JCM Accepted Manuscript Posted Online 30 October 2019
J. Clin. Microbiol. doi:10.1128/JCM.01037-19

Hash-based core genome multi-locus sequencing typing for *Clostridium difficile*

Dr David W Eyre,[1,2] Prof Tim EA Peto[2,3,4], Prof Derrick W Crook[2,3,4], Prof A Sarah Walker[2,3,4]*, Prof Mark H Wilcox[5]*

[1]Big Data Institute, University of Oxford
[2]National Institute for Health Research Oxford Biomedical Research Centre, Oxford, UK
[3]Nuffield Department of Medicine, University of Oxford
[4]National Institutes of Health Research Health Protection Unit on Healthcare Associated Infections and Antimicrobial Resistance, University of Oxford
[5]Healthcare Associated Infections Research Group, University of Leeds, Leeds, UK

* Profs Walker and Wilcox contributed equally

Running title: Hash-cgMLST for *C. difficile* surveillance

Corresponding author: David Eyre, david.eyre@bdi.ox.ac.uk

Abstract

Background. Pathogen whole-genome sequencing has huge potential as a tool to better understand infection transmission. However, rapidly identifying closely-related genomes among a background of thousands of other genomes is challenging.

Methods. We describe a refinement to core-genome multi-locus sequence typing (cgMLST) where alleles at each gene are reproducibly converted to a unique hash, or short string of letters (hash-cgMLST). This avoids the resource-intensive need for a single centralised database of sequentially-numbered alleles. We test the reproducibility and discriminatory power of cgMLST/hash-cgMLST compared to mapping-based approaches in *Clostridium difficile* using repeated sequencing of the same isolates (replicates) and data from consecutive infection isolates from six English hospitals.

Results. Hash-cgMLST provided the same results as standard cgMLST with minimal performance penalty. Comparing 272 replicate sequence pairs, using reference-based mapping there were 0, 1 or 2 SNPs between 262(96%), 5(2%) and 1(<1%) respectively. Using hash-cgMLST, 218(80%) replicate pairs assembled with SPAdes had zero gene differences, 31(11%), 5(2%) and 18(7%) pairs had 1, 2 and >2 differences respectively. False gene differences were clustered in specific genes and associated with fragmented assemblies, but reduced using the SKESA assembler. Considering 412 pairs of infections within ≤2 SNPS, i.e. consistent with recent transmission, 376(91%) had ≤2 gene differences and 16(4%) ≥4. Comparing a genome to 100,000 others took <1 minute using hash-cgMLST.

Conclusion. Hash-cgMLST is an effective surveillance tool for rapidly identifying clusters of related genomes. However, cgMLST/hash-cgMLST generates more false variants than mapping-based approaches. Follow-up mapping-based analyses are likely required to precisely define close genetic relationships.

Introduction

The rapid development of pathogen whole-genome sequencing offers huge potential for better understanding the epidemiology of many infections. When trying to intervene to stop transmission, it is often important to identify the most closely genetically-related organisms already sequenced, as these represent potential recent sources of infection or cases that share a common infection source. However, the rapidly growing scale of data generated makes identifying these closely-related genomes among a background of many thousands of other genomes very challenging.

Three main approaches can be taken to identify closely-related genomes. Comparing single nucleotide polymorphisms (SNPs) identified following mapping to a reference genome offers high precision, e.g.[1] but, despite efforts to optimise computational approaches[2], is relatively slow. In contrast, k-mer based approaches based on hash algorithms, e.g. MASH[3] and PopPUNK[4], are fast, but the inherent and unstructured dimensionality reduction (e.g. summarising the whole genome as 500 hash strings selected on the basis of sorted hash strings) can reduce precision in fine-scale transmission analyses. Core genome multi-locus sequencing typing (cgMLST)[5] potentially provides a solution; genomes are summarised as a list of ~2000-3000 numbers, with each number representing the unique sequence of each core gene, i.e. structured dimensionality reduction. This summary enables more rapid comparisons as, taking the example of *Clostridium difficile*, only 2270 gene allele numbers need be compared,[6] rather than having to compare 4.3 million base pairs of sequence data for SNPs. A drawback of cgMLST as described to date is that it requires a centralised database of alleles of each gene to be maintained, so that cgMLST profiles generated by different laboratories are comparable. This centralised support can potentially be provided by academic, public health or commercial organisations, but any given scheme's sustainability is potentially limited by the funding available to support it. Additionally, for some pathogens, including *C. difficile*, several competing cgMLST/whole-genome-MLST schemes (e.g. Enterobase [University of Warwick, UK], cgmlst.org [Ridom GmbH, Germany] and BioNumerics [BioMérieux, France]) containing different genes and profiles have been developed; the latter two being associated with a commercial platform for processing sequencing data.

We therefore propose an alternative to cgMLST as described to date. Instead of maintaining a database of alleles, each allele is reproducibly converted to a unique hash, or short string of letters. This compresses each item of identical data to the same smaller representation, based on the sequence of an allele alone. Therefore, this process can be undertaken independently in different laboratories without the need to maintain or subscribe to a central database, but still generates summary data in a reproducible form that can be exchanged by laboratories. This distributed approach avoids the potentially costly need maintain a central database.

This study has two main aims. Firstly, to demonstrate an implementation of hash-based cgMLST, and to test whether hash-cgMLST profiles can be compared without a significant performance penalty compared to standard cgMLST; and secondly to test the reproducibility and discriminatory power of cgMLST compared to SNP-based typing. The discriminatory power of cgMLST has been previously explored, e.g.[6–9], however how cgMLST gene differences relate to SNP distances has not been comprehensively assessed. Instead it is postulated that small numbers of SNPs are likely to fall in different genes, and so SNP distances and gene differences are likely to be similar for closely related isolates. We evaluate the extent to which this assumption holds. Related to this, only limited assessments of the reproducibility of cgMLST have been undertaken. The largest study to date involved the same *Staphylococcus aureus* DNA from 20 isolates undergoing sequencing in 5 laboratories.[10] In this setting, in 80 comparisons (i.e. 20 sequences from 4 laboratories compared with the baseline laboratory) only 3 false gene differences were identified. We investigate whether these results can be replicated in *C. difficile*.

Methods

Hash-cgMLST

Using the cgMLST scheme of Bletz *et al*,[6] the first allele for each of the 2270 genes was used to create a BLAST search query. Following previous descriptions,[6,10] BLAST searches for each gene required a 90% identity match, a matched length ≥99% of the query length and the matched gene to be free from ambiguous characters or premature truncation. To avoid apparent truncated genes arising from misassembly we checked the number of stop codons in the gene sequence, and only retained matches with a single stop codon. To avoid truncation arising from contig breaks we ensured that BLAST matches included the start and end of the query sequence. Other BLAST search parameters were: “evalue=0.01, word_size=11, penalty=-1, reward=1, gapopen=5, gapextend=2”. The resulting genes were either matched to the database available at cgmlst.org, i.e. standard cgMLST, or hashed using an md5 algorithm to create a 32-character hexadecimal string. Deletions relative to the search query, represented by dashes in the matched gene sequence were removed prior to generating the hash. This avoids false differences introduced by locally variable placement of these deletions introduced by BLAST. The resulting cgMLST and hash-cgMLST profiles were saved as json files, i.e. a format that could readily be exchanged between laboratories. Where no BLAST match was found for a gene in the scheme an empty value was recorded, and that gene excluded in pairwise comparisons.

The choice of md5 hash provides 16^{32}, i.e. 3.4×10^{38} possible hashes. There is a theoretical chance of hash collisions, i.e. different sequences resulting in the same hash, but as the number of viable sequences for each gene in cgMLST databases is typically only tens to hundreds this is very unlikely. Importantly if a hash collision occurred this would result in genomes appearing falsely more similar, rather than falsely excluding potential transmission.

Sequence data

During whole-genome sequencing of *C. difficile* undertaken in Oxford and Leeds, UK we have routinely re-sequenced a subset of isolates as part of our internal quality assurance. We searched our database for isolates sequenced more than once. For a subset of these replicate sequences, the same extracted DNA was used to generate both sequences; for the remainder it was not documented in our laboratory information management system whether the same DNA extract was re-sequenced, or whether a fresh DNA extract was made from the same frozen isolate (Table S1). Paired-end sequence data for both types of replicate were generated using Illumina technology, including on various iterations of the

HiSeq platform and the MiSeq platform, with read lengths varying from 100-150bp in the majority of sequences (two 50bp sequences were also included).

To compare the discriminatory power of hash-cgMLST compared to SNP-based typing we processed 973 genomes from a previously published study of consecutive *C. difficile* over one year in six English hospitals using our hash-cgMLST and SNP pipelines.[11]

Bioinformatic processing

For hash-cgMLST typing, raw sequence data underwent adapter trimming and quality trimming using bbduk.sh from the bbMap package (version 38.32).[12] Stringent quality trimming was applied following Mellmann *et al*,[10] both the left and right ends of each read were trimmed to a Q30 threshold (using bbduk parameters: "ktrim=r k=23 mink=11 hdist=1 tpe tbo qtrim=rl trimq=30"). Following this the number of bases remaining in the trimmed reads was divided by the length of the 630 reference genome[13] (4290252 bp) to provide the mean high quality coverage, this was required to be ≥50 for a sequence to be included in the study. Appropriate quality trimming and adapter removal was confirmed using FastQC.[14] To check for contamination with non-*C. difficile* DNA, the species origin of sequence reads was classified using Kraken2[15] using the MiniKraken2_v1 database (built from the refseq bacteria, archaea, and viral libraries).

Following Bletz *et al*,[6] reads were *de novo* assembled using SPAdes (version 3.11.1)[16], with the "--careful" flag to reduce misassembly by using bwa-based mapping to confirm variants. Assembly quality metrics were obtained using the stats.sh script from bbmap.[12] Samples with assembly sizes (base pairs in contigs) >10% above or below the median size were rejected. We also tested performance using SPAdes with an addition flag "--only-assembler" to disable SPAdes internal read correction procedure. As an additional comparison reads were also *de novo* assembled using SKESA (version 2.3)[17] with default settings.

Reads (without stringent quality trimming) were also mapped to the 630 reference genome as described previously,[1,11,18] using stampy[19] for mapping and mpileup[20] for variant calling, followed by quality filtering of variants. Variant calls were required to have a quality score of ≥30, be homozygous under a diploid model, be supported by ≥5 high quality reads including ≥1 read in each direction and a consensus of ≥90% of bases and not be within a repetitive region of the genome. See https://github.com/oxfordmmm/CompassCompact for example implementation. For inclusion, ≥70% of the reference genome needed to be called in the consensus sequence. Bases in the consensus sequence not passing quality filtering were denoted N rather than A, C, G or T.

The bioinformatic pipelines used in this study for assembly and hash-cgMLST were written as NextFlow workflows[21] and can be found at https://github.com/davideyre/hash-cgmlst. Information on required dependencies and system requirements are provided in the repository readme file.

Analysis

Sequences meeting all quality thresholds (high-quality average coverage, assembly size, proportion of reference genome called) were compared. For replicate sequences, when an isolate had been sequenced more than twice, a random sequence was chosen as the baseline sequence with which all other sequences from the same isolate were compared, in order to avoid multiple counting.

Pairwise observed SNP differences between replicates and recombination-corrected SNP differences between other *C. difficile* genomes were obtained using Python scripts, PhyML[22] and ClonalFrameML[23] as previously described[11] (https://github.com/davideyre/runListCompare). Whole-genome alignments were used as input for PhyML. Invariant sites, i.e. those called as the same base as the reference or an unknown base, N, across all genomes were set to be the same base as the reference for computational efficiency, given there was no evidence of variation at these sites. All other sites had evidence of variation in at least one genome and were included unchanged including any genomes with an N at that site. The maximum likelihood approach taken accounts the uncertainty in the phylogeny arising from some genomes having an N called at some variable sites.

The number of cgMLST loci differences and number loci compared were obtained using Python (https://github.com/davideyre/hash-cgmlst). Where no BLAST match was found for a gene in either (or both) of the genomes in a pairwise comparison this was not counted towards the total number of cgMLST gene differences.

Data availability

Short read archive accession numbers for analysed replicate genomes are provided in Supplementary Table S1 with explanatory notes in the accompanying legend. Data for the 973 genomes from six English hospitals can be found at NCBI BioProject PRJNA369188.

Results

Hash-cgMLST provided the same results as standard cgMLST with minimal performance penalty. Results are presented throughout using pairwise core-gene differences generated with hash-cgMLST as these were identical to standard cgMLST gene differences if novel alleles were accounted for.

Comparison of hash-cgMLST and SNP typing performance in replicate sequences

A total of 374 sequences from 104 isolates passed all quality checks and were available for comparison to investigate the reproducibility of sequencing followed by cgMLST for *C. difficile* transmission analyses. A median (interquartile range) [range] of 2 (2-3) [2-27] sequences were available per isolate. Comparing replicate sequences with a randomly selected baseline sequence for each isolate yielded 272 comparisons for analysis.

With perfect sequencing no variants would be expected between pairs of sequences from the same isolate (replicate pairs). Using reference-based mapping and variant calling there were 0 SNPs between 262 (96%) replicate pairs, 1 SNP between 5 (2%) pairs and 2 SNPs between 1 (<1%) pair, i.e. a mean 0.026 SNPs per pair which equates to 1 false SNP call per 39 sequences (Figure 1A). Based on the rate of *C. difficile* evolution and the extent of within host genetic diversity ≤2 SNPs are expected between >95% of cases related by recent transmission;[1] therefore it is unlikely that transmission would be falsely excluded on the basis of the error rates seen.

Using either hash-cgMLST or standard cgMLST following assembly using SPAdes, 218 (80%) replicates pairs had zero gene differences, 31 (11%) pairs 1 difference, 5 (2%) pairs 2 differences, and 18 (7%) pairs had >2 differences, with a mean of 0.64 false gene differences per genome (Figure 1B) (test for symmetry considering 0, 1, 2, >2 SNPs or gene differences, p=0.004). Applying a threshold of >2 gene differences to rule out transmission (by analogy with SNP-based metrics[1,6]), the observed error rate would result in 6.6% (95% binomial confidence interval, CI, 4.0-10.3%) of transmission pairs being falsely excluded. Restricting to the subset of sequences where sequencing was known to have been undertaken from the same pool of extracted DNA produced fewer gene differences (Figure 1). Of 190 pairs, 189 (>99%) had 0 SNPs and 1 (<1%) pair had 1 SNP. From cgMLST, 167 (88%) pairs had 0 gene differences, 19 (10%) had 1 difference, 4 (2%) had 2 differences, and none had >2 differences.

Predictors of false cgMLST gene differences

The observation of greater differences between replicates restricting to variation in the 2270 core genes versus considering SNPs across the whole genome is potentially counter-intuitive. However, it should be remembered that the whole-genome SNP approach depends on a different bioinformatic approach with sophisticated per variant quality filtering, whereas the cgMLST is based on *de novo* assembly with more limited quality filtering. We therefore investigated potential predictors of false cgMLST gene differences using the hash-cgMLST algorithm (which were identical to the standard cgMLST approach) to see if filtering could be improved. Although we had already restricted our analysis to only include sequences with a mean genome coverage of >50, we investigated whether a more stringent threshold would improve performance (Figure 2). There was no evidence that increased coverage was associated with fewer cgMLST gene differences (Spearman's rho -0.04, p=0.43). There were only 2 sequences in the dataset with 50bp reads, the remainder had 100 or 150bp reads. 14/222 (6%) sequence pairs where the minimum sequence length was 100bp contained >2 gene differences, compared to 4/48 (8%) in pairs with both 150bp reads (exact p=0.54).

The relationship between cgMLST gene differences and *de novo* assembly quality metrics is shown in Figure 3A-C. Given the filtering applied, there was still an association between the number of false gene differences and the maximum absolute percentage deviation from the overall median assembly size (4165590bp) within each replicate pair (which was constrained to be ≤10% for inclusion in the analysis) (Spearman's rho 0.21, p<0.001, Figure 3A, with both small and large assemblies contributing to this effect). L50 describes the minimum number of contigs required to achieve 50% of the assembly size, with higher values representing more fragmented lower quality assemblies. Higher values of L50 were associated with greater rates of false gene differences (Spearman's rho 0.37, p<0.001). 9 (2%) of 257 pairs with both L50 values ≤125 had >2 false gene differences compared to 9/15 (60%) with one or more sequences with an L50 >125 (Figure 3B). Another measure of assembly fragmentation is the total number of contigs; higher numbers of contigs were also associated with greater false gene differences (Spearman's rho 0.31, p<0.001, Figure 3C).

Figure 3D shows the impact of the proportion of reads classified as *C. difficile* by Kraken2 on cgMLST gene differences. Within the dataset there was no evidence of significant contamination with a bacterial species other than *C. difficile* and the most common species was *C. difficile* in all samples. However, the proportion of reads that could not be classified at all varied from 0-11% between sequences with the exception of one replicate pair (36% and 24%). Higher rates of unclassified sequences were associated with higher false gene differences, but without any clear separation of the data on this basis (Spearman's rho -0.23, p<0.001).

Distribution of cgMLST gene differences in replicate sequences

The gene differences observed between replicate sequences disproportionately affected a small number of genes (Supplementary Table S2). Only 82 (4%) of 2270 genes contained differences within the replicate sequences. To avoid multiple counting, we evaluated the number of isolates that contained at least a pair of replicates with gene differences: 16 genes contained differences in two or more isolates' replicates, and of these 15 were due to the same nucleotide differing in all replicate pairs. The reproducible location of the differences observed for a given gene across different isolates is compatible with consistent mis-assembly (Table S2). If the 15 genes with identical gene differences affecting ≥2 isolates were excluded, the number of the 272 replicate pairs with 0 gene differences increased from 218 (80%) to 236 (87%) and the number of pairs with >2 gene differences reduced from 18 (7%) to 14 (5%). (Figure S1B). Using the full 2270 gene set and disabling SPAdes internal read correction resulted in fewer false gene differences: 0 differences in 236 (87%) pairs and >2 differences in 14 (5%) (Figure S1C).

Alternative assembler, SKESA

Use of SKESA in place of SPAdes as the assembler used for hash-cgMLST resulted in the fewer differences between replicate pairs (Figure 1C), 241 (89%) pairs had 0 differences, 22 (8%) pairs 1 difference, 6 (2%) pairs 2 differences and 3 (1%) pairs 3 differences. This equates to 0.16 false gene differences per replicate pair sequenced. The median (IQR) number of genes compared between replicate pairs was 2225 (2187 – 2235) using SKESA and 2227 (2205 – 2242) using SPAdes out of a possible maximum 2270 genes.

Benchmarking

Samples were processed in parallel, with each sample using a single core from an Intel Xeon Gold 6150 2.70GHz 18-core CPU. For a single sample, the median (IQR) time to undertake quality control and read filtering was 3.6 (2.7-4.9) minutes and 27.4 (19.6-35.4) minutes to generate an assembly using Spades with read error correction and 16.3 (12.1-21.5) minutes without; SKESA took 19.4 (15.5-24.3) minutes. From the assemblies creating a hash-cgMLST profile took 44.1 (43.5-44.9) seconds. Having made hash-cgMLST profile files, running on a single CPU core, to compare a single genome to 100,000 others took 40.4 seconds. In contrast 100,000 comparisons using a standard cgMLST approach took marginally less time, 38.7 seconds, after loading the profiles into memory.

cgMLST profiles can also be rapidly compared using a laptop or desktop, e.g. using one core of Intel i7 2.6Ghz laptop processor, comparing the 973 samples from the six hospitals study required 467Mb of memory, and took 236 seconds for 472,879 comparisons, i.e. 49.9 seconds per 100,000 comparisons. Using the same laptop, creating hash-cgMLST profiles from existing assemblies typically took ~40 seconds and required <100Mb of memory.

Comparison of hash-cgMLST and SNP typing in data from six English hospitals

We analysed 973 genomes from a previous study of *C. difficile* transmission in six English hospitals[11] Of these, 56 failed the assembly size threshold and 20 the coverage threshold (one also failing the assembly threshold), leaving 898 (92%) genomes for analysis. We considered all pairs of genomes within ≤2 SNPs and tested the extent to which the numbers of hash-cgMLST gene differences, using SPAdes (with the --only-assembler flag) or SKESA assemblies, followed the number of SNPs (Figure 4A and 4C). Of 412 pairs of sequences within ≤2 SNPs, using SPAdes 376 (91%) were within ≤2 gene differences, 30 (7%) had 3 differences, 16 (4%) had ≥4 differences and using SKESA 406 (99%) had ≤2 gene differences, and the remainder all ≤5 differences. The median (IQR) number of genes called in each pair was 2143 (2084-2191) using SPAdes and 2003 (1891-2110) using SKESA.

To achieve ≥99% sensitivity for identifying genomes within ≤2 SNPs required a threshold of ≤9 gene differences using SPAdes and ≤3 gene differences using SKESA, with an associated positive predictive value (PPV) of 11% (410/3720) and 38% (410/1092) respectively. Specificity was >99% with both assemblers (399031/402341 and 401659/402341 respectively).

We also considered the distribution of SNPs within pairs of genomes with ≤2 gene differences using hash-cgMLST. Following assembly with SPAdes, of 590 pairs of genomes, 376 (64%) were within ≤2 SNPs, with the maximum number of SNPs observed 20 (Figure 4B). Using SKESA of 749 genome pairs, 406 (54%) were within ≤2 SNPs (Figure 4D).

Discussion

Here we present the concept of hash-cgMLST as a tool for rapid comparison of bacterial sequencing data. This is a significant development over standard cgMLST approaches as it removes the need for a central database of alleles. Such databases require resource-intensive curation to ensure they are maintained to a high standard. Additionally, allele numbering is currently done consecutively in a single location, which is problematic with large datasets that span many laboratories; hashes also overcome this limitation. We also provide the code to run the algorithms developed.

This manuscript also highlights important limitations of common implementations of cgMLST as a tool for high resolution outbreak detection. Stringent filtering done on the basis of mapped data allows the number of false variant calls to be controlled; here we obtained around 1 false SNP for every 39 genomes sequenced. In contrast, fine-grained per base quality control is typically not implemented in studies using *de novo* assembly tools. Using SPAdes we observed an mean of 0.64 false gene differences per replicate genome pair. The alternative assembler tested, SKESA, was able to better control false gene differences, with 0.16 per replicate pair, i.e. 1 error per every 6.3 genomes sequenced. The higher rates of false variation observed using cgMLST/hash-cgMLST led to the counter-intuitive observation in some samples of more differences comparing 2270 genes than comparing the whole genome. It should be noted that undertaking SNP-based analyses from alignments of *de novo* assemblies without further filtering of variants would be similarly affected. These errors can be reduced by ensuring the assemblies studied are of high quality. Our data suggest that the previously described read quality trimming and filtering based on assembly sizes[6,10] could be further improved by also only analysing samples with an L50 value of below ~125. However, this stringent filtering would have resulted in 30% of the previously published dataset studied being unavailable for analysis, questioning its practicability.

Although our approach does not depend on a database of alleles it is dependent of the development of a high quality cgMLST scheme, i.e. appropriate identification of core genes based on a large and diverse collection of genomes, and careful selection of problematic genes for exclusion. Despite such an approach being taken in developing the *C. difficile* cgMLST scheme used, we show that removing a small number of genes from this cgMLST scheme would likely improve performance if using SPAdes assemblies, as a small subset of genes contained higher numbers of false gene differences (Table S2, Figure S1). This highlights the importance of assessing the performance of each cgMLST scheme created on a per species and scheme basis using appropriate test datasets which include replicate and closely-related sequences.

Many of the apparent errors seen in replicate pairs appear to arise from mis-assembly. SPAdes based read correction did not improve accuracy and instead resulted in more rather than fewer differences between replicate pairs. Use of an alternative assembler SKESA[17] reduced the number of replicate pairs with >2 differences to just 1%, within minimal reduction in the number of genes compared between replicate pairs (median 2225 compared to 2227 with SPAdes). The reduction in genes compared was greater in the clinical dataset analysed (median 2143 and 2003), but this reduced discriminatory power for transmission studies will usually be more than offset by reduced error rates (and therefore reductions in erroneous exclusion of transmission).

Our data also highlight that extrapolating the ≤2 SNP threshold for identifying genetically plausible transmission events to two (or three[6]) gene differences may be inappropriate depending on the choice of assembler and settings. Using SPAdes, 4% of pairs of samples within ≤2 SNPs were >3 genes different by cgMLST, whereas with SKESA this was only 1%. For public health applications optimised to identify potential transmission, to be ≥99% sure of not missing pairs of sequences within ≤2 SNPs, a threshold of ≤9 gene differences was needed for SPAdes assemblies and ≤3 differences with SKESA. However these thresholds for SPAdes resulted in around 8 genome pairs >2 recombination-corrected SNPs apart being identified for every 1 pair within ≤2 SNPs (PPV 11%), and 1.6 pairs >2 SNPs apart for every pair within ≤2 SNPs using SKESA (PPV 38%). In this scenario further SNP-based analysis based on mapping and filtered variant calling is likely to be required to determine which genomes are potentially related by recent transmission and which are not. In other cases, higher numbers of SNPs were observed than gene differences (Figure 4B and 4D), which may arise from SNPs outside core genes, SNPs in uncalled genes, and imperfect correction of recombination events.

Hash-cgMLST allowed rapid comparison of many thousands of bacterial genomes within seconds, using a relatively unoptimized python script running on a single laptop or server CPU core. As comparisons with other genomes can be easily divided into independent parts, this task is readily parallelisable. Using hash-cgMLST, it is therefore potentially possible to compare each new sequence generated with millions of previous sequences. The summaries of each genome produced, a roughly 130kb json file, are readily exchangeable between laboratories and could potentially be hosted alongside raw reads in sequence read archives. As such, each laboratory could maintain its own database of hash-cgMLST profiles and distances, as well as this potentially being usefully provided as part of future web-based services based on publicly available data. Although without further refinements hash-cgMLST may not allow high-precision fine-scaled transmission studies, it has the potential to dramatically reduce the search space for closely-related genomes, which can then be followed by more precise SNP-based analyses on a much smaller subset of genomes.

Using SPAdes we observed a higher rate of ‘false’ gene differences between genomes where the sequences were potentially generated from separate DNA extractions of the same isolates, compared with genomes obtained from the same DNA extraction. It is therefore plausible that the differences observed represent true differences, but a form of variation that is much faster and more erratic than mutation/recombination rates based on filtered SNPs. The erratic nature of the variation observed is unlikely to be informative about recent transmission. We also did not see these differences to the same extent using an alternative assembler, SKESA.

This study is potentially limited by not being an exhaustive investigation of all the potential options for assembly and for filtering *de novo* assembly data, in particular further filtering of variants based on mapping reads back to assemblies may improve precision, e.g. as done by Enterobase.[24] Although we used Kraken2 to search for contamination with DNA from other species, contamination with *C. difficile* DNA from other samples processed concurrently may be an important contributor to some of the differences seen with hash-cgMLST, whereas resulting mixed calls can be filtered using mapped data.

In conclusion, appropriately quality controlled cgMLST can identify clusters of related genomes rapidly and is an appropriate tool for surveillance and reducing the search space in outbreaks. The SKESA assembler, compared to SPAdes, was associated with lower rates of gene differences between replicate sequences, and when used for hash-cgMLST more closely matched the number of SNPs between closely related samples. The approach we describe has potential to be deployed across a range of pathogens, including those where linkage across time and wide geographic space, i.e. involving very large sequencing datasets, may help resolve sources and routes of transmission, such as for food borne infections. Refined variant calling based on mapping is likely required to precisely define close genetic relationships. This study highlights the need for detailed quality assurance to determine the performance of algorithms used for comparing genomes. Our hash-cgMLST implementation is freely available and provides an effective database-free approach to cgMLST.

Funding Sources

This work was supported by the National Institute for Health Research Oxford Biomedical Research Centre and the Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at the University of Oxford in partnership with Public Health England (PHE) [HPRU-2012-10041]. DWC, TEAP and ASW are National Institute for Health Research Senior Investigators. DWE is a Big Data Institute Robertson Fellow. Computation was supported by an Academy of Medical Science starter grant for academic clinical lecturers awarded to DWE. Computation also used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Declaration of Interests

MHW has received consulting fees from Actelion, Astellas, MedImmune, Merck, Pfizer, Sanofi-Pasteur, Seres, Summit, and Synthetic Biologics; lecture fees from Alere, Astellas, Merck & Pfizer; and grant support from Actelion, Astellas, bioMerieux, Da Volterra, Merck and Summit. No other author has a conflict of interest to declare.

References

1. Eyre DW, Cule ML, Wilson DJ, Griffiths D, Vaughan A, O'Connor L, Ip C, Golubchik T, Batty EM, Finney JM, Wyllie DH, Didelot X, Piazza P, Bowden R, Dingle KE, Harding RM, Crook DW, Wilcox MH, Peto T, Walker SA. 2013. Diverse Sources of *C. difficile* Infection Identified on Whole-Genome Sequencing. New Engl J Medicine 369:1195–1205.

2. Mazariegos-Canellas O, Do T, Peto T, Eyre DW, Underwood A, Crook D, Wyllie DH. 2017. BugMat and FindNeighbour: command line and server applications for investigating bacterial relatedness. BMC Bioinformatics 18:477.

3. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 17:132.

4. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, Corander J, Bentley SD, Croucher NJ. 2019. Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome Res 29:304–316.

5. Maiden MC, van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND. 2013. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol 11:nrmicro3093.

6. Bletz S, Janezic S, Harmsen D, Rupnik M, Mellmann A. 2018. Defining and Evaluating a Core Genome Multilocus Sequence Typing Scheme for Genome-Wide Typing of *Clostridium difficile*. J Clin Microbiol 56:e01987-17.

7. Ruppitsch W, Pietzka A, Prior K, Bletz S, Fernandez H, Allerberger F, Harmsen D, Mellmann A. 2015. Defining and Evaluating a Core Genome Multilocus Sequence Typing Scheme for Whole-Genome Sequence-Based Typing of *Listeria monocytogenes*. J Clin Microbiol 53:2869–2876.

8. de Been M, Pinholt M, Top J, Bletz S, Mellmann A, van Schaik W, Brouwer E, Rogers M, Kraat Y, Bonten M, Corander J, Westh H, Harmsen D, Willems RJ. 2015. Core Genome Multilocus Sequence Typing Scheme for High-Resolution Typing of *Enterococcus faecium*. J Clin Microbiol 53:3788–3797.

9. Cody AJ, Bray JE, Jolley KA, McCarthy ND, Maiden MC. 2017. Core Genome Multilocus Sequence Typing Scheme for Stable, Comparative Analyses of *Campylobacter jejuni* and *C. coli* Human Disease Isolates. J Clin Microbiol 55:2086–2097.

10. Mellmann A, Andersen P, Bletz S, Friedrich AW, Kohl TA, Lilje B, Niemann S, Prior K, Rossen JW, Harmsen D. 2017. High Interlaboratory Reproducibility and Accuracy of Next-Generation-Sequencing-Based Bacterial Genotyping in a Ring Trial. J Clin Microbiol 55:908–913.

11. Eyre DW, Fawley WN, Rajgopal A, Settle C, Mortimer K, Goldenberg SD, Dawson S, Crook

DW, Peto TE, Walker SA, Wilcox MH. 2017. Comparison of Control of *Clostridium difficile* Infection in Six English Hospitals Using Whole-Genome Sequencing. Clin Infect Dis 65:433 441.

12. Bushnell B. BBMap. http://sourceforge.net/projects/bbmap/

13. Sebaihia M, Wren BW, Mullany P, Fairweather NF, Minton N, Stabler R, Thomson NR, Roberts AP, Cerdeño-Tárraga AM, Wang H, Holden MT, Wright A, Churcher C, Quail MA, Baker S, Bason N, Brooks K, Chillingworth T, Cronin A, Davis P, Dowd L, Fraser A, Feltwell T, Hance Z, Holroyd S, Jagels K, Moule S, Mungall K, Price C, Rabbinowitsch E, Sharp S, Simmonds M, Stevens K, Unwin L, Whithead S, Dupuy B, Dougan G, Barrell B, Parkhill J. 2006. The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. Nat Genet 38:779 786.

14. Babraham Bioinformatics. FastQC. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

15. Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15:1.

16. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biology 19:455–77.

17. Souvorov A, Agarwala R, Lipman DJ. 2018. SKESA: strategic k-mer extension for scrupulous assemblies. Genome Biol 19:153.

18. Eyre DW, Davies KA, Davis G, Fawley WN, Dingle KE, Maio N, Karas A, Crook DW, Peto TE, Walker SA, Wilcox MH, study group E. 2018. Two Distinct Patterns of *Clostridium difficile* Diversity Across Europe Indicates Contrasting Routes of Spread. Clin Infect Dis 365:1693.

19. Lunter G, Goodson M. 2011. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res 21:936–939.

20. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

21. Tommaso P, Chatzou M, Floden EW, Barja P, Palumbo E, Notredame C. 2017. Nextflow enables reproducible computational workflows. Nat Biotechnol 35:316–319.

22. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biol 59:307–321.

23. Didelot X, Wilson DJ. 2015. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. Plos Comput Biol 11:e1004041.

24. Alikhan N-F, Zhou Z, Sergeant MJ, Achtman M. 2018. A genomic overview of the population structure of *Salmonella*. Plos Genet 14:e1007261.

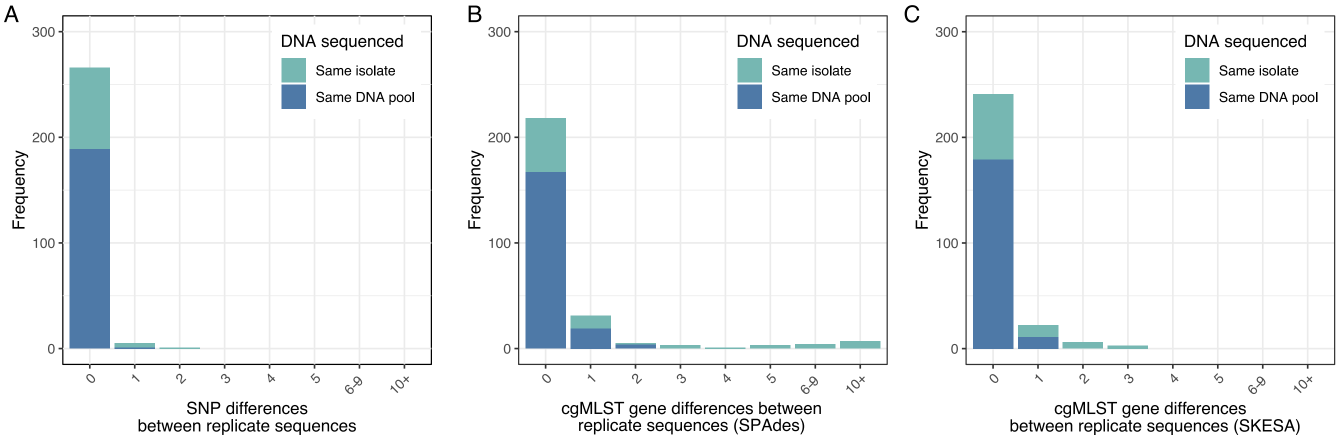
Figure Legends

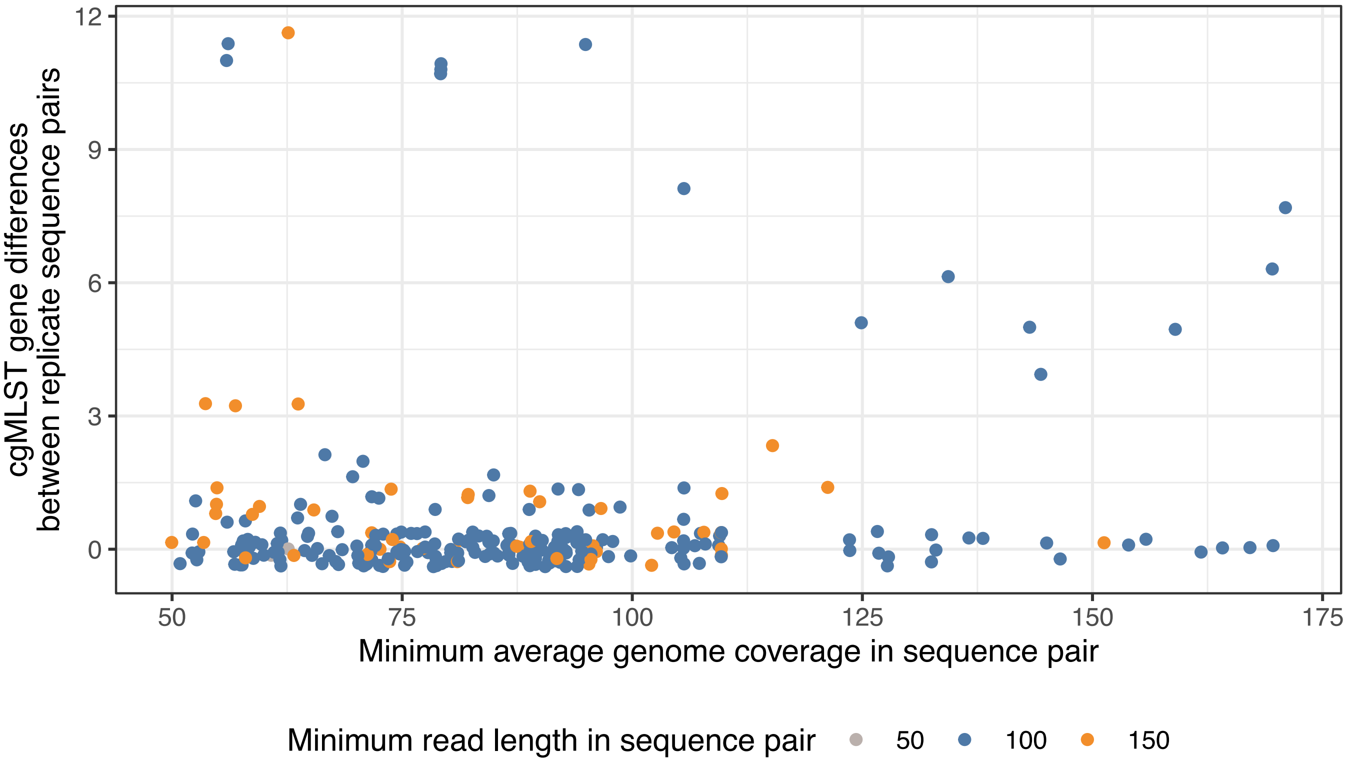
Figure 1. Observed differences using SNP typing (panel A) and hash-cgMLST based on SPAdes (panel B) and SKESA (panel C) assemblies in 272 replicate sequence pairs. With perfect sequencing no variants would be expected between pairs of sequences from the same isolate. Pairs of sequences known to have been obtained from the same pool of DNA are shown in dark blue. Where information was unavailable on whether the same pool of DNA was used or a fresh DNA extract was made from the same isolate, this is shown in light blue.

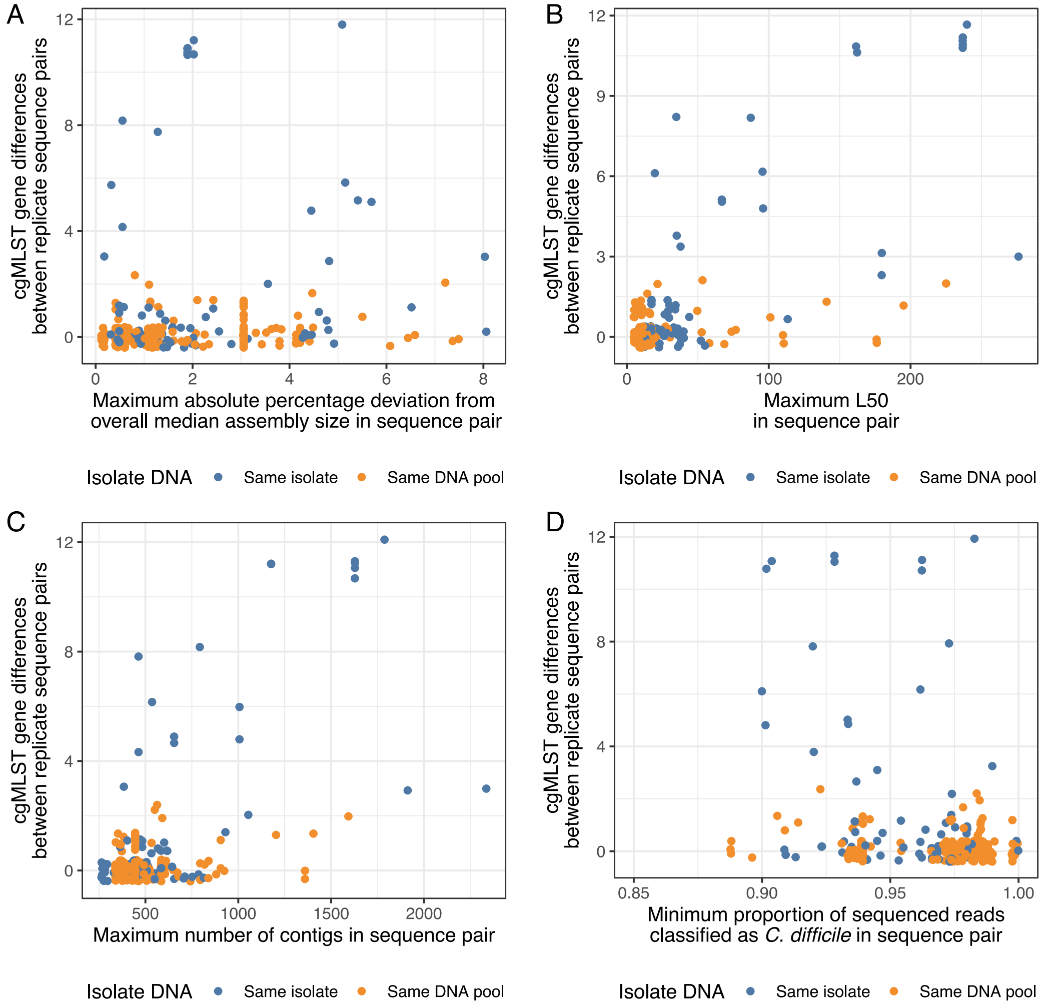
Figure 2. Relationship between hash-cgMLST gene differences in replicate sequence pairs and average genome coverage and read length. Jitter applied to points to assist visualisation. SPAdes with “--careful" flag used to generate assemblies.

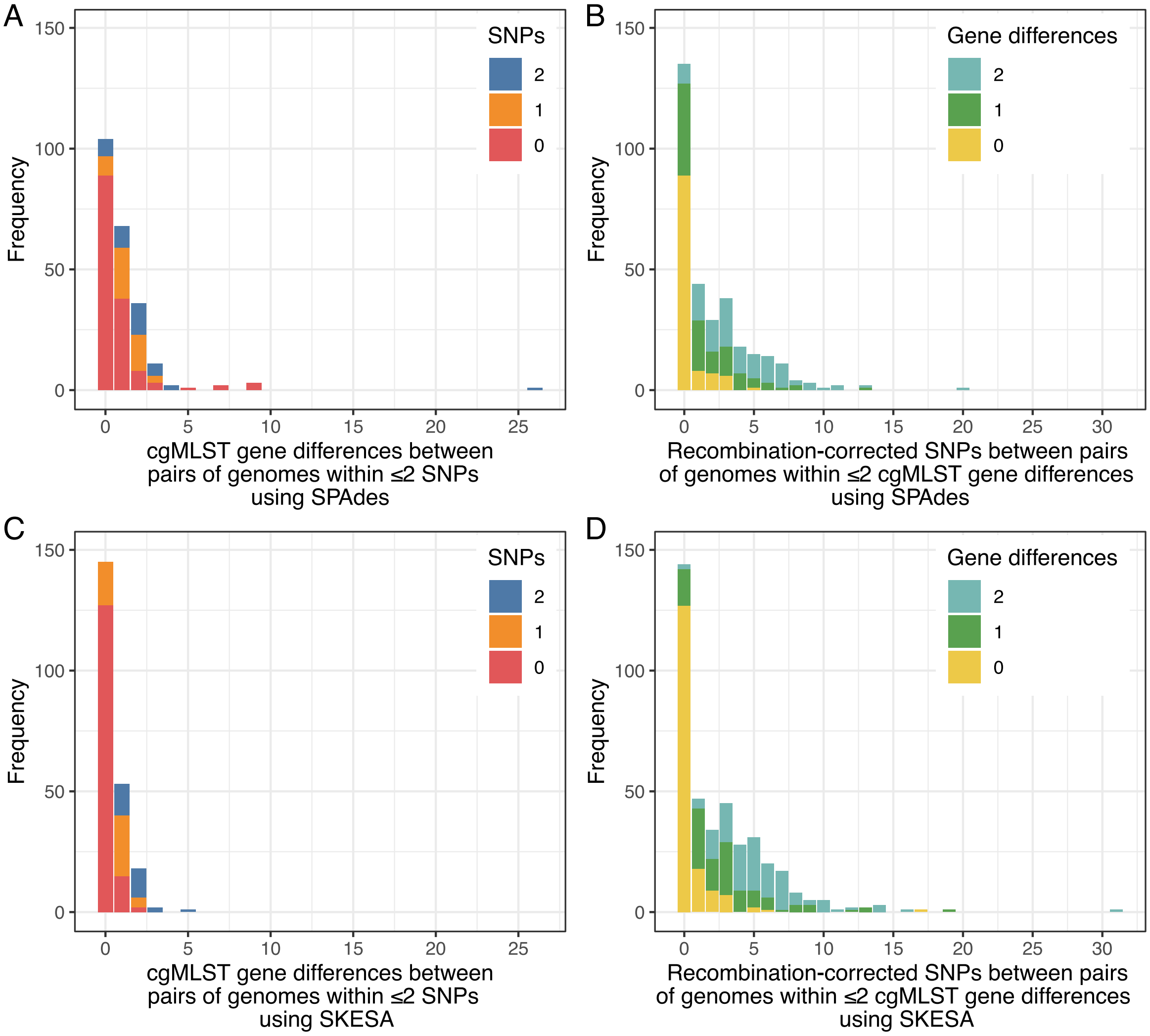
Figure 3. Relationship between hash-cgMLST gene differences in replicate sequence pairs and *de novo* assembly quality metrics (panels A-C) and Kraken2 read classification (panel D). Jitter applied to points to assist visualisation. One point is omitted from Figure 3D for ease of visualisation with the proportion of reads classified as *C. difficile* of 0.64 and 0 gene differences. SPAdes with “--careful" flag used to generate assemblies.

Figure 4. Relationship between hash-cgMLST gene differences and SNPS in *C. difficile* genomes from consecutive infections in six English hospitals. Panel A shows the distribution of hash-cgMLST gene differences between pairs of genomes within ≤2 SNPs. Panel B shows the distribution of SNPs within pairs of genomes within ≤2 gene differences. Panel A and B were generated using SPAdes assemblies with the “--careful --only-assembler” flags. Panel C and D show the same analysis using the SKESA assembler.


A
300
200
100
0
Frequency
DNA sequenced
Same isolate
Same DNA pool
0
1
2
3
4
5
6-9
10+
SNP differences
between replicate sequences
B
300
200
100
0
Frequency
DNA sequenced
Same isolate
Same DNA pool
0
1
2
3
4
5
6-9
10+
cgMLST gene differences between
replicate sequences (SPAdes)
C
300
200
100
0
Frequency
DNA sequenced
Same isolate
Same DNA pool
0
1
2
3
4
5
6-9
10+
cgMLST gene differences
between replicate sequences (SKESA)


cgMLST gene differences
between replicate sequence pairs
12
9
6
3
0
50
75
100
125
150
175
Minimum average genome coverage in sequence pair
Minimum read length in sequence pair
● 50
● 100
● 150


A
cgMLST gene differences
between replicate sequence pairs
0
4
8
12
0
2
4
6
8
Maximum absolute percentage deviation from
overall median assembly size in sequence pair
Isolate DNA
● Same isolate
● Same DNA pool
B
cgMLST gene differences
between replicate sequence pairs
0
3
6
9
12
0
100
200
Maximum L50
in sequence pair
Isolate DNA
● Same isolate
● Same DNA pool
C
cgMLST gene differences
between replicate sequence pairs
0
4
8
12
500
1000
1500
2000
Maximum number of contigs in sequence pair
Isolate DNA
● Same isolate
● Same DNA pool
D
cgMLST gene differences
between replicate sequence pairs
0
4
8
12
0.85
0.90
0.95
1.00
Minimum proportion of sequenced reads
classified as C. difficile in sequence pair
Isolate DNA
● Same isolate
● Same DNA pool


A
150
100
50
0
Frequency
SNPs
2
1
0
0
5
10
15
20
25
cgMLST gene differences between
pairs of genomes within ≤2 SNPs
using SPAdes
B
150
100
50
0
Frequency
Gene differences
2
1
0
0
5
10
15
20
25
30
Recombination-corrected SNPs between pairs
of genomes within ≤2 cgMLST gene differences
using SPAdes
C
150
100
50
0
Frequency
SNPs
2
1
0
0
5
10
15
20
25
cgMLST gene differences between
pairs of genomes within ≤2 SNPs
using SKESA
D
150
100
50
0
Frequency
Gene differences
2
1
0
0
5
10
15
20
25
30
Recombination-corrected SNPs between pairs
of genomes within ≤2 cgMLST gene differences
using SKESA