***TaxonSluice: A digital sluice box for taxonomy gene marker analyses***

Ramírez, GA & Garber, AI

Please cite this tool’s official release:

Ramírez, GA, Garber,AI, LecoeuvreA, Wheat, CG and Orcutt, BN. Ecology and Functional Potential of Subseafloor Crustal Biofilms: A MicrobialIncubation and Sequence Data Integration Assessment. ISME J. 2018, (xx): xxx-xxx.

Below we highlight the functionalities and limitations of *TaxonSluice* by providing a mock community example comprised of 3 samples and 3 sample-specific blanks. We stress that our program is not a “silver bullet” against contaminants, however, our algorithm helps investigators make better informed decisions regarding potential contaminants in their data.

**Algorithm rationale overview**

The *TaxonSluice* algorithm (Figure 1) analyzes OTUs present in environmental samples for overlap with i) sample-specific blanks and ii) non-sample-specific blanks. This is not an automated solution against contaminants. Our tool provides a software assisted analysis where the user ultimately decides, based on multiple lines of evidence, if a specific flagged OTU will be retained or discarded. Our key algorithm steps and their rationale are summarized below.

*Sample-specific blank overlap* (Figure 1, criteria 1 & 2)

For every comparison, if overlap between sample and blanks exists, a decision to “flag” the OTU as a potential contaminant is made based on a user-defined proportional abundance threshold. The program default is to flag an OTU if its proportional abundance in the blank is at or above 10% its abundance in the sample. We justify this arbitrary threshold by assuming that, following geometric amplification during PCR, legitimate kit contaminant sequences, present in both sample and blank reactions at equal template amounts, should not diverge in counts by more than an order of magnitude. If an OTU, present in a sample and its blank, has more than an order of magnitude sequence counts in the sample relative to the blank, it is likely that the environment contains closely related lineages to kit contaminants and that these sequences have been artificially merged by the OTU clustering step. Thus, in such a scenario, we would opt to retain this OTU, since it likely represents legitimate environmental data despite its similarity to lineages present in kit reagents. However, since this is not a quantitative assessment, our assumption is imperfect and the user is welcomed to explore other threshold cutoffs. Note: If an OTU is only present in blanks, it is automatically removed from the data set.

*Non-sample-specific blank Overlap* (Figure 1, criteria 3 & 4)

The presence of multiple blanks in a dataset allows higher order comparisons that take into account, in addition to potential reagent contamination, laboratory introduced contaminants. A single sample and blank pair may fail to catch laboratory introduced contamination following technical error (poor aseptic practice, pipetting error, *etc.*) or contaminant of laboratory disposables (microcentrifuge tubes, pipette tips, *etc.*). Therefore, if the environmental sample is inadvertently contaminated and the blank is not; comparisons of OTUs across multiple, non-sample-specific blanks, may catch such laboratory introduced contaminants. *TaxonSluice* checks every OTU not flagged as a contaminant based on its sample-specific blank against every available non-sample-specific blank to take into account laboratory introduced contamination.

*Inspection of closest relatives in the SILVA database*

With the exception of automatic elimination of the OTUs found only in blanks, the **user** is charged with making a final decision on the retention or removal of all flagged OTUs. To inform this decision, *TaxonSluice* outputs the ten closest relatives to each flagged OTU in the SILVA database. Identity, coverage and e-value metrics, in addition to environmental isolation metadata, if available, are also provided by *TaxonSluice*. This information, we hope, will help guide the user in making a retain or remove decision for each flagged OTU in the dataset. Thus, *TaxonSluice*, parses through the bulk of environmental sequence data, and based on 3 independent lines of evidence [i) sample-specific blanks, ii) non-sample-specific blanks and iii) environmental context of close sequence relatives], provides the user as much information as possible to robustly determine retention or removal of taxonomic lineages in the dataset that were also present in sample blanks.

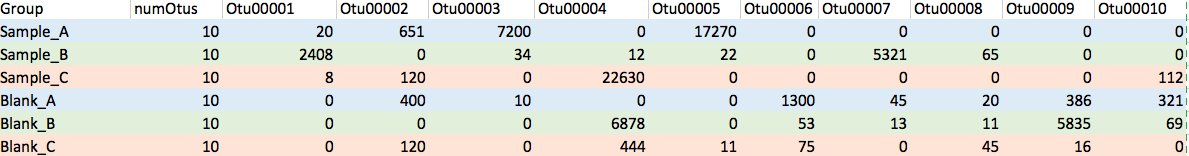
heuristicAlgo.pdf

**Figure 1.** *TaxonSluice* algorithm.

**Mock Community Example**

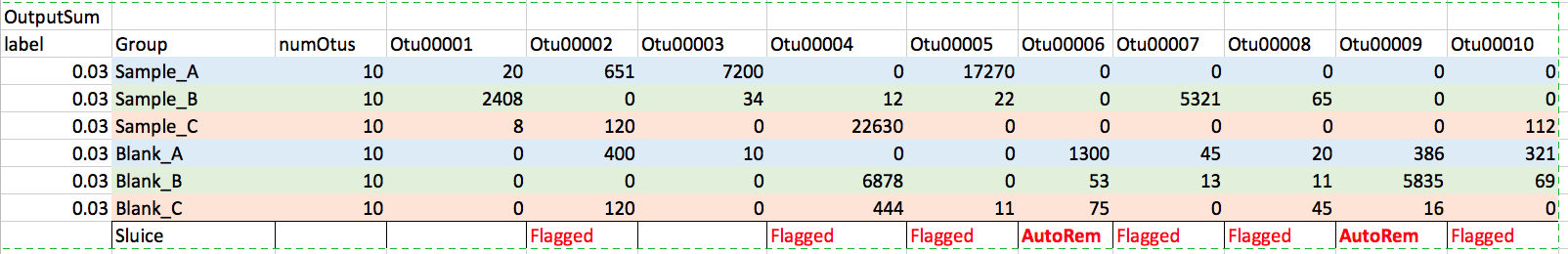
*Note: All code and data files below are found in the TaxonSluice GitHub page Mock Community folder.*

The mock “OTU table” that represents an analyses of a marine sediment environment. The dataset is comprised of three independent environmental samples (Sample A-C) along with 3 sample-specific blanks (Blank A-C). Each blank and sample are paired alphabetically (The bank for Sample\_A is Blank\_A, etc.) and are color-paired below (Table 1).



**Table 1. Mock dataset comprised of 3 independent marine sediment samples.**

Running this dataset through *TaxonSluice* produced the following results summarized below (Table 2). OTUs 00006 and 00009 are automatically removed because they are exclusive to blank samples. OTUs 00002, 00004, 00005, 00007, 00008, and 00010 are flagged as potential contaminants. Lastly, based on three independent criteria, only OTUs 00001 & 00003 in this mock dataset are cleared as very unlikely to represent kit or environmental contaminants. The algorithm at this stage has removed certain contaminants only recovered from blanks and, based on user defined thresholds, it has also identified (flagged) potential contaminants. The user is provided a summary of the ten closest matches to the SILVA database for each flagged OTU (Table 3). It is now up to the user to decide to retain or remove, based on i) the percent identity of and similarity of sequence alignments and ii) the environmental information, if any, available for the OTU’s closest relatives in the SILVA database, each flagged OTU. This approach is also imperfect since the database may be biased towards highly sampled environments; however, in every instance, this step provides as much information as is currently available for the user to make decisions and, ultimate, gage the strength of their data.

**

**Table 2. Summary OTUs flagged or removed by *TaxonSluice.***

Below we outline the rationale used to make a *TaxonSluice*-enabled decision on each of our flagged OTUs.

*Flagged OTU00002:* **Discard**- This OTU was present in two of three independent blank samples. In both instances its sequence abundance was within the same order of magnitude in both environmental samples and blanks. Close relatives were isolated from soil rather than marine habitats. The OTU was thus discarded from the dataset.

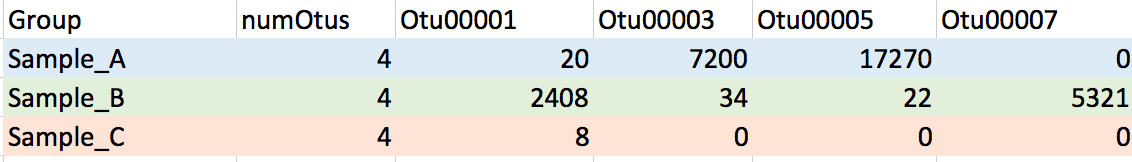
*Flagged OTU00004:* **Discard**- This OTU was present in two of three independent blank samples. In one instance (Sample-Blank Pair B) the blank was had two orders of magnitude sequence counts relative to the environmental sample. Close relatives were isolated from the mammalian oral cavity rather than marine habitats. The OTU was thus discarded from the dataset.

*Flagged OTU00005:* **Keep**-This OTU was recovered from one of three independent blanks and two of three independent environmental samples. The sequence abundance of this OTU is two orders less in the blank relative to one of the environmental samples. Further all close relatives for which environmental data exists inhabit marine environments. The OTU was thus kept in the dataset.

*Flagged OTU00007:* **Keep**- The sequence abundance of this OTU is two orders less in the blank relative to the environmental sample where it was detected. All close relatives for which environmental data exists inhabit marine environments. The OTU was thus kept in the dataset.

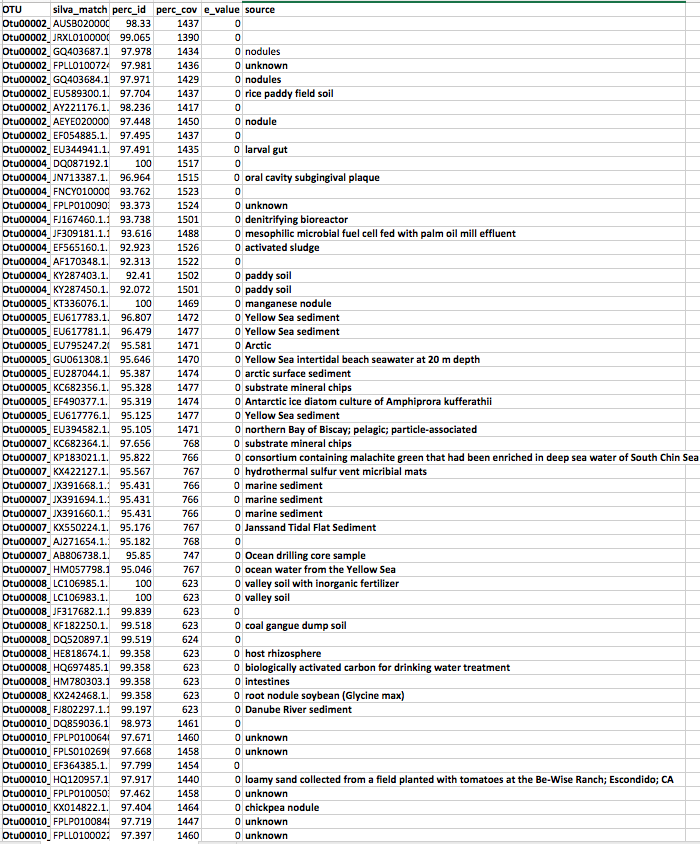
*Flagged OTU00010:* **Discard**- This OTU was present in two of three independent blank samples. In one instance, sequence abundance is within the same order of magnitude in an environmental samples and a blanks. Close relatives were isolated from soil rather than marine habitats. The OTU was thus discarded from the dataset.

Our final, *TaxonSluice* parsed and user pared, “clean” OTU table (omitting blanks) is show below (Table 4).



**Table 4. Final Clean mock dataset**

*In sum*: *TaxonSluice* 1) eliminated every OTU only detected in blanks automatically, 2) identified and provided information on each potential contaminant as determined by user-defined thresholds, and 3) based on the user’s final decisions, provided a *mothur* compatible final “clean” OTU table which may be used for performing downstream ecological analyses.



**Table 3.** *TaxonSluice* output for the Mock Community (Table 1) dataset.