** INSERT TITLE **

Patrick D. Schloss†

† To whom correspondence should be addressed: pschloss@umich.edu

Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI

Abstract

150 words

- Add new Swarm reference

The ability to assign 16S rRNA gene sequences to operational taxonomic units (OTUs) allows microbial ecologists to overcome the inconsistencies and biases within bacterial taxonomy and provides a strategy for clustering similar sequences that do not have representatives in a reference database. These sequences are clustered into OTUs based on their distance from (or similarity to) each other. Numerous algorithms for solving this seemingly simple problem have blossomed in recent years and were recently the subject of benchmarking studies performed by Westcott and myself [ref x 2], He et al. [ref], and Kopylova et al [ref]. These studies provide a thorough review of the sequencing clustering landscape, which can be divided into three general approaches: (i) *de novo* clustering where sequences are clustered without first mapping sequences to a reference database, (ii) closed-reference clustering where sequences are clustered based on the references that the sequences map to, and (iii) open reference clustering where sequences that do not map adequately to the reference are then clustered using a *de novo* approach. My studies have highlighted a persistent problem in the development of clustering algorithms, which is assessing the quality of the clustering assignments.

The most recent analysis of Kopylova et al (**???**) repeats many of the benchmarking strategies employed by previous researchers. First, they and others have compared the time and memory required to cluster sequences in a dataset (1–4). These are valid parameters to assess when judging a clustering method, but have little to say about the quality of the OTU assignments. In the extreme, we could easily develop a toy a clustering algorithm could randomly assign sequences to a predetermined number of OTUs and be efficient, but also poorly reflect the genetic diversity within the community. Second, they and others have compared the number of OTUs generated by various methods for a common dataset (**???**). Again, the toy clustering algorithm could cluster to a target number of OTUs, but the clusterings would likely be meaningless. Third, because many of the methods are sensitive to the initial order of the sequences, a metric of OTU stability has been proposed as a way to assess algorithms (5). Although it is important that the methods generate reproducible OTU assignments, this approach ignores the possability that the variation in assignments may be equally robust or that the assignemnts by a highly reproducible algorithm

may be quite poor. Fourth, the method that Kopylova et al. (**???**) relied upon the most was to cluster sequences from simulated data or data from synthetic communities of cultured organisms and quantify how well the OTU assignments matched the organisms' taxonomy (2, 4, 6–13). This approach is flawed because bacterial taxonomy often reflects those inconsistencies and biases within bacterial taxonomy that OTU-based methods strive to overcome. Furthermore, this benchmarking strategy can only be applied when the actual taxonomy of the organisms are known and so it is unclear how the methods scale to sequences from the novel organisms we are likely to encounter in deep sequencing surveys. In a final approach, Kopylova and others have assessed the quality of clustering based on their ability to generate the same OTUs generated by other methods (3, 14). Unfortunately, without the ability to ground truth any method, such comparisons are tenuous. Westcott and I have proposed an unbiased and objective method for assessing the quality of OTU assignments that can be applied to any collection of sequences (**???**, 15).

Our approach uses the observed dissimilarity between pairs of sequences and information about whether sequences were clustered together to quantify how well similar sequences are clustered together and dissimilar sequences are clustered apart. To quantify the correlation between the observed and expected OTU assignments, we synthesize the relationship between OTU assignments and the distances between sequences using the Matthew's correlation coefficient (16). In the most recent application of this approach (**???**), we found that closed-reference clustering algorithms could be sensitive to the order of the sequences in the reference database and frequently clustered sequences together that were more than 3% different from each other. We also discussed that because open-reference clustering was dependent on closed-reference clustering it had sensitivity to the order of the database. Furthermore, we described how it had a nebulous threshold for OTUs since sequences are clustered based on a radius of 3% under the closed-reference phase and a diameter of 3% under the open reference phase. Finally, we showed that *de novo* clustering algorithms generated the most robust OTU assignments and confirmed our previous analysis that the average neighbor algorithm consistently performed the best (15). Given the observation that the best algorithm may vary by dataset we concluded that researchers should quantify the MCC for several *do novo* algorithms before selecting an algorithm.

To revisit these results, I have expanded the analysis to incorporate additional *de novo* and closed-reference algorithms and two simulated datasets following the approach described by Kopylova et al (**???**). I evaluated three heirarchical and XXXXX greedy *de novo* algorithms, one open-reference clustering algorithm, and three closed-reference algorithms (Figure 1). To test these approaches I applied each of them to four datasets from soil (17), mouse feces (18), and two simulated communities. The simulated communities were generated by randomly selecting 10,000 16S rRNA sequences that were unique within the V4 region from the SILVA non-redundant database (**???**). Next, an even community was generated by specifying that each sequence had a frequency of 100 reads and a staggered community was generated by specifying that the abundance of each sequence was a randomly drawn a uniform distribution between 1 and 200. The benchmarking approach is the same as the approach used previously (**???**). A reproducible version of this manuscript and analysis has been added to the repository available at https://github.com/SchlossLab/Schloss_Cluster_PeerJ_2015.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure

dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Although the goal of Kopylova et al. (**???**) was to compare various clustering algorithms, they also studied these algorithms in the broader context of raw sequence processing, screening for chimeras, and removal of singletons. Each of these are critical decisions in a comprehensive pipeline; however, they confound the analysis of how best to cluster sequences into OTUs that reflect a specific distance threshold. Through the use of objective criteria that measure the quality of the clusterings, independent of taxonomy or database, researchers will be able to evaluate which clustering algorithm is the best fit for their data.

Figure 1. Comparison of OTU quality generated by multiple algorithms applied to four datasets. The nearest, average, and furthest neighbor clustering algorithms were used as implemented in mothur (v.1.37). Abundance (AGC) and Distance-based greedy clustering (DGC) were implemented using USEARCH (v.6.1) and VSEARCH (v.1.5.0). Other *de novo* clustering algorithms included Swarm (v.2.1.1), OTUCLUST (v.0.1), and SUMACLUST (v.1.0.20). The MCC values for Swarm were determined by selecting the distance threshold that generated the maximum MCC value for each dataset. The USEARCH and SortMeRNA (v.2.0) closed-reference clusterings were performed using QIIME (v.1.9.1) and the VSEARCH closed-reference clusterings were performed substituting VSEARCH for USEARCH in the QIIME workflows. The order of the sequences in each dataset was randomized thirty times and the intra-method range in MCC values was smaller than the plotting symbol. MCC values were calculated in mothur.

Nearest neighbor Average neighbor Furthest neighbor USEARCH AGC VSEARCH AGC USEARCH DGC VSEARCH DGC Swarm OTUClust SumaClust Open-reference USEARCH closed-reference VSEARCH closed-reference SortMeRNA closed-reference

References

1. **Sun Y**, **Cai Y**, **Liu L**, **Yu F**, **Farrell ML**, **McKendree W**, **Farmerie W**. 2009. ESPRIT: Estimating species richness using large collections of 16S rRNA pyrosequences. Nucleic Acids Research **37**:e76–e76. doi:10.1093/nar/gkp285.

2. **Cai Y**, **Sun Y**. 2011. ESPRIT-tree: Hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time. Nucleic Acids Research **39**:e95–e95. doi:10.1093/nar/gkr349.

3. **Rideout JR**, **He Y**, **Navas-Molina JA**, **Walters WA**, **Ursell LK**, **Gibbons SM**, **Chase J**, **McDonald D**, **Gonzalez A**, **Robbins-Pianka A**, **Clemente JC**, **Gilbert JA**, **Huse SM**, **Zhou H-W**, **Knight R**, **Caporaso JG**. 2014. Subsampled open-reference clustering creates

consistent, comprehensive OTU definitions and scales to billions of sequences. PeerJ **2**:e545. doi:10.7717/peerj.545.

4. **Mahé F**, **Rognes T**, **Quince C**, **Vargas C de**, **Dunthorn M**. 2014. Swarm: Robust and fast clustering method for amplicon-based studies. PeerJ **2**:e593. doi:10.7717/peerj.593.

5. **He Y**, **Caporaso JG**, **Jiang X-T**, **Sheng H-F**, **Huse SM**, **Rideout JR**, **Edgar RC**, **Kopylova E**, **Walters WA**, **Knight R**, **Zhou H-W**. 2015. Stability of operational taxonomic units: An important but neglected property for analyzing microbial diversity. Microbiome **3**. doi:10.1186/s40168-015-0081-x.

6. **Edgar RC**. 2013. UPARSE: Highly accurate OTU sequences from microbial amplicon reads. Nature Methods **10**:996–998. doi:10.1038/nmeth.2604.

7. **Barriuso J**, **Valverde JR**, **Mellado RP**. 2011. Estimation of bacterial diversity using next generation sequencing of 16S rDNA: A comparison of different workflows. BMC Bioinformatics **12**:473. doi:10.1186/1471-2105-12-473.

8. **Bonder MJ**, **Abeln S**, **Zaura E**, **Brandt BW**. 2012. Comparing clustering and pre-processing in taxonomy analysis. Bioinformatics **28**:2891–2897. doi:10.1093/bioinformatics/bts552.

9. **Chen W**, **Zhang CK**, **Cheng Y**, **Zhang S**, **Zhao H**. 2013. A comparison of methods for clustering 16S rRNA sequences into OTUs. PLoS ONE **8**:e70837. doi:10.1371/journal.pone.0070837.

10. **Huse SM**, **Welch DM**, **Morrison HG**, **Sogin ML**. 2010. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environmental Microbiology **12**:1889–1898. doi:10.1111/j.1462-2920.2010.02193.x.

11. **May A**, **Abeln S**, **Crielaard W**, **Heringa J**, **Brandt BW**. 2014. Unraveling the outcome of 16S rDNA-based taxonomy analysis through mock data and simulations. Bioinformatics **30**:1530–1538. doi:10.1093/bioinformatics/btu085.

12. **Sun Y**, **Cai Y**, **Huse SM**, **Knight R**, **Farmerie WG**, **Wang X**, **Mai V**. 2011. A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. Briefings in Bioinformatics **13**:107–121. doi:10.1093/bib/bbr009.

13. **White JR**, **Navlakha S**, **Nagarajan N**, **Ghodsi M-R**, **Kingsford C**, **Pop M**. 2010. Alignment and clustering of phylogenetic markers - implications for microbial diversity studies. BMC Bioinformatics **11**:152. doi:10.1186/1471-2105-11-152.