Patient-Related Metadata Reported in Sequencing Studies of SARS-CoV-2: Protocol for a Scoping Review and Bibliometric Analysis

Karen O'Connor[1], Davy Weissenbacher[2], Amir Elyaderani[3], Matthew Scotch[3,4], Graciela Gonzalez-Hernandez[2]

[1]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[2]Department of Computational Biomedicine, Cedars-Sinai Medical Center, West Hollywood, CA, USA

[3]Biodesign Center for Environmental Health Engineering, Arizona State University, Tempe, AZ, USA

[4]College of Health Solutions, Arizona State University, Tempe, AZ, USA

Corresponding Author:

Karen O'Connor

Senior Data Analyst

Department of Biostatistics, Epidemiology and Informatics

Perelman School of Medicine

University of Pennsylvania

Email: karoc@pennmedicine.upenn.edu

Abstract

Background: Since the onset of the COVID-19 pandemic, there has been an unprecedented effort in genomic epidemiology to sequence the SARS-CoV-2 virus and examine its molecular evolution. This has been facilitated by the availability of publicly accessible databases, GISAID and GenBank, which collectively hold millions of SARS-CoV-2 sequence records. However, genomic epidemiology seeks to go beyond phylogenetic analysis by linking genetic information to patient demographics and disease outcomes, enabling a comprehensive understanding of transmission dynamics and disease impact.

While these repositories include some patient-related information, such as the location of the infected host, the granularity of this data and the inclusion of demographic and clinical details are inconsistent. Additionally, the extent to which patient-related metadata is reported in published sequencing studies remains largely unexplored. Therefore, it is essential to assess the extent and quality of patient-related metadata reported in SARS-CoV-2 sequencing studies.

Moreover, there is limited linkage between published articles and sequence repositories, hindering the identification of relevant studies. Traditional search strategies based on keywords may miss relevant articles. To overcome these challenges, this study proposes the use of an automated classifier to identify relevant articles.

Objective: This study aims to conduct a systematic and comprehensive scoping review, along with a bibliometric analysis, to assess the reporting of patient-related metadata in SARS-CoV-2 sequencing studies.

Methods: The NIH's LitCovid collection will be used for the machine learning classification, while an independent search will be conducted in PubMed. Data extraction will be conducted using Covidence, and the extracted data will be synthesized and summarized to quantify the availability of patient metadata in the published literature of SARS-CoV-2 sequencing studies. For the bibliometric analysis, relevant data points, such as author affiliations, journal information, and citation metrics, will be extracted.

Results: The study will report findings on the extent and types of patient-related metadata reported in genomic viral sequencing studies of SARS-CoV-2. The scoping review will identify gaps in the reporting of patient metadata and make recommendations for improving the quality and consistency of reporting in this area. The bibliometric analysis will uncover trends and patterns in the reporting of patient-related metadata, such as differences in reporting based on study types or geographic regions. Co-occurrence networks of author keywords will also be presented to highlight frequent themes and their associations with patient metadata reporting.

Conclusion: This study will contribute to advancing knowledge in the field of genomic epidemiology by providing a comprehensive overview of the reporting of patient-related metadata in SARS-CoV-2 sequencing studies. The insights gained from this study may help improve the quality and consistency of reporting patient metadata, enhancing the utility of sequence metadata and facilitating future research on infectious diseases. The findings may also inform the development of machine learning methods to automatically extract patient-related information from sequencing studies.

Keywords: SARS-CoV-2; COVID-19; genomic epidemiology; scoping review; protocol

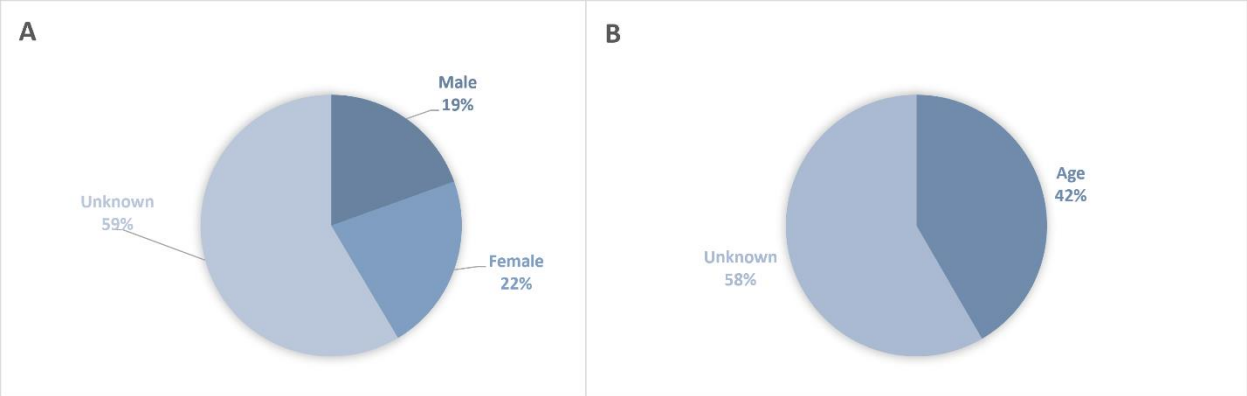
Introduction

Since the onset of the COVID-19 pandemic, there has been an unprecedented effort in genomic epidemiology to sequence the virus, study its transmission, and examine molecular evolution. Public repositories, such as the Global Initiative on Sharing Avian Influenza Data (GISAID)[1] and NCBI's GenBank[2] host millions of SARS-CoV-2 sequence records. As of July 2023, GISAID contains 15.7 million sequences, while 7.7 million have been deposited in GenBank.

The availability of this vast amount of genomic data has facilitated significant discoveries, particularly in phylogenetic and phylodynamic studies [3–5]. Beyond phylogenetic studies, genomic epidemiology aims to understand the transmission dynamics, evolution, and impact of infectious diseases by analyzing the genetic information of pathogens and linking it to patient demographics and disease outcomes [6,7]. Thus enabling the tracking of the spread of pathogens, identifying high-risk populations, and discovering genetic factors that influence disease transmission, severity, and treatment response[6,8]. This knowledge can, in turn, inform public health strategies, guide the development of targeted interventions, and improve the overall understanding of infectious diseases[9].

Ideally, patient geographic, demographic, and clinical information should be included in the sequence metadata upon its submission to the repository. Both GISAID and GenBank provide the location of the infected host (LOIH) information in their sequence metadata, with almost 100% coverage for both databases. However, the reported location granularity may vary and often lacks important details such as patient travel history. Patient demographic and clinical information is rarely complete. Approximately, 60% of sequences in GISAID have the age or gender for the infected host of the sequence entered as unknown (e.g., 'Not Available', 'Declined', 'Not Reported', etc.) (Figure 1), while GenBank lacks any standardized fields to include this information with sequence submissions.

Figure 1: Percent of sequences with reported gender (A) or age (B) in GISAID*.



*data downloaded April 3, 2023, representing 15.3 million sequences

This patient information, or at least a subset of it, may be reported in the published studies of those who obtained and performed the genomic sequencing. Previous studies demonstrated the potential enrichment of LOIH information for GenBank sequences through the automated

extraction of location information from publications[10,11]. The extent to which location information as well as patient demographic or clinical information is reported in SARS-CoV-2 sequencing studies remains largely unexplored. Our review aims to bridge this gap in understanding by quantifying the extent and types of patient-related metadata reported in genomic viral sequencing studies of SARS-CoV-2.

In addition to the vast sequencing efforts of SAR-CoV-2, a substantial number of research articles related to SARS-C0V-2 and the pandemic have been published. The NCBI SARS-CoV-2 resource page lists over 573,000 articles in PubMed Central [12]. However, there is sparse linkage between the two resources making it difficult to identify publications relevant to the sequences in the databases. While submissions to GenBank allow the linkage of related publications indexed in PubMed, this process may be incomplete or require subsequent updates. The number of sequences linked to a publication in NCBI data remains low, currently of the 7.7 million sequences in the SARS-CoV-2 Data Hub[13] only 9057 sequences (0.001%) are linked to xx publications.

Traditionally, identifying studies for a review requires the development of a search strategy of databases, such as Medline or Scopus, using keywords and index terms. The selection of keywords greatly influences search results, leading to potential missed or irrelevant studies. Moreover, discussions of sequencing are often confined to the methods section of papers, rendering title and abstract screening less informative. Additionally, mentions of the repositories, GISAID or GenBank, or sequence identifiers do not reliably indicate that the article is discussing original sequencing. To overcome these limitations, we propose using an automated classifier to identify relevant studies for review.

A bibliometric analysis uses different methods and data points to quantify the trends and assess the impact of publications in a specific field [14]. While several bibliometric analyses have investigated COVID-19 related research trends, in general, [15–17] and in specific fields such as neurology[18], long Covid [19] and medical imaging [20], or specific geographic locations such as Africa[21], no analysis has specifically focused on the publication trends related to the reporting of patient metadata related to SARS-CoV-2 genomic sequences. Our aims with this review and analysis are to identify reporting and publication trends as well as highlight the gaps in reporting that may hinder the advancement of genomic epidemiology studies of the COVID-19 pandemic.

Primary Research Objectives

1. To quantitatively assess the extent and quality of patient-reported metadata, including demographic, clinical, and geographic information, in articles reporting original whole genome sequencing of the SARS-CoV-2 virus.

2. To perform a comprehensive bibliometric analysis to ascertain differences and discernible patterns between articles that include patient metadata and those that do not, thereby providing insights into the characteristics and factors associated with the reporting of patient data in the literature.

3. To evaluate the efficacy and reliability of a machine learning classifier in accurately identifying relevant articles for inclusion in the scoping review, enhancing the efficiency and effectiveness of the study selection process.

Methods

Our review will follow the methodological framework identified by Arksey and O'Malley[22]. The scoping review will be reported in line with the PRISMA-SrC checklist [23].

Data Sources

We will utilize the NIH's LitCovid collection 24 for our machine learning classification. LitCovid is a curated collection of scholarly articles related to the Coronavirus Disease 2019 (COVID-19). The collection contains over 319,000 publications from 8,000 journals and is updated daily. LitCovid includes published articles as well as preprints. Additionally, we will independently search PubMed using a two-faceted search strategy and the NCBI E-utilities program to find publications linked to sequences. This combined approach will help ensure a comprehensive coverage of the literature for our study.

Search Strategy

Classification Model

Our classification model was trained using manually annotated data. A full-text search strategy was developed to filter the LitCovid collection resulting in a corpus of targeted articles for annotation.

The papers identified through the pipeline were annotated by two experienced annotators using the Inception annotation tool [24]. The annotators reviewed the full text of the articles and labeled sentences which confirmed the study's performance of SAR-CoV-2 sample sequencing from human specimens.

The classifier is instantiated as a pre-trained neural network, specifically a transformer model called BERT-base-uncased. We fine-tuned the model to perform our task requirements. We trained the classifier using the corpus of 245 annotated articles to detect sentences indicating sequencing and disregarding the negative ones. The initial model was trained for 20 epochs, with the best performing model, based on F1-score on the validation set, being selected and evaluated on our test set. The model achieved moderate performance with 0.480 F1-score, 0.492 precision, and 0.469 recall.

Search Strategy

To evaluate our classifier and identify studies that may have been missed due to classification errors or lack of full text in the LitCovid collection, we will create a search strategy to independently search PubMed. We will develop a two-faceted search strategy to find "SARS-

CoV-2" and "whole genome sequencing" related publications. We will utilize the search strategy developed for the LitCovid collection with additional keywords added to identify studies that report whole genome sequencing. Additionally, we will search for publications linked to SARS-CoV-2 sequences using the NCBIs E-utilities eLink programming API.

A publication date restriction of December 2019 onwards will be used in the searches as this review is focused on SARS-CoV-2 sequencing studies. No language restrictions will be placed on the searches, although financial and logistical restraints will not allow translation from all languages. All results will be uploaded to a Zotero [25] library where duplicate results will be removed.

Inclusion/Exclusion Criteria

Papers positively identified by our classifier and our search results will be reviewed for inclusion in the review based on the criteria outlined in Table 1.

Table 1. Inclusion and exclusion criteria for the scoping review.

Facet	Inclusion Criteria	Exclusion Criteria
Sample Origin	Individual human subject	• Non-human sources (e.g., mice, bats, ferrets) • Wastewater • Microbiome • Cloned/Cell Culture Virus
Sequencing Type	Whole genomic sequencing	Studies will be excluded if the following sequencing methods were exclusively performed: • PCR or LAMP for viral detection • Single-cell sequencing • Gene expression studies • Protocol validation studies on cell culture virus • Exome sequencing
Study Design	Any type of peer-reviewed or preprint study reporting on the original sequencing of SARS-CoV-2 samples.	Any other study design.
Publication Dates	December 2019, or later	Before December 2019
Language	All	None

Two reviewers will perform title and abstract screening using the Covidence [26] systematic review management tool with any disagreements resolved by discussion. Two independent reviewers will also conduct full-text screening in Covidence.

Data Extraction

Data extraction will be conducted in Covidence. The reviewers will examine the full text of the articles, including any supplementary files, for data extraction. The customizable interface will be designed to prompt the reviewer to extract various details, such as general publication information, study characteristics, sequencing specifics, and the presence or absence of patient demographic, clinical, or location information. Furthermore, the location of this information within the articles will be noted. An example of the data extraction form can be found in Table 2.

Table 2: Example of data that will be extracted from included studies.

Prompt	**Response**
Publication Information	
Study Name	Free text
Article Title	Free text
Year of Publication	YYYY
Publication Type	Journal, Conference, Preprint
Study and Sequence Information	
Study Objective	Free text
Location of Study (country)	Free text
Number of patients	Free text
Number of samples sequenced	Free text
Repository sequences deposited to	GISAID, GenBank, Other, NR
For studies with >1 patient, are sequences linked to a patient?	Yes|No
Patient Demographic Information Reported	
Age	Yes|No
Gender	Yes|No
Race/Ethnicity	Yes|No
Patient Clinical Information Reported	
Symptoms	Yes|No
Severity	Yes|No
Inpatient or Outpatient	Yes|No
Treatments	Yes|No
Outcomes	Yes|No

Patient Geographic Information Reported	
Location of Residence	Yes|No
Travel Information	Yes|No

We will test the initial extraction form on a subset of articles and revise it as needed.

For bibliometric analysis, all pertinent data points will be extracted for studies included in our review including, author location and institution information, journal, study type, citation metrics, and author keywords.

Data Analysis

The extracted data will be synthesized and summarized to quantify the availability of patient metadata in the published literature of SARS-CoV-2 sequencing studies using an exported spreadsheet from Covidence. For the bibliometric analysis, data will be analyzed and visualized using the VOSviewer[27] software or the bibliometrix[28] package for R. Other software tools may be used as needed for analysis.

Results

We will summarize and narratively describe our findings, using tables, graphs, and charts when applicable regarding the number of sequences covered in our included studies, the distribution of the sequences in the respective repositories, and the quantity and type of reported patient metadata in the studies. We will also present the geographical location of the study's authors using maps and report our findings, including the most frequent journals and article types, as well as analyze differences between studies that reported patient data from those that did not. Co-occurrence networks of author keywords will be presented to highlight the frequency and differences in themes and study focus between the reporting groups.

Discussion

There has been an unprecedented effort in the sequencing and sharing of the viral genomes of SARS-CoV-2 through publicly available databases, such as GISAID and GenBank, during the COVID-19 pandemic. However, the utility of these sequences for genomic epidemiology may not be fully realized due to the unavailability of relevant metadata about the patient from whom the specimen was obtained [29]. The purpose of our study is to conduct a scoping review and bibliometric analysis focusing on patient metadata reporting in genomic viral sequencing studies of SARS-CoV-2.

This scoping review will provide valuable insights into the current state of reporting of patient-related metadata in SARS-CoV-2 sequencing studies. The findings of the review will be used to identify gaps in the reporting of patient metadata and to make recommendations for improving

the quality and consistency of reporting of patient-related metadata in SARS-CoV-2 sequencing studies.

In addition to the findings of our scoping review, the bibliometric analysis will likely identify several other important trends and patterns in the reporting of patient-related metadata. For example, the analysis may find that the reporting of patient-related metadata is more common in certain types of studies, or that it is more likely to be reported in studies from certain geographic regions. The findings of the scoping review and bibliometric analysis will provide valuable insights into the factors that influence the reporting of patient-related metadata and will help to inform future research on this topic. Furthermore, the identification and quantification of the metadata in literature may aid in advancing other research, such as the development of machine learning methods to extract this information and enhance sequence data through automatic methods.

Strengths/Limitations

This study will conduct a systematic and comprehensive scoping review, encompassing a large number of articles from various databases ensuring a thorough examination of the current state of reporting patient-related metadata in SARS-CoV-2 sequencing studies, and will provide a comprehensive overview of the available literature. The inclusion of bibliometric analysis will go beyond the scoping review to analyze publication trends, author affiliations, citation metrics, and other bibliographic information to provide several insights into the broader landscape of research that includes patient metadata reporting in genomic viral sequencing studies of SARS-CoV-2.

While some relevant studies may be missed due to search limitations and the classification model, our study's strength lies in providing valuable insights into the current state of reporting patient-related metadata. Although certain limitations exist, such as potential limitations in reported patient metadata[30,31] and the focus on SARS-CoV-2 sequencing studies, our findings will contribute to improving the quality and consistency of reporting in genomic epidemiology. Future research can build upon our study to address these gaps and enhance reporting practices in this field.

Conclusion

This protocol outlines the steps that we will take in our scoping review which will be supported by an automated classifier and bibliometric analysis. We will fill the knowledge gap regarding the extent and types of patient-related metadata reported in genomic viral sequencing studies of SARS-CoV-2 and will provide valuable insights by identifying themes and trends in the published literature. The results of this study may encourage improved and standardized reporting practices which will significantly enhance the utility of sequence metadata and aid in advancing our understanding of the SARS-CoV-2 or any future pandemic.

Funding Statement

Research reported in this publication was supported by the National Institute of Allergy And Infectious Diseases of the National Institutes of Health under Award Number R01AI164481

1. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance*. 2017;22(13):1-1. doi:10.2807/1560-7917.ES.2017.22.13.30494

2. Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. GenBank. *Nucleic Acids Res*. 2019;47(D1):D94-D99. doi:10.1093/nar/gky989

3. Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc Natl Acad Sci*. 2020;117(17):9241-9243. doi:10.1073/pnas.2004999117

4. van Dorp L, Acman M, Richard D, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect Genet Evol*. 2020;83:104351. doi:10.1016/j.meegid.2020.104351

5. Tang X, Wu C, Li X, et al. On the origin and continuing evolution of SARS-CoV-2. *Natl Sci Rev*. 2020;7(6):1012-1023. doi:10.1093/nsr/nwaa036

6. Hill V, Ruis C, Bajaj S, Pybus OG, Kraemer MUG. Progress and challenges in virus genomic epidemiology. *Trends Parasitol*. 2021;37(12):1038-1049. doi:10.1016/j.pt.2021.08.007

7. Tang P, Croxen MA, Hasan MR, Hsiao WWL, Hoang LM. Infection control in the new age of genomic epidemiology. *Am J Infect Control*. 2017;45(2):170-179. doi:10.1016/j.ajic.2016.05.015

8. National Academies of Sciences, Engineering, and Medicine. *Genomic Epidemiology Data Infrastructure Needs for SARS-CoV-2: Modernizing Pandemic Response Strategies.* The National Academies Press; 2020.

9. Worlds Health Organization. *Genomic Sequencing of SARS-CoV-2 A Guide to Implementation for Maximum Impact on Public Health*.; 2021.

10. Tahsin T, Weissenbacher D, O'Connor K, Magge A, Scotch M, Gonzalez-Hernandez G. GeoBoost: accelerating research involving the geospatial metadata of virus GenBank records. *Bioinforma Oxf Engl*. 2018;34(9):1606-1608. doi:10.1093/bioinformatics/btx799

11. Magge A, Weissenbacher D, O'Connor K, Tahsin T, Gonzalez-Hernandez G, Scotch M. GeoBoost2: A Natural Language Processing Pipeline for GenBank Metadata Enrichment for Virus Phylogeography. *Bioinformatics*. Published online July 2020. doi:10.1093/bioinformatics/btaa647

12. NCBI SARS-CoV-2 Resources. Accessed July 10, 2023. https://www.ncbi.nlm.nih.gov/sars-cov-2/

13. NCBI Virus. Accessed July 11, 2023. https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=SARS-CoV-2,%20taxid:2697049

14. Gutiérrez-Salcedo M, Martínez MÁ, Moral-Munoz JA, Herrera-Viedma E, Cobo MJ. Some bibliometric procedures for analyzing and evaluating research fields. *Appl Intell*. 2018;48(5):1275-1287. doi:10.1007/s10489-017-1105-y

15. Hossain MM. Current Status of Global Research on Novel Coronavirus Disease (COVID-19): A Bibliometric Analysis and Knowledge Mapping. Published online May 18, 2020. doi:10.2139/ssrn.3547824

16. Nasab FR, Rahim F. *Bibliometric Analysis of Global Scientific Research on SARS-CoV-2 (COVID-19)*. Health Informatics; 2020. doi:10.1101/2020.03.19.20038752

17. Yu Y, Li Y, Zhang Z, et al. A bibliometric analysis using VOSviewer of publications on COVID-19. *Ann Transl Med*. 2020;8(13):816-816. doi:10.21037/atm-20-4235

18. Zhang Q, Li J, Weng L. A bibliometric analysis of COVID-19 publications in neurology by using the visual mapping method. *Front Public Health*. 2022;10. Accessed July 11, 2023. https://www.frontiersin.org/articles/10.3389/fpubh.2022.937008

19. Kim TH, Jeon SR, Kang JW, Kwon S. Complementary and Alternative Medicine for Long COVID: Scoping Review and Bibliometric Analysis. *Evid-Based Complement Altern Med ECAM*. 2022;2022:7303393. doi:10.1155/2022/7303393

20. Wen R, Zhang M, Xu R, et al. COVID-19 imaging, where do we go from here? Bibliometric analysis of medical imaging in COVID-19. *Eur Radiol*. 2023;33(5):3133-3143. doi:10.1007/s00330-023-09498-z

21. Guleid FH, Oyando R, Kabia E, Mumbi A, Akech S, Barasa E. A bibliometric analysis of COVID-19 research in Africa. *BMJ Glob Health*. 2021;6(5):e005690. doi:10.1136/bmjgh-2021-005690

22. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol*. 2005;8(1):19-32. doi:10.1080/1364557032000119616

23. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann Intern Med*. 2018;169(7):467-473. doi:10.7326/M18-0850

24. Klie JC, Bugert M, Boullosa B, de Castilho RE, Gurevych I. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation.

25. Zotero | Your personal research assistant. Accessed July 14, 2023. https://www.zotero.org/

26. Covidence - Better systematic review management. Covidence. Accessed July 14, 2023. https://www.covidence.org/

27. VOSviewer - Visualizing scientific landscapes. VOSviewer. Accessed July 11, 2023. https://www.vosviewer.com//

28. Aria M, Cuccurullo C. bibliometrix: An R-tool for comprehensive science mapping analysis. *J Informetr*. 2017;11(4):959-975. doi:10.1016/j.joi.2017.08.007

29. Grad YH, Lipsitch M. Epidemiologic data and pathogen genome sequences: a powerful synergy for public health. *Genome Biol*. 2014;15(11):538. doi:10.1186/s13059-014-0538-4

30. Hernandez MM, Gonzalez-Reiche AS, Alshammary H, et al. Molecular evidence of SARS-CoV-2 in New York before the first pandemic wave. *Nat Commun*. 2021;12(1):3463. doi:10.1038/s41467-021-23688-7

31. Page AJ, Mather AE, Le-Viet T, et al. Large-scale sequencing of SARS-CoV-2 genomes from one region allows detailed epidemiology and enables local outbreak management. *Microb Genomics*. 2021;7(6):000589. doi:10.1099/mgen.0.000589