BMC Genomics

RESEARCH

Open Access

Unraveling the impact of genome assembly on bacterial typing: a one health perspective

Déborah Merda[1*], Meryl Vila-Nova[1], Mathilde Bonis[2], Anne-Laure Boutigny[3], Thomas Brauge[4], Marina Cavaiuolo[2], Amandine Cunty[3], Antoine Regnier[4], Maroua Sayeb[5], Noémie Vingadassalon[2], Claire Yvon[5] and Virginie Chesnais[1]

Abstract

Background In the context of pathogen surveillance, it is crucial to ensure interoperability and harmonized data. Several surveillance systems are designed to compare bacteria and identify outbreak clusters based on core genome MultiLocus Sequence Typing (cgMLST). Among the different approaches available to generate bacterial cgMLST, our research used an assembly-based approach (chewBBACA tool).

Methods Simulations of short-read sequencing were conducted for 5 genomes of 27 pathogens of interest in animal, plant, and human health to evaluate the repeatability and reproducibility of cgMLST. Various quality parameters, such as read quality and depth of sequencing were applied, and several read simulations and genome assemblies were repeated using three tools: SPAdes, Unicycler and Shovill. In vitro sequencing were also used to evaluate assembly impact on cgMLST results, for six bacterial species: *Bacillus thuringiensis*, *Listeria monocytogenes*, *Salmonella enterica*, *Staphylococcus aureus*, *Vibrio parahaemolyticus* and *Xylella fastidiosa*.

Results The results highlighted variability in cgMLST, which not only related to the assembly tools, but also induced by the intrinsic composition of the genomes themselves. This variability observed in simulated sequencing was further validated with real data for six of the bacterial pathogens studied.

Conclusion This highlights that the intrinsic genome composition affects assembly and resulting cgMLST profiles, and that variability in bioinformatics tools can induce a bias in cgMLST profiles. In conclusion, we propose that the completeness of cgMLST schemes should be considered when clustering strains.

Keywords cgMLST, Genome assembly, Bioinformatics tool variability, Data comparability

Introduction

In a One Health perspective, it is essential to maintain a global system of surveillance to better perceive and understand transmission events between animals, humans, and the environment. These surveillance systems need to be harmonized and to ensure interoperability between all the data generated so that they may be shared among all surveillance players, such as public health authorities, research institutions, and laboratories. These systems also involve several scientific domains, such as plant pathology or veterinary, medical, and food safety. The importance of such sharing of

*Correspondence:
Déborah Merda
deborah.merda@anses.fr
[1]Université Paris Est, ANSES, Laboratory for Food Safety, SPAAD unit, Maisons-Alfort F-94701, France
[2]Université Paris Est, ANSES, Laboratory for Food Safety, SBCL unit, Maisons-Alfort F-94701, France
[3]ANSES, Plant Health Laboratory, Bacteriology Virology GMO Unit, 7 rue Jean Dixméras, Angers cedex 01 49044, France
[4]ANSES, Laboratory for Food Safety, Bacteriology and Parasitology of Fishery and Aquaculture Products Unit (B3PA), Boulevard du Bassin Napoléon, Boulogne-sur-Mer, France
[5]Université Paris Est, ANSES, Laboratory for Food Safety, SEL unit, Maisons-Alfort F-94701, France

data has recently been proven for real-time monitoring of outbreaks or pandemics, as highlighted during the SARS-CoV-2 pandemic or other recent virus outbreaks [1]. Such systems are already used in bacteria monitoring systems to identify the origins and transmission routes of antimicrobial resistance [2, 3], or to monitor food-associated pathogens. Recommendations have thus been proposed to facilitate collaboration around data [4, 5]. These recommendations suggested in particular (i) defining quality criteria so as to ensure data trustworthiness, and (ii) providing guidelines and reference analytical tools for data processing while limiting the impact of their storage. To implement these recommendations, current systems for bacteria surveillance are primarily based on typing results [6].

The reference method for bacterial typing is multilocus sequence typing (MLST), based on seven housekeeping genes. It was developed for the first time in 1998 with *Neisseria meningitidis* and since then, the number of schemes available in the PubMLST database (Public database for molecular typing and microbial genome diversity) has steadily increased to over 130, demonstrating the ongoing growth and diversification of this typing method over time [7]. In the last few decades, the development of whole genome sequencing (WGS) has opened the path to gene-by-gene approaches to extend the MLST concept to all genes composing the core genome (cg) of bacterial species. This method, called cgMLST, is more discriminating than MLST due to its higher genome coverage level.

Zoonotic and foodborne pathogen surveillance based on these new approaches is increasingly, and most of the surveillance initiative tools published recently recommend using cgMLST outputs for comparing bacterial strains and identifying clusters of genetically-related strains (PulseNet USA [8], GenoSalmSurv [9], European Food Safety Authority (EFSA) [10]). Recently, an outbreak caused by *Listeria monocytogenes* ST1247 was investigated in five European countries (Denmark, Estonia, Finland, France, and Sweden), using the cgMLST approach [11]. In this study, only three allelic differences were found out of the 1744 loci detected from the 1748-loci cgMLST scheme [12]. Likewise, this method was used to investigate the global outbreak caused by *Salmonella Typhimurium* ST34 in chocolate-based products between 2021 and 2022. Cases were reported in 12 European Union countries, the UK, Switzerland, USA, and Canada [13].

Unlike methods based on read mapping, a variant that requires a reference genome to which reads are aligned, the gene-by-gene approach is reference-free, enabling better consideration of genetic variability among bacterial strains. Moreover, cgMLST appears to be less affected by homologous recombination than Single Nucleotide Polymorphims (SNP) analysis, and can be used to investigate outbreaks from highly recombinant pathogens like *Pseudomonas aeruginosa* [14], *Salmonella enterica* [15] or *Xylella fastidiosa* [16]. Furthermore, it is straightforward to establish nomenclature systems that can be shared among multiple institutes and/or analyses, facilitating the creation of a global monitoring system. These schemes and sequence variants are publicly available in several databases, e.g., PubMLST (https://pubmlst.org/), BIGSdb-Pasteur (https://bigsdb.pasteur.fr), EnteroBase (https://enterobase.warwick.ac.uk/), cgmlst.org (https://cgmlst.org/ncs) from Ridom SeqSphere and Chewie-NS (https://chewie-ns.readthedocs.io/en/latest/) [17]. There are different approaches to calling alleles and obtaining cgMLST profiles. One of them maps raw reads to a scheme to call genes, as implemented in Mentalist [18]. A second approach, implemented in ChewBBACA [19], is assembly-based, and requires genome assembly before calling cgMLST profiles. Various systems use it, like INNUENDO [10]. ChewBBACA is also implemented in an interoperable system shared by the European Food Safety Authority (EFSA) and the European Centre for Disease Prevention and Control (ECDC), which was set up in 2019 to analyze foodborne outbreaks caused by *Salmonella enterica*, *Listeria monocytogenes*, and *Escherichia coli* [5].

De novo assembly is a crucial step after sequencing to reconstruct the genomes of pathogens. Several pipelines designed to harmonize genome assembly have been published based on specific pathogens or institutes. These pipelines use de novo assembly tools like SPAdes [20], Shovill [21] or Unicycler [22], and short reads as the data input. SPAdes is a bacterial genome assembly algorithm based on de Bruijn graph published in 2012. Shovill and Unicycler are two assembly tools based on SPAdes that offer improvements over this first tool. Shovill, developed in 2016, is a pipeline designed to optimize the assembly runtimes by adding steps before and after SPAdes step. Unicycler, developed in 2017, is also an enhancement of SPAdes aimed at reducing the number of misassemblies at the end of the assembly process. One of the significant challenges in bacterial genome assembly is the use of short reads produced by next generation sequencing (NGS). Indeed, NGS tools can be easily impacted by genome composition, for example the occurrence of repeated sequences such as insertion sequences (IS), variable number tandem repeats (VNTRs), or homopolymers, which are very difficult to assemble. In addition, regions that vary greatly in GC composition have a poor sequencing coverage, leading to genome fragmentation [23]. Few is known about the impact of these genomic features on these different assembly tools.

The aim of this study was to evaluate the impact of assembly tools on bacteria to highlight the need for pipeline harmonization and to share cgMLST profiles with the EFSA/ECDC system, where cgMLST analyses are performed with ChewBBACA. Twenty-seven bacterial species corresponding to significant pathogens from a One Health perspective were examined in this study. These species encompass foodborne, plant, and animal pathogens. We compared the three tools most frequently used for assembly purposes: SPAdes [20], Unicycler [22] and Shovill [21]. The effect of the quality and depth of sequenced reads was evaluated on cgMLST results. The repeatability and reproducibility of analyses were also tested using both in silico and in vitro sequencing. We observed a major bioinformatics variability in the cgMLST profiles obtained, and therefore proposed recommendations to enhance interoperability between genomic results and to decrease the risk of excluding strains linked to each other in epidemic clusters.

Materials and methods

Experimental scheme

The genomes of 27 bacterial pathogen species—*Bacillus cereus*, *Bacillus thuringiensis*, *Bacillus cytotoxicus*, *Brucella melitensis*, *Burkholderia mallei*, *Campylobacter spp.*, *Citrobacter spp.*, *Clostridium botulinum*, *Clostridium difficile*, *Clostridium perfringens*, *Escherichia coli*, *Klebsiella aerogenes*, *Leptospira interrogans*, *Listeria monocytogenes*, *Mycobacterium bovis*, *Mycobacterium tuberculosis*, *Neisseria meningitides*, *Pseudomonas aeruginosa*, *Ralstonia solanacearum*, *Salmonella enterica*, *Staphylococcus argenteus*, *Staphylococcus aureus*, *Taylorella equigenitalis*, *Vibrio cholera*, *Vibrio parahaemolyticus*, *Xylella fastidiosa*, and *Yersinia enterocolitica* were used to perform these analyses (Table S1). The species were chosen according to the interest in these pathogens for public health, and their risk in food safety. A minimum of five circularized genomes were randomly chosen from the public NCBI database, resulting in 138 genomes being analyzed. All strain accession numbers are available in the supplementary data (Table S1).

The experimental design is presented in Fig. 1a. The short read paired end of 150 bp was simulated using ART v. 2.3.7 [24] to mimic Illumina sequencing. Phred quality scores (Q) for Illumina sequencing are guaranteed to be at least 95% above Q30 for all platforms, such as MiSeq, HiSeq and NextSeq. Two quality scores were then simulated: greater than Q40 to simulate high-quality reads and less than Q40 to estimate the impact of low-quality reads. The depth of sequencing can also differ depending on the multiplexing and sequencing platforms chosen. Because sequencing depth can affect genome assembly results, five different depths were simulated: 25x, 50x, 75x, 100x and 150x. The reproducibility of assembly, tested by comparing assembly following independent read simulations and cgMLST typing, was evaluated for three different simulated datasets of high-quality reads. Thus, a total of 2800 reads were simulated, with each genome undergoing 20 simulations. Read simulations were verified using fastp v. 0.20.1 [25].

Real dataset

In vitro sequencing data were used to validate simulation results for six bacterial species. The experimental design is presented in Fig. 1b. We used 28 different strains: five for *Bacillus thuringiensis*, five for *Listeria monocytogenes*, five for *Salmonella enterica*, five for *Staphylococcus aureus*, four for *Vibrio parahaemolyticus*, and four for *Xylella fastidiosa*. (Table S2). DNA was extracted from all these strains and sequenced independently twice. Quality was assessed and reads were trimmed using fastp v. 0.20.1 [25]. Finally, a total of 56 sequencing results were analyzed.

Assembly

In order to evaluate the impact of assembly tools on cgMLST typing, three tools were selected: SPAdes v.3.14.1 [20], Shovill v.1.0.9 [21], and Unicyler v.0.4.8 [22] using default settings. All the simulated and real sequenced reads were assembled with these three tools. To validate the repeatability of genome assembly by comparing assemblies obtained with the same tool and the same dataset simulation, each tool was used independently three times on high-quality simulated reads with a Phred score above 40 and a depth exceeding 75x. Real sequenced reads were also assembled independently three times. In all, 12,558 assemblies (138 genomes x 2 read quality simulations x 5 sequencing depth x 3 replicates for good quality of reads and for 75x, 100x and 150x of depth x 3 tools) were generated for simulated data and 504 assemblies for in vitro data (28 strains sequenced in duplicates, and analyzed thrice by 3 assembly tools).

Typing

All assemblies listed in Table S1 ($n=140$) were analyzed to generate the corresponding ST using mlst v2.23.0 and cgMLST profiles using chewBBACA v. 2.8.5, as recommended by the EFSA/ECDC system. Whenever possible, we used publicly available schemes from cgmlst.org or Big-SDB (Table S3). For *Taylorella equigenitalis* and *Xylella fastidiosa*, unpublished schemes were used to obtain cgMLST profiles with chewBBACA. The EFSA/ECDC system recommends using chewBBACA v. 2.8.5 or more recent versions [5]. In our study, cgMLST profiles were computed using chewBBACA v. 2.8.5 tools after assembly annotation using Prodigal [19].

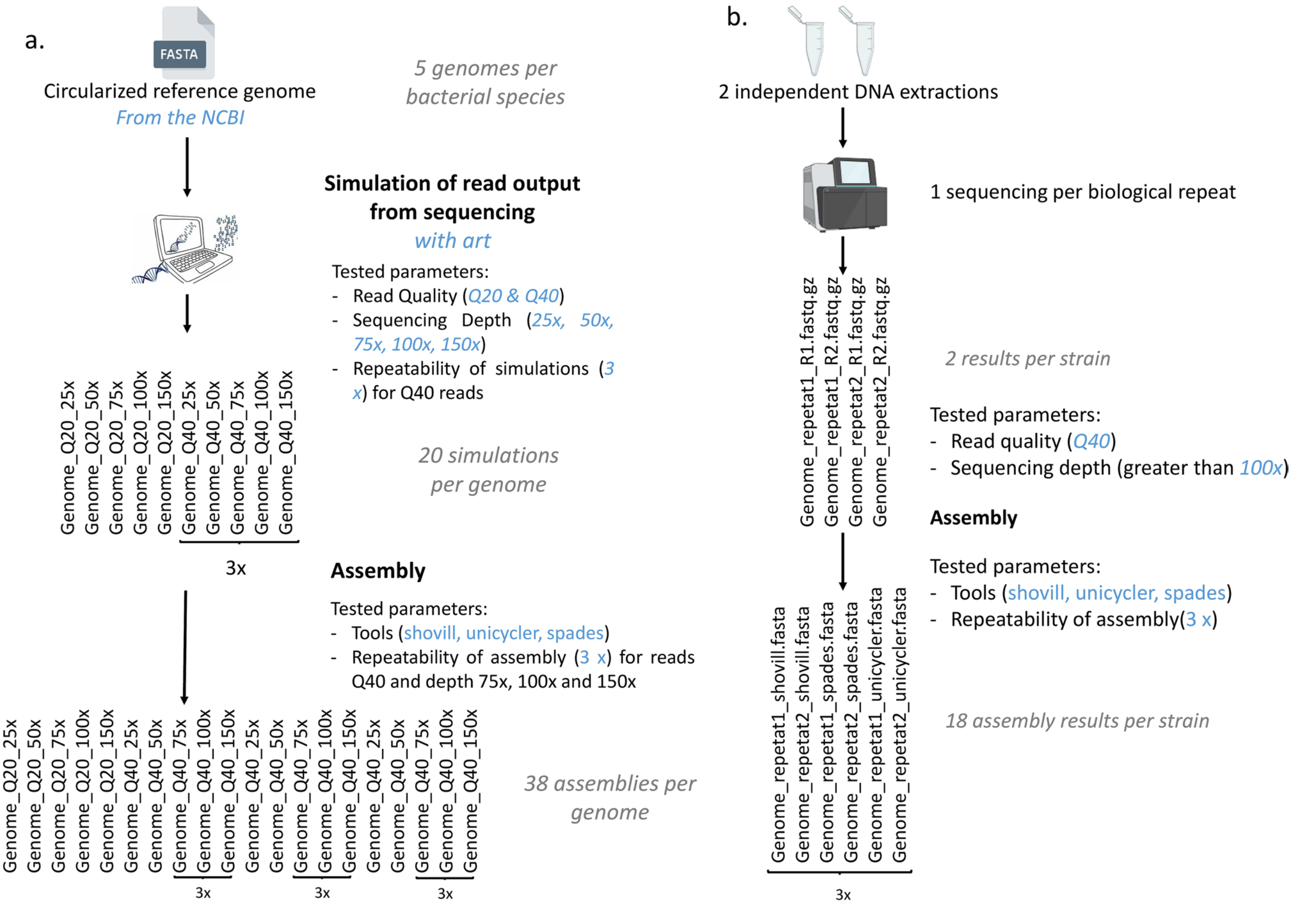
a.
FASTA
Circularized reference genome
From the NCBI
5 genomes per bacterial species
Simulation of read output from sequencing
with art
Tested parameters:
- Read Quality (Q20 & Q40)
- Sequencing Depth (25x, 50x, 75x, 100x, 150x)
- Repeatability of simulations (3 x) for Q40 reads
Genome_Q20_25x
Genome_Q20_50x
Genome_Q20_75x
Genome_Q20_100x
Genome_Q20_150x
Genome_Q40_25x
Genome_Q40_50x
Genome_Q40_75x
Genome_Q40_100x
Genome_Q40_150x
3x
20 simulations per genome
Assembly
Tested parameters:
- Tools (shovill, unicycler, spades)
- Repeatability of assembly (3 x) for reads Q40 and depth 75x, 100x and 150x
Genome_Q20_25x
Genome_Q20_50x
Genome_Q20_75x
Genome_Q20_100x
Genome_Q20_150x
Genome_Q40_25x
Genome_Q40_50x
Genome_Q40_75x
Genome_Q40_100x
Genome_Q40_150x
Genome_Q20_25x
Genome_Q40_50x
Genome_Q40_75x
Genome_Q40_100x
Genome_Q40_150x
Genome_Q40_25x
Genome_Q40_50x
Genome_Q40_75x
Genome_Q40_100x
Genome_Q40_150x
3x
3x
3x
38 assemblies per genome
b.
2 independent DNA extractions
1 sequencing per biological repeat
Genome_repetat1_R1.fastq.gz
Genome_repetat1_R2.fastq.gz
Genome_repetat2_R1.fastq.gz
Genome_repetat2_R2.fastq.gz
2 results per strain
Tested parameters:
- Read quality (Q40)
- Sequencing depth (greater than 100x)
Assembly
Tested parameters:
- Tools (shovill, unicycler, spades)
- Repeatability of assembly(3 x)
Genome_repetat1_shovill.fasta
Genome_repetat2_shovill.fasta
Genome_repetat1_spades.fasta
Genome_repetat2_spades.fasta
Genome_repetat1_unicycler.fasta
Genome_repetat2_unicycler.fasta
3x
18 assembly results per strain

Fig. 1 Representation of our study's experimental design **a**. Experimental design for simulated data. From circularized reference genome, sequencing was simulated. Quality parameters, including read quality and sequencing depth, were assessed. Reproducibility was evaluated, only on high read quality, through three simulated read datasets, from same reference genomes. Then, 20 raw reads were obtained for on genome. Each raw reads were assembled using three assembly tools: (SPAdes, Shovill and Unicycler). The repeatability of each assembler was tested with three repetitions of an assembly. From each genome, 38 assemblies were obtained ; **b**. Experimental design for real data. Two independent DNA extractions from each strain were sequenced independently. The three assemblers were compared for each sequencing dataset, and repeatability was assessed through three repetitions of an assembly. Finally, 18 assemblies were obtained per strain

Assembly quality parameters and visualization of cgMLST results

In order to compare assembly quality, four parameters from Quast results were analyzed [26]. To evaluate genome fragmentation, we compared contig numbers, N50 and largest contig sizes in all the assemblies. To assess assembly truthfulness, the number of misassemblies were detected by comparison with the initial genome and NGA50.

For each strain of all 27 species, assembly results were aligned with minimap2 [27] implemented in Quast to the initial reference genome used for reads simulation. Alignment was used to visualize contig fragmentation and evaluate assembly reproducibility and repeatability. The python library (seaborn v. 0.11.2 [28] and Circos v. 0.1.3 [29]) were used for all visualizations.

The cgMLST profiles of simulated datasets were compared by computing the allelic differences between genomes from NCBI and assembly results with GrapeTree v. 2.1 [30] after normalization. To obtain a completeness percentage for each scheme, this normalization step focused on the gene number in the scheme for each species analyzed (Table S3). The completeness was calculated on the basis of genes found by cgMLST analysis compared with the total number of genes in each scheme. The cgMLST results from real data were analyzed using the minimum spanning tree calculated with GrapeTree [30] and the MSTreeV2 method. These trees were visualized using the GrapeTree web application (achtman-lab.github.io/GrapeTree/MSTree_holder.html).

Data analysis

Analysis were performed on a linux server (Ubuntu 20.04.6 LTS) with 190 processors (Intel(R) Xeon(R) Gold 6348 H CPU @ 2.30 GHz) and 700GB RAM. Steps were parallelized with the help of Snakemake (v 7.24.2) to optimize jobs computing, and tools were managed using Anaconda environment (https://www.anaconda.com/). All assembly steps were paralyzed on 4 cores using tools specific options, and chewBBACA was performed for every species, dataset and assembly tools on a single core. Runtimes of steps varies according to the studied pathogens and the size of their genomes, as well as the complexity of cgMLST schema. The median runtime per species of main tools compared is reported in Supplemental table (Table S5). All scripts are available (SPAAD-ANSES/BenToAs (github.com)).

Results

Evaluation of assembly reproducibility according to sequencing quality and depth using simulated data

A key requirement for sharing data between interoperable surveillance systems is to evaluate the repeatability and reproducibility of analysis and to propose quality criteria for data inclusion. The assembly tools chosen (SPAdes, Unicycler and Shovill) were selected because they have been frequently used in recently published workflows dedicated to bacterial WGS. We evaluated the impact of read quality on sequencing simulations for 27 bacterial species, and observed that poor data quality (Q<40) decreases the quality of assembly: Assemblies were impossible to draft with Shovill, because the tool did not accept input data, or were shorter and more fragmented with SPAdes and Unicycler (Supplementary data S1). For *Vibrio parahaemolyticus*, the maximum number of contigs was 80 with high-quality data (Q>40) but increased to 120 with poor-quality data. For some species, such as *Bacillus cereus*, *Clostridium perfringens*, *Taylorella*, *Mycobacterium tuberculosis*, and *Ralstonia solanacearum*, some genome parts were even missing from the final assembly obtained with a poor read quality (Supplementary data S2), in position 0 Mb for *Bacillus cereus*, 0.1 Mb for *Clostridium perfringens*, 4.0 Mb for *Mycobacterium tuberculosis*, and 2.8 Mb for *Ralstonia solanacearum*.

Furthermore, the poor quality of reads also increased genome misassemblies compared with results obtained with a high read quality. Indeed, in *Klebsiella aerogenes*, at a depth of 75x, the maximum percentage of misassemblies was 40% with poor-quality reads whereas with high-quality reads this percentage would drop as far as 0%. For example, in *Mycobacterium bovis*, there were 20% of misassemblies with poor-quality reads vs. 0% with high-quality reads; in *Neisseria meningitides* these figures were 40% (poor quality) vs. 20% (high quality); in *Staphylococcus argenteus* they were 20% (poor quality) vs. 0%; and in *Bacillus cereus*, 20% (poor quality) vs. 7%. For *Clostridium perfringens*, the rate of misassemblies obtained with a poor read quality could reach 60% in some assemblies. For other species such as *Campylobacter spp.*, *Listeria monocytogenes*, *Escherichia coli* or *Vibrio cholerae*, assembly results appeared to be less affected by a poor read quality (Supplementary data S1).

When we compared the impact of various sequencing depths, we observed an optimal threshold at 75x. At this value, parameters representing high-quality assembly are maximized, i.e., the number of contigs and misassemblies decreases, and both N50 and total length increase. Mahn-Whitney tests used to compare the four-parameter distribution obtained at different sequencing depths were significant (Table S4). Results with 150x and 100x were identical. When comparing 25x with 100x sequencing depth, contig number distributions were significantly different for 10/27 species, N50 distributions were significantly different for 21/27 species, misassemblies for 25/27 species, and largest contig for 16/27 species. For 50x, no difference was observed in contig number, N50 and largest contig, while misassembly distributions were different for 10/27 species. For 75x, no difference was observed in contig number, N50, and largest contig, while misassembly distributions were different for 6/27 species. Therefore, for the subsequent analyses, we present results derived from high-quality reads at a depth of 75x (Supplementary data S3).

Comparison of assembly tools with a high read quality and sufficient depth using simulated data

To determine which tool performs better in genome assembly, SPAdes, Shovill and Unicycler were compared using simulated sequencing data with a high quality and mean depth of 75x. Firstly, we observed a major difference in tools runtime as Shovill is the fastest tool with a median runtime of 15 min per genome assembly and Unicycler the longest with a median runtime of 42 min for each assembly (Table S5). Our results indicated that assembly repeatability depends on the tools but seems to be also genome-dependent. An alignment of the generated assemblies to the reference used for the sequencing simulation revealed that both Shovill and Unicycler performed better for *Listeria monocytogenes* and *Ralstonia solanacearum* than for most of the 27 bacterial species (Fig. 2a). Interestingly, these tools fragmented the genome into similar genomic regions, which seem to correlate with variations in GC content across the genome. However, assembling the genome of *Mycobacterium bovis* and *Xylella fastidiosa* with the same assembly tool led to different results (Fig. 2b). Specifically, for these two species, the assembly differed for each sequencing simulation dataset (i.e., read simulations obtained from the same genome).

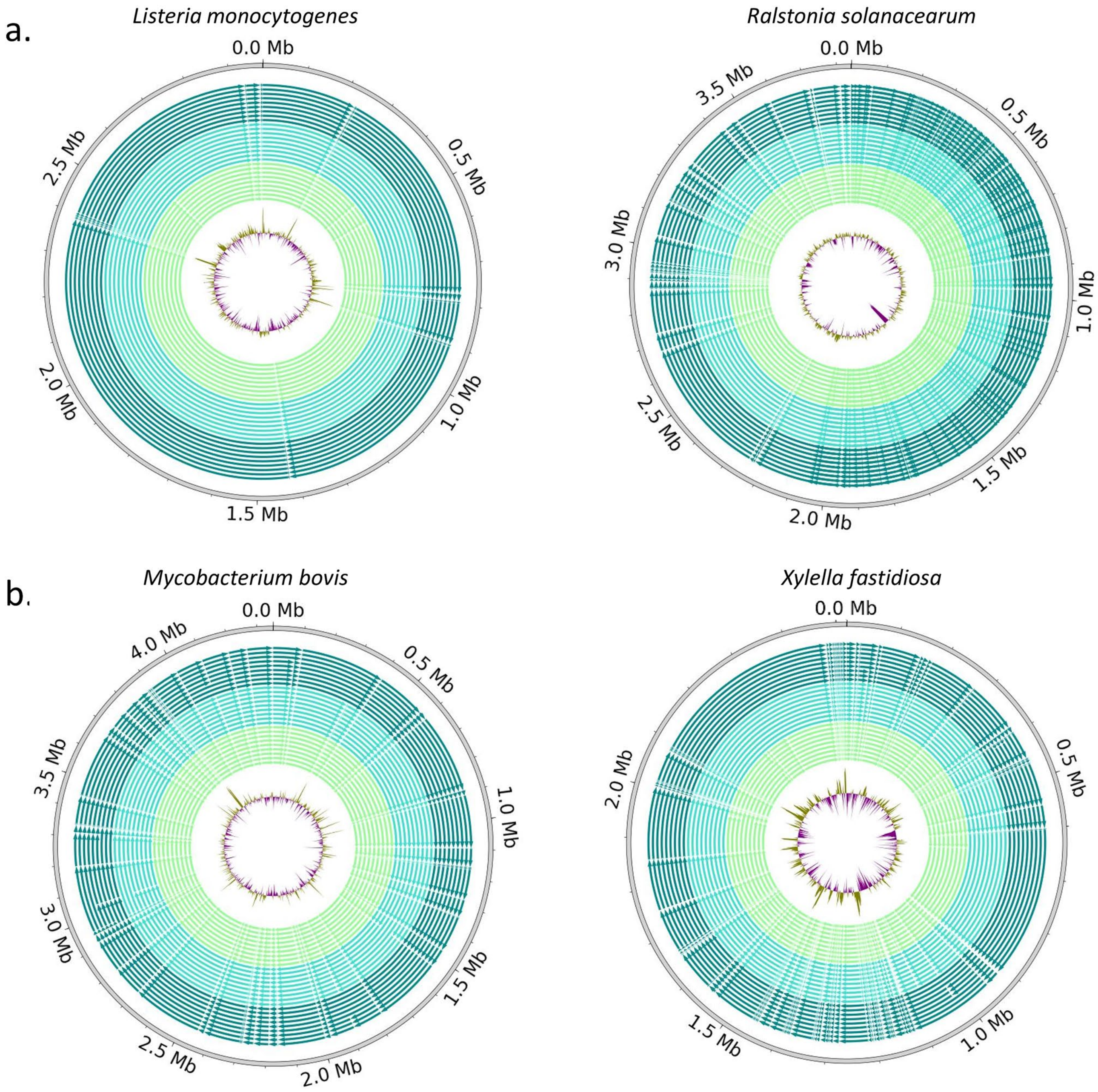

Fig. 2 Circos plots of assembled contig alignments to a reference genome used for simulations of high quality read. The GC variation along the genome is represented at center of the circle. The three simulated sequencing datasets and the three replicates for each assembly tool are represented (27 assemblies per genome) for a depth of 75x. The results from SPAdes are in green, those from Shovill in turquoise, and those from Unicycler in dark turquoise. **a**. Two represent genomes, one of *Listeria monocytogenes* and one of *Ralstonia solanacearum* showing identical results between Shovill and Unicycler. **b**. Two represent genomes, one of *Mycobacterium bovis* and one of *Xylella fastidiosa* showing different results between Unicycler replicates

Impact of assembly tools on cgMLST profiles using simulated data

Once the optimum quality criteria for sequencing were determined, the impact of cgMLST analyses was evaluated for 23 species for which a cgMLST scheme was available. The impact of sequencing quality was also studied on MLST, but no differences between the various simulations were observed (data not shown), which is why the impact on all genes involved in the cgMLST scheme was studied. The cgMLST profiles obtained from high-quality sequencing (i.e., Q>40) with sufficient depth (i.e., depth=75X) classified bacterial species into two categories based on the allelic difference rates observed between the reference genome and the assemblies obtained (Fig. 3). Results from SPAdes consistently exhibited higher assembly fragmentation and misassemblies

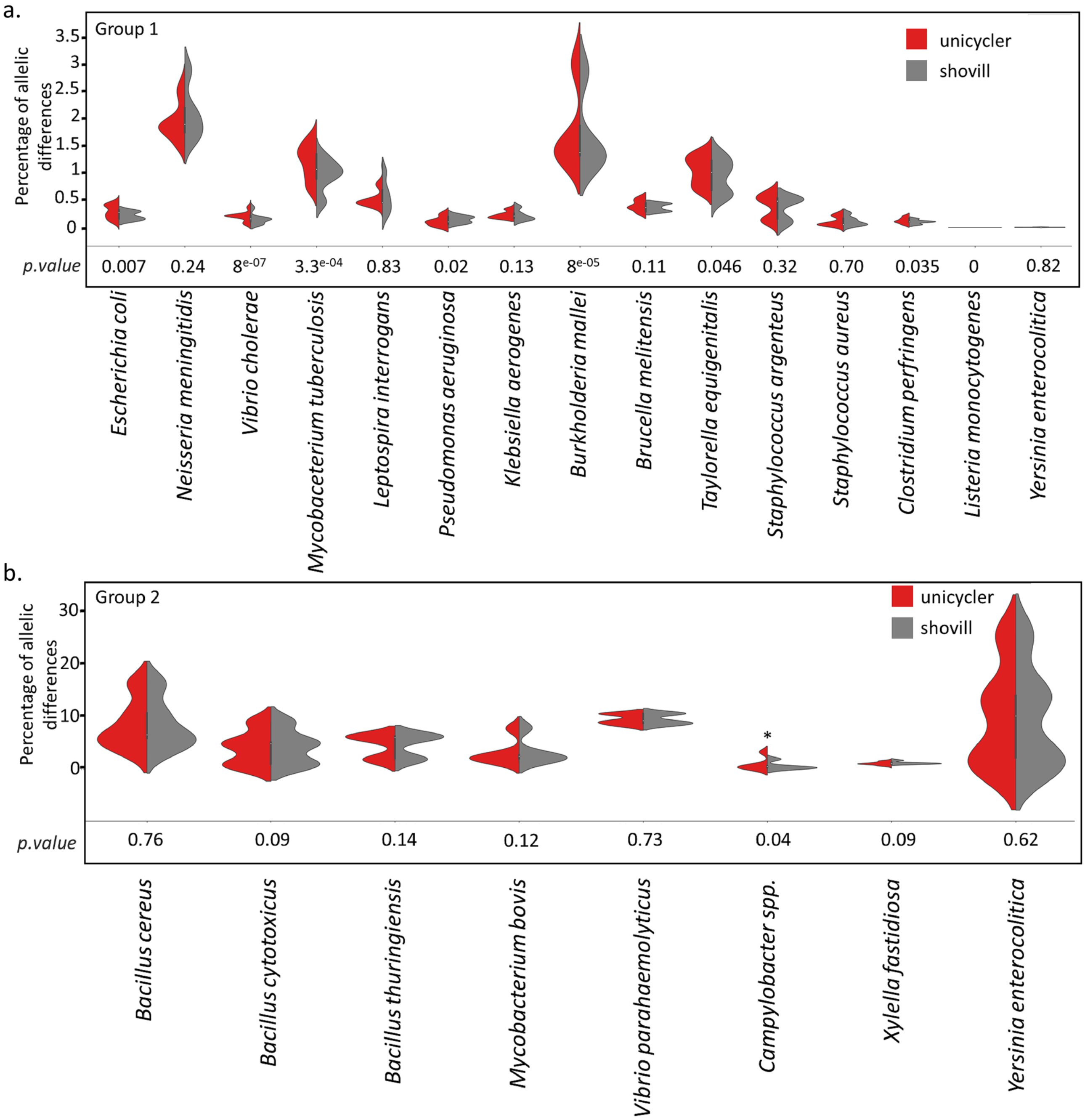

Fig. 3 Violin plot of allelic distribution rates according to the gene number in the cgMLST scheme. The results obtained with Unicycler (red) and with Shovill (grey) assemblies were obtained using simulated reads with a Phred score greater than Q40 and a depth of 75x. **a**: species for which the distribution of allelic difference rates is less than 5%. **b**: species for which the distribution of allelic difference rates is greater than 5%. The *p*-values were calculated with the non-parametric Mann-Whitney test, and significance is represented by *

than those obtained with Shovill and Unicycler, and are not therefore presented here. The first category (group 1) comprised 15 out of 23 bacterial species that had less than 5% of errors between the reference and the assembly obtained. For group 1, results suggested that the choice of assembler should vary according to the species studied (Fig. 3a). Indeed, for *Escherichia. coli*, *Mycobacterium tuberculosis*, *Vibrio cholerae*, and *Taylorella equigenitalis*, a significant difference (*p-value*<5% for Mann-Whitney test) was observed between Shovill and Unicycler results, suggesting that Shovill gave cgMLST profiles closest to the reference. However, for *Neisseria meningitidis* and *Leptospira interrogans*, the allelic profiles were closest to the reference when Unicycler was used, although no significant difference was observed when checked with the Mann-Whiney test.

The second category (group 2) comprised 8 out of 23 bacterial species for which the number of allelic differences between the reference and the assembly obtained was greater than 5% (Fig. 3b), with a maximum of 30% for *Salmonella enterica*. Within group 2, few differences were observed between the results obtained from Shovill and Unicycler assemblies, suggesting that the choice of assembly tool may be negligible compared with the intrinsic genome composition, except for *Campylobacter spp.* for which a significant difference was observed between distribution results from the two tools.

Comparison of cgMLST profiles obtained with different sequencing depths using simulated data

Related strains were identified by clustering cgMLST profiles obtained with different data quality and depth combinations. In open-source surveillance systems or applications, various data qualities can be shared with the science community with diverse internal sequencing capacities and/or quality thresholds. To evaluate the impact of various sequencing depths on cgMLST results, we compared simulated sequencing data associated with mean depths of 25x, 50x, and 75x. The number of allelic differences between reference cgMLST profiles and cgMLST profiles obtained significantly increased for assemblies with a sequencing depth less than 75x for all species belonging to group 1 (Fig. 4a). Only four out of 23 bacterial species, all belonging to group 2 previously described (i.e., greater than 5%), appeared not to be impacted by the quality of sequenced data: *Bacillus cereus*, *Bacillus cytotoxicus*, *Bacillus thuringiensis*, and *Vibrio parahaemolyticus* (Fig. 4b), as no significant difference was observed. However, for other species—regardless of whether they belong to the first or second group previously described—the number of allelic differences was significantly higher with poor depth (Q<40) using simulated sequencing data. These results underscored the importance of performing genomic typing on harmonized, high-quality data with a sufficient sequencing depth to investigate outbreaks.

Confirmation of reproducibility and repeatability when sequencing real data

To confirm the poor repeatability and reproducibility of cgMLST results obtained using simulated sequencing data and evaluate the impact on real data, we analyzed biological replicates of bacterial strains from six species. The cgMLST profiles were computed for each biological replicate to evaluate reproducibility, and bioinformatics analyses were performed in triplicate to investigate repeatability.

The cgMLST profiles obtained using real data showed that the results were repeatable between analyses, as also observed with simulated sequencing. Indeed, the cgMLST profiles resulting from SPAdes and Unicycler assemblies were comparable between each replicate, indicating 100% repeatability, as no distance was observed between assemblies obtained from the same raw data (Fig. 5). However, poor reproducibility was observed between the biological replicates, with distances observed between the same strains for which raw data were provided from two independent extractions. This finding suggests that the wet lab part has a major impact on cgMLST profiles, despite using the same DNA extraction protocol for *Salmonella enterica*, *Staphylococcus aureus*, and *Xylella fastidiosa*. Indeed, only four out of 28 strains had identical profile results with Unicycler. With Shovill, repeatability seemed to be dependent on the species. For instance, for *Listeria monocytogenes* all analyses were 100% identical, whereas for *Staphylococcus aureus*, *Vibrio parahaemolyticus*, and *Xylella fastidiosa* the strains had different cgMLST profiles resulting from distinct assemblies. For *Salmonella enterica* and *Bacillus thuringiensis*, one and two strains, respectively, gave different cgMLST profiles between analyses, but only one gene was systematically affected.

The cgMLST profiles for biological replicates were found to be identical for eight out of 28 analyzed strains (Fig. 5). These eight strains belong to *Bacillus thuringiensis* (two out of five strains), *Listeria monocytogenes* (four out of five strains), *Vibrio parahaemolyticus* (one out of four strains), and *Salmonella enterica* (one out of five strains). This level of reproducibility was mainly observed for the results generated by SPAdes and Unicycler, although only the Unicycler results maximized the completeness of the cgMLST scheme, i.e., more genes in the cgMLST scheme were found after Unicycler assembly. Conversely, with Shovill, only five strains had the same cgMLST profiles for biological replicates (one *Bacillus thuringiensis*, and four *Listeria monocytogenes*), and only four strains gave profiles that were identical to the Unicycler results (one *Bacillus thuringiensis* and three *Listeria monocytogenes*).

The number of allelic differences between biological replicates was found to be elevated (22 allelic differences between two *Listeria monocytogenes* replicates or 184 between two *Staphylococcus aureus* replicates), suggesting potential ambiguity for closely-related strains (Fig. 5). Depending on the species and assembly tools used, the number of allelic differences between biological replicates varied significantly, ranging from 10 allelic differences for *Bacillus thuringiensis*, to 138 for *Salmonella enterica* with Unicycler. Results obtained for two closely-related strains of *Xylella fastidiosa* subsp. *multiplex*, both belonging to ST6 based on the MLST of seven housekeeping genes (Amandine Cunty, personal communication), were mixed for cgMLST results, whereas they were found to be distinguishable in SNP analyses (data

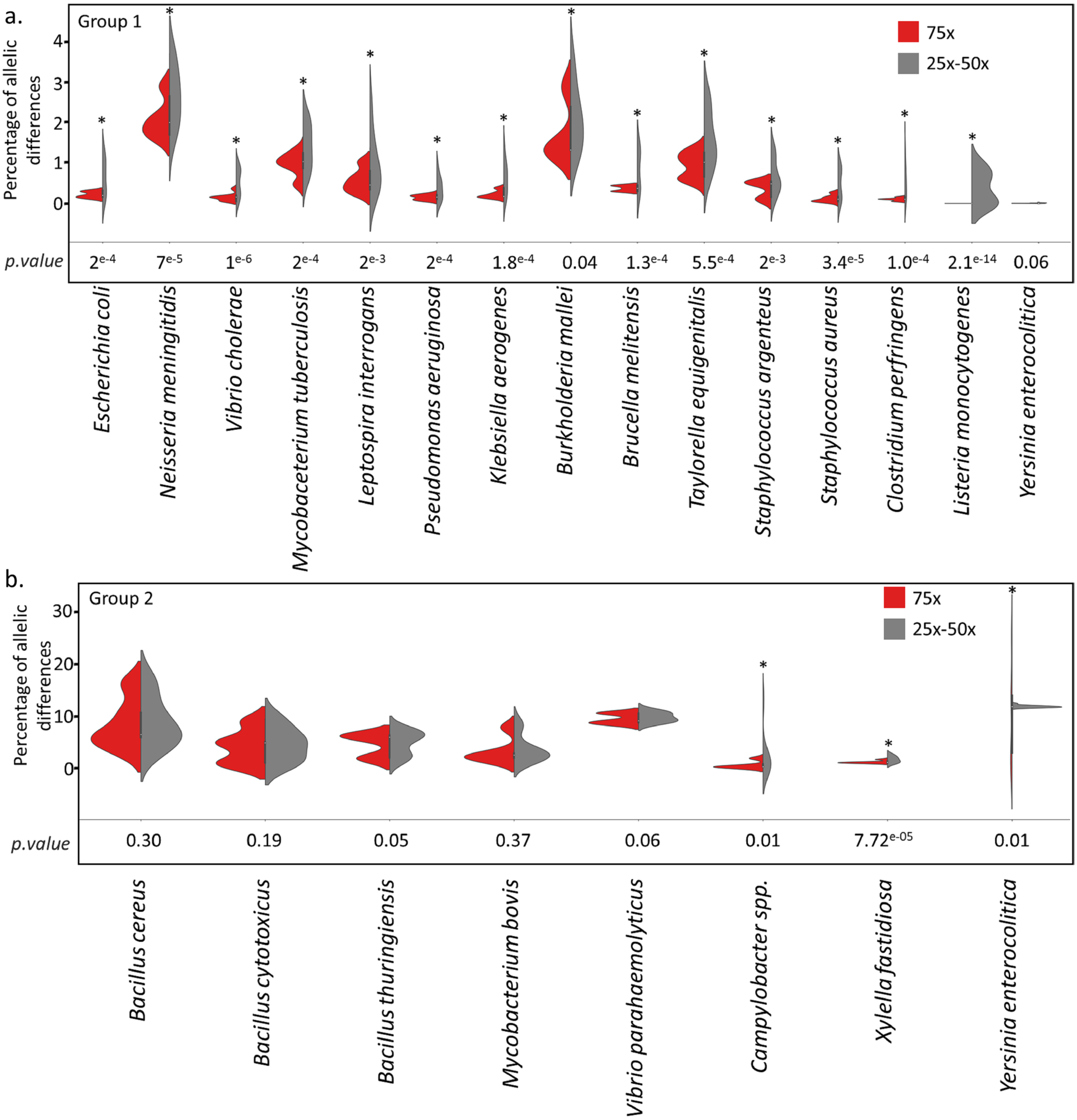

Fig. 4 Violin plot of distribution rates of allelic differences to genes number in the cgMLST scheme. Results obtained from simulated data with a Phred score greater than Q40 and a sequencing depth of 75x in red, or a depth lower than 75x in grey. Shovill was used for assemblies **a**: for which the distribution of allelic difference rates is less than 5%. **b**: for species whose distribution of allelic difference rates is greater than 5%. The *p* values were calculated according to the non-parametric Mann-Whitney test, and significance is denoted by *

not shown). These results suggested that for outbreak investigations using this method, it may be challenging to discriminate the strain responsible for the outbreak and consequently determine its source.

Discussion

cgMLST typing is one of the most widely used genomic methods for surveillance of bacterial pathogens. Our study aimed to investigate how the assembly step influences cgMLST profiles. Our results indicated that assembly-based cgMLST analyses, considering the entire scheme, may vary depending on the assembly method

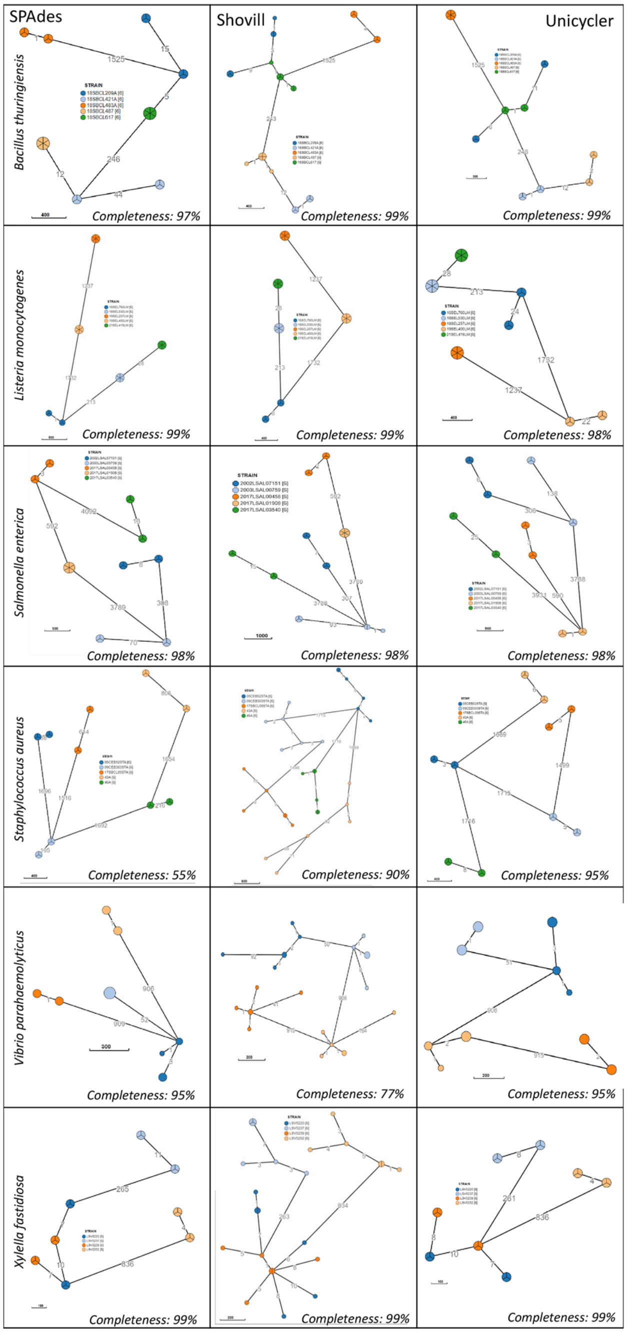

Fig. 5 Minimum spanning tree (MST) obtained from cgMLST profiles using real data. From left to right: results from SPAdes, Shovill and Unicycler. Each color represents one strain, for which two biological replicates were performed; the circle size indicates the number of assemblies sharing the same cgMLST profile, and allelic differences are indicated on the branches. The completeness value corresponds to the percentage of the gene scheme used to perform analyses

used. This represents a significant limitation for the gene-by-gene approach in interoperable systems, which aggregates data from various analytical pipelines. However, the observed differences, often referred to as false negatives, primarily involve genes that are missing rather than allelic differences potentially resulting in different allelic combinations.

The results obtained in this study highlight an impact of assembly on cgMLST profiles that is greater for particular bacterial species. Indeed, genomic composition may influence assembly quality, leading to possible contig fragmentation within a cgMLST gene. Repeat sequences such as insertion sequences (IS) or VNTRs can influence assembly quality, among other factors. A previous study demonstrated that the number of contigs obtained after assembly was correlated with the number of repeat elements in genomes [31]. The variability in GC content can also lead to non-reproducible analyses [32] due to biases introduced during sequencing, which alter sequencing depth in these regions [23]. Moreover, increased variability in a genome leads to a higher degree of bias observed during sequencing. This bias affects all assembly methods using short reads, since the corresponding tools are not capable of effectively handling inconsistent sequencing depths. Although Unicycler showed better performance in reducing misassemblies than SPAdes [22]and Shovill, as previously observed [22], it produced higher fragmented assembly than Shovill. Finally, as SPAdes seems to produce assemblies with poorer quality in most of the cases, the choice between Shovill and Unicycler should be made according to the parameter to optimize (i.e.: reduction of missassemblies or fragmentation).

The ability of a pathogen to capture external DNA by homologous recombination can directly impact GC content in recombination hotspots [33]. Thus, the difficulty in assembling genomes could be more pronounced for bacterial species with more frequent homologous recombination. Our results revealed two distinct groups with less than or more than 5% of allelic differences, respectively. Group 1, for which an allelic variation lower than 5% was described, included *Listeria monocytogenes*, *Staphylococcus aureus*, and *Brucella melitensis*, among others. For these species, mutations were identified as the primary evolutionary force responsible for polymorphism [34–36]. In contrast, within the second group—exemplified by *Xylella fastidiosa* and *Salmonella enterica*—strains had cgMLST results that were significantly different from those of the reference, indicating that recombination was the main evolutionary force [16, 37].

In addition to intrinsic genomic composition, our results showed that sequencing quality affected cgMLST-typing. A recent study conducted with four food pathogens: *Campylobacter spp.*, *Listeria monocytogenes*, *Salmonella enterica*, and *Escherichia coli*, demonstrated variability induced by the wet lab part of WGS analyses [38]. In our study, we observed that bioinformatics analyses could also introduce variability in results. In a precedent study based on read simulations, the authors proposed a depth threshold at 50x based on analyses carried out on food pathogens *Escherichia coli*, *Listeria monocytogenes*, and *Salmonella enteric*a [38]. It should be noted that the analyses were conducted on a single strain per species, using a single tool (SPAdes) to compare typing results. However, by increasing the number of strains and the diversity of species investigated, our results showed that the quality of assembly obtained from 50x affected the typing result, and this bias decreased with depths equal to or greater than 75x. In the global monitoring systems, the diversity analyzed is even greater, and it is essential to evaluate these criteria for several distinct genomes per species. For this reason, we extended the study to 27 pathogens and included several genomes per species, allowing us to evaluate both the intra- and interspecies variability. This is why we proposed a minimum depth threshold of 75x for all pathogens.

Our results also showed that wet lab and bioinformatic variabilities can artificially increase the distance between related strains and thus impact outbreak investigations, potentially resulting in false negatives with unrelated strains. Indeed, when analyzing an epidemiological cluster, it is crucial to identify both the strains within the cluster and those excluded. This is based on a computation of allelic distance between strains (i.e., the number of differences between two profiles). Below a specific threshold, strains are considered related [39, 40]. Thresholds for cgMLST clustering have been proposed for several bacterial species, including *Listeria monocytogenes* [39], *Escherichia coli* [41], *Staphylococcus aureus* [42], and *Pseudomonas aeruginosa* [43], and several methods to estimate them have been developed based on modeling [39] or nonparametric statistics [40]. However, in monitoring systems, such as Chewie-NS or GenoSalmSurv, the thresholds are applied exclusively to allelic differences, with the number of undiscovered loci frequently not taken into consideration. Yet, as we have shown in this study, the genome quality can highly affect the completeness of cgMLST results (i.e., the number of genes that are found during analysis). This parameter increases the weight for allelic differences. For example, the established threshold for *Staphylococcus aureus* is 24 different alleles to define a cluster of related strains [42], with a complete cgMLST scheme comprising 1861 genes. However, our results were obtained using only 1005 genes. So, based on the reduction in the scheme's completeness, the threshold should be reduced to 13 different alleles for this specific clustering analysis.

Consequently, for outbreak investigations, it may be beneficial to include the value of scheme completeness (as defined by Palma et al. (2022) [44]), and to propose quality criteria, which maximizes this value in monitoring systems. Other parameters—such as homologous recombination and GC content—could be taken into account by a gene-by-gene approach to scheme definition, as the GC bias could lead to major genome fragmentation in assembly analyses. However, these propositions should be balanced against the need to consider some of the evolutionary history of outbreaks, given that GC and recombination represent horizontal gene transfers (HGTs). Yet, these transfers are very important for the evolution of virulence among bacteria, as shown for *Yersinia enterocolitca* [45]. As recently proposed by Duval et al. (2023)., these thresholds should not be defined by species but rather by either outbreak, taking into account evolutionary parameters (such as mutation, duration, etc.) specific to outbreaks [39], or by specific lineages that could have a specific evolutionary mechanism (such as being highly clonal) compared with other lineages. Furthermore, the development of assembly-free methods like SNP approaches at pangenome level could facilitate outbreak investigations using the pangenome graph method. In this study, only the benchmark of the assembly-based cgMLST method was performed, as this method is implemented in the EFSA system. However, other methods based on raw reads are available, as such as Mentalist (https://github.com/WGS-TB/MentaLiST/tree/master) and were compared on *Listeria monocytogenes* [44]. They show, as previously observed, that the tool used for cgMLST analysis have a significant impact on profiling, and that assembly-free tools can outperformed assembly-based tools like EFSA system recommended tool ChewBBACA. However, it could be interesting to investigate if reads quality and quantity affect these assembly-free approaches as well as they affect ChewBBACA output and if there is significant differences regarding pathogens [4].

Conclusion

Our study assessed the bioinformatic variability induced in bacterial typing analyses using the cgMLST method. By including foodborne and clinical pathogens, and using simulated and real data, our findings led us to propose new practices when implementing this method in surveillance systems, such as integrating the notion of completeness for outbreak investigation, and establishing minimum quality criteria for sequencing.

Consequently, our study allows us to establish some recommendations for cgMLST analyses on genome assembly using ChewBBACA based on the identified parameters that affect the most cgMLST results:

- Use sequencing data with at least 75x depth ;
- Take into account the scheme completeness in outbreak investigation to ensure quality of reporting of relative strains using a predetermined reference threshold of pairwise allele differences ;

- Transform allelic profiles with hash script in order to facilitate data comparability and sharing, as well as standardize allelic distances as nomenclature of missing genes and various alleles are harmonized independently of an international reference database.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-10982-z.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Supplementary Material 5

Supplementary Material 6

Supplementary Material 7

Supplementary Material 8

Acknowledgements
We thank Laurent Guillier for discussion about analyses, and we thank Delphine Libby-Claybrough for English editing.

Author contributions
V.C. conceived and designed the experiments. D. M. designed the analytical strategy and performed analyses. M.V.N. participated in analytical strategy and revised the paper. M.B., A.L.B., T. B., M.C., A.C., A.R., M.S., N.V., and C.Y. collected the samples and extracted the DNA for whole genome sequencing. D.M. and V.C. wrote and revised the paper. All the authors read and approved the final manuscript.

Funding
This work was supported by the SPAAD unit's internal resources.

Data availability
Sequence data that support the findings of this study have been deposited in the NCBI with the primary accession code PRJNA1129992.Please find the link to the dataset: https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA1129992.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 5 July 2024 / Accepted: 30 October 2024
Published online: 08 November 2024

References

1. Oude Munnink BB, Sikkema RS, Nieuwenhuijse DF, Molenaar RJ, Munger E, Molenkamp R, et al. Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. Sci 8 janv. 2021;371(6525):172–7.
2. Chakraborty T, Barbuddhe S. Enabling one health solutions through genomics. Indian J Med Res. 2021;153(3):273.
3. Wheeler NE, Price V, Cunningham-Oakes E, Tsang KK, Nunn JG, Midega JT, et al. Innovations in genomic antimicrobial resistance surveillance. Lancet Microbe 1 déc. 2023;4(12):e1063–70.
4. Timme RE, Wolfgang WJ, Balkey M, Venkata SLG, Randolph R, Allard M, et al. Optimizing open data to support one health: best practices to ensure interoperability of genomic data from bacterial pathogens. One Health Outlook. 2020;2(1):20.
5. Costa G, Di Piazza G, Koevoets P, Iacono G, Liebana E, Pasinato L et al. Guidelines for reporting whole genome sequencing-based typing data through the EFSA One Health WGS System. EFSA Support Publ juin 2022;19(6).
6. Gerner-Smidt P, Hise K, Kincaid J, Hunter S, Rolando S, Hyytiä-Trees E, et al. PulseNet USA: a five-year update. Foodborne Pathog Dis mars. 2006;3(1):9–19.
7. Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russel JE, Urwin R, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A. 1998;95:3140–5.
8. Scharff RL, Besser J, Sharp DJ, Jones TF, Peter GS, Hedberg CW. An economic evaluation of PulseNet: A Network for Foodborne Disease Surveillance. Am J Prev Med mai. 2016;50(5 Suppl 1):S66–73.
9. Uelze L, Becker N, Borowiak M, Busch U, Dangel A, Deneke C, et al. Toward an Integrated Genome-based surveillance of Salmonella enterica in Germany. Front Microbiol [Internet]. 2021. https://doi.org/10.3389/fmicb.2021.626941. https://www.frontiersin.org/journals/microbiology/articles/. 12. Disponible sur.
10. Llarena AK, Ribeiro-Gonçalves BF, Nuno Silva D, Halkilahti J, Machado MP, Da Silva MS, et al. INNUENDO: a cross-sectoral platform for the integration of genomics in the surveillance of food-borne pathogens. EFSA Support Publ. 2018;15(11):1498E.
11. Mäesaar M, Mamede R, Elias T, Roasto M. Retrospective use of whole-genome sequencing expands the Multicountry Outbreak Cluster of Listeria monocytogenes ST1247. Int J Genomics 1 avr. 2021;2021:1–5.
12. Moura A, Tourdjman M, Leclercq A, Hamelin E, Laurent E, Fredriksen N, et al. Real-time whole-genome sequencing for Surveillance of Listeria monocytogenes, France. Emerg Infect Dis Sept. 2017;23(9):1462–70.
13. EFSA. Multi-country outbreak of monophasic Salmonella Typhimurium sequence type 34 linked to chocolate products – first update – 18 May 2022. EFSA Support Publ juin 2022;19(6).
14. Blanc DS, Magalhães B, Koenig I, Senn L, Grandbastien B (2020) Comparison of whole genome (wg-) and Core Genome (cg-) MLST (BioNumericsTM) Versus SNP variant calling for Epidemiological Investigation of Pseudomonas aeruginosa. Front Microbiol 11.
15. Didelot X, Bowden R, Street T, Golubchik T, Spencer C, McVean G, et al. Recombination and population structure in Salmonella enterica. PLoS Genet Juill. 2011;7(7):e1002191.
16. Vanhove M, Retchless AC, Sicard A, Rieux A, Coletta-Filho HD, De La Fuente L et al. Genomic diversity and recombination among Xylella fastidiosa subspecies. Appl Environ Microbiol Juill 2019;85(13).
17. Mamede R, Vila-Cerqueira P, Silva M, Carriço JA, Ramirez M. Chewie nomenclature server (chewie-NS): a deployable nomenclature server for easy sharing of core and whole genome MLST schemas. Nucleic Acids Res 8 janv. 2021;49(D1):D660–6.
18. Feijao P, Yao HT, Fornika D, Gardy J, Hsiao W, Chauve C et al. MentaLiST – a fast MLST caller for large MLST schemes. Microb Genomics 1 févr 2018;4(2).
19. Silva M, Machado MP, Silva DN, Rossi M, Moran-Gilad J, Santos S et al. chewBBACA: a complete suite for gene-by-gene schema creation and strain identification. Microb Genomics 1 mars 2018;4(3).
20. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a New Genome Assembly Algorithm and its applications to single-cell sequencing. J Comput Biol Mai. 2012;19(5):455–77.
21. Seemann T. Shovill: faster SPAdes assembly of Illumina reads. 2017.
22. Wick RR, Judd LM, Gorrie CL, Holt KE, Unicycler. Resolving bacterial genome assemblies from short and long sequencing reads. PLOS Comput Biol 8 juin. 2017;13(6):e1005595.
23. Chen YC, Liu T, Yu CH, Chiang TY, Hwang CC. Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly. PLoS ONE. 29 avr. 2013;8(4):e62856.
24. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. Bioinf 15 févr. 2012;28(4):593–4.
25. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinf 1 sept. 2018;34(17):i884–90.

26. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinf 15 avr. 2013;29(8):1072–5.
27. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinf Mai. 2018;34(18):3094–100.
28. Waskom M. Seaborn: statistical data visualization. J Open Source Softw 6 avr. 2021;6(60):3021.
29. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information esthetic for comparative genomics. Genome Res. 2009;19(604):1639–45.
30. Zhou Z, Alikhan NF, Sergeant MJ, Luhmann N, Vaz C, Francisco AP, et al. GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. Genome Res sept. 2018;28(9):1395–404.
31. Acuña-Amador L, Primot A, Cadieu E, Roulet A, Barloy-Hubler F. Genomic repeats, misassembly and reannotation: a case study with long-read resequencing of Porphyromonas gingivalis reference strains. BMC Genomics. 16 déc. 2018;19(1):54.
32. Mavromatis K, Land ML, Brettin TS, Quest DJ, Copeland A, Clum A, et al. The fast changing Landscape of sequencing technologies and their impact on Microbial Genome assemblies and Annotation. PLoS ONE 12 déc. 2012;7(12):e48837.
33. Lassalle F, Périan S, Bataillon T, Nesme X, Duret L, Daubin V. GC-Content evolution in bacterial genomes: the biased gene Conversion Hypothesis expands. PLOS Genet 6 févr. 2015;11(2):e1004941.
34. den Bakker HC, Didelot X, Fortes ED, Nightingale K, Wiedmann M. Lineage specific recombination rates and microevolution in Listeria monocytogenes. BMC Evol Biol. 2008;8(1):277.
35. Fraser C, Hanage WP, Spratt BG. Neutral microepidemic evolution of bacterial pathogens. Proc Natl Acad Sci 8 févr. 2005;102(6):1968–73.
36. Vishnu US, Sankarasubramanian J, Sridhar J, Gunasekaran P, Rajendhran J. Identification of recombination and positively selected genes in Brucella. Indian J Microbiol 29 déc. 2015;55(4):384–91.
37. Park CJ, Andam CP. Distinct but Intertwined Evolutionary Histories of Multiple Salmonella enterica Subspecies. mSystems. 11 févr. 2020;5(1).
38. Forth LF, Brinks E, Denay G, Fawzy A, Fiedler S, Fuchs J et al. Impact of wet-lab protocols on quality of whole-genome short-read sequences from foodborne microbial pathogens. Front Microbiol 29 nov 2023;14.
39. Duval A, Opatowski L, Brisse S. Defining genomic epidemiology thresholds for common-source bacterial outbreaks: a modelling study. Lancet Microbe mai. 2023;4(5):e349–57.
40. Radomski N, Cadel-Six S, Cherchame E, Felten A, Barbet P, Palma F et al. A simple and robust statistical method to define genetic relatedness of samples related to outbreaks at the genomic scale – application to Retrospective Salmonella Foodborne Outbreak investigations. Front Microbiol. 24 oct 2019;10.
41. Schürch AC, Arredondo-Alonso S, Willems RJL, Goering RV. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches. Clin Microbiol Infect avr. 2018;24(4):350–4.
42. Lagos AC, Sundqvist M, Dyrkell F, Stegger M, Söderquist B, Mölling P. Evaluation of within-host evolution of methicillin-resistant Staphylococcus aureus (MRSA) by comparing cgMLST and SNP analysis approaches. Sci Rep 22 juin. 2022;12(1):10541.
43. Martak D, Meunier A, Sauget M, Cholley P, Thouverez M, Bertrand X, et al. Comparison of pulsed-field gel electrophoresis and whole-genome-sequencing-based typing confirms the accuracy of pulsed-field gel electrophoresis for the investigation of local Pseudomonas aeruginosa outbreaks. J Hosp Infect août. 2020;105(4):643–7.
44. Palma F, Mangone I, Janowicz A, Moura A, Chiaverini A, Torresi M, et al. In vitro and in silico parameters for precise cgMLST typing of Listeria monocytogenes. BMC Genomics 26 déc. 2022;23(1):235.
45. Karlsson PA, Tano E, Jernberg C, Hickman RA, Guy L, Järhult JD, et al. Molecular characterization of Multidrug-Resistant Yersinia enterocolitica from Foodborne outbreaks in Sweden. Front Microbiol. 2021;12:664665.

Publisher's note