# Appendix 1B: Analysis of the metabarcoding sequence data from 1,001 samples.

## Introduction

Two types of analysis were conducted on the 16S metabarcode sequence data.

Refer to Appendix 1A for a description of the 16S rRNA gene sequence reference database on which all of these analyses are founded.

In the general case, 16S rRNA sequences (even full-length gene sequences) are sufficient to resolve to genus level but not to species level (there are numerous exceptions). The resolving power of a segment of the gene sequence (such as the V4 region) is still less than that of the full-length, but in general is still sufficient for genus-level resolution. Accordingly, the expectation of standard methods of analysing a set of 16S rRNA amplicon sequences is that a large proportion will be identifiable at genus level, with relatively few at species level, while a not insignificant proportion will be identifiable only at family level or higher.

### Microbial Community Analysis

We refer to the attempted aim of assigning taxa to all such amplicon sequences - or as many as possible - as the **Microbial Community Analysis** of 16S metabarcode data (the amplicon sequences are the 'query sequences'). This is described in section 2.5.1 of the main Report.

* In essence, this "forward" analysis uses all the amplicon sequences as queries, and attempts to answer for each: what is the taxonomic origin of this sequence?

### Priority Taxon Detection

* Rationale for a targeted method for priority taxa
* Priority taxa list
* Methods: Determination of ambiguous and distinguishable reference 16S V4 sequences
* Methods: Comparing the priority taxa 16S V4 sequences to the samples' metabarcode data

#### Rationale for a targeted method for priority taxa

In addition to the more conventional analysis described above (Microbial Community Analysis), a targeted approach to detecting the priority taxa was undertaken. **This does not constitute part of the AMR study itself**, and was for statutory food-safety considerations (see 1.2 Aims and objectives, in main Report).

The reason for supplementing the standard, general approach (2.5.1) is essentially that no such methods are flawless, and we took a parallel approach, targeted towards the taxa of concern. The general methods can quite rarely be expected to provide taxonomic identification finer than the genus level. Those methods also employ some trade-offs which may result in some sequences being mis-assigned. For example, it is possible for the methods which aim to correct error-containing sequences to occasionally change a sequence which was in fact already correct.

Therefore our targeted approach aimed to identify any original sequence data which is **consistent** with any of the priority taxa.

* This is effectively a "reverse" search, which uses the known priority taxa sequences as queries, and attempts to answer for each: do these occur in the samples?

There is however a prior question for each taxon in the priority list:

* Is it even in theory possible to distinguish the 16S V4 sequences of this taxon from all of **those which are not on the priority list**?

This analysis takes into account the distinct nature of groups of V4 sequences in the context of whether they originate from taxa which are of interest or not, in a way absent from the general methods employed for the Microbial Community Analysis. For example, within one genus, there may be no 16S rRNA V4-region sequences unique to a particular species, but if a particular sequence occurs only in several species which are **all** taxa of concern, then the sequence provides useful discriminating information.

#### Priority taxa list

The list consists of foodborne pathogens of humans, including all species in the following genera (using Silva database taxonomic designations);

* *Campylobacter*
* *Escherichia-Shigella*
* *Listeria*
* *Salmonella*
* *Yersinia*

(note the grouping of *Escherichia* and *Shigella* into a single genus in this taxonomy).

Furthermore, the following species were of interest;

* *Klebsiella pneumoniae*
* *Staphylococcus aureus*
* *Enterococcus faecalis* and *Enterococcus faecium*
* *Clostridium perfringens, Clostridium botulinum, Clostridium sporogenes, Clostridium butyricum,* and *Clostridium baratii* (these species are included in the **5 genera** *Clostridium sensu stricto* 1, 3, 7, 13 and 18 in the Silva database).

There are therefore 13 genera of relevance.

The total number of named distinct named taxa is effectively 16, given the two *Enterococcus* taxa, and that *C. botulinum* specifies 3 taxa (one in each of *Clostridium* sensu stricto 1, *Clostridium* sensu stricto 7 and *Clostridium* sensu stricto 18; the *C. botulinum* annotations in *Clostridium* sensu stricto 3 and *Clostridium* sensu stricto 13 are of more dubious provenance).

#### Methods: Determination of ambiguous and distinguishable reference 16S V4 sequences

The first analysis required, prior to any analysis of the sequence data obtained from the samples, was an *a priori* assessment of the reference database (SILVA) sequences relevant to the taxa listed in the previous section.

In essence, this procedure is as follows:

* Determine the V4-region segments of all of the reference sequences within all of the 13 **relevant genera**.
* Group together all of those V4-region sequences, such that every group represents a unique V4 sequence (irrespective of the taxonomic annotations of the member sequences).
  + This was done for each genus independently.
* For species-level priority taxa, identify all of the **named species** within each group.
* Identify all of the groups where the **named species** consist **only of one or more priority taxa**. The V4 sequence of each such group thus represents, in principle, a sequence which signals concern if an exactly identical sequence appears in the metabarcoding data of a sample.

The grouping by genus independently means that for any V4 sequences which occur in **more than one genus** (which is possible), the occurrence of such a sequence in a sample would flag the presence of two or more genera. However, inter-genus cases could conceivably occur where one genus was not a priority taxon (so would not yet be known via the above procedure). If so, then this would later become apparent in the more detailed analysis triggered by these events.

Reliably compiling named-species annotations is not straightforward; more details are provided in Appendix 1A. In brief, the reference sequences are **classified by the database curators to genus-level only**. The source of the SILVA sequences is a third-party database which does contain notionally species-level annotations ("organism names"). In many cases these names are formal bacterial species names, but in many cases they are not, such as "unclassified bacterium", "environmental sample", etc. In other cases, they may be minority annotations greatly at odds with the species names of the majority of identical or near-identical sequences, indicating that they are unreliable.

#### Methods: Comparing the priority taxa 16S V4 sequences to the samples' metabarcode data

In contrast to the Microbial Community Analysis which compares all the metabarcode sequences with all sequences in a reference database, the priority taxa search is narrower, with a relatively small number of query taxa involved. A traditional sequence similarity search (using Blast) is tractable and efficacious. We therefore created Blast databases of all of the samples' metabarcode sequences. The unique sequences (each representing a group of 1 or more different taxa which share that identical sequence) were used as queries in a BlastN search of these databases.

##### Search stringency

The overall aim was to detect amplicon sequences identical or near-identical to any of the priority taxa query 16S V4 sequences. Matching sequences would thus be **consistent** with, though **not necessarily diagnostic** of, the presence of DNA originating from the taxa of interest (this would depend on the discriminatory power of each query sequence, as explained previously).

The rationale was that for any given priority taxon, the *a priori* expectation was that the majority of samples would yield no positives. Any sequences consistent with the priority taxa would then be analysed on a case-by-case basis if so directed by the FSA, according to a contingency response flowchart.

The approach was therefore intentionally liberal, with even single instances of a positive (priority taxon-matching) amplicon sequence treated as a match worthy of further investigation. (In contrast, when matches occur at extremely low abundance in the Microbial Community Analysis, these may not register as positives if an insufficient number meets various other criteria). The search aimed to observe all explicit evidence of one or more sequence matches, even if this involved only one read of a pair (referring to the paired-end amplicon sequencing library we used, in which each fragment is in effect sequenced twice, once from each end; the two reads of a pair are usually identical but sequencing errors mean that this is not always the case). However, a read pair where both partner reads match the reference provides more convincing evidence than when only one does so.

##### Assessing the strength of sequence matches

Only sequence matches where the alignment length was the same or very close to the full length of the V4 region, and the sequence identity of the alignment was at least 97%, were of interest. The V4 sequence of the bacterial 16S is usually 253 DNA base pairs (bp) long. Alignments shorter than 250 bp were ignored. In practice, even those matches lower than 98% identity are are barely reliable enough to attempt to form conclusions with this particular methodology, but the 97% limit ensures that nothing of potential interest is lost.

We noted that the E-value Blast metric was invariably lower than 10-100 for such matches in our databases (E-value ranges depend partly on the composition of the database). We ran the BlastN searches with an E-value threshold of 10-90 which was easily sufficient to capture all matches with a sequence identity of somewhat lower than 97%. The results were then filtered to discard matches of worse than 97% identity. Statistics were then compiled for progressively more stringent combinations of % identity and alignment length.

**We emphasise that for these purposes, even 97%-98% identity over the full length of the V4 region is a liberal matching criterion.** However, screening at this level means that no genuine matches of concern escape attention. In practice, the matches at 99% and 100% sequence identity in the alignment would be those of interest - and even some of those may be of less interest depending on the alignment length.

##### Interpreting % sequence identities of alignments

The result for a given sample and given taxon is thus expressed as a percentage identity between 97 and 100 (which is the rounded-down integer of the exact percentage value). These should be interpreted as follows:

**100% sequence identity of alignment** : **This does not always represent a perfect match**, because a minority of matches omit one or a very few DNA bp from the ends of the alignment (sequence comparison): that is, the alignment is shorter than the V4 region. However, for most taxa, the great majority (well over 90%) of 100%-identical matches are full-length (253 bp). Some particular taxa are more variable (reflecting variability in the reference sequence lengths). However, the results presented here ignore alignments < 250 bp long. As stated, even a perfect match to a particular species does not necessarily identify that species’ DNA, because multiple species can have identical V4 sequences**. In some branches of the bacterial taxonomy, identical full-length V4 sequences can occur even in species of different genera**; this is particularly the case for some taxa of interest here, such as some genera of the Enterobacteriaceae.

**99% sequence identity of alignment**: This represents cases of the best match between any DNA sequence read in the sample differing by either 1 or 2 bp from the best-matching reference. Clearly, this is enough to represent different species in the general case (since even 0 bp differing can represent multiple species). However, such cases should be given serious consideration as identifying the DNA sequence of interest, especially given error rates in the DNA sequencer (although of high fidelity, the number of reads with 1 or 2 bp in error will be high in absolute terms; the majority of reads can be expected to be error-free).

**98% sequence identity of alignment**: These matches can differ by up to 5 bp from the reference. Some caution is required when evaluating these, and conclusions can differ depending on the taxa concerned.

**97% sequence identity of alignment:** These matches must be treated with great caution as they can be expected to include a substantial number of false positives**. Two V4 sequences 97% identical can easily be different species**. We consider a lack of any matches even at this lowest level, to constitute a clear negative.

**We emphasise that these values 97-100% do not represent ‘confidence intervals’**. They are a literal statement of the **best-case sequence identity** (in some cases **not even representing the full length** of the V4 region) found in each sample for a given taxon.

##### The analysis is not quantitative

The analysis is essentially qualitative, and only semi-quantitative, in terms of incidence within each sample. The qualitative aspect is the identification of samples as positive or negative for the V4 sequences in question. We do not present the results as quantitative (incidence of each V4 sequence per sample) *per se*. This is partly because we do not consider more than the best 1,000 sequence reads from a sample for any given reference sequence. Further, some taxa of interest, e.g. the whole *Escherichia-Shigella* genus, contain vastly more reference sequences than others, especially the single-species taxa. In some situations, a large number of sample-to-reference matches result from many different sample reads matching multiple references (many combinations). Alternatively, these can result from the same small number of reads matching multiple references.

The results of the conventional and high priority taxon analyses were reported to the FSA, with any further, more detailed sequence analysis conditional upon their consideration.

## Results: Microbial Community Analysis

The results of the standard "forward" analysis (Microbial Community Analysis) are described in section 3.3.1 of the main Report.

## Results: Priority Taxon Detection

### Determination of ambiguous and distinguishable reference 16S V4 sequences

#### Groups of species indistinguishable by the 16S rRNA gene V4 region

The headed lists below (‘Named species of interest’) refer solely to the *individual species* highlighted by the Food Standards Agency ('notifiable species'), and not the additional genus-wide taxa. The bulleted lists of species are, in general, indistinguishable from the notifiable species using the V4 region. It cannot be completely ruled out that some species in a given list have a unique variant, present in only that species, which might enable species-level identification. However, assuming the veracity of the reference-database annotations, such cases are exceptions.

Those species marked ‘\*’ appear relatively likely to be the result of mis-annotations (see minority annotations, Appendix 1A). E.g. the V4 sequences annotated as *Klebsiella oxytoca* could possibly have been mis-annotated, because there are very few instances of this whereas other V4 sequences in the database are much more commonly annotated as this species.

In contrast, there is a single V4 sequence which occurs in more than 17,000 databases sequences annotated as *Staphylococcus aureus*, and also in almost 900 sequences annotated as *S.* *epidermidis*. This V4 sequence is, by a huge margin, the V4 sequence most commonly annotated as *S.* *epidermidis* (and indeed many of the other species listed in the group, including *S. aureus, S. haemolyticus, S. hominis* etc). The conclusion is that these are all correctly annotated, and that these different species all have identical V4 regions. This is referred to as the ‘principal’ V4 sequence for all of these species.

Of the species highlighted as notifiable taxa, there appear to be a few which are represented in the database mainly by a unique V4 sequence. That is, of all the annotations of the traditional binomial form (i.e. of the form "*Genus-name specific-name*"), there is only one species (often there are multiple cases of non-binomial names occurring once each, e.g. ‘*Clostridium* sp. TM147’; also it is not unusual for the most common annotation of all to be ‘uncultured bacterium’, i.e. the sequences arose from a 16S metabarcoding study).

These potentially uniquely identifiable species are *Clostridium perfringens;* a particular *Clostridium botulinum* species which is a member of the*Clostridium sensu stricto 1* genus (this species includes B and E toxin types); *Clostridium butyricum* (which is unusual among the species of interest in that there are two almost-identical variants, both of which are distinguishable from other species). There is additionally a unique *C. botulinum* V4 sequence in the genus *Clostridium sensu stricto 3,* which appears to be dubious because it is vaguely annotated (no toxin type or strain information at all) and occurs only once.

#### Named species of interest

Those species marked ‘\*’ appear relatively likely to be the result of mis-annotations; it cannot be ruled out for certain that they are correct, however. Refer to previous section.

***Staphylococcus aureus***

Generally indistinguishable by V4 sequence:

* *Staphylococcus aureus*
* *Staphylococcus epidermidis*
* *Staphylococcus haemolyticus*
* *Staphylococcus warneri*
* *Staphylococcus hominis*
* *Staphylococcus pasteuri*
* *Staphylococcus capitis*
* *Staphylococcus argenteus*
* *Staphylococcus lugdunensis*
* *Staphylococcus caprae*
* *Staphylococcus petrasii*
* *Staphylococcus devriesei*
* *Staphylococcus saccharolyticus*
* *Staphylococcus schweitzeri*
* *Staphylococcus simiae*

***Klebsiella pneumoniae***

Generally indistinguishable by V4 sequence:

* *Klebsiella pneumoniae*
* *Klebsiella aerogenes*
* *Klebsiella variicola*
* *Klebsiella quasipneumoniae*
* *Klebsiella oxytoca\**
* *Klebsiella granulomatis*

***Enterococcus faecalis*** *and* ***Enterococcus faecium***

Generally indistinguishable by V4 sequence:

* *Enterococcus faecalis*
* *Enterococcus faecium*
* *Enterococcus durans*
* *Enterococcus hirae*
* *Enterococcus mundtii*
* *Enterococcus lactis*
* *Enterococcus dispar*
* *Enterococcus villorum*
* *Enterococcus rivorum*
* *Enterococcus canis*
* *Enterococcus pernyi*
* *Enterococcus canintestini*
* *Enterococcus ratti*
* *Enterococcus olivae*
* *Enterococcus saigonensis*
* *Enterococcus azikeevi*
* *Enterococcus thailandicus\**
* *Enterococcus raffinosus \**

***Clostridium perfringens***

Appears to have a unique V4 sequence among the traditional binomial-style names in genus *Clostridium sensu stricto 1.*

***Clostridium botulinum*** (B and E toxin types; genus *Clostridium sensu stricto 1*)

Possibly, one sequence (may be misannotated as *C. botulinum*) may be indistinguishable from *Clostridium septum.*

Conversely, a possibly misannotated *Clostridium taeniosporum* is indistinguishable from one of the botulinum sequences (the principal cluster, 98% of which are annotated as *C. botulinum*).

***Clostridium baratii***

Generally indistinguishable by V4 sequence:

* *Clostridium baratii*
* *Clostridium nitritogenes* (synonym of *Eubacterium nitritogenes)*

***Clostridium butyricum***

This appears to have two unique V4 sequences (both with large weight of evidence) among the traditional binomial-style names in genus *Clostridium sensu stricto 1.*

***Clostridium botulinum*** (no type stated; *Clostridium sensu stricto 3*)

Only a single sequence, which appears that it may well be a mis-annotation of *C. amylolyticum* (and possibly *C.* *polynesience*). The lack of a stated toxin type or other strain information weakens confidence in this annotation.

***Clostridium botulinum*** (C, D, C/D, D/C toxin types; *Clostridium sensu stricto 7*)

Generally indistinguishable by V4 sequence:

* *Clostridium botulinum*
* *Clostridium haemolyticum*
* *Clostridium novyi*
* *Clostridium massiliodielmoense*

***Clostridium botulinum*** (no type stated; *Clostridium sensu stricto 13*)

Generally indistinguishable by V4 sequence:

* *Clostridium argentinense*
* *Clostridium subterminale*
* *Clostridium botulinum\**

Again, no toxin type or strain information is provided for the *C. botulinum* records (which number only two, both identical in sequence to 17 and 3 records respectively annotated as the other two species).

***Clostridium botulinum*** (A, B, F toxin types; *Clostridium sensu stricto 18*)

Generally indistinguishable by V4 sequence:

* *Clostridium botulinum*
* *Clostridium sporogenes*
* *Clostridium butyricum\**

### Comparing the priority taxa 16S V4 sequences to the samples' metabarcode data

#### Numbers of samples with at least one metabarcode match to a priority taxon reference sequence

The numbers of samples which contain at least one metabarcode sequence match with at least one priority taxon reference sequence are shown for five stringency levels, in Table B1.

A stringency level is defined as a combination of minimum alignment (sequence match) length and the minimum % sequence identity over the length of the alignment. The % identity is the number of positions in the alignment where the sample metabarcode sequence and the reference sequence have the same nucleotide).

Only combinations of minimum identity = 98%, 99% or 100% with minimum length = 250 bp or 253 bp are considered.

For each taxon and at each stringency level

after which the presence of sequences consistent with Salmonella spp. were queried by FSA. Upon further investigation it was concluded that while the sequences were consistent with Salmonella spp., they were also consistent with several other taxa in the Enterobacteriaceae, and as such were not subjected to further investigation. This exercise served as a learning opportunity for all participants around deliverable scheduling and the potential additional responses required.

**Table B1.** Numbers of samples with at least one 16S rRNA gene V4 amplicon read aligning with at least one reference sequence representing the named priority taxa, to various levels of stringency. The stringency increases from left to right – from least to most stringent.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| minimum alignment length (bp) | 250 | | | 250 | | | 253 | | | 250 | | | 253 | | |
| minimum alignment % identity | 98 | | | 99 | | | 99 | | | 100 | | | 100 | | |
| matching read | fwd | rev | both | fwd | rev | both | fwd | rev | both | fwd | rev | both | fwd | rev | both |
| *Staphylococcus aureus* | 688 | 630 | **622** | 640 | 583 | **575** | 639 | 579 | **571** | 561 | 488 | **481** | 558 | 484 | **477** |
| *Listeria* | 221 | 193 | **172** | 15 | 10 | **10** | 15 | 10 | **10** | 9 | 7 | **7** | 9 | 7 | **7** |
| *Enterococcus faecalis* 1 | 435 | 407 | **376** | 281 | 254 | **249** | 281 | 251 | **247** | 175 | 154 | **149** | 170 | 151 | **146** |
| *Enterococcus faecium* 1 | 488 | 458 | **438** | 337 | 309 | **302** | 337 | 306 | **300** | 288 | 262 | **259** | 288 | 262 | **259** |
| *Clostridium perfringens* | 15 | 13 | **13** | 12 | 10 | **10** | 12 | 9 | **9** | 8 | 7 | **7** | 8 | 6 | **6** |
| *Clostridium butyricum* | 105 | 102 | **102** | 100 | 96 | **94** | 100 | 94 | **93** | 44 | 45 | **43** | 44 | 44 | **42** |
| *Clostridium baratii* | 24 | 22 | **21** | 3 | 1 | **1** | 3 | 0 | **0** | 1 | 0 | **0** | 1 | 0 | **0** |
| *Clostridium botulinum* (s.s. 1) | 221 | 212 | **212** | 32 | 18 | **4** | 32 | 17 | **4** | 0 | 0 | **0** | 0 | 0 | **0** |
| *Clostridium botulinum* (s.s. 3) 2 | 0 | 0 | **0** | 0 | 0 | **0** | 0 | 0 | **0** | 0 | 0 | **0** | 0 | 0 | **0** |
| *Clostridium botulinum* (s.s. 7) | 2 | 2 | **2** | 1 | 1 | **1** | 1 | 1 | **1** | 1 | 0 | **0** | 1 | 0 | **0** |
| *Clostridium botulinum* (s.s. 13) 2 | 4 | 6 | **2** | 0 | 0 | **0** | 0 | 0 | **0** | 0 | 0 | **0** | 0 | 0 | **0** |
| *Clostridium botulinum* (s.s. 18) 3 | 6 | 6 | **6** | 6 | 5 | **5** | 6 | 5 | **5** | 6 | 3 | **3** | 6 | 3 | **3** |
| *Clostridium sporogenes* 3 | 6 | 6 | **6** | 6 | 5 | **5** | 6 | 5 | **5** | 6 | 3 | **3** | 6 | 3 | **3** |
| *Campylobacter* | 18 | 19 | **18** | 17 | 17 | **17** | 17 | 17 | **17** | 17 | 17 | **17** | 17 | 17 | **17** |
| *Klebsiella pneumoniae* 4 | 701 | 650 | **641** | 640 | 590 | **587** | 638 | 588 | **580** | 551 | 484 | **476** | 550 | 483 | **475** |
| *Escherichia-Shigella*4 | 872 | 823 | **819** | 795 | 761 | **750** | 794 | 758 | **746** | 697 | 630 | **618** | 696 | 628 | **617** |
| *Salmonella* 4 | 805 | 756 | **753** | 719 | 662 | **654** | 718 | 660 | **652** | 449 | 407 | **396** | 449 | 404 | **395** |
| *Yersinia* | 687 | 620 | **596** | 504 | 446 | **434** | 504 | 442 | **430** | 433 | 363 | **358** | 432 | 361 | **356** |

Notes:

1. *E. faecalis* and *E. faecium* are not only generally not distinguishable from several other *Enterococcus* species, but in many cases are indistinguishable from each other; that is, some reference 16S V4 sequences are identical in both, and may account for some of the positive samples in both rows. Some others occur only in instances of one but not the other (but also in other *Enterococcus* species).
2. The reference sequences annotated as *Clostridium botulinum* in the SILVA genera *Clostridium sensu stricto 3* and *Clostridium sensu stricto* 13 are considered here to be questionable and need extra caution, but have been included since they are representatives of the named species in those genera.
3. *Clostridium botulinum* of genus *Clostridium sensu stricto* 18 and *C. sporogenes* (same genus) are in fact represented by some reference V4 sequences which are identical, and one of those is the sole reference sequence involved here. The two taxa have been identified in separate rows since these are two named species in the FSA list.
4. Genera *Klebsiella, Escherichia-Shigella* and *Salmonella* belong to family Enterobacteriaceae, and some reference V4 sequences may occur in more than one genus (including the other priority taxa in this family and also non-priority genera). Such identities have not been pre-determined. Therefore some of these reference sequences may account for some of the positive samples in one, two or three of these rows.

The alignments are the results of BlastN searches using the **references sequences** (SILVA database), corresponding to the taxa, as queries versus databases of the 1,001 food samples 16S metabarcoding sequences.

**Many reference sequences do not discriminate between the named taxon of origin and other taxa**, because the same sequence occurs in the reference database in non-trivial frequencies from multiple taxonomic origins.

The named taxa (listed in order of major taxonomic group) are those in the FSA priority list. Colour scheme:

* The cells shaded in grey are species-level taxa which cannot be distinguished from non-priority species in the same genus. These are *S. aureus*, *E. faecalis, E. faecium, C. botulinum* (s.s.7 and s.s. 13)*, K. pneumoniae*.
* Those in blue and green appear to be distinguishable from **non-priority** species, in principle (if taking only reference sequences with formal binomial species annotations into account). Blue – *C. botulinum* (s.s. 18), *C. sporogenes*. Green – *C. perfringens, C. butyricum, C botulinum* (s.s.1 and s.s.3)
* Those in green appear to be distinghishable from other species (including other priority species).
* The two blue species are not distinguishable from each other.
* The genera shaded in pink belong to family Enterobacteriaceae, and some reference V4 sequences may occur in more than one genus (including the other priority taxa in this family and also non-priority genera). Pink – *Escherichia-Shigella*, *Salmonella*
* For each of the remaining genus-level taxa, in white (*Listeria, Campylobacter, Yersinia*) it cannot be guaranteed that all V4 sequences are restricted solely to that genus.

"s.s. 1" etc refer to genera named "*Clostridium sensu stricto* 1" etc of the SILVA database taxonomy.

"**fwd**", "**rev**" refer to corresponding forward and reverse reads of a pair. As a result of the sequencing library, each 16S amplicon is sequenced twice. Many pairs return identical sequences, but some pairs can differ in length or in a small number of nucleotide base-calls. Normally, these are merged (as in the Microbial Community Analysis performed separately), which in effect "corrects" discrepancies. However, the correct sequence is not always certain, and the forward and reverse reads have been analysed separately in the interests of exposing any explicit sequence evidence which might be compatible with the reference sequences (i.e. a liberal analysis). Conversely, the "**both**" column (yellow, orange) represents the pairs where both (unmerged) reads must pass the criteria. For the most stringent case (100% identical over 253 base-pairs; orange), this is thus even more stringent than the merged case (where at least one of the two reads might differ from the reference, but the merged sequence is identical), since both reads must be identical to each other and to the reference. The orange column represents the most stringent criteria possible for assessing the alignment.

In all cases of species-level annotations in the **reference database**, these have arisen from the original "organism name" annotations in the third-party database (European Nucleotide Archive, ENA) from which SILVA draws its data. **These are not the result of SILVA curation.** Some of these species-level annotations must be treated with caution (e.g. see note 3 below) while others appear very sound due to the weight of frequency with which they occur in the reference database.

#### Numbers of unique priority taxa reference sequences matched exactly

Table B2 is restricted to the most stringent match criteria (final column in table Jx1). It shows the numbers of unique reference sequences, originating from the sequences annotated as each taxon, which are involved in one or more exact matches with sample amplicon(s).

**Table B2.** Total numbers of samples and unique reference sequences involved in the alignments passing the most stringent criteria in Table Jx1 (100% sequence identity of both metabarcode read sequences with the reference sequence, over the 253 bp length).

|  |  |  |  |
| --- | --- | --- | --- |
| taxon | Number of samples | Number of unique reference sequences | Notes on reference sequences |
| *Staphylococcus aureus* | **477** | 63 |  |
| *Listeria* | **7** | 2 |  |
| *Enterococcus faecalis* 1 | **146** | 19 | 6 are also reference sequences in the *E. faecium* alignments |
| *Enterococcus faecium* 1 | **259** | 35 | 6 are also reference sequences in the *E. faecalis* alignments |
| *Clostridium perfringens* | **6** | 1 |  |
| *Clostridium butyricum* | **42** | 4 |  |
| *Clostridium baratii* | **0** | 0 |  |
| *Clostridium botulinum* (s.s. 1) | **0** | 0 |  |
| *Clostridium botulinum* (s.s. 3) 2 | **0** | 0 |  |
| *Clostridium botulinum* (s.s. 7) | **0** | 0 |  |
| *Clostridium botulinum* (s.s. 13) 2 | **0** | 0 |  |
| *Clostridium botulinum* (s.s. 18) 3 | **3** | 1 | the same reference sequence as for *C. sporogenes* |
| *Clostridium sporogenes* 3 | **3** | 1 | the same reference sequence as for *C. botulinum, Clostridium*  *sensu stricto* 18 |
| *Campylobacter* | **17** | 3 |  |
| *Klebsiella pneumoniae* 4 | **475** | 32 |  |
| *Escherichia-Shigella*4 | **617** | 309 |  |
| *Salmonella* 4 | **395** | 5 |  |
| *Yersinia* | **356** | 3 |  |

Refer to Table B1 for footnotes and explanation of shading.

#### Matching reference sequences (selected taxa only) found in each sample, grouped by food category

For the species-level priority taxa which can be putatively discriminated from all non-priority taxa (green and blue cells in Tables B1, B2), a full list of occurrence of each unique reference sequence in each individual sample is shown in Table B3. The two taxa negative in all samples (*Clostridium botulinum* toxin types of genera *Clostridium* *sensu stricto 1* and *Clostridium sensu stricto* 3) are omitted.Again this is restricted to the most stringent match criteria.The number of read pairs each reference sequence exactly matched within each sample is also shown.

Additionally, the same data is shown for the two genus-level priority taxa (*Listeria*, *Campylobacter*), as the total numbers of samples are small enough to list each individually. For total numbers of positive samples for each of the *Salmonella* and *Yersinia* reference sequences, see Table B4.

The table is in several parts for convenience, with samples grouped by food category:

1. Whole milk (10 samples)
2. Semi-skimmed and skimmed milk (12 samples)
3. Miscellaneous: Ice cream (2 samples), yogurt (1), soya milk (3), apple juice (1)
4. Apples and pears (17 samples)
5. Various fruit (15 samples)
6. Various fruit and vegetables (5 samples)
7. Meat (2 samples)

For ease of comparison, where one of the taxa is negative for all samples in the group, the (empty) column is retained.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| sample ID | food category | *Listeria* | *Clostridium perfringens* | *Clostridium butyricum* | *Clostridium sporogenes/ C. botulinum s.s.*18 | *Campylobacter* |
| 540950 | whole milk | MJSU01000018.28598.30158 (23) |  |  |  |  |
| 343524 | whole milk | MJSU01000018.28598.30158 (1) |  |  |  |  |
| 2672591 | whole milk |  |  | CP016332.3479095.3480607 (279)  JXBT01000001.134164.135664 (6) |  |  |
| 2672731 | whole milk |  |  | CP016087.5908981.5910481 (20) |  |  |
| 558123 | whole milk |  |  |  |  | EU773268.1.1377 (156) |
| 2672655 | whole milk |  |  |  |  | EU773268.1.1377 (98) |
| 2672770 | whole milk |  |  |  |  | EU773268.1.1377 (5) |
| 2672585 | whole milk |  |  |  |  | EU773268.1.1377 (3) |
| 2686046 | whole milk |  |  |  |  | EU773268.1.1377 (2) |
| 2672639 | whole milk |  |  |  |  | EU773268.1.1377 (2) |

**Table B3(i).** Whole milk samples with at least one pair of metabarcode reads which aligns perfectly (100% sequence identity over 253 base pairs) with at least one reference 16S V4 sequence representing each of the selected priority taxa. Refer to text for a description of limitations of the reference sequences. Matches cannot be considered as confirmation that DNA from the named taxon is present. The matching sequence identifier(s) from the reference database (SILVA) are shown. Numbers of pairs of reads which match exactly are shown in parentheses. Both reads of each pair match exactly. Columns for negative taxa are retained for comparison with the following tables.

**Table B3(ii).** Semi-skimmed milk and skimmed milk samples with at least one pair of metabarcode reads which aligns perfectly (100% sequence identity over 253 base pairs) with at least one reference 16S V4 sequence representing each of the selected priority taxa. Refer to Table Jx3(i) for details. Columns for negative taxa are retained for comparison with the other tables.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| sample ID | food category | *Listeria* | *Clostridium perfringens* | *Clostridium butyricum* | *Clostridium sporogenes/C. botulinum s.s.*18 | *Campylobacter* |
| 540471 | semi- skimmed milk | MJSU01000018.28598.30158 (287) |  |  |  |  |
| 2672491 | semi- skimmed milk | MJSU01000018.28598.30158 (1) |  |  |  | EU773268.1.1377 (38) |
| 343415 | semi-skimmed milk |  |  | CP016087.5908981.5910481 (6) |  |  |
| 343416 | semi-skimmed milk |  |  | CP016332.3479095.3480607 (185) |  |  |
| 343496 | semi-skimmed milk |  |  | CP016332.3479095.3480607 (1) |  |  |
| 343541 | semi-skimmed milk |  |  | CP016332.3479095.3480607 (2) |  | EU773268.1.1377 (1) |
| 380924 | semi-skimmed milk |  |  |  |  | EU773268.1.1377 (41) |
| 343560 | semi-skimmed milk |  |  |  |  | EU773268.1.1377 (37) |
| 2672454 | semi-skimmed milk |  |  |  |  | EU773268.1.1377 (13) |
| 2672486 | semi-skimmed milk |  |  |  |  | EU773268.1.1377 (8) |
| 343573 | semi-skimmed milk |  |  |  |  | EU773268.1.1377 (1) |
| 343555 | semi-skimmed milk |  |  |  |  | EU773268.1.1377 (1) |
| 2664722 | skimmed milk |  |  | CP016087.5908981.5910481 (91) |  |  |
| 2672595 | skimmed milk |  |  |  |  | EU773268.1.1377 (1) |

**Table B3(iii).** Ice cream, yogurt, soya milk, apple juice samples with at least one pair of metabarcode reads which aligns perfectly (100% sequence identity over 253 base pairs) with at least one reference 16S V4 sequence representing each of the selected priority taxa. Refer to Table Jx3(i) for details. Columns for negative taxa are retained for comparison with the other tables.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| sample ID | food category | *Listeria* | *Clostridium perfringens* | *Clostridium butyricum* | *Clostridium sporogenes/C. botulinum s.s.*18 | *Campylobacter* |
| 2672480 | vanilla ice cream |  |  | CP016332.3479095.3480607 (14) |  |  |
| 2672583 | vanilla ice cream |  |  | CP016087.5908981.5910481 (8) |  |  |
| 2672520 | unsweetened yogurt |  | FM865913.1.1506 (1) |  |  |  |
| 343616 | soya milk unsweetened |  |  | CP016087.5908981.5910481 (12) |  |  |
| 343527 | soya milk sweetened |  |  | CP016087.5908981.5910481 (1) |  |  |
| 343617 | soya milk sweetened |  |  | CP016087.5908981.5910481 (15) |  |  |
| 343461 | apple juice pasteurised |  |  | CP016087.5908981.5910481 (1) |  |  |

**Table B3(iv).** Apples and pears samples with at least one pair of metabarcode reads which aligns perfectly (100% sequence identity over 253 base pairs) with at least one reference 16S V4 sequence representing each of the selected priority taxa. Refer to Table Jx3(i) for details. Columns for negative taxa are retained for comparison with the other tables.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| sample ID | food category | *Listeria* | *Clostridium perfringens* | *Clostridium butyricum* | *Clostridium sporogenes/ C. botulinum s.s.*18 | *Campylobacter* |
| 343460 | apples | MJSU01000018.28598.30158 (1) |  |  |  |  |
| 6395 | apples | JNFA01000014.181.1733 (14) |  |  |  |  |
| 2664620 | apples |  |  | CP016332.3479095.3480607 (6) |  |  |
| 2664660 | apples |  |  | CP016332.3479095.3480607 (16) |  |  |
| 343305 | apples |  |  | CP016087.5908981.5910481 (4) |  |  |
| 6571 | apples |  |  | CP016087.5908981.5910481 (6) |  |  |
| 2664730 | apples |  |  |  | MWIY01000025.33.1461 (9) |  |
| 2664755 | apples |  |  |  | MWIY01000025.33.1461 (3) |  |
| 2664748 | apples |  |  |  | MWIY01000025.33.1461 (4) |  |
| 2672830 | pears | MJSU01000018.28598.30158 (2) |  |  |  |  |
| 6671 | pears |  | FM865913.1.1506 (2) | CP016087.5908981.5910481 (3) |  | DQ174182.1.1339 (2) |
| 2685918 | pears |  | FM865913.1.1506 (1) |  |  |  |
| 2664675 | pears |  |  | CP016332.3479095.3480607 (3) |  |  |
| 6457 | pears |  |  | CP016332.3479095.3480607 (4) |  |  |
| 6507 | pears |  |  | CP016332.3479095.3480607 (2) |  |  |
| 6639 | pears |  |  | CP016087.5908981.5910481 (1) |  |  |
| 6657 | pears |  |  | CP016332.3479095.3480607 (7) |  |  |

**Table B3(v).** Various fruit produce samples with at least one pair of metabarcode reads which aligns perfectly (100% sequence identity over 253 base pairs) with at least one reference 16S V4 sequence representing each of the selected priority taxa. Refer to Table Jx3(i) for details. Columns for negative taxa are retained for comparison with the other tables.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| sample ID | food category | *Listeria* | *Clostridium perfringens* | *Clostridium butyricum* | *Clostridium sporogenes/ C. botulinum s.s.*18 | *Campylobacter* |
| 2664702 | plums |  |  | CP016332.3479095.3480607 (1) |  |  |
| 2664703 | plums |  |  | CP016087.5908981.5910481 (30)  CP016332.3479095.3480607 (1) |  |  |
| 6409 | peaches |  |  | CP016332.3479095.3480607 (1) |  |  |
| 6522 | peaches |  |  | CP016332.3479095.3480607 (2) |  |  |
| 2664789 | nectarines |  |  | CP016332.3479095.3480607 (2) |  |  |
| 2672537 | nectarines |  |  | CP016332.3479095.3480607 (1) |  |  |
| 6406 | nectarines |  |  | CP016332.3479095.3480607 (5)  FJ753821.1.1272 (5) |  |  |
| 2664692 | black grapes |  |  | CP016332.3479095.3480607 (8)  JXBT01000001.134164.135664 (2) |  |  |
| 2685916 | black grapes |  |  | CP016332.3479095.3480607 (12)  FJ753821.1.1272 (4) |  |  |
| 343459 | black grapes |  |  | CP016332.3479095.3480607 (1) |  |  |
| 6566 | black grapes |  |  | CP016332.3479095.3480607 (5) |  |  |
| 2664731 | white grapes |  |  | CP016332.3479095.3480607 (6) |  |  |
| 2664764 | white grapes |  |  | CP016087.5908981.5910481 (3) |  |  |
| 6297 | white grapes |  |  | CP016332.3479095.3480607 (8) |  |  |
| 2664781 | blueberries |  |  | CP016087.5908981.5910481 (6) |  |  |

**Table B3(vi).** Various fruit and vegetable produce samples with at least one pair of metabarcode reads which aligns perfectly (100% sequence identity over 253 base pairs) with at least one reference 16S V4 sequence representing each of the selected priority taxa. Refer to Table Jx3(i) for details. Columns for negative taxa are retained for comparison with the other tables.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| sample ID | food category | *Listeria* | *Clostridium perfringens* | *Clostridium butyricum* | *Clostridium sporogenes/C. botulinum s.s.*18 | *Campylobacter* |
| 2664735 | tomatoes |  | FM865913.1.1506 (1) |  |  |  |
| 2664583 | olives in brine |  |  | CP016087.5908981.5910481 (41) |  |  |
| 2664570 | white onions |  |  | CP016332.3479095.3480607 (3) |  |  |
| 2664594 | white onions |  |  | CP016087.5908981.5910481 (1) |  |  |
| 2664765 | white onions |  |  | CP016087.5908981.5910481 (1) |  |  |

**Table B3(vii).** Meat produce samples with at least one pair of metabarcode reads which aligns perfectly (100% sequence identity over 253 base pairs) with at least one reference 16S V4 sequence representing each of the selected priority taxa. Refer to Table Jx3(i) for details. Columns for negative taxa are retained for comparison with the other tables.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| sample ID | food category | *Listeria* | *Clostridium perfringens* | *Clostridium butyricum* | *Clostridium sporogenes/C. botulinum s.s.*18 | *Campylobacter* |
| 2664738 | corned beef |  | FM865913.1.1506 (1) |  |  |  |
| 2664776 | corned beef |  | FM865913.1.1506 (1) |  |  |  |
| 2664618 | ham not smoked |  |  |  |  | EU773268.1.1377 (1) |

#### Matching reference sequences (selected taxa only): total numbers of matching samples

Table B4 summarises the total numbers of samples with which an exact match was made by each of the individual reference sequences (where the number of matching samples was at least 1). Priority taxa which were specified at the species level but which are not considered to be putatively distinguishable from other taxa, have been omitted. The genus-level priority taxon *Escherichia-Shigella* has also been omitted due to the large number of reference matching reference sequences involved. In contrast, the large numbers of samples matching *Salmonella* and *Yersinia* involve a very few distinct references in each case.

**Table B4.** Reference 16S V4 database sequences of selected priority taxa which align perfectly (100% sequence identity over 253 base pairs) with both reads of at least one read-pair of at least one food sample. Priority taxa which were specified at the species level but which are not considered to be putatively distinguishable from other taxa, have been omitted. The genus-level priority taxon *Escherichia-Shigella* has also been omitted due to the large number of reference sequences which make such alignments (309).

\* ENA database sequence LZDY01000158 is annotated as a genome sequence of *E. coli*; in the SILVA database, the gene sequence specified by the segment coordinates 145 - 1,615 has been classified as *Salmonella.*

|  |  |  |
| --- | --- | --- |
| priority taxa | SILVA database reference sequence identifiers | number of samples |
| *Listeria* | MJSU01000018.28598.30158 | 6 |
| *Listeria* | JNFA01000014.181.1733 | 1 |
| *Clostridium perfringens* | FM865913.1.1506 | 6 |
| *Clostridium butyricum* | CP016087.5908981.5910481 | 18 |
| *Clostridium butyricum* | CP016332.3479095.3480607 | 25 |
| *Clostridium butyricum* | FJ753821.1.1272 | 2 |
| *Clostridium butyricum* | JXBT01000001.134164.135664 | 2 |
| *Clostridium botulinum* (genus: *Clostridium* *sensu stricto* 18) / *Clostridium sporogenes* | MWIY01000025.33.1461 | 3 |
| *Campylobacter* | EU773268.1.1377 | 15 |
| *Campylobacter* | DQ174182.1.1339 | 1 |
| *Campylobacter* | AGYD01000001.688.2187 | 1 |
| *Salmonella* | CP016565.295709.297270 | 67 |
| *Salmonella* | EF604056.1.1485 | 1 |
| *Salmonella* | LZDY01000158.145.1615\* | 237 |
| *Salmonella* | MXOF01000008.49667.51188 | 44 |
| *Salmonella* | MXZI01000185.3731.5252 | 163 |
| *Yersinia* | FN668383.1.1523 | 250 |
| *Yersinia* | KC776764.1.1450 | 250 |
| *Yersinia* | X75278.1.1495 | 7 |

It is essential to regard the reference query sequence identifiers in the context that necessarily only the part of their V4 regions corresponding to the amplified segment (generally 253 base pairs) were used as search queries.

For example, the SILVA sequence MXOF01000008.49667.51188 consists of the segment specified by base-pair coordinates 49,667 - 51,188 of the ENA sequence accession MXOF01000008 - this is the 1,522 bp sequence of the 16S rRNA gene. The amplicon sequence is a 253 bp fragment of this, i.e. only about 17% of the gene length. Indeed, many other sequences in the SILVA database share this exact sequence; the MXOF01000008.49667.51188 has effectively been randomly selected (by us in our analysis) to represent all of these sequences. The ENA record MXOF01000008 is indeed annotated as a *Salmonella bongori* genome, but this exact 253 bp segment occurs in the 16S rRNA gene sequences of numerous other species, including those in other genera; for example, a check of the ENA database in December 2019 indicated that the segment sequence occurs in around four times as many *Enterobacter* database records than *Salmonella*. This is an illustration of the inherent limitations of the V4 sequence and 16S metabarcoding more generally, at least for the Enterobacteriaceae.

Thus, all other parts of the 1,522 bp *S. bongori* 16S rRNA gene, and the wider genome sequence, have necessarily not been observed in the metabarcode data from these samples. The *S. bongori* provenance of the original sequence is assumed to be completely correct, but the reference provenance does not in this case imply the provenance of the experimental metabarcoding sequences which match it.

In contrast, other reference sequences' V4 amplicon segments have much narrower provenance, which is why they are deemed as putative identifiers of a genus, or in some cases a species. For example, among formally named bacteria (i.e. ignoring "uncultured bacterium" etc) in the major sequence databases, the 253-bp segment of FM865913.1.1506 appears to be an exact match only with sequences annotated as *C. perfringens* (312 at the time we compiled the data; at that time, 17 non-binomially named species such as "*Clostridium* sp. AG07-4", "*Clostridium* sp. H2-3", "*Clostridium* sp. AL05-11", etc. were also present as annotations, only once each).

### Detailed analysis of *Salmonella* matches in within-expiry date samples

#### Introduction

When we reported the results of the priority taxa sequence analysis in December 2019, the Food Standards Agency requested us to further investigate the data of 20 specified food samples whose metabarcodes matched at least one of the reference *Salmonella* sequences. These were mostly defined by the specified expiry date, which had not yet passed.

The matching criteria for the purpose of identifying samples was 100% identity over an alignment of at least 250 bp, and not insisting on both the partner reads of a pair matching; i.e. the 'fwd' and 'rev' columns were permitted to flag a positive independently (Table B1).

19 of the 20 samples also fulfilled the 253-bp alignment criterion. We also examined the within-sample frequencies (i.e. read-counts).

Even using only the strictest criteria, many other samples samples matched at least one *Salmonella* reference (Table B1), but had either no specified expiry date, or the expiry data had already passed by that time.

#### Summary of results

Four different *Salmonella* references were implicated in the matches for these 20 samples. We analysed these in detail and determined that in fact none of these particular reference sequences' 253-bp amplicon regions are unique to *Salmonella*, but also occur in the V4 regions of several other Enterobacteriaceae genera, including some **not** on the priority list. These are therefore ambiguous and do not constitute clear evidence that a *Salmonella* species is the origin of the sampled DNA. As a result, no further analysis of these samples was requested.

Furthermore, in sequence databases generally, three of these four 253-bp sequences occur in many more database records for other genera than for *Salmonella*.

#### Relevance to other samples matching Salmonella reference sequences

As noted (Table B4), only five different reference sequences derived from *Salmonella-*annotated sequences account for all 375 samples involving an exact match. In fact, the four investigated references, now concluded to be ambiguous, account for 374 of those.

The fifth reference sequence (EF604056.1.1485) exactly matched a single sample only (#2664680, tomatoes; expiry date 2-10-2019, prior to completion of the sequence analysis; 7 read pairs matched exactly). The 253-bp amplicon segment of this appears slightly more convincing in terms of *Salmonella* provenance, but occurs in very few sequences in any public databases (for any species, *Salmonella* or otherwise); these include *Salmonella enterica* subsp. *houtenae* serovar Houten strain NCTC10401, and at least one unnamed bacterium ("uncultured bacterium"). Furthermore, this 253-bp reference sequence differs by a single base pair from a reference sequence which occurs in numerous other species (including other *Salmonella,* including 6 instances elsewhere in the same strain's genome). It must be assumed that it may be a genuine and correct sequence unique to *Salmonella* (albeit a sequencing error could be another cause), but unlike the others has quite weak weight of evidence in terms of database incidence.

#### Detailed results of the analysis of the matches in 20 within-expiry samples

The detailed analysis was motivated by a request from the Food Standards Agency as a result of putative evidence of the presence of a notifiable food-borne pathogen in samples from within-expiry batches. Irrespective of whether the outcome was establishment (at the DNA sequence evidence level) of positives or not, it is therefore appropriate to include full details of the sequence accessions, names of databases and dates on which they were searched.

As defined by use-by date or otherwise, there were 20 samples to investigate, as per the FSA's request. These all had at least one sequence read (either forward or reverse alone) which was marked ‘100%’ in comparison to the V4 region of one of the SILVA reference sequences. Some samples match more than one reference sequence.

Among these 20 samples, **only four different reference sequences** were involved, in terms of being identical over the **full 253 bp** length to one or more reads in one or more samples, referred to here as 1 to 4. The IDs in the SILVA database were:

1. CP016565.295709.297270
2. LZDY01000158.145.1615
3. MXZI01000185.3731.5252
4. MXOF01000008.49667.51188

Each of those sequences represents the full-length (roughly 1,500 bp), or most of the length, of a 16S rDNA gene sequence. Only a relatively small segment (253 bp), referred to here as the **V4 region**, is relevant to metabarcoding. Table B5 shows the incidence in each sample of metabarcoding sequences **identical** to one or more of the four V4 sequences, over the **full V4 amplicon region length of 253 bp**, i.e. **exact matches.**

Table B5. Numbers of amplicon sequence reads (*forward*; *reverse* – this refers to a pair of reads), in each of the 20 samples, which are exact matches to the V4 region of each of the four SILVA reference sequences (i.e. 100% identical over the full 253 bp length). Food types: ic, ice cream; ob, olives in brine; ra, raisins; oj, pasteurised orange juice; so, soya milk unsweetened. Y = at least 1 read present. \*Sample 2664571 is included here even though it has no perfect matches to any of these sequences. It makes one 100% identity match over a length of 252 bp with one of the sequences (sequence 3). Therefore, in terms of literally perfect matches, this sample is negative for all reference sequences derived from the *Salmonella* genus-annotated records in SILVA.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sample ID | Food Type | (1) CP016565.  295709.  297270 | | (2) LZDY01000158  145.1615 | | (3) MXZI01000185  .3731.5252 | | (4) MXOF01000008.  49667.51188 | |
|  |  | *fwd* | *rev* | *fwd* | *rev* | *fwd* | *rev* | *fwd* | *rev* |
| 2672480 | ic | 2 | 2 | 136 | 80 | 641 | 398 | - | - |
| 2672583 | ic | 1 | 2 | 47 | 30 | 1 | 0 | - | - |
| 2672648 | ic | - | - | 6 | 5 | - | - | - | - |
| 343470 | ic | - | - | 80 | 28 | - | - | - | - |
| 540915 | ic | - | - | 49 | 34 | 34 | 26 | - | - |
| 6412 | ic | 1 | 2 | 244 | 113 | 105 | 51 | 1 | 0 |
| 6413 | ic | 34 | 22 | 93 | 68 | - | - | - | - |
| 2664555 | ob | - |  | - | - | 1 | 1 | - | - |
| 2664583 | ob | 2 | 11 | - | - | 6,276 | 3,914 | - | - |
| 2664701 | ob | 2 | 1 | - | - | - | - | - | - |
| 2664782 | ob | - | - | - | - | 0 | 1 | - | - |
| 2672534 | ob | - | - | - | - | 21 | 14 | - | - |
| 2664571\* | ra | - | - | - | - | - | - | - | - |
| 2664592 | ra | - | - | - | - | 10 | 4 | - | - |
| 2664717 | ra | - | - | - | - | 1 | 1 | - | - |
| 6228 | oj | 1 | 1 | - | - | 17 | 10 | 172 | 80 |
| 6386 | oj | 2 | 0 | - | - | 32 | 11 | 375 | 131 |
| 6655 | oj | 56 | 18 | - | - | 80 | 23 | 1,744 | 602 |
| 343616 | so | - | - | - | - | 7 | 4 | - | - |
| 343617 | so | - | - | - | - | 12 | 8 | - | - |

The identifiers ("accessions") in the SILVA reference database originate from the source data that was incorporated into the database by the curators. Usually the first part of the identifier is the same as that of a record in the primary sequence database (ENA; accessions are the same in NCBI and DDBJ databases). The numbers which follow it can be assumed to represent coordinates of the reference sequence.

For example, CP016565.295709.297270 indicates that the sequence in this record corresponds to the nucleotides spanning coordinates 295,709 to 297,270 of the ENA/NCBI/DDBJ sequence with ID ‘CP016565’, i.e. 1,562 nucleotides, the whole 16S rRNA gene. The V4 region which would be amplified corresponds to nucleotides 544-796 of this gene sequence.

The (assumed) origin of each of the four sequences are as follows:

1. “*Salmonella enterica* subsp. *enterica* serovar Heidelberg strain AMR588-04-00318, complete sequence” <https://www.ncbi.nlm.nih.gov/nuccore/CP016565>
2. “*Escherichia coli* strain EPEC 1316 EPEC-1316\_contig\_95, whole genome shotgun sequence” <https://www.ncbi.nlm.nih.gov/nuccore/LZDY01000158>
3. “*Salmonella enterica* subsp. *salamae* serovar 55:k:z39 strain BCW\_2788 NODE\_185\_length\_5400\_cov\_4.98082, whole genome shotgun sequence” <https://www.ncbi.nlm.nih.gov/nuccore/MXZI01000185>
4. “*Salmonella bongori* serovar 44:z39:- strain BCW\_1554 NODE\_8\_length\_172755\_cov\_0.434789, whole genome shotgun sequence” <https://www.ncbi.nlm.nih.gov/nuccore/MXOF01000008>

These all appear to be from genome sequencing projects, including (2) which is from an *E. coli* strain. There is therefore a discrepancy between the two databases.

The SILVA curators classify external sequences to genus level, and retain the original species (and finer-level, if present) annotation, **irrespective** of whether this agrees or disagrees with their classification.

Thus, the annotation of the SILVA reference sequence (2) is:

>LZDY01000158.145.1615 Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Salmonella;Escherichia coli

We observe that the *E. coli* classification of (2) appears to be of sound provenance – the genome sequence is associated with a publication named “*Virulence, Antimicrobial resistance properties and phylogenetic background of E. coli O157 isolates that do not carry the H7 flagellar antigen*” (see hyperlink above).

Irrespective of the above, it was worth checking each of the four sequences (V4 253-bp segment only, since that is what occured in the food sample metabarcode sequences) against sequences of closely-related genera, in third party databases (and also double-checking within SILVA itself).

Table B6 shows the number of **exact** (completely identical over 253 bp) matches which each of those four sequences make to sequences annotated as members of *Salmonella, Klebsiella, Escherichia, Shigella, Enterobacter, Raoultella*, in the NCBI nonredundant nucleotide sequence database (‘nt’) and in SILVA release 132 itself. Note that in SILVA, *Escherichia* and *Shigella* are collectively treated as a single genus.

The crucial point is that **according to the annotations in both NCBI nt and SILVA, *none* of the 253-bp reference sequences are unique to *Salmonella*.** (According to the NCBI annotations, the four sequences occur respectively in four to six different genera (of those considered here), and occur in two, four or five different genera in SILVA.)

**That is sufficient to conclude that no sequences uniquely identifiable as originating from *Salmonella* were present in the metabarcode DNA sequence data in any of these food samples.**

Also in terms of database records, sequence (2) is far more common in *Escherichia* (or *Escherichia-Shigella*); (3) likewise in *Klebsiella* (also in *Enterobacter* in SILVA) and (4) is most common in sequences from *Enterobacter* in nt (but in *Salmonella* in SILVA). That does not necessarily mean “more biologically common”, nor indicate that the SILVA annotations for sequences 2, 3, 4 are incorrect. Note though that for (2), these results are consistent with the origin of this reference sequence (*E. coli* genome).

Sequence (1) alone is more commonly associated with *Salmonella* 16S sequences than any other genus, according to both databases. Although, as shown above, this does not mean that it could not have arisen from non-*Salmonella* bacteria, it is worth noting the incidence of this reference sequence’s matches in Table B5.

This sequence (1) had exact matches in 9 of the 20 samples, in all cases in very low read counts, and at trace levels in 7 cases. It is possible that these trace reads would be removed by normal quality-control methods (as has been emphasised, the priority taxon-detection had the aim of being a liberal search to minimise false negatives). Indeed, only four of the 9 samples have at least one completely-accordant read **pair** which make an exact match, with the numbers of such pairs being 20, 18, 1, 1 (not shown in the table).

It cannot be proven or disproven from the available data alone, but it also seems likely that in some samples, some or all of the trace-levels of reads matching sequence (1) result from sequencing errors, given that an almost identical sequence (3) is present in much larger counts. Table B7 shows the differences between each of the four reference sequences; these two references differ by only a single bp. This does not apply to all samples, however.

The facts that sequences (1) and (3) differ in only one position, that sequence (1) is most commonly associated in the databases with *Salmonella* and (3) with *Klebsiella* (or *Klebsiella* and *Enterobacter* according to SILVA) further reinforces the high levels of similarity between these Enterobacteriaceae genera, and that the ‘database-majority’ genus cannot be assumed to be the probable origin of an observed instance of the 16S V4 sequence.

#### Conclusion

For all of the samples concerned, the matches with the four reference sequences neither ruled out nor confirmed that DNA from members of one or more genera (including *Salmonella*) of the family Enterobacteriaceae was present.

**Table B6.** The number of sequences present in the NCBI nt and SILVA databases which are annotated as the named genus and which exactly match (over the full 253 bp V4 length) the 253 bp V4 region of each of the four SILVA reference sequences. The NCBI nt matches refer to that database on 21st Dec 2019. SILVA refers to release 132.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | (1) CP016565.  295709.  297270 | (2) LZDY01000158  145.1615 | (3) MXZI01000185  .3731.5252 | (4) MXOF01000008.  49667.51188 |
| NCBI nt *Salmonella* | 1,767 | 20 | 20 | 25 |
| NCBI nt  *Klebsiella* | 23 | 12 | 923 | 4 |
| NCBI nt  *Escherichia* | 2 | 4,633 | 0 | 30 |
| NCBI nt  *Shigella* | 0 | 587 | 0 | 0 |
| NCBI nt  *Enterobacter* | 103 | 6 | 99 | 102 |
| NCBI nt  *Raoultella* | 0 | 1 | 3 | 0 |
| SILVA *Salmonella* | 10,557 | 14 | 8 | 38 |
| SILVA *Klebsiella* | 11 | 0 | 1,718 | 0 |
| SILVA *Escherichia-Shigella* | 5 | 23,997 | 4 | 0 |
| SILVA *Enterobacter* | 43 | 0 | 1,275 | 2 |
| SILVA *Raoultella* | 0 | 0 | 1 | 0 |

**Table B7.** Number of non-identical positions in the 253 bp V4 region of each of the four reference sequences.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1. CP016565.   295709.  297270 | (2) LZDY01000158  145.1615 | (3) MXZI01000185  .3731.5252 | (4) MXOF01000008.  49667.51188 |
| 1. CP016565.   295709.  297270 |  | 9 | 1 | 3 |
| (2) LZDY01000158  145.1615 | 9 |  | 10 | 6 |
| (3)MXZI01000185  .3731.5252 | 1 | 10 |  | 4 |
| (4) MXOF01000008.  49667.51188 | 3 | 6 | 4 |  |