A core-genome multilocus sequence typing scheme for the detection of genetically related Streptococcus pyogenes clusters

Toorop, M.M.A.; Kraakman, M.E.M.; Hoogendijk, I.V.; Prehn, J. van; Claas, E.C.J.; Wessels, E.; Boers, S.A.

Journal of Clinical Microbiology

Clinical Microbiology | Full-Length Text

A core-genome multilocus sequence typing scheme for the detection of genetically related *Streptococcus pyogenes* clusters

Myrthe M. A. Toorop,[1] Margriet E. M. Kraakman,[1] Irene V. Hoogendijk,[1] Joffrey van Prehn,[1] Eric C. J. Claas,[1] Els Wessels,[1] Stefan A. Boers[1]

AUTHOR AFFILIATION See affiliation list on p. 7.

ABSTRACT The recently observed increase in invasive *Streptococcus pyogenes* infections causes concern in Europe. However, conventional molecular typing methods lack discriminatory power to aid investigations of outbreaks caused by *S. pyogenes*. Therefore, there is an urgent need for high-resolution molecular typing methods to assess genetic relatedness between *S. pyogenes* isolates. In the current study, we aimed to develop a novel high-resolution core-genome multilocus sequence typing (cgMLST) scheme for *S. pyogenes* and compared its discriminatory power to conventional molecular typing methods. The cgMLST scheme was designed with the commercial Ridom SeqSphere+ software package. To define a cluster threshold, the scheme was evaluated using publicly available data from nine defined *S. pyogenes* outbreaks in the United Kingdom. The cgMLST scheme was then applied to 23 isolates from a suspected *S. pyogenes* outbreak and 117 *S. pyogenes* surveillance isolates both from the Netherlands. MLST and *emm*-typing results were used for comparison to cgMLST results. The allelic differences between isolates from defined outbreaks ranged between 6 and 31 for isolates with the same *emm*-type, resulting in a proposed cluster threshold of ≤5 allelic differences out of 1,095 target loci. Seven out of twenty-three (30%) isolates from the suspected outbreak had an allelic difference of ≤2, thereby identifying a potential cluster that could not be linked to other isolates. The proposed cgMLST scheme shows a higher discriminatory ability when compared to conventional typing methods. The rapid and simple analysis workflow allows for extended detection of clusters of potential outbreak isolates and surveillance and may facilitate the sharing of sequencing results between (inter)national laboratories.

KEYWORDS core-genome multilocus sequence typing, cgMLST, next-generation sequencing, NGS, *Streptococcus pyogenes*, group A *Streptococcus*

Streptococcus pyogenes [group A *Streptococcus* (GAS) according to the Lancefield classification (1)] is a bacterium that causes a large range of diseases, varying from superficial skin diseases and pharyngitis to more severe diseases such as necrotizing fasciitis and is associated with significant morbidity and mortality (2). Although severe GAS infections occur sporadically, community outbreaks (OBs) of GAS have been frequently reported (3–5). In the Netherlands, a post-COVID-19 increase in invasive group A streptococcal disease (iGAS) was noted by the national surveillance program (6, 7). This trend is also observed in other European countries where recent reports show an increase in outbreaks of iGAS, especially among children (8).

Identification of an outbreak has traditionally been made based on the combination of molecular typing of isolates and the presence of an epidemiological link. Commonly used molecular typing methods to differentiate among *S. pyogenes* isolates include single-locus sequence typing of the *emm*-gene and a multilocus sequence typing (MLST) scheme published by Tewodros and Kronvall (9). Sequence analysis of 180 bp of the

Editor Alexander Mellmann, Westfalische Wilhelms-Universitat Munster, Münster, Germany

Address correspondence to Stefan A. Boers, s.a.boers@lumc.nl.

The authors declare no conflict of interest.

Received 4 May 2023
Accepted 7 August 2023

Published 10 October 2023

emm-gene, which encodes the M protein, has resulted in differentiation of more than 200 *S. pyogenes* M genotypes that are associated with varying levels of virulence (9). In MLST, nucleotides of seven housekeeping genes are used to analyze the genetic relationships of *S. pyogenes*, resulting in an allelic profile (sequence type or "ST") (10). At least 1,367 different STs have been identified over the years (11). Although these conventional sequence-based typing methods are highly robust and the data achieved by different laboratories can be reliably compared using online databases, these methods may lack sufficient discriminatory power to distinguish different isolates within the same lineage that may be responsible for an outbreak (12). Therefore, there is a need for higher-resolution molecular typing methods to assess genetic relatedness between *S. pyogenes* strains.

Whole-genome sequencing (WGS) has the ability to produce more discriminatory power to differentiate isolates and evaluate outbreaks by including a higher number of target genes (13). Next-generation sequencing (NGS) has enabled the cost-effective implementation of WGS in the diagnostic laboratory. Genome similarities can be investigated by single-nucleotide polymorphism (SNP)-based mapping, in which sequence reads are compared to a reference genome to identify SNP variations and define clusters (14, 15). A disadvantage of this method is that it is difficult to standardize between laboratories due to differences in quality assurance criteria and reference genomes (16).

For several bacterial species, gene-by-gene-based approaches, such as whole-genome MLST (wgMLST) or core-genome MLST (cgMLST) schemes have been used for strain differentiation (17).

CgMLST schemes are developed using a fixed set of target genes (i.e., core genes) allocated throughout the genome that can be identified in most strains of a particular species (18). For different bacterial species, this technique proved to be a highly discriminative, efficient, and reliable tool for the differentiation of strains (17). Using a fixed set of target genes, a standardized comparison between laboratories can be achieved. A recent study by Friães et al. found that wg/cgMLST has a higher discriminatory power for the detection of *S. pyogenes* isolate variations than the conventional MLST schema and could, therefore, assist in further discriminating within STs (19). To our knowledge, there are no other studies available that used cgMLST for *S. pyogenes* typing. So far, cgMLST schemes for *S. pyogenes* strains using widely available analyzing software such as Ridom SeqSphere+ are lacking. In the current study, we aimed to develop such a novel cgMLST scheme using a local cluster and publicly available data sets.

MATERIALS AND METHODS

Development of the *S. pyogenes* cgMLST scheme

A novel cgMLST scheme was designed with the commercially available Ridom SeqSphere+ (version 8.3.5) software package (20). We used WGS data of 66 publicly available *S. pyogenes* genomes, comprising 38 different *emm*-types and 45 different sequence types, adopted from the NCBI RefSeq database (21) that was downloaded on 19 August 2022. A total of 1,095 common target genes within the genome of each strain were identified and used for developing the cgMLST scheme. The 66 genomes that were used represent a broad spectrum of S. *pyogenes* strains (Table S1).

Evaluation of the *S. pyogenes* cgMLST scheme (Data set 1)

The cgMLST scheme was evaluated using Data set 1, which included publicly available data (52 isolates from nine different *S. pyogenes* outbreaks that were independently reported to Public Health England with available epidemiology data), which were documented in England (2010–2015) (22). Previous analyses using wg/cgMLST confirmed that these isolates were indeed part of an outbreak (19). These data have been used to determine a cluster threshold to accurately distinguish isolates from

different clonal lineages. FastA-files were downloaded from an open data source (23) and processed further using Ridom SeqSphere+ (version 8.3.5) and PubMLST (24) for data analysis. For comparability, the same OB numbers as in the initial publication have been used (22).

Application of the *S. pyogenes* cgMLST scheme (Data sets 2 and 3)

The cgMLST scheme was applied to isolates from a potential *S. pyogenes* outbreak and WGS-data from (un)related *S. pyogenes* isolates provided by the Netherlands Reference Laboratory for Bacterial Meningitis (NRLBM). Data set 2 consisted of 23 possible outbreak-related clinical *S. pyogenes* isolates collected between December 2021 and October 2022 at Leiden University Medical Center (LUMC), the Netherlands. The included *S. pyogenes* isolates were identified for clinical diagnostic purposes using matrix-assisted laser desorption/ionization- time of flight (MALDI-TOF) mass spectrometry (25) and Streptex agglutination tests, after overnight incubation of the clinical specimen on blood agar (BioMérieux SA). To examine whether the isolates were outbreak-related, the samples were analyzed by *emm*-typing, MLST, and cgMLST. For this, DNA was extracted and purified from cultured *S. pyogenes* strains using the QIAamp DNA Blood Mini Kit (QIAGEN Benelux BV). From the purified DNA extracts, NGS libraries were prepared with the Illumina DNA Prep (Illumina, San Diego, CA, USA), which was bidirectionally sequenced using the Illumina MiniSeq platform with 2 × 150 bp chemistry (Illumina, San Diego, CA, USA). FastQ-formatted sequences were extracted from the MiniSeq machine and processed further using Ridom SeqSphere+ and PubMLST (https//pubmlst.org) for data analysis.

The third data set (Data set 3) included WGS data of 117 samples provided by the NRLBM. Medical microbiology laboratories from the Netherlands send their *S. pyogenes* isolates cultured from a normally sterile site to the NRLBM for national surveillance purposes (*emm*-typing) (26). No detailed epidemiological information was available for these isolates. The included samples were collected between 2009 and 2022. FastA-files were provided by the NRLBM and processed at the LUMC for data analysis. All samples included in this study have been anonymized and are not traceable to individuals, omitting the need for approval by an ethical committee.

Distances between target genes (allelic differences) were presented using minimum-spanning trees. In the scheme, the minimum depth of coverage was 30 and minimum base Q values was set at 120. The number of contigs was <1,000 and contig N50 was >15,000. The assembler was Velvet 1.1.04. Genome size was 1.83 kb. The lowest percentage of included cgMLST targets was 99%. All missing loci were removed as a whole from the analysis. Standard MLST scheme (seven target genes) (10) and *emm*-typing (27, 28) were determined to compare with cgMLST results.

RESULTS

Conventional typing methods and cgMLST scheme

The core genome for the cgMLST scheme was defined as a standard of 1,095 target genes using the default setting of the cgMLST target definer in combination with one seed genome and 66 different S. *pyogenes* genomes (Table S1). Using conventional typing methods, 34 different sequence-types and 20 *emm*-types were identified in Data sets 1–3.

Performance of cgMLST on publicly available WGS data

The proposed scheme was applied to publicly available (outbreak) WGS data from previous publications. When applying the scheme to isolates from previously defined outbreaks, the maximum allelic difference between isolates within one outbreak was two. The allelic differences between isolates from different outbreaks ranged between 6 and 31 for isolates with the same *emm*-type and between 958 and 982 for outbreaks with

a different *emm*-type (Fig. 1). Based on these results, we propose a cluster threshold of ≤5 allelic differences for the current cgMLST scheme.

Performance of cgMLST with clinical isolates

The developed cgMLST scheme was challenged with different strains from 16 different *emm*-types. Based on the previous analyses using outbreak data, the cluster threshold was defined at ≤5. The percentages of included cgMLST genes were >99%. The 23 isolates from the suspected outbreak (Data set 2) are shown in Fig. 2. Of those, seven isolates had an allelic difference of ≤2, thereby identifying a potential cluster. The cgMLST scheme was able to discriminate between isolates that belonged to the same ST or *emm*-types. From the NLRBM isolates (Data set 3), allelic differences between isolates belonging to the same *emm*-type ranged between 0 and 118. The number of isolates that are within potential clusters was 60, identifying 15 potential outbreak clusters.

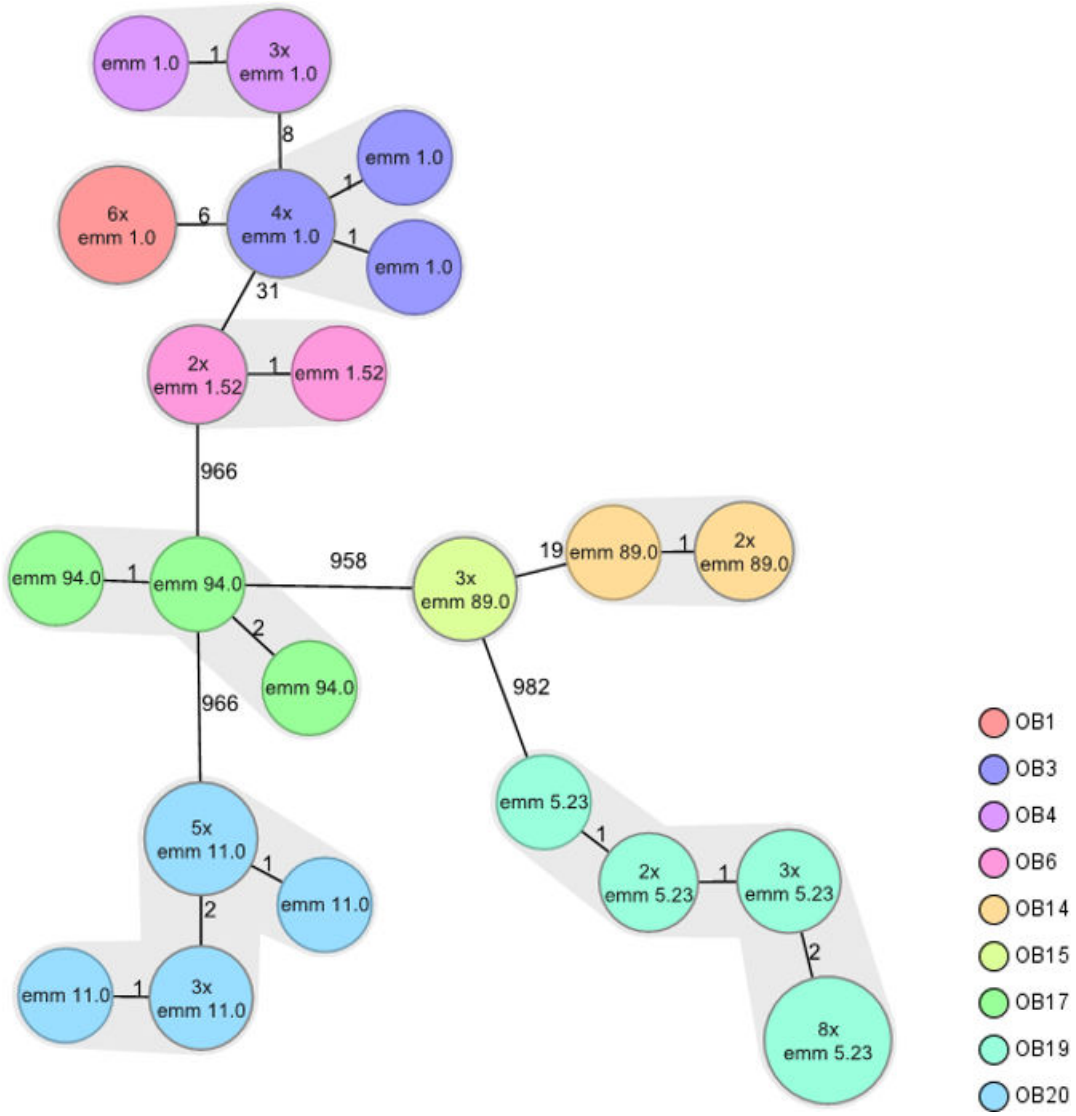

FIG 1 Minimum-spanning tree constructed using cgMLST profiles of 52 *S. pyogenes* isolates that were part of nine confirmed *S. pyogenes* outbreaks in England, UK (2010–2015). (i) OB = outbreak. (ii) The numbers at the connecting lines indicate the number of allelic differences between isolates. The numbers within the circles indicate the *emm*-type and the number of isolates with the same genetic profile (without allelic differences). The cluster threshold was set at <5 allelic differences.

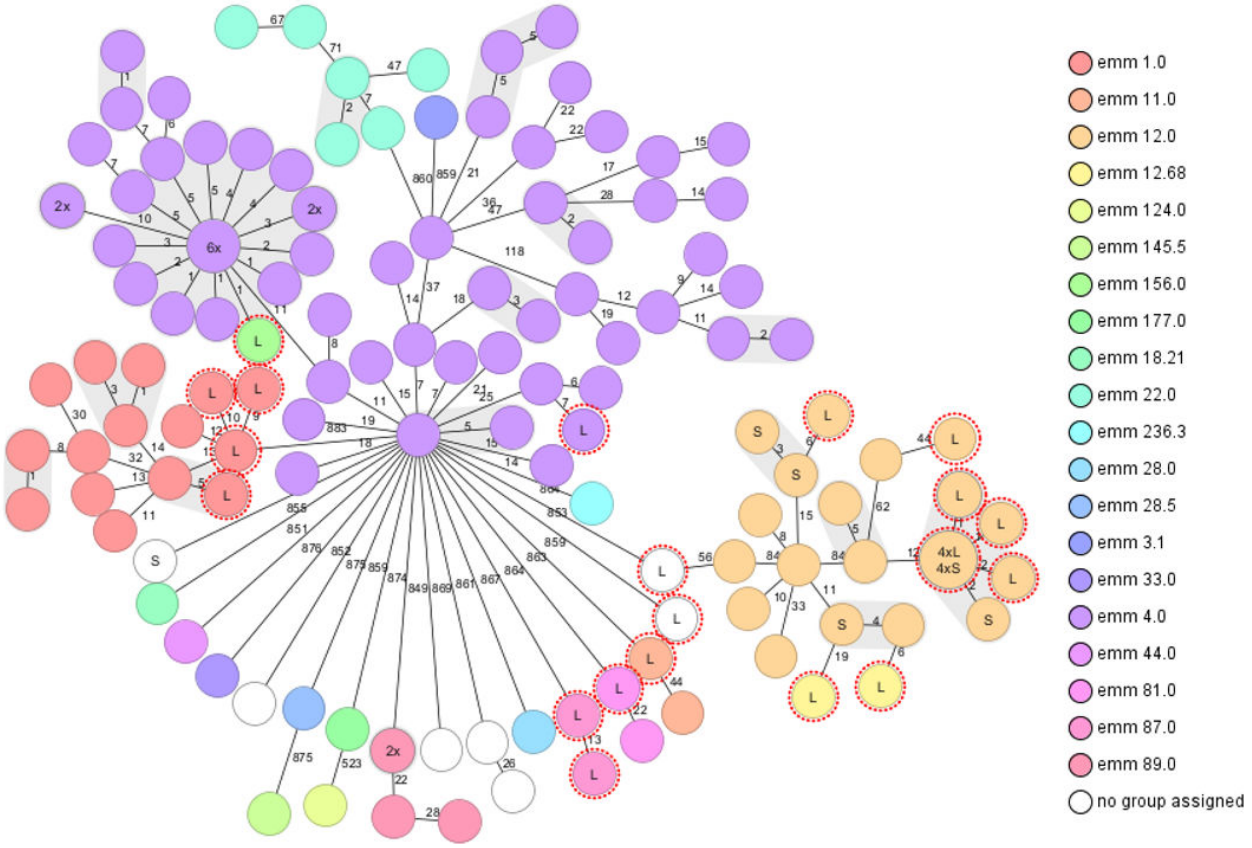

FIG 2 Minimum-spanning tree constructed using cgMLST profiles of 140 clinical isolates from the Netherlands. (i) OB = outbreak. (ii) Of the 140 profiles, 23 were derived from clinical isolates of patients and health-care workers from the Leiden University Medical Center, indicated with L and a red dotted line (Data set 2). The remaining 117 profiles were obtained from the NRLBM (Data set 3) and included nine profiles (S) of patient isolates with a geographic link to the Leiden region. The cluster threshold was set at <5; the gray shading indicates a cluster.

DISCUSSION

The proposed cgMLST scheme shows a higher discriminatory ability when compared to conventional typing methods. When using the cgMLST scheme on WGS data from previous publications containing *S. pyogenes* isolates from defined outbreaks, the maximum allelic difference between isolates from an outbreak was two, and the minimum allelic distance between isolates from different outbreaks was six. The proposed cluster definition for the cgMLST scheme is, therefore, ≤5 allelic differences. When the cgMLST scheme was applied to 140 clinical isolates from the Netherlands, 15 clusters could be identified based on the proposed cluster threshold. However, since no epidemiological data were available, it was not possible to define these clusters as outbreaks. Consistent across all isolates analyzed, cgMLST was able to identify allelic differences between isolates that had identical ST or *emm*-types, showing a higher discriminatory ability.

This study used a selection of WGS data (22) from isolates published in a previous study by Friães et al. (19), in which a different wg/cgMLST scheme for *S. pyogenes* was proposed. This group reports that their cgMLST scheme results were comparable to SNP-based methods, with a higher discriminatory power when compared to conventional typing methods. Our proposed cgMLST algorithm technique was developed with the commercially available Ridom SeqSphere+ software, whereas Friães et al. used chewBBACA (BSR-Based Allele Calling Algorithm) software. ChewBBACA software is freely available, but the software is command line-based, making it more complex to use compared to SeqSphere+ software. Friães et al. report a maximum link distance of

six allelic differences within *S. pyogenes* outbreaks similar to our findings (Data set 1, maximum link distance two). Outbreaks containing isolates that were excluded based on wg/cgMLST results in the study by Friães et al. were not analyzed in our study.

CgMLST schemes for several other bacterial species have been established and evaluated. They were developed in recent years and can be used for the evaluation of strain differentiation and help to identify outbreaks. For example, in a recent study, cgMLST was reliably used for high-resolution typing of outbreaks with *Brucella* strains (29). For many bacterial species, commercially available software such as Ridom SeqSphere+ or BioNumerics published fixed cgMLST bacterial gene schemes for the standardization of WGS-based bacterial genotyping (30). For some of these species (including *S. pyogenes*), the maximum allelic distance or cluster type threshold within an outbreak is unknown. Previous studies have observed that allelic distances within outbreaks were less than 10 alleles for species such as *Klebsiella pneumoniae, Listeria monocytogenes, Mycobacterium tuberculosis,* and *Legionella pneumophila* (31–34). For *Staphylococcus aureus*, the cluster threshold distance is estimated to be 10–25 (35, 36). In the outbreak data used in our study (Fig. 1), the maximum distance between target genes within an outbreak was <6. We, therefore, propose a threshold of ≤5 to identify *S. pyogenes* isolates that are likely to belong to the same outbreak. For *Clostridioides difficile*, a similar low threshold (≤6) is suggested to define the maximum distance between epidemiologically linked clusters (37).

The proposed cgMLST scheme for *S. pyogenes* has several advantages. It has the potential to enable widespread improvement of genomic surveillance and outbreak detection of *S. pyogenes*. The standardized analysis workflow of this cgMLST scheme is performed using easy-to-use software (Ridom SeqSphere+) and is made available for users of this software (https://www.cgmlst.org/ncs/schema/30585223/). Especially at a time when there is an increase in outbreaks of invasive GAS, and NGS becomes more widely available in many laboratories, a standardized cgMLST typing scheme with high discriminatory power is needed whereby the results can be easily exchanged between (inter)national laboratories. Of note, Ridom SeqSphere + is not free of charge and although it can be used for MLST and genomic antimicrobial resistance determination, *emm*-typing is currently not possible with this software. Future updates to include *emm*-typing would be a meaningful update.

The current study also has some limitations. Most importantly, detailed epidemiological information was lacking for most isolates (Data sets 2 and 3), which complicates reliably identifying outbreaks. Second, the cgMLST technique by design only detects allele variations that have been previously defined as the core genome (18). Techniques such as wgMLST or SNP include more coverage of the complete genome (including the accessory genome) and potentially achieve a higher resolution, which further increases discriminatory power. For example, in a recent article from the Netherlands [currently available in preprint (26)], whole-genome sequence analysis of *emm*4 isolates identified a novel S. *pyogenes* lineage, accounting for 85% of the *emm*4 invasive S. *pyogenes* cases in 2022. However, for reproducibility and to compare results between laboratories, it is desirable to use a standardized cgMLST method with easy-to-use software. Due to the lack of available studies with epidemiological data, the proposed cluster threshold of ≤5 allelic differences is based on a single outbreak study only, and more data are necessary to support this.

In conclusion, the robustness of this cgMLST scheme that is made publicly available via Ridom, in combination with easy-to-use software, enables widespread improvement of genomic surveillance of *S. pyogenes*, allowing increased detection of transmission and highlighting opportunities for intervention.

ACKNOWLEDGMENTS

We would like to thank Boas van der Putten and Nina van Sorge from the Netherlands Reference Laboratory for Bacterial Meningitis (NRLBM) for sharing their data and providing feedback on the manuscript.

M.M.A.T., M.E.M.K., and S.A.B. designed the research and collected and analyzed the data. M.E.M.K. and S.A.B. developed the cgMLST scheme. M.M.A.T. and S.A.B. wrote the manuscript. M.E.M.K., I.V.H., J.v.P., E.C.J.C., and E.W. revised the paper for important intellectual content.

AUTHOR AFFILIATION

[1]Department of Medical Microbiology, Leiden University Medical Center, Leiden, the Netherlands

AUTHOR ORCIDs

Myrthe M. A. Toorop http://orcid.org/0000-0002-3348-2419
Els Wessels http://orcid.org/0000-0002-6707-2311
Stefan A. Boers http://orcid.org/0000-0002-1560-6799

AUTHOR CONTRIBUTIONS

Myrthe M. A. Toorop, Conceptualization, Data curation, Formal analysis, Methodology, Writing – original draft | Margriet E. M. Kraakman, Formal analysis, Methodology, Software, Writing – review and editing | Irene V. Hoogendijk, Writing – review and editing | Joffrey van Prehn, Resources, Writing – review and editing | Eric C. J. Claas, Software, Writing – review and editing | Els Wessels, Software, Writing – review and editing | Stefan A. Boers, Conceptualization, Data curation, Methodology, Writing – review and editing

DATA AVAILABILITY

Raw sequence reads for Data set 1 can be found in the Sequence Read Archive (SRA) under project number PRJEB49967 (38). Raw sequence reads for Data set 2 have been deposited in the SRA under project accession number PRJNA966900 (39). Raw sequence data for Data set 3 have been deposited in the SRA under project accession number PRJNA967239 (40).

ETHICAL STATEMENT

All samples included in this study have been anonymized and are not traceable to individuals, omitting the need for approval by an ethical committee.

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Table S1 (JCM00558-23-S0001.docx). List of Emm-types and STs from the 66 penetration query genomes

REFERENCES

1. Lancefield RC. 1933. A serological differentiation of human and other groups of hemolytic streptococci. J Exp Med 57:571–595. https://doi.org/10.1084/jem.57.4.571
2. Björck V, Påhlman LI, Bodelsson M, Petersson A-C, Kander T. 2020. Morbidity and mortality in critically ill patients with invasive group A *Streptococcus* infection: an observational study. Crit Care 24:302. https://doi.org/10.1186/s13054-020-03008-z
3. Deutscher M, Schillie S, Gould C, Baumbach J, Mueller M, Avery C, Van Beneden CA. 2011. Investigation of a group A streptococcal outbreak among residents of a long-term acute care hospital. Clin Infect Dis 52:988–994. https://doi.org/10.1093/cid/cir084
4. Nanduri SA, Metcalf BJ, Arwady MA, Edens C, Lavin MA, Morgan J, Clegg W, Beron A, Albertson JP, Link-Gelles R, Ogundimu A, Gold J, Jackson D, Chochua S, Stone N, Van Beneden C, Fleming-Dutra K, Beall B. 2019. Prolonged and large outbreak of invasive group A *Streptococcus* disease within a nursing home: repeated intrafacility transmission of a single strain. Clin Microbiol Infect 25:248. https://doi.org/10.1016/j.cmi.2018.04.034
5. Ahmed SS, Diebold KE, Brandvold JM, Ewaidah SS, Black S, Ogundimu A, Li Z, Stone ND, Van Beneden CA. 2018. The role of wound care in 2 group A streptococcal outbreaks in a Chicago skilled nursing facility, 2015–2016. Open Forum Infect Dis 5:fy145. https://doi.org/10.1093/ofid/ofy145
6. Rijksinstituut voor Volksgezondheid en Milieu (RIVM). 2022. Meldingen van invasieve GAS-infecties in Nederlan. Available from: https://

wwwrivmnl/groep-a-streptokokkeninfecties-gas/meldingen-van-invasieve-gas-infecties-in-nederland. Accessed 19 Dec 2022.
7. de Gier B, Marchal N, de Beer-Schuurman I, Te Wierik M, Hooiveld M, ISIS-AR Study Group, GAS Study group, de Melker HE, van Sorge NM, Members of the GAS study group, Members of the ISIS-AR study group. 2023. Increase in invasive group A streptococcal (*Streptococcus pyogenes*) infections (iGAS) in young children in the Netherlands, 2022. Euro Surveill 28:2200941. https://doi.org/10.2807/1560-7917.ES.2023.28.1.2200941
8. World Health Organization (WHO). Increase in invasive group A streptococcal infections among children in Europe, including fatalities. Available from: https://wwwwhoint/europe/news/item/12-12-2022-increase-in-invasive-group-a-streptococcal-infections-among-children-in-europe--including-fatalities. Accessed December 19, 2022
9. Tewodros W, Kronvall G. 2005. M protein gene (emm type) analysis of group A beta-hemolytic streptococci from Ethiopia reveals unique patterns. J Clin Microbiol 43:4369–4376. https://doi.org/10.1128/JCM.43.9.4369-4376.2005
10. Enright MC, Spratt BG, Kalia A, Cross JH, Bessen DE. 2001. Multilocus sequence typing of *Streptococcus pyogenes* and the relationships between emm type and clone. Infect Immun 69:2416–2427. https://doi.org/10.1128/IAI.69.4.2416-2427.2001
11. Pubmlst scheme information *S. Pyogenes*. Available from: https://pubmlstorg/bigsdb?db=pubmlst_spyogenes_seqdef&page=schemeInfo&scheme_id=1. Accessed April 1, 2023
12. Turner CE, Bedford L, Brown NM, Judge K, Török ME, Parkhill J, Peacock SJ. 2017. Community outbreaks of group A *Streptococcus* revealed by genome sequencing. Sci Rep 7:8554. https://doi.org/10.1038/s41598-017-08914-x
13. Maiden MCJ, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND. 2013. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol 11:728–736. https://doi.org/10.1038/nrmicro3093
14. Uelze L, Grützke J, Borowiak M, Hammerl JA, Juraschek K, Deneke C, Tausch SH, Malorny B. 2020. Typing methods based on whole genome sequencing data. One Health Outlook 2:3. https://doi.org/10.1186/s42522-020-0010-1
15. Köser CU, Holden MTG, Ellington MJ, Cartwright EJP, Brown NM, Ogilvy-Stuart AL, Hsu LY, Chewapreecha C, Croucher NJ, Harris SR, Sanders M, Enright MC, Dougan G, Bentley SD, Parkhill J, Fraser LJ, Betley JR, Schulz-Trieglaff OB, Smith GP, Peacock SJ. 2012. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. N Engl J Med 366:2267–2275. https://doi.org/10.1056/NEJMoa1109910
16. Meehan CJ, Goig GA, Kohl TA, Verboven L, Dippenaar A, Ezewudo M, Farhat MR, Guthrie JL, Laukens K, Miotto P, Ofori-Anyinam B, Dreyer V, Supply P, Suresh A, Utpatel C, van Soolingen D, Zhou Y, Ashton PM, Brites D, Cabibbe AM, de Jong BC, de Vos M, Menardo F, Gagneux S, Gao Q, Heupink TH, Liu Q, Loiseau C, Rigouts L, Rodwell TC, Tagliani E, Walker TM, Warren RM, Zhao Y, Zignol M, Schito M, Gardy J, Cirillo DM, Niemann S, Comas I, Van Rie A. 2019. Whole genome sequencing of *Mycobacterium tuberculosis*: current standards and open issues. Nat Rev Microbiol 17:533–545. https://doi.org/10.1038/s41579-019-0214-5
17. Ghanem M, Wang L, Zhang Y, Edwards S, Lu A, Ley D, El-Gazzar M. 2018. Core genome multilocus sequence typing: a standardized approach for molecular typing of *Mycoplasma gallisepticum*. J Clin Microbiol 56:e01145-17. https://doi.org/10.1128/JCM.01145-17
18. Sheppard SK, Jolley KA, Maiden MCJ. 2012. A gene-by-gene approach to bacterial population genomics: whole genome MLST of Campylobacter. Genes (Basel) 3:261–277. https://doi.org/10.3390/genes3020261
19. Friães A, Mamede R, Ferreira M, Melo-Cristino J, Ramirez M. 2022. Annotated whole-genome multilocus sequence typing schema for scalable high-resolution typing of *Streptococcus pyogenes*. J Clin Microbiol 60:e0031522. https://doi.org/10.1128/jcm.00315-22
20. Jünemann S, Sedlazeck FJ, Prior K, Albersmeier A, John U, Kalinowski J, Mellmann A, Goesmann A, von Haeseler A, Stoye J, Harmsen D. 2013. Updating benchtop sequencing performance comparison. Nat Biotechnol 31:294–296. https://doi.org/10.1038/nbt.2522
21. NCBI RefSeq database. Available from: https://wwwncbinlmnihgov/refseq. Accessed December 19, 2022
22. Coelho JM, Kapatai G, Jironkin A, Al-Shahib A, Daniel R, Dhami C, Laranjeira AM, Chambers T, Phillips S, Tewolde R, Underwood A, Chalker VJ. 2019. Genomic sequence investigation *Streptococcus pyogenes* clusters in England (2010-2015). Clin Microbiol Infect 25:96–101. https://doi.org/10.1016/j.cmi.2018.04.011
23. Friães A, Mamede R, Ferreira M, Melo-Cristino J, Ramirez M, Diekema DJ. 2022. Supplemental material of "an annotated whole-genome multilocus sequence typing schema for scalable high resolution typing of *Streptococcus pyogenes*" Zenodo.
24. Pubmlst public databases for molecular typing and microbial genome diversity. 2023. *Streptococcus pyogenes* typing database. Available from: https://pubmlst.org/bigsdb?db=pubmlst_spyogenes_seqdef. Accessed 16 Feb 2023.
25. Wang J, Zhou N, Xu B, Hao H, Kang L, Zheng Y, Jiang Y, Jiang H, VertesA. 2012. Identification and cluster analysis of *Streptococcus pyogenes* by MALDI-TOF mass spectrometry. PLoS One 7:e47152. https://doi.org/10.1371/journal.pone.0047152
26. van der Putten BCL, Bril-Keijzers WCM, Rumke LW, Vestjens SMT, Koster LAM, Willemsen M, van Houten MA, Rots NY, Vlaminckx BJM, de Gier B, van Sorge NM. 2023. Novel emm4 lineage associated with an upsurge in invasive group A streptococcal disease in the Netherlands, 2022. bioRxiv. https://doi.org/10.1101/2022.12.31.522331
27. McGregor KF, Spratt BG, Kalia A, Bennett A, Bilek N, Beall B, Bessen DE. 2004. Multilocus sequence typing of *Streptococcus pyogenes* representing most known emm types and distinctions among subpopulation genetic structures. J Bacteriol 186:4285–4294. https://doi.org/10.1128/JB.186.13.4285-4294.2004
28. McMillan DJ, Drèze P-A, Vu T, Bessen DE, Guglielmini J, Steer AC, Carapetis JR, Van Melderen L, Sriprakash KS, Smeesters PR. 2013. Updated model of group A *Streptococcus* M proteins based on a comprehensive worldwide study. Clin Microbiol Infect 19:E222–E229. https://doi.org/10.1111/1469-0691.12134
29. Abdel-Glil MY, Thomas P, Brandt C, Melzer F, Subbaiyan A, Chaudhuri P, Harmsen D, Jolley KA, Janowicz A, Garofolo G, Neubauer H, Pletz MW. 2022. Core genome multilocus sequence typing scheme for improved characterization and epidemiological surveillance of pathogenic brucella. J Clin Microbiol 60:e0031122. https://doi.org/10.1128/jcm.00311-22
30. cgMLST.org Nomenclature Server. Available from: https://-wwwcgmlstorg/ncs. Accessed December 20, 2022
31. Zhou H, Liu W, Qin T, Liu C, Ren H. 2017. Defining and evaluating a core genome multilocus sequence typing scheme for whole-genome sequence-based typing of *Klebsiella pneumoniae*. Front Microbiol 8:371. https://doi.org/10.3389/fmicb.2017.00371
32. Ruppitsch W, Pietzka A, Prior K, Bletz S, Fernandez HL, Allerberger F, Harmsen D, Mellmann A. 2015. Defining and evaluating a core genome multilocus sequence typing scheme for whole-genome sequence-based typing of *Listeria monocytogenes*. J Clin Microbiol 53:2869–2876. https://doi.org/10.1128/JCM.01193-15
33. Kohl TA, Diel R, Harmsen D, Rothgänger J, Walter KM, Merker M, Weniger T, Niemann S. 2014. Whole-genome-based *Mycobacterium tuberculosis* surveillance: a standardized, portable, and expandable approach. J Clin Microbiol 52:2479–2486. https://doi.org/10.1128/JCM.00567-14
34. Moran-Gilad J, Prior K, Yakunin E, Harrison TG, Underwood A, Lazarovitch T, Valinsky L, Luck C, Krux F, Agmon V, Grotto I, Harmsen D. 2015. Design and application of a core genome multilocus sequence typing scheme for investigation of Legionnaires' disease incidents. Euro Surveill 20:21186. https://doi.org/10.2807/1560-7917.es2015.20.28.21186
35. *Staphylococcus aureus* cgMLST. Available from: https://wwwcgmlstorg/ncs/schema/141106. Accessed December 20, 2022
36. Leopold SR, Goering RV, Witten A, Harmsen D, Mellmann A. 2014. Bacterial whole-genome sequencing revisited: portable, scalable, and standardized analysis for typing and detection of virulence and antibiotic resistance genes. J Clin Microbiol 52:2365–2370. https://doi.org/10.1128/JCM.00262-14
37. Bletz S, Janezic S, Harmsen D, Rupnik M, Mellmann A. 2018. Defining and evaluating a core genome multilocus sequence typing scheme for genome-wide typing of *Clostridium difficile*. J Clin Microbiol 56:e01987-17. https://doi.org/10.1128/JCM.01987-17
38. An annotated whole-genome multilocus sequence typing schema for *Streptococcus pyogenes* (dataset 1). Available from: https://www.ncbi.nlm.nih.gov/bioproject/PRJEB49967. Accessed July 12, 2023

39. Development of a novel cgMLST scheme for *Streptococcus pyogenes* (dataset 2). Available from: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA966900. Accessed May 4, 2023
40. Development of a novel cgMLST scheme for *Streptococcus pyogenes* (dataset 3). Available from: https://www.ncbi.nlm.nih.gov/bioproject/967239. Accessed May 4, 2023