# Appendix 4: Analysis of metagenomics data for evidence of priority taxa

## Purpose

This part of the analysis does not relate to the AMR study, and was performed on request by the FSA for food-safety reasons. Potentially, the results could be used in future to relate to findings regarding ARGs in the samples concerned, but that was not an objective here.

The object was to identify priority taxa of concern in any samples still within the use-by date. Naturally, at the time of purchase all samples were within the expiry date (if stated), and were processed for DNA extraction promptly. The expiry date of only a few samples (such as ice cream) had not yet passed by the time the sequence data was available for analysis. The analysis used the short-read NovaSeq metagenomics data, and the contig sequences assembled from them.

## Methods

### Organisms of concern, as defined by FSA:

* *Campylobacter* (any)
* *Clostridium botulinum*
* *Clostridium perfringens*
* *Enterococcus faecalis*
* *Enterococcus faecium*
* *Escherichia coli*
* *Klebsiella pneumoniae*
* *Listeria monocytogenes*
* *Salmonella* (any)
* *Shigella* (any)
* *Staphylococcus aureus*
* *Yersinia enterocolitica*

### Resources and Procedures

The steps of the procedure were as follows:

1. An initial processing of all samples with an established, relatively fast tool for profiling the composition of microbial communities from unassembled metagenomic sequencing data, (MetaPhlAn3 (Beghini et al., 2020)).
2. For those in-date samples which were positive for at least one of the priority taxa according to MetaPhlAn3, an analysis of the assembled contig sequences (to attempt to confirm any priority taxa, irrespective of which were positive in (1)). This is because closely-related taxa can have highly similar sequences, and differences may be missed if considering only individual read pairs. Complementary analyses were performed on these samples:
   1. More in-depth analysis using sequences of MLST (Multi Locus Sequence Typing) loci (where available) of all the priority taxa. This included:
      1. Comparison of all contig sequences with the MLST sequences using BlastN (Camacho et al., 2009)
      2. Comparison of the resulting MLST-positive contigs only versus the NCBI nt database (to provide confirmation of the MLST result; since as few as 1 MLST sequence may be involved, it is useful to compare the full-length positive contigs with a more general database).
   2. Irrespective of the MLST-sequence results, processing and inspection of the results of BlastN of all the sample's contigs versus the NCBI nt database (a release from March 2020), again with regard to all priority taxa and not just those positive in (1).

#### The MetaPhlAn3 database

The reference database was checked to ensure that all of the taxa of concern are present. All taxa were present explicitly by name, except for *Shigella -* potentially this is due to alternative taxonomies treating these bacteria differently. For example, in some cases it is grouped with *Escherichia*. However, even if the relevant markers are present in the database, this does mean that all samples will necessarily be negative for *Shigella* by that name.

#### Identification of any well-characterised "classical" marker genes present

The marker genes involved in the MetaPhlAn3 analysis are essentially derived computationally, by a large-scale analysis which aims to identify genes which are unique to all and any taxonomic clades, including species.

Some or even all of these may be very reliable (and may well include "classical" marker genes) but in order to confirm gene-sequence matches by inspection, it is easier to use genes which are well characterised in the field; these are known by experts to characterise particular organisms (i.e. different sequence variants are known to be unique to a particular species, etc). Positively identifying any such markers thus provides a greater weight of evidence compared to identifying parts of the genome which may currently be diagnostically "unique" according to computational methods; potentially, they might also occur in different closely-related species, where the sequences in question are not in any databases (yet). This is especially a concern for Enterobacteriaceae species. In contrast, the classical markers have been intensively studied for many years.

A further difference is that MetaPhlAn3 uses the samples' unassembled read sequences, whereas the methods described here use the assembled "contig" sequences (these had been assembled from the reads using MEGAHIT, following removal of read pairs which matched host genomes; refer to main text of Report).

Sequences of alleles of MLST loci were obtained for all of the taxa of concern where possible, or for the most suitable super-taxon otherwise - i.e., not just for the 4 species implicated above.

##### MLST schemes used

Sequences of alleles of MLST loci were obtained for all of the taxa of concern where possible, or for the most suitable super-taxon.

* *Campylobacter:* PubMLST (<https://pubmlst.org>) provides sequences for two relevant schemes: one is for *C. jejuni* and *C. coli* collectively, the other for other species.
* *Clostridium botulinum:* obtained from PubMLST.
* *Clostridium perfringens:* in brief, it appears that there are resources at PubMLST but these are "protected". At the front-page portal, this species is not listed.
* *Enterococcus faecalis:* obtained from PubMLST.
* *Enterococcus faecium:* obtained from PubMLST.
* *Escherichia coli*: Sequences of the Achtman and Pasteur schemes were obtained from PubMLST. Although the description at PubMLST is of "***Escherichia* spp.**", and via the front-page PubMLST databases portal there is nothing described as *E. coli-*specific, the Achtman sequences appear to be the same (including in almost exactly the same numbers) as those for *E. coli* (obtainable via a search of the PubMLST site). This is the same situation as for the Pasteur scheme sequences (also identical to those obtainable from the Pasteur Institute, where they are described as "*E. coli"*)*.*
* *Klebsiella pneumoniae:* sequences of the scheme collectively for *K. pneumoniae, K. quasipneumoniae* and *K. variicola* were obtained from the [Pasteur Institute](https://bigsdb.pasteur.fr/).
* *Listeria monocytogenes:* obtained from the Pasteur Institute.
* *Salmonella:* obtained from PubMLST.
* *Shigella*: none obtained specifically by that name alone. (Enterobase has a scheme "*Escherichia/Shigella"* but we did not use that here).
* *Staphylococcus aureus:* obtained from PubMLST.
* *Yersinia enterocolitica:* obtained sequences of a scheme for "*Yersinia* spp. (legacy)" from PubMLST. The Pasteur Institute appears to publicly distribute only **cg**MLST sequences (for *Yersinia,* for *Y. enterocolitica*, and for *Y. pseudotuberculosis*).

In summary, the compiled database includes scheme sequences for all of the taxa except *Clostridium perfringens* and (possibly) *Shigella.* Some taxa are represented by more than one scheme, and some schemes represent a wider range of taxa than the one named in the of-concern list.

##### Summary of database sequences

The MLST locus allele sequence database totals 37,004 sequences in all, as shown in Table 1.

Table - MLST locus allele sequence database entries

|  |  |  |
| --- | --- | --- |
| *Campylobacter coli + jejuni* | 7 loci | 5,126 |
| *Campylobacter* other | 12 loci | 3,013 |
| *Clostridium botulinum* | 7 loci | 257 |
| *Enterococcus faecalis* | 7 loci | 659 |
| *Enterococcus faecium* | 7 loci | 767 |
| *Escherichia* (Achtman) | 7 loci | 6,832 |
| *Escherichia* (Pasteur) | 8 loci | 1,728 |
| *Klebsiella* | 6 loci | 2,378 |
| *Listeria monocytogenes* | 7 loci | 2,785 |
| *Salmonella* | 7 loci | 7,674 |
| *Staphylococcus aureus* | 7 loci | 5,240 |
| *Yersinia* | 7 loci | 545 |

#### Search methods

The comparison with the database of MLST locus allele sequences was performed with BLASTN of the Blast 2.9.0+ package. Default parameters were used other than a maximum E-value of 10-3.

#### General sequence database

We used a local copy of the NCBI nt database, dating from March 2020.

#### Interpretation of results

The terms 'exact/perfect matches' need to be precisely defined, because otherwise they may be ambiguous. Also, even given such a definition, the conclusions which can/cannot be made from such matches also need to be carefully described.

Definition: **HSP**. High-scoring Segment Pairs (also known as Maximal-scoring Segment Pairs, MSPs) is the formal term for the matching alignment between a full, or partial, query sequence with a full, or partial, database sequence (the "hit"), arising from a Blast search. There may be more than one HSP for a given query sequence-hit sequence pair. Such alignments can include mismatches (which might be referred to as "substitutions" in an evolutionary context) and insertions/deletions (indels). An HSP exhibits 100% sequence identity if there are no mismatches nor indels. An HSP may be as long as, or shorter (sometimes much shorter) than either the full-length query sequence, or the full-length database (hit) sequence, or both. Some HSPs can be very short (e.g. < 30 b.p.).

Definition: **perfect match.** For the purposes here, a perfect match with a database sequence consists of a query contig sequence which is at least as long as the database sequence, and **which is identical to it over the full length of the database sequence**. Clearly, database sequences in this case are limited to gene (allele) sequences, so where such a sequence occurs in a query contig that is longer than a gene, the full length of the contig can never be included in the match.

**Lack of perfect matches**: If the **best** match between a given query sequence (or part thereof) covers the whole database sequence but is not identical to it, then for these purposes, it cannot be concluded with certainty that the query represents the organism of provenance of the database sequence (call this organism X).

One rationale for this is that MLST loci are well characterised, with in some cases hundreds of different sequence variants identified and curated by experts (see numbers above). Therefore even if the difference between the query sequence and the best match is only a single b.p., it would be questionable to assume that the query sequence constitutes a previously unseen variant of this organism X's gene.

How *likely* it is for that assumption to be true depends on the existence of reference sequences which are also (in this example) 1 b.p. different to the X reference, but annotated as originating from a different organism; and/or annotated as originating from X. But irrespective of that, in this example the query sequence would be considered as no more than an unconfirmed (novel) instance of an allele from organism X.

Definitions: **Exact matches (with respect to query)**. Some query contig sequences are shorter than the MLST sequences and so can never make a perfect match. In the event that a contig sequence represents a genuine instance of a fragment (**sub-sequence**) of a gene sequence from organism X, then the full-length of the query will form an HSP, of the same length. A different configuration is that the query contig can represent a "fragment of the reference" and still *not* be a sub-sequence, if the full-length query does not include the full length of the reference, but extends beyond either the 5' or 3' end (but not both ends) of the reference. This is still considered as an exact match w.r.t. query if the overlapping region is 100% identical, because it represents the best possible match, given that the query sequence is too short to cover the whole reference. "Too short" can still mean a query far longer than the reference; for example, if the query is 10,000 bp long, and the last 250 bp match exactly with the first 250 bp of a 400 bp-reference, then the query is still effectively 150 bp too short to be able to make a perfect match. But over the 250 bp region in common between the query and reference, the match is exact. (Compare and contrast this with an internal HSP - see below).

The converse term "exact match w.r.t. reference" is avoided because it might suggest that the full-length of the reference is included in the HSP, when it is not. All cases of an HSP encompassing the full-length reference are deemed "perfect" (see above).

Definitions: **all other matches are deemed partial/inexact**.

Definitions: an "**internal HSP**" is one whose query- and hit-segments are each flanked on either or both sides by a significant number of base pairs (i.e. both the query and the hit sequence extend beyond the HSP, yet they do not match). "Significant" refers to the allowance of a small number of bp at either/both ends to be missing simply because they do not match; these are effectively mismatches which do not appear in the HSP. E.g. if a query segment starting at bp 3 aligned with bp 1,000 of a reference sequence, that would not be considered significant mismatch. If a query segment starting at bp 103 aligned with bp 1,000 of the reference, then that means that there are ~ 100 bp of the query and reference which are too different to be aligned in the HSP. The query could not then be considered to be an observed instance of the reference sequence, even if there was a substantial HSP, even of thousands of bp, with an identity of 100% and no indels. Misassembled contigs could thus be falsely treated as negatives, but the contig sequences here are treated as-is.

#### Consequences of perfect and exact matches

##### Perfect matches

Even if the full-length of a reference sequence matches all (or part) of a query exactly (i.e. a perfect match) this does not necessarily constitute a positive identification. If there are other reference sequences in other organisms which are also identical over the full length of this region, then if their provenance is other species of concern then these should be present in the MLST database (see above). Irrespective of that, it is still possible that reference sequence(s) from other (non-concern) organisms are also identical over this full-length, but to confirm that possibility necessitates an additional analysis (i.e. sequence similarity search versus a larger, more general sequence database such as NCBI's nonredundant nucleotide sequence database, "nt").

In summary, a query might be found (after additional search(es) versus database(s) external to the MLST) to make a perfect match with (i) organism of concern X alone; or (ii) with more than one organism of concern, e.g. X, Y; or (iii) with organism of concern X (and possibly others) and at least one non-concern organism.

Given that we are dealing with MLST locus sequences (well-characterised and covering a wide range of alleles) then it can be assumed that close but non-identical matches are also reliably annotated. E.g. if a query makes a perfect match only with a reference from organism X and is 1 b.p. different from a reference from organism Z, it is concluded that the query represents X and not Z. Admittedly, this conclusion might be considered still firmer if the closest non-X match to the query differs in many more b.p. than just 1.

Outcomes (i) and (ii) mean that at the bioinformatics level, the sample is positive for organisms of concern, by dint of this query. Outcome (ii) would mean that (for this query at least), it is unknown which of two or more organisms of concern are present (both/all could be).

Outcome (iii) means that the sample cannot be concluded as positive from this query.

##### Exact matches with respect to the query

The conclusion from these exact matches are essentially the same as for the perfect matches, i.e. they require confirmation/refutation by searching other databases. The only difference is one of expectation, i.e. that because some of the reference sequence is in essence "missing" from the query - i.e., the query is shorter- then it is more likely that sequences completely identical to the query (or that part of it which aligns with the reference) will occur in other organisms.

#### Summary of analysis steps

For each sample:

Compare query sequences (all contigs) versus the MLST database.

For each of the taxa, if there are any perfect or exact matches, which match only that taxon or only taxa within the list of confirm these must be confirmed by comparing the **putatively positive query sequences** against a more general database (nt); that is, to rule out that they might also make as good matches with other organisms.

If there are no perfect or exact matches for any taxa, then that constitutes a negative as far as these MLST sequences are concerned. For all of the taxa (of those which tested positive by MetaPhlAn3) still not shown to be positive, then **all of the contigs** must then be searched versus a general database (nt).

## Results

### First-pass screen: analysis of all 256 samples with MetaPhlan3

First, the reference database was checked to confirm that reference sequences are present for all of the taxa of concern. This was true with the exception of *Shigella*.

Metaphlan3 was run on all 256 samples, and the output profiles for each sample were searched for the taxa of interest.

4 taxa (*Enterococcus faecalis, Enterococcus faecium, Escherichia coli, Klebsiella pneumoniae*) were found across 17 samples (2664788, 2664542, 2672455, 2672484, 343624 2664541, 2664571, 2664691, 2672480, 2672792, 343541, 343554, 6297, 6412, 6473, 6628, 6647); Table 2.

Note that MetaPhlAn3 uses unassembled read sequences as input.

Table 2 Incidence of one or more marker sequences in each sample's metagenomic read data as identified by MetaPhlan3. Expiry dates which had not yet passed at the time of the analysis are marked with a \*. Sample 2664571 was due to expire a matter of days after this analysis began.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Sample ID | Food category | Expiry date | *E. faecalis* | *E. faecium* | *E. coli* | *K. pneumoniae* |
| 6297 | White grapes | 30/09/2019 | - | - | - | Y |
| 6412 | Vanilla ice cream | 01/08/2021\* | Y | Y | Y | Y |
| 6473 | Orange juice pasteurised | 30/11/2019 | - | - | - | Y |
| 6628 | Bananas | not provided | - | - | - | Y |
| 6647 | Apples | 06/09/2019 | - | - | - | Y |
| 343541 | Semi skimmed milk | 14/07/2019 | - | - | - | Y |
| 343554 | Semi skimmed milk | 14/07/2019 | - | - | - | Y |
| 343624 | Whole milk | 11/07/2019 | - | - | Y | - |
| 2664541 | Grapefruit | not provided | - | - | - | Y |
| 2664542 | Corned beef | 11/10/2019 | - | Y | - | - |
| 2664571 | Raisins | 31/07/2020 | - | - | - | Y |
| 2664691 | Blueberries | 13/09/2019 | - | - | - | Y |
| 2664788 | Peaches | 04/10/2019 | Y | - | - | - |
| 2672455 | Whole milk | 25/06/2019 | - | - | Y | - |
| 2672480 | Vanilla ice cream | 01/05/2021\* | - | - | - | Y |
| 2672484 | Whole milk | 02/08/2019 | - | - | Y | - |
| 2672792 | Bananas | not provided | - | - | - | Y |

### Identifying in-date samples among the putative positives

EH determined that of the 17 samples, two are from products with a use-by date specified which is yet to expire: sample IDs 2672480 and 6412.

SM determined that the MetaPhlAn3 results are positive for taxa as follows in these samples:

* 2672480: *Klebsiella pneumoniae*
* 6412: *Enterococcus faecalis*, *Enterococcus faecium, Escherichia coli, Klebsiella pneumoniae*

### Query sequences of the two positive samples

The sets of contig query sequences used were as described in Table 3.

Table 3 Characteristics of the contigs used as query sequences for the sequence similarity search of the MLST and nt databases.

|  |  |  |
| --- | --- | --- |
| Sample ID | #2672480 | #6412 |
| Number of sequences | 19,070 | 100,752 |
| Number of bases in sequences | 20,857,779 | 90,551,047 |
| Mean sequence length | 1,093.7 | 898.8 |
| Minimum sequence length | 201 | 200 |
| Maximum sequence length | 88,535 | 102,349 |

### Summary of results

Table 4 summarises the results of the three analyses performed on the metagenomics data of the two in-date samples.

Table 4 Summary of results for the Species of Concern analyses for the two positive samples. The unassembled short-read metagenomics was analysed with MetaPhlAn3. The assembled contig sequences were used as queries for sequence similarity searches of MLST and nt databases.

|  | Sample 2672480 vanilla ice cream | | | Sample 6412 vanilla ice cream | | |
| --- | --- | --- | --- | --- | --- | --- |
| Species of Concern | MetaPhlAn3 | MLST | BlastN nt+ inspection | MetaPhlAn3 | MLST | BlastN nt+ inspection |
| *Campylobacter* (any) | n | n | n | n | n | n |
| *Clostridium botulinum* | n | n | n | n | n | n |
| *Clostridium perfringens* | n | N/A | n | n | N/A | n |
| *Enterococcus faecalis* | n | n | n | Y | y3 | Y |
| *Enterococcus faecium* | n | n | n | Y | Y | Y |
| *Escherichia coli* | n | n | n | Y | y4 | Y |
| *Klebsiella pneumoniae* | Y | y1 | Y2 | Y | n | Y5 |
| *Listeria monocytogenes* | n | n | n | n | n | n |
| *Salmonella* (any) | n | n | n | n | n | n6 |
| *Shigella* (any) | n | N/A | n | n | N/A | n6 |
| *Staphylococcus aureus* | n | n | n | n | n | n |
| *Yersinia enterocolitica* | n | n | n | n | n | Y7 |

Notes.

1. Strictly speaking, the MLST sequence comparisons are consistent only with the contig in question (there is only one positive) originating from either *Klebsiella,* or *Escherichia-Shigella.* However, the single *Escherichia-Shigella* match appears to be quite an extreme outlier, i.e. in short, it looks very likely to be wrong. However, note that the MLST scheme is for '*Klebsiella'* and without a more in-depth check of the species of origin whose ST includes the allele variants in question (e.g. the only ones which match at 100% identity are mdh = 2, 31, 81, 153, 159, 198), it is not certain that these are all *K. pneumoniae*; they could also be *K. quasipneumoniae* or *K. variicola*.
2. BlastN analysis of the *mdh*-positive contig (see (1)) also does not strictly confirm *K. pneumoniae*, even though it appears to rule out *K. quasipneumoniae* and *K. variicola*; however, it does not rule out *K. aerogenes.* However, analysis of all contigs (not just the MLST-containing) does appear to confirm *K. pneumoniae* for some contigs. Of course, this does not rule out that other *Klebsiella* species may *also* be present
3. There are no 100%-identical matches, but 2 matches of 99.25% identity over 267 bp (*gyd* gene).
4. This exclusively matches *E. coli recA* at 100% identity over 262 bp.
5. The only *Klebsiella* matches at 100% identity are with *K. pneumoniae.* The vast majority of matches which are ≥ 99% identical are also from *K. pneumoniae*, but there are also two *K. aerogenes* sequences. This applies to multiple contigs. Again, it is quite possible that *K. aerogenes* is also present.
6. Among some of the 100% identity matches with *E. coli* sequences, some were also 100% identical to *Salmonella* or *Shigella* sequences. However, no sequences were found which exclusively matches these genera (in contrast to *E. coli*, where numerous 100% identical matches were found).
7. Several instances were found of contigs which exclusively made a 100% identical match with *Y. enterolitica*. However, all but one of those appear to be quite likely to result from erroneous annotations, because the *Y. enterolitica* matches are outliers compared to all other *Yersinia* sequences. In one case however, the match identity profile is extremely convincing. Very recent re-runs of the BlastN search at NCBI WWW confirm that the contig sequence concerned exclusively matches *Y. enterolitica* at 100% and many other matches are made to this species alone at gradually lower percentages, with the first instance of a different species (also *Yersinia*) some way down.

### Results of searching MLST database

Within the assembled metagenomic sequences of the ice cream samples, only #2672480 contained identifiable (perfectly-matching segments of) MLST sequences of *K. pneumoniae.*

That requires qualification, since this sequence segment was also identical to a single *Escherichia-Shigella* sequence, However, we concluded that this latter sequence is an extreme outlier compared to others of that group, and so may be a mis-annotation (see next subsection). This sample contained no other MLST sequences.

In sample #6412, we identified MLST sequences for *Enterococcus faecalis, Enterococcus faecium* and *Escherichia coli.*

Details of the identification of MLST sequences are now provided for each sample separately.

#### MLST Search Results: Sample 2672480

There is only one query contig sequence which makes any matches (of any length) of 100% identity to any MLST sequences.

Contig ID: k141\_5046, length 444 bp.

The following are the MLST sequences, and their total lengths, which make a 100%-identical, 444-bp HSP with it:

* Klebsiella\_mdh\_81 477
* Klebsiella\_mdh\_31 477
* Klebsiella\_mdh\_2 477
* Klebsiella\_mdh\_198 477
* Klebsiella\_mdh\_159 477
* Klebsiella\_mdh\_153 477
* Escherichia\_Shigella\_Achtman\_mdh\_636 452

A further 60 hits, all Klebsiella\_mdh\_*n* sequences, form a 444-bp HSP with only 1 differing bp (99.8% identity). A further 105 differ by 2 bp (99.6% identity). Others follow, with an increasing number of mismatches/indels, up to the 390th hit, which differs by 32 substitutions over 444 bp (92.8% identity). The next hit differs by more than 30 bp inlcuding indels and has an identity of 92.2%. All these top 391, with the exception of the single Escherichia\_Shigella sequence above, are *Klebsiella* sequences (not necessarily *K. pneumoniae*, as these could be from *K. pseudopneumoniae* or *K. variicola*).

Notably the query contig sequence exactly matches the beginning of the reference (i.e. the full length of the query is identical to the first 444 b.p. of the reference).

The 392nd hit is 84.7% identical over a 437-b.p. HSP (Escherichia\_Shigella\_Achtman\_mdh\_392), and all the remaining hits are also Escherichia\_Shigella\_Achtman\_mdh\_*n*, ranging from 84.4% to 82.3% identical (HSP lengths are either 437 or 441 b.p.). It is thus noticeable that the 100%-identical Escherichia\_Shigella\_Achtman\_mdh\_636 hit is an extreme outlier compared to all of the other Escherichia\_Shigella\_Achtman\_mdh\_*n*, i.e. they are all < 85% identical to it (over this segment).

Aside from this contig, the next highest sequence identity observed is 95.1% (k141\_5462, which is 970 b.p. long, versus Escherichia\_Shigella\_Achtman\_gyrB\_395, 460 b.p.); but the HSP length is only 41 b.p. Of all HSPs of length > 50 b.p., the highest identity is 88.2%, and for those > 100 b.p. it is 87.7% (length 179 b.p.). Therefore the only matches worth any consideration are those with the k141\_5046 contig described above.

#### MLST Search Results: Sample 6412

Contig ID: k141\_49447, length 357 bp.

The following are the MLST sequences, and their total lengths, which make a 100%-identical, 357-bp HSP with it:

* Enterococcus\_faecium\_purK\_92 492
* Enterococcus\_faecium\_purK\_88 492
* Enterococcus\_faecium\_purK\_86 492
* Enterococcus\_faecium\_purK\_73 492
* Enterococcus\_faecium\_purK\_70 492
* Enterococcus\_faecium\_purK\_6 492
* Enterococcus\_faecium\_purK\_57 492
* Enterococcus\_faecium\_purK\_53 492
* Enterococcus\_faecium\_purK\_35 492
* Enterococcus\_faecium\_purK\_3 492

A further 3 hits form a 356-bp HSP of 100.00% identity, i.e. effectively 99.72% identical over the full query length:

* Enterococcus\_faecium\_purK\_26 492
* Enterococcus\_faecium\_purK\_2 492
* Enterococcus\_faecium\_purK\_121 492

A further 68 match with an HSP of 356 or 357 bp, with identity of between 99.72% and 99.16%, with 41 more 94.96% and 93.84% identity. All matches to this contig are Enterococcus\_faecium\_purK\_*n* sequences.

Contig ID: k141\_11026, length 408 bp.

This is the only contig which makes any 100%-identical matches with any *E. coli* sequences. This exclusively matches *E. coli/Escherichia* recA sequences (all 500), which are all 510 bp long. These are perfect 'overhang' matches of 262-bp HSPs: the 3'-most 262 bp (147-408) of the query is a perfect match with bp 1-262 of the reference.

There are only two 100%-identical matches with any *Klebsiella* MLST sequences, and again these are overhangs, but they are relatively very short, both 98-bp HSPs involving the same contig k141\_61057 (444 bp long): the first 98 bp of the query matches the final 98 bp (353-450) of the reference sequences (Klebsiella\_gapA\_146 and Klebsiella\_gapA\_140, both 450 bp long). The same region of this contig makes numerous other matches, < 100% identical, with other *Klebsiella* gapA sequences, and also a shorter sub-sequence of that region matches *E. faecalis* sequences (< 80% identical).

There are no 100%-identical matches with any *E. faecalis* sequences. There are only 2 HSPs with a sequence identity ≥ 88%, which are both 99.25% identical: they are 267 b.p. HSPs between k141\_68769 (which is 337 bp long) and each of Enterococcus\_faecalis\_gyd\_33 and Enterococcus\_faecalis\_gyd\_11 (both 395 bp long). These represent near-exact matches, because they involve the first 267 bp of the query aligning with the last 267 bp (129-395) of the reference. 265/267 bp are identical (99.25%).

#### Conclusions from comparison with MLST database sequences

#### Sample 2672480

Taking the MLST reference sequences as "ground truth", it can be concluded that sample 2672480 contains DNA which (among those organisms represented in the database) originate from *Klebsiella or Escherichia coli.* This conclusion arises from the nature of a single contig in relation exclusively to *mdh* locus sequences.

Although the rationale is to not speculate as to the reliability of the MLST references, it has to be noted that *E. coli* is represented by a single *mdh* sequence which is an exact match to the query contig but differs by 15% identity to the next most similar *E. coli mdh* locus sequence in the database. In contrast, several *Klebsiella* reference *mdh* sequences make exact matches and there is a gradation of non-identical hits, all *Klebsiella*, from 1, 2, 3... b.p. different to the query.

The observation that the contig sequence in question begins exactly at the first coordinate of the *mdh* sequence suggests that the gene sequence may be significantly more common in the sample to the nucleotide sequence 5' to it, i.e. an identical gene sequence may occur in different locations in genomes of different strains perhaps. This would lead to the gene sequence itself having the greatest depth of coverage, resulting in the assembled contig beginning exactly at its 5' end.

#### Sample 6412

From the MLST locus comparisons, this sample clearly matches *Enterococcus faecium*.

### Comparing positive contig sequences with more general database sequences (in NCBI nt)

Since only the contigs matching MLST sequences were used as queries for these searches, these constitute an attempt to confirm the MLST matches and demonstrate that equally good matches do not occur in other taxa. The rationale that the MLST database only contains sequences of the priority taxa; finding matches to them does not in theory rule out that they may also match non-priority taxa (although this is unlikely).

#### General database (nt) Search Results: Sample 2672480

The results of BLASTN of the k141\_5046contig sequence versus the nt database are completely *consistent with* the k141\_5046 contig having originated from *K. pneumoniae*, with 191 database hits which are 100% identical over the full length of the query contig (444 b.p.)*.* The results are not consistent with the origin being *E. coli* or any other non-*Klebsiella* species. They are also not consistent with either *K. quasipneumoniae or K. variicola* (the other two *Klebsiella* species represented in the *Klebsiella* MLST sequences).

However, aside from the 191 *K. pneumoniae*, there is a single other database match which is 100% identical over the full length of the query contig, which is from *K. aerogenes* strain NCTC9652. Thus, the result is also *consistent with* the k141\_5046 contig having originated from this species.

There are a further 480 *K. pneumoniae* near-perfect matches (differ by 1 b.p., = 99.77% identical). The only other hit from any species of the same identity is from *K. aerogenes* strain NCTC9644.

There are 2 sequences, both *K. pneumoniae*, which align with a single b.p. insertion in what is otherwise a completely identical sequence to the query (444/445 identical b.p. = 99.78%).

A further 95 sequences, all *K. pneumoniae*, align with no indels but have 2 single-b.p. mismatches (identity = 442/444 = 99.55%). Another sequence, also from the same species, has effectively the same identity but the mismatching b.p. are at the end, resulting in a 100%-identical alignment over 442 b.p.

A further 10 sequences, all *K. pneumoniae*, have between 99.77% and 97.75% over varying lengths of the query (mostly 100% coverage).

Beyond those are various sequences mostly from *K. quasipneumoniae*, also from *K. pneumoniae* (also one from Enterobacteriaceae bacterium S05 and *Klebsiella* sp. LY, at 96.17% and 94.99% identity (usually covering the full 444 b.p. of the query).

These are followed by the best *K. variicola* sequences, at 94.82% identity over the full-length query, with others at 94.14% and 93.69% identity (there is also a single *K.* sp. P1CD1). A *K. quasivariicola* sequence, along with a *K.* sp. PO552, are at 93.69% over the full length.

The worst-matching *K. variicola* sequence is 93.02% identical over the full length. There are several *K. pneumoniae* sequences with slightly better identities.

Following the above *K. variicola* sequence, the next highest identity is 89.41% (over the full 444 b.p. length, i.e. 47 b.p. mismatches), which is the best-matching *K. aerogenes* aside from the NCTC9652 (identical) and NCTC9644 (1 b.p. different) sequences. Many other *K. aerogenes* follow, from 88.96% to 87.72% identity.

This does make the two NCTC strains appear to be extreme outliers among *K. aerogenes*. The suspicion is then that either these two strains are mis-identified, or else incorporate some contaminant sequence from *K. pneumoniae*, which matches.

Taken at face-value, however, the results indicate that the k141\_5046 contig from sample 2672480 originates either from one of numerous *K. pneumoniae* strains, or from *K. aerogenes* strain NCTC9652.

When considered with the MLST-sequence results (origin is one of *K. pneumoniae, K. quasipneumoniae, K. variicola* or *E. coli*), the strong suspicion is that the origin of this contig is *K. pneumoniae.* But strictly speaking, this cannot be deemed as confirmed. Regarding the sample as a whole, other contigs' results would need to be checked to see if any confirm the *K. pneumoniae* result.

#### General database (nt) Search Results: Sample 6412

The contig k141\_49447, length 357 bp (which matches the *purK* MLST locus).

The top 250 matches with nt are exclusively with *Enterococcus faecium* sequences:

* 56 sequences: 357/357 bp = 100% identity
* 25 sequences: 356/356 bp = 100% identical over 356 bp; the first bp of the query is absent from the HSP, i.e. this is effectively 356 / 357 bp = 99.72% identity
* 23 sequences: 356/357 bp = 99.72% identity
* 118 sequences: 355/356 bp (bp 1 is absent; effectively 355 / 357 = 99.44% identity)
* 1 sequence: 355/357 bp = 99.44% identity
* 22 sequences: 354/356 bp (bp 1 is absent; effectively 354 / 357 = 99.16% identity)
* 1 sequence: 339/357 bp = 94.96% identity
* 4 sequences: 338/357 bp = 94.68% identity

Thus, this sample is confirmed as *E. faecium* positive at the DNA sequence level. The whole of this contig makes an identical match, and the most similar any non-*E. faecium* sequence could be is 94.68% identical.

### Comparing all contigs with the nt database

In contrast to the use of only the putative positive contigs from the MLST locus sequence comparisons to confirm that non-concern taxa do not match equally well, the aim here is to use *all* of the other contigs as the queries of BLASTN searches versus the NCBI nt database.

Note that any of the species on the priority taxa list - not solely those which had thus far been determined as matches by MetaPhlAn3 or the MLST database search - are noted below. One of these species (*Yersinia enterolitica*) resulted as a positive, despite it being negative by the previous methods.

#### Sample 2672480: results of BLASTN versus nt (all contigs)

##### *Sample 2672480 Klebsiella pneumoniae*

Strictly speaking, the MLST locus results do not confirm a positive for *Klebsiella pneumoniae (*see above, with reference to the single *E. coli* match), albeit with caveats to the non-*K. pneumoniae* match. Therefore, this sample's contigs were checked for matches to *K. pneumoniae* in the general search.

There are examples of contigs which can clearly be assigned to *K. pneumoniae* as they are clearly distinguishable even from closely-related species (*K. variicola*) and especially from other genera.

k141\_14577 (420 bp):

* score 420, evalue 0, identity 420 / 420 = 100.000 % : 56 x Klebsiella pneumoniae
* score 411, evalue 0, identity 417 / 420 = 99.286 % : 4 x Klebsiella pneumoniae
* score 302, evalue 7e-155, identity 358 / 386 = 92.746 % : 1 x Klebsiella pneumoniae
* score 299, evalue 3e-153, identity 357 / 386 = 92.487 % : 2 x Klebsiella pneumoniae
* score 299, evalue 3e-153, identity 357 / 386 = 92.487 % : 3 x Klebsiella variicola \*
* score 296, evalue 2e-151, identity 356 / 386 = 92.228 % : 71 x Klebsiella pneumoniae
* score 296, evalue 2e-151, identity 356 / 386 = 92.228 % : 1 x Klebsiella variicola
* score 293, evalue 7e-150, identity 355 / 386 = 91.969 % : 5 x Klebsiella pneumoniae
* score 293, evalue 7e-150, identity 355 / 386 = 91.969 % : 1 x Klebsiella variicola
* score 289, evalue 1e-147, identity 351 / 382 = 91.885 % : 1 x Klebsiella quasivariicola
* score 289, evalue 1e-147, identity 351 / 382 = 91.885 % : 1 x Klebsiella sp.
* score 227, evalue 4e-113, identity 327 / 376 = 86.968 % : 1 x Escherichia coli
* score 212, evalue 8e-105, identity 321 / 375 = 85.600 % : 1 x Enterobacter hormaechei
* score 212, evalue 8e-105, identity 321 / 375 = 85.600 % : 18 x Escherichia coli
* score 212, evalue 8e-105, identity 321 / 375 = 85.600 % : 1 x Shigella flexneri
* score 209, evalue 4e-103, identity 320 / 375 = 85.333 % : 1 x Escherichia coli
* score 206, evalue 2e-101, identity 318 / 374 = 85.027 % : 21 x Escherichia coli
* score 206, evalue 2e-101, identity 321 / 377 = 85.146 % : 5 x Escherichia coli
* score 203, evalue 8e-100, identity 317 / 374 = 84.759 % : 26 x Escherichia coli
* score 203, evalue 8e-100, identity 320 / 377 = 84.881 % : 6 x Escherichia coli
* score 200, evalue 4e-98, identity 319 / 377 = 84.615 % : 24 x Escherichia coli

The best non-*K. pneumoniae* match is highlighted/marked with a \*. There are also other cases where the best hits are non-identical but very nearly identical to *K. pneumoniae* over a lengthy HSP, with the next best hits somewhat distant:

k141\_17662 (1,258 bp): top 2 hits are 1257/1258 = 99.92% identical: *K. pneumoniae* strain 47733 plasmid p47733\_COL\_PVC and *K. pneumoniae* strain DA48896 plasmid p48896\_7. Joint 3rd best hits are 1202/1259 = 95.47% identical: (i) *K. pneumoniae* strain C17KP0040 plasmid pC17KP0040-5, (ii) *Proteus mirabilis* strain T18 plasmid pT18, (iii) Enterobacter cloacae strain SZECL1 plasmid pSZECL\_a. Next best hit is 1176/1232 = 95.45% identity, 97.85% coverage, 'Uncultured prokaryote from Rat gut metagenome metamobilome, plasmid pRGRH0299'; next is 1077/1129 = 95.39% identity, 89.67% coverage, *K. pneumoniae* strain C16KP0164 plasmid pC16KP0164-4.

There are others where all of the hits (given the search parameters used) match or almost match *K. pneumoniae* exactly, but to rule out whether non-concern taxa could match equally as well and have dropped off the list, would require a repeat search with a higher hit limit.

k141\_4668 (849 bp): 100% match over full-length with all 250 best hits, all *K. pneumoniae*.

k141\_16616 (647 bp): 100% match over full-length with all 250 best hits, all *K. pneumoniae*.

k141\_13739 (523 bp): all 250 best hits are *K. pneumoniae*; 100% match over full-length with best 28 hits; next best 11 at 522/523 bp = 99.81% identity; next 211 at 251/253 b.p. = 99.21% identity.

k141\_559 (473 bp): all 250 best hits are *K. pneumoniae*; 100% match over full-length with best 11 hits; next best 239 at 472/473 bp = 99.79% identity.

k141\_5020 (444 bp): all 250 best hits are *K. pneumoniae*; 100% match over full-length with best 3 hits;

k141\_6348 (371 bp): all 250 best hits are *K. pneumoniae*; 100 match over-full-length with top hit only (strain RHBSTW-00843); remainder are all 370/371 bp = 99.73% identical.

The (possibly doubtful) *K. aerogenes* genomes appear again with some otherwise exclusively *K. pneumoniae* sequences:

k141\_13235 (483 bp): 249 of 250 best hits are *K. pneumoniae*; 100% match over full-length with best 3 hits; next best 3 at 482/483 bp = 99.79% identity; next 244 at 481/483 b.p. = 99.59% identity, which includes a single *K. aerogenes* NCTC9644 sequence.

##### Sample 2672480 *Escherichia coli*

There are no contigs which make any 100%-identical HSPs (of at least 200 bp) with *E. coli* sequences where there are not also equally good matches with other species. Some of these HSPs are internal. k141\_16648 and k141\_17662 (see above) both make 100%-identical HSPs; but for example, the latter makes a 100% identical 555-bp long ID but this is less than half the length of the HSP which the same contig makes with *K. pneumoniae* which has only a single bp mismatch.

##### Sample 2672480 *Enterococcus faecalis*

There are very few hits and there are none which pass the minimum criteria regarding % identity.

##### Sample 2672480 *Enterococcus faecium*

There are even fewer hits in *E. faecium*, and again none pass the minimum criteria.

#### Sample 6412

##### Sample 6412 *Klebsiella pneumoniae*

Although there are many cases of contig query sequences which form 100%-identical, ≥ 200 bp HSPs only with sequences annotated as *K. pneumoniae*, there are none where the difference with the best non-*K. pneumoniae* sequence is as large as in the 2672480 sample. However, the difference is sufficient to confirm the sequence is exclusively representing *K. pneumoniae*.

For this contig, the nearest non- *K. pneumoniae* match is a single case of a 4 bp different hit, whereas there are 31 instances of exact matches. The non- *K. pneumoniae* match is from one of the extreme-outlier *K. aerogenes* genomes (NCTC9644), with the other one (NCTC9652) differing by a further 2 bp. Indeed the next most similar non- *K. pneumoniae* match is 10 bp different (*K. variicola*, a close relative of *K. pneumoniae*) which once more provides a perspective on these two *K. aerogenes* genomes.

k141\_71065 (603 bp):

* score 1088, evalue 0, identity 603 / 603 = 100.000 % : 31 x Klebsiella pneumoniae
* score 1084, evalue 0, identity 602 / 603 = 99.834 % : 34 x Klebsiella pneumoniae
* score 1079, evalue 0, identity 601 / 603 = 99.668 % : 3 x Klebsiella pneumoniae
* score 1075, evalue 0, identity 600 / 603 = 99.502 % : 69 x Klebsiella pneumoniae
* score 1070, evalue 0, identity 599 / 603 = 99.337 % : 1 x Klebsiella aerogenes
* score 1070, evalue 0, identity 599 / 603 = 99.337 % : 84 x Klebsiella pneumoniae
* score 1067, evalue 0, identity 597 / 603 = 99.005 % : 1 x Klebsiella pneumoniae
* score 1067, evalue 0, identity 597 / 601 = 99.334 % : 1 x Klebsiella pneumoniae
* score 1066, evalue 0, identity 598 / 603 = 99.171 % : 137 x Klebsiella pneumoniae
* score 1062, evalue 0, identity 598 / 603 = 99.171 % : 1 x Klebsiella pneumoniae
* score 1061, evalue 0, identity 597 / 603 = 99.005 % : 1 x Klebsiella aerogenes
* score 1061, evalue 0, identity 597 / 603 = 99.005 % : 307 x Klebsiella pneumoniae
* score 1060, evalue 0, identity 599 / 604 = 99.172 % : 1 x Klebsiella pneumoniae
* score 1058, evalue 0, identity 599 / 605 = 99.008 % : 1 x Klebsiella pneumoniae
* score 1057, evalue 0, identity 596 / 603 = 98.839 % : 60 x Klebsiella pneumoniae
* score 1055, evalue 0, identity 597 / 604 = 98.841 % : 1 x Klebsiella pneumoniae
* score 1053, evalue 0, identity 596 / 603 = 98.839 % : 2 x Klebsiella pneumoniae
* score 1053, evalue 0, identity 594 / 601 = 98.835 % : 4 x Klebsiella pneumoniae
* score 1052, evalue 0, identity 595 / 603 = 98.673 % : 2 x Klebsiella pneumoniae
* score 1049, evalue 0, identity 597 / 605 = 98.678 % : 1 x Klebsiella pneumoniae
* score 1049, evalue 0, identity 593 / 601 = 98.669 % : 3 x Klebsiella pneumoniae
* score 1043, evalue 0, identity 593 / 603 = 98.342 % : 1 x Klebsiella sp.
* score 1043, evalue 0, identity 593 / 603 = 98.342 % : 1 x Klebsiella variicola
* score 1040, evalue 0, identity 591 / 601 = 98.336 % : 1 x Klebsiella pneumoniae
* score 1040, evalue 0, identity 591 / 601 = 98.336 % : 2 x Klebsiella variicola
* score 1039, evalue 0, identity 592 / 603 = 98.176 % : 10 x Klebsiella variicola
* score 1034, evalue 0, identity 591 / 603 = 98.010 % : 6 x Klebsiella variicola
* score 1031, evalue 0, identity 589 / 601 = 98.003 % : 1 x Klebsiella variicola
* score 1030, evalue 0, identity 590 / 603 = 97.844 % : 1 x Klebsiella variicola
* score 1026, evalue 0, identity 588 / 601 = 97.837 % : 1 x Klebsiella variicola
* score 1025, evalue 0, identity 589 / 603 = 97.678 % : 1 x Klebsiella variicola
* score 1021, evalue 0, identity 588 / 603 = 97.512 % : 1 x Klebsiella pneumoniae
* score 1012, evalue 0, identity 585 / 601 = 97.338 % : 5 x Klebsiella variicola
* score 1008, evalue 0, identity 584 / 601 = 97.171 % : 2 x Klebsiella variicola
* score 1004, evalue 0, identity 585 / 603 = 97.015 % : 1 x Klebsiella variicola
* score 958, evalue 0, identity 573 / 601 = 95.341 % : 1 x Klebsiella quasivariicola
* score 958, evalue 0, identity 573 / 601 = 95.341 % : 1 x Klebsiella sp.
* score 939, evalue 0, identity 570 / 603 = 94.527 % : 3 x Klebsiella quasipneumoniae
* score 935, evalue 0, identity 569 / 603 = 94.362 % : 1 x Enterobacteriaceae bacterium
* score 935, evalue 0, identity 569 / 603 = 94.362 % : 6 x Klebsiella quasipneumoniae
* score 935, evalue 0, identity 569 / 603 = 94.362 % : 1 x Klebsiella sp.
* score 930, evalue 0, identity 568 / 603 = 94.196 % : 1 x Klebsiella pneumoniae
* score 930, evalue 0, identity 568 / 603 = 94.196 % : 11 x Klebsiella quasipneumoniae
* score 926, evalue 0, identity 567 / 603 = 94.030 % : 1 x Klebsiella pneumoniae
* score 926, evalue 0, identity 567 / 603 = 94.030 % : 9 x Klebsiella quasipneumoniae
* score 921, evalue 0, identity 566 / 603 = 93.864 % : 6 x Klebsiella quasipneumoniae
* score 917, evalue 0, identity 565 / 603 = 93.698 % : 4 x Klebsiella quasipneumoniae
* score 785, evalue 0, identity 447 / 455 = 98.242 % : 1 x Klebsiella variicola
* score 559, evalue 1e-154, identity 481 / 601 = 80.033 % : 1 x Klebsiella grimontii
* ... and others of identity of ~ 80% or lower.

There is a similar situation with this contig:

k141\_88307 (349 bp):

* score 630, evalue 1e-176, identity 349 / 349 = 100.000 % : 11 x Klebsiella pneumoniae
* score 626, evalue 6e-175, identity 348 / 349 = 99.713 % : 2 x Klebsiella pneumoniae
* score 621, evalue 7e-174, identity 347 / 349 = 99.427 % : 3 x Klebsiella pneumoniae
* score 617, evalue 3e-172, identity 346 / 349 = 99.140 % : 1 x Klebsiella aerogenes
* score 617, evalue 3e-172, identity 346 / 349 = 99.140 % : 607 x Klebsiella pneumoniae
* score 617, evalue 3e-172, identity 346 / 349 = 99.140 % : 5 x Klebsiella quasipneumoniae
* score 612, evalue 4e-171, identity 345 / 349 = 98.854 % : 102 x Klebsiella pneumoniae
* score 610, evalue 1e-170, identity 346 / 350 = 98.857 % : 1 x Klebsiella aerogenes
* score 609, evalue 5e-170, identity 345 / 349 = 98.854 % : 3 x Klebsiella pneumoniae
* score 608, evalue 2e-169, identity 344 / 349 = 98.567 % : 9 x Klebsiella pneumoniae
* score 608, evalue 2e-169, identity 344 / 349 = 98.567 % : 1 x Klebsiella quasipneumoniae
* score 604, evalue 2e-168, identity 344 / 349 = 98.567 % : 2 x Klebsiella pneumoniae
* score 598, evalue 8e-167, identity 346 / 352 = 98.295 % : 1 x Klebsiella pneumoniae
* score 594, evalue 1e-165, identity 345 / 352 = 98.011 % : 1 x Klebsiella pneumoniae
* score 585, evalue 5e-163, identity 339 / 349 = 97.135 % : 1 x Klebsiella quasipneumoniae
* score 581, evalue 6e-162, identity 337 / 347 = 97.118 % : 4 x Klebsiella quasipneumoniae
* score 578, evalue 8e-161, identity 335 / 345 = 97.101 % : 3 x Klebsiella variicola
* score 573, evalue 3e-159, identity 336 / 347 = 96.830 % : 1 x Klebsiella pneumoniae
* score 573, evalue 3e-159, identity 336 / 347 = 96.830 % : 2 x Klebsiella quasipneumoniae
* score 573, evalue 3e-159, identity 334 / 345 = 96.812 % : 1 x Klebsiella sp.
* score 573, evalue 3e-159, identity 334 / 345 = 96.812 % : 21 x Klebsiella variicola
* score 572, evalue 3e-159, identity 335 / 347 = 96.542 % : 7 x Klebsiella quasipneumoniae
* etc

There are other contigs where a sequence from one of these two *K. aerogenes* genomes is even closer, such as only 1 bp different (and 0 bp different in some cases).

In conclusion: even if the two *K. aerogenes* genomes are completely correct, and for some reason contain multiple large tracts of genome sequence which are identical to some *K. pneumoniae* sequences and near-identical to many others, and indeed thus more similar to some *K. pneumoniae* genomes than various *K. pneumoniae* strains to each other; then it does not affect the result that there are multiple contig sequences in the 6412 sample which are distinguishable as exact matches to *K. pneumoniae*, not *K. aerogenes.*

##### Sample 6412 *Escherichia coli*

There are many cases of sequences making exact matches with *E. coli* but also making equally as good matches with other species, which are always the minority of matches - usually *Shigella*, and also *Salmonella.* Admittedly, these are also taxa of concern, and are closely related to *Escherichia* (some taxonomies treat *Escherichia* and *Shigella* as a collective genus). But ideally a species-level identification to confirm the *E. coli* presence would be observed (note also that the samples were negative for *Shigella* and *Salmonella* by MetaPhlAn3).

There are some clear exceptions, however; these put some distance between the best hit (*E. coli*) and the best non-*E. coli* hit:

k141\_93529 (393 bp):

* score 393, evalue 0, identity 393 / 393 = 100.000 % : 25 x Escherichia coli
* score 390, evalue 0, identity 392 / 393 = 99.746 % : 68 x Escherichia coli
* score 387, evalue 0, identity 391 / 393 = 99.491 % : 35 x Escherichia coli
* score 387, evalue 0, identity 392 / 394 = 99.492 % : 1 x Escherichia coli
* score 386, evalue 0, identity 391 / 393 = 99.491 % : 1 x Escherichia coli
* score 384, evalue 0, identity 390 / 393 = 99.237 % : 10 x Escherichia coli
* score 383, evalue 0, identity 390 / 393 = 99.237 % : 4 x Escherichia coli
* score 380, evalue 0, identity 389 / 393 = 98.982 % : 1 x Escherichia coli
* score 378, evalue 0, identity 389 / 394 = 98.731 % : 1 x Escherichia coli
* score 376, evalue 0, identity 391 / 397 = 98.489 % : 1 x Escherichia coli
* score 375, evalue 0, identity 390 / 396 = 98.485 % : 1 x Escherichia coli
* score 375, evalue 0, identity 387 / 393 = 98.473 % : 6 x Escherichia coli
* score 372, evalue 0, identity 386 / 393 = 98.219 % : 1 x Escherichia coli
* score 371, evalue 0, identity 386 / 393 = 98.219 % : 1 x Escherichia coli
* score 357, evalue 0, identity 381 / 393 = 96.947 % : 5 x Escherichia coli
* score 354, evalue 0, identity 380 / 393 = 96.692 % : 1 x Escherichia coli
* score 345, evalue 8e-179, identity 371 / 384 = 96.615 % : 6 x Escherichia coli
* score 345, evalue 8e-179, identity 373 / 387 = 96.382 % : 1 x Escherichia coli
* score 345, evalue 8e-179, identity 373 / 387 = 96.382 % : 38 x Shigella sonnei
* score 343, evalue 1e-177, identity 371 / 385 = 96.364 % : 30 x Escherichia coli
* score 343, evalue 1e-177, identity 374 / 389 = 96.144 % : 2 x Escherichia coli
* score 342, evalue 4e-177, identity 372 / 387 = 96.124 % : 11 x Escherichia coli

k141\_11061 (302 bp):

* score 302, evalue 5e-155, identity 302 / 302 = 100.000 % : 20 x Escherichia coli
* score 299, evalue 2e-153, identity 301 / 302 = 99.669 % : 14 x Escherichia coli
* score 299, evalue 2e-153, identity 299 / 299 = 100.000 % : 1 x Escherichia coli
* score 296, evalue 1e-151, identity 300 / 302 = 99.338 % : 4 x Escherichia coli
* score 293, evalue 5e-150, identity 299 / 302 = 99.007 % : 13 x Escherichia coli
* score 290, evalue 2e-148, identity 298 / 302 = 98.675 % : 3 x Escherichia coli
* score 287, evalue 1e-146, identity 297 / 302 = 98.344 % : 1 x Escherichia coli
* score 284, evalue 5e-145, identity 296 / 302 = 98.013 % : 1 x Escherichia coli
* score 272, evalue 2e-138, identity 292 / 302 = 96.689 % : 5 x Escherichia coli
* score 266, evalue 5e-135, identity 284 / 293 = 96.928 % : 1 x Escherichia coli
* score 258, evalue 1e-130, identity 278 / 288 = 96.528 % : 1 x Escherichia coli
* score 224, evalue 1e-111, identity 272 / 296 = 91.892 % : 1 x Escherichia coli
* score 181, evalue 9e-88, identity 239 / 268 = 89.179 % : 1 x Escherichia coli
* score 178, evalue 4e-86, identity 238 / 268 = 88.806 % : 1 x Escherichia coli
* score 177, evalue 2e-85, identity 253 / 291 = 86.942 % : 3 x Escherichia coli
* score 166, evalue 2e-79, identity 228 / 259 = 88.031 % : 1 x Escherichia coli
* score 163, evalue 9e-78, identity 227 / 259 = 87.645 % : 1 x Escherichia coli
* score 143, evalue 1e-66, identity 219 / 257 = 85.214 % : 1 x Escherichia coli
* score 143, evalue 1e-66, identity 203 / 233 = 87.124 % : 1 x Escherichia coli
* score 119, evalue 3e-53, identity 231 / 285 = 81.053 % : 2 x Escherichia albertii
* score 119, evalue 3e-53, identity 227 / 281 = 80.783 % : 7 x Escherichia coli
* score 116, evalue 1e-51, identity 202 / 243 = 83.128 % : 1 x Escherichia coli
* score 116, evalue 1e-51, identity 230 / 285 = 80.702 % : 1 x Escherichia coli
* score 116, evalue 1e-51, identity 226 / 281 = 80.427 % : 5 x Escherichia coli
* score 113, evalue 6e-50, identity 229 / 285 = 80.351 % : 1 x Escherichia albertii
* score 113, evalue 6e-50, identity 231 / 287 = 80.488 % : 10 x Escherichia coli
* score 113, evalue 6e-50, identity 225 / 281 = 80.071 % : 2 x Escherichia marmotae
* score 110, evalue 3e-48, identity 174 / 206 = 84.466 % : 1 x Escherichia coli
* score 109, evalue 1e-47, identity 201 / 244 = 82.377 % : 1 x Escherichia coli
* score 102, evalue 8e-44, identity 226 / 286 = 79.021 % : 1 x Escherichia marmotae

Also worth mentioning is the 5,209 bp contig k141\_9652, the last 5,077 bp of which make a near-exact match (2 bp differ) with the first 5,077 bp of an *E. coli* sequence (FAM21845 plasmid pFAM21845\_3) (i.e. it is not an internal HSP), with the next best hits much worse:

* score 5071, evalue 0, identity 5075 / 5077 = 99.961 % : 1 x Escherichia coli
* score 4470, evalue 0, identity 4776 / 4926 = 96.955 % : 1 x Escherichia fergusonii
* score 3589, evalue 0, identity 3729 / 3798 = 98.183 % : 1 x Klebsiella pneumoniae
* score 3261, evalue 0, identity 3435 / 3520 = 97.585 % : 1 x Escherichia coli
* score 3189, evalue 0, identity 3351 / 3430 = 97.697 % : 1 x Salmonella enterica
* score 3018, evalue 0, identity 3149 / 3213 = 98.008 % : 1 x Enterobacter cloaceae
* score 2748, evalue 0, identity 3058 / 3209 = 95.294 % : 1 x Escherichia coli
* score 2545, evalue 0, identity 2567 / 2578 = 99.573 % : 1 x Plasmid P4
* score 2543, evalue 0, identity 2590 / 2613 = 99.120 % : 1 x Escherichia fergusonii
* score 2441, evalue 0, identity 2503 / 2532 = 98.855 % : 1 x Salmonella enteritidis
* etc

For illustrative purposes, the following case is much more dubious because it is an internal HSP; the last 70 bp of the query are absent from the HSP, even though the hit sequences always extend beyond the end of the HSP:

k141\_81985 (489 bp):

* score 419, evalue 0, identity 419 / 419 = 100.000 % : 6 x Escherichia coli
* score 416, evalue 0, identity 418 / 419 = 99.761 % : 1 x Escherichia coli
* score 371, evalue 0, identity 404 / 420 = 96.190 % : 1 x Erwinia amylovora
* score 371, evalue 0, identity 404 / 420 = 96.190 % : 1 x Escherichia coli
* score 368, evalue 0, identity 403 / 420 = 95.952 % : 1 x Klebsiella pneumoniae

##### Sample 6412 *Enterococcus faecalis*

There are several contig queries which are clearly identifiable as exact matches over their full length with exclusively *E. faecalis* sequences. The best non-*E. faecalis* matches are highlighted/marked with a \*.

k141\_82730 (326 bp):

* score 326, evalue 3e-168, identity 326 / 326 = 100.000 % : 1 x Enterococcus faecalis
* score 102, evalue 8e-44, identity 178 / 216 = 82.407 % : 1 x Streptococcus milleri \*
* score 99, evalue 4e-42, identity 177 / 216 = 81.944 % : 1 x Streptococcus anginosus
* score 93, evalue 8e-39, identity 189 / 237 = 79.747 % : 1 x Streptococcus parasanguinis
* score 90, evalue 4e-37, identity 188 / 237 = 79.325 % : 1 x Streptococcus parasanguinis

k141\_85385 (333 bp):

* score 333, evalue 3e-172, identity 333 / 333 = 100.000 % : 1 x Enterococcus faecalis
* score 288, evalue 3e-147, identity 316 / 330 = 95.758 % : 6 x Enterococcus faecalis
* score 237, evalue 8e-119, identity 288 / 313 = 92.013 % : 1 x Enterococcus phage
* score 204, evalue 2e-100, identity 278 / 315 = 88.254 % : 1 x Enterococcus faecalis
* score 188, evalue 1e-91, identity 264 / 302 = 87.417 % : 2 x Enterococcus phage
* score 125, evalue 1e-56, identity 125 / 125 = 100.000 % : 1 x Enterococcus faecalis

k141\_34417 (322 bp):

* score 322, evalue 4e-166, identity 322 / 322 = 100.000 % : 3 x Enterococcus faecalis
* score 319, evalue 2e-164, identity 322 / 323 = 99.690 % : 3 x Enterococcus faecalis
* score 319, evalue 2e-164, identity 321 / 322 = 99.689 % : 1 x Enterococcus faecalis
* score 317, evalue 3e-163, identity 322 / 324 = 99.383 % : 5 x Enterococcus faecalis
* score 284, evalue 6e-145, identity 321 / 336 = 95.536 % : 3 x Enterococcus faecalis
* score 276, evalue 2e-140, identity 320 / 338 = 94.675 % : 2 x Enterococcus faecium
* score 273, evalue 7e-139, identity 319 / 338 = 94.379 % : 2 x Enterococcus faecium
* score 181, evalue 1e-87, identity 225 / 243 = 92.593 % : 1 x Enterococcus faecium
* score 180, evalue 4e-87, identity 214 / 228 = 93.860 % : 4 x Enterococcus faecium
* score 179, evalue 1e-86, identity 211 / 224 = 94.196 % : 1 x Enterococcus durans
* score 179, evalue 1e-86, identity 211 / 224 = 94.196 % : 1 x Enterococcus faecium
* score 178, evalue 5e-86, identity 215 / 230 = 93.478 % : 1 x Enterococcus faecium
* score 177, evalue 2e-85, identity 212 / 226 = 93.805 % : 6 x Enterococcus faecium
* score 176, evalue 6e-85, identity 211 / 225 = 93.778 % : 2 x Enterococcus durans
* score 176, evalue 6e-85, identity 211 / 225 = 93.778 % : 11 x Enterococcus faecium
* score 176, evalue 6e-85, identity 211 / 225 = 93.778 % : 1 x Lactobacillus sp.
* score 175, evalue 2e-84, identity 210 / 224 = 93.750 % : 2 x Enterococcus faecium
* score 174, evalue 8e-84, identity 211 / 226 = 93.363 % : 3 x Enterococcus faecium
* score 173, evalue 3e-83, identity 210 / 225 = 93.333 % : 3 x Enterococcus faecium
* score 172, evalue 1e-82, identity 209 / 224 = 93.304 % : 1 x Enterococcus faecium
* score 171, evalue 4e-82, identity 210 / 226 = 92.920 % : 1 x Enterococcus faecium
* score 171, evalue 4e-82, identity 212 / 229 = 92.576 % : 5 x Enterococcus faecium
* score 170, evalue 1e-81, identity 209 / 225 = 92.889 % : 1 x Enterococcus faecium
* score 110, evalue 3e-48, identity 113 / 114 = 99.123 % : 1 x Enterococcus faecium
* score 110, evalue 3e-48, identity 113 / 114 = 99.123 % : 1 x Enterococcus sp.
* score 100, evalue 1e-42, identity 109 / 113 = 96.460 % : 1 x Enterococcus gallinarum
* score 100, evalue 1e-42, identity 109 / 113 = 96.460 % : 1 x Enterococcus sp.
* score 94, evalue 2e-39, identity 107 / 113 = 94.690 % : 1 x Enterococcus sp.
* score 53, evalue 1e-16, identity 53 / 53 = 100.000 % : 1 x Enterococcus durans
* score 53, evalue 1e-16, identity 53 / 53 = 100.000 % : 8 x Enterococcus faecium
* score 47, evalue 3e-13, identity 51 / 53 = 96.226 % : 1 x Enterococcus faecium

k141\_71843 (444 bp):

* score 444, evalue 0, identity 444 / 444 = 100.000 % : 1 x Enterococcus faecalis
* score 412, evalue 0, identity 432 / 442 = 97.738 % : 1 x Enterococcus faecalis
* score 406, evalue 0, identity 430 / 442 = 97.285 % : 1 x Enterococcus faecalis
* score 403, evalue 0, identity 429 / 442 = 97.059 % : 2 x Bacillus subtilis
* score 403, evalue 0, identity 429 / 442 = 97.059 % : 1 x Enterococcus faecalis
* score 403, evalue 0, identity 429 / 442 = 97.059 % : 2 x Enterococcus faecium
* score 403, evalue 0, identity 429 / 442 = 97.059 % : 1 x Enterococcus thailandicus
* score 403, evalue 0, identity 429 / 442 = 97.059 % : 1 x Staphylococcus aureus
* score 403, evalue 0, identity 429 / 442 = 97.059 % : 1 x Streptococcus pyogenes
* score 400, evalue 0, identity 428 / 442 = 96.833 % : 7 x Enterococcus faecalis
* score 400, evalue 0, identity 428 / 442 = 96.833 % : 2 x Enterococcus sp.
* score 400, evalue 0, identity 429 / 443 = 96.840 % : 1 x Enterococcus thailandicus
* score 400, evalue 0, identity 429 / 443 = 96.840 % : 1 x Vagococcus sp.
* score 397, evalue 0, identity 427 / 442 = 96.606 % : 1 x Enterococcus faecalis
* score 397, evalue 0, identity 427 / 442 = 96.606 % : 1 x Enterococcus gilvus
* score 385, evalue 0, identity 423 / 442 = 95.701 % : 1 x Enterococcus faecalis
* score 370, evalue 0, identity 418 / 442 = 94.570 % : 1 x Enterococcus casseliflavus
* score 352, evalue 0, identity 413 / 443 = 93.228 % : 1 x Enterococcus casseliflavus
* score 264, evalue 1e-133, identity 284 / 294 = 96.599 % : 4 x Streptococcus agalactiae
* score 155, evalue 4e-73, identity 163 / 167 = 97.605 % : 1 x Arcanobacterium pyogenes
* score 155, evalue 4e-73, identity 163 / 167 = 97.605 % : 1 x Bacteroides fragilis
* score 155, evalue 4e-73, identity 163 / 167 = 97.605 % : 1 x Bacteroides uniformis
* score 155, evalue 4e-73, identity 163 / 167 = 97.605 % : 1 x Butyricimonas faecalis
* score 155, evalue 4e-73, identity 163 / 167 = 97.605 % : 1 x Campylobacter coli
* score 155, evalue 4e-73, identity 163 / 167 = 97.605 % : 1 x Clostridium perfringens

##### Sample 6412 *Enterococcus faecium*

This has already been confirmed from the contig matching the *purK* MLST locus alone, and no further analysis is necessary. See 'Comparing positive contig sequences with more general database sequences (in NCBI nt)'.

In the interests of demonstrating that there are multiple contig sequences in this sample which are resolvable to *E. faecium,* i.e. not just the MLST-matching contig, a few are highlighted/marked with a \* here:

k141\_71870 (377 bp):

* score 377, evalue 0, identity 377 / 377 = 100.000 % : 4 x Enterococcus faecium
* score 367, evalue 0, identity 374 / 377 = 99.204 % : 1 x Enterococcus faecium

k141\_82214 (332 bp):

* score 332, evalue 1e-171, identity 332 / 332 = 100.000 % : 1 x Enterococcus faecium
* score 314, evalue 1e-161, identity 326 / 332 = 98.193 % : 79 x Enterococcus faecium
* score 311, evalue 6e-160, identity 325 / 332 = 97.892 % : 3 x Enterococcus faecium
* score 308, evalue 3e-158, identity 324 / 332 = 97.590 % : 4 x Enterococcus faecium
* score 104, evalue 7e-45, identity 146 / 166 = 87.952 % : 45 x Enterococcus faecium
* score 104, evalue 7e-45, identity 144 / 163 = 88.344 % : 8 x Enterococcus faecium
* score 101, evalue 3e-43, identity 145 / 166 = 87.349 % : 11 x Enterococcus faecium
* score 101, evalue 3e-43, identity 143 / 163 = 87.730 % : 4 x Enterococcus faecium
* score 100, evalue 1e-42, identity 145 / 166 = 87.349 % : 1 x Enterococcus faecium
* score 98, evalue 1e-41, identity 144 / 166 = 86.747 % : 6 x Enterococcus faecium
* score 98, evalue 1e-41, identity 144 / 166 = 86.747 % : 1 x Enterococcus sp. \*

k141\_19663 (313 bp):

* score 313, evalue 4e-161, identity 313 / 313 = 100.000 % : 6 x Enterococcus faecium
* score 310, evalue 2e-159, identity 312 / 313 = 99.681 % : 56 x Enterococcus faecium
* score 307, evalue 9e-158, identity 311 / 313 = 99.361 % : 77 x Enterococcus faecium
* score 304, evalue 4e-156, identity 310 / 313 = 99.042 % : 1 x Enterococcus faecium
* score 192, evalue 7e-94, identity 272 / 312 = 87.179 % : 17 x Enterococcus faecium
* score 192, evalue 7e-94, identity 272 / 312 = 87.179 % : 1 x Enterococcus sp. \*
* score 188, evalue 1e-91, identity 268 / 308 = 87.013 % : 1 x Enterococcus faecium

k141\_54714 (536 bp):

* score 536, evalue 0, identity 536 / 536 = 100.000 % : 64 x Enterococcus faecium
* score 533, evalue 0, identity 535 / 536 = 99.813 % : 34 x Enterococcus faecium
* score 530, evalue 0, identity 534 / 536 = 99.627 % : 3 x Enterococcus faecium
* score 478, evalue 0, identity 517 / 536 = 96.455 % : 4 x Enterococcus faecium
* score 477, evalue 0, identity 517 / 536 = 96.455 % : 1 x Enterococcus faecium
* score 391, evalue 0, identity 393 / 394 = 99.746 % : 1 x Enterococcus faecium
* score 386, evalue 0, identity 388 / 389 = 99.743 % : 1 x Enterococcus faecium
* score 366, evalue 0, identity 368 / 369 = 99.729 % : 3 x Enterococcus faecium
* score 366, evalue 0, identity 366 / 366 = 100.000 % : 14 x Enterococcus faecium
* score 362, evalue 0, identity 367 / 369 = 99.458 % : 1 x Enterococcus faecium
* score 336, evalue 1e-173, identity 474 / 540 = 87.778 % : 1 x Enterococcus faecium
* score 334, evalue 2e-172, identity 474 / 540 = 87.778 % : 4 x Enterococcus faecium
* score 334, evalue 2e-172, identity 473 / 540 = 87.593 % : 11 x Enterococcus faecium
* score 331, evalue 7e-171, identity 473 / 541 = 87.431 % : 2 x Enterococcus faecium
* score 330, evalue 3e-170, identity 472 / 540 = 87.407 % : 2 x Enterococcus faecium
* score 328, evalue 3e-169, identity 472 / 540 = 87.407 % : 1 x Enterococcus sp. \*
* score 310, evalue 3e-159, identity 310 / 310 = 100.000 % : 2 x Enterococcus faecium
* score 246, evalue 1e-123, identity 285 / 303 = 94.059 % : 4 x Enterococcus faecium
* score 243, evalue 6e-122, identity 284 / 303 = 93.729 % : 9 x Enterococcus faecium
* score 86, evalue 1e-34, identity 99 / 105 = 94.286 % : 2 x Enterococcus hirae \*
* score 83, evalue 5e-33, identity 98 / 105 = 93.333 % : 1 x Enterococcus faecium
* score 77, evalue 1e-29, identity 85 / 89 = 95.506 % : 3 x Enterococcus durans
* score 76, evalue 4e-29, identity 88 / 94 = 93.617 % : 2 x Enterococcus hirae
* score 63, evalue 7e-22, identity 112 / 134 = 83.582 % : 4 x Enterococcus durans
* score 63, evalue 7e-22, identity 112 / 134 = 83.582 % : 1 x Lactobacillus sp.

k141\_45263 (400 bp):

* score 400, evalue 0, identity 400 / 400 = 100.000 % : 1 x Enterococcus faecium
* score 364, evalue 0, identity 388 / 400 = 97.000 % : 4 x Enterococcus faecium
* score 352, evalue 0, identity 384 / 400 = 96.000 % : 1 x Enterococcus faecium
* score 349, evalue 0, identity 383 / 400 = 95.750 % : 3 x Enterococcus faecium
* score 267, evalue 2e-135, identity 353 / 396 = 89.141 % : 9 x Enterococcus faecium
* score 267, evalue 2e-135, identity 351 / 393 = 89.313 % : 6 x Enterococcus faecium
* score 264, evalue 9e-134, identity 352 / 396 = 88.889 % : 5 x Enterococcus faecium
* score 261, evalue 4e-132, identity 351 / 396 = 88.636 % : 1 x Enterococcus faecium
* score 258, evalue 2e-130, identity 350 / 396 = 88.384 % : 102 x Enterococcus faecium
* score 258, evalue 2e-130, identity 348 / 393 = 88.550 % : 25 x Enterococcus faecium
* score 255, evalue 9e-129, identity 347 / 393 = 88.295 % : 1 x Enterococcus faecium
* score 255, evalue 9e-129, identity 349 / 396 = 88.131 % : 1 x Enterococcus faecium
* score 253, evalue 1e-127, identity 350 / 398 = 87.940 % : 1 x Enterococcus faecium
* score 249, evalue 2e-125, identity 347 / 396 = 87.626 % : 1 x Enterococcus faecium
* score 246, evalue 9e-124, identity 344 / 393 = 87.532 % : 1 x Enterococcus sp. \*
* score 125, evalue 2e-56, identity 135 / 140 = 96.429 % : 1 x Enterococcus faecium

k141\_86966 (329 bp):

* score 329, evalue 5e-170, identity 329 / 329 = 100.000 % : 9 x Enterococcus faecium
* score 326, evalue 3e-168, identity 328 / 329 = 99.696 % : 1 x Enterococcus faecium
* score 323, evalue 1e-166, identity 327 / 329 = 99.392 % : 83 x Enterococcus faecium
* score 322, evalue 4e-166, identity 327 / 329 = 99.392 % : 2 x Enterococcus faecium
* score 319, evalue 2e-164, identity 326 / 329 = 99.088 % : 1 x Enterococcus faecium
* score 315, evalue 3e-162, identity 319 / 321 = 99.377 % : 45 x Enterococcus faecium
* score 311, evalue 6e-160, identity 318 / 321 = 99.065 % : 1 x Enterococcus faecium
* score 209, evalue 3e-103, identity 294 / 334 = 88.024 % : 1 x Enterococcus faecium
* score 206, evalue 1e-101, identity 293 / 334 = 87.725 % : 4 x Enterococcus faecium
* score 206, evalue 1e-101, identity 293 / 334 = 87.725 % : 1 x Enterococcus sp.
* score 205, evalue 5e-101, identity 293 / 334 = 87.725 % : 7 x Enterococcus faecium
* score 203, evalue 6e-100, identity 293 / 335 = 87.463 % : 2 x Enterococcus faecium
* score 197, evalue 1e-96, identity 290 / 334 = 86.826 % : 5 x Enterococcus faecium
* score 172, evalue 1e-82, identity 227 / 253 = 89.723 % : 1 x Enterococcus faecium
* score 62, evalue 1e-21, identity 85 / 96 = 88.542 % : 7 x Enterococcus durans
* score 62, evalue 1e-21, identity 85 / 96 = 88.542 % : 1 x Lactobacillus sp.

##### Sample 6412 *Yersinia enterocolitica*

Below are detailed instances of contigs which fulfil these criteria:

* make a 100% match (non-internal HSP) with at least one sequence annotated as *Yersinia* *enterocolitica*
* make no 100% matches with any other species.

There were four such contigs.

The results for the first three of these are unconvincing, because they involve 1 or 2 *Y. enterolitica*-annotated sequences as the best matches, with all of the subsequent, highly similar hits being from various Enterobacteriaceae genera, not *Yersinia.* In all three cases, the next closest *Yersinia* matches are considerably different and further down the hit list (in two cases not high enough to show).

The conclusion is that the 100%-identical *Yersinia* annotations in those three cases are incorrect.

The fourth case (k141\_93443 ; 309 bp) is markedly different, with only *Yersinia* appearing in the hit list as far down as 82% identity. There are three 100%-identical *Y. enterolitica* matches, 17 others differing by 2, 3 or 4 b.p. and more by 14 bp (95% identity). All of these matvhes occur over the full length of the contig. No more *Y. enterolitica* appear in the list, with the next best match being 90% identical (*Y. canariae*). The weight of evidence is thus strong that this contig sequence is uniquely associated with *Y. enterolitica*.

k141\_13472 (504 bp):

* score 504, evalue 0, identity 504 / 504 = 100.000 % : 1 x Yersinia enterocolitica
* score 483, evalue 0, identity 497 / 504 = 98.611 % : 1 x Salmonella enterica
* score 476, evalue 0, identity 494 / 503 = 98.211 % : 1 x Salmonella enterica
* score 461, evalue 0, identity 489 / 503 = 97.217 % : 3 x Salmonella enterica
* score 459, evalue 0, identity 489 / 504 = 97.024 % : 1 x Salmonella enterica
* score 455, evalue 0, identity 487 / 503 = 96.819 % : 8 x Salmonella enterica
* score 448, evalue 0, identity 485 / 503 = 96.421 % : 1 x Salmonella enterica
* score 431, evalue 0, identity 477 / 500 = 95.400 % : 1 x Serratia marcescens
* score 431, evalue 0, identity 477 / 500 = 95.400 % : 1 x Serratia nematodiphila
* score 414, evalue 0, identity 474 / 504 = 94.048 % : 2 x Salmonella enterica
* score 336, evalue 1e-173, identity 450 / 506 = 88.933 % : 1 x Citrobacter freundii
* score 321, evalue 2e-165, identity 447 / 508 = 87.992 % : 1 x Citrobacter freundii
* score 315, evalue 5e-162, identity 443 / 506 = 87.549 % : 1 x Klebsiella pneumoniae
* score 312, evalue 2e-160, identity 442 / 506 = 87.352 % : 19 x Escherichia coli
* score 312, evalue 2e-160, identity 441 / 505 = 87.327 % : 3 x Escherichia coli
* score 312, evalue 2e-160, identity 442 / 506 = 87.352 % : 1 x Klebsiella variicola
* score 311, evalue 9e-160, identity 442 / 506 = 87.352 % : 1 x Serratia marcescens
* score 309, evalue 1e-158, identity 441 / 506 = 87.154 % : 1 x Citrobacter freundii
* score 309, evalue 1e-158, identity 441 / 506 = 87.154 % : 15 x Escherichia coli
* score 309, evalue 1e-158, identity 440 / 505 = 87.129 % : 1 x Escherichia coli
* score 306, evalue 5e-157, identity 440 / 506 = 86.957 % : 3 x Citrobacter freundii
* score 306, evalue 5e-157, identity 440 / 506 = 86.957 % : 1 x Citrobacter sp.
* score 306, evalue 5e-157, identity 440 / 506 = 86.957 % : 58 x Escherichia coli
* score 306, evalue 5e-157, identity 439 / 505 = 86.931 % : 16 x Escherichia coli
* score 306, evalue 5e-157, identity 440 / 506 = 86.957 % : 1 x Klebsiella variicola
* score 306, evalue 5e-157, identity 440 / 506 = 86.957 % : 5 x Salmonella enterica
* score 306, evalue 5e-157, identity 439 / 505 = 86.931 % : 1 x Salmonella enterica
* score 306, evalue 5e-157, identity 440 / 506 = 86.957 % : 2 x Salmonella sp.
* score 306, evalue 5e-157, identity 439 / 505 = 86.931 % : 1 x Shigella dysenteriae
* score 303, evalue 2e-155, identity 439 / 506 = 86.759 % : 2 x Citrobacter freundii
* score 303, evalue 2e-155, identity 439 / 506 = 86.759 % : 1 x Citrobacter sp.
* score 303, evalue 2e-155, identity 438 / 505 = 86.733 % : 1 x Citrobacter sp.
* score 303, evalue 2e-155, identity 439 / 506 = 86.759 % : 6 x Escherichia coli
* score 303, evalue 2e-155, identity 439 / 506 = 86.759 % : 1 x Klebsiella michiganensis
* score 302, evalue 9e-155, identity 442 / 510 = 86.667 % : 1 x Serratia marcescens
* score 300, evalue 1e-153, identity 437 / 505 = 86.535 % : 1 x Escherichia coli
* score 300, evalue 1e-153, identity 438 / 506 = 86.561 % : 1 x Pectobacterium carotovorum
* score 299, evalue 4e-153, identity 435 / 502 = 86.653 % : 1 x Citrobacter braakii
* score 299, evalue 4e-153, identity 435 / 502 = 86.653 % : 1 x Klebsiella pneumoniae
* score 298, evalue 1e-152, identity 437 / 505 = 86.535 % : 1 x Escherichia coli
* score 297, evalue 5e-152, identity 437 / 506 = 86.364 % : 1 x Citrobacter freundii
* score 296, evalue 2e-151, identity 434 / 502 = 86.454 % : 3 x Citrobacter sp.
* score 296, evalue 2e-151, identity 434 / 502 = 86.454 % : 1 x Klebsiella quasipneumoniae
* score 294, evalue 2e-150, identity 436 / 506 = 86.166 % : 1 x Citrobacter freundii
* score 294, evalue 2e-150, identity 436 / 506 = 86.166 % : 1 x Enterobacter kobei
* score 294, evalue 2e-150, identity 436 / 506 = 86.166 % : 12 x Klebsiella pneumoniae
* score 291, evalue 1e-148, identity 436 / 507 = 85.996 % : 1 x Citrobacter freundii
* score 291, evalue 1e-148, identity 436 / 507 = 85.996 % : 1 x Escherichia coli
* score 288, evalue 5e-147, identity 434 / 506 = 85.771 % : 1 x Citrobacter sp.
* score 285, evalue 3e-145, identity 434 / 507 = 85.602 % : 1 x Citrobacter sp.
* score 285, evalue 3e-145, identity 434 / 507 = 85.602 % : 1 x Escherichia coli
* score 282, evalue 1e-143, identity 433 / 507 = 85.404 % : 1 x Serratia marcescens
* score 267, evalue 3e-135, identity 426 / 505 = 84.356 % : 5 x Shigella boydii
* score 267, evalue 3e-135, identity 426 / 505 = 84.356 % : 9 x Shigella dysenteriae
* score 267, evalue 3e-135, identity 426 / 505 = 84.356 % : 1 x Shigella flexneri
* score 264, evalue 1e-133, identity 425 / 505 = 84.158 % : 8 x Shigella boydii
* score 264, evalue 1e-133, identity 425 / 505 = 84.158 % : 2 x Shigella dysenteriae
* score 261, evalue 5e-132, identity 424 / 505 = 83.960 % : 1 x Shigella boydii

k141\_83377 (328 bp):

* score 328, evalue 2e-169, identity 328 / 328 = 100.000 % : 2 x Yersinia enterocolitica
* score 324, evalue 3e-167, identity 327 / 328 = 99.695 % : 1 x Enterobacter hormaechei
* score 312, evalue 2e-160, identity 322 / 327 = 98.471 % : 1 x Enterobacteriaceae sp.
* score 312, evalue 2e-160, identity 322 / 327 = 98.471 % : 157 x Escherichia coli
* score 312, evalue 2e-160, identity 322 / 327 = 98.471 % : 1 x Proteus mirabilis
* score 312, evalue 2e-160, identity 322 / 327 = 98.471 % : 3 x Salmonella enterica
* score 312, evalue 2e-160, identity 322 / 327 = 98.471 % : 3 x Shigella boydii
* score 309, evalue 7e-159, identity 321 / 327 = 98.165 % : 22 x Escherichia coli
* score 309, evalue 7e-159, identity 321 / 327 = 98.165 % : 1 x Shigella boydii
* score 308, evalue 3e-158, identity 321 / 327 = 98.165 % : 1 x Escherichia coli
* score 306, evalue 3e-157, identity 320 / 327 = 97.859 % : 4 x Escherichia coli
* score 306, evalue 3e-157, identity 320 / 327 = 97.859 % : 1 x Shigella boydii
* score 306, evalue 3e-157, identity 320 / 327 = 97.859 % : 1 x Shigella dysenteriae
* score 305, evalue 1e-156, identity 319 / 326 = 97.853 % : 1 x Escherichia coli
* score 303, evalue 2e-155, identity 319 / 327 = 97.554 % : 13 x Escherichia coli
* score 303, evalue 2e-155, identity 317 / 324 = 97.840 % : 2 x Escherichia coli
* score 303, evalue 2e-155, identity 319 / 327 = 97.554 % : 1 x Shigella boydii
* score 303, evalue 2e-155, identity 319 / 327 = 97.554 % : 4 x Shigella flexneri
* score 301, evalue 2e-154, identity 319 / 327 = 97.554 % : 1 x Escherichia coli
* score 300, evalue 7e-154, identity 318 / 327 = 97.248 % : 5 x Escherichia coli
* score 299, evalue 3e-153, identity 318 / 327 = 97.248 % : 1 x Escherichia coli
* score 297, evalue 3e-152, identity 315 / 324 = 97.222 % : 12 x Escherichia coli
* score 297, evalue 3e-152, identity 317 / 327 = 96.942 % : 11 x Escherichia coli
* score 297, evalue 3e-152, identity 315 / 324 = 97.222 % : 1 x Klebsiella pneumoniae

k141\_42437 (493 bp):

* score 493, evalue 0, identity 493 / 493 = 100.000 % : 2 x Yersinia enterocolitica
* score 424, evalue 0, identity 470 / 493 = 95.335 % : 3 x Citrobacter freundii
* score 424, evalue 0, identity 470 / 493 = 95.335 % : 1 x Citrobacter sp.
* score 399, evalue 0, identity 457 / 486 = 94.033 % : 1 x Citrobacter freundii
* score 391, evalue 0, identity 459 / 493 = 93.103 % : 2 x Citrobacter freundii
* score 391, evalue 0, identity 459 / 493 = 93.103 % : 36 x Klebsiella pneumoniae
* score 391, evalue 0, identity 453 / 484 = 93.595 % : 1 x Serratia liquefaciens
* score 384, evalue 0, identity 457 / 493 = 92.698 % : 1 x Klebsiella pneumoniae
* score 380, evalue 0, identity 458 / 496 = 92.339 % : 1 x Klebsiella pneumoniae
* score 379, evalue 0, identity 455 / 493 = 92.292 % : 1 x Citrobacter amalonaticus
* score 379, evalue 0, identity 455 / 493 = 92.292 % : 8 x Citrobacter freundii
* score 379, evalue 0, identity 455 / 493 = 92.292 % : 1 x Citrobacter portucalensis
* score 379, evalue 0, identity 455 / 493 = 92.292 % : 6 x Citrobacter sp.
* score 379, evalue 0, identity 455 / 493 = 92.292 % : 1 x Citrobacter youngae
* score 379, evalue 0, identity 455 / 493 = 92.292 % : 1 x Enterobacter roggenkampii
* score 376, evalue 0, identity 454 / 493 = 92.089 % : 2 x Citrobacter freundii
* score 376, evalue 0, identity 449 / 485 = 92.577 % : 1 x Klebsiella oxytoca
* score 376, evalue 0, identity 449 / 485 = 92.577 % : 2 x Klebsiella pneumoniae
* score 373, evalue 0, identity 441 / 475 = 92.842 % : 1 x Yersinia rohdei
* score 361, evalue 0, identity 444 / 485 = 91.546 % : 1 x Lelliottia jeotgali
* score 361, evalue 0, identity 445 / 487 = 91.376 % : 1 x Yersinia canariae
* score 360, evalue 0, identity 444 / 486 = 91.358 % : 1 x Enterobacter lignolyticus
* score 358, evalue 0, identity 450 / 495 = 90.909 % : 1 x Escherichia coli
* score 352, evalue 0, identity 448 / 495 = 90.505 % : 1 x Citrobacter freundii
* score 352, evalue 0, identity 434 / 475 = 91.368 % : 1 x Enterobacter sp.
* score 343, evalue 1e-177, identity 439 / 487 = 90.144 % : 1 x Yersinia pseudotuberculosis
* score 337, evalue 3e-174, identity 437 / 486 = 89.918 % : 1 x Citrobacter freundii
* score 337, evalue 3e-174, identity 441 / 493 = 89.452 % : 1 x Enterobacter ludwigii
* score 334, evalue 1e-172, identity 440 / 493 = 89.249 % : 1 x Serratia fonticola
* score 331, evalue 7e-171, identity 439 / 493 = 89.047 % : 1 x Citrobacter amalonaticus
* score 331, evalue 7e-171, identity 439 / 493 = 89.047 % : 1 x Citrobacter freundii
* score 331, evalue 7e-171, identity 435 / 487 = 89.322 % : 1 x Xenorhabdus bovienii
* score 322, evalue 7e-166, identity 436 / 493 = 88.438 % : 4 x Citrobacter freundii
* score 316, evalue 1e-162, identity 434 / 493 = 88.032 % : 1 x Leclercia adecarboxylata
* score 311, evalue 9e-160, identity 427 / 484 = 88.223 % : 1 x Enterobacter roggenkampii
* score 310, evalue 3e-159, identity 432 / 493 = 87.627 % : 1 x Klebsiella sp.
* score 307, evalue 1e-157, identity 430 / 491 = 87.576 % : 1 x Pantoea ananatis
* score 304, evalue 7e-156, identity 426 / 487 = 87.474 % : 1 x Klebsiella pneumoniae
* score 304, evalue 7e-156, identity 425 / 485 = 87.629 % : 1 x Pectobacterium parmentieri
* score 301, evalue 3e-154, identity 429 / 493 = 87.018 % : 2 x Citrobacter freundii
* score 301, evalue 3e-154, identity 429 / 493 = 87.018 % : 1 x Citrobacter sp.
* score 301, evalue 3e-154, identity 429 / 493 = 87.018 % : 1 x Enterobacter ludwigii
* score 301, evalue 3e-154, identity 429 / 493 = 87.018 % : 2 x Escherichia coli
* score 298, evalue 1e-152, identity 428 / 493 = 86.815 % : 1 x Cedecea lapagei
* score 298, evalue 1e-152, identity 428 / 493 = 86.815 % : 1 x Rahnella aquatilis
* score 298, evalue 1e-152, identity 424 / 487 = 87.064 % : 1 x Raoultella ornithinolytica
* score 296, evalue 2e-151, identity 426 / 490 = 86.939 % : 1 x Enterobacter sp.
* score 295, evalue 7e-151, identity 413 / 471 = 87.686 % : 1 x Klebsiella pneumoniae
* score 292, evalue 3e-149, identity 422 / 487 = 86.653 % : 2 x Citrobacter sp.
* score 292, evalue 3e-149, identity 427 / 494 = 86.437 % : 1 x Enterobacter sp.
* score 292, evalue 3e-149, identity 421 / 485 = 86.804 % : 1 x Erwinia tracheiphila
* score 292, evalue 3e-149, identity 414 / 475 = 87.158 % : 1 x Klebsiella aerogenes
* score 292, evalue 3e-149, identity 414 / 475 = 87.158 % : 1 x Klebsiella grimontii
* score 292, evalue 3e-149, identity 431 / 499 = 86.373 % : 1 x Yersinia intermedia
* score 289, evalue 1e-147, identity 419 / 484 = 86.570 % : 1 x Pantoea alhagi
* score 289, evalue 1e-147, identity 413 / 475 = 86.947 % : 1 x Pantoea ananatis
* score 289, evalue 1e-147, identity 421 / 487 = 86.448 % : 1 x Yersinia enterocolitica
* score 286, evalue 7e-146, identity 424 / 493 = 86.004 % : 1 x Citrobacter freundii
* score 286, evalue 7e-146, identity 419 / 485 = 86.392 % : 1 x Enterobacter hormaechei
* score 286, evalue 7e-146, identity 420 / 487 = 86.242 % : 2 x Klebsiella pneumoniae
* score 286, evalue 7e-146, identity 420 / 487 = 86.242 % : 1 x Morganella morganii
* score 284, evalue 9e-145, identity 421 / 489 = 86.094 % : 1 x Klebsiella oxytoca
* score 283, evalue 3e-144, identity 411 / 475 = 86.526 % : 7 x Erwinia amylovora
* score 283, evalue 3e-144, identity 423 / 493 = 85.801 % : 2 x Erwinia amylovora
* score 280, evalue 1e-142, identity 422 / 493 = 85.598 % : 1 x Erwinia amylovora
* score 280, evalue 1e-142, identity 419 / 488 = 85.861 % : 1 x Tatumella ptyseos
* score 274, evalue 3e-139, identity 422 / 495 = 85.253 % : 1 x Klebsiella variicola
* score 268, evalue 7e-136, identity 414 / 487 = 85.010 % : 1 x Pectobacterium odoriferum
* score 265, evalue 3e-134, identity 419 / 495 = 84.646 % : 1 x Enterobacter sp.
* score 265, evalue 3e-134, identity 419 / 495 = 84.646 % : 1 x Mixta intestinalis
* score 245, evalue 4e-123, identity 279 / 296 = 94.257 % : 1 x Escherichia coli
* score 238, evalue 3e-119, identity 404 / 487 = 82.957 % : 3 x Enterobacter cloacae
* score 231, evalue 3e-115, identity 365 / 432 = 84.491 % : 1 x Dickeya aquatica
* score 229, evalue 3e-114, identity 287 / 316 = 90.823 % : 1 x Yersinia kristensenii
* score 228, evalue 1e-113, identity 364 / 432 = 84.259 % : 1 x Serratia marcescens
* score 225, evalue 6e-112, identity 274 / 298 = 91.946 % : 1 x Klebsiella pneumoniae
* score 144, evalue 6e-67, identity 263 / 321 = 81.931 % : 2 x Lelliottia amnigena
* score 95, evalue 1e-39, identity 286 / 381 = 75.066 % : 1 x Pectobacterium wasabiae
* score 87, evalue 3e-35, identity 305 / 412 = 74.029 % : 1 x Serratia surfactantfaciens
* score 62, evalue 2e-21, identity 172 / 225 = 76.444 % : 1 x Salmonella enterica

k141\_93443 (309 bp):

* score 309, evalue 7e-159, identity 309 / 309 = 100.000 % : 3 x Yersinia enterocolitica
* score 303, evalue 1e-155, identity 307 / 309 = 99.353 % : 8 x Yersinia enterocolitica
* score 300, evalue 7e-154, identity 306 / 309 = 99.029 % : 3 x Yersinia enterocolitica
* score 297, evalue 3e-152, identity 305 / 309 = 98.706 % : 6 x Yersinia enterocolitica
* score 267, evalue 1e-135, identity 295 / 309 = 95.469 % : 2 x Yersinia enterocolitica
* score 212, evalue 6e-105, identity 276 / 308 = 89.610 % : 1 x Yersinia canariae
* score 185, evalue 6e-90, identity 267 / 308 = 86.688 % : 1 x Yersinia kristensenii
* score 167, evalue 6e-80, identity 257 / 302 = 85.099 % : 1 x Yersinia frederiksenii
* score 167, evalue 6e-80, identity 257 / 302 = 85.099 % : 2 x Yersinia kristensenii
* score 164, evalue 3e-78, identity 256 / 302 = 84.768 % : 1 x Yersinia kristensenii
* score 152, evalue 1e-71, identity 250 / 299 = 83.612 % : 1 x Yersinia bercovieri
* score 152, evalue 1e-71, identity 256 / 308 = 83.117 % : 1 x Yersinia sp.
* score 148, evalue 2e-69, identity 230 / 271 = 84.871 % : 1 x Yersinia aldovae
* score 146, evalue 3e-68, identity 250 / 302 = 82.781 % : 1 x Yersinia frederiksenii
* score 146, evalue 3e-68, identity 254 / 308 = 82.468 % : 2 x Yersinia massiliensis
* score 139, evalue 2e-64, identity 233 / 280 = 83.214 % : 1 x Yersinia intermedia
* score 139, evalue 2e-64, identity 233 / 280 = 83.214 % : 1 x Yersinia sp.
* score 133, evalue 5e-61, identity 231 / 280 = 82.500 % : 1 x Yersinia intermedia

#### Conclusions from comparison of all contigs with the MLST database

For sample 2672480, the general database (nt) search concurs entirely with both the MetaPhlan3 results and the MLST database search: the results are positive for identification of *K. pneumoniae* and negative for all other species of concern in the list.

For sample 6412, the general database (nt) search concurs entirely with both the MetaPhlan3 results and the MLST database search regarding positive results for *Enterococcus faecium, Enterococcus faecalis and E. coli.* The general search also concurs with MetaPhlan3 regarding the positive for *K. pneumoniae* (the MLST search result was ruled negative for this, but was not clear-cut). For all other species except *Y. enterolitica,* the general search results concur with both of the other two methods, on a negative result.

Thus, the automated MetaPhlAn3 method appears to have been largely reliable, with a single additional positive species yielded by the additional search coupled with detailed inspection. False negatives for the MLST-based method can be expected to be quite common, because the chances of at least one of the MLST sequences of a genome being matched, are relatively low unless there is particularly high abundance of the particular strain(s) present.

## Overall conclusions

Conclusions regarding the positives at the DNA-sequence level of any of the specified species of concern in the two ice cream samples with IDs 2672480 and 6412 are in Table 5. Presence of DNA does not constitute confirmation of viable bacteria.

Table 5 Identification of DNA sequences unique to the species of the priority taxa list in the two ice cream samples.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| sample ID | *Enterococcus faecalis* | *Enterococcus faecium* | *Klebsiella pneumoniae* | *Echerichia coli* | *Yersinia enterolitica* | All other species |
| 2672480 | - | - | positive | - | - | - |
| 6412 | positive | positive | positive | positive | positive | - |