Identifying Composition Novelty in Microbiome Studies: Improvement for Prediction Accuracy

Yu Sun,[a,b] Yanling Li,[a,b] Qianqian Yuan,[a,b] Xi Fu[c]

aGuangdong Provincial Key Laboratory of Protein Function and Regulation in Agricultural Organisms, College of Life Sciences, South China Agricultural University, Guangzhou, Guangdong, People's Republic of China
bKey Laboratory of Zoonosis of Ministry of Agriculture and Rural Affairs, South China Agricultural University, Guangzhou, Guangdong, People's Republic of China
cSchool of Public Health, Sun Yat-sen University, Guangzhou, People's Republic of China

We read with interest the paper recently published by Su et al. entitled "Identifying and Predicting Novelty in Microbiome Studies" (1). The paper presents the novel Microbiome Search Engine (MSE; http://mse.single-cell.cn/index.php/mse), which enables rapid searches of the composition similarity of query samples against sequences in Qiita (http://qiita.ucsd.edu), a well-curated and regularly updated microbiome reference database currently housing 177,022 bacterial 16S rRNA samples from humans, seawater, freshwater, soil, buildings, plants, foods, and many other environments (2, 3). Su et al. introduce a microbiome novelty score (MNS), which calculates bacterial composition difference between the query samples and the top 10 most matched microbiomes in the Qiita database. A high MNS score indicates low composition similarity to previously sampled microbiomes and high novelty. The MNS score is an easy-to-use and quantitative index to evaluate compositional uniqueness for new samples, which has the potential to be an essential analysis for future microbiome studies. However, due to technique limitation from low taxonomic resolution of 16S rRNA sequences, unexpected high similarity can be observed between unrelated samples.

We searched 83 newly sequenced settled-dust samples from dormitories in Shanxi University, China, with MSE and found that the majority of samples (67.5%) matched best with mosquito tissue samples; another 6% of samples matched birds' egg shell sequences. The query samples also had low MNS scores (<0.09) due to the high similarities. We observed similar results in our continental hotel microbiome study, in which one sample matched mosquito sequences and one sample matched birds' sebum sequences. Although mosquitos are present in dormitories, especially in summer, it is not possible that they contribute dominantly to this indoor environment. Many studies show that indoor microbes are mainly from outdoor environmental sources, such as air and soil, as well as human sources, such as from respiration and shedding (4–6). For both dormitory and mosquito samples, a high percentage of *Pseudomonas* spp. have been identified (Fig. 1). *Pseudomonas* is one of the most diverse and ubiquitous bacterial genera in many environments, such as sediments, water, soil, desert, plants, and animals (7). The *Pseudomonas* species in mosquito tissue certainly differs from the *Pseudomonas* sequences in dormitories, but as the 16S rRNA technique can resolve taxonomy only to the genus level, the composition similarity between the two samples is high.

To solve this issue, both the query samples and the reference database should use high-resolution taxonomic data, such as metagenomics data, to evaluate community similarity, but this approach is not economically practical for most current microbiome survey studies. Alternative solutions, such as adding a new score algorithm, can be more practical. The previous MNS score searches sequences from all environments; the new MNS score searches only a subset of environments, the query-related environ-
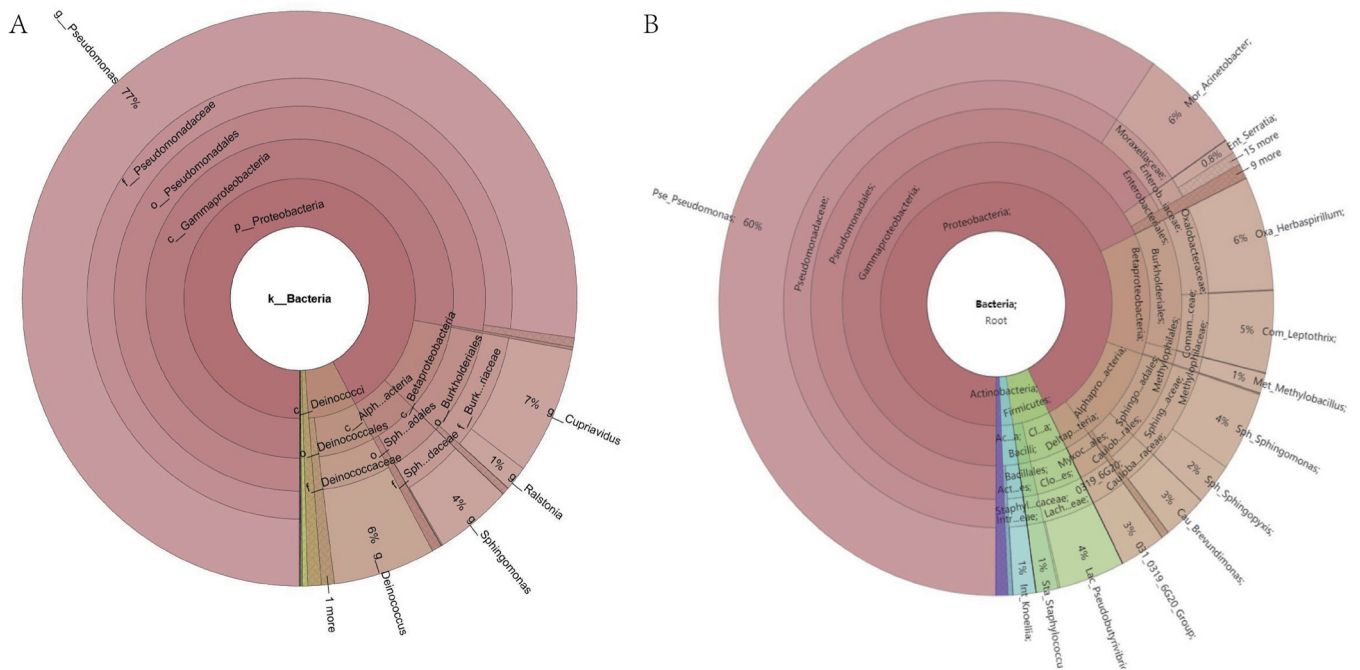
FIG 1 Bacterial compositions for one representative sample from a university dorm (A) and mosquito tissue (B). g_, genus; f_, family; o_, order; c_, class; p_, phylum; Alpha...acteria, *Alphaproteobacteria*; Sph...adales, *Sphingomonadales*; Sph...daceae, *Sphingomonadaceae*; Burk...riaceae, *Burkholderiaceae*; Pse_, *Pseudomonadaceae*; Mor_, *Moraxellaceae*; Ent_, *Enterobacteriaceae*; Oxa_, *Oxalobacteraceae*; Com_, *Comamonadaceae*; Met_, *Methylophilaceae*; Sph_, *Sphingomonadaceae*; Cau_, *Caulobacteraceae*; Lac_, *Lachnospiraceae*; Sta_, *Staphylococcaceae*; Int_, *Intrasporangiaceae*; Alphapro...aceteria, *Alphaproteobacteria*; Deltap...teria, *Deltaproteobacteria*; Myxoc...ales, *Myxococcales*; Sphingo...adales, *Sphingomonadales*; Act...es, *Actinomycetales*; Cl...es, *Clostridiales*; Caulob...rales, *Caulobacterales*; Enterb...iaceae, *Enterobacteriaceae*; Staphyl...caceae, *Staphylococcaceae*; Sping...aceae, *Sphingomonadaceae*; Comam...ceae, *Comamonadaceae*; Intr...eae, *Intrasporangiaceae*; Lach...eae, *Lachnospiraceae*; Cauloba...raceae, *Caulobacteraceae*.

ments. For example, for the dormitory query samples, the new MNS score will be calculated from the building, air, soil, and other related environments, and erroneous identifications from mosquitos are thus not included. Then users can choose which MNS scores to use from their scientific perspectives.

REFERENCES

1. Su X, Jing G, McDonald D, Wang H, Wang Z, Gonzalez A, Sun Z, Huang S, Navas J, Knight R, Xu J. 2018. Identifying and predicting novelty in microbiome studies. mBio 9:e02099-18. https://doi.org/10.1128/mBio.02099-18.

2. Gonzalez A, Navas-Molina JA, Kosciolek T, McDonald D, Vázquez-Baeza Y, Ackermann G, DeReus J, Janssen S, Swafford AD, Orchanian SB, Sanders JG, Shorenstein J, Holste H, Petrus S, Robbins-Pianka A, Brislawn CJ, Wang M, Rideout JR, Bolyen E, Dillon M, Caporaso JG, Dorrestein PC, Knight R. 2018. Qiita: rapid, web-enabled microbiome meta-analysis. Nat Methods 15:796–798. https://doi.org/10.1038/s41592-018-0141-9.

3. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vázquez-Baeza Y, González A, Morton JT, Mirarab S, Zech Xu Z, Jiang L, Haroon MF, Kanbar J, Zhu Q, Jin Song S, Kosciolek T, Bokulich NA, Lefler J, Brislawn CJ, Humphrey G, Owens SM, Hampton-Marcell J, Berg-Lyons D, McKenzie V, Fierer N, Fuhrman JA, Clauset A, Stevens RL, Shade A, Pollard KS, Goodwin KD, Jansson JK, Gilbert JA, Knight R. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551:457–463. https://doi.org/10.1038/nature24621.

4. Prussin AJ, II, Marr LC. 2015. Sources of airborne microorganisms in the built environment. Microbiome 3:78. https://doi.org/10.1186/s40168-015-0144-z.

5. Adams RI, Bhangar S, Dannemiller KC, Eisen JA, Fierer N, Gilbert JA, Green JL, Marr LC, Miller SL, Siegel JA, Stephens B, Waring MS, Bibby K. 2016. Ten questions concerning the microbiomes of buildings. Build Environ 109:224–234. https://doi.org/10.1016/j.buildenv.2016.09.001.

6. Kelley ST, Gilbert JA. 2013. Studying the microbiology of the indoor environment. Genome Biol 14:202. https://doi.org/10.1186/gb-2013-14-2-202.

7. Peix A, Ramírez-Bahena M-H, Velázquez E. 2009. Historical evolution and current status of the taxonomy of genus Pseudomonas. Infect Genet Evol 9:1132–1147. https://doi.org/10.1016/j.meegid.2009.08.001.