Journal of Microbiological Methods 160 (2019) 11–19

Contents lists available at ScienceDirect

Journal of Microbiological Methods

journal homepage: www.elsevier.com/locate/jmicmeth

Identification of bacterial isolates from a public hospital in Australia using complexity-reduced genotyping

Berenice Talamantes-Becerra[a,*], Jason Carling[b], Karina Kennedy[c], Michelle E. Gahan[d], Arthur Georges[a]

[a] Institute for Applied Ecology, University of Canberra, ACT 2601, Australia
[b] Diversity Arrays Technology Pty Ltd, Canberra, ACT 2617, Australia
[c] Canberra Health Services, Departments of Microbiology and Infectious Diseases, Canberra Hospital, Yamba Drive, Garran 2605, Australia
[d] National Centre for Forensic Studies, University of Canberra, ACT, 2617, Australia

ARTICLE INFO

Keywords:
Bacterial pathogens
Genotyping
GBS
Bacterial identification
Strain typing

ABSTRACT

Bacterial identification methods used in routine identification of pathogens in medical microbiology include a combination approach of biochemical tests, mass spectrometry or molecular biology techniques. Extensive publicly-available databases of DNA sequence data from pathogenic bacteria have been amassed in recent years; this provides an opportunity for using bacterial genome sequencing for identification purposes. Whole genome sequencing is increasing in popularity, although at present it remains a relatively expensive approach to bacterial identification and typing. Complexity-reduced bacterial genome sequencing provides an alternative. We evaluate genomic complexity-reduction using restriction enzymes and sequencing to identify bacterial isolates. A total of 165 bacterial isolates from hospital patients in the Australian Capital Territory, between 2013 and 2015 were used in this study. They were identified and typed by the Microbiology Department of Canberra Public Hospital, and represented 14 bacterial species. DNA extractions from these samples were processed using a combination of the restriction enzymes *Pst*I with *Mse*I, *Pst*I with *Hpa*II and *Mse*I with *Hpa*II. The resulting sequences (length 30–69 bp) were aligned against publicly available bacterial genome and plasmid sequences. Results of the alignment were processed using a bioinformatics pipeline developed for this project, Currito3.1 DNA Fragment Analysis Software. All 165 samples were correctly identified to genus and species by each of the three combinations of restriction enzymes. A further 35 samples typed to the level of strain identified and compared for consistency with MLST typing data and *in silico* MLST data derived from the nearest sequenced candidate reference. The high level of agreement between bacterial identification using complexity-reduced genome sequencing and standard hospital identifications indicating that this new approach is a viable alternative for identification of bacterial isolates derived from pathology specimens. The effectiveness of species identification and in particular, strain typing, depends on access to a comprehensive and taxonomically accurate bacterial genome sequence database containing relevant bacterial species and strains.

1. Introduction

Methods for accurate bacterial identification are critical for medical diagnosis and treatment of infectious diseases. Traditional identification based on the study of phenotypic characteristics or biochemical testing provide an inexpensive option. However, phenotypic similarities between bacterial strains may lead to incorrect diagnosis (Buszewski et al., 2017). In the last decade, sequencing techniques such as 16S rRNA gene sequencing and other PCR methods have provided novel approaches for bacterial identification (Buszewski et al., 2017). The use of mass spectrometry technologies, such as matrix assisted laser desorption ionization-time of flight mass spectrometry (MALDI–TOF), have decreased the time taken for pathogen identification, reducing the probability of inadequate decisions in patient treatment (Florio et al., 2018; Maurer et al., 2017). Although these techniques are useful tools for identification, their resolution may be limited or biased. For example, multiple differing copies of the 16S rRNA gene in some bacterial genomes can cause difficulties in the interpretation of sequencing data (Conville and Witebsky, 2007; Louca et al., 2018). MALDI-TOF is considered to be useful in the clinical context due to low cost and rapid turnaround time. MALDI-TOF has some limitations for bacterial identification and sometimes produces unexpected results that are difficult

* Corresponding author.
E-mail address: Berenice.TalamantesBecerra@canberra.edu.au (B. Talamantes-Becerra).

https://doi.org/10.1016/j.mimet.2019.03.016
Received 10 January 2019; Received in revised form 16 March 2019; Accepted 17 March 2019
Available online 17 March 2019
0167-7012/ © 2019 Elsevier B.V. All rights reserved.

to interpret, requiring further testing for clarification (Van Belkum et al., 2017). The reference databases used for identification with MALDI-TOF may present limitations for uncommon organisms, and laboratory or even instrument-specific references have been used (Williams et al., 2003). Problems encountered in MALDI-TOF with uncommon organisms would be expected to decrease as the size and breadth of available reference databases continues to increase. Whole genome sequencing (WGS) for routine identification is a step forward in accurate diagnosis of bacterial disease, but its application in bacterial identification is cost-effective only in limited scenarios, and it also has a long turnaround time, making it unsuitable for routine clinical use. The general absence of expertise and software tools to transform WGS data into diagnostic results is a further limitation (Köser et al., 2012; Quainoo et al., 2017).

Existing options for bacterial identification vary in resolution, from that of 16S rRNA sequencing and MALDI-TOF to high-resolution whole genome sequencing, but there are few options between these extremes. DArTseq, a genotyping by sequencing (GBS) method, promises to fill this gap. It has been successfully applied to a wide range of plants, animals and fungi for measuring genetic diversity (Baloch et al., 2017; Egea et al., 2017; Garavito et al., 2016), in breeding trials of plants and animals (dos Santos et al., 2016; Valdisser et al., 2017), in ecological studies (Lambert et al., 2016), and in studies of phylogenetic relationships (Georges et al., 2018). The method has been seldom used in bacterial studies, yet it has the potential to deliver a novel and cost-effective approach for bacterial identification. The DArTseq method involves sequencing complexity-reduced genomic representations produced by digestion with a pair of restriction enzymes, followed by PCR amplification of a subset of restriction fragments (Ren et al., 2015). The advantage of complexity-reduced genomic representations is that they obtain a reproducible subset of the genome, which can then be sequenced at a considerably lower cost than a whole genome, whilst still providing a large amount of sequence information for identification *via* alignment against genome sequence databases (Al-Beyroutiová et al., 2016).

Here we compare DArTseq complexity-reduced genotyping by sequencing with established methods for bacterial identification in a public hospital in Australia. We show that DArTseq can provide an alternative to traditional methods for obtaining high resolution DNA sequence-based results for bacterial identification and strain typing at a lower cost than whole genome sequencing. A bioinformatics pipeline, Currito3.1 DNA Fragment Analysis Software, was developed to automate the data analysis. Automation enables time-savings, facilitates reproducibility and reduces the probability of human mistakes. The final report produced by the analytical pipeline indicates the closest identified bacterial reference sequence including genus, species and in some cases, strain typing information.

2. Methods

2.1. Bacterial strains

Bacterial isolates ($n = 165$) were provided by the Microbiology Department of Canberra Public Hospital, Australia from mainly clinical patient specimens ($n = 150$) and environmental samples ($n = 15$). All isolates were cultured and identified using standard clinical laboratory techniques, with identification confirmed by MALDI-TOF (Bruker Daltonics, Leipzig, Germany). These identification results were considered as a standard against which the experimental results were compared. There were 76 Gram negative isolates from a culture collection of carbapenem resistant *Enterobacteriales* (*Citrobacter amalonaticus, Citrobacter freundii, Enterobacter aerogenes, Enterobacter cloacae* complex, *Escherichia coli, Providencia rettgeri, Hafnia alvei, Klebsiella oxytoca, Klebsiella pneumoniae, Morganella morganii* and *Serratia marcescens)* isolated from patient clinical and surveillance specimens and environmental specimens between 2009 and 2015 (Table S1). The presence of the carbapenemase genotype was confirmed by commercial nucleic acid amplification techniques (Xpert Carba-R, Cepheid, Sunnyvale, CA, USA or CRE, AusDiagnostics, Mascot, NSW, Australia). Eighty-nine Gram positive isolates (*Enterococcus faecium* and *Enterococcus faecalis* clinical and surveillance specimens and *Staphylococcus aureus* blood culture isolates*)* from 2013 to 2015 were also included. All isolates were stored in glycerol at −80 °C within the Microbiology Department, until the time of the study, at which stage they were thawed and inoculated onto Horse Blood Agar (HBA) Columbia, then incubated at 35.5 ± 0.5 °C for 24 h prior to DNA extraction.

DNA extractions were performed on all bacterial cultures using a chloroform-isoamyl alcohol method (Green and Sambrook, 2017). This method was selected as a safety measure, because commercial DNA extraction kits may not inactivate spores of some pathogenic bacterial strains (Dauphin et al., 2009; Panning et al., 2007). The protocol used is a modified and simplified version from Moore et al. (2004). Samples of isolated bacterial colonies were taken with a bacteriological loop and suspended into 1.5 mL sterile tubes containing a lysis mix made with 467 μL EB Buffer (10 mM Tris-Cl, pH 8.5), 30 μL 10% SDS, 2 μL RNAse A (10 μg/μL) and 3 μL Proteinase K (20 mg/mL). Tubes were incubated for 1 h at 35.5 ± 0.5 °C. Subsequently, an equal volume of chloroform-isoamyl alcohol solution (24:1) was added into each tube, and mixed 40 times by inversion. Tubes were centrifuged at 15,000*g* for 20 min. The supernatant (approx. 200 μL to 400 μL) was transferred into 2 mL sterile tubes containing 700 μL of isopropanol 99.5% and mixed forty times by inversion. Tubes were centrifuged at 15,000*g* for 20 min. The supernatant was discarded carefully without removing the pellet, then, 700 μL of freshly prepared ethanol 70% *v*/v was added and mixed by vortex. Tubes were centrifuged at 15,000*g* for 20 min; the supernatant was discarded, and tubes were placed in a desiccator to remove the remaining ethanol. Tubes were frequently inspected to avoid excessive dryness of pellets. Finally, a volume between 30 and 200 μL EB Buffer was added to dissolve the pellet. Volumes added varied according to the size of the pellet. DNA quality and concentrations were determined by 0.8% agarose gel electrophoresis. DNA extractions showing a high molecular weight band in the gel were considered successful. DNA extractions that showed degradation were repeated by regrowing bacterial cultures and performing DNA extraction 24 h after inoculation. Samples that presented residues and heavily concentrated DNA bands were purified using a 96-well plate Zymo ® DNA clean and concentrate kit SKU D4017 (Integrated Sciences, Chatswood, NSW, Australia). All samples were assigned into 96-well skirted PCR plates and stored at −20 °C.

2.2. Library preparation and sequencing

Complexity-reduced genotyping was applied to all bacterial isolates. The restriction enzymes *Pst*I (5′-CTGCA|G-3′), *Mse*I (5′-TTA|A-3′) and *Hpa*II (5'-CCG|G-3′) were used in combination: *Pst*I with *Mse*I, *Pst*I with *Hpa*II and *Mse*I with *Hpa*II. The three combinations were tested in identification of all bacterial isolates, and for strain typing in a selected subset. The choice of restriction enzymes used for complexity reduction impacts the size and composition of the genomic fraction obtained. The number of bases in the recognition sites of the enzymes primarily determines the fragment quantity, and the base composition and GC balance impacts the spread and location of restriction fragments obtained. The enzyme *Pst*I is a primary factor in limiting the representation size because of its six base recognition sequence, in comparison to *Mse*I and *Hpa*II with four base recognition sites. Complexity-reduction methods used in DArTSeq for eukaryotic genomes are generally based around *Pst*I and a second enzyme. The two initial complexity reduction methods *Pst*I with *Hpa*II and *Pst*I with *Mse*I were chosen to provide GC or AT rich alternatives with different fragment set selections. A third complexity-reduction method based on *Mse*I as the primary enzyme was developed by designing additional oligo-nucleotide barcode adapter

sequences.

To evaluate the complexity-reduction process, genomic DNA of *Escherichia coli* O157 (EDL 933) IRMM449 Sigma-Aldrich (Castle Hill, NSW, Australia) certified reference standard, GenBank accession number AE005174.2 (Agarwala et al., 2018), genome size of 5,639,399 bp (Perna et al., 2001), was also processed using the same three restriction enzyme combinations.

Library construction methods followed the procedure described in Ren et al. (2015), but differing in the choice of restriction enzymes. Briefly, digestions were performed with the selected pairs of restriction enzymes and PCR adapters were ligated. Two adapters were used, one corresponding to each restriction enzyme. The adapter design included Illumina flow-cell specific sequences required for bridge PCR in cluster generation, as well as a barcode region to enable sample multiplexing. The adapters were designed such that only fragments with differing restriction sites at each end were capable of cluster generation (Ren et al., 2015). Equal volumes of PCR products were pooled together, purified with a QIAGEN QIAquick PCR Purification Kit Cat No./ID: 28106 (QIAGEN, Chatstone, Victoria, Australia) and added into sequencing lanes. The bacterial libraries can be 'spiked' on top of other multiplexed GBS libraries, representing only a small portion of the total sequencing flowcell capacity, since only approximately 100,000–150,000 reads per sample were required. Clustering was done according to Illumina protocols using a HiSeq SR Cluster Kit V4 recipe v9.0 and HiSeq SR Flow Cell v4 (Illumina Inc., San Diego, CA, US). For sequencing, the Flow Cell was loaded according to the Illumina protocols on a HiSeq 2500 sequencer, using HiSeq SBS kit v4 for a total of 77 cycles (Georges et al., 2018).

Technical replicate assays were performed on 11 samples: 6 samples of *E. faecium*, 4 samples of *S. aureus* and the single *E. coli* certified reference. Technical replicate assays were performed twice for all samples except the *E. coli* reference, which was assayed with six technical replicates for each enzyme combination. Separate processing was carried out for the technical replicates starting from genomic DNA, with independent library construction, sequencing, and data analysis.

2.3. Data analysis

Raw data obtained from the sequencer in the form of fastQ files were demultiplexed to produce one fastQ file for each sample assayed. The reads were filtered according to Phred scores (Ewing et al., 2005), with a higher stringency applied to the barcode region of the sequence read to ensure correct demultiplexing, following the methods described in Georges et al., (2018). Subsequently, the barcode region was removed from the reads, leaving 69-bp sequences. Each fastQ file was condensed into a fastQcol file which contained each unique sequence present in the original fastQ file, along with the respective read counts and the mean quality score at each base (Ren et al., 2015).

As a first stage for running the analytical pipeline, the set of fastQcol files for all samples to be analysed was grouped into a single tabular data file which contained all unique sequences present across the complete sample set, along with the read counts for these sequences across all samples. Each unique sequence was also represented by a SeqIndex providing a unique identifier. The reverse adapters which were present on sequences derived from fragments shorter than 69 bp were identified and trimmed, resulting in sequences of variable length. Sequences which were less than 30 bp were removed, as they are less suitable for BLAST alignment.

To process the data, we developed a bioinformatic analytical pipeline, Currito3.1 DNA Fragment Analysis Software, which we describe here. The pipeline was developed to process the DNA sequence tags obtained from complexity reduced genotyping, in order to identify the bacterial isolates to genus and species level, and select the closest matching strain from among a bacterial genome sequence database. Currito3.1 uses as input complexity-reduced genotyping data in the format described above, proceeding first with a BLAST alignment (McGinnis and Madden, 2004) of the sample sequences against all complete bacterial genome assemblies and plasmids in the NCBI nt collection database (Agarwala et al., 2018) to identify the best candidate bacterial genomes for each sample. The following BLAST parameters were used: word size 12, bitscore 50, evalue 0.000001, percentage identity 80%, and percentage query cover 80%. Candidates are identified according to the number of sequence tags obtaining a best or equal best BLAST hit to each reference, as measured by bit score. After identifying candidate genomes, the sequences derived from each sample are used to BLAST against the top three closest identified genomes individually. The individual BLAST against each candidate reference genome is computationally less intensive than the global BLAST against the full sequence database, allowing the BLAST parameters to be better optimised for short sequence queries (word size 10). The output produced by the pipeline describes the results obtained from the BLAST of all sequences from a sample against each candidate reference genome, including the following values: sample name, sample ID, subject accession, subject title, subject sequence length, sum of aligned sequence lengths, number of alignment positions, coverage length, coverage percentage, mean percentage identity, maximum percentage identity, minimum percentage identity, mean of gap length, maximum gap length, number of gaps, sum of gaps, number of overlaps, sum of overlaps, number of zero length gaps, nucleotide sequence distance (NSD), number of alignments for each of the top three candidates, size of the intersection between the top three candidates, size of the union between top three candidates, and Nei and Li distance (Nei and Li, 1979) between the candidates. Additionally, the analysis provides the total number of sequences derived from the sample with and without hits to each candidate reference, and those which did not obtain BLAST hits to the reference but gave a BLAST hit for a plasmid are included for comparison. The Currito3.1 pipeline uses the NSD to determine the best matching candidate for each sample. The NSD calculation is:

$$NSD = -\frac{3}{4}\ln\left[1 - \frac{4}{3}\left(\frac{S}{I+S}\right)\right]\left[1 - \frac{G}{T}\right] + \frac{G}{T}$$

NSD is a DNA sequence distance measurement which considers the number of identical nucleotides (I), substitutions (S) and gap openings (G) across all aligned sequences (Dams et al., 1987; Huysmans, 1986; Jukes and Cantor, 1969; Van de Peer et al., 1990), and the sum of these three variables (T) to produce a global distance value (Van de Peer et al., 1990). Lower NSD values in samples are associated with closer relatedness to the reference genome.

Although the analytical pipeline is designed for use with isolated bacterial strains, in some instances more than a single organism may be present in a sample. To test for this, the relationship between the top candidate reference genomes is described by calculating the Nei and Li distance (Nei and Li, 1979) derived from the proportion of sequences with BLAST hits in common between the candidate references. When the presence of more than a single species is indicated, the identity and alignment statistics for the additional organisms are considered.

The output of the analytical pipeline provides the identity of the selected candidate reference genome with the lowest NSD value, and also includes a number of descriptive plots providing results for the top three candidates identified from the first stage BLAST analysis such as: a) circular genome alignment plots indicating the size of the reference genome, position of fragments aligned and percentage identity of each fragment alignment; b) Bar plots to show the total number of sequences obtained, the total number of sequences with and without BLAST hits to each candidate reference or BLAST hits to plasmids; c) histograms for BLAST %Identity values.

Examples of figures generated in the report output by Currito3.1 DNA Fragment Analysis Software for a single sample of *S. aureus* processed using the *Mse*I with *Hpa*II complexity reduction method are shown. Sequence tags with BLAST alignments to the candidate reference genomes were plotted using Circos (Krzywinski et al., 2009) as

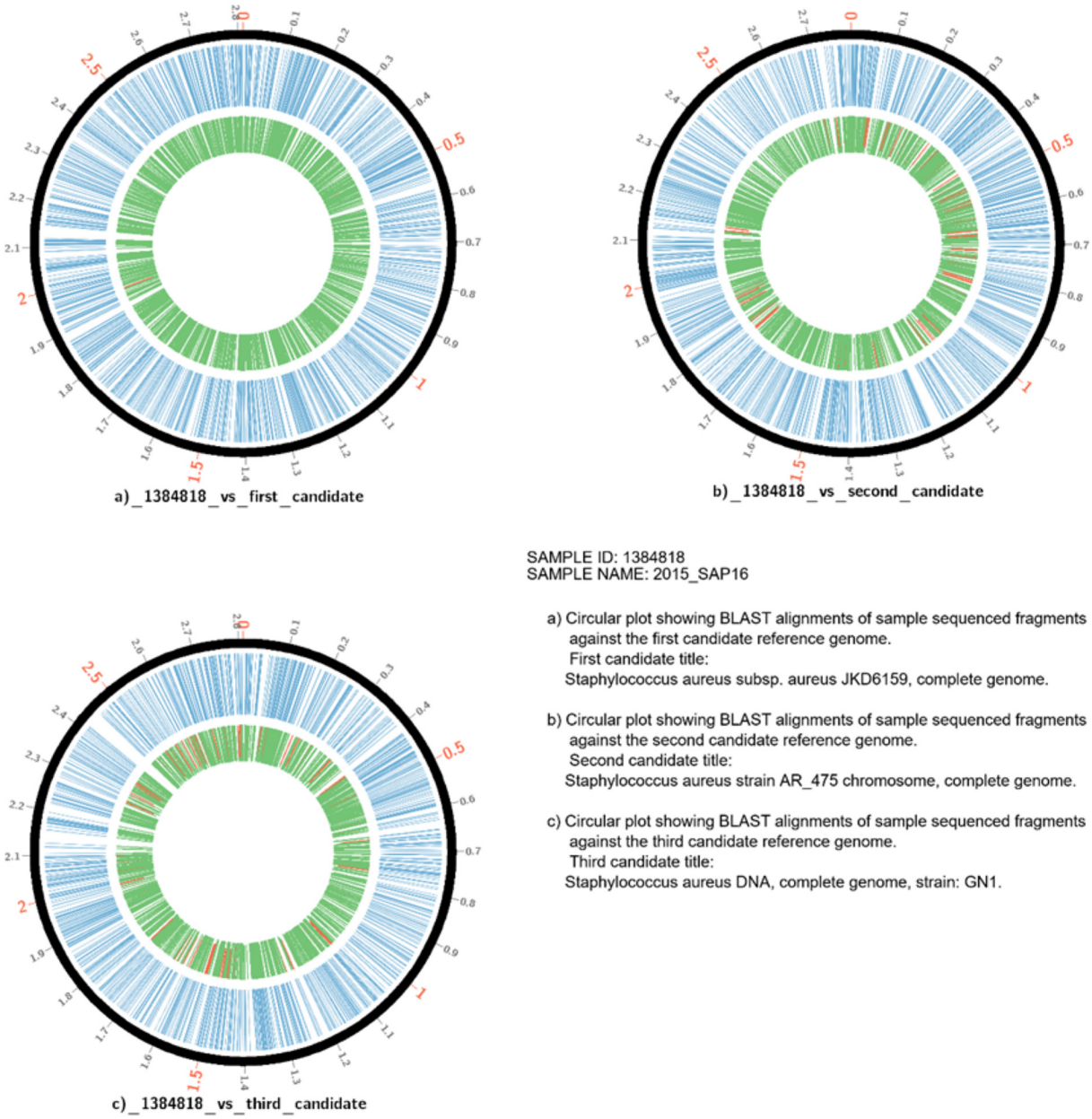

Fig. 1. Extract of report generated by bioinformatics pipeline Currito3.1 DNA Fragment Analysis Software. Shown here are the circular alignment plots of a single sample of *S. aureus* using the *Mse*I with *Hpa*II method and indicates BLAST alignment positions of sequence fragments obtained, for the top three candidate references (a, b c). The outer black circle represents the candidate reference genome with size indicated in megabases (Mb). The middle blue circle shows aligned sequenced fragments obtained by complexity reduced genotyping. The inner green/red circle shows the percentage identity of the alignments. Values below 95% are red, values greater than or equal to 95% are green. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

displayed in Fig. 1. For this sample, most of the sequences obtained for the best candidate (labelled as 'best candidate') have a BLAST alignment %identity above 95% and there are no large gaps between aligned fragments. A histogram of the distribution of BLAST alignment % identity values for each of the candidates is shown in Fig. 2. Additionally, the total number sequences with and without BLAST alignments against the selected top candidate reference, and those with BLAST hits to a plasmid sequence but not to the candidate reference, is shown in Fig. 3.

2.4. Strain typing

Reduced representation genotyping provides coverage of the genome up to 10%, which potentially enables some sequence-based strain typing information to be derived from the sequences, however in practice, this level of coverage would be insufficient to reliably capture allele-specific information from targeted loci. Existing widely used typing schemes such as core gene MLST rely on highly targeted sequence information which will only be partially represented in reduced representation sequence data. More recently, whole genome sequence data has been used to derive typing information such as MLST profiles *in silico* (Schürch et al., 2018). Potentially, reduced representation sequence data can be used indirectly to infer *in silico* typing information when a fully matching bacterial strain is found in the bacterial genome sequence database (Lüth et al., 2018). To test this, *in silico* MLST profiles were produced from the best matching candidate reference genome for each sample, and then compared with the MLST typing data produced directly from the isolates by standard PCR methods. The publicly available software package MLST2.0 (Larsen et al., 2012), which uses bacterial genome sequence data, in conjunction with MLST

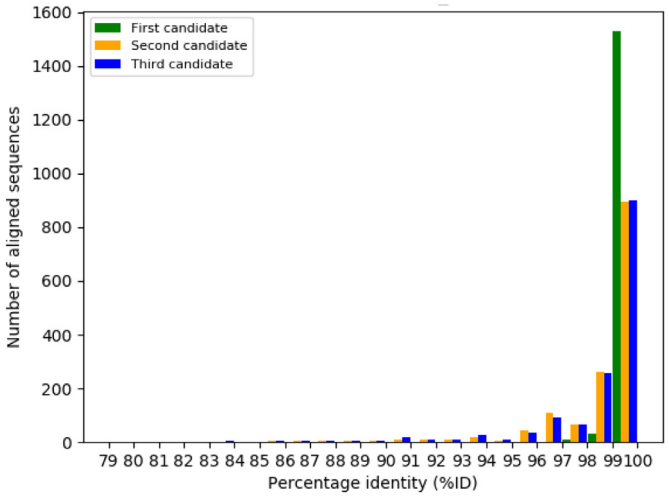

Fig. 2. Extract of report. Histogram showing BLAST alignment %identity of all BLAST alignments against the candidate references, where the X axis shows the BLAST alignment %identity and the Y axis shows the number of aligned sequences.

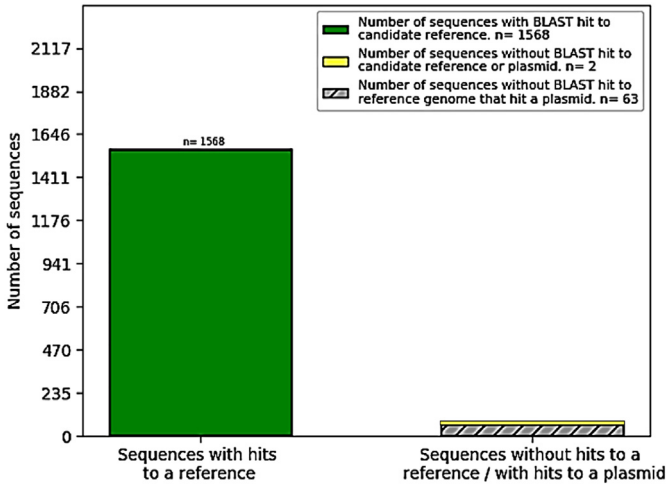

Fig. 3. Extract of report. Bar plot showing total sequences with and without BLAST alignments against the selected best candidate reference, and those with BLAST hits to a plasmid sequence but not to the candidate reference. The X axis shows the sequence classification. The Y axis shows the total number of sequences.

profile databases, was integrated into the Currito3.1 analytical pipeline to perform this task, adding to the final report the MLST type and alleles obtained for each sample.

3. Results

3.1. Comparison of complexity-reduction methods

The complexity-reduction enzyme combination of *Mse*I with *Hpa*II produced the highest average genome coverage of 4.78% and the highest average number of obtained complexity-reduced fragments. The other enzyme combinations showed an average genome coverage of 2.48% for *Pst*I with *Hpa*II and 2.25% for *Pst*I with *Mse*I. The percentage of successful identification, average genome coverage and number of sequence fragments are shown in Table 1. For all enzyme combinations, all samples could be identified to species level.

Despite the differing number of fragments, all three enzyme combinations identified correctly the bacterial isolates at the genus and species level (Table 2). There is a high level of agreement among results

Table 1
Results of identification up to genus, species and genome coverage per complexity-reduction method tested for all bacterial isolates.

	*Pst*I-*Hpa*II	*Pst*I-*Mse*I	*Mse*I-*Hpa*II
Genus level % ID	100.00%	100.00%	100.00%
Species level % ID	100.00%	100.00%	100.00%
Average of % genome coverage	2.48%	2.25%	4.78%
Average number of restriction fragments per method	1944	1430	4092

Table 2
Tally of identification success to species level for three complexity reduction methods.

		*Pst*I-*Hpa*II	*Pst*I-MseI	MseI-*Hpa*II
Species name	N			
Enterococcus faecium	59	59	59	59
Staphylococcus aureus	29	29	29	29
Enterobacter cloacae complex	17	17	17	17
Citrobacter freundii	16	16	16	16
Klebsiella pneumoniae	15	15	15	15
Klebsiella oxytoca	12	12	12	12
Enterobacter asburiae	5	5	5	5
Escherichia coli	4	4	4	4
Morganella morganii	2	2	2	2
Citrobacter amalonaticus	1	1	1	1
Enterobacter aerogenes	1	1	1	1
Enterococcus faecalis	1	1	1	1
Hafnia alvei	1	1	1	1
Providencia rettgeri	1	1	1	1
Serratia marcescens	1	1	1	1
TOTAL	**165**	**165**	**165**	**165**

obtained for the three enzyme combinations, which in many cases identified the same strain from among the reference genome assemblies as the closest matching candidate (Table 2).

3.2. Technical replication and E. coli certified reference

The results of the 11 samples performed with technical replication showed complete agreement for identification to the genus and species level, giving 100% repeatability for the identification assay.

Identification results obtained from the genomic DNA of *E. coli* O157 (EDL 933) IRMM449 Sigma-Aldrich certified reference (Perna et al., 2001) showed a 100% match for genus, species and strain level for all methods. Among 6 technical replicates, the enzyme pair *Mse*I with *Hpa*II showed an average genome coverage of 10.41% and an average of total sequences obtained of more than 10,000 fragments. The BLAST alignment %Identity and NSD values obtained against the *E. coli* O157 (EDL 933) genome sequence provide a benchmark for the amount of sequence error in the assay results. For example, the average NSD values showed less than 1 bp of difference per 10,000 bp of sequence fragments obtained for the *Pst*I with *Hpa*II enzyme combination. The results obtained for each method using the DNA of *E. coli* O157 (EDL 933) are shown in Table 3. A circular plot providing the alignment locations of sequence tags from the *Mse*I with *Hpa*II enzyme combination against *E. coli* O157 (EDL 933) is shown in Fig. 4.

3.3. Strain typing

A total of 35 samples of the species *C. freundii, K. pneumoniae* and *E. coli* were examined for consistency between the standard MLST typing data and *in silico* MLST data produced from the nearest sequenced candidate references. Results of this comparison showed that *in silico* derived MLST data was accurate when a matching strain was present in the genome reference database, as indicated by the NSD value (Table 4). For the *Pst*I with *Mse*I enzyme combination, NSD values

Table 3
Results of identification up to genus, species and genome coverage for *E. coli* O157 (EDL 933) IRMM449 Sigma-Aldrich certified reference (Perna et al., 2001), for six technical replicates.

	*Pst*I-*Hpa*II	*Pst*I- *Mse*I	*Mse*I-*Hpa*II
% genus level ID	100.00%	100.00%	100.00%
% species level ID	100.00%	100.00%	100.00%
Average genome coverage	2.64%	2.34%	10.41%
BLAST alignment % ID	99.9915%	99.9974%	99.9846%
Sequence tags in common between replicates	99.09%	99.73%	98.84%
Average number of restriction fragments	2433	1836	10,602
Average NSD	0.000103	0.000040	0.000136

associated with correctly matching MLST profiles were ≤ 0.000241, and ≤ 0.000384, for *C. freundii* and *K. pneumoniae* respectively. Values below these NSD thresholds resulted in accurate MLST prediction (Table 4). Correctly matching MLST allele profiles were obtained in all four samples of *E. coli* tested. The results confirm that accurate *in silico* MLST prediction can be achieved when the genome sequence database contains a matching strain. The distribution of NSD values for samples of *C. freundii* and *K. pneumoniae* with matching and non-matching *in silico* MLST profiles for three complexity-reduction enzyme combinations is shown in Fig. 5.

4. Discussion

The bacterial identification from the complexity-reduced genome sequencing used in this study agreed to species level with the identifications performed by the Microbiology Department of Canberra Public Hospital for all isolates, indicating that this approach is a viable alternative for identification of bacterial isolates derived from pathology specimens. Improvements can be made in establishing a comprehensive and taxonomically accurate bacterial genome sequence database containing relevant bacterial strains. Were such a resource available, we believe that the complexity-reduced genome sequencing would also be effective in typing at the level of strain.

Standardisation and curation of taxonomic information is essential for reliable sample identification. For this study, genome assemblies lacking species level taxonomic declarations were disregarded. Correct and consistent species level identification of samples belonging to certain species complexes within the *Enterbacteriales* required the recognition of species synonyms and potential inconsistency in species assignments among the genome accessions present in the NCBI nt collection sequence database. It is widely recognized that the taxonomy of strains belonging to the *E. cloacae* species complex contains a high level of confusion and synonymous species naming. The identification results for samples from the *E. cloacae* complex were considered correct if the best candidate was among the six species known to be part of the complex: *Enterobacter cloacae, Enterobacter asburiae, Enterobacter hormaechei, Enterobacter kobei, Enterobacter ludwigii* and *Enterobacter nimipressuralis* (Mezzatesta et al., 2012). The genus *Klebsiella* also presents taxonomic complexities. Identification of *Klebsiella quasipneumoniae* and *Klebsiella variicola* remains a challenge in general, and phenotypic methods, biochemical tests or mass spectrometry are unable to differentiate these species accurately (Fonseca et al., 2017). For identification purposes, *K. quasipneumoniae* (Arena et al., 2015) is considered a synonym of the opportunistic pathogen *K. pneumoniae*. *Enterobacter aerogenes* is considered as a synonym of *K. aerogenes* (Iyer et al., 2017). Similarly, *K. michiganensis* is considered synonym of *K. oxytoca* (Dantur et al., 2018, 2015; Saha et al., 2013).

In the study, some samples were found to contain more than one bacterial species and were not pure strain isolates. This could be detected by the presence of taxonomically distinct bacteria within the genome assemblies indicated as candidate matches. This was measured by applying the Nei and Li (1979) distance to determine the relative proportion of aligned sequences in common for each candidate, where a low proportion of sequences in common between candidates is indicative of a likely non-clonal culture. When presence of multiple organisms in a sample is indicated, each organism is considered separately. In this study, seven samples were seen to contain two identifiable bacterial species (Table S2). The reason for the presence of more than a single organism in these cultures could not be determined in this study, although one possible explanation is contamination of the isolates. For example, all isolates were initially identified and stored at −80 °C, and subsequently regrown for this study. The presence of a second organism in these cultures may be caused by contamination introduced at storage or regrowth of the isolates. Owing to the time elapsed between the initial isolation and identification, and subsequent regrowth, it was not possible to perform new isolations from the original samples. For the purpose of determining successful identification rates, the presence of the expected species within the sample was

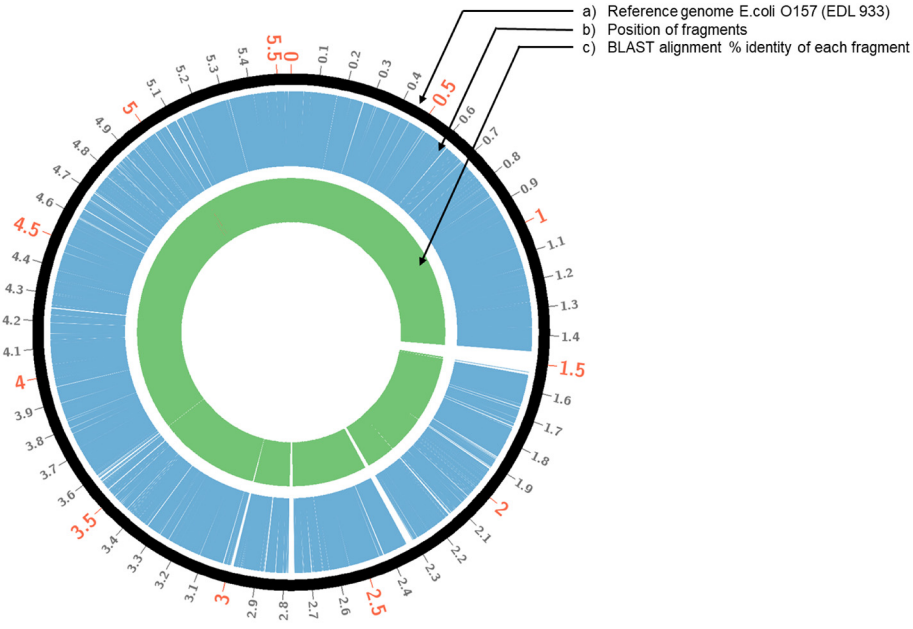

Fig. 4. Circular plot showing restriction fragments obtained for the method *Mse*I with *Hpa*II against the certified reference model genome. a) Total length of 5.6 Mb of the reference genome *E. coli* O157 (EDL 933) IRMM449 Sigma-Aldrich (Perna et al., 2001). b) Position of fragments obtained for complexity-reduction method *Mse*I with *Hpa*II. c) Percentage identity of alignments in comparison to the reference. Green indicates a percentage identity greater than or equal to 95% and red indicates percentage identity below 95%. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Comparison of MLST typing with *in silico* derived MLST, considering the average NSD obtained for each complexity-reduced method.

Species name	Number of samples tested	*Pst*I-*Hpa*II		*Pst*I- *Mse*I		*Mse*I-*Hpa*II	
		Samples correctly typed	Avg NSD of correctly typed samples	Samples correctly typed	Avg NSD of correctly typed samples	Samples correctly typed	Avg NSD of correctly typed samples
C. freundii	16	3	0.000217	3	0.000241	3	0.000459
K. pneumoniae	15	5	0.000263	5	0.000384	5	0.000346
E. coli	4	4	0.001511	4	0.001642	4	0.000822
Total	**35**	**12**		**12**		**12**	

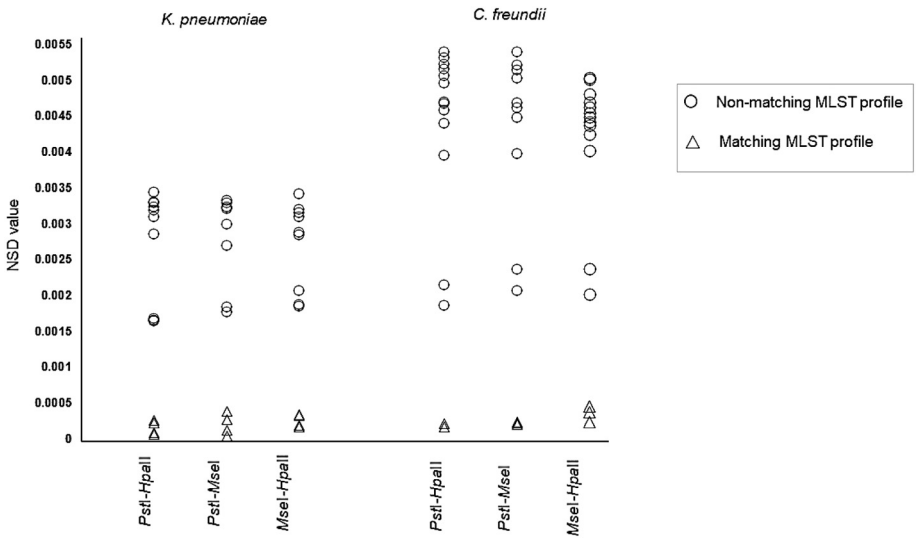

Fig. 5. Matching and non-matching MLST profiles for *C. freundii* and *K. pneumoniae*. The values of Nucleotide Sequence Distance (NSD) for best candidate reference assemblies is shown for 15 samples of *K. pneumoniae* and 16 samples of *C. freundii,* for three complexity reduction methods. All samples were correctly identified to the species level. NSD values for samples with correctly matching *in silico* MLST profiles are represented by triangles.

considered a correct identification, in spite of a potentially contaminating species being present.

The comparison of MLST data with *in silico* derived MLST data for *C. freundii, K. pneumoniae* and *E. coli* was accurate in those instances where a matching strain was present in the genome reference database, as indicated by the NSD value. The threshold NSD values associated with accurate *in silico* MLST profiling differed for each of the species. Values below the observed NSD thresholds for each species resulted in accurate MLST prediction for *C. freundii and K. pneumoniae*. In the case of *E. coli*, the NSD threshold value could not be identified as the *in silico* MLST profiles matched the standard profile in all samples tested. The results confirm that accurate *in silico* MLST prediction can be achieved from complexity reduced genotyping results.

For any identification or typing technology, turnaround time is an important consideration for clinical use. The components of the assay can be broken down as follows, DNA extraction, library preparation, DNA sequencing, primary data analysis, and secondary data analysis. In the current study, DNA extractions were performed manually, however, automated extraction protocols and systems for use with bacterial isolates are available (Bird et al., 2018), with DNA extraction taking approximately 1 h. Library preparation time for the assay was approximately 4 h. Sequencing was performed on an Illumina HiSeq 2500, although the use of an Illumina MiSeq sequencer would offer faster turnaround time, with an expected run time of 6 h for the current assay (Quick et al., 2015). Primary sequence data analysis involves the conversion of raw sequencer output into quality filtered FastQ files, followed by demultiplexing, requiring approximately 1 h to process the MiSeq run data. Lastly, the secondary data analysis, performed using the Currito3.1 pipeline, requires 2 h to process the results for 96 samples to completion. The total turnaround time of 14 h cannot match the speed of standard benchtop microbiological tests and MALDI-TOF but would be comparable or better than other sequence-based techniques, offering the higher resolution afforded by DNA sequence-based methods. The DNA barcoding system allows for multiplexing of up to 2300 samples into a single sequencing lane. For a HiSeq 2500 v4 sequencer, with each of the 8 lanes producing approximately 220 million reads, up to 1465 samples could be multiplexed per lane at the read depths used in this study. Conversely, for a MiSeq v3 producing approximately 25 million reads per flow-cell, up to 165 samples could be processed per run. The cost per sample including library preparation and sequencing would be expected to be 7 dollars (USD). This technology would be particularly well suited for use in outbreak management and infection control programs as a cost-effective alternative to whole genome sequencing.

5. Conclusions

Complexity-reduced genotyping by is a viable alternative for identification of bacterial isolates derived from pathology specimens. The three restriction enzyme pairs *Pst*I with *Mse*I, *Pst*I with *Hpa*II and *Mse*I with *Hpa*II tested for this project agreed with all of the identifications performed by the hospital pathology department using standardised methods. In comparison with the other complexity reduction combinations, the *Mse*I with *Hpa*II method yielded the expected higher number of restriction fragments and provided greater genome coverage. In spite of the differences in the number and composition of restriction fragments, the three methods identified all samples correctly. The methods also gave the same results when tested for strain typing *via in silico* MLST profiling. Owing to the differences in the number of sequence tags produced, the subsequent genome alignment coverage differed approximately two-fold between the *Pst*I and *Mse*I based methods. This difference in coverage did not impact identification and typing results for the samples used in this study. When using bacterial sequence information for identification and particularly strain typing, the question arises of where to draw distance thresholds for determining clonality and what level of sequence dissimilarity is

meaningful in the context of providing clinical results. It is clear that these questions need to be answered on a per species basis at least and will require careful consideration of clinical phenotypic and genotypic information together. These issues are faced by all of the sequence based bacterial identification technologies whether using core gene sequences, whole genome sequences or complexity-reduced genome sequences (Schürch et al., 2018). The use of a comprehensive bacterial genome database, which has been curated for taxonomic accuracy is a key to use of this technology. The NCBI nt collection database provided a large and readily available source of genome sequences for testing the use of DArTseq in bacterial identification and typing. Routine use of complexity-reduced genome sequence data to identify and type bacterial isolates will require the development of a properly curated database, excluding problems caused by taxonomic errors and ambiguities. A curated database of this type is being developed for use with sequence-based identification of food borne pathogens, with international co-operation for the establishment of a data-sharing framework (Lüth et al., 2018). The Pathosystems Resource Integration Center (PATRIC) is another example, which provides a growing database of genomic information, linked to phenotypic data such as anti-microbial resistance (Wattam et al., 2017). Continuing improvements in the size and scope of the available genome databases is on-going. The inclusion of genome sequences from strains of local origin would help to improve the chances of finding an exact match, and deriving typing information.

Acknowledgements

The author B. Talamantes-Becerra, would like to acknowledge Consejo Nacional de Ciencia y Tecnología (CONACYT) for providing a scholarship “Becas CONACYT al extranjero 2015” to pursue postgraduate studies. We would also like to thank Susan Bradbury, Chief Scientist Microbiology Dept of ACT Pathology for providing samples used in this study, in addition to identification and typing data. We thank Prof. Dennis McNevin and Dr. Andrzej Kilian for their suggestions on project development and methods and for co-supervising the PhD project from which this work arises.

Data accessability

Data used for this paper are contained in Supplementary Materials and sequencing data has been lodged with Data in Brief.

Author contributions

BTB and JC conceived the project. KK selected and provided bacterial isolates from a public hospital in Australia. BTB undertook laboratory work. BTB and JC created bioinformatic pipeline for data analysis. AG contributed expertise in statistical analysis and AG and MEG supervised the PhD project from which this work arises. BTB led the writing of the manuscript to which all authors contributed.

Conflicts of interest

JC is a full-time employee of DArT and BTB is enrolled in a PhD which involves the use of DArT Technology. The remaining authors declare no conflict of interest.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.mimet.2019.03.016.

References

Agarwala, R., Barrett, T., Beck, J., Benson, D., Bollin, C., Bolton, E., Bourexis, D., Brister, R., Bryant, S., Canese, K., Charowhas, C., Clark, K., DiCuccio, M., Dondoansky, I., Federhen, S., Feolo, M., Funk, K., Geer, L., Gorelenkov, V., Hoeppner, M., Holmes, B., Johnson, M., KarschMizrachi, I., Khotomlianski, V., Kimchi, A., Kimelman, M., Kitts, P., Klimke, W., Krasnov, S., Kuznetsov, A., Landrum, M., Landsman, D., Lee, J., Lipman, D., Lu, Z., Madden, T., Madej, T., Marchler-Bauer, A., Murphy, T., O'Sullivan, C., Orris, R., Ostell, J., Panchenko, A., Phan, L., Preuss, D., Pruitt, K., Rodarmer, K., Rubinstein, W., Sayers, E., Schneider, V., Schuler, G., Sherry, S., Sirotkin, K., Siyan, K., Slotta, D., Soboleva, A., Soussov, V., Starchenko, G., Tatusova, T., Todorov, K., Trawick, B., Vakatov, D., Wang, Y., Ward, M., Wilbur, J., Yaschenko, E., Zbicz, K., 2018. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 46, D8–D13. https://doi.org/10.1093/nar/gkx1095.

Al-Beyroutiová, M., Sabo, M., Sleziak, P., Dušinský, R., Birčák, E., Hauptvogel, P., Kilian, A., Švec, M., 2016. Evolutionary relationships in the genus *Secale* revealed by DArTseq DNA polymorphism. Plant Syst. Evol. 302, 1083–1091. https://doi.org/10.1007/s00606-016-1318-2.

Arena, F., Henrici De Angelis, L., Pieralli, F., Di Pilato, V., Giani, T., Torricelli, F., D'Andrea, M.M., Rossolini, G.M., 2015. Draft genome sequence of the first Hypermucoviscous *Klebsiella quasipneumoniae* subsp. *quasipneumoniae* isolate from a bloodstream infection. Genome Announc. 3https://doi.org/10.1128/genomeA.00952-15. (e00952-15).

Baloch, F.S., Alsaleh, A., Shahid, M.Q., Çiftçi, V., Sáenz De Miera, L.E., Aasim, M., Nadeem, M.A., Aktaş, H., Özkan, H., Hatipoğlu, R., 2017. A whole genome DArTseq and SNP analysis for genetic diversity assessment in durum wheat from central fertile crescent. PLoS One 12, 1–18. https://doi.org/10.1371/journal.pone.0167821.

Bird, P., Benzinger, M.J., Bastin, B., Crowley, E., Agin, J., Goins, D., Armstrong, M., 2018. Evaluation of mericon *E. coli* O157 screen plus and mericon *E. coli* STEC O-type pathogen detection assays in select foods: collaborative study, first action 2017.05. J. AOAC Int. 101, 739–760. https://doi.org/10.5740/jaoacint.17-0301.

Buszewski, B., Rogowska, A., Pomastowski, P., Złoch, M., Railean-Plugaru, V., 2017. Identification of microorganisms by modern analytical techniques. J. AOAC Int. 100, 1607–1623. https://doi.org/10.5740/jaoacint.17-0207.

Conville, P.S., Witebsky, F.G., 2007. Analysis of multiple differing copies of the 16S rRNA gene in five clinical isolates and three type strains of *Nocardia* species and implications for species assignment. J. Clin. Microbiol. 45, 1146–1151. https://doi.org/10.1128/JCM.02482-06.

Dams, E., Yamada, T., De Baere, R., Huysmans, E., Vandenberghe, A., De Wachter, R., 1987. Structure of 5S rRNA in actinomycetes and relatives and evolution of eubacteria. J. Mol. Evol. 25, 255–260.

Dantur, K.I., Enrique, R., Welin, B., Castagnaro, A.P., 2015. Isolation of cellulolytic bacteria from the intestine of *Diatraea saccharalis* larvae and evaluation of their capacity to degrade sugarcane biomass. AMB Express 5. https://doi.org/10.1186/s13568-015-0101-z.

Dantur, K.I., Chalfoun, N.R., Claps, M.P., Tórtora, M.L., Silva, C., Jure, Á., Porcel, N., Bianco, M.I., Vojnov, A., Castagnaro, A.P., Welin, B., 2018. The endophytic strain *Klebsiella michiganensis* Kd70 lacks pathogenic island-like regions in its genome and is incapable of infecting the urinary tract in mice. Front. Microbiol. 9, 1–14. https://doi.org/10.3389/fmicb.2018.01548.

Dauphin, L.A., Moser, B.D., Bowen, M.D., 2009. Evaluation of five commercial nucleic acid extraction kits for their ability to inactivate *Bacillus anthracis* spores and comparison of DNA yields from spores and spiked environmental samples. J. Microbiol. Methods 76, 30–37. https://doi.org/10.1016/j.mimet.2008.09.004.

dos Santos, J.P.R., Pires, L.P.M., de Castro Vasconcellos, R.C., Pereira, G.S., Von Pinho, R.G., Balestre, M., 2016. Genomic selection to resistance to *Stenocarpella maydis* in maize lines using DArTseq markers. BMC Genet. 17, 1–10. https://doi.org/10.1186/s12863-016-0392-3.

Egea, L.A., Mérida-García, R., Kilian, A., Hernandez, P., Dorado, G., 2017. Assessment of genetic diversity and structure of large garlic (*Allium sativum*) germplasm bank, by diversity arrays technology “genotyping-by-sequencing” platform (DArTseq). Front. Genet. 8, 1–9. https://doi.org/10.3389/fgene.2017.00098.

Ewing, B., Ewing, B., Hillier, L., Hillier, L., Wendl, M.C., Wendl, M.C., Green, P., Green, P., 2005. Base-calling of automated sequencer traces using phred. Genome Res. 175–185. https://doi.org/10.1101/gr.8.3.175.

Florio, W., Tavanti, A., Barnini, S., Ghelardi, E., Lupetti, A., 2018. Recent advances and ongoing challenges in the diagnosis of microbial infections by MALDI-TOF mass spectrometry. Front. Microbiol. https://doi.org/10.3389/fmicb.2018.01097.

Fonseca, E.L., Ramos, N. da V., Andrade, B.G.N., Morais, L.L.C.S., Marin, M.F.A., Vicente, A.C.P., 2017. A one-step multiplex PCR to identify *Klebsiella pneumoniae*, *Klebsiella variicola*, and *Klebsiella quasipneumoniae* in the clinical routine. Diagn. Microbiol. Infect. Dis. 87, 315–317. https://doi.org/10.1016/j.diagmicrobio.2017.01.005.

Garavito, A., Montagnon, C., Guyot, R., Bertrand, B., 2016. Identification by the DArTseq method of the genetic origin of the *Coffea canephora* cultivated in Vietnam and Mexico. BMC Plant Biol. 16, 1–12. https://doi.org/10.1186/s12870-016-0933-y.

Georges, A., Gruber, B., Pauly, G.B., White, D., Adams, M., Young, M.J., Kilian, A., Zhang, X., Shaffer, H.B., Unmack, P.J., 2018. Genome-wide SNP markers breathe new life into phylogeography and species delimitation for the problematic short-necked turtles (Chelidae: *Emydura*) of eastern. Australia. Mol. Ecol. 27, 5195–5213. https://doi.org/10.1111/mec.14925.

Green, M.R., Sambrook, J., 2017. Isolation of high-molecular-weight DNA using organic solvents. Cold Spring Harb Protoc 2017, 356–359. https://doi.org/10.1101/pdb.prot093450.

Huysmans, E., 1986. The distribution of 5S ribosomal RNA sequences in phenetic hyperspace. Implications for eubacterial, eukaryotic, archaebacterial and early biotic

evolution. Endocyt. C Res. 3, 133–155.
Iyer, R., Iken, B., Damania, A., 2017. Whole genome of *Klebsiella aerogenes* PX01 isolated from San Jacinto River sediment west of Baytown, Texas reveals the presence of multiple antibiotic resistance determinants and mobile genetic elements. Genomics Data 14, 7–9. https://doi.org/10.1016/j.gdata.2017.07.012.
Jukes, T.H., Cantor, C.R., 1969. Evolution of Protein Molecules, Mammalian Protein Metabolism. Academic Presshttps://doi.org/10.1016/B978-1-4832-3211-9.50009-7.
Köser, C.U., Ellington, M.J., Cartwright, E.J.P., Gillespie, S.H., Brown, N.M., Farrington, M., Holden, M.T.G., Dougan, G., Bentley, S.D., Parkhill, J., Peacock, S.J., 2012. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS Pathog. 8, e1002824. https://doi.org/10.1371/journal.ppat.1002824.
Krzywinski, M.I., Schein, J.E., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., Marra, M.A., 2009. Circos: an information aesthetic for comparative genomics. Genome Res. https://doi.org/10.1101/gr.092759.109.
Lambert, M.R., Skelly, D.K., Ezaz, T., 2016. Sex-linked markers in the North American green frog (Rana clamitans) developed using DArTseq provide early insight into sex chromosome evolution. BMC Genomics 17, 1–13. https://doi.org/10.1186/s12864-016-3209-x.
Larsen, M.V., Cosentino, S., Rasmussen, S., Friis, C., Hasman, H., Marvig, R.L., Jelsbak, L., Sicheritz-Pontén, T., Ussery, D.W., Aarestrup, F.M., Lund, O., 2012. Multilocus sequence typing of total-genome-sequenced bacteria. J. Clin. Microbiol. 50, 1355–1361. https://doi.org/10.1128/JCM.06094-11.
Louca, S., Doebeli, M., Parfrey, L.W., 2018. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. Microbiome 6, 1–12. https://doi.org/10.1186/s40168-018-0420-9.
Lüth, S., Kleta, S., Al Dahouk, S., 2018. Whole genome sequencing as a typing tool for foodborne pathogens like Listeria monocytogenes – the way towards global harmonisation and data exchange. Trends Food Sci. Technol. 73, 67–75. https://doi.org/10.1016/j.tifs.2018.01.008.
Maurer, F.P., Christner, M., Hentschke, M., Rohde, H., 2017. Advances in rapid identification and susceptibility testing of bacteria in the clinical microbiology laboratory: implications for patient care and antimicrobial stewardship programs. Infect. Dis. Rep. 9, 18–27. https://doi.org/10.4081/idr.2017.6839.
McGinnis, S., Madden, T.L., 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. 32, W20–W25. https://doi.org/10.1093/nar/gkh435.
Mezzatesta, M.L., Gona, F., Stefani, S., 2012. *Enterobacter cloacae* complex: clinical impact and emerging antibiotic resistance. Future Microbiol 7, 887–902. https://doi.org/10.2217/fmb.12.61.
Moore, E., Arnscheidt, A., Krüger, A., Strömpl, C., Mau, M., 2004. Simplified protocols for the preparation of genomic DNA from bacterial cultures. In: Akkrmans, A.D.L., Van Elsas, J.D., De Bruijn, F.J. (Eds.), Molecular Microbial Ecology Manual. Kluwer Academic Publishers, The Netherlands, pp. 3–18.
Nei, M., Li, W.H., 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. 76, 5269–5273. https://doi.org/10.1073/pnas.76.10.5269.
Panning, M., Kramme, S., Petersen, N., Drosten, C., 2007. High throughput screening for spores and vegetative forms of pathogenic *B. anthracis* by an internally controlled real-time PCR assay with automated DNA preparation. Med. Microbiol. Immunol. 196, 41–50. https://doi.org/10.1007/s00430-006-0029-7.
Perna, N.T., Plunkett, G., Burland, V., Mau, B., Glasner, J.D., Rose, D.J., Mayhew, G.F., Evans, P.S., Gregor, J., Kirkpatrick, H.A., Pósfai, G., Hackett, J., Klink, S., Boutin, A., Shao, Y., Miller, L., Grotbeck, E.J., Davis, N.W., Lim, A., Dimalanta, E.T., Potamousis, K.D., Apodaca, J., Anantharaman, T.S., Lin, J., Yen, G., Schwartz, D.C., Welch, R.A., Blattner, F.R., 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature 409, 529–533. https://doi.org/10.1038/35054089.
Quainoo, S., Coolen, J.P.M., Van Hijum, S.A.F.T., Huynen, M.A., Melchers, W.J.G., Van Schaik, W., Wertheim, H.F.L., 2017. Whole-genome sequencing of bacterial pathogens: the future of nosocomial outbreak analysis. Clin. Microbiol. Rev. 30, 1015 LP–1063. https://doi.org/10.1128/CMR.00016-17.
Quick, J., Ashton, P., Calus, S., Chatt, C., Gossain, S., Hawker, J., Nair, S., Neal, K., Nye, K., Peters, T., De Pinna, E., Robinson, E., Struthers, K., Webber, M., Catto, A., Dallman, T.J., Hawkey, P., Loman, N.J., 2015. Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella*. Genome Biol. 16, 1–14. https://doi.org/10.1186/s13059-015-0677-2.
Ren, R., Ray, R., Li, P., Xu, J., Zhang, M., Liu, G., Yao, X., Kilian, A., Yang, X., 2015. Construction of a high-density DArTseq SNP-based genetic map and identification of genomic regions with segregation distortion in a genetic population derived from a cross between feral and cultivated-type watermelon. Mol. Gen. Genomics. 290, 1457–1470. https://doi.org/10.1007/s00438-015-0997-7.
Saha, R., Farrance, C.E., Verghese, B., Hong, S., Donofrio, R.S., 2013. *Klebsiella michiganensis* sp. nov., a new bacterium isolated from a tooth brush holder. Curr. Microbiol. 66, 72–78. https://doi.org/10.1007/s00284-012-0245-x.
Schürch, A.C., Arredondo-Alonso, S., Willems, R.J.L., Goering, R.V., 2018. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches. Clin. Microbiol. Infect. 24, 350–354. https://doi.org/10.1016/j.cmi.2017.12.016.
Valdisser, P.A.M.R., Pereira, W.J., Almeida Filho, J.E., Müller, B.S.F., Coelho, G.R.C., de Menezes, I.P.P., Vianna, J.P.G., Zucchi, M.I., Lanna, A.C., Coelho, A.S.G., de Oliveira, J.P., Moraes, A. da C., Brondani, C., Vianello, R.P., 2017. In-depth genome characterization of a Brazilian common bean core collection using DArTseq high-density SNP genotyping. BMC Genomics 18, 1–19. https://doi.org/10.1186/s12864-017-3805-4.
Van Belkum, A., Welker, M., Pincus, D., Charrier, J.P., Girard, V., 2017. Matrix-assisted laser desorption ionization time-of-flight mass spectrometry in clinical microbiology: what are the current issues? Ann. Lab. Med. 37, 475–483. https://doi.org/10.3343/alm.2017.37.6.475.
Van de Peer, Y., Neefs, J.M., De Wachter, R., 1990. Small ribosomal subunit RNA sequences, evolutionary relationships among different life forms, and mitochondrial origins. J. Mol. Evol. 30, 463–476. https://doi.org/10.1007/BF02101118.
Wattam, A.R., Davis, J.J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., Conrad, N., Dietrich, E.M., Disz, T., Gabbard, J.L., Gerdes, S., Henry, C.S., Kenyon, R.W., Machi, D., Mao, C., Nordberg, E.K., Olsen, G.J., Murphy-Olson, D.E., Olson, R., Overbeek, R., Parrello, B., Pusch, G.D., Shukla, M., Vonstein, V., Warren, A., Xia, F., Yoo, H., Stevens, R.L., 2017. Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. Nucleic Acids Res. 45, D535–D542. https://doi.org/10.1093/nar/gkw1017.
Williams, T.L., Andrzejewski, D., Lay, J.O., Musser, S.M., 2003. Experimental factors affecting the quality and reproducibility of MALDI TOF mass spectra obtained from whole bacteria cells. J. Am. Soc. Mass Spectrom. 14, 342–351. https://doi.org/10.1016/S1044-0305(03)00065-5.