**BMC
Proceedings**

The consequences of denoising marker-based metagenomic data

John M Gaspar[*], W Kelley Thomas

Background

Early marker-based metagenomic studies, such as those of the human microbiome, were performed without properly accounting for the effects of noise (pyrosequencing errors, PCR single-base errors, and PCR chimeras). One popular solution to address these issues is to utilize Amplicon-Noise [1]. This collection of algorithms was validated on mock community datasets in which the 'correct' result, such as the number of operational taxonomic units (OTUs), was known. However, when conducting a real study, one will not know the correct result, but still must consider how the data has been transformed by denoising.

Materials and methods

We applied AmpliconNoise to several real metagenomic datasets. At each stage of the pipeline, we reconstituted the reads and determined how they had been affected. The changes were quantified as substitutions, insertions, deletions and '3' gap', which is the number of bases removed from (or added to) the 3' end of a read. We further analyzed the effects of the related denoising programs in QIIME (Denoiser [2]) and in mothur [3].

Results

The preliminary filtering steps of AmpliconNoise caused most of the sequence reads to be eliminated or truncated. Following this, the algorithm PyroNoise caused changes to the reads that were inconsistent with the known spectrum of pyrosequencing errors, until one of the parameters was increased substantially. Additionally, because PyroNoise mapped reads onto longer representatives, sequences were added to the 3' ends of reads that were often dissimilar from those that were removed by the truncations of the filtering steps. After this, SeqNoise, which was designed to remove PCR single-base errors, further clustered the reads

and caused even more changes to the reads with little justification.

Denoiser, which is based on an earlier version of AmpliconNoise, caused far more changes to the data. The evaluation of the changes was not as clear here, since they were not clearly delineated as to which type of errors they were correcting, but we found some of the same flawed methodology that produced much of the negative effects seen in AmpliconNoise. This was also true of the denoising programs in mothur, which were recoded directly from the AmpliconNoise algorithms.

Conclusions

While reducing the effects of noise in the analysis of marker-based metagenomic data is important, the algorithms of AmpliconNoise make changes to sequence reads that are inconsistent with simply removing noise. We recommend that those using AmpliconNoise be cognizant of the possible side effects and, at a minimum, consider adjusting the parameters of the algorithms accordingly.

References
1. Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ: **Removing noise from pyrosequenced amplicons.** *BMC Bioinformatics* 2011, **12**:38.
2. Reeder J, Knight R: **Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions.** *Nat Methods* 2010, **7**:668-669.
3. Schloss PD, Gevers D, Westcott SL: **Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies.** *PLoS One* 2011, **6**:e27310.

Department of Molecular, Cellular, and Biomedical Sciences, University of New Hampshire, Durham, NH, USA

BioMed Central