# Appendix 1A: The reference 16S rRNA gene sequence database

## Purpose of the database

The reference 16S rRNA gene sequence database is relevant to analysis of the 16S metabarcode sequence data obtained from 1,001 samples, in both of these contexts:

1. The Microbial Community Analysis (section 2.5.1 of the main Report), which identifies the sequences in each sample which are of bacterial origin (and in the process determines the likely bacterial taxonomy of each sequence, to one level or another), as opposed to a food-organism DNA origin.
2. The targeted search for DNA sequences of Priority Taxa (foodborne pathogens of concern; section 2.5.2 of main Report).

The reference database used was SILVA (release 132).

## Limitations of the database

By the inherent nature of 16S rRNA genes, no reference database can be expected to be able to identify the detailed origin of 16S sequences obtained from metabarcoding.

Key aspects are as follows, expanded on in the later sections:

* 16S rRNA V4-region sequences are generally classifiable to genus level but no further.
* The reference database we use for 16S rRNA gene sequences is SILVA SSU.
* SILVA SSU bacterial sequences are sourced by the SILVA maintainers from the European Nucleotide Archive (ENA) database.
* The SILVA curators subject the sequences to intensive analysis, to classify them taxonomically; this is generally to genus level, as expected.

### Standard metabarcode-classification methods

Most taxonomic-assignment methods for 16S metabarcode sequence reads (context 1 above; such as the QIIME2 pipeline, which we used) involve identification to the finest level at which the taxon is unique. For example if a metabarcode sequence occurs in multiple species within one genus but no other genera, then a genus-level assignment is possible, but no species assignment can be made. That is the case whether the metabarcode sequence narrows down the matching taxa to 1, 2, 10 or 100 different species.

### Targeted approaches for small lists of taxa of interest

For the purpose of identifying priority taxa of concern, however (context 2 above), a more detailed treatment can be appropriate. 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.

Therefore, **where reliable species-level annotations are available in the reference database sequences themselves,** it is worth determining which species (or groups of species) can be identified, and checking for the presence of the corresponding sequences in the metabarcode data obtained from the food samples.

However, this is not straightforward. Considerations are as follows.

* The sequence records in the original database (ENA) have "organism name" annotations for each sequence.
* Many of these specify a named species; others are (much) less specific; e.g. "unclassified bacterium").
* The SILVA classifications of the ENA-derived sequences amount to a re-classification in many cases (differ from the annotation in the originally deposited sequences in ENA).
* The original organism name (effectively, species name; but can be less specific) is however retained by the SILVA maintainers (even if this conflicts with the genus to which they have classified the sequence).
* **The original organism name is therefore the only statement** (correct or otherwise) **of the species from which the reference sequence originated** (in those cases where it is itself to species level); given the above points, **it must be treated with caution.**
* Where an identical V4 sequence occurs with the same original organism (species) name multiple times in SILVA, this lends credence to the organism name.
* There are however many cases of **single instances** of an original organism name for a particular V4 sequence; these are by nature less convincing.
* The **single name**-instance cases are still less credible when additionally, **multiple instances of an identical V4 sequence occur in the database with a different original organism name**, especially where these are taxonomically at odds dramatically with the single-instance name.

These points are expanded in the following sections and in Appendix 1B.

## The reference database and variable region (V4) sequences therein

Reference 16S rRNA gene sequences ("16S sequences") have been obtained from the SILVA database, release 132. It is important to note that the same 16S sequence can occur multiple times in the reference database. More importantly, many reference sequences can be **identical in the metabarcode sequence** (the amplified segment of the V4 variable region of the 16S gene) **even if their full-length 16S sequences are different**.

Therefore there are many cases of groups of different, closely-related species which are indistinguishable using this metabarcoding technique, because they have identical V4 segment regions.

Conversely, it is possible for different strains of a single species to differ slightly in their V4 sequence. However, it appears likely that many such cases in the reference database are the result of sequencing errors producing a variant which has been incorporated into the database (they only occur once). In other cases, by sheer weight of evidence (many instances of the variant sequence in the reference database) it is assumed that these are genuine.

## Taxonomic annotations by third parties and by the database curators

The SILVA database curators obtain sequence data from an external database, the European Nucleotide Archive (ENA), along with the organism names which are present in those database records. These third-party annotations are notionally at the species level but can be very unspecific – e.g. ‘uncultured bacterium’ is extremely common. At the other extreme, some third-party sequences are annotated at a much finer level than species, while simple species names are also very common.

It is crucial to note that the SILVA curators (re-)classify the sequences at the *genus level and above*; the original annotations at species-level and below appear in the SILVA database *unchanged*, even if they conflict with the SILVA genus classification.

Thus, when we mention apparent ‘mis-annotation’, we refer to the original, third-party annotations of the sequences. In all cases, we take the **genus-level** annotation by the SILVA curators to be ground truth.

## The interpretation of minority and high-frequency taxonomic annotations of 16S rRNA gene sequences

### Minority annotations contrasting with high-frequency annotations

#### Example 1: A database annotation overwhelmingly outweighed by other evidence in the same database

As an example, one particular V4 sequence occurs 73,559 times in SILVA. 72% of these identical sequences were annotated with an organism name of 'uncultured bacterium' or 'uncultured organism' in the original ENA database, while 24% were annotated as '*Staphylococcus* *aureus*' (one of the named species of concern in this study). 1.2% (885 sequences) were annotated as '*Staphylococcus* *epidermidis*', with several other *Staphylococcus* species occurring more than 100 times (*S. haemolyticus, S. warneri, S. hominis, S. pasteuri, S. capitis*, with numerous other named species at < 0.1%, which is still tens of instances. These can be considered to be **high-frequency annotations**, and it is reasonable to assume that they are all correct.

A single instance of this identical V4 sequence is also present which was originally annotated as '*Yersinia* *pestis*'. It is fair to assume that this is an incorrect original annotation (*Yersinia* belongs to a different phylum to *Staphylococcus*).

Therefore, if an amplicon sequence occurring in a food sample microbiome were identical to the above V4 sequence, it can be concluded that it originated from a *Staphylococcus* species, but could be from any of quite a large number of different species. It would however be consistent with, but not demonstrative of, the presence of *S. aureus*. This illustrates the genus-level limit (in this particular case). In contrast, it would not be viewed as consistent with *Y. pestis*, even though it is identical to a sequence originally annotated as such.

#### Example 2: lack of expected context casts additional doubt on a minority annotation

A less clear-cut example occurs in the one of the SILVA database's *Clostridium* genera. Some essential background information is that the traditional genus of this name is generally thought to be inconsistent and inappropriate, with reclassifications since resulting. In SILVA, there are numerous genera with names '*Clostridium sensu stricto* *n',* containing species retaining the *Clostridium* genus name*.* The traditional species name *C. botulinum* is itself held to encapsulate several true species instead of just one. For important medical reasons, *C. botulinum* strains are traditionally characterised by their toxin type (which has a strong relationship to the notionally separate species).

The following example concerns genus *Clostridium* *sensu stricto* 3. A particular V4 sequence is present in the database 14 times, with 8 of these originally annotated as 'uncultured bacterium' or similar, while four instances were annotated as '*Clostridium* *amylolyticum*'. The other two instances were originally annotated as '*Clostridium* *polynesiense*' and '*Clostridium* *botulinum*'. The *C. botulinum* -annotation is treated with scepticism, not only because it is the only instance for this V4 sequence, but also the only instance in this *sensu stricto* genus (unlike *C. polynesiense*), and also lacks any more detailed information (such as toxin-type or strain name). Also it is in effect in the wrong genus, since all of the taxon-type characterised *C. botulinum* sequences are in the three genera *Clostridium* *sensu stricto* 1, *Clostridium* *sensu stricto* 7 and *Clostridium* *sensu stricto* 18.

Therefore, the *C. botulinum* annotation of this reference sequence is treated as unreliable. Any sample amplicon sequence which exactly matched the above V4 reference would not be treated as genuinely consistent with the presence of *C. botulinum*, a priority taxon of concern. It would be considered to most likely originate from either *C. amylolyticum* or *C. polynesiense*.

### Some species are represented by unique V4 sequences

An example of a taxon of concern which appears generally identifiable to species level is *Clostridium butyricum*. This occurs as an original organism name-annotation at a high frequency for several V4 sequences. For example, one sequence occurs 247 times in SILVA (classified to genus *Clostridium* sensu stricto 1). The majority of these were unidentified in the original annotation ('uncultured bacterium' or similar), but 110 cases were '*Clostridium* *butyricum*'. This is the only binomial name which occurs for this sequence. There were however 6 other original species-level annotations, but these each occur only once, and have names such as '*Clostridium* sp. IODB-O3', '*Clostridium* sp. IBUN62F', '*Clostridium* sp. SH012', etc. Without further information, it might be assumed that these unnamed species are in fact *C. butyricum*, although this would not be certain without further investigation. A second V4 sequence occurring 101 times in the database had an original organism annotation of '*Clostridium* *butyricum*' in the majority of cases (58). Again, it is the only binomially-named species which occurs, but there are four non-binomially named species ('*Clostridium* sp. SP3' etc) which again only occur once each. There are further V4 sequences where '*Clostridium butyricum*' occurs more than once and is the only binomial name, albeit these constitute only 2 of 4, or 3 of 4 total instances of the particular sequence.

Therefore if an amplicon sequence matched any of the above sequences exactly, it would be reasonable to treat this as positive evidence for the presence of DNA originating from *C. butyricum* (not merely evidence consistent with it). Although '*Clostridium* sp. IODB-O3' etc might in fact be a distinct species, a safer assumption might be that this is not the case, but it depends on the provenance. Many of the organisms with this form of name were themselves likely first identified by metagenomics or 16S amplicon sequencing (while many are isolates which have not yet been formally identified/named).

### Single-instance unique V4 sequences are commonplace in the database

Strictly speaking, there are many species which are represented by unique V4 sequence(s). That is because there are many sequences which occur only once in the database (and thus necessarily have a 100% consensus annotation). In real terms, many of these will be genuine - a sequence variant which has been accurately sequenced, but observed only once. However, a great many are likely to be the result of sequencing error(s), creating an artefactual unique sequence. Both causes will result in sequences near-identical to other sequences present in higher frequency. In general, such single-instance sequences cannot be assumed to be reliable, and therefore any matches in a metabarcode data set would be treated with caution.