**Identification of secondary sequences in the MMETSP dataset (preliminary and ongoing)**

**Zhenfeng Liu (Caron lab)**

zhenfeng.liu1@gmail.com

**Background**

 We (Sarah Hu and I at Caron lab) have been using MMETSP dataset as reference for taxonomic assignments in our metatranscriptomic analyses. We have found in several occasions that some sequences in the dataset are not from the organisms listed for specific MMETSP samples. Most of the cultures in MMETSP are non-axenic, and bacterial or algal prey were often added to the cultures, which could be the primary source of these sequences. The presence of these sequences from other organisms would often lead to incorrect taxonomic assignments. I conducted a search for these sequences with the intention to remove them from the database.

**Method**

**a) Starting point**

It is known that the original MMETSP assembly produced by NCGR was not optimal. Lisa Cohen from Titus Brown’s lab has conducted a re-assembly of the entire MMETSP dataset (<https://monsterbashseq.wordpress.com/2016/09/13/mmetsp-re-assemblies/>). All protein sequences from her assembly were downloaded. Additional protein sequences from other genome/transcriptome projects were added (see attached file for a list of these organisms). This is for adding more diversity to the database, and is largely unrelated to this effort but I thought it should be noted anyway. Sequences of the same genus were combined and cdhit was used to reduce redundancy (>95% similarity) in all genera.

**b) Search for bacterial sequences**

All sequences were searched against Uniref90 (release September 19, 2017) using diamond aligner (with --sensitive setting, bit score cutoff of 100). A script was used to find all sequences whose best hit in the database is more than 90% similar to a bacterial sequence. The source (MMETSP ID) of these sequences and the taxonomy of the corresponding bacterial sequences were analyzed.

**c) Search for protistan secondary sequences**

An all-vs-all protein search was conducted against the entire protein dataset itself using diamond with the same cutoffs. Only hits with more than 90% similarity were kept. Hits between sequences from the same phylum (taxonomy information can be found in attached file) were not further analyzed. Highly similar sequences from samples of different phyla were counted.

**Findings**

**a) Bacterial sequences**

The assumption here is a sequence is probably of bacterial origin if its best hit in the Uniref90 database is a bacterial sequence and it is more than 90% similar to that (since Uniref90 is generated by clustering all Uniref sequences at 90% similarity). Under this assumption, I found 21 MMETSP samples with more than 3% bacterial sequences (Table 1, a more complete list is in the attached files). Many of the bacterial sources are known marine bacteria, but strangely, Pelomonas, which is not known to be found in marine environments, were found to be in many samples. About half of these samples were of Heterosigma akashiwo, but others have no relation with H. akashiwo, and were produced by different labs. Not all H. akashiwo samples have this Pelomonas. Furthermore, these Pelomonas sequences were often found in many samples (on average 5.5 samples). Later, I would also find that these samples with Pelomonas all share a large number of sequences with each other (see Fig. 1B below). Therefore, it is highly likely these Pelomonas sequences are from a systemic contamination at NCGR rather than native bacteria in the cultures.

**b) Protistan secondary sequences**

It is unlikely for two organisms from different phyla to have a large number of highly similar (>90%, usually found in organisms from the same genus) sequences. Among more than 230,000 possible cross-phylum sample pairs in our dataset, only 1% had more than 15 proteins shared between samples, and only <0.1% had more than 250 proteins shared between samples. The more proteins two samples from different phyla shared, the more likely a secondary source existed in either or both of the samples. I started my analysis from those sample pairs (172 pairs) that share more than 250 proteins, since they are the most “suspicious” and they had the most secondary sequences. Identifying these sample pairs is not enough, because a large number of shared proteins between samples could be caused by culture of organism A containing organism B, or the reverse, or both cultures of A and B containing organism C. The actual cause should determine which sequences should be removed from the database. However, it is difficult to bioinformatically determine the cause. For each sample pairs, I searched the MMETSP metadata and literature to try to find a documented predator/prey or symbiosis relationship to determine the likely cause of shared proteins, with some success. The findings are summarized and visualized in Fig. 1 (for table of actual data, see attached file).

 Fig. 1A shows several known predator/prey relationship, for example, Favella, Oxyrrhis, and Strombidinopsis all feed on Isochrysis, and therefore the connection between the three; Strombidinopsis and Mesodinium feed on Heterocapsa; Strombidinopsis and Tiarina feed on Rhodomonas. These are all documented in MMETSP metadata. However, the data also indicates other undocumented relationships. MMETSP0851 Psuedo-nitzschia likely had dinoflagellate(s) in it. MMETSP0472 Tiarina likely had some dinoflagellate(s) in it as well, given its similarities to Alexandrium and Lingulodinium.

 In Fig. 1B, I see 13 samples that are interconnected to each other (marked by yellow). These are the same samples that were found to contain Pelomonas bacterial sequences (Table 1). This finding further suggests that these samples share a same secondary source, and given how unrelated some of these samples are to each other, it is likely that bacterial source was originated at NCGR. Fig. 1B also shows several known connections between samples such as Oxyrrhis feeding on Heterosigma, Sorites harboring a Symbiodinium symbiont.

 Fig. 1C shows the rest of these sample pairs. It suggests a previously undocumented dinoflagellate contamination likely existed in MMETSP1098 Spumella. An undocumented Micromonas likely existed in MMETSP0384 Alexandrium.

**Future work and challenges**

 Though I was able to determine (or with educated guess) the cause of shared proteins between many of these sample pairs, I still don’t know how to explain relationship between many sample pairs. These might be solved based on knowledge of those samples. I started on sample pairs that shared more than 250 proteins, but it is likely that sample pairs that shared less than 250 but more than 50 or 100 proteins also had secondary sources. Expanding that would require more work, most of which is manual.

 Another very important limitation of this work is that it only identifies secondary sequences that are already in the database. If a sample contained a secondary source that is poorly represented by the database, it would not be identified by this method. Other methods exist, for example, GC content based separation, but their success and reliability is uncertain. I suggest creating a pipeline to insert future data into the database, which would also use the new data to screen the database for secondary sequences.

Table 1 Samples with more than 3% bacterial sequences and the most frequent bacterial taxa in these samples.



Figure 1. A network view of sample pairs sharing large numbers of proteins. Panels A, B, C are just zoomed in views of the entire network. Each node is a sample, identified by MMETSP sample number and the genus name of the sample. Each edge indicates that more than 250 proteins were shared by the two samples of different phyla at more than 90% similarity. Red edges indicate known preditor/prey or symbiosis relationship. Blue edges indicate that two samples likely shared the same prey. Yellow nodes indicate a set of samples that likely contained a systematic bacterial contamination.

**A**



**B**

****

**C**

****

**Appendix A. Removal of proteins from EukZoo v.0.1 to create EukZoo v.0.2.**

Bacterial sequences and protistan secondary sequences from EukZoo v.0.1 (described in methods: starting point) were removed from the database to create a “cleaner” version, which is v.0.2. All sequences whose best hits in Uniref90 are bacterial sequences and the identity of the hits are more than 90% are considered bacterial sequences, and thus removed from EukZoo. Sample pairs who share more than 250 proteins at more than 90% identity (pictured in Fig. 1, and listed in supplementary tables) were examined manually to find known prey-predator relationships, and to predict contaminated samples. For all known prey-predator relationships (red lines in Fig. 1), shared sequences were removed from the predator sample, but not the prey sample. For others, if a sample is located in a central position in Fig. 1 (for example, 0851 and 0472 in Fig. 1), we consider it is highly likely that the central sample contained contaminations rather than that the samples connected to it had contaminations. Therefore, we only removed shared proteins from those central samples, but not the connected samples. We consider the following samples to be samples that probably contained contamination:

0123, 0470, 0126, 0467, 0851, 0463, 0472, 0191, 0924, 0469, 1098, 0384, 1465, 0798, 0098, 1426, 0449.

Proteins shared with these samples were only removed from these samples, but not others. Proteins shared between samples not listed above were removed from both samples. In total, 303,167 sequences were removed from 10,929,182 sequences of EukZoo v.0.1.