Shortcomings of SARS-CoV-2 Genomic Metadata

Landen Gozashti (lgozashti@g.harvard.edu)
Harvard University https://orcid.org/0000-0001-6023-3138
Russell Corbett-Detig
University of California Santa Cruz

Research note

Keywords: SARS-CoV-2, metadata, genomics, databases, data quality, COVID-19

Posted Date: March 18th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-322287/v1

Abstract

Objective

The SARS-CoV-2 pandemic has prompted one of the most extensive and expeditious genomic sequencing efforts in history. Each viral genome is accompanied by a set of metadata which supplies important information such as the geographic origin of the sample, age of the host, and the lab at which the sample was sequenced, and is integral to epidemiological efforts and public health direction. Here, we interrogate some shortcomings of metadata within the GISAID database to raise awareness of common errors and inconsistencies that may affect data-driven analyses and provide some possible avenues for solutions.

Results

Our analysis reveals a startling prevalence of spelling errors and inconsistent naming conventions, which together occur in an estimated ~9.8% and ~11.6% of “originating labs” and “submitting labs” GISAID metadata categories respectively. We also find numerous ambiguous entries which provide very little information about the actual source of a sample and could easily associate with multiple sources worldwide. Importantly, all of these issues can impair the ability and accuracy of association studies by deceptively causing a group of samples to identify with multiple sources when they truly all identify with one source, or vice versa.

Introduction

Metadata, or “data about data,” [1] is an essential component of science: informing both data-driven analyses and decisions with regards to public health [2–6]. Consequently, inadequate metadata quality can inhibit the discoverability of relevant data and hinder epidemiological research efforts and the development of clinical policy [3, 7, 8]. In spite of this, metadata standards remain neglected, and databases critical to public health related research efforts including Dryad, Genbank, BioSample (managed by the National Center for Biotechnology Information), BioSamples (managed by the European Bioinformatics Institute), the Electronic Health Record (EHR) and various repositories maintained by the CDC are plagued by inconsistencies and erroneous metadata entries [9–18].

As some groups have previously mentioned, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic has shed light on metadata inadequacies, which have inhibited studies relevant to both epidemiology and viral population dynamics [18–21]. Databases such as the global initiative on sharing avian influenza data (GISAID) [22] and Nextstrain [23] have empowered an impressive array of SARS-CoV-2 studies by maintaining SARS-CoV-2 genomic sequences and corresponding metadata. The GISAID database is the largest and most widely used database of SARS-CoV-2 genomic variation. Here, we specifically highlight inconsistencies and erroneous entries in “originating lab” and “submitting lab” descriptions within GISAID, which maintained 223,024 SARS-CoV-2 genomic sequences as of November

27th 2020, to exemplify where improvements in metadata quality are needed and to raise awareness to data submitters and maintainers alike.

Results And Discussion

We used a manual string comparison approach to estimate the prevalence of spelling errors and naming inconsistencies in “originating lab” and “submitting lab” metadata categories for all GISAID SARS-CoV-2 sequences as of November 27th 2020. Our analysis reveals that an alarmingly large proportion of lab names are misspelled or exhibit inconsistent naming conventions among samples at least once: ~9.8% and ~11.6% for “originating labs” and “submitting labs” respectively. Furthermore, we observe many instances in which lab names are misspelled or named inconsistently multiple times across samples, and cases of highly ambiguous lab names such as “Hospital” or “Biology Dpt” that could be associated with multiple sources (Figure 1A-C).

One of the primary consequences of spelling errors and inconsistent naming conventions in these particular categories (and more generally) is the appearance that a group of samples identifies with multiple sources, when they all truly identify with one particular source (Figure 1D). The opposite effect, where samples from disparate sources are erroneously associated with the same source, is also possible. Both of these deceptions can impair association studies. Notably, “originating lab” and “submitting lab” metadata categories are pertinent to the ability to accurately identify systematic sequencing errors associated with specific sequencing groups in SARS-CoV-2 genomes and the sources and causes of erroneous variants in SARS-CoV-2 genomic data [20, 24]. The challenges with accurate interpretation of these metadata fields has led to onerous workarounds such as using “country” as an imprecise proxy for the likely origin of a sequence [25]. Concerningly, the same metadata errors we describe have been propagated into downstream analysis platforms (*e.g., [26]*), further highlighting a need for improved metadata quality.

There are three possible solutions to the challenges of inconsistent and inaccurate metadata. First, we urge producers of SARS-CoV-2 genomic data to proceed with caution when submitting their metadata and advocate that maintainers of genomic databases be aware of possible errors in incoming metadata (such as those we show) and attentively promote metadata standardization. A second solution is to completely ignore samples with suspected corresponding metadata errors [18]. However, this solution can result in a significant decrease in sample size, limiting the power of statistical analyses [18]. On another hand, the development of new reliable methods for metadata correction could serve as an alternative and could likely be applied across multiple disciplines [1, 27, 28]. Methods for metadata quality evaluation and subsequent correction are in active development [4, 16, 28]. However, automated metadata correction is a nontrivial task, and future work is required to evaluate current algorithms for metadata correction and the feasibility of their application to large genomic databases like GISAID.

Conclusion

The SARS-CoV-2 pandemic has prompted an unprecedented response from the scientific and public health community, and the development and maintenance of databases such as GISAID have permitted epidemiological and comparative studies of unparalleled power. However, a brief analysis reveals that the quality of metadata accompanying such datasets remains unreliable. A study conducted by McMahon and Denaxas in 2016 concluded that “one of the main challenges in assessing quality in epidemiological and public health research is a lack of awareness of the issue of poor quality metadata” [4]. The SARS-CoV-2 pandemic is an unfortunate source of enlightenment to metadata shortcomings. Here we primarily focus on errors and inconsistencies, but metadata completeness and detail are of equivalent importance [21]. The importance of quality metadata with regard to our ability as a species to combat this pandemic and future pandemics is now more important than ever, and we must strive for a higher standard.

Limitations

This work primarily focuses on issues within the GISAID database and does not consider other SARS-CoV-2 genomic databases. It is possible that GISAID exemplifies an extreme case of metadata inconsistencies and that these observations are less prevalent across SARS-CoV-2 metadata as a whole.

Abbreviations

SARS-CoV-2, GISAID

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The datasets analyzed during the current study are available in the GISAID repository, https://www.gisaid.org/.

Competing interests

The authors declare that they have no competing interests.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions

LG and RCD originally noticed inconsistencies in metadata entries. LG conceived the idea to estimate the abundance of inconsistencies. LG performed the analysis. LG and RCD verified the analytical methods. Both authors discussed the results and contributed to the final manuscript.

Acknowledgements

The authors thank the GISAID database and all labs who contributed SARS-CoV-2 sequence data. A full acknowledgement table for all GISAID authors can be found at https://github.com/lgozasht/SARS-COV-2_COLLECTIVE_ANALYSIS/blob/master/gisaid_acknowledgements.tsv.gz. We also thank Hopi Hoekstra for advice and David Haussler and William P Hanage for feedback.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Goble C, Corcho O, Alper P, De Roure D. e-Science and the Semantic Web: A Symbiotic Relationship. In: Discovery Science. Springer Berlin Heidelberg; 2006. p. 1–12.
2. Matters MD, Lekiachvili A, Savel T, Zheng Z-J. Developing metadata to organize public health datasets. AMIA Annu Symp Proc. 2005;:1047.
3. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, et al. The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol. 2008;26:541–7.
4. McMahon C, Denaxas S. A novel framework for assessing metadata quality in epidemiological and public health research settings. AMIA Jt Summits Transl Sci Proc. 2016;2016:199–208.
5. Martin MA, VanInsberghe D, Koelle K. Insights from SARS-CoV-2 sequences. Science. 2021;371:466–7.
6. Bernasconi A, Canakoglu A, Masseroli M, Ceri S. META-BASE: a Novel Architecture for Large-Scale Genomic Metadata Integration. IEEE/ACM Trans Comput Biol Bioinform. 2020;PP. doi:10.1109/TCBB.2020.2998954.
7. Embi PJ, Richesson R, Tenenbaum J, Kannry J, Friedman C, Sarkar IN, et al. Reimagining the research-practice relationship: policy recommendations for informatics-enabled evidence-generation across the US health system. JAMIA Open. 2019;2:2–9.
8. Wurtz R. The role of public health in health information exchanges. J Public Health Manag Pract. 2013;19:485–7.
9. Fabreau GE, Minty EP, Southern DA, Quan H, Ghali WA. A Meta-Data Manifesto: The Need for Global Health Meta-Data. Int J Popul Data Sci. 2018;3:436.

10. Field D, Amaral-Zettler L, Cochrane G, Cole JR, Dawyndt P, Garrity GM, et al. The Genomic Standards Consortium. PLoS Biol. 2011;9:e1001088.
11. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018.
12. Hoffman S, Podgurski A. Big bad data: law, public health, and biomedical databases. J Law Med Ethics. 2013;41 Suppl 1:56–60.
13. National Research Council (US) Board on Biology, Pool R, Esnayra J. Maintaining the Integrity of Databases. National Academies Press (US); 2000.
14. Ozkaynak H, Glenn B, Qualters JR, Strosnider H, McGeehin MA, Zenick H. Summary and findings of the EPA and CDC symposium on air pollution exposure and health. J Expo Sci Environ Epidemiol. 2009;19:19–29.
15. Gonçalves RS, Musen MA. The variable quality of metadata about biological samples used in biomedical experiments. Sci Data. 2019;6:190021.
16. Schmedes SE, King JL, Budowle B. Correcting Inconsistencies and Errors in Bacterial Genome Metadata Using an Automated Curation Tool in Excel (AutoCurE). Front Bioeng Biotechnol. 2015;3:138.
17. Rousidis D, Garoufallou E, Balatsoukas P, Sicilia M-A. Metadata for Big Data: A preliminary investigation of metadata quality issues in research data repositories. Inf Serv Use. 2014;34:279–86.
18. Velazquez A, Bustria M, Ouyang Y, Moshiri N. An analysis of clinical and geographical metadata of over 75,000 records in the GISAID COVID-19 database. 2020. doi:10.1101/2020.09.22.20199497.
19. Kaiser KA, Chodacki J, Habermann T, Kemp J, Paglione L, Urberg M, et al. Metadata: The accelerant we need. Inf Serv Use. 2020;40:181–91.
20. Turakhia Y, De Maio N, Thornlow B, Gozashti L, Lanfear R, Walker CR, et al. Stability of SARS-CoV-2 phylogenies. PLoS Genet. 2020;16:e1009175.
21. Schriml LM, Chuvochina M, Davies N, Eloe-Fadrosh EA, Finn RD, Hugenholtz P, et al. COVID-19 pandemic reveals the peril of ignoring metadata standards. Sci Data. 2020;7:188.
22. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill. 2017;22. doi:10.2807/1560-7917.ES.2017.22.13.30494.
23. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018;34:4121–3.
24. De Maio N, Walker C, Borges R, Weilguny L, Slodkowicz G, Goldman N. Issues with SARS-CoV-2 sequencing data. 2020. https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473/1.
25. Gozashti L, Walker C, Goldman N, Corbett-Detig R, De Maio N. Issues with SARS-CoV-2 sequencing data: Updated analysis with data from 13th November 2020. 2020. https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473/14.
26. Canakoglu A, Pinoli P, Bernasconi A, Alfonsi T, Melidis DP, Ceri S. ViruSurf: an integrated database to investigate viral sequences. Nucleic Acids Res. 2021;49:D817–24.

27. Michener WK. Ecological data sharing. Ecol Inform. 2015;29:33–44.
28. Assaf A, Senart A, Troncy R. Roomba: Automatic Validation, Correction and Generation of Dataset Metadata. In: Proceedings of the 24th International Conference on World Wide Web. New York, NY, USA: Association for Computing Machinery; 2015. p. 159–62.

Figures

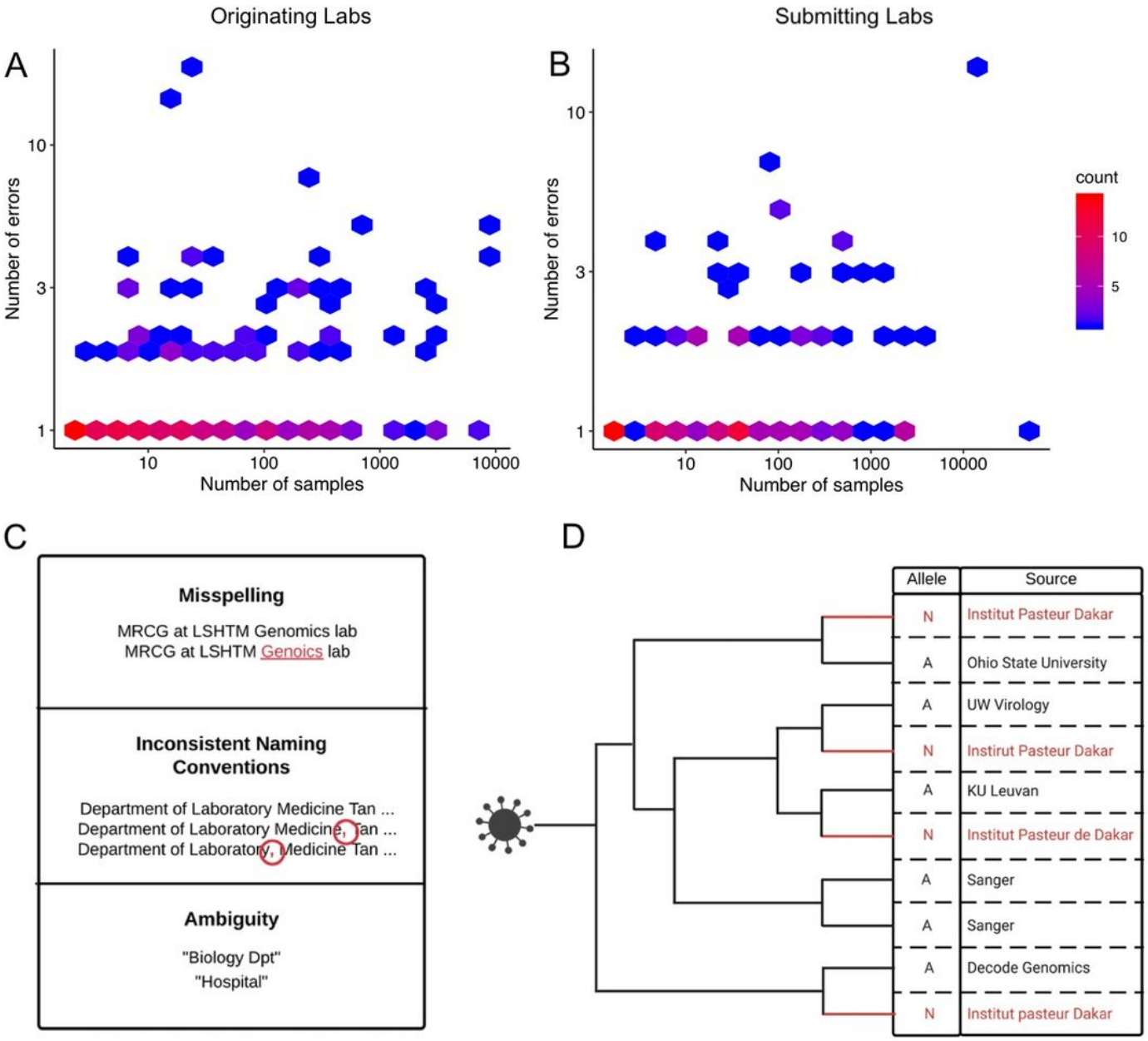

Figure 1

The number of samples produced by each (A) “originating lab” and (B) “submitting lab” and the corresponding number of errors (or inconsistencies) for that respective source. Color encodes the respective number of data points at a given position on the plot, with positions with fewer points shaded blue and positions with more points shaded red. (C) Some observed examples of misspellings, inconsistent naming conventions, and highly ambiguous entries. (D) A phylogenetic tree displaying an example of a case in which errors in “originating lab” metadata might impede association studies with regard to SARS-CoV-2 genomic data. In this case, ambiguous “N” alleles occur multiple times across a phylogeny at a given site and all stem from the same source. However, metadata errors (shown in red) cause this ambiguous “N” allele to appear as if it is associated with 4 different sources (rather than 1). Such a site could impair phylogenetic inference and should be flagged in the SARS-CoV-2 masking recommendations but could be overlooked as a result of these errors [24].