A Core Genome Multilocus Sequence Typing Scheme for *Pseudomonas aeruginosa*

Romário Oliveira de Sales\(^1\), Leticia Busato Migliorini\(^1\), Renato Puga\(^1\), Bela Kocsis\(^2\) and Patricia Severino\(^*\)

\(^1\) Hospital Israelita Albert Einstein, Albert Einstein Research and Education Institute, São Paulo, Brazil, \(^2\) Institute of Medical Microbiology, Semmelweis University, Budapest, Hungary

*Pseudomonas aeruginosa* is a ubiquitous microorganism and an important opportunistic pathogen responsible for a broad spectrum of infections mainly in immunosuppressed and critically ill patients. Molecular investigations traditionally rely on pulsed field gel electrophoresis (PFGE) and multilocus sequence typing (MLST). In this work we propose a core genome multilocus sequence typing (cgMLST) scheme for *P. aeruginosa*, a methodology that combines traditional MLST principles with whole genome sequencing data. All publicly available complete *P. aeruginosa* genomes, representing the diversity of this species, were used to establish a cgMLST scheme targeting 2,653 genes. The scheme was then tested using genomes available at contig, chromosome and scaffold levels. The proposed cgMLST scheme for *P. aeruginosa* typed over 99% (2,314/2,325) of the genomes available for this study considering at least 95% of the cgMLST target genes present. The absence of a certain number gene targets at the threshold considered for both the creation and validation steps due to low genome sequence quality is possibly the main reason for this result. The cgMLST scheme was compared with previously published whole genome single nucleotide polymorphism analysis for the characterization of the population structure of the epidemic clone ST235 and results were highly similar. In order to evaluate the typing resolution of the proposed scheme, collections of isolates belonging to two important STs associated with cystic fibrosis, ST146 and ST274, were typed using this scheme, and ST235 isolates associated with an outbreak were evaluated. Besides confirming the relatedness of all the isolates, earlier determined by MLST, the higher resolution of cgMLST denotes that it may be suitable for surveillance programs, overcoming possible shortcomings of classical MLST. The proposed scheme is publicly available at: https://github.com/BioinformaticsHIAEMolecularMicrobiology/cgMLST-Pseudomonas-aeruginosa.

**Keywords:** *Pseudomonas aeruginosa*, core genome multilocus sequence typing, sequence type, genome sequencing analysis, mlst

**INTRODUCTION**

*Pseudomonas aeruginosa* is an opportunistic human pathogen responsible for nosocomial infections worldwide, such as ventilator-associated pneumonia and burn wound infections, mainly in immunosuppressed patients (De Bentzmann and Plesiat, 2011; Gellatly and Hancock, 2013; Cabot et al., 2016; Sousa et al., 2018; Yin et al., 2018). Moreover, it is highly involved in chronic...
pneumonia in patients with cystic fibrosis (CF) (Lyczak et al., 2000; Breidenstein et al., 2011; Ballarini et al., 2012; Lopez-Causa et al., 2017). Intrinsic and acquired antibiotic resistance limits treatment of these infections (Gellatly and Hancock, 2013; Gomila et al., 2013; Arzanlou et al., 2017). Also, due to its ability to form biofilm, *P. aeruginosa* can remain viable for a long time on hospital surfaces and medical fomites, contributing to its spread in clinical settings and possible outbreaks (Lister et al., 2009; Veessenmeyer et al., 2009; Breathnach et al., 2012; Gill et al., 2016; Alipour et al., 2017; Moradali et al., 2017). As such, *P. aeruginosa*, especially those belonging to high-risk clonal complexes which are often resistant, are a threat to public health worldwide (Woodford et al., 2006; Mulet et al., 2013; Treepong et al., 2018).

High-resolution sequence typing is important to evaluate epidemiological links and to characterize possible transmission sources during outbreaks. Currently, two major subtyping techniques are used to identify the clonal relationship between *P. aeruginosa* isolates: Pulsed-field Gel Electrophoresis (PFGE) and multilocus sequence typing (MLST) (Curran et al., 2004; Johnson et al., 2007; Kidd et al., 2011). PFGE is one of the most useful DNA fingerprinting techniques and a method widely used in hospitals to identify *P. aeruginosa* outbreaks. However, PFGE is a time-consuming and laborious method, and often not easily reproducible between clinical laboratories (Tenover et al., 1995; Fothergill et al., 2010; Lila et al., 2018; Martak et al., 2019). In contrast, MLST, developed in 2004 for *P. aeruginosa*, is based on the sequencing and evaluation of the allelic variation of seven housekeeping genes, generating the so-called sequence types (STs) to characterize isolates (Curran et al., 2004). However, the discriminatory power of MLST does not always allow enough resolution during outbreaks. Therefore, both PFGE and MLST methods are often simultaneously required (Li et al., 2009; Kidd et al., 2011; Sabat et al., 2013).

Recently, methods based on whole genome sequencing (WGS) have become viable due to lower sequencing costs; enabling high discrimination between isolates in epidemiological investigation. Nevertheless, suitable bioinformatics tools are necessary to handle and interpret sequence data (Tang et al., 2017; Bogaerts et al., 2019). The aim of Core Genome MLST (cgMLST) is to extend the MLST concept to a larger number of genes of the core genome. It relies on WGS but allows for the examination of a fixed number of genome *loci* and the creation of a systematic allele numbering system. cgMLST schemes may be developed and locally implemented using commercial softwares such as BioNumerics (Applied Maths, Sint-Martens-Latem, Belgium), SeqSphere+ (Ridom, Münster, Germany), and BIGSdb3 (Schürch et al., 2018). The design may lead to maximum resolution for both epidemiological and surveillance analyses (Mellmann et al., 2011; de Been et al., 2015; Zhou et al., 2017; Neumann et al., 2019). The cgMLST has already shown good results for bacterial pathogens such as *Mycobacterium tuberculosis*, *Acinetobacter baumannii*, *Klebsiella pneumoniae*, and *Enterococcus faecalis* (Kohl et al., 2014; Higgins et al., 2017; Zhou et al., 2017; Neumann et al., 2019).

Publicly available cgMLST schemes for *P. aeruginosa* do not exist yet. Two ad hoc schemes have been constructed using SeqSphere+ (Ridom, Münster, Germany) but they have not been validated or used outside the context proposed by the authors (Mellmann et al., 2016; Royer et al., 2020). In this study, we created a cgMLST scheme for *P. aeruginosa* using the open source ChewBBACA algorithm (Silva et al., 2018). The scheme was created taking in consideration the core genome of all available *P. aeruginosa* genome sequences at the time of this publication and it is freely available2. We evaluated this scheme for its ability to discriminate isolates within the same STs (ST146, ST235, and ST274), both in an outbreak setting and using unrelated isolates. Taken together our results indicate that the scheme may be adequate for epidemiology and surveillance approaches and outbreak investigations and we aim to foment discussions and possibly help in establishing a cgMLST consensus for *P. aeruginosa*.

**MATERIALS AND METHODS**

**Core Genome Multilocus Sequence Typing Scheme**

*A. P. aeruginosa* cgMLST scheme was set up with the Comprehensive and Highly Efficient Workflow for a Blast Score Ratio Based Allele Calling Algorithm (ChewBBACA) (Silva et al., 2018). Completed genome sequences (FASTA files) of *P. aeruginosa* publicly available at the NCBI Reference Sequence database (RefSeq)2 in September 2018 were used (Supplementary Table S1). A total of 141 complete sequences were annotated with Prodigal v2.6.3 (Hyatt et al., 2010) in the first step of the algorithm (CreateSchema). The genome sequence of *P. aeruginosa* PAO1 (RefSeq assembly accession: GCF_000060675.1) was used only as the reference genome to predict the wgMLST *loci* and then removed from cgMLST analysis. Briefly, in the first step, the algorithm defined coding sequences (CDs) for each genome, compared them in a pairwise way and generated a single FASTA file containing the selected CDs. In the next step, the allele-calling algorithm (AlleleCall), identified and excluded possible paralogous *loci*. The remaining list of *loci*, now called the wgMLST, was then used to define the cgMLST scheme. We chose to select candidate *loci* for the cgMLST scheme present in 100% of the available complete genomes (141 genomes). This choice minimizes the exclusion of core genome *loci* due to misassembling in incomplete unfinished sequences. In the validation step, we lowered the stringency and kept candidate *loci* common to 99% of the isolates. Since the sequences used during this step were unfinished genomes, and due to the genetic plasticity of *P. aeruginosa*, we believe this new cut-off was necessary and sufficient to make the typing scheme suitable for most isolates of various origins. Publicly available unfinished genomes of *P. aeruginosa* (contigs, scaffold, or chromosome level genome sequences) were considered

---

1https://pubmlst.org/software/database/bigsdb/

2https://github.com/BioinformaticsHIAEMolecularMicrobiology/cgMLST-Pseudomonas-aeruginosa

3https://www.ncbi.nlm.nih.gov/assembly
for the validation of the cgMLST scheme (see text footnote 3) (Supplementary Table S1). In order to ensure sufficient quality of the sequence data, all genomes for which no MLST could be assigned using the sanger-pathogens/mlst_check script1 2 (J. Page et al., 2016) were filtered out. Additionally, unfinished genomes with ≥200 contigs were not included. New STs were assigned when no correspondent STs were found in the pubMLST database for the purpose of this study. All the steps followed for the creation of the cgMLST scheme as well as input and output files are available at GitHub https://github.com/BioinformaticsHIAEMolecularMicrobiology/cgMLST-Pseudomonas-aeruginosa.

**Graphic Representation of cgMLST Results: Minimum-Spanning Trees**

Using the allelic profiles obtained by the cgMLST scheme for each isolate, minimum-spanning trees were constructed using the software GrapeTree (version 1.5.0) with parameters implemented in MSTree v2 ignoring missing values for the entire collection. Furthermore, similarity trees were calculated using the neighbor-joining algorithm (StandardNJ) with default parameters implemented in GrapeTree (version 1.5.0) (Zhou et al., 2018). Trees were visualized using Interactive Tree Of Life (iTOL, version 4.2.3) (Letunic and Bork, 2016).

**cgMLST Analysis Compared With Core Genome SNP-Based Analysis for ST235**

All publicly available genomes of *P. aeruginosa* ST235 were selected (128 isolates, Supplementary Table S1). These were collected over a 28-year period and came from different locations: Africa (n = 2), Asia (n = 11), Europe (n = 36), North America (n = 48), South America (n = 16), Oceania (n = 3), and unknown origin (n = 12). An unrooted similarity tree based on cgMLST targets was built using the neighbor-joining algorithm (StandardNJ) with default parameters implemented in GrapeTree (version 1.5.0) (Zhou et al., 2018). The resulting similarity tree from the cgMLST-based approach was visualized by iTOL (version 4.2.3) (Letunic and Bork, 2016). For the core genome SNP-based analysis, the core genome was aligned using Parsnp, part of the Harvest software package (Treangen et al., 2014), using NCGM2.S1 (RefSeq assembly accession: GCF_000284555.1) as the reference genome. All sequences near the end of contigs (near possible gaps) were excluded for this analysis in order to avoid using false SNPs for similarity reconstruction (default settings in Parsnp). This tree was also visualized with iTOL. The concordance of the two methods, cgMLST and core genome SNP-based analysis, was discussed based on similarities and differences in clustering (similarity tree).

**Typing Resolution of the Proposed cgMLST-Based Analysis**

Two STs were selected for comparing the typing resolution between MLST and our cgMLST scheme: ST146 and ST274.

Forty-eight isolates belonging to ST146, 40 of which isolated from a single patient in United Kingdom in 2009 (Williams et al., 2015) (NCBI Bioproject PRJEB6642) and 8 unrelated isolates reported in the United Kingdom (n = 5) and Canada (n = 3) were selected. For isolates belonging to ST274, 197 isolates, most of which (n = 167) collected from 12 sputum specimens obtained over a 1-year period from a single patient in Canada (Diaz Caballero et al., 2015) (NCBI Bioproject PRJNA282164), and 30 unrelated ST274 isolates coming from the United States (n = 14), Denmark (n = 4), France (n = 2), Spain (n = 1), Germany (n = 2), Brazil (n = 1), China (n = 1), Canada (n = 1), and from unknown origin (n = 4), were included.

We also used 4 isolates of ST235 *P. aeruginosa* (NCBI Bioproject PRJEB32170) involved in a local outbreak of extended-spectrum β-lactamase SHV2a-producing *P. aeruginosa* ST235 (Royer et al., 2020) to investigate the performance of the cgMLST scheme proposed here during an outbreak setting. In the work, the authors developed an ad hoc cgMLST scheme using SeqSphere + software (Ridom, Münster, Germany) to evaluate the ST235 isolates involved in the outbreak and we compared the result of our scheme to theirs.

A description of isolates from bioprojects NCBI PRJEB6642, PRJNA282164, and PRJEB32170 used in this section can be found in Supplementary Table S2.

**RESULTS**

**Characterization of the P. aeruginosa Population Used in This Study**

In total, 2,901 genome sequences were available at the RefSeq database. Of these genomes, 73 were not included due to the absence of MLST loci and 502 because they contained ≥200 contigs. Thus, from the 2,901 genomes originally available, 2,326 genomes were included in this study. Most isolates were collected from a single patient in United Kingdom in 2009 (Williams et al., 2015) (Figure 1A). They were distributed across 54 countries (Figure 1B), and the period of sampling extended across 80 years (Figure 1C).

A total of 518 different STs characterized the available population: 402 STs previously reported in the pubMLST database and 116 new STs determined for the purpose of this study (Supplementary Table S1). The most frequent STs were ST274 (n = 197), due to a specific project (NCBI Bioproject PRJNA282164, Supplementary Table S2) that contributed with 167 isolates, ST253 (n = 137) and ST235 (n = 128) (Figure 1D).

Among the 2,326 MLST-typed genome sequences, 141 were considered complete genomes and were considered for the initial selection of candidate targets for the cgMLST scheme. They comprised 72 distinct STs (Supplementary Table S1).

To evaluate the distribution of the isolates selected to create the cgMLST scheme within the *P. aeruginosa* population with publicly available sequenced genomes, maximum-likelihood similarity trees based on the concatenated conserved sequences used for MLST were calculated using FastTree 2 (v. 2.1.7, Jukes-Cantor_CAT model) (Price et al., 2010). The analysis of the tree for the total population of 2,326 genomes (518 distinct STs) (Figure 2A) and the distribution of the 141 genomes (72 STs)
used for the creation of the cgMLST scheme (Figure 2B) indicate that the selected isolates reflect the diversity of the *P. aeruginosa* population with genome sequences currently available in RefSeq.

**P. aeruginosa** cgMLST Scheme

For the scheme creation (Figure 3), the 141 complete genome sequences were retrieved and, using PAO1 as the reference genome (with 5,570 predicted ORFs) (Stover et al., 2000), a total of 13,588 target *loci* were annotated and generated the wgMLST data set. Following filtering steps, 282 *loci* were considered paralogous and discarded and genome quality test filtered-out additional 10,138 *loci* targets. A total of 3,168 gene targets were retained as candidates for the cgMLST scheme and they were present in 130 complete genome sequences.
FIGURE 3 | Detailed flowchart describing the development of the cgMLST scheme for *P. aeruginosa* using ChewBBACA (https://github.com/B-UMMI/chewBBACA).

(Supplementary Figure S1). For the evaluation step, 2,184 additional unfinished genome sequences were retrieved from the RefSeq database (Supplementary Table S1), thoroughly filtered following the same steps used for the scheme creation and analyzed using the candidate list of 3,168 gene targets. Following this step, a cgMLST scheme consisting of 2,653 gene targets (Supplementary Figure S2 and Supplementary Table S3) was defined, covering 47.63% of the 5,570 ORFs predicted for the reference strain PAO1. At least 90% of this cgMLST target list was found in 99% of all queried unfinished *P. aeruginosa* genomes. During the creation of the scheme, 11 complete genomes (fully assembled) had been discarded because they did not harbor the 3,168 candidate loci, and another 5 unfinished genomes were removed in the validation step. We reanalyzed those 16 isolates considering the final set of 2,653 loci and obtained similar results: with the exception of one isolate, all of the genomes contained at least 90% of the loci.

Finally, it has been proposed that, ideally, 95% of the cgMLST gene targets should be present in the great majority of isolates tested so that it is considered well-defined (de Been et al., 2015; Ruppitsch et al., 2015; Ghanem and El-Gazzar, 2018). The isolate could then be considered as “typed” by the cgMLST scheme. We verified our results and, considering the 2,325 genomes (141 fully assembled sequences and 2,184 unfinished genomes) only 11 isolates did not harbor at least 95% of the selected loci (1 fully assembled genome and 10 unfinished genome sequences). Genome sequence quality is possibly the main reason for the exclusion of the unfinished genome sequences, but for the fully assembled sequence, specific characteristics of the isolate would have to be analyzed in order to understand this result.

In summary, the proposed cgMLST scheme for *P. aeruginosa* typed over 99% (2,314/2,325) of the isolates available for this study when we considered that at least 95% of all selected loci are present in the genome.

**cgMLST Analysis Compared With Core Genome SNP-Based Analysis for ST235**

The cgMLST scheme was used to investigate the relatedness of *P. aeruginosa* isolates assigned to ST235. Using the proposed cgMLST scheme we were able to type all 128 ST235 isolates...
which remained after the genome quality assessment filters and an unrooted similarity tree was built based on the cgMLST gene targets proposed here (Figure 4A). Two major clusters are readily seen for these isolates, corroborating data previously shown by Treepong et al. (2018) when they evaluated 79 ST235 isolates using core genome analysis of SNPs. Of these 79 isolates, 62 were included in the cgMLST analysis and all clustered as originally proposed by Treepong et al. (highlighted with a green background in the tree). In fact, when we replicate the core genome typing approach proposed by Treepong et al. (using Parsnp for alignment) for the 128 isolates, the unrooted similarity tree reports that the isolates are clustered, identically, within two major groups (Figure 4B). This result indicates that the scheme proposed in this study may be used with the same purpose as core genome SNP analysis, with the advantage of using fewer genome regions.

Comparison of Typing Resolution Between MLST and cgMLST

A selection of isolates belonging to ST146 and ST274 was used to evaluate the typing resolution of the proposed cgMLST scheme. In order to determine epidemiological relationships between isolates, specific thresholds in cgMLST have been applied depending on the species. For instance, during outbreaks, isolates of *Listeria monocytogenes* (Ruppitsch et al., 2015), *Mycobacterium tuberculosis* (Kohl et al., 2014), *Legionella pneumophila* (Moran-Gilad et al., 2015) and *Klebsiella pneumoniae* have been shown to differ in fewer than 10 cgMLST alleles (Zhou et al., 2017). For *P. aeruginosa*, an ad hoc cgMLST scheme using SeqSphere+ (Ridom, Münster, Germany) considered fewer than 14 alleles for epidemiologically related isolates (Mellmann et al., 2016). The dataset for ST146 comprised 40 isolates of *P. aeruginosa* Liverpool Epidemic Strain (LES), collected from a cystic fibrosis patient during 2009 (Williams et al., 2015). Using a SNP-based analysis and a reference genome from an isolate collected in 1988 (LESB58), Williams et al. (2015) divided this set into two major lineages, A and B, with 13 and 27 isolates, respectively, separated by 79 SNPs. Our cgMLST-based analysis also proposed two major clusters, A and B, differing in 42 alleles (Figure 5A and Supplementary Figure S3). However, six isolates (dashed branches) were not grouped within the 2 major groups (Figure 5A). We do not have access to epidemiological data that could explain this separation. However, samples were collected during 2009 but were not associated with an outbreak, and, in this scenario, differences are not unexpected and could have arisen from microevolution.

Eight additional ST146 isolates from the United Kingdom and Canada were added to this analysis, and they were clearly distinct from the LES isolates (Figure 5B). The minimum-spanning tree shows that the two clusters comprising the 40 LES and the 8 additional isolates differ in at least 25 alleles (Supplementary Figure S4).

Another important strain associated with both colonization and infection in cystic fibrosis is ST274 (Fernández-Olmos et al., 2013; López-Causapé et al., 2013; Lopez-Causape et al., 2017). From a total of 197 available ST274 genome sequences, we were able to analyze all isolates using the proposed cgMLST scheme. Of these, 167 genome sequences belonged to *P. aeruginosa* isolates collected in Canada within 1 year and previously analyzed by Diaz Caballero et al. (2015) (NCBI Bioproject PRJNA282164, Supplementary Table S2), and 30 genome sequences were from unrelated isolates from

---

**FIGURE 4** | Similarity trees for *P. aeruginosa* ST235 isolates. (A) Similarity tree based on cgMLST analysis. (B) Similarity tree based on core genome SNP analysis. Branches in black and orange represent groups 1 and 2, respectively, previously described by Treepong et al. (2018). The 62 isolates previously evaluated by Treepong et al. are highlighted in green in the tree.
de Sales et al.

FIGURE 5 | Similarity trees based on cgMLST analysis of ST146 strains. (A) ST146 P. aeruginosa Liverpool Epidemic Strain (LES) from a single cystic fibrosis patient (NCBI Bioproject PRJEB6642, n = 40). Two distinct clusters are highlighted in red and blue, and 6 more divergent isolates are indicated by dashed branches. (B) ST146 P. aeruginosa LES and 8 additional unrelated ST146 isolates from the United Kingdom and Canada. Blue and red branches represent the 2 clusters originally described for this set of isolates (dashed branches indicate the more divergent isolates) and branches in green represent the 8 additional isolates from United Kingdom and Canada.

DISCUSSION

Simple, standardized, accurate and portable typing methods are needed for global surveillance programs and for outbreak investigations in clinical settings. Currently, PFGE and MLST are the two most frequently the used typing methods to investigate outbreaks (Maiden et al., 1998; Higgins et al., 2012). Most important drawbacks for these typing approaches include the low discriminatory power of MLST within outbreaks and difficulties for standardization of PFGE for inter laboratory comparisons (Higgins et al., 2012; Pérez-Losada et al., 2013).

Recent technical advances in WGS and lower costs have brought new possibilities for both global surveillance and local clinical investigations (de Been et al., 2015; Ruppitsch et al., 2015; Mellmann et al., 2016; Higgins et al., 2017; Nadon et al., 2017; Zhou et al., 2017; Janowicz et al., 2018; Kimura, 2018). Both wgMLST and cgMLST solutions are being implemented for a diverse set of microorganisms, such as those contemplated within the Global Platform for Pathogen Surveillance6. cgMLST is being considered the method of choice for typing pathogens for epidemiological surveillance of infectious diseases in the European Union and European Economic Area countries, overcoming SNP-based typing (Revez et al., 2017). However, there is no current consensus for a cgMLST scheme for P. aeruginosa despite two ad hoc cgMLST schemes previously

6https://pathogen.watch/
FIGURE 6 | Similarity trees based on cgMLST analysis of ST274 isolates. (A) *P. aeruginosa* ST274 from a single cystic fibrosis patient (NCBI Bioproject PRJNA282164, n = 167). Blue and red branches represent clades A and B, respectively, corroborating data reported by Diaz Caballero et al. (2015). Dashed and solid branches identify subgroups in clade A. (B) *P. aeruginosa* ST274 isolates from different geographic regions were added to the dataset reported in (A). Orange branches represent the 30 additional ST274 isolates from different geographic regions.

proposed (Mellmann et al., 2016; Royer et al., 2020) (see section “Discussion” later).

cgMLST is a typing technique that relies on the availability of precisely sequenced full genomes for generating typing schemes. Using only CDs present in core genome, cgMLST reduces the number of sites typed in the analysis, when compared with core genome SNP-based analysis, and would also allow for a standardized, allele-based, reproducible nomenclature (Janowicz et al., 2018). For reliable results in cgMLST the quality of genome sequences is fundamental (Janowicz et al., 2018).

We propose a public cgMLST scheme for *P. aeruginosa* based on 130 whole genome sequences and a total of 2,653 target loci were validated in 2,184 unfinished genome sequences from diverse geographical regions, hosts, and human body sites. In our study we were able to successfully type over 99% of the evaluated genomes with a cgMLST scheme considering that at least 95% of the target loci were present. The initial number of candidate genes from full genome data was 3,168, and the reduction to 2,653 gene targets relates not only to the genomic variability of *P. aeruginosa*, but also to technical aspects: assembling issues such as gaps or incomplete assembling, inadequate coverage during sequencing with impact in consensus sequence quality, and absence of genes from the core genome. In all core-genome based methods, the size of the core genome inevitably decreases when the number of analyzed isolates increases (Denton et al., 2014; Van Tonder et al., 2014; de Been et al., 2015). The most commonly used metric for assessing assembling quality is N50 (Salzberg and Yorke, 2005; Yandell and Ence, 2012; Lantz et al., 2018). Although it is a popular metric, it is questionable and there is no consensus on an ideal N50 value for genomes: hard-hitting assemblers can produce large contigs when compared to conservative assemblers, but are more prone to errors (Salzberg and Yorke, 2005; Lischer and Shimizu, 2017; Lantz et al., 2018).

Thus, in this study we chose to use lower number of contigs (< 200) for the selection of unfinished genomes for the validation of the proposed scheme.

During the scheme creation 11 fully assembled genomes were discarded due to lack of the candidate loci. Two of these sequences have been previously considered as outlier strains: PA7 (RefSeq assembly accession: GCF_000017205.1.) and CR1 (RefSeq assembly accession: GCF_000302534.5.) (Roy et al., 2010; Sood et al., 2019). In the above-mentioned study by Sood et al., the authors evaluated 14 genomes considered as outliers and 64 classical *P. aeruginosa* genomes. When analyzed together, these sequences shared 2,885 core genome genes. However, when the 64 classical genomes were analyzed independently from the outliers, the number of common genes rose to 3,199, and the outliers alone shared 4,708 genes. When we look at the list of genomes we kept for the cgMLST scheme we see that the algorithm correctly discarded both PA7 and CR1 due to lack of target genes7 (see file analysis_cg/removedGenomes.txt and analysis_cg/GenomeRemoved120thr.txt). When using the determined set of cgMLST gene targets proposed here, 10 of the fully assembled genomes were successfully typed (contained at least 95% of the selected loci).

It has been proposed that an isolate should harbor at least 95% of the set of cgMLST gene targets (de Been et al., 2015; Ruppitsch et al., 2015; Ghanem and El-Gazzar, 2018). We were able to find 95% of our gene set in all but 11 genomes analyzed. Only one of the genomes lacking 95% of gene targets was considered a fully assembled sequence (RefSeq assembly accession: GCF_000473745.2). This isolate, *P. aeruginosa* VRFPA04, was collected in India from a human

7https://github.com/BioinformaticsHIAEMolecularMicrobiology/cgMLST-Pseudomonas-aeruginosa
loci were not present. The unfinished sequences that had less of this sequence would be necessary to understand why these cornea sample (Murugan et al., 2016) and a more deep analysis of this sequence would be necessary to understand why these loci were not present. The unfinished sequences that had less than 95% of the targets did not belong to any particular origin or location.

Two ad hoc cgMLST schemes proposed for *P. aeruginosa* (Mellmann et al., 2016; Royer et al., 2020) used the commercial software Ridom SeqSphere+ (Ridom GmbH, Muenster, Germany) to generate the scheme. The above-mentioned schemes used the PAO1 genome as the seed genome and defined 3,842 and 4,162 target genes, respectively. In our scheme we propose 2,653 target genes, but we used 130 complete genomes as seed genomes. When we compare our gene targets to the one proposed by Mellmann et al. (2016) 2,081 of our gene targets (78.43%) match with the gene targets proposed by Mellmann et al. (data not shown). It is also important to mention that, in the literature, we find variation both in the total number of genomes included to define the core genome as well as parameters used for the delimitation of the core genome: sets between 4 and 1488 genomes have been used and between 665 and 5233 genes have been reported as constituting the core genome of *P. aeruginosa* (Valot et al., 2015; Vincent et al., 2017; Freschi et al., 2019; Sood et al., 2019; Subedi et al., 2019; Vasquez-Ríño et al., 2019). Besides the strict cut-off for the core genome gene selection used in this work (target loci were present in 100% of the 130 genomes at the creation of the scheme), other factors are worth consideration when comparing our gene set with that of Mellmann et al., we used a larger population for this study and a more varied one in terms of origin (human, ocean, animal, plants and industry) than the population they used, and we considered all unfinished genomes with acceptable quality at the validation phase, a factor that may have led to the exclusion of targets. For *K. pneumoniae*, similar results are seen, with two schemes, one proposing 1,143 gene targets and a second one 2,365 (Zhou et al., 2017; Neumann et al., 2019), even though both used SeqSphere+ (Ridom, Münster, Germany) to generate the scheme. A third cgMLST scheme was proposed for *K. pneumoniae* using the open source software Bacterial Isolate Genome Sequence Database (BIGSdb) and 634 target genes were included in their cgMLST scheme (Bialek-Davenet et al., 2014).

We did not compare our scheme directly with what we would obtain with the same set of isolates and using SeqSphere+ (Ridom, Münster, Germany), but we evaluated isolates from an outbreak setting characterized using the software (Royer et al., 2020). The cgMLST scheme that we propose identified few allelic differences (0 to 2) between the isolates, a similar result to the one proposed by the authors, but using our reduced number of target genes when compared to theirs. Few differences were observed between isolates involved in an outbreak, an expected result due to limited time for intraoutbreak evolution (Ruppitsch et al., 2015; Zhou et al., 2017). For instance, in outbreak settings involving *L. monocytogenes* (Ruppitsch et al., 2015), *M. tuberculosis* (Kohl et al., 2014), *L. pneumophila* (Moran-Gilad et al., 2015) and *K. pneumoniae*, isolates have been shown to differ in fewer than 10 cgMLST alleles (Zhou et al., 2017). de Been et al. (2015) proposed an interpretation for cgMLST of *Enterococcus faecium*: (i) isolates differing from 0 to 20 alleles are considered undistinguishable or closely related, and possibly involved in an outbreak; (ii) isolates differing between 21 and 40 alleles possibly belong to the same outbreak; and (iii) isolates differing in more than 40 alleles are unrelated. However, similarly to PFGE and MLST, thresholds should not substitute epidemiological investigation and, as previously suggested, microevolutionary events within each outbreak and the threshold of ≤10 different alleles warrant further validation (Ruppitsch et al., 2015; Zhou et al., 2017).

The cgMLST scheme proposed here helped to discriminate isolates belonging to the same ST, while clearly grouping isolates that had been epidemiologically linked in the original publications. Even though MLST is widely used in epidemiological studies, the fact that it addresses less than 0.1% of the genome, often limits its discriminatory power (Li et al., 2009; Sabat et al., 2013; Davis et al., 2015; Tang et al., 2017; Chen et al., 2018; Ghanem and El-Gazzar, 2018; Kimura, 2018). Our results are in agreement with recent publications for pathogens such as *Enterococcus faecium*, *L. monocytogenes*, *A. baumannii*, *K. pneumoniae*, and *E. faecalis* (de Been et al., 2015; Ruppitsch et al., 2015; Higgins et al., 2017; Zhou et al., 2017; Neumann et al., 2019).

Additionally, the cgMLST results presented here also corroborate with *P. aeruginosa* population studies for ST235 which used core genome SNP-based methodologies, with the clear advantage of targeting a smaller set of loci and the possibility of implementing a database with unique profiles for *P. aeruginosa* typing based on its core genome (de Been et al., 2015; Ruppitsch et al., 2015; Ghanem and El-Gazzar, 2018; Pearce et al., 2018; Schürch et al., 2018; Neumann et al., 2019).

In conclusion we present a highly discriminatory cgMLST scheme for WGS-based typing of *P. aeruginosa* developed using an open access platform. Different from the two available publications that used cgMLST to analyze *P. aeruginosa* isolates, we have made our scheme and associated files available at GitHub (https://github.com/BioinformaticsHAEMolecularMicrobiology/cgMLST-Pseudomonas-aeruginosa) aiming to foment discussions and possibly help in establishing a cgMLST consensus for *P. aeruginosa*. Gene target lists and selected loci are available at this location. In the GitHub link we also present a step-by-step explanation on how new genome sequences may be analyzed. The remaining challenge is to establish an Internet-based nomenclature server to facilitate universal global nomenclature for any user, as currently available for MLST.

**DATA AVAILABILITY STATEMENT**

All datasets generated for this study are included in the article/Supplementary Material.

**AUTHOR CONTRIBUTIONS**

PS conceived and supervised the project. RS and RP performed the analysis with support from LM. RS, LM, and PS wrote the manuscript with contributions from BK. RS and PS interpreted the data.
ACKNOWLEDGMENTS

We thank our colleagues from the clinical laboratory from Hospital Israelita Albert Einstein for insightful discussions, especially Dr. Marines Dalla Valle Martino, Dr. André Mario Doi, and Dr. Paula Celina Mariko Koga. We also thank Dr. Laura Leaden and Elisangela de Paula Silva for assistance in evaluating gene targets.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.01049/full#supplementary-material
molecular typing of Pseudomonas aeruginosa. J. Clin. Microbiol. 45, 3707–3712. doi: 10.1128/JCM.00560-07
Kidd, T. J., Grimwood, K., Ramsay, K. A., Rainey, P. B., and Bell, S. C. (2011). Comparison of three molecular techniques for typing Pseudomonas aeruginosa isolates in sputum samples from patients with cystic fibrosis. J. Clin. Microbiol. 49, 263–268. doi: 10.1128/JCM.01421-10
Kimura, B. (2018). Will the emergence of core genome MLST end the role of in silico MLST? Food Microbiol. 75, 28–36. doi: 10.1016/j.fm.2017.09.003
Kohl, T. A., Diel, R., Harmsen, D., Rothgänger, J., Meywald, W. K., Merker, M., et al. (2014). Whole-genome-based Mycobacterium tuberculosis surveillance: a standardized, portable, and expandable approach. J. Clin. Microbiol. 52, 2479–2486. doi: 10.1128/JCM.00567-14
Lantz, H., Dominguez Del Angel, V., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., et al. (2018). Ten steps to get started in genome assembly and annotation. F1000Res. 7:148. doi: 10.2807/1560-7917.ES.2015.20.10
Lazarovitch, T., et al. (2015). Design and application of a core genome multilocus sequence typing scheme for Enteroctococcus facialis. J. Clin. Microbiol. 57, e1686–e1618. doi: 10.1128/JCM.01686-18
Page, J. A., Taylor, B., and Keane, A. (2016). Multilocus sequence typing by blast from de novo assemblies against PubMed. J. Open Source Softw. 1:118. doi: 10.21105/joss.00118
Pearce, M. E., Alikhani, N. F., Dallman, T. J., Zhou, Z., Grant, K., and Maiden, M. C. J. (2018). Comparative analysis of core genome MLST and SNP typing within a European Salmonella serovar Enteritidis outbreak. Int. J. Food Microbiol. 274, 1–11. doi: 10.1016/j.ijfoodmicro.2018.02.023
Pérez-Losada, M., Cabezas, P., Castro-Nallar, E., and Crandall, K. A. (2013). Pathogen typing in the genomics era: MLST and the future of molecular epidemiology. Infect. Genet. Evol. 16, 38–53. doi: 10.1016/j.meegid.2013.01.009
Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS One 5:e9940. doi: 10.1371/journal.pone.0009490
Rezv, J. Espinosa, L., Albigler, B., Lettymeyer, K. C., and Struelens, M. J. (2017). Survey on the use of whole-genome sequencing for infectious diseases surveillance: rapid expansion of European national capacities, 2015-2016. Front. Public Health 5:347. doi: 10.3389/fpubh.2017.00347
Roy, P. H., Tetu, S. G., Larouche, A., Elbourne, L., Tremblay, S., Ren, Q., et al. (2010). Complete genome sequence of the multiresistant taxonomic outlier Pseudomonas aeruginosa PA7. PLoS One 5:8842. doi: 10.1371/journal.pone.008842
Royer, G., Fourreau, F., Boulanger, B., Mercier-Darty, M., Ducellier, D., Cizeau, J., et al. (2020). Local outbreak of extended-spectrum β-lactamase SHV2-producing Pseudomonas aeruginosa reveals the emergence of a new specific sub-lineage of the international ST235 high-risk clone. J. Hosp. Infect. 104, 33–39. doi: 10.1016/j.jhin.2019.07.014
Ruppitsch, W., Pietzka, A., Prior, K., Bletz, S., Fernandez, H. L., Allerberger, F., et al. (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. doi: 10.1128/JCM.01193-15
Sabat, A. J., Budimir, A., Nashev, D., Sá-Leão, R., van Dijl, J. M., Laurent, F., et al. (2013). Overview of molecular typing methods for outbreak detection and epidemiological surveillance. Eur. Commun. Dis. Bull. 18:20380. doi: 10.2807/ese.18.04.20380-en
Salzberg, S. L., and Yorke, J. A. (2005). Beware of mis-assembled genomes. Bioinformatics 21, 4320–4321. doi: 10.1093/bioinformatics/bti429
Schürch, A. C., Arredondo-Alonso, S., Willems, R. J. L., and Goering, R. V. (2015). Unraveling genomic and phenotypic nature of multidrug-resistant (MDR) Pseudomonas aeruginosa VRFPA04 isolated from keratitis patient. Microbiol. Res. 193, 140–149. doi: 10.1016/j.micres.2016.10.002
Sousa, D., Ceniceros, A., Galeiras, R., Pertega-Diaz, S., Gutierrez-Urbon, J. M., Rodriguez-Mayo, M., et al. (2018). Microbiology in burns patients with blood
stream infections: trends over time and during the course of hospitalization. Infect. Dis. 50, 289–296. doi: 10.1080/23744235.2017.1397738

Stover, C. K., Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warren, P., Hickey, M. J., et al. (2000). Complete genome sequence of Pseudomonas aeruginosa PA01, an opportunistic pathogen. Nature 406, 959–964. doi: 10.1038/35023079

Subedi, D., Kohli, G. S., Vijay, A. K., Wilcox, M., and Rice, S. A. (2019). Accessory genome of the multi-drug resistant ocular isolate of Pseudomonas aeruginosa PA34. PLoS One 14:e0215038. doi: 10.1371/journal.pone.0215038

Tang, P., Croxen, M. A., Hasan, M. R., Hsiao, W. W., and Hoang, L. M. (2017). Infection control in the new age of genomic epidemiology. Am. J. Infect. Control 45, 170–179. doi: 10.1016/j.ajic.2016.05.015

Tenover, F. C., Arbeit, R. D., Goering, R. V., Mickelsen, P. A., Murray, B. E., and Persing, D. H. (1995). Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. J. Clin. Microbiol. 33, 2233–2239.

Treepong, T. J., Ondov, B. D., Koren, S., and Phillippy, A. M. (2014). The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol. 15:524. doi: 10.1186/s13059-014-0524-x

Treepong, P., Kos, V. N., Guyeux, C., Blanc, D. S., Bertrand, X., Valot, B., et al. (2018). Global emergence of the widespread Pseudomonas aeruginosa ST235 clone. Clin. Microbiol. Infect. 24, 258–266. doi: 10.1016/j.cmi.2017.06.018

Valot, B., Guyeux, C., Rolland, J. Y., Mazouzi, K., Bertrand, X., and Hocquet, D. (2015). What it takes to be a Pseudomonas aeruginosa? The core genome of the opportunistic pathogen updated. PLoS One 10:e013715. doi: 10.1371/journal.pone.0126468

Van Tonder, A. J., Mistry, S., Bray, J. E., Hill, D. M. C., Cody, A. J., Farmer, C. L., et al. (2014). Defining the estimated core genome of bacterial populations using a bayesian decision model. PLoS Comput. Biol. 10:e1003788. doi: 10.1371/journal.pcbi.1003788

Vasquez-Rifo, A., Veksler-Lublinsky, I., Cheng, Z., Ausubel, F. M., and Ambros, V. (2019). The Pseudomonas aeruginosa accessory genome elements influence virulence towards Caenorhabditis elegans. Genome Biol. 20:270. doi: 10.1186/s10059-019-1896-1

Veesemeyer, J. L., Hauser, A. R., Lisboa, T., and Rello, J. (2009). Pseudomonas aeruginosa virulence and therapy: evolving translational strategies. Crit. Care Med. 37, 1777–1786. doi: 10.1097/CCM.0b013e31819ff137

Vincent, A. T., Freschi, L., Jeukens, J., Kukavica-Ibrulj, I., Jean-Guillaume, E.-R., Leduc, A., et al. (2017). Genomic characterisation of environmental Pseudomonas aeruginosa isolated from dental unit waterlines revealed the insertion sequence ISPa11 as a chaotropic element. FEMS Microbiol. Ecol. 93:fix106. doi: 10.1093/femsec/fix106

Williams, D., Evans, B., Haldenby, S., Walshaw, M. J., Brockhurst, M. A., Winstanley, C., et al. (2015). Divergent, coexisting Pseudomonas aeruginosa lineages in chronic cystic fibrosis lung infections. Am. J. Respir. Crit. Care Med. 191, 775–785. doi: 10.1164/rccm.201410-1646OC

Woodford, N., Ellington, M. J., Coelho, J. M., Turton, J. F., Ward, M. E., Brown, S., et al. (2006). Multiplex PCR for genes encoding prevalent OXA carbapenemases in Acinetobacter spp. Int. J. Antimicrob. Agents 27, 351–353. doi: 10.1016/j.ijantimicag.2006.01.004

Yandell, M., and Ence, D. (2012). A beginner’s guide to eukaryotic genome annotation. Nat. Rev. Genet. 13, 329–342. doi: 10.1038/nrg3174

Yin, S., Chen, P., You, B., Zhang, Y., Jiang, B., Huang, G., et al. (2018). Molecular typing and carbapenem resistance mechanisms of Pseudomonas aeruginosa isolated from a Chinese burn center from 2011 to 2016. Front. Microbiol. 9:1135. doi: 10.3389/fmicb.2018.01135

Zhou, H., Liu, W., Qin, T., Liu, C., and 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. doi: 10.3389/fmicb.2017.00371

Zhou, Z., Alkilan, N. F., Sergeant, M. J., Luhmann, N., Vaz, C., Francisco, A. P., et al. (2018). Grapepetree: visualization of core genomic relationships among 100,000 bacterial pathogens. Genome Res. 28, 1395–1404. doi: 10.1101/gr.232397.117

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 de Sales, Migliorini, Puga, Kocsis and Severino. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.