# Appendix 4b: Priority Taxon Detection and Identification

## Determination of in-principle distinguishable taxa.

We determined which priority taxa are likely to be distinguishable by 16S rRNA V4 sequences.

We first systematically examined the **reference database sequences** corresponding to the taxa in the FSA's priority list, to determine which are represented only by unique 16S V4 sequences. There are various uncertainties surrounding this treatment, due to the presence of many reference sequences whose original organism name is unclassified ("uncultured bacterium" etc) and others which appear to be of unconvincing provenance. Species names of a non-binomial form (e.g. "*Campylobacter* sp. RM8964") also complicate. Methodology is described further in Appendix 1A.

The results are described in Appendix 1B. Briefly, these are that potentially uniquely identifiable species are *Clostridium perfringens;* a particular *Clostridium botulinum* group (includes B and E toxin types) which is a member of the*Clostridium sensu stricto 1* genus (in the reference SILVA database taxonomy); and *Clostridium butyricum*. 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.

Additionally of interest is a reference V4 sequence which is identical between one *C. botulinum* group (includes annotated A, B and F toxin types) and *C. sporogenes* (both in *Clostridium sensu stricto* 18); although the two species cannot be distinguished from each other, both are in the priority taxa list and the V4 sequence is not shared with named non-listed taxa. This therefore represents a finer taxonomic resolution than is possible using standard methodology (3.3.1), which would classify such sequences only to genus level.

To place the above in context, we note that the taxonomy of the traditionally-named genus *Clostridium* is complex and has been subject to various reclassifications in a number of taxonomies, including the SILVA database (thus the numbered *'sensu stricto'* genus names). So too is the taxonomy of *C. botulinum,* which in effect represents several different species defined largely by toxin type, which have been placed in different genera in the SILVA classification.

In contrast to the above few cases, the following species in the FSA priority list cannot be distinguished from one or more non-priority taxa in the same genus: *Klebsiella pneumoniae, Staphylococcus aureus, Enterococcus faecalis* and *Enterococcus faecium, Clostridium baratii* and the *Clostridium botulinum* C, D, C/D toxin type group (genus *Clostridium sensu stricto* 7). We nonetheless reported on the samples containing metabarcode sequences which match those taxa, while noting that limitation.

Further, the priority list includes several taxa specified at the genus level: *Listeria, Campylobacter, Salmonella, Escherichia-Shigella.* Some of the genera of the family Enterobacteriaceae are closely related, reflected by the fact that the two traditional genera *Escherichia* and *Shigella* are merged into a single genus in SILVA. This family also contains *Salmonella* and *Klebsiella*. In bacteria generally, many V4 sequences are confined to one genus but there are numerous exceptions, which we examined on a case by case basis only if they arose as positives in the samples.

## Frequency of matching sequences in samples

Full details of the frequencies of matches between one or more reads of each sample and any of the reference sequences of these taxa, are presented in Appendix 1B.

These matches must be viewed in the dual context of (i) how discriminating the information provided would be even if an exact sequence match arose (refer to previous subsection) and (ii) how close to exact the sequence match is. The matches are manifest as pairwise sequence alignments, which will not necessarily span the full length of the amplicon sequences (which is usually 253 bp). The alignment is also characterised by the percentage of positions which are identical between the two sequences (reference and sample metabarcode), i.e. the % identity. Only alignments of at least 250 bp are considered.

We present results where identity is a minimum of 98% (Appendix 1B Table B1). Within the 98%, 250 bp threshold, the great majority of the 1,001 samples match at least one of the priority taxa reference sequences, in terms of at least 1 read/read pair.

The highest frequency of alignments occurs with sequences of the *Escherichia-Shigella* genus, with more than 800 samples passing this threshold.

We note that this threshold is by nature passed even where as many as 5 bp of a 250 bp alignment differ. Further, an alignment of this length will mean that several bp have been omitted from it (because they do not match); in total, 8 bp may fail to match from a full amplicon length of 253 bp. **These criteria are considered very liberal.**

As stringency is increased by raising the % identity and length threshold, the numbers of samples matching the query taxa reference sequences falls, dramatically for some taxa. The most stringent criteria are 100% sequence identity over 253 bp. We consider this to be the appropriate threshold for production data, but the range of stringencies provides useful context.

For example, using the least stringent criteria, 172 samples match *Listeria* references, but the strictest measures reduce this to 7 samples. There are several taxa with matching samples at the 98%/250bp level which match none at the two strictest thresholds. *Escherichia-Shigella* references still match more than 600 samples, however. Refer to the next section for details of *Salmonella*.

## More detailed analysis of selected taxa

In response to the results (3.3.2.2), the FSA requested that the sequence data of samples that exhibited amplicon matches with one taxon of concern (*Salmonella*) and which had not yet passed their expiry date at the time (late December 2019), be examined on a case-by-case basis. The full details are in Appendix 1B.

Using the most stringent matching criteria (100% sequence identity over 253 bp for both reads), 395 samples matched at least one of the *Salmonella*-derived V4 reference regions. Of these, 20 samples were still within the expiry date. Only four different references, i.e. four unique 253-bp sequences, were involved. We analysed these in detail and determined that in fact none of these four particular sequences are unique to *Salmonella*, but also occur in the V4 regions of several other Enterobacteriaceae genera, including some not on the priority list. As a result, no further analysis of these samples was requested.

In fact, these four reference sequences account for all but one of the 395 samples' matches with *Salmonella*-derived V4 references, illustrating some of the limitations of taxonomic resolution by 16S metabarcoding; especially involving closely-related genera such as in this particular bacterial family (where alternative genes may be more appropriate for this purpose).

## First screen by MetaPhlAn3

Only four of the priority taxa of concern (TOCs) were found to have positive sequences by this method, which uses unassembled short reads as input: *Enterococcus faecalis, Enterococcus faecium, Escherichia coli* and *Klebsiella pneumoniae*. Seventeen of the samples were positive for one or more of these. Only three of these had expiry dates which had not yet lapsed, but one of these was due to expire imminently as we commenced this analysis, and was therefore excluded from further analysis. That left two samples for further investigation by the subsequent steps, both vanilla ice cream, due to expire in 2021 (respectively May and August): #2672480 and #6412. Refer to Appendix 4 for more details of this stage of the analysis.

## Identification of MLST sequences

Within the assembled metagenomic sequences of the ice cream samples, only #2672480 contained identifiable (perfectly-matching segments of) MLST sequences of *K. pneumoniae.* This sample contained no other MLST sequences.

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

Further details are in Appendix 4.

## Sequence similarity search of NCBI nt

The BlastN search confirmed our findings regarding the positive taxa inferred from MLST sequences. We found no evidence for additional species of concern in sample #2672480.

In #6412, we additionally found clear evidence supporting the presence of DNA from *Enterococcus faecalis, Enterococcus faecium, Escherichia* *coli* and *Klebsiella pneumoniae;* i.e. independent confirmation at the sequence level of the positives identified by MetaPhlAn3.

However, we also found convincing evidence in #6412 of a sequence uniquely identifying *Yersinia enterocolitica,* along with several other sequences matching this species (but which appeared to be less convincing). Details are in Appendix 4. This species was negative by all the other sequence analysis methods.