BMC Genomics

RESEARCH

Open Access

In vitro and in silico parameters for precise cgMLST typing of *Listeria monocytogenes*

Federica Palma[1†], Iolanda Mangone[2†], Anna Janowicz[3], Alexandra Moura[4,5], Alexandra Chiaverini[6], Marina Torresi[6], Giuliano Garofolo[3], Alexis Criscuolo[7], Sylvain Brisse[1,8], Adriano Di Pasquale[2], Cesare Cammà[2] and Nicolas Radomski[2*]

Abstract

Background: Whole genome sequencing analyzed by core genome multi-locus sequence typing (cgMLST) is widely used in surveillance of the pathogenic bacteria *Listeria monocytogenes*. Given the heterogeneity of available bioinformatics tools to define cgMLST alleles, our aim was to identify parameters influencing the precision of cgMLST profiles.

Methods: We used three *L. monocytogenes* reference genomes from different phylogenetic lineages and assessed the impact of in vitro (i.e. tested genomes, successive platings, replicates of DNA extraction and sequencing) and in silico parameters (i.e. targeted depth of coverage, depth of coverage, breadth of coverage, assembly metrics, cgMLST workflows, cgMLST completeness) on cgMLST precision made of 1748 core loci. Six cgMLST workflows were tested, comprising assembly-based (BIGSdb, INNUENDO, GENPAT, SeqSphere and BioNumerics) and assembly-free (i.e. kmer-based MentaLiST) allele callers. Principal component analyses and generalized linear models were used to identify the most impactful parameters on cgMLST precision.

Results: The isolate's genetic background, cgMLST workflows, cgMLST completeness, as well as depth and breadth of coverage were the parameters that impacted most on cgMLST precision (i.e. identical alleles against reference circular genomes). All workflows performed well at ≥40X of depth of coverage, with high loci detection (> 99.54% for all, except for BioNumerics with 97.78%) and showed consistent cluster definitions using the reference cut-off of ≤7 allele differences.

Conclusions: This highlights that bioinformatics workflows dedicated to cgMLST allele calling are largely robust when paired-end reads are of high quality and when the sequencing depth is ≥40X.

Keywords: cgMLST, Comparability of workflows, *Listeria monocytogenes*, Principal component analysis, Generalized linear model

Introduction

A key component of the surveillance of microbial pathogens is the recognition of closely related strains, so that clusters of infection cases can be identified, and further investigations (e.g., identification of the source of contamination) and control measures taken [1]. Multi-locus sequence typing (MLST) was developed in 1998 and provided high reproducibility in the characterization of isolates, enabling to identify the same clones within bacterial populations [2]. However, it lacks discrimination at the

*Correspondence: n.radomski@izs.it
†Federica Palma and Iolanda Mangone contributed equally to this work.
[2] Istituto Zooprofilattico Sperimentale dell'Abruzzo e del Molise "Giuseppe Caporale" (IZSAM), National Reference Centre (NRC) for Whole Genome Sequencing of microbial pathogens: data-base and bioinformatics analysis (GENPAT), via Campo Boario, 64100 Teramo, TE, Italy
Full list of author information is available at the end of the article

strain level [3–5]. With the advances in whole genome sequencing (WGS) [6–9], core genome MLST (cgMLST) tools and schemes have been proposed for several bacterial pathogens, expanding the advantages of MLST at the genomic scale and providing a high level of bacterial strain discrimination. cgMLST relies on defining alleles for thousands of gene loci, translating sequence variation into numerical profiles, which are computationally easier and faster to handle and analyze, as compared with genome-based sequence alignments [10, 11].

Different commercial and open-source solutions have been proposed for cgMLST, differing in the type of input data (i.e. reads and/or assemblies), in the allele definition strategies (i.e. algorithms based on nucleotide alignments, protein-coding genes predictions, or kmer counting) and in settings used to generate cgMLST profiles [12–20]. Multi-center ring trials focusing on reproducibility and comparability of cgMLST-based bacterial typing and clustering showed discrepancies due to non-harmonized bioinformatic workflows that may affect the precision of WGS-based surveillance and outbreaks investigation [21, 83].

Distinct core genome-based MLST schemes have been proposed for high resolution typing of the foodborne pathogen *Listeria monocytogenes*, ranging from 1013 to 1827 loci [12, 23–25], including an open-source reference cgMLST scheme of 1748 gene loci that is used worldwide [26, 28, 28, 30, 31, 31] and curated in the open-source Bacterial Isolate Genome Sequence database (BIGSdb) [26].

Several parameters such as genetic background of tested strains [32], successive platings [33, 38], replicates of DNA extraction and sequencing [35], targeted depth of coverage [24, 28, 36], estimated depth and breadth of coverage [13, 38, 42] and assembly quality [39], may impact alleles called, compromising cgMLST profiles reproducibility and the definition of outbreak clusters.

We therefore aimed to identify in vitro and in silico parameters impacting the precision of cgMLST profiles from six cgMLST complete workflows while assessing clustering concordance using the cut-off of 7 alleles mismatches [24]. Our study represents a substantial extension in terms of number of assessed allele callers and parameters of the study recently published by Lüth et al. (2021) [46].

Results

The experimental plan set-up for this study (Fig. 1A) allowed us to build an accurate dataset of paired-end reads controlling the depth of coverage (Fig. 1B-i), statistically identify parameters explaining the cgMLST precision among a large set of in vitro and in silico parameters (Fig. 1B-ii), and illustrate graphically those parameters explaining the cgMLST precision (Fig. 1B-iii). Here, we focused on the precision (i.e. identical alleles against reference circular genomes (IAAR)) and completeness (i.e. identified alleles against schema (IAAS)) of cgMLST profiles, rather than accuracy, because allele differences were observed when comparing cgMLST profiles of reference circular genomes from compared cgMLST workflows (i.e. BIGSdb, INNUENDO, GENPAT, SeqSphere, BioNumerics and MentaLiST) (Fig. 2).

Benchmarking dataset of downsampled reads

Paired-end reads used for downsampling ($n=42$) contained enough reads ($3.77 \pm 0.71 \times 10^6$) to prepare a dataset of downsampled paired-end reads to process with the selected cgMLST workflows. This dataset presented the highest expected read depth of coverage (i.e. 100X), as well as high and stable average Phred quality scores (34.64 ± 0.07) and percentages of Phred quality scores higher than 30 ($93.00 \pm 1.29\%$). No single nucleotide variant (SNV) was detected during Confindr-based exogenous DNA contamination screening [41] in the dataset used for downsampling. Regardless of the targeted read depth of coverage (Dr) defined according to kmer depth (Dk) with BBNorm (ranging from 10X to 100X) [42], the breadth of coverage of the downsampled paired-end reads estimated with BBMap ($n=420$) [42] was very high (>99.3%) for each of the tested reference genomes (Table 1, Fig. 3A and Additional file 1). The accuracy of this downsampled reads was corroborated by the concordance (i.e. linear dependencies with slopes close to one) observed between the read depth of coverage estimated with BBMap [42] and INNUca [49] ($R^2 > 99.7\%$; Pearson test: $p < 2 \times 10^{-16}$) for the three reference genomes of interest (Fig. 3B).

Principal component analysis

Principal component analyses (PCAs) were built according to investigated categorical parameters, namely tested genomes (A) successive platings (B), replicates of DNA extraction (C) and sequencing (D), targeted depth (E) and cgMLST workflows (F) (Fig. 4 and Additional file 4). PCAs showed that the investigated in vitro parameters (i.e. successive platings, DNA and sequencing replicates), did not impact the precision (i.e. IAAR) and completeness (i.e. IAAS) of cgMLST profiles (Fig. 4B-Fig. 4D; Additional file 4B-Additional file 4D). In contrast, the tested reference genomes (Fig. 4A), targeted depth (Fig. 4E) and cgMLST workflows (Fig. 4F) may influence the precision and completeness of cgMLST profiles. More precisely, high targeted read (Dr) and kmer (Dk) depth (DrDk) were associated with high IAAR values (Fig. 4E), depth and breadth of coverage, as well as LA, N50 and NA50 for the assembly-based workflows

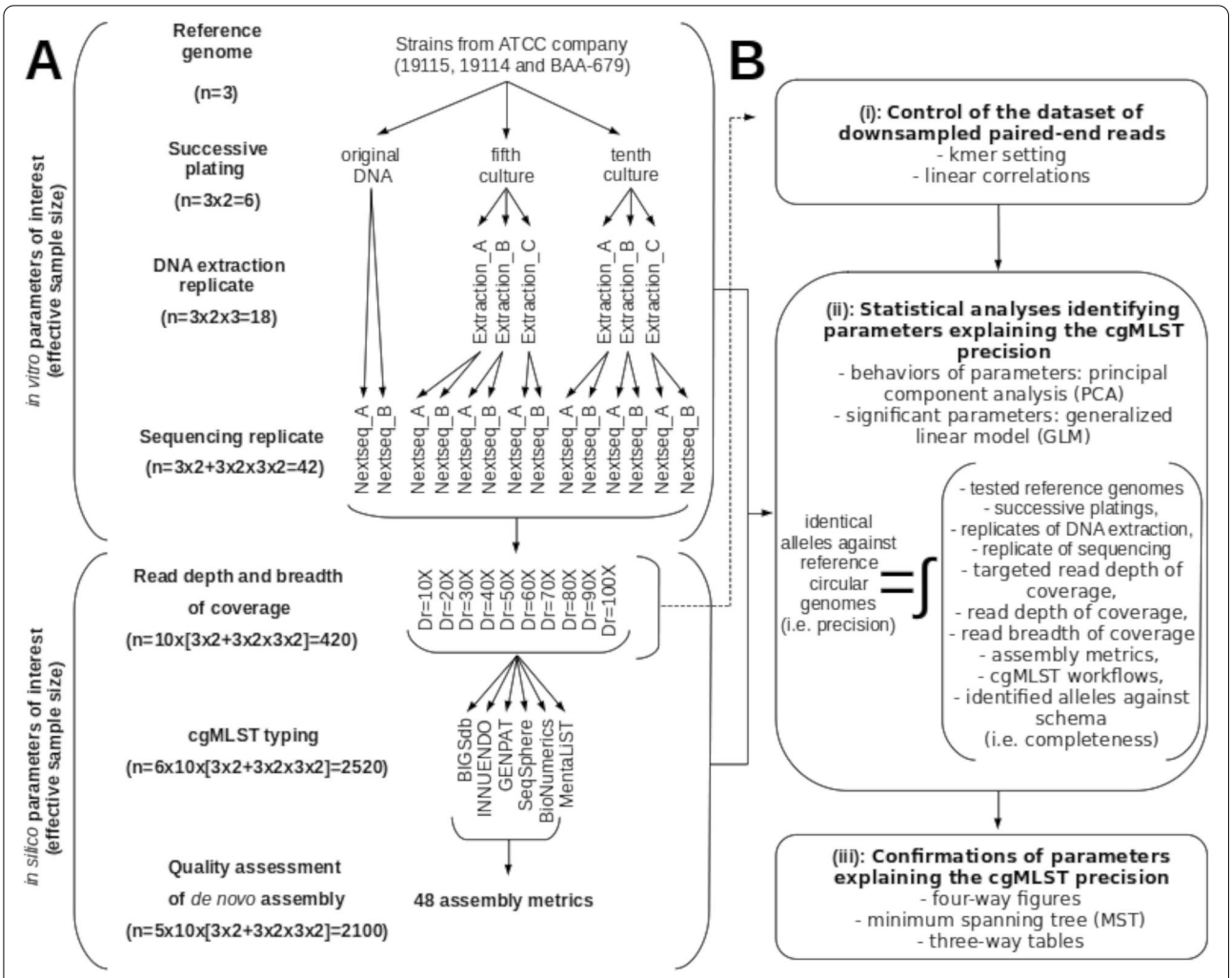

Fig. 1 Experimental plan (A) and data analyses (B) aiming at controlling a dataset of downsampled paired-end reads (i: $n=420$ paired-end reads), as well as identifying statistically (ii) and confirming significant parameters (iii) explaining the *Listeria monocytogenes* cgMLST precision (identical alleles against circular reference genomes: $n=2520$ cgMLST typing) among in vitro (i.e. tested reference genomes, successive platings, as well as replicates of DNA extraction and sequencing) and in silico (i.e. targeted depth of coverage, read depth and breadth of coverage, assembly metrics, cgMLST workflows, identified alleles against schema) parameters ($n=57$ parameters of interest)

(Additional file 4E-Additional file 4F). Otherwise, low values of DrDk (Fig. 4E) were mainly associated with the workflows BioNumerics and MentaLiST (Fig. 4F; Additional file 4F). Low values of IAAS were associated with the reference genome ATCC19114 (Fig. 4A; Additional file 4A). Overall for assembly-based workflows, the decrease of cgMLST precision (i.e. IAAR) was associated with high values of MA, GC, TL1000 and TL10000 or high values of L50, LA50, C1000 and C10000 (Additional file 4).

Generalized linear model

Generalized linear models (GLMs) were performed including all cgMLST workflows (A) or focusing on BIGSdb (B), INNUENDO (C), GENPAT (D), SeqSphere (E) and BioNumerics (F) (Additional file 5). The assembly metrics were not linearly correlated to cgMLST precision through GLMs ($p>1.0\times10^{-3}$) (Additional file 5). The GLM (Table 2) globally showed that IAAR (i.e. precision) was significantly explained by the workflow MentaLiST ($p=2.0\times10^{-16}$), breadth ($p=5.3\times10^{-12}$) and depth ($p=1.5\times10^{-12}$) of coverage, tested reference genome ATCCBAA679 ($p=2.0\times10^{-16}$), as well as amount of any base (N) per 100 kb (N100) ($p=2.0\times10^{-16}$) and IAAS (i.e. completeness) ($p=3.7\times10^{-6}$) for assembly-based cgMLST workflows (Additional file 5A). Looking at these workflows individually, the GLMs showed that IAAR was significantly explained by N100 ($p=2.0\times10^{-16}$)

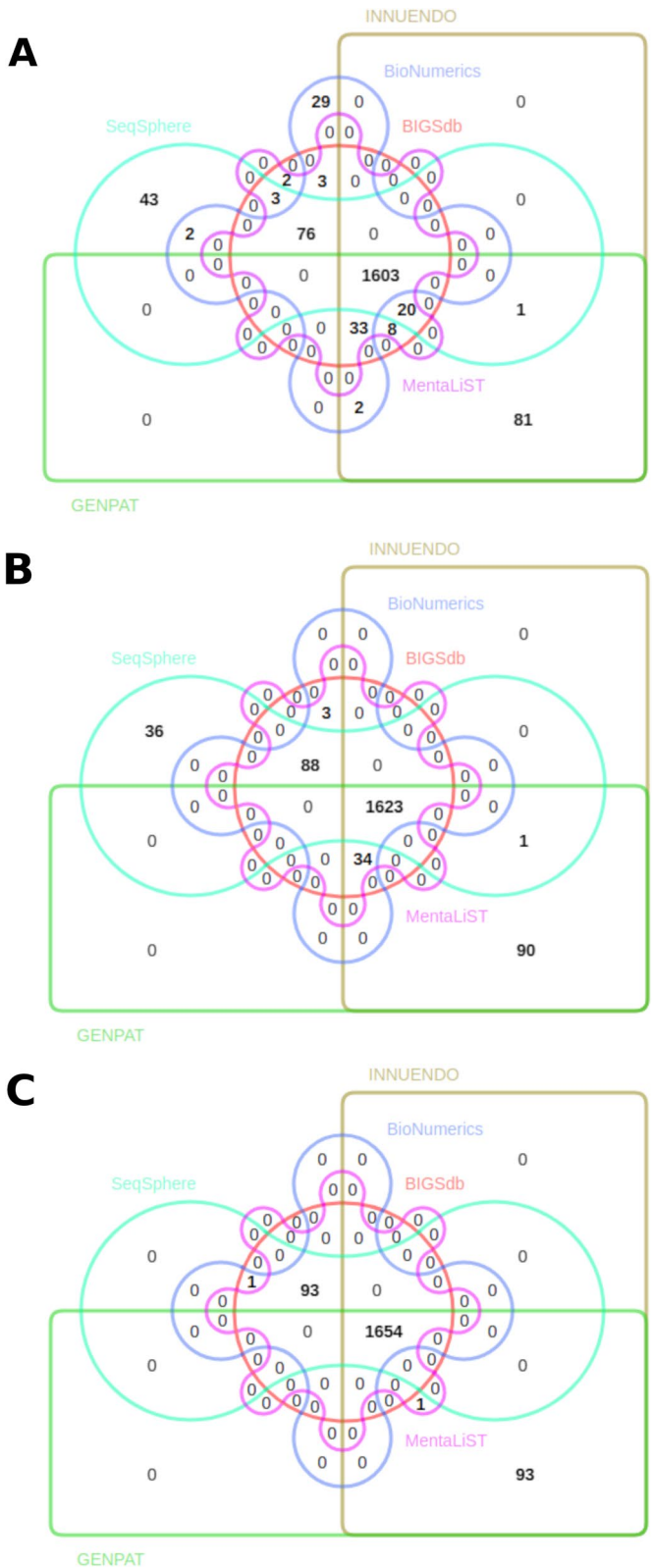

Fig. 2 Edward's Venn diagrams representing the identical alleles between the cgMLST workflows BIGSdb, INNUENDO, GENPAT, SeqSphere, BioNumerics and MentaLiST for the *Listeria monocytogenes* reference circular genomes ATCC19114 (A), ATCC19115 (B) and ATCCBAA679 (C)

Table 1 Mean and standard deviation of read depth of coverage estimated from BBMap (version February 13, 2020) or INNUca (version 4.2.2) with constant high read breadth of coverage (99.34 ± 0.07%) according to targeted read (Dr) and kmer (Dk) depth (X) from BBNorm downsampling (read length $R = 150$ and kmer size K = 30) of *Listeria monocytogenes* paired-end reads from tested reference genomes ATCC19114, ATCC19115 and ATCCBAA679 ($n = 420$)

Targeted depth of coverage	ATCC19114		ATCC19115		ATCCBAA679	
	BBMap	INNUca	BBMap	INNUca	BBMap	INNUca
Dr100-Dk75	101.6 ± 1.6	98.2 ± 1.5	101.9 ± 1.4	96.2 ± 1.8	101.0 ± 2.1	97.6 ± 2.6
Dr90-Dk68	91.9 ± 1.5	89.3 ± 1.7	92.3 ± 1.4	87.2 ± 2.3	92.0 ± 1.3	89.2 ± 2.1
Dr80-Dk60	80.9 ± 1.3	78.7 ± 1.8	81.3 ± 1.2	78.3 ± 1.7	81.4 ± 1.1	79.9 ± 2.2
Dr70-Dk53	71.4 ± 1.1	69.5 ± 2.0	71.7 ± 1.1	67.8 ± 1.6	72.0 ± 1.3	70.4 ± 1.7
Dr60-Dk45	60.5 ± 0.9	58.7 ± 2.0	60.7 ± 0.9	58.1 ± 1.5	61.1 ± 1.2	59.3 ± 2.2
Dr50-Dk38	50.9 ± 0.8	49.5 ± 1.2	51.1 ± 0.8	48.6 ± 1.6	51.5 ± 1.1	49.9 ± 2.0
Dr40-Dk31	41.3 ± 0.6	40.1 ± 1.0	41.5 ± 0.6	38.7 ± 1.3	41.9 ± 0.9	41.3 ± 1.7
Dr30-Dk23	30.7 ± 0.4	29.9 ± 1.6	30.8 ± 0.4	29.5 ± 1.1	31.1 ± 0.6	29.8 ± 1.1
Dr20-Dk16	21.5 ± 0.2	20.9 ± 0.9	21.5 ± 0.2	20.7 ± 0.9	21.7 ± 0.4	21.8 ± 0.8
Dr10-Dk8	10.9 ± 0.1	11.3 ± 0.1	10.9 ± 0.1	11.2 ± 0.1	11.0 ± 0.2	11.2 ± 0.2

for BioNumerics (Additional file 5F), while poorly correlated linearly with the other parameters for BIGSdb (Additional file 5B: $p > 9.1 \times 10^{-1}$), INNUENDO (Additional file 5C: $p > 9.8 \times 10^{-1}$), GENPAT (Additional file 5D: $p > 9.4 \times 10^{-1}$) and SeqSphere (Additional file 5E: $p > 9.3 \times 10^{-1}$).

Graphically confirmations

The graphical representation in four-way figures were built including IAAS (A, B, C, D) or IAAR at extended (E, F, G, H) or restricted (I, J, K, L) scales, according to reference genomes (A, E, I), successive platings (B, F, J), DNA extraction replicate (C, G, K) and sequencing replicate (C, H, L) (Additional file 6). These four-way figures clearly showed that IAAS (i.e. completeness) were impacted by tested reference genomes and cgMLST workflows (Additional file 6A) but not by in vitro parameters (Additional file 6B-Additional file 6D). In fact, BioNumerics profiles showed the higher number of unidentified alleles (38 over 1748 loci) for ATCC19114 compared to the other workflows (5 over 1748) (Additional file 7). For ATCC19115 and ATCCBAA679, INNUENDO and GENPAT showed 3 unidentified alleles over 1748 loci (Additional file 7), while the other workflows identified all the loci of the schema. IAAR (i.e. precision) is impacted by DrDk, cgMLST workflows and tested reference genomes (Fig. 5). More precisely, IAAR of BioNumerics and MentaLiST sharply dropped down at depth of coverage of ~30X (up to 1686) and ~40X (up to 1614), respectively. While INNUENDO showed almost 100% of identical allele calls at ≥30X (as it filters out reads at ≤25X), BIGSdb, GENPAT and SeqSphere called almost 100% of IAAR at lower depth of coverage (≥20X) (Fig. 5). At this depth of coverage (~ 20X) the number of misidentified alleles against reference (MIAAR) of BioNumerics and MentaLiST was remarkably higher (i.e. >7) compared to BIGSdb, GENPAT and SeqSphere that showed similar misidentified alleles against reference (MIAAR) only at ~10X coverage (Fig. 6). As reported in Table 3, all workflows reached ~100% precision at ≥40X depth of coverage excepted BioNumerics with ~98%.

Clustering of cgMLST profiles

Minimum spanning tree (MST)-based clustering showed that the minimum depth of coverage of 40X consistently grouped the cgMLST profiles from each reference genomes into clusters with up to 7 pairwise allele differences (Fig. 7A-Fig. 7F). Below 40X, cluster discrepancies were identified for each cgMLST workflows (Additional file 8A-Additional file 8F). The major increase of pairwise allele differences according to decreasing of targeted depth was observed with

(See figure on next page.)

Fig. 3 Boxplot-based distributions of targeted read (Dr) and kmer (Dk) depth (X) from BBNorm downsampling (read length $R = 150$ and kmer size K = 30) of *Listeria monocytogenes* paired-end reads from reference genomes ATCC19114, ATCC19115 and ATCCBAA679 ($n = 420$) according to estimated read depth (X) from BBMap (version February 13, 2020) or INNUca (version 4.2.2) with constant high read breadth of coverage (99.34 ± 0.07%) (A) and linear correlations between read depth of downsampled paired-end reads ($n = 420$) estimated with BBMap or INNUca for each reference genome (B)

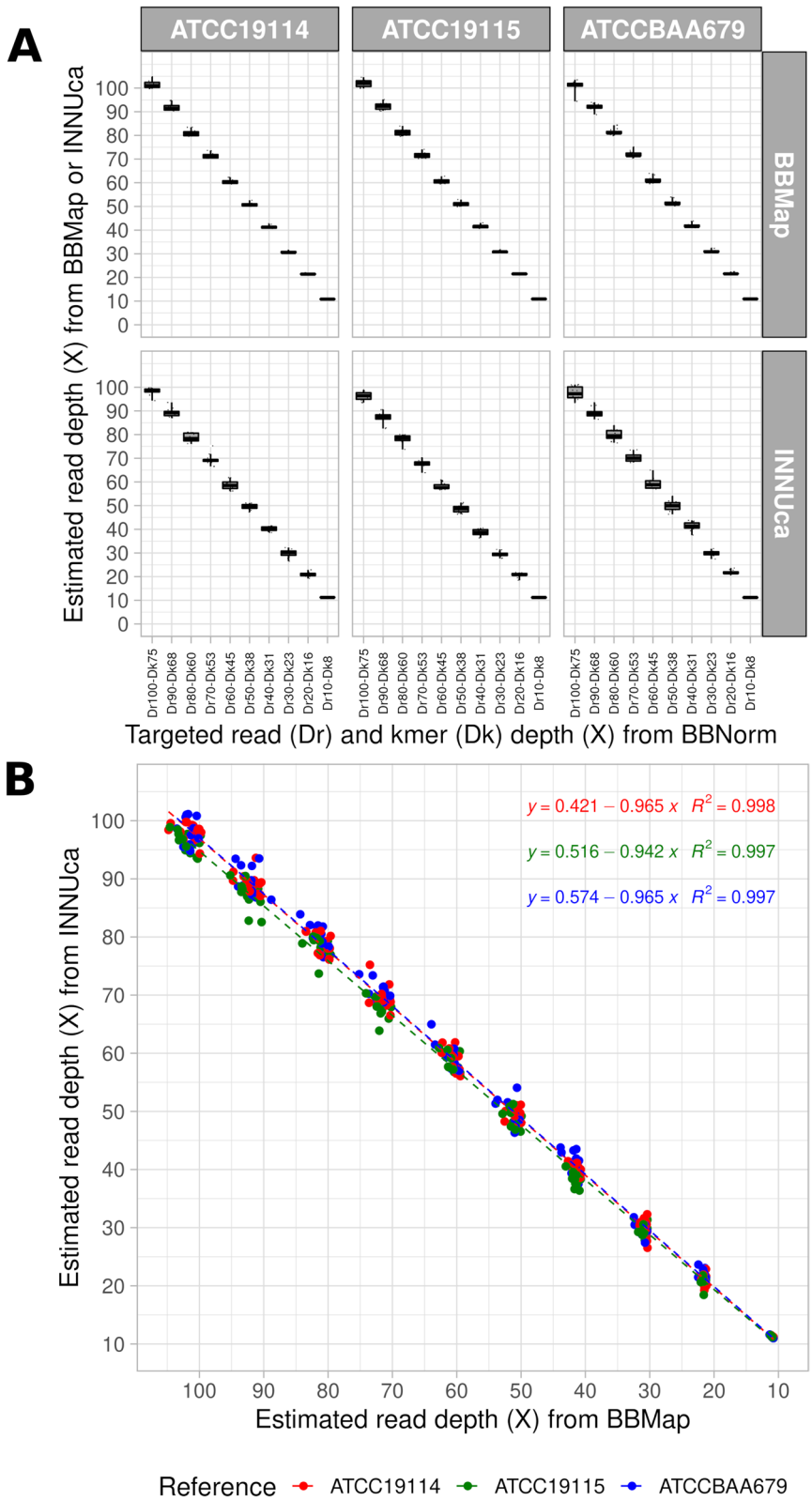

Fig. 3 (See legend on previous page.)

MentaLiST (Additional file 8F), and to a lesser extent with BIGSdb (Additional file 8A), INNUENDO (Additional file 8B), GENPAT (Additional file 8C), SeqSphere (Additional file 8D) and BioNumerics (Additional file 8E). The effect of downsampling on MST-clustering was observed at 10X depth of coverage for all workflows (Additional file 8A-Additional file 8F) excepted MentaLiST that poorly clustered profiles from reads downsampled at ≤40X.

Discussion

Internationally accepted validation of cgMLST typing workflows contributes to enhance routine surveillance of bacterial pathogens [21] by promoting the application of standards and benchmarking data sets [17]. Here, we focused on cgMLST precision and completeness between workflows rather than overall accuracy as the latter would refer to the ability to call the "right" alleles based on commonly assumed reference alleles. The comparison between different cgMLST workflows based on accuracy is hampered by the absence of a common strategy for definition of alleles, due to cgMLST approaches (i.e. assembly-based [12, 14–17] or -free [13, 18, 19], or combination of both [20]), as well as implemented algorithmic steps and related parameters (e.g. BLAST-based or -free algorithms, BLASTN or BLASTP, detection of open reading frames (ORFs) before BLAST step, coverage and identity of aligned sequences [12–20]).

Allele differences between cgMLST workflows

In the present study, we did not assess the cgMLST precision with schemes presenting missing alleles because it would have decreased the completeness and precision of all cgMLST workflows, while minimizing differences of completeness and precision observed between these workflows. The allele differences observed between cgMLST workflows (Fig. 2) are induced by algorithmic differences of the definition of alleles and reflect the impossibility of direct comparisons of cgMLST profiles generated by different workflows (i.e. accuracy), delaying the multi-centers surveillance of strain variants. In the present study, the cgMLST allele calling of six workflows was assessed, using the 1748-loci *L. monocytogenes* schema [24]. All workflows successfully detected ~100% loci of the schema in the reference circular genomes with up to ~95% of common alleles showing exact match with alleles from the schema (Fig. 2). Overall, the main differences resulting from different profiles were either alleles uniquely found in a workflow, up to ~2% for SeqSphere or BioNumerics, or due to a different allele calling strategy, up to ~5% for INNUENDO and GENPAT (i.e. chewBBACA allele caller) (Fig. 2). While BIGSdb found an exact match in the reference schema for each allele as expected (because the schema was built based on this workflow), other cgMLST workflows (e.g. MentaLiST, INNUENDO and GENPAT) inferred new alleles (not presented in the reference schema) based on the implemented algorithms. These divergences hamper the comparison of profiles generated using different workflows, even when using a common scheme aiming at supporting interoperability of genomic data [44].

In vitro parameters and cgMLST precision

Our overall results showed that in vitro parameters such as successive platings, replicates of DNA extraction and sequencing did not impact cgMLST precision, demonstrating that these wet-lab steps are very reproducible ($p > 1.0 \times 10^{-3}$). Indeed, the improvements during several years of documentation, validation, quality check and quality monitoring of wet-lab steps, from growth of isolates to sequencing through DNA extraction and library preparation, allowed to obtain nowadays a stable and repeatable wet-lab process [45].

In silico parameters and cgMLST precision

In contrast to the absence of effect from wet-lab parameters, the depth and breadth of coverages, as well as cgMLST workflows, tested reference strains and completeness (i.e. IAAS), were the main factors explaining cgMLST precision (i.e. IAAR), based on PCAs, GLMs and graphical confirmations. Indeed, the incapability to call alleles against schema (i.e. IAAS) impacts directly the number of identical alleles against reference genomes (i.e. IAAR), and consequently cgMLST precision (IAAR linearly correlated with IAAS; $p = 3.7 \times 10^{-6}$). This underlines the necessity to keep cgMLST schemes regularly updated through synchronized systems (e.g. BIGSdb-*Lm* [24, 46] and chewieNS [47]). Recently proposed Hash-based nomenclature servers may circumvent the need of schema synchronization, and likely facilitate interlaboratories data comparability and sharing when confidentiality concerns apply (chewieSnake [48]). In terms of precision, we would expect chewieSnake having the same outputs

(See figure on next page.)

Fig. 4 Principals component analyses (PCAs) of the numerical parameters IAAR, IAAS, DEPTH and BREADTH (defined in the section abbreviations) according to the categorical parameters "reference genome" (A), "successive platings" (B), "DNA extraction replicate" (C), "sequencing replicate" (D), "targeted depth" (E), " cgMLST workflows" (F) including BIGSdb ($n = 420$), INNUENDO (n = 336), GENPAT ($n = 420$), SeqSphere ($n = 420$), BioNumerics ($n = 420$) and MentaLiST ($n = 420$) applied to downsampled paired-end reads from 3 reference genomes of *Listeria monocytogenes* (i.e. ATCC19114, ATCC19115 and ATCCBAA679). The PCA outcomes from the workflows BIGSdb and SeqSphere are overlapped

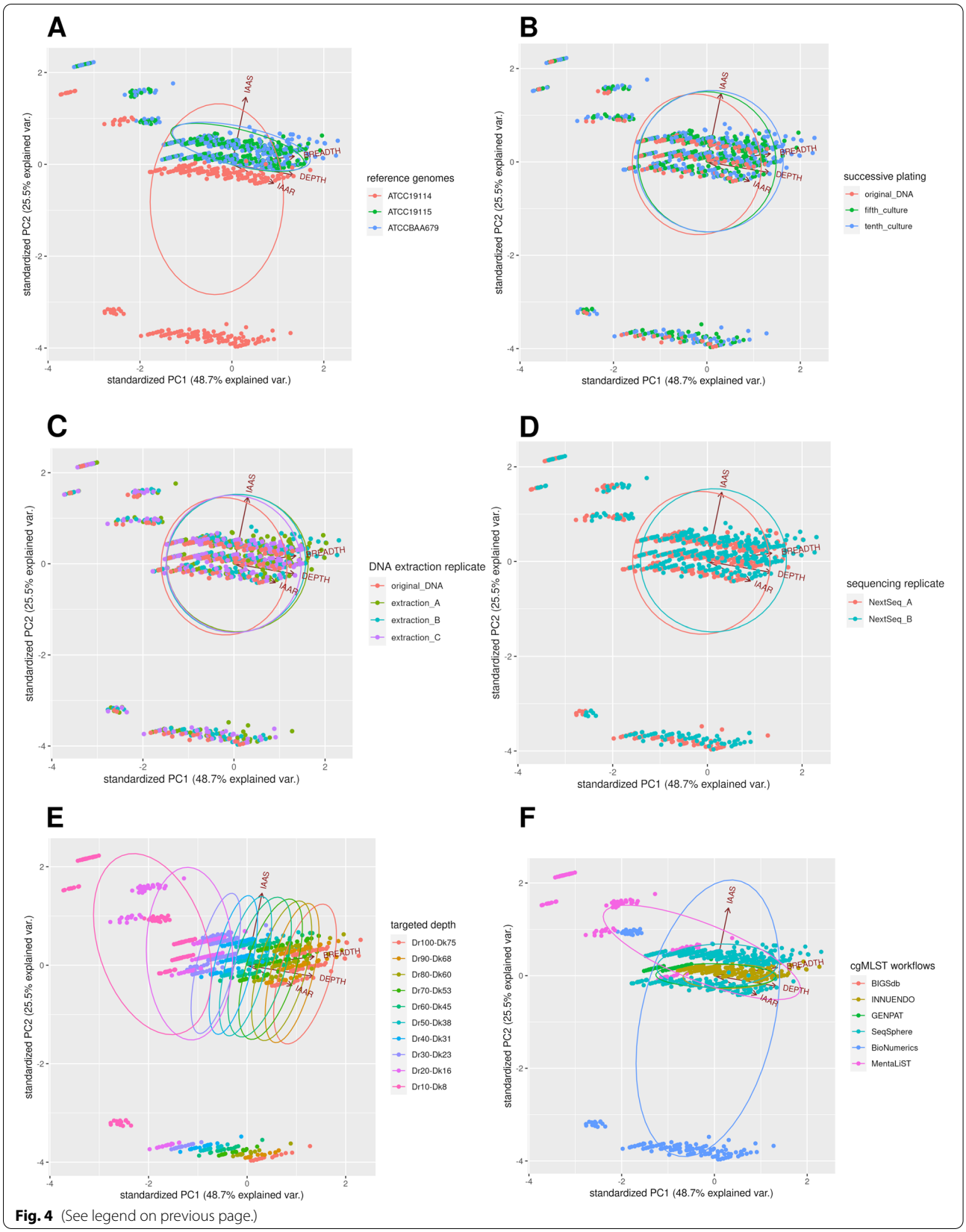

Fig. 4 (See legend on previous page.)

Table 2 Coefficients (Coef.) of the generalized linear model (GLMs with quasi Poisson distribution and with overdispersion) comparing the parameters "identical alleles against circular reference genomes" (IAAR) with the parameters of interest "tested reference genomes" (REFERENCE), "successive platings" (PLATING) (B), "DNA extraction replicate" (DNA), "sequencing replicate" (SEQUENCING), read depth (DEPTH), read breadth (BREADTH), identified alleles against schema (IAAS) and cgMLST workflows (WORKFLOW) including BIGSdb ($n=420$), INNUENDO ($n=336$), GENPAT ($n=420$), SeqSphere ($n=420$), BioNumerics ($n=420$) and MentaLiST ($n=420$) applied to downsampled paired-end reads from 3 tested reference genomes of *Listeria monocytogenes* (i.e. ATCC19114, ATCC19115 and ATCCBAA679). Few parameters are not defined because of singularities

Parameters	Coef. estimate	Coef. standard error	Coef. t value	Coef. *P*-value(>|t|)
WORKFLOW: MentaLiST	-1.9E-01	1.1E-02	-1.7E+01	2.0E-16
BREADTH	5.6E-01	8.1E-02	6.9E+00	5.3E-12
DEPTH	8.1E-04	1.9E-04	4.3E+00	1.5E-05
REFERENCE: ATCCBAA679	−3.8E-02	1.1E-02	−3.6E+00	3.5E-04
SEQUENCING: NextSeq_B	−2.3E-02	7.0E-03	−3.2E+00	1.3E-03
REFERENCE: ATCC19115	−3.0E-02	1.0E-02	−3.0E+00	2.9E-03
WORKFLOW: BioNumerics	−2.9E-02	1.2E-02	−2.4E+00	1.7E-02
PLATING: tenth_culture	−2.4E-02	1.1E-02	−2.1E+00	3.3E-02
WORKFLOW: INNUENDO	−2.0E-02	1.1E-02	−1.8E+00	7.5E-02
PLATING: fifth_culture	−1.8E-02	1.1E-02	−1.6E+00	1.1E-01
IAAS	6.3E-04	5.1E-04	1.2E+00	2.2E-01
DNA: extraction_A	−7.4E-03	8.4E-03	−8.9E-01	3.7E-01
DNA: extraction_B	−4.5E-03	8.3E-03	−5.4E-01	5.9E-01
WORKFLOW: GENPAT	1.7E-04	1.1E-02	1.6E-02	9.9E-01
WORKFLOW: SeqSphere	4.5E-05	1.1E-02	4.3E-03	1.0E+00
Model Intercept	**−4.9E+01**	**8.1E+00**	**−6.1E+00**	**1.1E-09**

than the workflow chewBBACA as both allele callers are based on the chewBBACA suite. However, such decentralized and nomenclature-free approach requires further developments to be integrated in global surveillance systems where common language and genotypes naming are essential. Indeed, when using the reference threshold of 7 pairwise allele differences, commonly used for WGS-based surveillance of *L. monocytogenes* to define clusters of isolates likely sharing an epidemiological link, the negative effect of incomparable profiles from different workflows became negligible, with all workflows leading to the same clusters when read depth of coverage was ≥40X (Fig. 7 and Additional file 8). These findings are consistent with previous studies on viruses [49] and bacteria [50], that did not observe improvement of the breadth of coverage above specific values of depth of coverage. Few studies recommended minimal depth of coverage for precise cgMLST typing of *L. monocytogenes* (40X with BIGSdb) [24, 28, 36], *Yersinia* (50X with BIGSdb) [42], *Mycoplasma* (47X with SeqSphere) [38], *Campylobacter*, *Chlamydia*, *Neisseria* and *Streptococcus* (20X with STing) [13]. For the first time in the present study, we recommend 40X as a suitable read depth of coverage for the highest cgMLST precision across 6 different assembly-based and -free workflows. This recommendation of minimal depth of coverage for precise cgMLST typing has been defined based on Illumina short reads (i.e. NextSeq) sequencing. Other short reads (IonTorrent) and long reads (PacBio SMRT and Oxford Nanopore) sequencing technologies may require higher depth of coverage than Illumina to reach similar quality of base

(See figure on next page.)

Fig. 5 Box-plots representing the impact of downsampled paired-end reads (i.e. 2x150bp) of *Listeria monocytogenes* on identical alleles against reference at extended (A) or restricted (B) scales, according to reference genomes (i.e. ATCC19114, ATCC19115 and ATCCBAA679) and cgMLST workflows including BIGSdb ($n=420$), INNUENDO ($n=336$), GENPAT ($n=420$), SeqSphere ($n=420$), BioNumerics ($n=420$) and MentaLiST ($n=420$). The targeted read depth (Dr: 10X, 20X, 30X, 40X, 50X, 60X, 70X, 80X, 90X and 100X) were prepared according to kmer depth (Dk): 8X, 15X, 23X, 30X, 38X, 45X, 52X, 60X, 67X, 75X) setting of BBNorm (read length $R=150$ and kmer size $K=30$). Because of internal firewall, the INNUca assembler integrated into the cgMLST workflow INNUENDO cannot not perform assemblies of paired-end reads with read depth of coverage of 20X ($n=42$) and 10X ($n=42$)

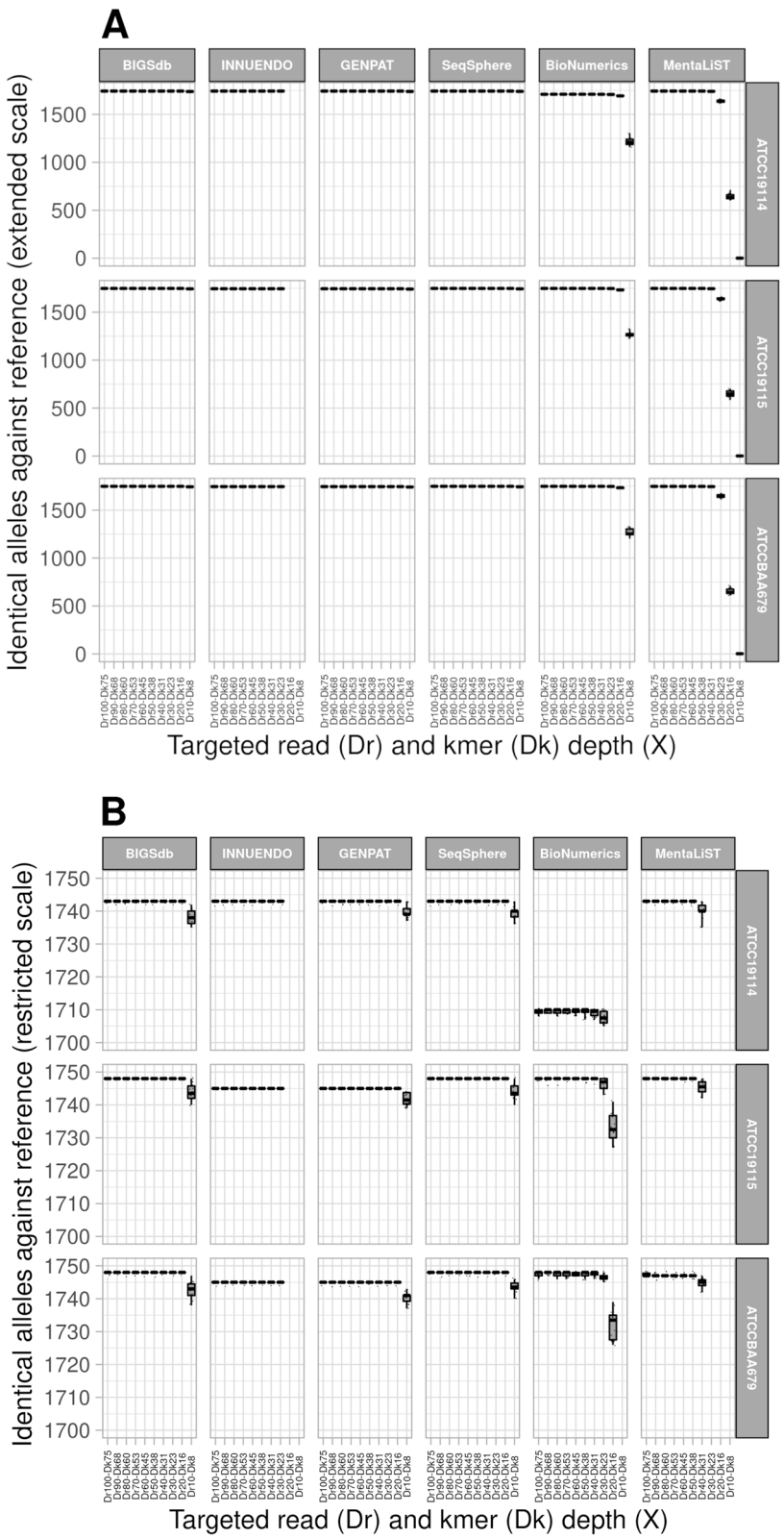

Fig. 5 (See legend on previous page.)

calling, independently of GC-content and repeated region biases which are inherent in sequencing technologies based on short reads [51–53].

Performances of assembly-based and -free cgMLST workflows

Among the 6 compared cgMLST workflows, BIGSdb, INNUENDO, GENPAT and SeqSphere did not show obvious effect of depth of coverage on precision contrary to BioNumerics and MentaLiST (Table 3). In particular, BIGSdb and SeqSphere performed well also at very low coverage values. This is probably due to refinement steps of assembly pipelines used for the workflow BIGSdb (i.e. fq2dna) and SeqSphere (i.e. average quality >30 with a window of 20 bases), as well as similar allele definitions between BIGSdb and SeqSphere (i.e. BLASTN; nucleotide identity >70%; coverage >70%). In contrast, BioNumerics and MentaLiST were poorly precise for depth of coverage ≤30X and 40X (Fig. 6B) according to PCAs (Fig. 4) and GLMs (Table 2). Differences of precision between the cgMLST workflows are consequently induced by their respective de novo assemblers and/or allele callers. Even though MentaLiST requires more reads to achieve adequate precision compared to the assembly-based workflows, its precision is slightly impacted by tested reference genomes for high read depth of coverage (i.e. ≤ 30X and 40X) (Fig. 6B). This result highlights that MentaLiST precision is overall less impacted by in vitro and in silico parameters compared to assembly-based workflows, whose precision also depend on de novo assembly. Further comparisons with other assembly-free cgMLST workflows would confirm the supposed absence of strain effect on precision observed with MentaLiST [54]. However, MentaLiST outperformed other workflows in terms of percentage of correct allele predictions for cgMLST in a recent benchmarking of different assembly-free approaches [13]. Here we observed that both assembly-free and -based cgMLST workflows reach ~100% of identical allele predicted in the processed reads with coverage ≥40X compared to reference circular genomes.

Performances of assembly-based cgMLST workflows

The decrease of cgMLST precision from assembly-based workflows may reflect the fragmentation of de novo assembly potentially induced by the GC bias [55] and/or and repetitive regions [56] (Additional file 4). This was particularly evident for BioNumerics workflow where the decreasing of cgMLST precision (i.e. IAAR) was linearly correlated with high amount of N100 through GLMs ($p = 2.0 \times 10^{-16}$). This is probably induced by the absence of assembly refinement steps and/or an old version of SPAdes implemented in BioNumerics, in comparison with the other workflows (Table 4) [57]. In this study, no linear correlations between cgMLST precision and GC%, or cgMLST precision and duplication ratio were identified. Nevertheless, significant differences were observed (Wilcoxon rank sum tests: $p < 2.2 \times 10^{-16}$) between GC% of references genomes draft assemblies (38.081 ± 0.007% for ATCC19114, 37.879 ± 0.006% for ATCC19115 and 37.865 ± 0.006% for ATCCBAA679), while duplication ratios were not significantly different (Wilcoxon rank sum tests: $p > 1.5 \times 10^{-2}$) between these references genomes draft assemblies (1.0001 ± 0.0003 for ATCC19114, 1.00018 ± 0.0003 for ATCC19115 and 1.0000 ± 0.0008 for ATCCBAA679). Other statistical approaches would be necessary to test non-linear correlations [63, 64] between cgMLST precision and assembly metrics.

Future analytical prospects

The analytical approach (Fig. 1) here applied to *L. monocytogenes* can be easily fine-tuned for the analysis other bacterial species and taxa, assuming that a species-specific cgMLST scheme is established.

In the present study, the read depth of coverage was identified as one of the most impactful parameters on cgMLST precision. We thus proposed a minimal read depth of coverage of 40X for precise cgMLST typing and consistent MST clustering. We did not assess an upper limit of read depth but we showed that increasing the sequencing depth up to 100X did not effectively improve cgMLST precision. Sequencing at very high depth of coverage may promote errors on the assembly graph and confuse error correction algorithms, in addition to increase the computational burden [65]. Further studies may be needed to assess precision at higher coverage, yet 100X is enough high for *L. monocytogenes* cgMLST typing. Indeed, bacterial genomes sequences deposited in public databases (e.g. RefSeq, independently of the considered assembly surveillance project) are mostly

(See figure on next page.)

Fig. 6 Box-plots representing the impact of downsampled paired-end reads (i.e. 2x150bp) of *Listeria monocytogenes* on misidentified alleles against reference at extended (A) or restricted (B) scales, according to reference genomes (i.e. ATCC19114, ATCC19115 and ATCCBAA679) and cgMLST workflows including BIGSdb ($n = 420$), INNUENDO ($n = 336$), GENPAT ($n = 420$), SeqSphere ($n = 420$), BioNumerics ($n = 420$) and MentaLiST ($n = 420$). The targeted read depth (Dr: 10X, 20X, 30X, 40X, 50X, 60X, 70X, 80X, 90X and 100X) were prepared according to kmer depth (Dk): 8X, 15X, 23X, 30X, 38X, 45X, 52X, 60X, 67X, 75X) setting of BBNorm (read length $R = 150$ and kmer size $K = 30$). Because of internal firewall, the INNUca assembler integrated into the cgMLST workflow INNUENDO cannot not perform assemblies of paired-end reads with read depth of coverage of 20X ($n = 42$) and 10X ($n = 42$)

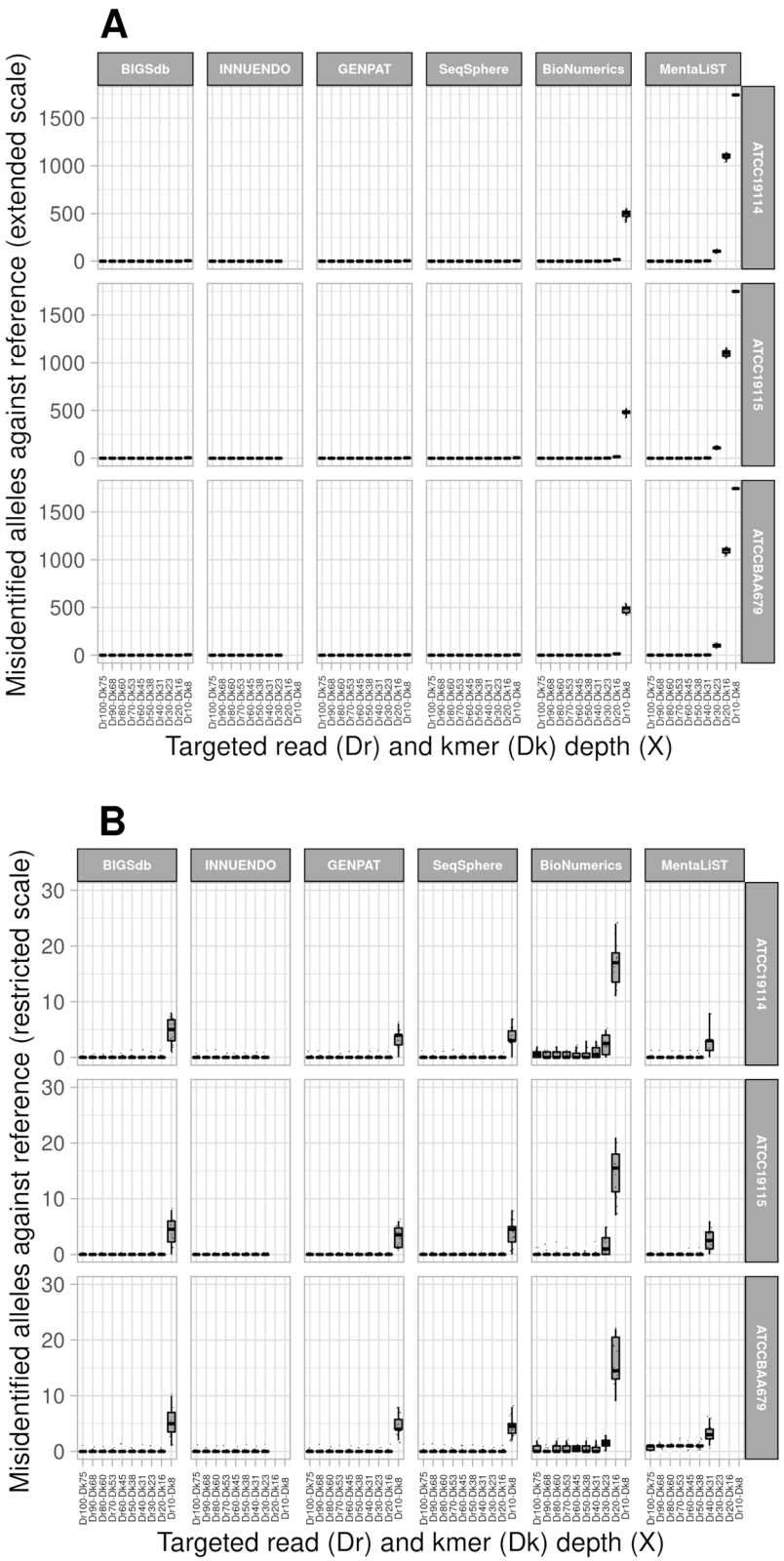

Fig. 6 (See legend on previous page.)

Table 3 cgMLST precision (i.e mean percentage ± standard deviation) of the workflows BIGSdb ($n=420$), INNUENDO ($n=336$), GENPAT ($n=420$), SeqSphere ($n=420$), BioNumerics ($n=420$) and MentaLiST ($n=420$) according to targeted read (Dr) and kmer (Dk) depth (X) from BBNorm downsampling (read length $R=150$ and kmer size K = 30) of *Listeria monocytogenes* paired-end reads from reference genomes ATCC19114, ATCC19115 and ATCCBAA679. The cgMLST schema harbors 1748 loci. NA means not applicable: Because of internal firewall, the INNUca assembler integrated into the cgMLST workflow INNUENDO cannot not perform assemblies of paired-end reads with read depth of coverage of 20X ($n=42$) and 10X ($n=42$)

Reference	Targeted depth of coverage	BIGSdb	INNUENDO	GENPAT	SeqSphere	BioNumerics	MentaLiST
ATCC19114	**Dr100-Dk75**	99.71 ± 0.02	99.71 ± 0.02	99.71 ± 0.02	99.71 ± 0.02	97.79 ± 0.04	99.71 ± 0.02
	Dr90-Dk68	99.71 ± 0.02	99.71 ± 0.02	99.71 ± 0.02	99.71 ± 0.02	97.80 ± 0.03	99.71 ± 0.02
	Dr80-Dk60	99.71 ± 0.02	99.71 ± 0.02	99.71 ± 0.02	99.71 ± 0.02	97.80 ± 0.04	99.71 ± 0.02
	Dr70-Dk53	99.71 ± 0.02	99.71 ± 0.02	99.71 ± 0.02	99.71 ± 0.02	97.80 ± 0.03	99.71 ± 0.02
	Dr60-Dk45	99.71 ± 0.02	99.71 ± 0.02	99.71 ± 0.02	99.71 ± 0.02	97.80 ± 0.04	99.71 ± 0.02
	Dr50-Dk38	99.71 ± 0.02	99.71 ± 0.02	99.71 ± 0.02	99.71 ± 0.02	97.79 ± 0.06	99.70 ± 0.02
	Dr40-Dk31	99.71 ± 0.02	99.71 ± 0.02	99.71 ± 0.02	99.71 ± 0.02	97.78 ± 0.06	99.54 ± 0.12
	Dr30-Dk23	99.71 ± 0.02	99.71 ± 0.02	99.71 ± 0.02	99.71 ± 0.02	97.69 ± 0.10	93.77 ± 0.82
	Dr20-Dk16	99.71 ± 0.02	NA	99.71 ± 0.02	99.71 ± 0.02	96.87 ± 0.22	36.67 ± 1.70
	Dr10-Dk8	99.44 ± 0.13	NA	99.51 ± 0.02	99.51 ± 0.09	69.63 ± 2.34	0.07 ± 0.07
ATCC19115	**Dr100-Dk75**	100 ± 0.00	99.83 ± 0.00	99.83 ± 0.00	100 ± 0.00	99.99 ± 0.02	100 ± 0.00
	Dr90-Dk68	100 ± 0.00	99.83 ± 0.00	99.83 ± 0.00	100 ± 0.00	99.99 ± 0.03	100 ± 0.00
	Dr80-Dk60	100 ± 0.00	99.83 ± 0.00	99.83 ± 0.00	100 ± 0.00	99.99 ± 0.03	100 ± 0.00
	Dr70-Dk53	100 ± 0.00	99.83 ± 0.00	99.83 ± 0.00	100 ± 0.00	100 ± 0.02	100 ± 0.00
	Dr60-Dk45	100 ± 0.00	99.83 ± 0.00	99.83 ± 0.00	100 ± 0.00	100 ± 0.00	100 ± 0.02
	Dr50-Dk38	100 ± 0.00	99.83 ± 0.00	99.83 ± 0.00	100 ± 0.00	99.99 ± 0.02	100 ± 0.02
	Dr40-Dk31	100 ± 0.00	99.83 ± 0.00	99.83 ± 0.00	100 ± 0.00	99.99 ± 0.03	99.85 ± 0.11
	Dr30-Dk23	100 ± 0.00	99.83 ± 0.00	99.83 ± 0.00	100 ± 0.00	99.91 ± 0.09	93.76 ± 0.84
	Dr20-Dk16	100 ± 0.00	NA	99.83 ± 0.00	100 ± 0.00	99.15 ± 0.26	37.04 ± 2.01
	Dr10-Dk8	99.77 ± 0.14	NA	99.64 ± 0.11	99.78 ± 0.13	72.41 ± 1.40	0.09 ± 0.05
ATCCBAA679	**Dr100-Dk75**	100 ± 0.02	99.82 ± 0.02	99.82 ± 0.02	100 ± 0.02	99.98 ± 0.04	99.96 ± 0.03
	Dr90-Dk68	100 ± 0.02	99.82 ± 0.02	99.82 ± 0.02	100 ± 0.02	99.99 ± 0.02	99.95 ± 0.03
	Dr80-Dk60	100 ± 0.02	99.82 ± 0.02	99.82 ± 0.02	100 ± 0.02	99.98 ± 0.04	99.95 ± 0.02
	Dr70-Dk53	100 ± 0.02	99.82 ± 0.02	99.82 ± 0.02	100 ± 0.02	99.97 ± 0.04	99.94 ± 0.03
	Dr60-Dk45	100 ± 0.02	99.82 ± 0.02	99.82 ± 0.02	99.99 ± 0.02	99.97 ± 0.03	99.95 ± 0.03
	Dr50-Dk38	100 ± 0.02	99.82 ± 0.02	99.82 ± 0.02	100 ± 0.02	99.97 ± 0.04	99.94 ± 0.04
	Dr40-Dk31	100 ± 0.02	99.82 ± 0.02	99.82 ± 0.02	100 ± 0.02	99.98 ± 0.04	99.81 ± 0.08
	Dr30-Dk23	100 ± 0.02	99.82 ± 0.02	99.82 ± 0.02	100 ± 0.02	99.92 ± 0.05	94.19 ± 1.02
	Dr20-Dk16	100 ± 0.02	NA	99.82 ± 0.02	100 ± 0.02	99.09 ± 0.25	37.18 ± 1.77
	Dr10-Dk8	99.69 ± 0.14	NA	99.56 ± 0.10	99.75 ± 0.10	72.49 ± 2.33	0.13 ± 0.06

generated at ≤100X sequencing depth (range: 30-150X) [66].

Our main goal here was to provide guidance concerning the "standalone" solutions that can be adopted today for assembly and allele calling following developers' recommendations. Our results suggests that the assembly pipelines may impact the cgMLST precision to a greater extent than the allele calling pipelines. This hypothesis should be further confirmed assessing the impact of allele callers on cgMLST precision pipeline. However, results

(See figure on next page.)

Fig. 7 Minimum spanning trees (MSTs) representing the impact on clustering of cgMLST workflows BIGSdb (A: n = 339), INNUENDO (B: $n=339$), GENPAT (C: $n=339$), SeqSphere (D: $n=339$), BioNumerics (E: $n=339$) and MentaLiST (F: $n=339$), of *Listeria monocytogenes* reference genomes (i.e. ATCC19114, ATCC19115 and ATCCBAA679) and targeted depth of coverage (Dr: 30X, 40X, 50X, 60X, 70X, 80X, 90X and 100X) prepared according to kmer depth (Dk): 8X, 15X, 23X, 30X, 38X, 45X, 52X, 60X, 67X, 75X) setting of BBNorm (read length $R=150$ and kmer size K = 30) from downsampled paired-end reads (i.e. 2x150bp). The MSTs were built with BioNumerics ignoring missing data. The MST clusters of at least two genomes, one node and allele differences ≤7, were highlighted in grey

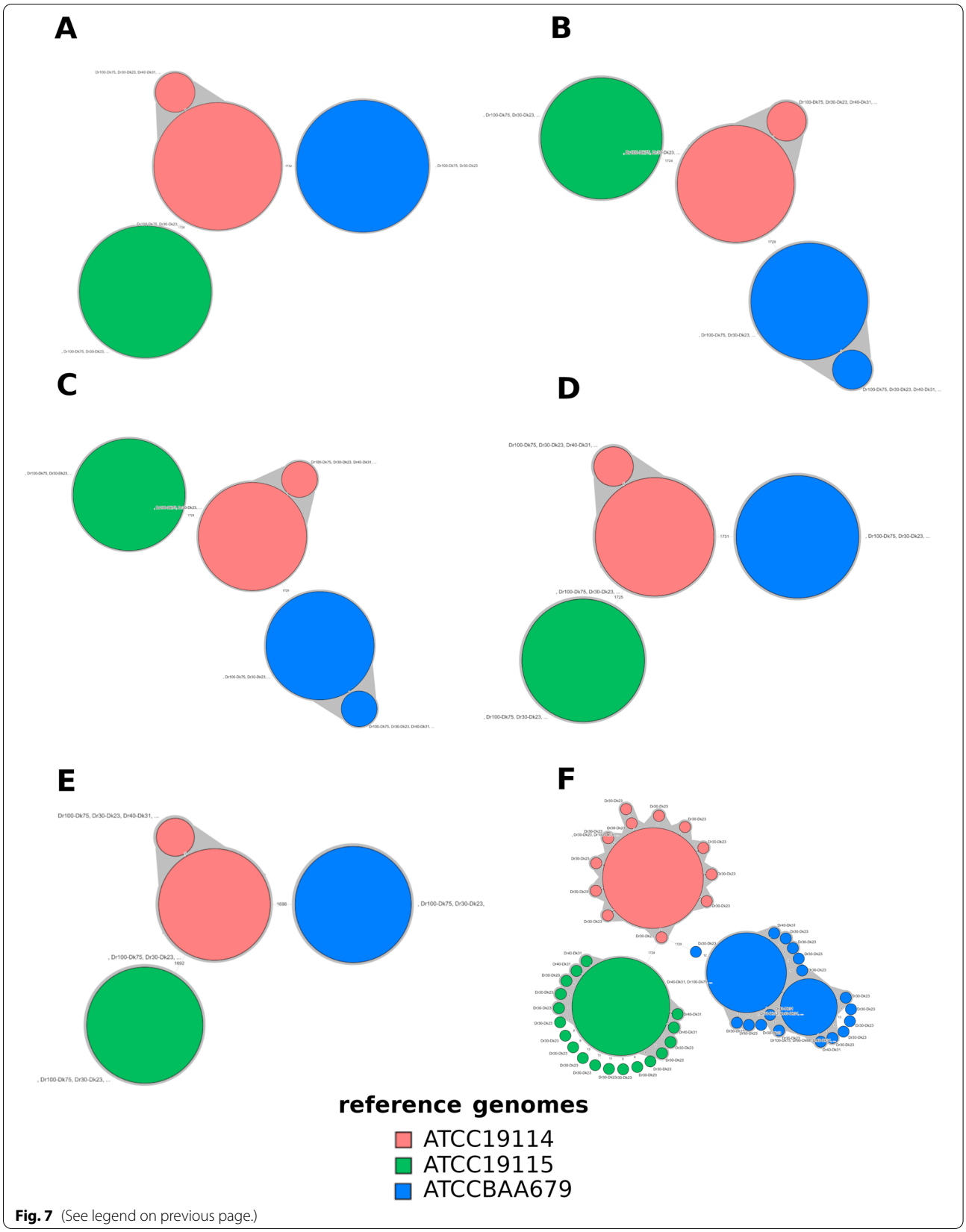

Fig. 7 (See legend on previous page.)

Table 4 License type, as well as de novo assembly and allele calling pipelines recommended by developers of cgMLST workflows compared in the present study to assess precision of *Listeria monocytogenes* cgMLST typing. N/A stands for not applicable

cgMLST workflow (version)	License type	Recommended assembly pipeline (version)	Recommended allele calling pipeline (strategy or version)	Reference
BIGSdb (N/A)	open source	AlienTrimmer (2.0)-, Musket (1.1)-and SPAdes (3.15.0)-based fq2dna (21.06)	BLASTN-based BIGSdb (alignment)	[24]
INNUENDO (N/A)	open source	Trimmomatic (0.36)-, Pilon (1.18)- and SPAdes (3.9.0)- based INNUca (4.2.2)	Prodigal- (ORF discovery) and BLASTP-based (alignment) chewBBACA (2.6.0)	[14, 49]
GENPAT (N/A)	open source	Trimmomatic (0.36)- and SPAdes (3.11.1)-based pipeline	Prodigal- (ORF discovery) and BLASTP-based (alignment) chewBBACA (2.6.0)	[14, 57, 58]
SeqSphere (6.0.2)	commercial	FastQC (0.11.7)- and SPAdes (3.11.1)-based pipeline	BLASTN-based SeqSphere (alignment)	[12, 59, 67]
Bionumerics (7.6.3)	commercial	SPAdes (3.7.1)-based pipeline	BLASTN-based assembly-based and -free algorithms (alignments)	[17, 20, 61, 62]
MentaLiST (1.0.0)	open source	N/A (i.e. assembly free)	stringMLST principle-based MentaLiST (kmer counting)	[13, 19]

from Lüth et al. (2021) showed a ~ 100% correlation between matrices of cgMLST profile distances providing identical *L. monocytogenes* assemblies to different allele callers (e.g. Ridom SeqSphere versus chewBBACA) [46].

To foster interoperability between the tested cgMLST solutions, the impact of different allele calling settings on cgMLST precision and nonidentical calls (i.e. missing data, partial alleles and new alleles) should also be investigated. In view of the main differences between the cgMLST allele calling algorithms, such studies should assess settings, such as BLASTN nucleotide identity, BLAST coverage, word size (i.e. BIGSdb, SeqSphere, BioNumerics), allele size threshold, minimum BLASTP score ratio (i.e. chewBBACA implemented in GENPAT and INNUENDO), mutation threshold and kmer threshold (i.e. MentaLiST).

The definition of new alleles is not centralized between allele calling pipelines. This inevitably leads to a drift of allele identifiers in the scheme adopted by each system and consequently hinders profiles' comparability and communication on *L. monocytogenes* genotypes across laboratories. A common effort of developers, curators and users of such cgMLST systems will allow the implementation of novel functionalities (e.g. application programming interfaces, nomenclature mapping) to ensure that an universal language is adopted by the scientific community.

Conclusion

cgMLST precision was mainly impacted by the tested reference strains, cgMLST workflows, cgMLST completeness, as well as depth and breadth of coverage. Successive platings, DNA extraction and sequencing replicates did not show an impact on cgMLST precision. Overall loci detection was > 99% for assembly-free and assembly-based workflows and had no impact on cluster definitions, for read depth of coverage ≥40X. This study highlights the importance of high sequencing depth to ensure reproducibility of profiles in genomic surveillance and outbreak investigations.

Material and methods

After a review about the cgMLST principles and approaches, the experimental plan, cgMLST workflows of interest, statistical analyses and confirmations of relevant parameters are presented successively.

Review about cgMLST principles and approaches

The MLST method aims at assigning arbitrary numbers to each allele of a small set of DNA fragments from different loci (typically < 10 gene fragments with ~ 500 bp) presenting up- and downstream conserved sites for hybridization of forward and reverse oligonucleotides during PCR amplifications of housekeeping genes of interest [2]. The combination of these MLST allele numbers from a single strain allows assignment of a MLST sequence type (ST) already shared between laboratories or a new one [67]. The cgMLST is an extension of the MLST principle allowing screening of alleles from several hundreds of core genes. More precisely, after steps related to potential read trimming (usually with Trimmomatic [58]) and mandatory de novo assembly (usually with SPAdes [57]), the assembly-based cgMLST workflows include (i.e. chewBBACA [14]) or not (i.e. SeqSphere^{+} [12], MLSTar [15], BIGSdb-Pasteur [16], BioNumerics [17]) a step to detect open reading frames (ORFs) from drafts de novo assembly (i.e. Prodigal [68] implemented in chewBBACA [14]). Then, these assembly-based cgMLST workflows aligne alleles from schema to sequences from drafts de novo assembly (ORFs or not)

based on the BLASTN (i.e. SeqSphere [12], MLSTar [15], BIGSdb-Pasteur [16], BioNumerics [17]) or BLASTP (i.e. chewBBACA [14]) algorithms [69], as well as different parameters related coverage and identity of aligned sequences. In addition, recently published assembly-free cgMLST workflows process reads independently of de novo assembly based on heuristic kmer mapping (i.e. KMA [18]) or counting and voting of kmers (MentaLiST [19] and STing [13]). Some cgMLST workflows may combine de novo assembly-free and -based allele calling (e.g. BioNumerics [20]). This review drove the selection of the 6 workflows of interest and related settings recommended by developers (BIGSdb, INNUENDO, GENPAT, SeqSphere, BioNumerics and MentaLiST), in order to cover the different genomics-based cgMLST typing approaches (Table 4).

Experimental plan

The experimental plan was built to take into account a large range of in vitro and in silico parameters potentially explaining the cgMLST precision (i.e. identical alleles against reference circular genomes $\times 100$ / 1748). The in vitro parameters include the tested reference genomes, successive platings, as well as replicates of DNA extraction and sequencing. The in silico parameters include the targeted read/kmer depth of coverage, read depth of coverage, read breadth of coverage, assembly metrics, cgMLST workflows and identified alleles against schema. For the sake of clarity, acronyms of this large set of parameters were defined in the section abbreviations.

In vitro parameters of interest

Three *L. monocytogenes* strains and three original genomic DNA (gDNA) were obtained from the American Type Culture Collection Global Bioresource Center (ATCC: https://www.atcc.org): ATCC 19114 (cgMLST type L3-SL69-ST201-CT996, serotype 4a), ATCC 19115 (L1-SL2-ST145-CT375, serotype 4b) and ATCC BAA-679 (L2-SL9-ST35-CT637, serotype 1/2a), which corresponds to the reference EGD-e strain. The original gDNA of each of the three ATCC strains was sequenced in two different batches (i.e. $n=3\times2=6$ paired-end reads) (Fig. 1). The three ATCC strains were grown 5 and 10 times through successive plating (i.e. 4 and 9 platings, respectively), leading to two subcultures for each strain. Each of the subcultures was extracted three times, and each extract was then sequenced in two different batches ($n=3\times2\times3\times2=36$ paired-end reads) (Fig. 1). For bacterial culture, DNA was extracted using previously described procedures [70]. All gDNA samples were quantified by Qubit dsDNA HS Assay Kit using the Qubit fluorometer 2.0 (Thermo Fisher Scientific, Waltham, Massachusetts, United States). gDNA quality was estimated based on the Eppendorf BioSpectrometer® fluorescence (Eppendorf, Hamburg, Germany), whereas gDNA integrity was assessed using the Agilent 4200 TapeStation system (Agilent Technologies, Santa Clara, CA, United States). The sequencing libraries were prepared with 30 µl of Illumina DNA Prep kit and 100–500 ng of input gDNA. These libraries were sequenced with a NextSeq500 sequencer (Illumina). In total, a set of 42 paired-end reads ($n=3\times2+3\times2\times3\times2=42$ paired-end reads) were produced to assess the impacts on cgMLST precision of in vitro parameters of interest: tested reference genomes, successive platings, as well as replicates of DNA extraction and sequencing (Fig. 1).

In silico parameters of interest

The in silico parameters of interest include the targeted read/kmer depth of coverage, read depth of coverage, read breadth of coverage, assembly metrics, cgMLST workflows and identified alleles against schema. The number of reads, average Phred quality scores and percentages of Phred quality scores higher than 30 were checked for each 42 paired-end reads with FastQC (version 0.11.5) [79]. In addition, the absence of exogenous DNA contamination was confirmed with ConFindr (version 0.7.4) [41]. After quality assessment, downsampling of paired-end reads was performed with BBNorm (version February 13, 2020) in parallel with the estimation of depth and breadth of coverages of reads through BBMap-based mapping (version February 13, 2020) [42]. BBNorm-based downsampling was performed from paired-end reads (i.e. duplicated DNA samples of each 3 tested reference genomes) at 10 different kmer depth of coverage (Dk: 8X, 16X, 24X, 32X, 40X, 48X, 56X, 64X, 72X and 80X) fixing read length ($R=150$) and kmer size ($K=30$). Then, the corresponding read depth of coverage (Dr) measured with BBMap allowed estimation of the correlation with kmer depth of coverage (Dr = 1.3502 x Dk - 0.2923; R2 = 99.98%; $n=60$) based on the 'stats' R library [72]. After this standard curve building, the setting of kmer depth of coverage (Dk: 8X, 15X, 23X, 30X, 38X, 45X, 52X, 60X, 67X, 75X) during another BBNorm-based downsampling (i.e. argument 'target') allowed preparation of 420 paired-end reads with different read depth of coverage (Dr: 10X, 20X, 30X, 40X, 50X, 60X, 70X, 80X, 90X and 100X). Finally, the high read breadth of coverage and expected read depth of coverage (Dr) of the 420 prepared paired-end reads were double checked independently with BBMap [42], and the INNUca (version 4.2.2) [49] internal module based on Bowtie2 (version 2.2.9) [73] and Samtools (version 1.3.1) [74]. In total, 420 paired-end reads were produced to assess the impacts on cgMLST precision of in silico parameters (Fig. 1). Following de novo assembly steps recommended by developers

detailed below, the 420 paired-end reads were processed through the six cgMLST workflows of interest ($n = 6 \times 10$ x $[3 \times 2 + 3 \times 2 \times 3 \times 2] = 2520$ cgMLST results) (Fig. 1). Then, de novo assembly metrics of the assembly-based cgMLST workflows were assessed with Quast (version 5.0.2) [75] and combined with MultiQC (version 1.9) [76] ($n = 5 \times 10$ x $[3 \times 2 + 3 \times 2 \times 3 \times 2] = 2100$ quality results assessing 48 assembly metrics) (Fig. 1).

cgMLST workflows of interest

Six different cgMLST workflows were tested: BIGSdb [24], INNUENDO [14, 49], GENPAT [14, 57, 58], SeqSphere [12, 59, 67], BioNumerics [17, 20, 61, 62] and MentaLiST [19] (Fig. 1). The open-source workflows (MentaLiST, INNUENDO and GENPAT), based on Docker images (version 19.03.4) (https://www.docker.com/) which are hosted in the in-house GENPAT system (IZSAM, Italy), and commercial workflows (BioNumerics and SeqSphere) were executed in IZSAM (Italy). The workflow GENPAT corresponds to the in-house cgMLST workflow implemented in the GENPAT system (IZSAM, Italy). The open-source workflow BIGSdb was executed using the genomic taxonomy platform of Institut Pasteur (France; https://bigsdb.pasteur.fr/). All cgMLST workflows included in the present study were assessed based on the same set of loci and alleles, using the *L. monocytogenes* schema of 1748 cgMLST loci [24] downloaded from BIGSdb-*Lm* [24, 46] on 8th March 2021.

BIGSdb

Paired-end reads were de novo assembled using fq2dna version 21.06 (https://gitlab.pasteur.fr/GIPhy/fq2dna; strategy B; default settings). The corresponding fq2dna pipeline consists of trimming and clipping of low-quality reads and adapters with AlienTrimmer (version 2.0) [77], sequencing error correction with Musket (version 1.1) [78], paired-end read merging with FLASh (version 1.2.11) [79], coverage homogenization with ROCK (version 1.9.3; https://gitlab.pasteur.fr/vlegrand/ROCK) [81, 82, 90], and de novo assembly with SPAdes (version 3.15.0) [57]. In brief, the paired-end reads were first pre-processed through deduplication, clipping, trimming (Phred score threshold: 15, minimum read length: 50 bp) and error correction. Second, two distinct sequence datasets were created for each paired-reads by merging or not the pre-processed paired-end reads. Third, the coverage depth of the two read datasets (i.e. merged or not) was homogenized to 60X (i.e. digital normalization procedure), and each of the two resulting subsets of paired-end reads was used to infer a de novo genome assembly. The most precise between the two assemblies was selected by maximizing the number of genes completely contained within assembled contigs (E-size) [83]. Finally, the selected assembly was used together with its corresponding paired-end reads to infer a genome coverage profile (GCP) (i.e. distribution of the number of assembled bases per sequencing depth value) [84]. Based on the coverage profile, sufficiently long (> 1000 bp) and significantly covered scaffold sequences were finally selected. Contigs smaller than 300 bp were ignored. Draft assemblies were uploaded in a dedicated project in BIGSdb-*Lm* (https://bigsdb.pasteur.fr/listeria) powered by the BIGSdb software (version 1.31.0) [46]. cgMLST allele calling [46] was performed therein based on the BLASTN algorithm [69], with minimum of 70% of nucleotide identity and 70% of coverage and word size of 10. The missing data (0) and mismatches (empty set) from BIGSdb were considered as nonidentical calls in the present study. For the record, the mismatches (empty set) correspond to potential new alleles which are quality-checked by the Institute Pasteur curator before designation of new identifiers.

INNUENDO

As proposed by the cross-sectoral platform for the integration of genomics in the surveillance of food-borne pathogens (INNUENDO), the cgMLST workflow INNUENDO was based on de novo assembly and allele calling using INNUca (version 4.2.2) [49] and chewBBACA (version 2.6.0; default setting) [14] pipelines, respectively. More precisely, the INNUca assembler performs successively read control with FastQC (version 0.11.5) [79], trimming with Trimmomatic (version 0.36; clipping 3:30:10:6; sliding window 5:20; leading 3; trailing 3; minimum length 55) [58], coverage estimation with the internal module based on Bowtie2 (version 2.2.9) [73] and Samtools (version 1.3.1) [74], de novo assembly with SPAdes (version 3.9.0, careful; only assembler: coverage cutoff 2; k 21,33,55,67 and 77) [57], pearl-based filtering of contigs presenting at least 200 bp, kmer coverage of 2 and CG content between 5.0 and 95.0% (version 0.9.10), and correction of draft assembly with Pilon (version 1.18) [85], as well as MLST assessment based on MLST (version 2.4) [16]. The default parameters of chewBBACA (including allele size threshold = 0.2, BLASTP score ratio ≥ 0.6 and the recommended prodigal training file Listeria_monocytogenes.trn: https://chewbbaca.online/stats [69]) were applied in the present study considering exact match with known alleles (encoded EXC) as identical calls, as well as new inferred allele (INF), locus not found (LNF), possible locus on the tip of contigs (PLOT), non-informative paralogous hits (NIPH), alleles larger (ALM) and smaller (ASM) than mode, as nonidentical calls.

GENPAT

The GENPAT workflow is constituted of the NGSmanager de novo assembly pipeline implemented in GENPAT and chewBBACA allele caller with identical setting and version described above (see INNUENDO) [14]. More precisely, the NGSmanager assembly pipeline performs read trimming with Trimmomatic (version 0.36; clipping 2:30:10; leading 25; trailing 25 sliding window 20:25 minimal length 36) [58], de novo assembly with SPAdes (version 3.11.1; only assembler; careful; −k 21, 33, 55 and 77) [57], and filtering of contigs lower than 200 bp with a homemade Python script AssemblyFilter.py (i.e. version 2.7.8). The chewBBACA-based definitions of identical and nonidentical calls of the GENPAT workflow were identical to those described above (see INNUENDO).

SeqSphere

A new task template was created in Ridom SeqSphere+ (version 6.0.2), so-called SeqSphere in the present study, by importing allele library constructed using *L. monocytogenes* 1748 loci schema of cgMLST alleles [25] downloaded from BIGSdb-Pasteur, as described above. The first allele of each target was indicated as a reference sequence (ref-seq) and ref-seq alignment gap penalty was set to default. In the default "Target QC Procedure", the warnings were issued for alleles with breadth of coverage <75% and read depth of coverage <5X, as well as in cases of frameshift detected in translatable target and consensus length varying by more than 6 triplets compared to the ref-seq. Moreover, ambiguities were not allowed in the target sequences. The target scan procedure was set according to the guidelines for *L. monocytogenes* cgMLST typing from the Institute Pasteur (https://bigsdb.pasteur.fr/listeria/cgMLST_guidelines.pdf) with the minimum required allele identity and minimum percentage aligned to re-seq of 70% [24]. The best matching allele was forced when multiple gene matches were identified. In order to assess the full workflow of Ridom SeqSphere+, the sequencing reads were assembled de novo using the integrated assembly pipeline. Briefly, the paired sequencing reads were quality-trimmed with FastQC (version 0.11.7) at 5′ and 3′ end until average quality was 30 in a window of 20 bases [79]. The trimmed reads were assembled with SPAdes using default settings (−-careful option enabled) [57]. The assembled scaffolds were scanned for the presence of targeted genes and the alleles were assigned using the established parameters. The unidentified (? (not found)) and new alleles (? (new)) from SeqSphere+ were considered as nonidentical calls in the present study.

BioNumerics

BioNumerics (Applied Maths NV: bioMérieux company, Sint-Martens-Latem, Belgium) offers a fully automated workflow for cgMLST, the so-called WGS tools plugin (version 7.6.3). By default, the WGS tools plugin (i.e. AWS environment) proposes assembly-based (i.e. BLASTN algorithm from de novo assembly [69]) and/or -free workflows (i.e. kmer-based detection of alleles from unassembled reads) [34, 46, 61]. The BioNumerics assembly-based workflow can detect new alleles in addition to allele calling, while the assembly-free workflow cannot identify new alleles (https://www.applied-maths.com/news/bionumerics-version-763-released). By default, the BioNumerics outputs of the free-assembly workflow correspond to cgMLST alleles identically identified by assembly-based and -free workflows, in addition to alleles identified only through assembly-free workflow. Consequently, the output of the assembly-based workflow alone (BioNumericsAB), or in combination with the assembly-free workflow (BioNumericsAF), were firstly compared to each other in the present study in order to compare secondly the most precise one to the other cgMLST workflow of interest. More precisely, the reads were assembled using SPAdes (version 3.7.1) implemented in BioNumerics (version 7.6.2) without specifying any parameter, then the sequences obtained were scanned with the "assembly-based calls" and "assembly-free calls" algorithms successively. The minimum similarity to call new alleles (i.e. 80%), kmer size (35 bases), minimum coverage (3X), minimum forward coverage (1X) and minimum reverse coverage (1X) were set following BioNumerics recommendations. The unidentified alleles from BioNumerics (labeled with a question mark '?') were considered as nonidentical calls in the present study.

Even though few differences of identical alleles against reference circular genomes (IAAR) were observed at extended scales between the workflows BioNumericsAB and BioNumericsAF (Additional file 9A and Additional file 9B), the workflow BioNumericsAF identified significantly (Wilcoxon signed rank tests: $p < 1 \times 10^{-6}$) more IAAR that the workflow BioNumericsAB for each targeted depth of coverage (Additional file 9C and Additional file 9D). Consequently, the BioNumerics workflow combining assembly-based and-free approaches was retained to be compared to the other cgMLST workflow. In the interests of simplification, this BioNumerics workflow combining assembly-based and-free approaches (i.e. BioNumericsAF) will be named BioNumerics workflow in the present study.

MentaLiST

Working directly with the raw paired-end reads, MentaLiST does not require prior genome assembly (i.e. de novo assembly or reference genome mapping) [19]. In brief, the workflow MentaLiST (version 1.0.0) implements the

principle of kmer counting [54] and data compression to decrease dataset sizes and execution duration based on the construction of a coloured de Bruijn graph [87]. After assessment of all kmers present on the schema of alleles for each locus stored as a kmer hash map, all alleles that contain kmers from reads of a given sample will receive one vote, and the called alleles are those with the most votes for each locus [19]. The argument "--fasta" of MentaLiST was used to perform cgMLST of the three ATCC reference assemblies used in the present study. The default parameters of MentaLiST were applied in the present study considering multiple possible alleles (+) and partially covered alleles (−) as identical calls, as well as missing loci (0 or 0?) and new allele (N) as nonidentical calls.

Statistical analyses

The differences of alleles between cgMLST workflows applied to reference circular genomes were represented through Edward's Venn diagrams [88] built with jvenn (http://jvenn.toulouse.inra.fr/app/example.html) [89]. The results from paired-end read downsampling (Additional file 1), cgMLST typing (Additional file 2) and the parameters of interest (Fig. 1) were compiled into a single dataframe (Additional file 3) to perform statistical analyses. With the objective to explain the precision of cgMLST workflows, the amount of identical alleles against reference genomes (i.e. the parameter to explain, also called the response variable) was compared to several in vitro and in silico parameters of interest (i.e. the parameters potentially explaining the response variable, also called the explanatory variables) based on two independent statistical analyses, namely PCA and GLM. The PCA and GLM were selected because of their abilities to manage together categorical and numerical parameters. The in vitro parameters of interest include 4 categorical parameters (i.e. tested reference genomes, successive platings, as well as replicates of DNA extraction and sequencing). The in silico ones include 2 categorical (i.e. cgMLST workflows and targeted read/kmer depth of coverage) and 51 numerical parameters (i.e. read depth of coverage, read breadth of coverage, 48 assembly metrics and number of identified alleles against schema) (Additional file 3). The R-scripts dedicated to statistical analyses are available in GitHub (https://github.com/Nicolas-Radomski/DownsampledReads and https://github.com/Nicolas-Radomski/cgMLSTcomparison).

Principal component analyses

The exploratory PCAs aimed at increasing interpretability and minimizing information loss at the same time, by reducing the dimensional of the large dataset of numerical parameters through projection of data points on the first few principal components [90]. Two different PCAs were performed in the present study. The first PCA assessed the behavior of the response variable (i.e. the parameter to explain: IAAR) together with the explanatory variables corresponding to in silico numerical parameters of interest estimated through all assembly-based and assembly-free cgMLST workflows (i.e. the parameters potentially explaining the response variable: DEPTH, BREADTH and IAAS). The PCA was repeated excluding the assembly-free workflow MentaLiST (i.e. DEPTH, BREADTH, assembly metrics and IAAS) to additionally evaluate the impact of 48 assembly metrics on cgMLST precision for a total of 52 numerical parameters (i.e. DEPTH, BREADTH, 48 assembly metrics, IAAS and IAAR). For readability of the illustrations, these numerical parameters were grouped together according to PCA outcomes and only one parameter from each group was represented (Additional file 4). These PCAs were systematically performed in comparison to the in vitro and in silico categorical parameters of interest (i.e. tested reference genomes, successive platings, as well as replicates of DNA extraction and sequencing, targeted read/kmer depth of coverage and cgMLST workflows). These PCAs were performed with the ggplot2-based biplot R library [91] called "ggbiplot" (https://github.com/vqv/ggbiplot) requiring R libraries "usethis" and "devtools" [72].

Generalized linear models

Extending the concept of the linear regression model, the GLMs integrate link functions around the linear combinations of the explanatory variables in order to bypass the restriction to linearity from the linear models [92]. As described above concerning the PCAs, two different GLMs were performed in the present study. The first GLM aimed at explaining the response variable (i.e. the parameter to explain: IAAR) by explanatory variables corresponding to in vitro and in silico parameters of interest (i.e. numerical and categorical) estimated through all assembly-based and assembly-free cgMLST workflows (i.e. the parameters potentially explaining the response variable: tested reference genomes (REFERENCE), successive platings (PLATING), DNA extraction replicates (DNA), sequencing replicates (SEQUENCING), read depth of coverage (DEPTH), read breadth of coverage (BREADTH) and IAAS). Following the same design, the second GLM aimed at explaining the response variable by explanatory variables from assembly-based cgMLST workflows (i.e. the parameters potentially explaining

the response variable: REFERENCE, PLATING, DNA, SEQUENCING, DEPTH, BREADTH, assembly metrics and IAAS). Before to perform these GLMs, the distributions of the response variable were assessed through statistical tests Shapiro-Wilk (Gaussian distribution), Chi-square (uniform distribution), two side Poisson (two side Poisson distribution), one side Poisson with upper hypothesis (one side Poisson distribution with upper hypothesis) and one side Poisson with lower hypothesis (one side Poisson distribution with upper hypothesis) implemented in the R library "stats" [72].

Including or excluding MentaLiST from the cgMLST comparison, the IAAR did not follow Gaussian (Shapiro-Wilk, $p < 2.2 \times 10^{-16}$), uniform (Chi-square, $pp < 2.2 \times 10^{-16}$), two side Poisson (two side Poisson, $p < 2.2 \times 10^{-16}$) or one side Poisson with upper hypothesis (one side Poisson with upper hypothesis, $p < 2.2 \times 10^{-16}$) distributions, in the favor of one side Poisson with lower hypothesis (one side Poisson with upper hypothesis, $p = 1$). The presence (including MentaLiST) and absence (excluding MentaLiST) of GLM overdispersions, implemented in the R library "AER" [93], allowed retainment of quasiPoisson- (dispersion test, $p < 2.2 \times 10–16^{-16}$ and alpha > 1) and Poisson- (dispersion test, $p = 1$ and alpha $\approx$ 1) distributions for GLMs, respectively, for the R function "glm" from the R library "stats" [72].

Confirmations of parameters explaining the cgMLST precision

In order to confirm results from PCA- and GLMs-based statistical analyses, the parameters explaining the cgMLST precision (i.e. IAAR × 100 / 1748) were presented trough four-way figures, MST-based clustering and three-way tables.

Four-way figures

The IAAS (Additional file 6A-Additional file 6D) and IAAR (Additional file 6E-Additional file 6H) were presented according to the parameters of interest focusing on the parameters explaining the cgMLST precision through four-way figures with the R library "ggplot2" [91]. These four-way figures were prepared with y-axis presenting broad (i.e. extended scale) or narrow (i.e. restricted scale) range of units.

MST-based clustering

The cgMLST clustering was represented through MSTs according to parameters explaining the cgMLST precision. The cgMLST profiles from each workflow (Additional file 2) were used to build MSTs using Bionumerics software (version 7.6.3). Missing alleles calls were ignored in the MST differences calculations. The MST clusters containing at least two genomes and allele differences ≤7, were highlighted in grey.

Three-way tables

The cgMLST precision (i.e. IAAR × 100 / 1748) was presented according to parameters explaining it through three-way tables with the R functions 'subset' and 'dcast' from the R library "base" and "reshape2", respectively [72].

Abbreviations

ALM: Alleles larger than mode; ASM: Alleles smaller than mode; ATCC: American Type Culture Collection Global Bioresource Center; BIGSdb-*Lm*: Bacterial Isolate Genome Sequence database (BIGSdb) dedicated to *L. monocytogenes*; BREADTH: Read breadth of coverage; C0: Contigs > 0 bp; C1000: Contigs > 1 000 bp; C5000: Contigs > 5 000 bp; C10000: Contigs > 10 000 bp; C25000: Contigs > 25 000 bp; C50000: Contigs > 50 000 bp; cgMLST: Core genome multi-locus sequence typing; CRBIP: Biological Resources Center of the Institut Pasteur; DEPTH: Read depth of coverage; DNA: DNA extraction replicate; DR: Duplication ratio; DrDk: Targeted read (Dr) and kmer (Dk) depth; E-size: Number of genes completely contained within assembled contigs; EXC: Exact match with known alleles; GC: GC%; GCP: Genome coverage profile; GENPAT: Whole Genome Sequencing of microbial pathogens: data-base and bioinformatics analysis; GF: Genome fraction; GLM: Generalized linear model; IAAR : Identical alleles against reference circular genomes; IAAS: Identified alleles against schema; ID100: Indels per 100 kb; INF: New inferred allele; INNUENDO: Cross-sectoral platform for the integration of genomics in the surveillance of food-borne pathogens; IT: Information technology; K: Kmer size; L50: Smallest number of contigs whose length sum makes up 50% of the total genome length; L75: Smallest number of contigs whose length sum makes up 75% of the total genome length; LA: Largest alignment; LA50: Smallest number of aligned blocks whose length sum makes up 50% of the total genome length; LA75: Smallest number of aligned blocks whose length sum makes up 75% of the total genome length; LC: Largest contig; LG50: Smallest number of contigs whose length sum makes up 50% of the total reference genome length; LG75: Smallest number of contigs whose length sum makes up 75% of the total reference genome length; LGA50: Smallest number of aligned blocks whose length sum makes up 50% of the total reference genome length; LGA75: Smallest number of aligned blocks whose length sum makes up 75% of the total reference genome length; LMA: Local misassemblies; LNF: Locus not found; MA: Misassemblies; MAC: Misassembled contigs; MACL: Misassembled contigs length; MIAAR: Misidentified alleles against reference; MLST: Multi-locus sequence typing; MM100: Mismatches per 100 kb; MST: Minimum spanning tree; N: Any base; N100: N per 100 kb; N50: Sequence length of the shortest contig at 50% of the total genome length; N75: Sequence length of the shortest contig at 75% of the total genome length; NA50: Sequence length of the shortest aligned blocks at 50% of the total genome length; NA75: Sequence length of the shortest aligned blocks at 75% of the total genome length; NG50: Sequence length of the shortest contig at 50% of the total reference genome length; NG75: Sequence length of the shortest contig at 75% of the total reference genome length; NGA50: Sequence length of the shortest aligned blocks at 50% of the total reference genome length; NGA75: Sequence length of the shortest aligned blocks at 75% of the total reference genome length; NIPH: Non-informative paralogous hits; ORFs: Open reading frames; PCA: Principal component analysis; PLATING: Successive platings; PLOT: Possible locus on the tip of contigs; R: Read length; REFERENCE: Tested reference genomes s; SEQUENCING: Sequencing replicate; SNV: Single nucleotide variant; SQEM: Scaffold gap extensive misassembly; SQLM: Scaffold gap local misassembly; ST: Sequence type; TAL: Total aligned length; TL: Total length; TL0: Total length>0bp; TL1000: Total length > 1000 bp; TL5000: Total length > 5 000 bp; TL10000: Total length > 10 000 bp; TL25000: Total length > 25 000 bp; TL50000: Total length > 50 000 bp; UAC: Unaligned contigs; UACP: Unaligned contigs partial; UAL: Unaligned length; UAMC: Unaligned and misassembly contigs; UIAAS: Unidentified alleles against schema; wgMLST: Whole genome MLST; WGS: Whole genome sequencing.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-022-08437-4.

Additional file 1 Read depth (X) and breadth (%) coverages estimated with BBMap (version February 13, 2020) or INNUca (version 4.2.2) of *Listeria monocytogenes* paired-end reads from reference genomes ATCC19114, ATCC19115 and ATCCBAA679 ($n=420$) downsampled at different targeted read (Dr: 10X, 20X, 30X, 40X, 50X, 60X, 70X, 80X, 90X and 100X) and kmer (Dk: 8X, 15X, 23X, 30X, 38X, 45X, 52X, 60X, 67X, 75X) depth (X) with BBNorm (read length $R=150$ and kmer size K = 30). (TSV 65 kb)

Additional file 2 Standardized matrices of the cgMLST workflows BIGSdb ($n=423$), INNUENDO ($n=339$), GENPAT ($n=423$), SeqSphere ($n=423$), BioNumericsAB ($n=423$), BioNumericsAF ($n=423$) and MentaLiST ($n=423$) applied to downsampled paired-end reads from 3 reference genomes of *Listeria monocytogenes* (i.e. ATCC19114, ATCC19115 and ATCCBAA679). The terms ATCC19114, ATCC19115 and ATCCBAA679 from the field FILE correspond to cgMLST profiles of the corresponded circular de novo assemblies from ATCC company. The empty sets represent mismatches. Because of internal firewall, the INNUca assembler integrated into the cgMLST workflow INNUENDO cannot not perform assemblies of paired-end reads with read depth of coverage of 20X ($n=42$) and 10X ($n=42$). (TSV 12262 kb)

Additional file 3 Standardized outcomes of the cgMLST workflows BIGSdb ($n=420$), INNUENDO ($n=336$), GENPAT ($n=420$), SeqSphere (n = 420), BioNumerics ($n=420$) and MentaLiST ($n=420$) applied to downsampled paired-end reads from 3 reference genomes of *Listeria monocytogenes* (i.e. ATCC19114, ATCC19115 and ATCCBAA679) with associated de novo assembly parameters assessed with Quast (version 5.0.2) and MultiQC (version 1.9). The targeted read depth (Dr: 10X, 20X, 30X, 40X, 50X, 60X, 70X, 80X, 90X and 100X) were prepared according to kmer depth (Dk): 8X, 15X, 23X, 30X, 38X, 45X, 52X, 60X, 67X, 75X) setting of BBNorm (read length $R=150$ and kmer size K = 30). Because of internal firewall, the INNUca assembler integrated into the cgMLST workflow INNUENDO cannot not perform assemblies of paired-end reads with read depth of coverage of 20X ($n=42$) and 10X ($n=42$). (TSV 1207 kb)

Additional file 4 Principals component analyses (PCAs) of the numerical parameters C1000, C10000, DR, GC, IAAR, IAAS, ID100, L50, LA50, LA, LMA, MA, N50, NA50, DEPTH, BREADTH, SQEM, SQLM, TL1000, TL10000, UACP and UAMC (defined in the section abbreviations) according to the categorical parameters "reference genome" (A), "successive platings" (B), "DNA extraction replicate" (C), "sequencing replicate" (D), "targeted depth" (E), "cgMLST workflows" (F), including assembly-based cgMLST workflows BIGSdb ($n=420$), INNUENDO ($n=336$), GENPAT ($n=420$), SeqSphere ($n=420$) and BioNumerics ($n=420$) applied to downsampled paired-end reads from 3 reference genomes of *Listeria monocytogenes* (i.e. ATCC19114, ATCC19115 and ATCCBAA679). The PCA parameters C0-C1000-C5000-C10000-C25000-C50000, GC-TL0-TL1000-TL5000-TL10000-TL25000-TL50000-TL-TAL-MACL, N50-NG50-N75-NG75-SQEM-NA50-NGA50-NA75-NGA75-LA, L50-LG50-L75-LG75, LA50-LGA50-LA75-LGA75, DEPTH-GF, LMA-UAL-MM100-SQLM, DR-N100-UAC and MA-MAC were overlapped and are consequently not presented together.

Additional file 5 Coefficients (Coef.) of the generalized linear models (GLMs with Poisson distribution and without overdispersion) comparing the parameters "identical alleles against reference circular genomes" (IAAR) with the parameters of interest "reference genome" (REFERENCE), "successive platings" (PLATING) (B), "DNA extraction replicate" (DNA), "sequencing replicate" (SEQUENCING), "read depth" (DEPTH), "read breadth" (BREADTH), assembly parameters (defined in the section abbreviations: C0, C1000, C10000, C25000, C5000, C50000, DR, GC, GF, ID100, L50, L75, LA, LA50, LA75, LC, LG50, LG75, LGA50, LGA75, LMA, MA, MAC, MACL, MM100, N100, N50, N75, NA50, NA75, NG50, NG75, NGA50, NGA75, SQEM, SQLM, TAL, TL, TL0, TL1000, TL10000, TL25000, TL5000, TL50000, UAC, UACP, UAL, UAMC), cgMLST workflows (WORKFLOW) together (A) and cgMLST workflows independently including BIGSdb (B: $n=420$), INNUENDO (C: $n=336$), GENPAT (D: $n=420$), SeqSphere (E: $n=420$) and BioNumerics (F: $n=420$), applied to downsampled paired-end reads from 3 reference genomes of *Listeria monocytogenes* (i.e. ATCC19114, ATCC19115 and ATCCBAA679). Few parameters are not defined because of singularities.

Additional file 6 Box-plots representing the impact of downsampled paired-end reads (i.e. 2x150bp) of reference genomes of *Listeria monocytogenes* (i.e. ATCC19114, ATCC19115 and ATCCBAA679), on cgMLST outcomes (BIGSdb: $n=420$, INNUENDO: $n=336$, GENPAT: $n=420$, SeqSphere: $n=420$, BioNumerics: $n=420$ and MentaLiST: $n=420$), including identified alleles against schema (A, B, C, D) or identical alleles against reference circular genomes at extended (E, F, G, H) or restricted (I, J, K, L) scales, according to reference genomes (A, E, I), successive platings (B, F, J), DNA extraction replicate (C, G, K) and sequencing replicate (C, H, L). The targeted read depth (Dr: 10X, 20X, 30X, 40X, 50X, 60X, 70X, 80X, 90X and 100X) were prepared according to kmer depth (Dk): 8X, 15X, 23X, 30X, 38X, 45X, 52X, 60X, 67X, 75X) setting of BBNorm (read length $R=150$ and kmer size K = 30). Because of internal firewall, the INNUca assembler integrated into the cgMLST workflow INNUENDO cannot not perform assemblies of paired-end reads with read depth of coverage of 20X ($n=42$) and 10X ($n=42$).

Additional file 7 Box-plots representing the impact of downsampled paired-end reads (i.e. 2x150bp) of *Listeria monocytogenes* on unidentified alleles against schema at extended (A) or restricted (B) scales, according to reference genomes (i.e. ATCC19114, ATCC19115 and ATCCBAA679) and cgMLST workflows including BIGSdb ($n=420$), INNUENDO ($n=336$), GENPAT ($n=420$), SeqSphere ($n=420$), BioNumerics ($n=420$) and MentaLiST ($n=420$). The targeted read depth (Dr: 10X, 20X, 30X, 40X, 50X, 60X, 70X, 80X, 90X and 100X) were prepared according to kmer depth (Dk): 8X, 15X, 23X, 30X, 38X, 45X, 52X, 60X, 67X, 75X) setting of BBNorm (read length $R=150$ and kmer size K = 30). Because of internal firewall, the INNUca assembler integrated into the cgMLST workflow INNUENDO cannot not perform assemblies of paired-end reads with read depth of coverage of 20X ($n=42$) and 10X ($n=42$).

Additional file 8 Minimum spanning trees (MSTs) representing the impact on clustering of cgMLST workflows BIGSdb (A: $n=423$), INNUENDO (B: $n=339$), GENPAT (C: $n=423$), SeqSphere (D: $n=423$), BioNumerics (E: $n=423$) and MentaLiST (F: $n=423$), of *Listeria monocytogenes* reference genomes (i.e. ATCC19114, ATCC19115 and ATCCBAA679 on the left of each workflow) and targeted depth of coverage (i.e. on the right of each workflow) from downsampled paired-end reads (i.e. 2x150bp). The MSTs were built with BioNumerics ignoring missing data. The MST clusters of at least two genomes, one node and allele differences ≤ 7, were highlighted in grey. The targeted read depth (Dr: 10X, 20X, 30X, 40X, 50X, 60X, 70X, 80X, 90X and 100X) were prepared according to kmer depth (Dk): 8X, 15X, 23X, 30X, 38X, 45X, 52X, 60X, 67X, 75X) setting of BBNorm (read length $R=150$ and kmer size K = 30). Because of internal firewall, the INNUca assembler integrated into the cgMLST workflow INNUENDO cannot not perform assemblies of paired-end reads with read depth of coverage of 20X ($n=42$) and 10X ($n=42$).

Additional file 9 Box-plots representing the impact of downsampled paired-end reads (i.e. 2x150bp), at extended (A and B) or restricted (C and D) scales of identical alleles against reference circular genomes, spiting (A and C) or merging (B and D) reference genomes of *Listeria monocytogenes* (i.e. ATCC19114, ATCC19115 and ATCCBAA679), on cgMLST outcomes from the assembly-based workflow alone (BioNumericsAB: $n=420$), or in combination with the assembly-free workflow implemented in BioNumerics (version 7.6.2) (BioNumericsAF: $n=420$).

Acknowledgments

We thank the Italian Ministry of Health for supporting in the acquisition of high-performance computing resources. This work used the computational and storage services (Maestro cluster) provided by the information technology (IT) department at the Institut Pasteur, Paris.

Authors' contributions

All authors have made substantial contributions to the conception and design of the work, as well as to the interpretation of data. C.C. designed the wet-lab experimental plan and managed the corresponding analyses in GENPAT (IZSAM). A.C.H. and M.T. performed the dry-lab experiments. A.D.P. and

S.B. managed the dry-lab activities of GENPAT (IZSAM) and BIGSdb-Pasteur, respectively. A.C.R. developed the de novo assembly workflow recommended for BIGSdb related analyses (Pasteur). I.M. and A.J.-G.G. performed BioNumerics and SeqSphere+ workflows in IZSAM, respectively. F.P. and A.M. performed the analyses related to BIGSdb-Pasteur. F.P. performed the MST-based analyses. N.R. designed the dry-lab experimental plan, performed the open source bioinformatics analyses, collected genomic data, combined dry- and wet-lab outcomes, and developed statistical analyses in GENPAT (IZSAM). N.R. drafted the manuscript and integrated comments from F.P., I.M., A.J., A.M., A.C.H., M.T., G.G., A.C.R., S.B., A.D.P. and C.C.. All authors commented and approved the final manuscript including the author's contribution to the study, and have agreed both to be personally accountable for the author's contributions and to ensure that questions related to the accuracy or integrity of any part of the work are appropriately investigated, resolved, and documented.

Funding
The study was funded by the European Joint Programme (EJP) dedicated to One Health Structure In Europe (COHESIVE) under Grant Agreement No 773830.

Availability of data and materials
The paired-end reads are available in the European Nucleotide Archive (ENA) under the BioProject PRJEB45560 (https://www.ebi.ac.uk/ena/browser/view/PRJEB45560).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Institut Pasteur, Université de Paris, Biological Resources Center of Institut Pasteur, 75015 Paris, France. [2]Istituto Zooprofilattico Sperimentale dell'Abruzzo e del Molise "Giuseppe Caporale" (IZSAM), National Reference Centre (NRC) for Whole Genome Sequencing of microbial pathogens: data-base and bioinformatics analysis (GENPAT), via Campo Boario, 64100 Teramo, TE, Italy. [3]Istituto Zooprofilattico Sperimentale dell'Abruzzo e del Molise "Giuseppe Caporale" (IZSAM), Bacteriology Unit, via Campo Boario, 64100 Teramo, TE, Italy. [4]Institut Pasteur, National Reference Center and WHO Collaborating Center Listeria, 75015 Paris, France. [5]Institut Pasteur, Université de Paris, Inserm U1117, Biology of Infection Unit, 75015 Paris, France. [6]Istituto Zooprofilattico Sperimentale dell'Abruzzo e del Molise "Giuseppe Caporale" (IZSAM), National Reference Labororatory (LNR) for Listeria monocytogenes, via Campo Boario, 64100 Teramo, TE, Italy. [7]Institut Pasteur, Université de Paris, Bioinformatics and Biostatistics Hub, 75015 Paris, France. [8]Institut Pasteur, Université de Paris, Biodiversity and Epidemiology of Bacterial Pathogens, 75015 Paris, France.

Received: 6 December 2021 Accepted: 28 February 2022
Published online: 26 March 2022

References

1. Payne M, Kaur S, Wang Q, Hennessy D, Luo L, Octavia S, et al. Multilevel genome typing: genomics-guided scalable resolution typing of microbial pathogens. Euro Surveill Bull Eur Sur Mal Transm Eur Commun Dis Bull. 2020;25:1900519.
2. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell 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.
3. Patiño LH, Camargo M, Muñoz M, Ríos-Chaparro DI, Patarroyo MA, Ramírez JD. Unveiling the multilocus sequence typing (MLST) schemes and Core genome phylogenies for genotyping Chlamydia trachomatis. Front Microbiol. 2018;9:1854.
4. Pitondo-Silva A, Santos ACB, Jolley KA, Leite CQF, Darini AL da C. Comparison of three molecular typing methods to assess genetic diversity for mycobacterium tuberculosis. J Microbiol Methods 2013;93:42–48.
5. Yan S, Zhang W, Li C, Liu X, Zhu L, Chen L, et al. Serotyping, MLST, and Core genome MLST analysis of Salmonella enterica from different sources in China during 2004–2019. Front Microbiol. 2021;12:688614.
6. O'Connor M, Peifer M, Bender W. Construction of large DNA segments in Escherichia coli. Science. 1989;244:1307–12.
7. Zhang J-H, Wu L-Y, Zhang X-S. Reconstruction of DNA sequencing by hybridization. Bioinforma Oxf Engl. 2003;19:14–21.
8. Istrail S, Sutton GG, Florea L, Halpern AL, Mobarry CM, Lippert R, et al. Whole-genome shotgun assembly and comparison of human genome assemblies. Proc Natl Acad Sci U S A. 2004;101:1916–21.
9. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008;452:872–6.
10. Soucy SM, Huang J, Gogarten JP. Horizontal gene transfer: building the web of life. Nat Rev Genet. 2015;16:472–82.
11. Maiden MCJ, van Rensburg MJJ, Bray JE, Earle SG, Ford SA, Jolley KA, et al. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol. 2013;11:728–36.
12. Ruppitsch W, Pietzka A, Prior K, Bletz S, Fernandez HL, Allerberger F, et al. Defining and evaluating a Core genome multilocus sequence typing scheme for whole-genome sequence-based typing of Listeria monocytogenes. J Clin Microbiol. 2015;53:2869–76.
13. Espitia-Navarro HF, Chande AT, Nagar SD, Smith H, Jordan IK, Rishishwar L. STing: accurate and ultrafast genomic profiling with exact sequence matches. Nucleic Acids Res. 2020;48:7681–9.
14. 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. 2018;4:e000166.
15. Ferrés I, Iraola G. MLSTar: automatic multilocus sequence typing of bacterial genomes in R. PeerJ. 2018;6:e5098.
16. Ragon M, Wirth T, Hollandt F, Lavenir R, Lecuit M, Le Monnier A, et al. A new perspective on Listeria monocytogenes evolution. PLoS Pathog. 2008;4:e1000146.
17. 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. 2019;10:2413.
18. Clausen PTLC, Aarestrup FM, Lund O. Rapid and precise alignment of raw reads against redundant databases with KMA. BMC Bioinformatics. 2018;19:307.
19. Feijao P, Yao H-T, Fornika D, Gardy J, Hsiao W, Chauve C, et al. MentaLiST – a fast MLST caller for large MLST schemes. Microb. Genomics. 2018;4:e000146.
20. Blanc DS, Magalhães B, Koenig I, Senn L, Grandbastien B. Comparison of whole genome (wg-) and Core genome (cg-) MLST (BioNumericsTM) versus SNP variant calling for epidemiological investigation of Pseudomonas aeruginosa. Front Microbiol. 2020;11:1729.
21. Coolen JPM, Jamin C, Savelkoul PHM, Rossen JWA, Wertheim HFL, Matamoros SP, et al. Centre-specific bacterial pathogen typing affects infection-control decision making. Microb. Genomics. 2021;7:000612.
22. Jamin C, De Koster S, van Koeveringe S, De Coninck D, Mensaert K, De Bruyne K, et al. Harmonization of whole-genome sequencing for outbreak surveillance of Enterobacteriaceae and enterococci. Microb. Genomics. 2021;7:000567.
23. Pightling AW, Petronella N, Pagotto F. The Listeria monocytogenes Core-genome sequence Typer (LmCGST): a bioinformatic pipeline for molecular characterization with next-generation sequence data. BMC Microbiol. 2015;15:224.
24. Moura A, Criscuolo A, Pouseele H, Maury MM, Leclercq A, Tarr C, et al. Whole genome-based population biology and epidemiological surveillance of Listeria monocytogenes. Nat Microbiol. 2016;2:16185.
25. Chen Y, Gonzalez-Escalona N, Hammack TS, Allard MW, Strain EA, Brown EW. Core genome multilocus sequence typing for identification of globally distributed clonal groups and differentiation of outbreak strains of Listeria monocytogenes. Appl Environ Microbiol. 2016;82:6258–72.

26. 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. 2017;23:1462–70.
27. Schjørring S, Gillesberg Lassen S, Jensen T, Moura A, Kjeldgaard JS, Müller L, et al. Cross-border outbreak of listeriosis caused by cold-smoked salmon, revealed by integrated surveillance and whole genome sequencing (WGS), Denmark and France, 2015 to 2017. Eurosurveillance. 2017;22:17-00762.
28. Van Walle I, Björkman JT, Cormican M, Dallman T, Mossong J, Moura A, et al. Retrospective validation of whole genome sequencing-enhanced surveillance of listeriosis in Europe, 2010 to 2015. Eurosurveillance. 2018:23.
29. Painset A, Björkman JT, Kiil K, Guillier L, Mariet J-F, Félix B, et al. LiSEQ – whole-genome sequencing of a cross-sectional survey of Listeria monocytogenes in ready-to-eat foods and human clinical cases in Europe. Microb. Genomics. 2019;5:e000257.
30. Kurpas M, Osek J, Moura A, Leclercq A, Lecuit M, Wieczorek K. Genomic characterization of Listeria monocytogenes isolated from ready-to-eat meat and meat processing environments in Poland. Front Microbiol. 2020;11:1412.
31. Rivas L, Paine S, Dupont P-Y, Tiong A, Horn B, Moura A, et al. Genome typing and epidemiology of human Listeriosis in New Zealand, 1999 to 2018. J Clin Microbiol. 2021;59:e00849–21.
32. Orsi RH, Bakker HC den, Wiedmann M. Listeria monocytogenes lineages: genomics, evolution, ecology, and phenotypic characteristics. Int J Med Microbiol 2011;301:79–96.
33. Heisick JE, Rosas-Marty LI, Tatini SR. Enumeration of viable Listeria species and Listeria monocytogenes in foods. J Food Prot. 1995;58:733–6.
34. Sabol A, Joung YJ, VanTubbergen C, Ale J, Ribot EM, Trees E. Assessment of genetic stability during serial in vitro passage and in vivo carriage. Foodborne Pathog Dis 2021;18:894-901.
35. Pasquali F, Do Valle I, Palma F, Remondini D, Manfreda G, Castellani G, et al. Application of different DNA extraction procedures, library preparation protocols and sequencing platforms: impact on sequencing results. Heliyon. 2019;5:e02745.
36. Larsonneur E, Criscuolo A, Moura A, Rocha EPC, Glaser P, Brisse S. Evaluation of de novo assemblies in view of creating automated pipelines dedicated to core-genome bacterial typing; 2017. https://doi.org/10.7490/F1000RESEARCH.1114831.1.
37. Savin C, Criscuolo A, Guglielmini J, Le Guern A-S, Carniel E, Pizarro-Cerdá J, et al. Genus-wide Yersinia core-genome multilocus sequence typing for species identification and strain characterization. Microb. Genomics. 2019;5:e000301.
38. Ghanem M, El-Gazzar M. Development of Mycoplasma s ynoviae (MS) core genome multilocus sequence typing (cgMLST) scheme. Vet Microbiol. 2018;218:84–9.
39. Liu Y-Y, Chen B-H, Chen C-C, Chiou C-S. Assessment of metrics in next-generation sequencing experiments for use in core-genome multilocus sequence type. PeerJ. 2021;9:e11842.
40. Lüth S, Deneke C, Kleta S, Al DS. Translatability of WGS typing results can simplify data exchange for surveillance and control of Listeria monocytogenes. Microb. Genomics. 2021;7:mgen000491.
41. Low AJ, Koziol AG, Manninger PA, Blais B, Carrillo CD. ConFindr: rapid detection of intraspecies and cross-species contamination in bacterial whole-genome sequence data. PeerJ. 2019;7:e6995.
42. Bushnell B. BBMap: A Fast, Accurate, Splice-Aware Aligner: Berkeley Lab; 2014. Report Number: LBNL-7065E
43. Llarena A, Ribeiro-Gonçalves BF, Nuno Silva D, Halkilahti J, Machado MP, Da Silva MS, et al. INNUENDO: a crosssectoral platform for the integration of genomics in the surveillance of food-borne pathogens. EFSA Support Publ. 2018;15:1-142.
44. 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;2:20.
45. Gargis AS, Kalman L, Lubin IM. Assuring the quality of next-generation sequencing in clinical microbiology and public health laboratories. J Clin Microbiol. 2016;54:2857–65.
46. Jolley KA, Maiden MCJ. BIGSdb: scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics. 2010;11:595.
47. 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. 2021;49:D660–6.
48. Deneke C, Uelze L, Brendebach H, Tausch SH, Malorny B. Decentralized investigation of bacterial outbreaks based on hashed cgMLST. Front Microbiol. 2021;12:649517.
49. Kubik S, Marques AC, Xing X, Silvery J, Bertelli C, De Maio F, et al. Recommendations for accurate genotyping of SARS-CoV-2 using amplicon-based sequencing of clinical samples. Clin Microbiol Infect. 2021;27:1036.e1–8.
50. Richardson MF, Weinert LA, Welch JJ, Linheiro RS, Magwire MM, Jiggins FM, et al. Population genomics of the Wolbachia endosymbiont in Drosophila melanogaster. PLoS Genet. 2012;8:e1003129.
51. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of ion torrent, Pacific biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012;13:1–13.
52. Uelze L, Borowiak M, Bönn M, Brinks E, Deneke C, Hankeln T, et al. German-wide Interlaboratory study compares consistency, accuracy and reproducibility of whole-genome short read sequencing. Front Microbiol. 2020;11:573972.
53. Magi A, Giusti B, Tattini L. Characterization of MinION nanopore data for resequencing analyses. Brief Bioinform 2016;:bbw077.
54. Gupta A, Jordan IK, Rishishwar L. stringMLST: a fast k-mer based tool for multilocus sequence typing. Bioinformatics. 2017;33:119–21.
55. Chen Y-C, Liu T, Yu C-H, Chiang T-Y, Hwang C-C. Effects of GC Bias in next-generation-sequencing data on De novo genome assembly. PLoS One. 2013;8:e62856.
56. Kuśmirek W, Nowak R. De novo assembly of bacterial genomes with repetitive DNA regions by dnaasm application. BMC Bioinformatics. 2018;19:273.
57. 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. 2012;19:455–77.
58. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinforma Oxf Engl. 2014;30:2114–20.
59. Pietzka A, Allerberger F, Murer A, Lennkh A, Stöger A, Cabal Rosel A, et al. Whole genome sequencing based surveillance of L. monocytogenes for early detection and investigations of Listeriosis outbreaks. Front Public Health 2019;7:139.
60. Halbedel S, Prager R, Fuchs S, Trost E, Werner G, Flieger A. Whole-genome sequencing of recent Listeria monocytogenes isolates from Germany reveals population structure and disease clusters. J Clin Microbiol. 2018;56:e00119-18.
61. Jagadeesan B, Baert L, Wiedmann M, Orsi RH. Comparative analysis of tools and approaches for source tracking Listeria monocytogenes in a food facility using whole-genome sequence data. Front Microbiol. 2019;10:947.
62. Camargo AC, Moura A, Avillan J, Herman N, McFarland AP, Sreevatsan S, et al. Whole-genome sequencing reveals *Listeria monocytogenes* diversity and allows identification of long-term persistent strains in Brazil. Environ Microbiol. 2019;21:4478–87.
63. Laarne P, Zaidan MA, Nieminen T. Ennemi: non-linear correlation detection with mutual information. SoftwareX. 2021;14:100686.
64. Wang Y, Li Y, Cao H, Xiong M, Shugart YY, Jin L. Efficient test for nonlinear dependence of two continuous variables. BMC Bioinformatics. 2015;16:260.
65. Lapidus AL, Korobeynikov AI. Metagenomic data assembly – the way of decoding unknown microorganisms. Front Microbiol. 2021;12:613791.
66. Segerman B. The Most frequently used sequencing technologies and assembly methods in different time segments of the bacterial surveillance and RefSeq genome databases. Front Cell Infect Microbiol. 2020;10:527102.
67. Tse CW, Curreem SO, Cheung I, Tang BS, Leung K-W, Lau SK, et al. A novel MLST sequence type discovered in the first fatal case of Laribacter hongkongensis bacteremia clusters with the sequence types of other human isolates. Emerg Microbes Infect. 2014;3:e41.
68. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.

69. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
70. Portmann A-C, Fournier C, Gimonet J, Ngom-Bru C, Barretto C, Baert L. A validation approach of an end-to-end whole genome sequencing workflow for source tracking of Listeria monocytogenes and Salmonella enterica. Front Microbiol. 2018;9:446.
71. FastQC AS. A quality control tool for high throughput sequence data. Babraham Bioinforma. 2018;1:1-1. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
72. R Development Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2015.
73. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012;9:357–9.
74. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
75. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29:1072–5.
76. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32:3047–8.
77. Criscuolo A, Brisse S. AlienTrimmer: a tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads. Genomics. 2013;102:500–6.
78. Liu Y, Schröder J, Schmidt B. Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data. Bioinforma Oxf Engl. 2013;29:308–15.
79. Magoč T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinforma Oxf Engl. 2011;27:2957–63.
80. Brown CT, Howe A, Zhang Q, Pyrkosz AB, Brom TH. A reference-free algorithm for computational normalization of shotgun sequencing data. ArXiv12034802 Q-Bio. 2012;1:1-18.
81. Wedemeyer A, Kliemann L, Srivastav A, Schielke C, Reusch TB, Rosenstiel P. An improved filtering algorithm for big read datasets and its application to single-cell assembly. BMC Bioinformatics. 2017;18:324.
82. Durai DA, Schulz MH. Improving in-silico normalization using read weights. Sci Rep. 2019;9:5133.
83. Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, et al. GAGE: a critical evaluation of genome assemblies and assembly algorithms. Genome Res. 2012;22:557–67.
84. Lindner MS, Kollock M, Zickmann F, Renard BY. Analyzing genome coverage profiles with applications to quality control in metagenomics. Bioinforma Oxf Engl. 2013;29:1260–7.
85. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9:e112963.
86. Chen Y, Luo Y, Carleton H, Timme R, Melka D, Muruvanda T, et al. Whole genome and Core genome multilocus sequence typing and single nucleotide polymorphism analyses of Listeria monocytogenes isolates associated with an outbreak linked to cheese, United States, 2013. Appl Environ Microbiol. 2017;83:e00633-17.
87. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet. 2012;44:226–32.
88. Edwards AWF. Cogwheels of the mind: the story of Venn diagrams. Baltimore: Johns Hopkins University Press; 2004.
89. Bardou P, Mariette J, Escudié F, Djemiel C, Klopp C. jvenn: an interactive Venn diagram viewer. BMC Bioinformatics. 2014;15:293.
90. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Transact A Math Phys Eng Sci. 2016;374:20150202.
91. Wickham H. ggplot2. Springer New York: New York, NY; 2009.
92. Müller M. Generealized Linear Models. In: Gentle JE, Härdle WK, Mori Y, editors. Handbook of Computational Statistics. Berlin: Springer Berlin Heidelberg; 2012. p. 681–709.
93. Cameron AC, Trivedi PK. Regression-based tests for overdispersion in the Poisson model. J Econom. 1990;46:347–64.

Publisher's Note