KmerAperture: Retaining *k*-mer synteny for alignment-free estimation of within-lineage core and accessory differences

Matthew P. Moore[1,2], Mirjam Laager[3], Xavier Didelot[1,2]

1. School of Life Sciences, University of Warwick, Coventry, UK
2. Department of Statistics, University of Warwick, Coventry, UK
3. Division of Transplant Immunology and Nephrology, University Hospital Basel, Basel, Switzerland

Abstract

By decomposing genome sequences into *k*-mers, it is possible to estimate genome similarity without alignment. Dimension-reduction techniques such as *k*-mer minimisers (MinHash), have been developed and are often accurate approximations of distances based on full *k*-mer sets. These and other alignment-free methods avoid the enormous temporal and computational expense of alignment or mapping. However, these *k*-mer set comparisons are not entirely accurate within-species and can be completely inaccurate within-lineage. This is due, in part, to their inability to distinguish core polymorphism from accessory differences. KmerAperture takes the relative complements of a pair of whole genome *k*-mer sets and matches back to the original enumerated *k*-mer lists to gain positional information. SNPs are expected to result in contiguous series of unique *k*-mers of length $L = k$. On the other hand, series of length $L > 2k$ are likely to be caused by accessory differences of length $L-k+1$; when the start and end of the sequence are contiguous with homologous sequence. KmerAperture was benchmarked against Jaccard similarity and 'split *k*-mer analysis' (SKA) using datasets including a diverse lineage (*Clostridioides difficile* RT017), a lower core diversity sub-lineage with a large accessory genome (*Escherichia coli* ST1193) and a very low core diversity simulated population with accessory content not associated with number of SNPs. We present a new algorithm that demonstrates that with the few available axioms of how core and accessory sequence diversity is represented in *k*-mers, we can accurately distinguish them and estimate both. By matching unique *k*-mers back to their original lists we regain their synteny and may make inferences about their likely cause.

Corresponding author: matthew.p.moore@warwick.ac.uk

Introduction

The increasing availability of whole bacterial genomes provides us with an opportunity to transform outbreak analysis[1,2]. This potential for genomic epidemiology is being investigated by public health organisations, such as the United Kingdom Health Security Agency (UKHSA), where routine surveillance involves whole genome sequencing, typing and detailed comparative and phylogenetic analysis[3]. However, once collected and sequenced, the comparison of thousands of bacterial genomes poses a number of important difficulties and limitations, particularly the computational time and expense required to compile alignments.

The most intensive and detailed of comparative genomic methods involve the comparison of exclusive pairs whereby 5,000 genomes will require $>12x10^6$ pairwise alignments. This is often shortcut to 5,000 alignments by aligning each genome to a single reference genome and then comparing the genomes at each reference position, using string or matrix counting methods. This is much more efficient than comparing all pairs of genomes, but it still involves a large number of alignments. Furthermore, it can introduce a bias depending on which reference genome is used, and also loses any information about accessory regions not found in the reference.

A number of efficient alignment-free approaches have been developed. The most efficient of these involve decomposition of genomes into *k*-mers (*k*-mers/kmers/*n*-grams), unique sequences of length *k*, subsampling into 'sketches' and performing set analysis[4–7]. Dimensionality reduction methods such as the MinHash algorithm, accurately approximate the full-set differences across genomes and are extremely fast[4]. However, the accuracy in estimating relatedness decreases the more closely related the genomes are.

Alignment-free distance estimates tend to correlate primarily with quantities of differentially present sequences rather than SNP distances, especially for closely related genomes up to the point of re-sequenced isolates[8]. These closely related genomes are those that are most relevant for discovery and further analysis for public health[9–11]. For instance, MinHash-based distance measures will not accurately discern clusters at a lower diversity level than an MLST-defined lineage. Ultimately, this is due to the inability to know whether mismatching *k*-mers result from polymorphism or presence/absence of the sequence. PopPUNK is one method that improves on MinHash for within-species differences by modelling the change in mismatches across different *k* sizes to differentiate diversity that is core to the population and diversity that is accessory[12]. Two approaches have

demonstrated[13] an ability to estimate some within-lineage SNPs: kSNP3.0[14] and Split K-mer Analysis (SKA)[15], with similar benchmark performance[13,15]. These approaches successfully identify what, in set analysis-based methods, would be represented by mismatched *k*-mers resulting from SNPs indistinguishable from those resulting from accessory sequence.

Here, we propose a new algorithm that takes advantage of efficient set analysis followed by differentiation of mismatched *k*-mers based on their original synteny. The *k*-mers present in the relative set complement of query genomes compared with a reference that form a contiguous position series equal to the length of *k* are taken as an estimate of SNP-only diversity between pairs of genomes. Additionally, we can leverage this data to estimate the size of pairwise accessory difference and identify accessory regions in the original sequence, that may then be removed prior to alignment. This primarily set-based approach provides a novel proof-of-concept for inexpensive, within-lineage comparisons.

Methods

Using synteny to distinguish core and accessory variation

The data for analysis consists of a single reference genome and several query genomes, all of which have been assembled *de novo* and may therefore be made of multiple contigs. The reference and query genomes are decomposed into *k*-mers, without ambiguous bases, of length *k*, with a step (*s*) of 1. All *k*-mers are canonicalised to avoid mismatches due to strandedness, using screed[16] reverse complement and selecting the alternate of the lower lexicographical order. The original list of *k*-mers is retained for all genomes, including the genomic location of each *k*-mer. Next, each query genome is considered in turn and compared to the reference genome as follows. The reference and query genomes unique elements are constructed into sets, A and B respectively. Next the relative complements of each set are extracted by subtracting from one another, resulting in A^c and B^c. Each element of A^c and B^c are matched against the original, enumerated list of *k*-mers. All matches are retained, including duplicates, each *k*-mer match position is stored and sorted. Contiguous series of *k*-mer positions are then examined for their overall length. Each series of length *k* is taken to result from a SNP. Sets were then constructed of the middle *k*-mer stored without the middle base for each genome. The intersection of these sets (number of unique matching middle *k*-mers) is taken to be the SNP count. As a result, *k* must be an odd number. The full *k*-mer set Jaccard similarity is computed as the intersection of *k*-mers over the union ($A \cap B \,/\, A \cup B$) for comparison. Accessory regions are determined by taking *k*-mer series in A^c or B^c of length greater or equal than $2k$ to be sequence regions

present in only one genome (pairwise accessory). This cut-off of $2k$ is chosen to be conservative. Relative presence for each genome is calculated as L-k+1 where L is the series length.

Implementation and benchmarking of KmerAperture

KmerAperture is written in python 3 and depends on the python packages numpy[17] and the screed module of the khmer package[16]. Visualisations were generated in python with the use of seaborn[18] and matplotlib[19] in the Jupyter Notebook environment[20]. Conda was used for package management[21]. KmerAperture was benchmarked against full *k*-mer set Jaccard similarity and SKA *map* SNP estimations. KmerAperture was run on assembled genomes across a range of values of *k* (*k*=19, 31 and 49), whereby *k* must be an odd number. SKA requires that the *k* size be generated by an integer divisible by 3, that represents the length of each sequence flanking a flexible base. As such, a range of flanking lengths was selected of 9, 15 and 24 to generate *k*=19, 31 and 49 respectively, so as to reflect *k* sizes close to the range used by KmerAperture.

Data acquisition and processing

Two medium- to low-diversity populations were selected for analysis. The first was a PCR ribotype-defined lineage: *C. difficile* RT017 (n=282)[22]. The RT017 dataset resulted in a final set of 261 genomes for analysis after exclusion of low-quality assemblies and reads that did not assemble (n=21). Exclusion by assembly quality was determined by RT017 genomes with >205 contigs (≥5x the median of 41). An RT017 within-lineage reference genome was used for mapping and variant calling, *C. difficile* M68 (M68, FN668375.1)[23].

The second was a subset of the pandemic *E. coli* (MLST-defined) lineage *E.* coli ST1193. The ST1193 population records were further subset to the HierCC HC20-level cluster 571 (n=1,072). Within cluster HierCC,[24] HC20 cluster: 571, reference genome *E. coli* MCJCHV-1 chromosomal sequence (CP030111.1), was selected for mapping and variant calling. EnteroBase[25] was accessed and searched for these criteria on 22/08/2022. Those with assembled genomes available were downloaded for analysis (n=506), genomes that were not ST1193 were removed (n=8) and those that failed quality checking (>2x the median number of contigs, 128) (n=62) and those too diverse to belong to a HC20-level cluster (n=18) resulting in 407 genomes for analysis.

Finally, a population of 250 genomes was simulated with between 1-50 SNPs, 1-5 indels ranging from 1-10bp and 1-15 accessory sequences of 100-1500bp. The reference genome *Klebsiella pneumoniae* HS11286 (NC_016845)[26] was downloaded and used as the template for each mutated

genome. Number of insertions, insertion length, number of genes and genomic position to mutate were randomly sampled while number of SNPs were sampled from a log10 distribution. As a result, even those with the lowest SNP diversity will be assigned between 100 and 22,500bp of accessory sequence and up to 50bp of smaller insertions.

All genomes were either downloaded from the NCBI short read archive (SRA) using the sra-toolkit[27] or pre-assembled from the NCBI genomes database (Supplementary Table 1). For those downloaded unassembled, sequencing adapters were trimmed from sequencing reads and they were quality filtered to a Phred Score of Q30 with TrimGalore![28]. Genome assemblies were performed with SPAdes[29]. QUAST[30] was used for all contig metrics. Multi locus sequence types (MLST) and species identity were checked with mlst[31]. SNP counts to the references were determined by snippy[32] and a sequence presence/absence matrix obtained from Panseq[33], from which to calculate real pairwise accessory difference.

Results

Illustrative example

Figure 1 provides a toy example of the KmerAperture algorithm compared with full set Jaccard similarity estimates. The reference sequence and query are separated by a single SNP and a further 5bp sequence present only in the query. The sequences are decomposed into *k*-mers where $k=3$. When there is an isolated unshared sequence of length L we can anticipate this to produce $L-k+1$ unique *k*-mers covering the unshared sequence. In the case, as in Figure 1, that the unshared sequence is also contiguous with shared sequence it will generate a further $k-1$ unique *k*-mers for each contiguous side. Alternatively, we anticipate a SNP to generate k unique *k*-mers in each set except when the SNP is within k of another SNP or sequence start or end (such as the end of a contig).

In this example, KmerAperture would correctly discern mismatched *k*-mers resulting from SNPs vs accessory sequence. Both set complements would result in contiguous *k*-mers of serial length k. As a result, the mean estimated SNPs would be calculated and it would be determined that in the comparable sequence regions there is a single SNP. KmerAperture would additionally discern, from the five contiguous *k*-mers unique to the query sequence, that there was an accessory difference of 5bp (since these *k*-mers are contiguous with shared sequence on one side, they are expected to create $(L-k+1)+(k-1)=L$ *k*-mers).

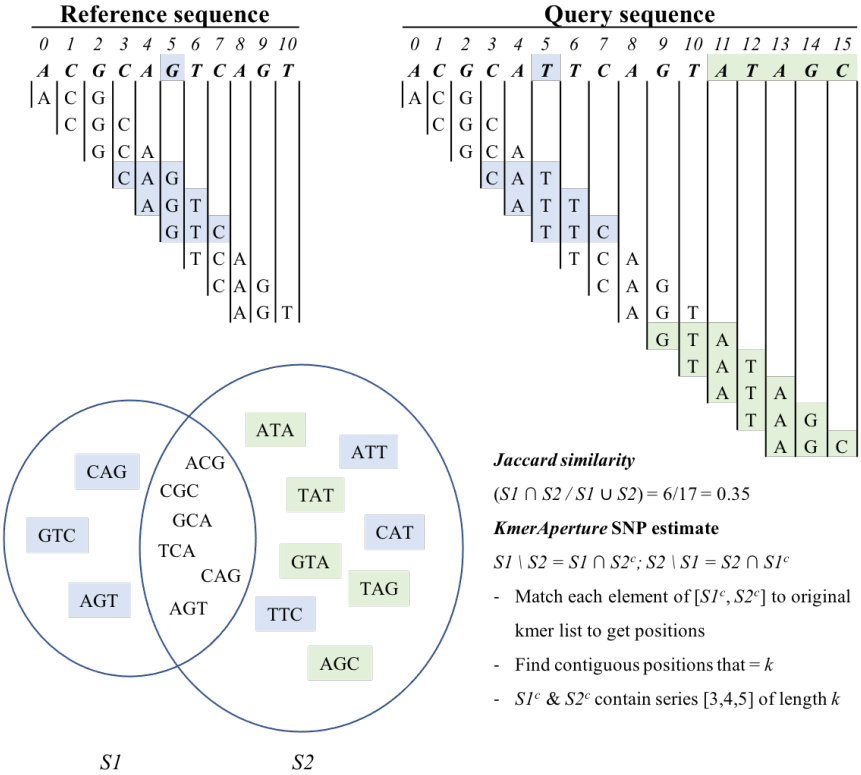

Figure 1. Blue shading highlights unique *k*-mers generated as the result of a SNP and green shading indicates unique *k*-mers as the result of an accessory sequence. In the Jaccard similarity these types of mismatches are not differentiated. *KmerAperture* starts by considering the genomic position of the elements unique to *S2*. SNPs are then estimated by counting contiguous *k*-mers of serial length *k.* The intended outcome of comparing the two genomes' *k*-mers would be to discover a single SNP between them (query position 5). This is a toy example where *KmerAperture* would estimate a single SNP whereas full set analysis would calculate distance including from sequence presence/absence (query positions 11-15)

Application to simulated genomes

With a range of real pairwise SNPs of 1-50 this population represents the lowest SNP diversity sample. KmerAperture estimated a median (range) of 6 (0-48) SNPs with *k*=19 and 7 (0-49) SNPs with *k*=31 and *k*=51. The correlations between KmerAperture estimates and SNPs ranged from 0.998 for *k*=19, 0.999 for *k*=31 and 1.0 for *k*=49. For comparison, the correlation between Jaccard distance and real SNPs ranged from 0.186 for *k*=19 to 0.346 for *k*=49. SKA produced a median (range) of 7 (0-49) and correlation of 1.0 across all sizes of *k*. Further, the ability of KmerAperture and SKA of identifying SNPs below a threshold was assessed. KmerAperture the produced high positive predictive values across SNP thresholds of ≤25,20,15,10 and 5. For instance, at a SNP threshold of ≤10 SNPs the (153 true positives) KmerAperture with *k*=31 and k=49 included only a single false positive, predicting 154 positives (PPV>0.99) (table 1).

Table 1. Performance of KmerAperture and SKA for identifying real SNP thresholds. True positives (*TP*) are those ≤SNP thresholds (25, 20, 15, 10, 5), false positives (*FP*) those under the KmerAperture or SKA threshold required for 100% sensitivity that are >SNP threshold and true negatives (*TN*), *n* samples - FP. Performance was judged on specificity: *TN*/(*TN*+*FP*) and positive predictive value (PPV): *TP*/(*TP*+*FP*).

Algorithm	***k*-mer size**	**Accuracy**	**SNP thresholds**				
			≤25 (n=208)	***≤20 (n=194)***	***≤15 (n=177)***	***≤10 (n=153)***	***≤5 (n=115)***
KmerAperture	*k*=19	Specificity	0.98	0.97	0.95	0.98	0.97
		PPV	0.99	0.99	0.98	0.99	0.98
	k=31	Specificity	0.95	0.97	0.96	0.99	0.99
		PPV	0.99	0.99	0.98	0.99	0.99
	k=49	Specificity	0.95	0.97	0.97	0.99	0.99
		PPV	0.99	0.99	0.99	0.99	0.99
SKA	*k*=19	Specificity	0.96	0.97	0.97	0.99	0.99
		PPV	0.99	0.99	0.99	0.99	0.99
	k=31	Specificity	0.95	0.97	0.97	0.99	0.99
		PPV	0.99	0.99	0.99	0.99	0.99
	k=49	Specificity	0.98	0.97	0.97	0.99	0.99
		PPV	0.99	0.99	0.99	0.99	0.99

For *k*=19, 31 and 49, KmerAperture estimated the accessory difference for each genome compared with the reference. Relative presence was estimated for each genome in each pair. In all cases, it was correctly estimated that there was no relative additional sequence in the reference genome (0bp estimated). Correlation coefficient was calculated between real and estimated accessory difference (bp). For all *k*-mer sizes the correlation was >0.999. The smallest absolute median (range) error accessory difference (bp) between estimated and real accessory size was 290.5 (31-554) at *k*=19. Increasing *k* size resulted in increasing amounts of overestimation. The Jaccard distance also correlates well with the accessory size ranging from 0.953-0.991 for *k*=49 and *k*=19 respectively (figure 2).

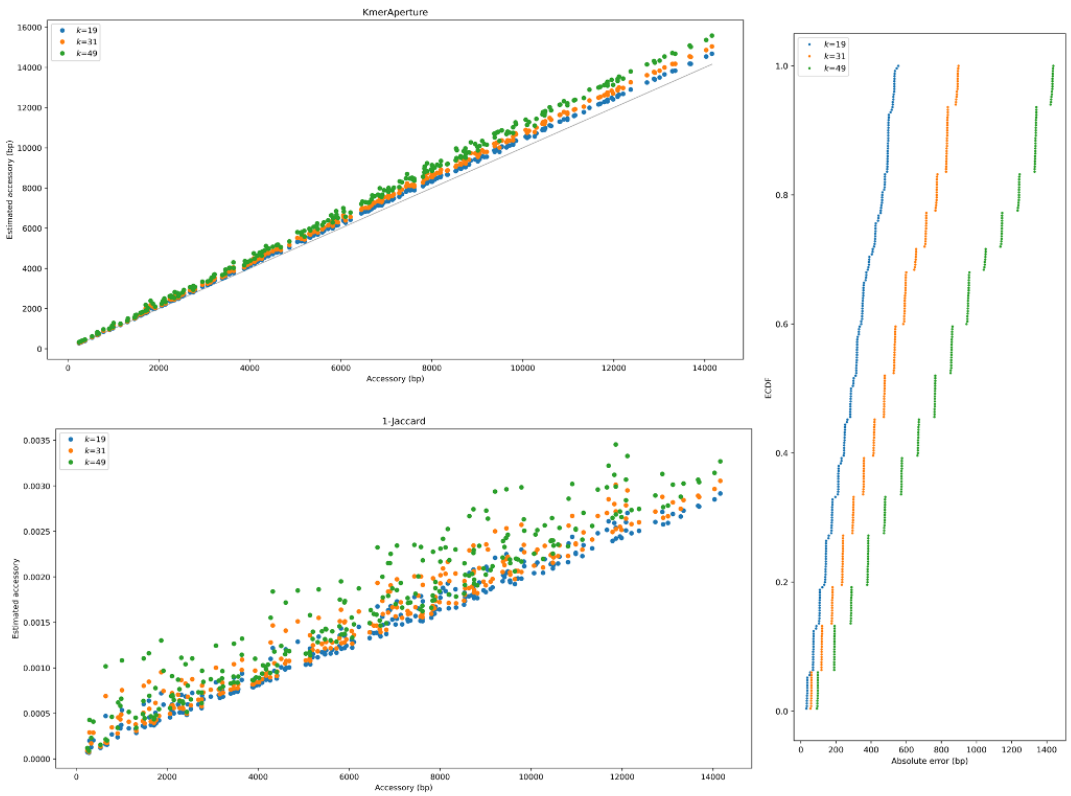

Figure 2. Scatterplot of KmerAperture accessory estimates (bp) for *k*=19,31,49 vs true accessory (bp) (top left) and Jaccard distance (bottom left). Grey horizontal line indicates a perfect correlation with true accessory size. The empirical cumulative distribution function (ECDF) of absolute error of KmerAperture estimates from the true values is displayed for each *k* (right)

Application to *E. coli* ST1193

The *E. coli* ST1193 population genomes had a great number of low-diversity query genomes compared with the reference. Overall, 275 genomes were within 100 SNPs from the reference including 45 that were within 50 SNPs. KmerAperture was run for all 407 genomes, with full Jaccard distances and SKA for comparison (Figure 3). There was a median (range) of 75 (29-514) SNPs compared with the reference in the population. From querying each *E. coli* genome against the reference, MCJCHV-1, KmerAperture predicted a median (range) of 41 (9-194), 73 (28-257), 73 (29-245) and 73 (29-222) SNPs for *k*=15, 21, 31 and 51 respectively. SKA produced SNP estimates with a median (range) of 78 (32-323), 75 (29-271), 75 (29-249) and 74 (29-223) for *k*=15, 21, 30 and 51 respectively. Correlation coefficients were calculated between true numbers of SNPs and Jaccard distances, SKA SNP estimates and KmerAperture SNP estimates. The strongest correlation between KmerAperture SNP estimates and real SNPs was when *k*=19 (0.971) while the lowest was when *k*=49 (0.892). For comparison, the correlation between Jaccard distance was <0.4 for all sizes of *k*, increasing from *k*=19 to *k*=49 (0.317 to 0.362). The strongest correlations with true SNPs was SKA

SNP estimates at *k*=19 (0.984) and *k*=31 (0.983) followed by KmerAperture at *k*=19 with 0.971. KmerAperture further correlated 0.945 at *k*=31 and 0.892 for *k*=49 and SKA 0.963 at *k*=49.

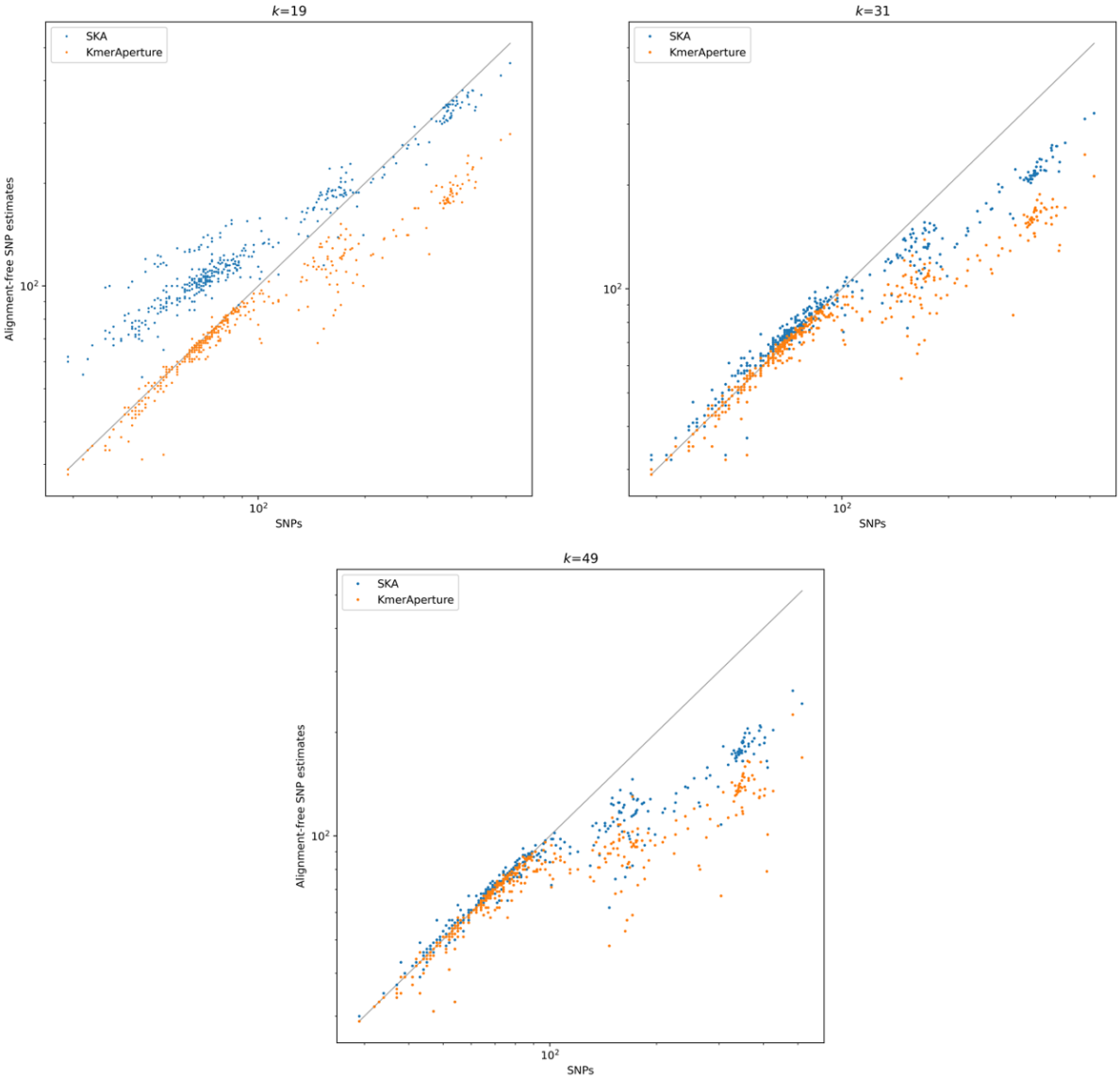

Figure 3. Results on *E. coli* ST1193 SNP distances. Scatterplot of KmerAperture and SKA SNP estimates vs real SNPs, for *k*=19, 31 and 49. *x* and *y* axis *log10* scaled

The accessory size difference between all *E. coli* query genomes and the reference had a median (range) of 487,000 (149,000-998,000), making the maximum size difference almost 1Mbp. The correlation coefficient was calculated between KmerAperture accessory size difference estimates and Jaccard distance against the real accessory size differences (Figure 4). The correlations were similar with 0.953, 0.954 and 0.954 for KmerAperture and 0.957, 0.956 and 0.954 with Jaccard distance for *k*=19, *k*=31 and *k*=49, respectively. The smallest absolute error was with KmerAperture *k*=51 with a

median (range) of 213,546bp (45,108-494,306), or as a percent of the real accessory size: 43.4% (27.5%-60.1%).

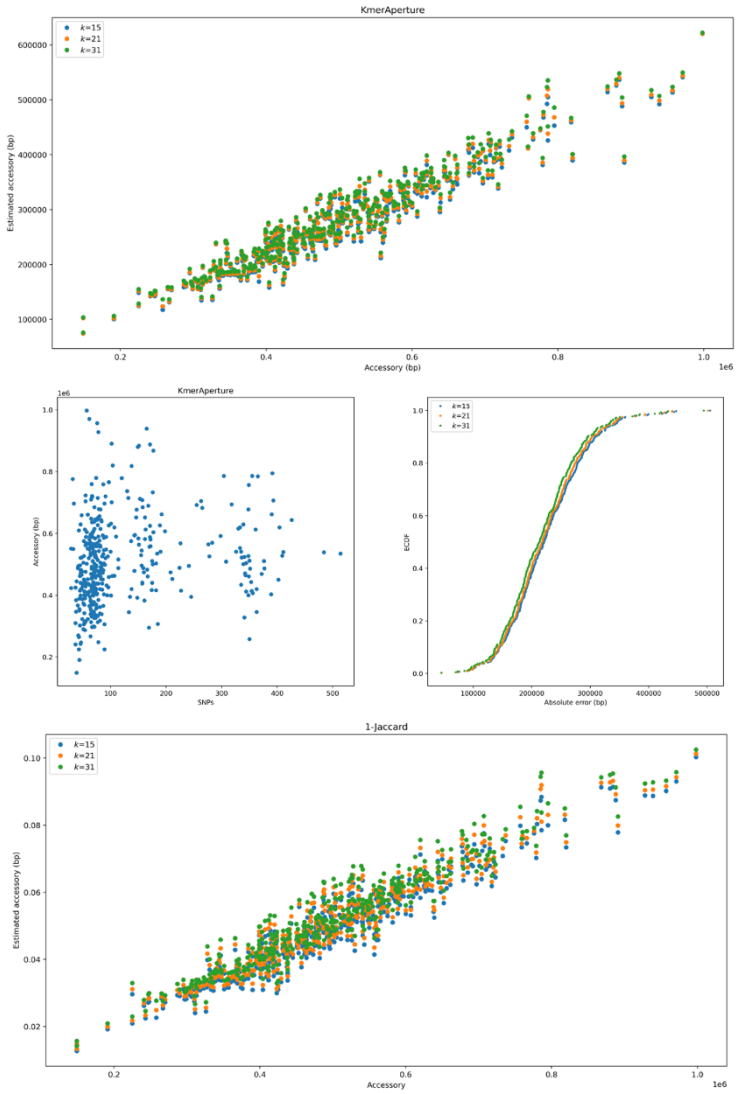

Figure 4. Results on *E. coli* ST1193 accessory distance. Scatterplot of KmerAperture and Jaccard distance accessory difference estimates across *k*=19,31,49 (top and bottom respectively). The empirical cumulate distribution function of absolute error (centre right). Scatterplot of real accessory difference vs real SNP difference (centre left)

Application to *C. difficile* RT017

The *C. difficile* within-lineage population is the most diverse under study, more diverse than either KmerAperture or SKA are built towards or recommended for. It is crucial that the sensitivity in clustering closely related genomes is not decreased by including very diverse genomes also. However, without alignment this lineage-diversity level is approximately the level that can be clustered with a median (range) of 70 (66-2,464) SNPs.

The correlation coefficient was determined for the entire population and separately for pairs separated by ≤1000 SNPs. KmerAperture SNP estimates correlated strongly with real SNPs with 0.969, 0.927 and 0.904 for *k*=19, *k*=31 and *k*=49 respectively. Similarly, SKA correlated 0.990, 0.979 and 0.952 for *k*=19, 31 and 49 respectively. There was a reduction for genomes separated by ≤1,000 SNPs with 0.967 for *k*=19 and with larger reductions for *k*=31 and *k*=49 to <0.9. The strongest correlation was with SKA at *k*=19 which reduced only to 0.975 for ≤1,000 SNPs. The strongest correlation for Jaccard distance was with *k*=51 (0.331) and weakest with *k*=15 (0.223).

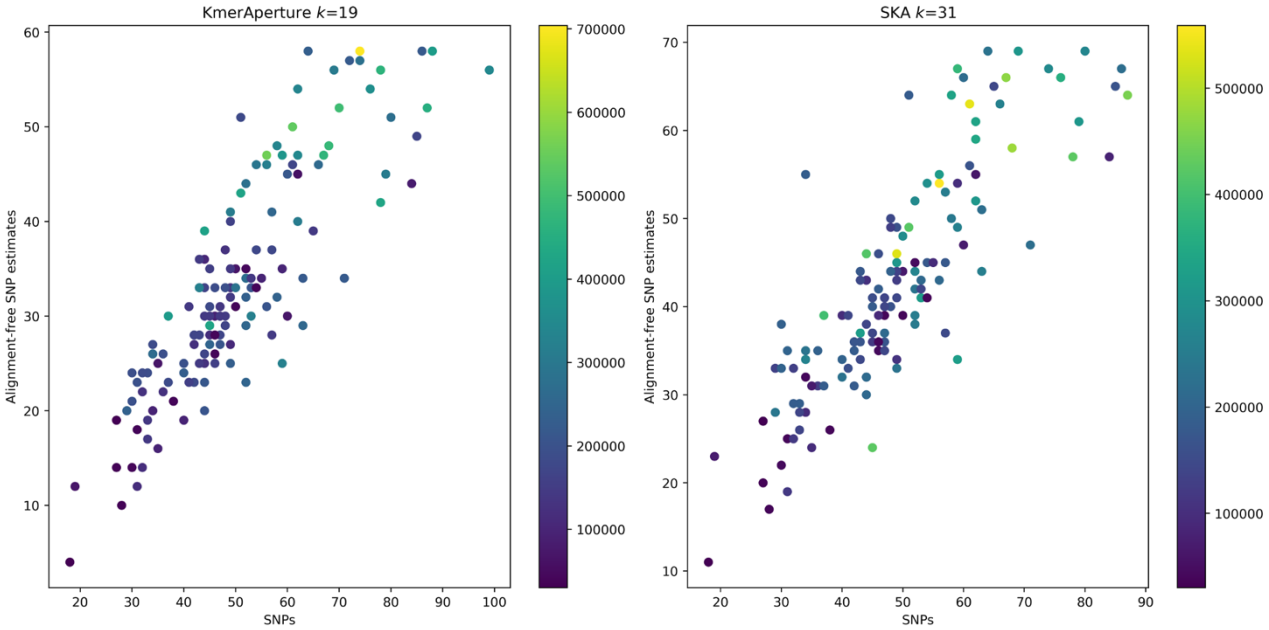

Figure 5. Scatterplot of SNP estimates vs SNPs for SKA *k*=31 (right) and KmerAperture *k*=19 (left). Each show the largest possible real SNP threshold that can be used and corresponding estimate threshold while excluding false positives >100 SNPs in dissimilarity. All points are colored by pairwise accessory difference.

The largest SNP threshold possible while excluding diverse false positives and retaining a sensitivity of 100% was investigated (Figure 5). Diverse false positives are defined here as those with >100 SNPs. With KmerAperture, the largest SNP threshold possible was ≤65 with *k*=19. This corresponded to a KmerAperture estimate SNP threshold of 58, resulting in a largest false positive SNP distance of 99 (*TP*=125, *FP*=20, specificity=0.87, PPV=0.86). With SKA it was possible to have the same largest SNP threshold of ≤65 with *k*=31, whereby the largest false positive SNP difference was 87, corresponding to an SKA estimate of 69 SNPs. Additionally, there were fewer false positives with SKA *k*=31 than with KmerAperture at *k*=19 (*TP*=125, *FP*=14, specificity=0.91, PPV=0.89).

The correlation with real accessory difference (Figure 6) was strongest with KmerAperture *k*=19 (0.970) followed by KmerAperture *k*=31 (0.964) and finally, *k*=49 (0.958) and Jaccard distance at *k*=15 (0.962). Similarly, Jaccard distances correlated 0.962, 0.954 and 0.944 for *k*=19, *k*=31 and *k*=49 respectively. The smallest median absolute error with KmerAperture resulted from *k*=51, with a median (range) 60,195bp (589-236,836). The median proportion of incorrect accessory bases was also smallest for *k*=49 at 28.9%. Proportional error was weakly correlated with an increase in real accessory size difference (0.37) for *k*=49.

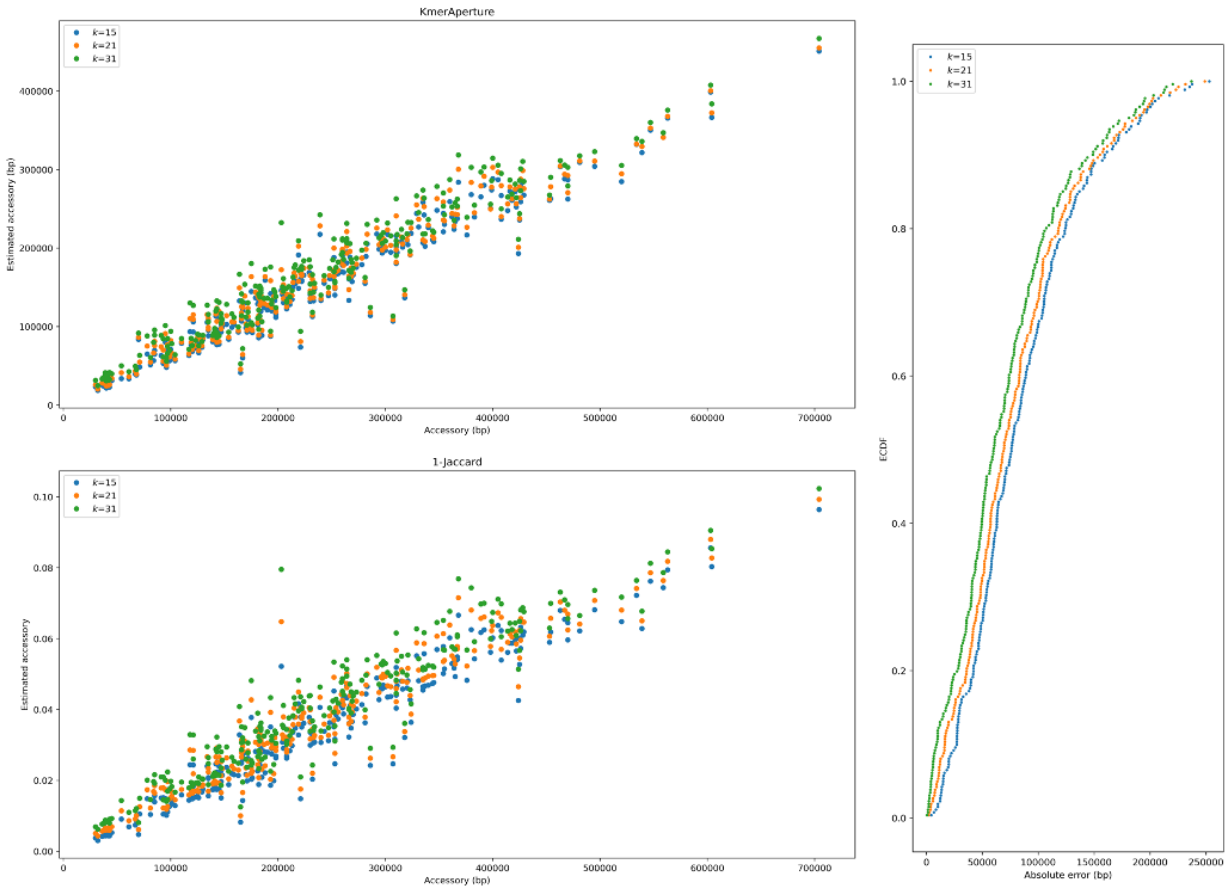

Figure 6. Results on *C. difficile* RT017 accessory distance. Scatterplot of KmerAperture and Jaccard distance accessory difference estimates across *k*=19,31,49 (upper and lower left respectively). The empirical cumulate distribution function of KmerAperture absolute error (bp) (right)

KmerAperture runtime

Across all datasets the median (range) time in seconds for KmerAperture analysis was 4.3 (2.6-8.9). Though the datasets varied in genome size, SNP and accessory diversity KmerAperture ranged in median analysis runtime from 3.6-4.7 seconds across all *k*. Median runtime increased with each increase in *k*, resulting in a 26% slowdown from *k*19 to *k*49 with a median of 4 seconds for *k*=19 and

5.4 seconds for *k*=49. We provide a mean runtime for SKA map as all genomes were run at once. Running individually would involve re-reading the reference genome each run artificially increasing the runtime. As such, we provide a mean as overall runtime / number of query genomes. The mean runtime for SKA map ranged from 13.1 for *k*=19 to 14.6 for *k*=31 and 15.9 seconds for *k*=49. This represents a 17.6% slowdown from the smallest to largest *k*-mer sizes. The majority of KmerAperture runtime was comprised of extracting the *k*-mers from sequences and, discerning the canonical *k*-mers including reverse complementing each. The entire process took a median (range) of 48.3 (29.1-71.5) seconds across datasets. SKA split *k*-mer file (.skf) creation meanwhile took a median (range) of 11.8 (7.7-17.9).

Discussion

The KmerAperture algorithm is based on the few axioms we have defined of how relative sequence differences are represented by series of *k*-mers. It was, however, possible to accurately estimate SNP differences. Initially the algorithm filters mismatched *k*-mers by their original order, nominating those that equal *k* in series length as being generated by a single SNP difference between the genomes. This is followed, comparable to SKA's methodology, by matching the middle *k*-mers without the middle base.

It is also possible with KmerAperture to identify accessory regions. It is possible therefore, to extract accessory sequence positions, prior to alignment. The computational and temporal challenge of alignment is related both to the combinatorial size of sequence comparisons but also the diversity of the sequences under study. First reducing this diversity, especially in genome samples of unknown diversity, could reduce the cost of alignment. KmerAperture could also be adapted to progressively remove accessory regions as genomes are added, compare the genomes to a core genome alignment, or, as in this study, compare to a reference. Proportion of error remained constant in the median underreporting of accessory size difference, which became more pronounced in basic (bp) accessory with larger accessory genomes. *E. coli* MCJCHV-1 contains a number of plasmids[34], which were not included for comparison but are likely present in the query genomes. However, the plasmids combined do not approach 1Mbp and as such members of the lineage have large accessory differences, of a similar length as was previously reported for *E. coli*[35].

Datasets were chosen to reflect real-life scenarios of the type of diversity we might cluster genomes to, by genomic typing, or by methods, such as MinHash, without alignment. If it could be known up-

front that genomes are within a 5 SNP single-linkage cluster, for instance, then there would be little need for KmerAperture. Both SKA and KmerAperture are designed such that they will perform best in low-diversity samples, but they also need to first differentiate between tens and hundreds or thousands of SNP distances before attempting to estimate the distances of those of low-diversity. Analysis of a lineage-level diversity population (RT017) showed that both KmerAperture and SKA were robust to very diverse false positive inclusion. As a result, low-diversity genomes can be clustered for further, alignment-based comparison.

In a scenario where there is a survey of bacterial genomes, for instance in a clinical setting it is important to rapidly differentiate species, lineage, sub-lineage and possible outbreak clusters based on genetic distance. The size and diversity of these datasets can render alignment-based methods difficult and possibly intractable with most available computational resources. KmerAperture runs on a laptop, requiring few resources as it exploits the efficiency of python set-analysis. The KmerAperture algorithm is also efficient without using a compiled language. With the logic of syntenic *k*-mers, it is possible to produce much faster runtimes and render this a scalable approach for genomic surveillance. Whilst scalable, the problem of discerning SNPs >100 and <1,000's (typical within-lineage ranges) remains unresolved by alignment-free approaches. With KmerAperture however, the entire relative diversity of genomes is encoded in the syntenic *k*-mer series. Methods may further be developed to discern multiple SNPs within *k* of one another, indels and smaller accessory fragments from *k*-mer series.

Bibliography

1. Didelot, X. & Parkhill, J. A scalable analytical approach from bacterial genomes to epidemiology. *Philos. Trans. R. Soc. B Biol. Sci.* **377**, 20210246 (2022).
2. Didelot, X., Bowden, R., Wilson, D. J., Peto, T. E. A. & Crook, D. W. Transforming clinical microbiology with bacterial genome sequencing. *Nat. Rev. Genet.* **13**, 601–612 (2012).
3. Dallman, T. *et al.* SnapperDB: a database solution for routine sequencing analysis of bacterial isolates. *Bioinformatics* (2018). doi:10.1093/bioinformatics/bty212
4. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* (2016). doi:10.1186/s13059-016-0997-x
5. Titus Brown, C. & Irber, L. sourmash: a library for MinHash sketching of DNA. *J. Open Source Softw.* (2016). doi:10.21105/joss.00027
6. Zhao, X. BinDash, software for fast genome distance estimation on a typical personal laptop. *Bioinformatics* (2019). doi:10.1093/bioinformatics/bty651
7. Baker, D. N. & Langmead, B. Dashing: Fast and Accurate Genomic Distances with HyperLogLog. *bioRxiv* (2019). doi:10.1101/501726
8. Moore, M. P., Wilcox, M. H., Walker, A. S. & Eyre, D. W. K-mer based prediction of Clostridioides difficile relatedness and ribotypes. *bioRxiv* 2021.05.17.444522 (2021). doi:10.1101/2021.05.17.444522
9. Harris, S. R. *et al.* Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: A descriptive study. *Lancet Infect. Dis.* (2013). doi:10.1016/S1473-3099(12)70268-2
10. Sherry, N. L. *et al.* Genomics for molecular epidemiology and detecting transmission of carbapenemase-producing enterobacterales in Victoria, Australia, 2012 to 2016. *J. Clin. Microbiol.* (2019). doi:10.1128/JCM.00573-19
11. Didelot, X. *et al.* Microevolutionary analysis of Clostridium difficile genomes to investigate transmission. *Genome Biol.* (2012). doi:10.1186/gb-2012-13-12-r118
12. Lees, J. A. *et al.* Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res.* (2019). doi:10.1101/gr.241455.118
13. van der Putten, B. C. L., Huijsmans, N. A. H., Mende, D. R. & Schultsz, C. Benchmarking the topological accuracy of bacterial phylogenomic workflows using in silico evolution. *Microb. Genomics* (2022). doi:10.1099/mgen.0.000799
14. Gardner, S. N., Slezak, T. & Hall, B. G. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics* (2015). doi:10.1093/bioinformatics/btv271
15. Harris, S. R. SKA: Split Kmer Analysis Toolkit for Bacterial Genomic Epidemiology. *bioRxiv* (2018). doi:10.1101/453142
16. Crusoe, M. R. *et al.* The khmer software package: enabling efficient nucleotide sequence analysis. *F1000Research* (2015). doi:10.12688/f1000research.6924.1
17. © Copyright 2017 NumPy developers. NumPy — NumPy. *NumPy Website* (2017).
18. Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* (2021). doi:10.21105/joss.03021
19. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* (2007). doi:10.1109/MCSE.2007.55
20. Kluyver, T. *et al. Jupyter Notebooks—a publishing format for reproducible computational workflows. Positioning and Power in Academic Publishing: Players, Agents and Agendas* (2016). doi:10.3233/978-1-61499-649-1-87
21. Anaconda Software Distribution. *Anaconda Documentation* (2020).
22. Imwattana, K. *et al.* Global evolutionary dynamics and resistome analysis of Clostridioides difficile ribotype 017. *Microb. Genomics* (2022). doi:10.1099/mgen.0.000792
23. He, M. *et al.* Evolutionary dynamics of Clostridium difficile over short and long time scales. *Proc. Natl. Acad. Sci.* (2010). doi:10.1073/pnas.0914322107
24. Zhou, Z., Charlesworth, J. & Achtman, M. HierCC: a multi-level clustering scheme for population assignments based on core genome MLST. *Bioinformatics* **37**, 3645–3646 (2021).

25. Zhou, Z., Alikhan, N. F., Mohamed, K., Fan, Y. & Achtman, M. The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. *Genome Res.* (2020). doi:10.1101/gr.251678.119
26. Liu, P. *et al.* Complete genome sequence of Klebsiella pneumoniae subsp. pneumoniae HS11286, a multidrug-resistant strain isolated from human sputum. *Journal of Bacteriology* (2012). doi:10.1128/JB.00043-12
27. Han, E. S. & goleman, daniel; boyatzis, Richard; Mckee, A. NCBI SRA Toolkit Technology for Next Generation Sequence Data Stephen. *J. Chem. Inf. Model.* (2019).
28. Krueger, F. Trim Galore!. *[http://www.bioinformatics.babraham.ac.uk/projects/trim galore/]* (2012). doi:http://www.bioinformatics.babraham.ac.uk/projects/trim galore/
29. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–77 (2012).
30. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* (2013).
31. Seemann, T. mlst.
32. Seeman, T. Snippy-2.0. Available at: http://www.vicbioinformatics.com/software.snippy.shtml.
33. Laing, C. *et al.* Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. *BMC Bioinformatics* **11**, 461 (2010).
34. Nielsen, D. W. *et al.* Complete Genome Sequence of the Multidrug-Resistant Neonatal Meningitis Escherichia coli Serotype O75:H5:K1 Strain mcjchv-1 (NMEC-O75). *Microbiol. Resour. Announc.* (2018). doi:10.1128/mra.01043-18
35. Didelot, X., Méric, G., Falush, D. & Darling, A. E. Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli. *BMC Genomics* (2012). doi:10.1186/1471-2164-13-256

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

Data availability

All data is publicly available and acquired from the National Centre for Biotechnology Information (NCBI) databases and Enterobase. All accession IDs are provided in supplementary information. Open-source code is available from: https://github.com/moorembioinfo/KmerAperture

Author contributions

All authors contributed to this research and resulting manuscript. Their contributions are listed here in accordance with the Contributor Roles Taxonomy (CRediT). ***Matthew P. Moore:*** conceptualisation, methodology, software, validation, formal analysis, data curation, writing - original draft, visualisation. ***Mirjam Laager:*** conceptualisation, methodology. ***Xavier Didelot:*** conceptualisation, methodology, writing - review and editing, resources, supervision, project administration, funding acquisition.

Acknowledgements

This study is funded by the National Institute for Health and Care Research (NIHR) Health Protection Research Unit in Gastrointestinal Infections, a partnership between the UK Health Security Agency, the University of Liverpool and the University of Warwick. The views expressed are those of the author(s) and not necessarily those of the NIHR, the UK Health Security Agency or the Department of Health and Social Care.