clustur: An R package for clustering features using sparse distances matrices

Running title: clustur

Gregory Johnson Jr., Sarah L. Westcott, Patrick D. Schloss†

†To whom correspondence should be addressed

pschloss@umich.edu

Department of Microbiology & Immunology

University of Michigan

Ann Arbor, MI 48109

Software Announcement

Abstract (needs to be under 50 words)

Veniam commodo eu ullamco in cupidatat. Labore exercitation incididunt occaecat ut ullamco ad velit laboris cupidatat velit reprehenderit excepteur commodo. Est in consequat in sit non cillum laborum aliqua do pariatur deserunt. Minim commodo commodo sint Lorem elit et adipisicing commodo aute officia officia. Dolore aliqua culpa id minim reprehenderit duis magna voluptate. Id laborum deserunt dolor dolore elit. Et est reprehenderit aute velit occaecat ipsum labore.

Announcement (needs to be around 500 words, currently 595)

- Problem definition
  - Taxonomic classification of 16S rRNA gene sequences is not great
  - Methods have been developed for de novo clustering of sequences
  - Heuristic methods that have been developed
  - Methods have been developed for reference-based clustering of sequences
  - These are available within mothur
- What does clustur do to solve problem
  - Package bundles together mothur's functionality
  - Hopes to help spurn further innovation in clustering by making functions accessible via R
  - Will make functionality available to other types of analysis beyond 16S rRNA gene sequences
- Design of clustur
  - Makes use of Rcpp
- How to install and use clustur
  - Users can install via CRAN or through the devtools package's install_github function
  - Users provide different inputs depending on desired function, output is a long shared file with columns indicating the sample, OTU, and count

(1–7)

Data availability

clustur is available through CRAN and developmental versions are available through the project's GitHub website (https://github.com/schlosslab/clustur). The package is available under the MIT open source license.

Acknowledgements

This project was supported, in part, by a grant from the US National Institutes of Health (U01CA264071) to PDS.

References

45 1. **Schloss PD**, **Handelsman J**. 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Applied and Environmental Microbiology **71**:1501–1506. doi:10.1128/aem.71.3.1501-1506.2005.

2. **Schloss PD**, **Westcott SL**, **Ryabin T**, **Hall JR**, **Hartmann M**, **Hollister EB**, **Lesniewski RA**, **Oakley BB**, **Parks DH**, **Robinson CJ**, **Sahl JW**, **Stres B**, **Thallinger GG**, **Van Horn DJ**, **Weber CF**. 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology **75**:7537–7541. doi:10.1128/aem.01541-09.

3. **Schloss PD**, **Westcott SL**. 2011. Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. Applied and Environmental Microbiology **77**:3219–3226. doi:10.1128/aem.02810-10.

4. **Schloss PD**. 2016. Application of a database-independent approach to assess the quality of operational taxonomic unit picking methods. mSystems **1**. doi:10.1128/msystems.00027-16.

5. **Sovacool KL**, **Westcott SL**, **Mumphrey MB**, **Dotson GA**, **Schloss PD**. 2022. OptiFit: An improved method for fitting amplicon sequences to existing OTUs. mSphere **7**. doi:10.1128/msphere.00916-21.

50 6. **Westcott SL**, **Schloss PD**. 2015. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. PeerJ **3**:e1487. doi:10.7717/peerj.1487.

7. **Westcott SL**, **Schloss PD**. 2017. OptiClust, an improved method for assigning amplicon-based sequence data to operational taxonomic units. mSphere **2**. doi:10.1128/mspheredirect.00073-17.