MICROBIAL GENOMICS

REVIEW

Page *et al.*, *Microbial Genomics* 2017;3

DOI 10.1099/mgen.0.000124

MICROBIOLOGY SOCIETY

OPEN DATA OPEN MICROBIOLOGY

Comparison of classical multi-locus sequence typing software for next-generation sequencing data

Andrew J. Page,[1,*] Nabil-Fareed Alikhan,[2] Heather A. Carleton,[3] Torsten Seemann,[4] Jacqueline A. Keane[5] and Lee S. Katz[3,6]

Abstract

Multi-locus sequence typing (MLST) is a widely used method for categorizing bacteria. Increasingly, MLST is being performed using next-generation sequencing (NGS) data by reference laboratories and for clinical diagnostics. Many software applications have been developed to calculate sequence types from NGS data; however, there has been no comprehensive review to date on these methods. We have compared eight of these applications against real and simulated data, and present results on: (1) the accuracy of each method against traditional typing methods, (2) the performance on real outbreak datasets, (3) the impact of contamination and varying depth of coverage, and (4) the computational resource requirements.

DATA SUMMARY

1. Simulated reads for datasets testing coverage and mixed samples have been deposited in Figshare; DOI: https://doi.org/10.6084/m9.figshare.4602301.v1.

2. Outbreak databases are available from GitHub; url – https://github.com/WGS-standards-and-analysis/datasets.

3. Docker containers used to run each of the applications are available from GitHub; url – https://github.com/andrewjpage/docker_mlst.

4. Accession numbers for the data used in this paper are available in the Supplementary Material.

INTRODUCTION

A small number of bacterial foodborne pathogens, such as *Salmonella*, *Campylobacter*, *Listeria* and *Escherichia*, cause a huge burden of disease in humans and animals. With *Listeria monocytogenes*, although the case count is small, the case-fatality rate is high at approximately 21 to 38 % [1, 2] and a high economic burden [3]. In the US, each foodborne illness can cost anywhere from hundreds to millions of US dollars depending on the organism. Therefore, investigating potential foodborne outbreaks and preventing any illness is advantageous from both economic and public health standpoints. In order to understand these bacteria in more depth, there have been many studies to describe their population structure using phylogenetic methods based on multi-locus sequence typing (MLST) [4, 5].

Additionally, there have been many large-scale surveillance efforts for these pathogens. One of the most successful programs has been PulseNet International [6], which aids in the detection of common source outbreaks. Recently, large numbers of isolates have been subjected to whole-genome sequencing (WGS) through an initiative between the Centers for Disease Control and Prevention (CDC), the US Food and Drug Administration (FDA), the US Department of Agriculture (USDA), and the National Center for Biotechnology Information (NCBI). Through this collaboration, every *L. monocytogenes* genome that is discovered in the food supply, or in clinical samples, is being sequenced and uploaded to the NCBI Sequence Read Archive (SRA) database. This collaboration has since started sequencing a large percentage of *Escherichia coli, Salmonella enterica, Campylobacter coli, Campylobacter jejuni* and many others, with the eventual goal of completely switching from pulsed-field gel electrophoresis to WGS. In Europe, Public Health England sequences every *Salmonella* and *Mycobacterium*

Received 15 March 2017; Accepted 7 June 2017

Author affiliations: [1]Pathogen Genomics, Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK; [2]Microbiology and Infection, University of Warwick, Coventry, UK; [3]Enteric Diseases Laboratory Branch, Centers for Disease Control and Prevention, Atlanta, GA, USA; [4]Doherty Applied Microbial Genomics, Department of Microbiology and Immunology, University of Melbourne, Peter Doherty Institute for Infection and Immunity, Melbourne, Australia; [5]Pathogen Informatics, Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, UK; [6]Center for Food Safety, College of Agricultural and Environmental Sciences, University of Georgia, Griffin, GA, USA.
***Correspondence:** Andrew J. Page, ap13@sanger.ac.uk

Keywords: MLST; multi-locus sequence typing; software comparison; next-generation sequencing.
Abbreviations: MLST, multi-locus sequence typing; NGS, next-generation sequencing; ST, sequence type; WGS, whole-genome sequencing.
Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Two supplementary tables are available with the online Supplementary Material.

tuberculosis isolate submitted to them and deposits the data in the SRA. Perceiving a future need for worldwide collaboration on these new methods, the Global Microbial Identifier (GMI) [7] partnership was initiated in 2011 to encourage data sharing among all nations for many purposes, including public health and research.

To aid in population structure studies and in epidemiological investigations, MLST has been used for nearly two decades [8] to categorize different clonal expansions of these pathogens into broad categories, based on allelic variation amongst seven highly conserved housekeeping genes. Sequence typing can be performed using both next-generation sequencing (NGS) and classical sequencing techniques. Whilst MLST is a low-resolution classification compared to what is possible from NGS data, the nomenclature is in common usage by microbiologists and clinicians. A number of software applications have been developed using a variety of fundamentally different techniques to calculate sequence types (STs) from NGS data. However, there has been no comprehensive review to date on the accuracy, computational performance, robustness and ease of use of these methods. In this paper, we have evaluated multiple MLST software applications on a variety of datasets, both real and simulated, such as: (1) standard sets of outbreak data from the Gen-FS WGS Standards and Analysis Working Group (available from https://github.com/WGS-standards-and-analysis/datasets) [9], which includes *C. jejuni, E. coli, L. monocytogenes* and *S. enterica*; (2) *Salmonella* isolates that have been typed using both traditional capillary sequencing and NGS; (3) simulated reads of varying coverage; and (4) simulated mixed strains. Here, we describe a comprehensive list of command-line tools for MLST analysis and benchmark them with these standardized datasets in terms of accuracy and computer resources required.

IMPACT STATEMENT

Sequence typing is rapidly transitioning from traditional sequencing methods to using whole-genome sequencing. A number of *in silico* prediction methods have been developed on an *ad hoc* basis and aim to replicate classical multi-locus sequence typing (MLST). This is believed to be the first study to comprehensively evaluate multiple MLST software applications on real validated datasets and on common simulated difficult cases. It will give researchers a clearer understanding of the accuracy, limitations and computational performance of the methods they use, and will assist future researchers to choose the most appropriate method for their experimental goals.

SOFTWARE OVERVIEW

MLST software can be categorized according to the input data they accept; there are tools that use raw sequence reads and tools that use *de novo* assemblies. Calling MLST from raw reads avoids the need to fully reconstruct the whole genome, theoretically allowing for a lower running time. However, in practice *de novo* assembly is routinely performed for bacteria [10] and assemblies may already be available for any given MLST analysis leading to faster sequence typing. The process of *de novo* assembly can introduce artefacts, particularly from short reads. For example, a gene may be fragmented over multiple contigs. A full overview is given in Table 1. In general, the desired characteristics of MLST software include:

(1) high specificity of calling STs,

(2) resilience in the face of mixed samples,

(3) tolerance with low sequencing coverage,

(4) efficient usage of computational and disk resources,

Table 1. Overview of MLST software

Software	Input	Algorithm	Licence	Source	Tests	Installation	Interface
ARIBA	Reads	Assembly	GPL3	GitHub	Yes	Pip, Apt, Docker	Command line
BigsDB [11]	Contigs	BLASTN	GPL3	GitHub	No	Manual	Website
BioNumerics	Reads/ contigs	Proprietary/BLASTN	Bespoke	Proprietary	NA	Manual	GUI
EnteroBase	Reads	UBLAST/USEARCH	NA	NA	NA	NA	Website
MOST [14]	Reads	Mapping	FreeBSD	GitHub	No	Manual	Command line
mlst*	Contigs	BLASTN	GPL2	GitHub	No	Brew	Command line
MLST-CGE [16]	Contigs	BLASTN	Apache 2	Bitbucket	No	Docker	Command line/Website
MLSTcheck [17]	Contigs	BLASTN	GPL3	GitHub	Yes	CPAN, Docker	Command line
SeqSphere+ [18]	Contigs	NA	Bespoke	Proprietary	NA	Manual	GUI
SRST2 (24)	Reads	Mapping	BSD	GitHub	Yes	Apt, pip	Command line
stringMLST [21]	Reads	*k*-mer	Bespoke	GitHub	No	Manual	Command line

*https://github.com/tseemann/mlst

(5) simple dependency management and installation,

(6) validated with automated tests to verify functionality works as intended,

(7) transparency of algorithm,

(8) and scalability to large numbers of isolates.

The interfaces to the software applications fall into two categories, those that operate on the command line and those that have a graphical interface. Command line input allows for high throughput analysis, but has a high barrier to entry for non-technical users. Graphical interfaces, such as websites, provide point and click interfaces that non-technical users find easier to use initially; however, they are often limited to the analysis of a few samples at a time. To reduce the impact of this limitation, some websites precompute results by downloading raw data directly from the short-read archives (EnteroBase: http://enterobase.warwick.ac.uk) [11].

Most of the software packages are available under open-source licences, with source code available in public repositories, such as GitHub (https://github.com). Source-code availability facilitates transparency for the underlying methods. Comprehensive automated tests, if designed correctly, ensure stability within software applications. Applications packaged for easy installation and dependency management such as: Apt (Debian), Homebrew, Docker, PyPy and CPAN allow for the software to be installed in one step, allowing for immediate use by a range of users. An overview of MLST software applications follows.

Antibiotic Resistance Identification By Assembly (ARIBA) (https://github.com/sanger-pathogens/ariba) takes raw reads as input on the command line, and uses a combination of mapping and local *de novo* assembly to calculate alleles. Like SRST2, it can be used more generally for gene detection and classification, allowing for antibiotic-resistance prediction, virulence-gene detection and plasmid replication gene classification. It is open source, has extensive unit tests and is packaged for easy installation.

Bacterial Isolate Genome Sequence Database (BigsDB) [11] is a web service whose primarily purpose is the management of sequence typing databases, as opposed to querying them. It is used by the majority of schemes as the backend for storing their typing data. The database can be queried in two ways, via a web interface or programmatically through a REST API. There is no described command line interface for queries; however, the mechanisms are in place to allow for it in the future. BigsDB can be used to create new MLST schemes.

BioNumerics (http://www.applied-maths.com/bionumerics) from Applied Maths is a commercial application that is widely used by public-health laboratories to calculate STs. Due to its proprietary nature, a full review is not possible; however, the authors described a reads-based *k*-mer sequence typing method in a patent [12] and do assembly-based sequence typing using BLASTN.

EnteroBase (http://enterobase.warwick.ac.uk) is a web resource that incorporates sequencing data from both public databases and directly from users for four genera (*Salmonella*, *Escherichia*, *Yersinia* and *Moraxella*), and assembles it *de novo* with an adjusted pipeline using SPAdes [13]. EnteroBase succeeds the University College Cork/Warwick MLST database (http://www.mlst.net/databases/), and maintains the database and assigns new alleles of MLST schemes for these genera. These data are mirrored through PubMLST via the EnteroBase API, which is available for all EnteroBase users. Alleles are called using nucleotide and amino acid sequence with USEARCH/UBLAST, which allows for high sensitivity for divergent allele variants. However, the source code is not publicly available.

Metric-Oriented Sequence Typer (MOST) [14] builds upon SRST (version 1) [15] and uses a mapping-based approach to align alleles to reads, with a traffic light system indicating the confidence in the ST calling. One major difference to SRST2 is that it takes a 100 base flanking region around the locus from a reference genome, reducing the impact of coverage drop off at the ends of the sequences. Additionally, it can assign predicted serovars to *Salmonella* isolates. It is used by Public Health England on clinical isolates and has strict, well-defined conservative criteria for calling STs to ensure accuracy. *mlst* (https://github.com/tseemann/mlst) takes *de novo* assemblies as input on the command line and uses BLASTN to align sequences to alleles. It is very fast and searches all databases on pubMLST to automatically detect the organism, then calculates the ST. Installation is very easy using *brew*.

MLST from the Center for Genomic Epidemiology (MLST-CGE) [16] is a web-based method for calculating MLST. It can take assembled genomes or raw sequencing reads. If raw sequencing reads are provided, it performs a *de novo* assembly. Alleles are called using a BLAST-based method.

MLSTcheck [17] takes *de novo* assemblies as input on the command line and uses BLASTn to align sequences to alleles. It is packaged for easy dependency installation, and has unit test coverage. It produces a multi-FASTA alignment of concatenated allele sequences for each sample, which allows for phylogenetic trees to be easily reconstructed. Novel allele sequences are saved to allow for them to be submitted to the MLST curators.

SeqSphere+ [18] from Ridom is a commercial application that is widely used by public-health laboratories. It uses assembled sequences to call STs. It is packaged for easy installation and consists of a large suite of analysis pipelines for automated sequence analysis. Due to its proprietary nature, a full review is not possible.

Short Read Sequence Typing 2 (SRST2) [19] takes raw reads as input on the command line and uses a mapping-based approach to align reads to the alleles. It is packaged for easy dependency installation and is widely used for a variety of applications in addition to MLST including: antibiotic-resistance prediction, virulence-gene detection and serotyping

[20]. The software licence is free for both commercial and non-commercial use, and it has unit tests.

stringMLST [21] takes raw reads as input on the command line and uses *k*-mers to detect MLST alleles. Instead of detecting allele coverage or parsing for potential SNPs, an allele call is made by identifying the allele with the most number of matching *k*-mers. The use of *k*-mers gives a substantial speed advantage, but at the expense of accuracy. This method is fast enough to detect STs in real time during sequencing, so it holds much promise for the future. It is free for non-commercial purposes and it has no automated tests.

The described applications were optimized to work with MLST. Their performance on higher resolution schemes, such as ribosomal MLST, core genome MLST, and whole genome MLST, is quite different, with most scaling poorly to schemes with hundreds or thousands of genes, as this was a case the applications were never fundamentally designed to handle. Alternative methods are required to cater for these cases; thus, extended schemes are not covered in this paper.

DATABASE AVAILABILITY

The availability of databases containing alleles and ST profiles for different species is an important aspect of any MLST software application as outlined in Table 2, since this dictates how easy it is to use the software. These databases also need to be kept up to date, as the underlying schemes are constantly being extended as new isolates are sequenced. Out of date databases can mean that rapidly emerging clonal expansions may be missed, impairing epidemiological investigations. ARIBA, BioNumerics, *mlst*, MLSTcheck, stringMLST, SeqSphere+ and SRST2 all provide automated scripts/methods to download all of the latest databases from pubMLST [11], which are immediately ready to use. This provides immediate access to schemes for over 125 species.

Table 2. Overview of the MLST databases available with each software application.

Software	Automated download	Bundled DBs	Age of bundled DBs*	DBs ready to use
ARIBA	Yes	0	–	Yes
BioNumerics	Yes	0	–	Yes
mlst	Yes	125	1 month	Yes
MLSTcheck	Yes	0	–	Yes
MOST	No	6	>1 year	Yes
SeqSphere+	Yes	0	–	Yes
SRST2	Yes	0	–	Yes
stringMLST	Yes	128	1 month	Yes

DB, Database.

*The age of the bundled databases was calculated on the 15 March 2017.

mlst and stringMLST go one step further and additionally bundle all available databases in their software repository, which are regularly updated. MOST does not provide an automated method for downloading new or updated databases, instead directing researchers to a set of manual steps. They do provide a small number of bundled databases (six and nine, respectively); however, these only represent a fraction of the currently available databases on pubMLST. The databases bundled with MOST were last updated in December 2015, so are missing all recent updates and additions to the schemes, including new STs, so researchers cannot be certain novel results are indeed novel.

EVALUATION

A full comparison could only be performed with the six open-source command line MLST software applications, ARIBA (v2.7.2), *mlst* (v2.8), MLSTcheck (v2.1.1630910), MOST (v 2e3da07), SRST2 (v0.2.0) and stringMLS*T* (v0.3.6). Comparisons of the accuracy of results were performed for the two commonly used commercial applications, BioNumerics (v7.6.2) and Ridom SeqSphere +v4.0.0 (2017–04); comparable computational performance evaluations were not possible; however, these are secondary to accuracy. BigsDB and EnteroBase were excluded as they are web services with extensively featured pipelines and the computational performance of the MLST calling component could not be measured independently. MLST-CGE was excluded because an essential internally hosted software repository was unavailable at the time of testing. Partial results are available for EnteroBase for some datasets, where relevant.

Each application was evaluated on four different datasets, two real and two simulated. Dataset 1 contained 85 samples from standard sets of outbreak data from the Gen-FS WGS Standards and Analysis working group (available from https://github.com/WGS-standards-and-analysis/datasets). Dataset 2 consisted of 72 *Salmonella* samples from EnteroBase, which represent samples that have both MLST data based using traditional capillary sequencing and using Illumina NGS technologies. Dataset 3 consisted of artificially generated reads with varying levels of coverage. From this, the minimum sequence depth required for each software application could be calculated. Dataset 4 consisted of artificially generated reads from two different *Salmonella* serovars where all alleles differ, mixed in different ratios out of a total depth of coverage of 50×. The accuracy of applications could then be determined with mixed samples (a common case) and the point at which the mixed samples became detectable.

The experiments for Dataset 1 were performed using the CDC compute infrastructure. For the rest of the experiments [2–4], we used the MRC CLIMB OpenStack cloud [22] as the base platform for the evaluations. Each of the applications was run in their own Docker container [23] available from GitHub (https://github.com/andrewjpage/docker_mlst). The Debian Testing distribution was used as the base operating system for all containers as it provides access to a

large range of up-to-date bioinformatics software. The host VM had four cores and 32 GB of RAM running Ubuntu 16.04 (LTS); however, only a single core was used for the evaluations. All datasets used for this analysis are available for download as described in the data bibliography or from the public archives using the accession numbers in the Supplementary Material. Where assemblies were required as input to MLST applications, the raw reads were *de novo* assembled with SPAdes (v3.9.0) [13] using the default parameters. SPAdes was chosen as it is widely used and consistently produces high-quality results on bacterial data [24]. All experiments using the two commercial applications, BioNumerics and SeqSphere+, were performed using the CDC compute infrastructure with default options and the SPAdes assemblies as described above.

REAL OUTBREAK DATASETS

Standard datasets (https://github.com/WGS-standards-and-analysis/datasets), covering *L. monocytogenes* from stone fruit [25], *E. coli* from sprouts [26], *C. jejuni* from raw milk (http://www.outbreakdatabase.com/details/hendricks-farm-and-dairy-raw-milk-2008/) and *S. enterica* from spicy tuna [27], comprising 85 samples, were analysed by each of the software applications. These are real outbreak datasets where there were substantive epidemiological investigations and full details are available [9]. No false positives were reported by any application, they made either the correct call, a low-confidence call or no call. A summary of the overall performance is provided in Table 3, with extended details available in Table S1 (available with the online Supplementary Material). There was a wide variation in the results, with only three applications (stringMLST, BioNumerics and MLSTcheck) correctly calling all of the STs. MOST failed to confidently call any of the spicy tuna *Salmonella* samples, but did identify the correct STs, flagged as low confidence (amber). There was a 29-fold variation in the running times between the applications (stringMLST vs SRST2) using raw reads as input (Table 1). This extra computation imposes financial costs and increases the analysis time after sequencing.

Table 3. Summary of performance of each algorithm on real outbreak data for four different species (85 samples)

Software	Total time (min)	Correct ST (%)	No call/low confidence (%)
ARIBA	109.5	98.8	1.2
BioNumerics	NA	**100**	**0**
mlst*	1.9 (+2873)	96.5	3.5
MOST†	1189.7	49.4	50.6
MLSTcheck*	63.8 (+2873)	**100**	**0**
SeqSphere+	NA	96.5	3.5
SRST2	2380.2	95.3	4.7
stringMLST	**80.8**	**100**	**0**

Values in bold indicate the best results in each column.

*The time to assemble with SPAdes before running the applications was 2873 min and is included separately.

†MOST identified the correct ST in 97.6 % of cases, but flagged 48.2 % of these calls as low confidence.

COMPARISON TO CAPILLARY DATA

This dataset consisted of 72 *Salmonella* samples that had been sequenced using traditional capillary sequencing (originally deposited in http://mlst.warwick.ac.uk, now available through EnteroBase) and sequenced using NGS. This allowed for technology independent validation of the NGS MLST software applications. The samples covered a wide range of *Salmonella*, from hosts including humans, reptiles, birds and farm/domestic animals, and from the environment, collected between 1940 and 2014. The dataset contained an estimated 32 different STs, with 38 of the samples predicted to have a serovar of Typhimurium, which causes severe disease in a wide range of hosts, including humans. Full details of the samples (including accession numbers) and results are in Table S2. The ST calls matched in 89 % (64/72) of cases between the capillary data and the NGS MLST software applications, which additionally includes MLST results from the EnteroBase website. Two samples (RKS1252 and RKS1256) were suspected sample swaps with each other. The sample E698 differed between the capillary sequencing results and all other methods with no overlapping alleles. It is possibly a sample swap with another unknown sample or the original sample contained multiple strains. For OLC-1602 and 556-59/192, six out of seven alleles matched in all of the results, but the capillary sequencing data reported a single different allele. Whilst capillary sequencing data is recognized by the community as a gold standard, it is not error free [28], with calls sometimes made using a single read, leaving little resilience to sequencing errors. As the NGS data had very high depth of coverage (over 30×) of this allele, it is likely that the NGS results were correct. Nearly all of the calls from MOST were low confidence (rated amber); however, they correlated with the results from the other applications, and it is just that MOST has very stringent, validated, criteria for calling an ST. Three samples were flagged by multiple applications as problematic; however, in every case the capillary sequencing data, stringMLST, EnteroBase, SeqSphere+ and BioNumerics confidently called an ST, indicating a contaminant has been missed. Eight applications flagged sample 139K as problematic; however, stringMLST confidently called an ST, indicating overconfidence in ST calling. MLSTcheck and BioNumerics called a different ST for 2 samples; however, this appears to be due to duplicate allele profiles in the underlying database at pubMLST. Overall, we conclude that whilst the MLST results between capillary sequencing data and NGS data are nearly identical, the MLST based on NGS data is more accurate and reliable when presented with edge cases.

IMPACT OF DEPTH OF COVERAGE

The impact of depth of coverage over the MLST genes was assessed by artificially generating perfect paired-ended reads with a length of 125 bases and a median insert size of 400 bases with varying levels of coverage using FASTAQ (v3.14.0).

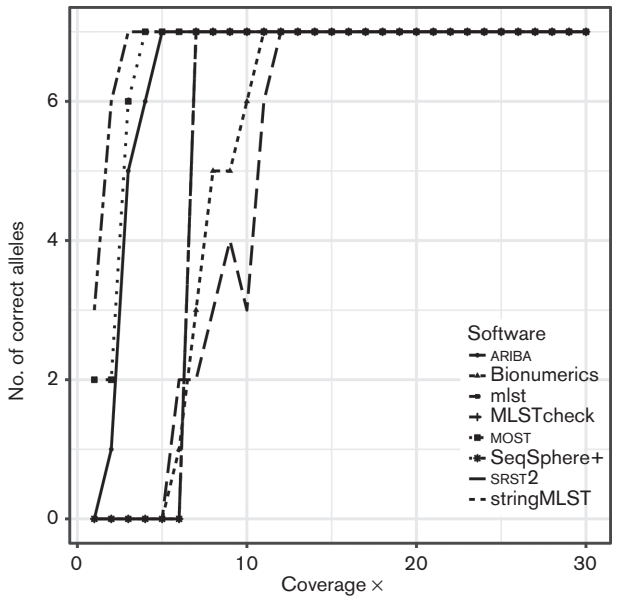

Fig. 1. Number of correct calls of each application as coverage increases. Each ST consists of seven alleles, and all seven must be correctly and confidently called to calculate a ST.

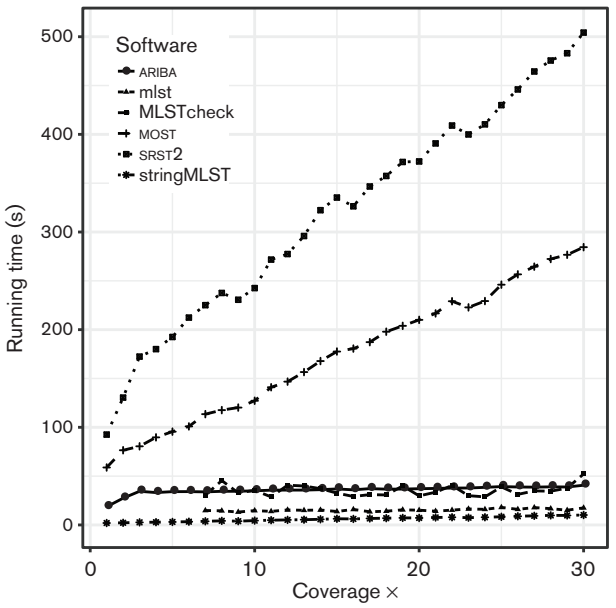

Fig. 2. Running time (s) of each application as the coverage increases to assess the impact of the depth of coverage. No assembled contiguous sequences could be generated where the coverage was less than 7×, as such no data was recorded for the reliant methods (*mlst* and MLSTcheck). No performance results are available for BioNumerics or SeqSphere+.

The allele sequences plus 500 base flanking regions were extracted from *S. enterica* Typhi CT18 [29], accession number AL513382, and artificial paired-end reads were generated with mean depths of coverage from 1× to 30×. The simulated reads were free from sequencing errors to allow for the effect of coverage alone to be measured. Therefore, the minimum effective depth of coverage for each application could be tested. All applications could accurately call STs when the coverage was greater than 12×; however, below this the minimum depth of coverage applications required varied greatly, as shown in Fig. 1. stringMLST correctly called the ST with just 3× coverage; however, it gave false-positive results for lower coverage alleles. ARIBA correctly called the ST from 5× with no false-positive results. SRST2 correctly called the ST from 12× coverage with no false-positive results; however, it did correctly identify the ST from 6× with low confidence.

The computational resources required varied greatly with stringMLST taking just 10 s to call an ST with 30× coverage, as shown in Fig. 2, and the final disk space requirements were negligible, as shown in Fig. 3. Whilst minimizing the disk space resources needed for the application is generally positive, stringMLS*T* does not output enough information about the allele calls to allow for further analysis, for example, to interrogate a false-positive result. The time to call an ST at 30× with ARIBA was 40 s with 0.1 Mbytes output data. The disk-space requirement is higher than stringMLST, but provides the allele assemblies used to call the ST, which is useful for further analysis. SRST2 is an order of magnitude slower, taking over 500 s to call an ST at 30×. The disk space required for the final output is also very substantial at 147 Mbytes, which equates to a storage cost of 475 bytes per base of sequencing as shown in Fig. 3. While MOST

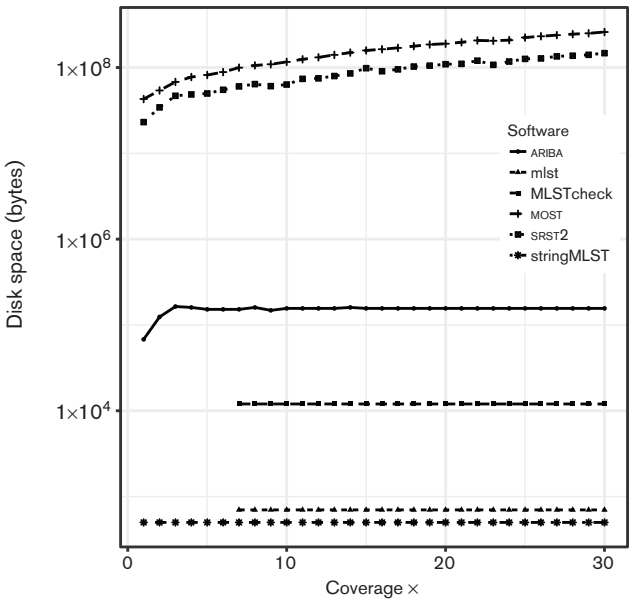

Fig. 3. Disk space requirements in bytes for each software application as the depth of coverage increases. Due to the large difference between applications, a log scale is used.

confidently correctly called each individual allele from 4×, the overall ST call was flagged as low confidence below 10× due to its inherently conservative nature. The running time given for *mlst* and MLSTcheck includes the *de novo* assembly time with SPAdes, which accounts for most of the running time. MLSTcheck takes on average four times longer (25 s per sample) to return a result than *mlst* (5.9 s per sample), with the final results between the two being identical.

IMPACT OF MIXED SAMPLES

Contamination and mixed colonies are a standard complexity in microbiology [30]. To understand the behaviour of the different MLST software applications in the presence of more than one strain, we constructed a simulated dataset consisting of two *Salmonella* samples with different alleles in varying ratios. This allowed us to see at what point contamination/mixed strains becomes detectable. Once detected, we would expect an MLST application to flag the results as low confidence or provide no result at all to avoid false positives. The flip side of this is that if algorithms are too sensitive to low level contamination and sequencing errors, they become less useful on real world applications, so need to be tolerant to some low-level noise.

The allele sequences plus 500 base flanking regions were extracted from *S. enterica* Typhi CT18 [29], accession number AL513382, and *S. enterica* Weltevreden 10 259 [31], accession number LN890518. Artificial paired-end reads were generated using FASTAQ to give a total coverage of 50×, beginning with CT18 at 1× and 10 259 at 49× in a single FASTQ file. The coverage of each sample was varied in steps of 1× to generate a dataset of 49 FASTQ files. Fig. 4 shows that the accuracy of the software varies, but follows a general pattern, calling the sample with the highest coverage at the highest levels, with uncertainty in the middle as the proportion of the two samples becomes similar. The worst case is where a software application calls an ST with high confidence that is not in the underlying data (false positive), and only occured with stringMLST. MOST and ARIBA are highly conservative, detecting that there are mixed samples when the samples are at very low levels of coverage (at 4–5×). MLSTcheck, *mlst*, SeqSphere+ and BioNumerics all performed identically, with the performance linked to how well SPAdes assembled the underlying genomes. There was no clear boundary with SRST2 and it varied between high-quality calls and low-confidence calls as the mixing of the samples changed.

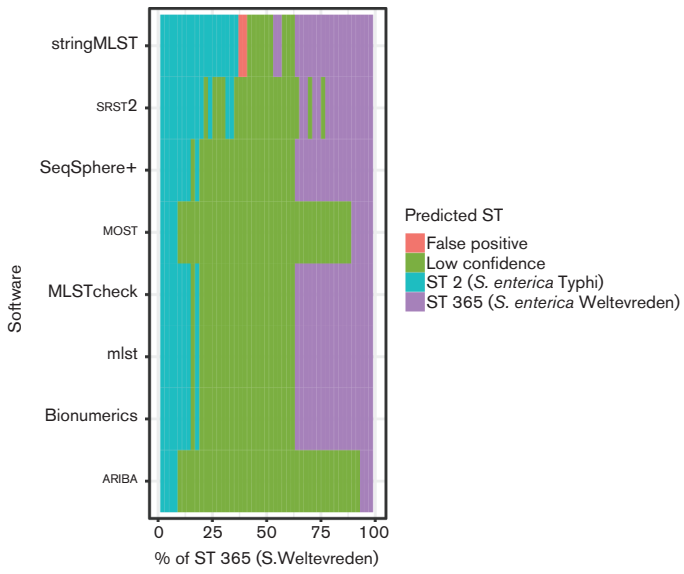

Fig. 4. STs called by each software application when given data containing two different *Salmonella* samples in varying ratios of abundance. Where there is no ST called, or where the ST has any ambiguity at all, it is marked as low confidence. A false positive is where an ST is called with high confidence and is not one of the two samples in the raw data.

CONCLUSION

It is clear that not all MLST calling applications function as expected. Problems with some software include: out of date databases, computationally inefficient methods, false-positive results, inability to call alleles at low coverage and variable performance in the presence of mixed samples. Therefore, there is scope for improvement. Overall though, these software applications' ST calls using NGS data are concordant with traditional MLST calling methods based on capillary sequencing data, perform moderately well with low mean genome coverage, and are sometimes able to report low confidence when faced with contamination.

Funding information
This work was made possible through support from the Advanced Molecular Detection (AMD) Initiative at the Centers for Disease Control and Prevention. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention. This work was supported by the Wellcome Trust (grant WT 098051). NFA has support from the Wellcome Trust (202792/Z/16/Z)

Acknowledgements
Thanks to João Carriço, Miguel Machado, Anthony Underwood, Martin Hunt, King Jordan, Lavanya Rishishwar, Peter Gerner-Smidt and Simon Harris for their helpful discussions and suggestions. We wish to thank Mark Achtman, Zhemin Zhou and Martin Sergeant for feedback on this manuscript.

Conflicts of interest
The authors declare that there are no conflicts of interest.

Data bibliography
1. Parkhill J. *Salmonella enterica* subsp. enterica serovar Typhi CT18, EMBL AL513382 (2002).

2. Andrew J. Page. *Salmonella enterica* subsp. enterica serovar Weltevreden 10259, EMBL LN890518 (2016).

References
1. **Silk BJ, Mahon BE, Griffin PM, Gould LH, Tauxe R V *et al.*** Vital signs: listeria illnesses, deaths, and outbreaks - United States, 2009–2011. *Morb Mortal Wkly Rep* 2013;62:448–452.

2. **Siegman-Igra Y, Levin R, Weinberger M, Golan Y, Schwartz D *et al.*** *Listeria monocytogenes* infection in Israel and review of cases worldwide. *Emerg Infect Dis* 2002;8:305–310.

3. Scharff RL, Besser J, Sharp DJ, Jones TF, Peter GS *et al.* An economic evaluation of PulseNet: a network for foodborne disease surveillance. *Am J Prev Med* 2016;50:S66–73.
4. Ragon M, Wirth T, Hollandt F, Lavenir R, Lecuit M *et al.* A new perspective on *Listeria monocytogenes* evolution. *PLoS Pathog* 2008;4:e1000146.
5. Achtman M, Wain J, Weill FX, Nair S, Zhou Z *et al.* Multilocus sequence typing as a replacement for serotyping in *Salmonella enterica*. *PLoS Pathog* 2012;8:e1002776.
6. Swaminathan B, Gerner-Smidt P, Ng LK, Lukinmaa S, Kam KM *et al.* Building PulseNet International: an interconnected system of laboratory networks to facilitate timely public health recognition and response to foodborne disease outbreaks and emerging foodborne diseases. *Foodborne Pathog Dis* 2006;3:36–50.
7. GMI Steering Committee. *Global Microbial Identifier Charter and Structure*. www.globalmicrobialidentifier.org/about-gmi/charter-and-structure. Global Microbial Identifier Steering Committee: Lyngby; 2013.
8. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE *et al.* Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci USA* 1998;95:3140–3145.
9. Timme RE, Rand H, Shumway M, Trees EK, Simmons M. Bacterial pathogen genome datasets for bioinformatics pipelines. *PLOS Currents: Tree of Life* 2017 (in press).
10. Page AJ, de Silva N, Hunt M, Quail MA, Parkhill J *et al.* Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data. *Microb Genom* 2016;2:e000083.
11. Jolley KA, Maiden MC. BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* 2010; 11:595.
12. Pouseele H, Janssens K 2016. Method of typing nucleic acid or amino acid sequences based on sequence analysis. WIPO. https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2016124600.
13. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.
14. Tewolde R, Dallman T, Schaefer U, Sheppard CL, Ashton P *et al.* MOST: a modified MLST typing tool based on short read sequencing. *PeerJ* 2016;4:e2308.
15. Inouye M, Conway TC, Zobel J, Holt KE. Short read sequence typing (SRST): multi-locus sequence types from short reads. *BMC Genomics* 2012;13:338.
16. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H *et al.* Multilocus sequence typing of total-genome-sequenced bacteria. *J Clin Microbiol* 2012;50:1355–1361.
17. Page AJ, Taylor B, Keane JA. Multilocus sequence typing by BLAST from de novo assemblies against PubMLST. *JOSS* 2016;1: JPage2016.
18. Jünemann S, Sedlazeck FJ, Prior K, Albersmeier A, John U *et al.* Updating benchtop sequencing performance comparison. *Nat Biotechnol* 2013;31:294–296.
19. Inouye M, Dashnow H, Raven LA, Schultz MB, Pope BJ *et al.* SRST2: rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med* 2014;6:90.
20. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA *et al.* Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. *Proc Natl Acad Sci USA* 2015;112:E3574–E3581.
21. Gupta A, Jordan IK, Rishishwar L. stringMLST: a fast k-mer based tool for multilocus sequence typing. *Bioinformatics* 2017;33:119–121.
22. Connor TR, Loman NJ, Thompson S, Smith A, Southgate J *et al.* CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community. *Microb Genom* 2016;2:e000086.
23. di Tommaso P, Palumbo E, Chatzou M, Prieto P, Heuer ML *et al.* The impact of Docker containers on the performance of genomic pipelines. *PeerJ* 2015;3:e1273.
24. Page AJ, de Silva N, Hunt M, Quail MA, Parkhill J *et al.* Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data. *Microb Genom* 2016;2:e000083.
25. Chen Y, Burall LS, Luo Y, Timme R, Melka D *et al.* Isolation, enumeration and whole genome sequencing of *Listeria monocytogenes* in stone fruits linked to a multistate outbreak. *Appl Environ Microbiol* 2016;82:7030–7040.
26. Multistate outbreak of Shiga toxin-producing *Escherichia coli* O121 infections linked to raw clover sprouts. 2014. www.cdc.gov/ecoli/2014/o121-05-14/index.html.
27. Hoffmann M, Luo Y, Monday SR, Gonzalez-Escalona N, Ottesen AR *et al.* Tracing origins of the *Salmonella* Bareilly strain causing a food-borne outbreak in the United States. *J Infect Dis* 2016;213: 502–508.
28. Liu L, Li Y, Li S, Hu N, He Y *et al.* Comparison of next-generation sequencing systems. *J Biomed Biotechnol* 2012;2012:1–11.
29. Parkhill J, Dougan G, James KD, Thomson NR, Pickard D *et al.* Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature* 2001;413:848–852.
30. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* 2014;12:87.
31. Makendi C, Page AJ, Wren BW, Le Thi Phuong T, Clare S *et al.* A phylogenetic and phenotypic analysis of *Salmonella enterica* serovar Weltevreden, an emerging agent of diarrheal disease in tropical regions. *PLoS Negl Trop Dis* 2016;10:e0004446.