (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18
24. World Health Organization (2022). Global genomic surveillance strategy for pathogens with pandemic and epidemic potential, 2022–2032. Accessed Jan 12 2023. https://www.who.int/initiatives/genomic-surveillance-strategy
25. Xiaoli L et al (2022). Benchmark datasets for SARS-CoV-2 surveillance bioinformatics. PeerJ. 10:e13821. doi: 10.7717/peerj.13821. PMID: 36093336; PMCID: PMC9454940.