Preprocessing make critical bias at microbial community

Kyuin Hwang[1,2], Soon Gyu Hong[1,2,*]

[1]Division of Life Sciences, Korea Polar Research Institute, KIOST, Incheon 406-840, Korea

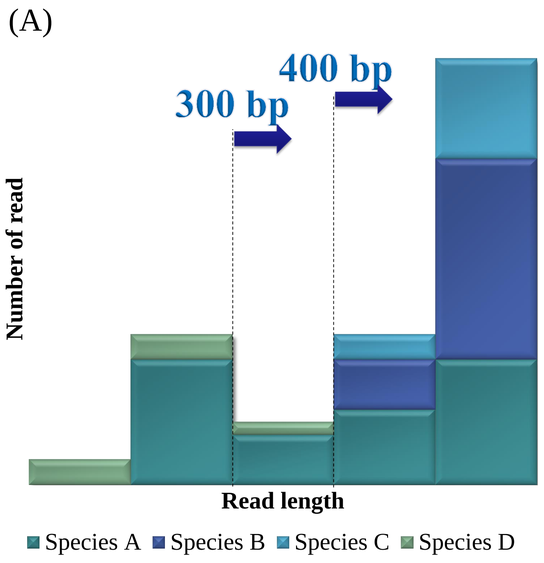
[2]Department of Polar Sciences, University of Science and Technology, Incheon 406-840, Korea

* Corresponding author:

Soon Gyu Hong, Tel: +82-32-760-5580; e-mail: polypore@kopri.re.kr

INTRODUCTION

Next-generation sequencing technology (NGS) is becoming a standard method to examine microbial diversity of environmental and human microbiome samples. Read number is used to estimate relative abundance of specific taxa in the samples as a key parameter to define microbial community structures. However, it is affected not only by relative abundance of the taxa in the original sample but also by various processes such as PCR amplification, sequencing reactions and sequence processing pipelines including trimming, filtering, noise removing, chimera detection, clustering, and DB search. Although sequence read number biases by PCR and sequencing reactions are well known, read number bias created by quality trimming and filtering is not well known. Trimming low-quality nucleotides and filtering out short sequence reads are included in most of the popular NGS processing pipelines to analyze microbial community based on high-quality sequence reads. These processes are based on the assumption that low-quality reads are evenly distributed among taxa. However, several publication report that sequencing error is highly dependent on the sequence context, which in turn can create biased sequence trimming and discarding sequences from specific taxa.



	All	300 bp*	400 bp*
Species A	46.15%	39.22%▼	34.78%▼
Species B	30.77%	39.22%▲	43.48%▲
Species C	15.38%	19.61%▲	21.74%▲
Species D	7.69%	1.96%▼	0.00%▽

Fig. 1. Principle of preprocessing bias. (A) When different species has different read length distribution after low-quality trimming. (B) Relative abundance of species changed by short read filtering. **300 bp*** and **400 bp***, filter out reads which short than 300 bp and 400 bp

MATERIALS & METHODS

Pyrosequencing data of synthetic microbial community samples achieved from NCBI Short Read Archive which involve in Human Microbiom Project (SRP002443). This data include more than 8 synthetic microbial community. And each community comprised by 21 clones (species). 3 variable region (V13,V36,V69) of 16S ribosomal DNA sequenced by forward and reverse direction from each community

From this data,

(1) Sort sequences according to the barcode primer sequences.
(2) Trim sequence reads using various trimming methods which included in popular pipelines (AmpliconNoise, mothur, PyroTrimmer, QIIME, UPARSE)
(3) Identify sequences by BLAST search.
(4) Calculate filtering rate of each species when filter out shortest 10%, 15%, 20%, 25% reads from community.
(5) Quantify amount of read number bias as standard deviation of filtering rate between species.

Above processes conducted by in-house python scripts.

Acknowledgement

This work was supported by Korea Polar Research Institute (Grant PE15020).

RESULT

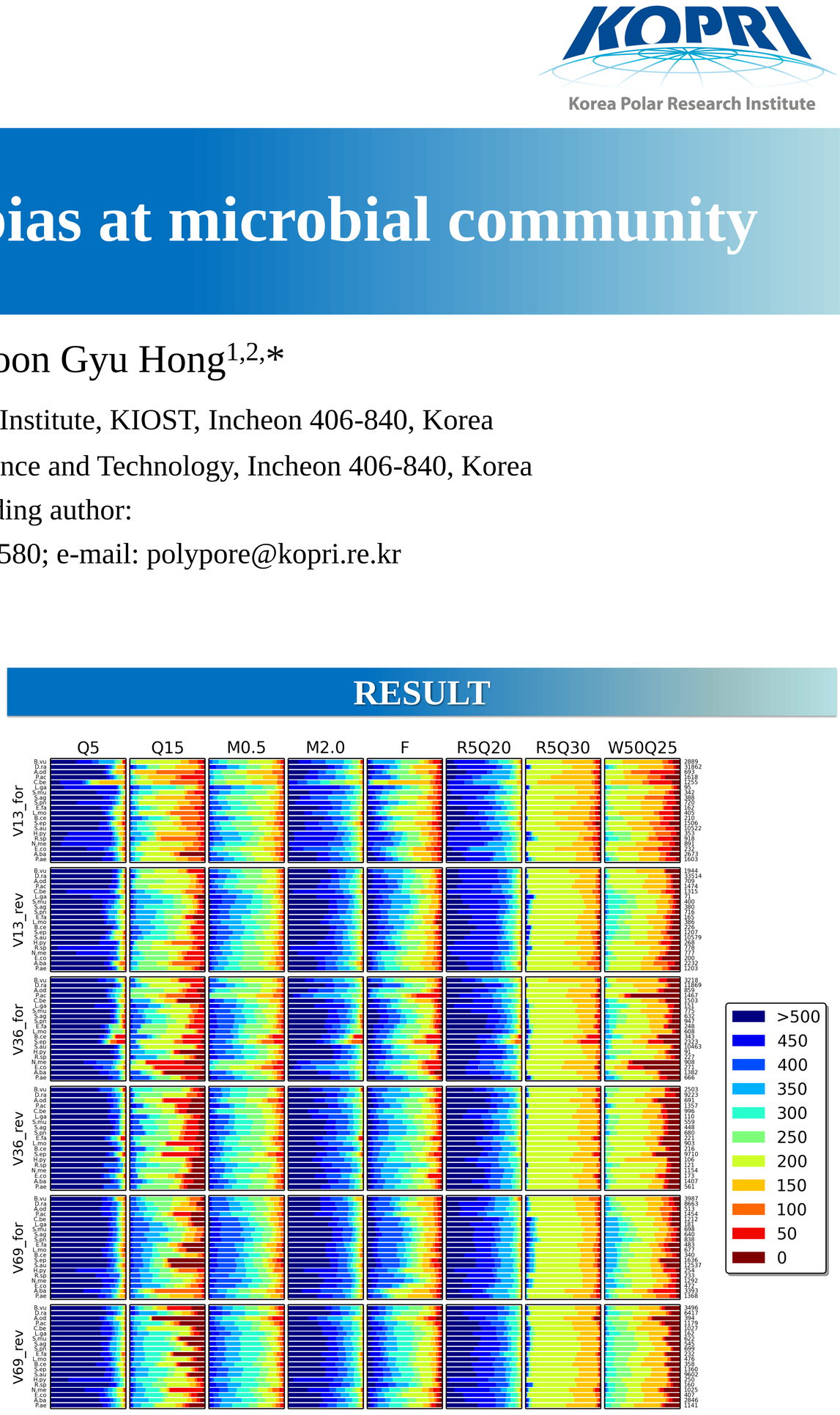


Fig. 2. Read length distribution of species after trimming. Six sequence sets for three domains (V13, V36, and V69 domains, forward and reverse direction) of bacterial 16S rRNA genes were processed by different trimming algorithms. Species name was abbreviated by first letter of genus name and first two letter of specific epithet (ex. *Acinetobacter baumannii* : A.ba). **Q5** and **Q15**, trimming at first low quality base with threshold 5 and 15; **M0.5** and **M2.0**, trimming at maximum expected error number with the threshold 0.5 and 2.0; **F**, trimming at noisy flow; **R5Q20** and **R5Q30**, trimming at the last high scored region with threshold 20 and 30 for 5 bp window; **W50Q25**, trimming at first low scored region with threshold 25 for 50 bp window.

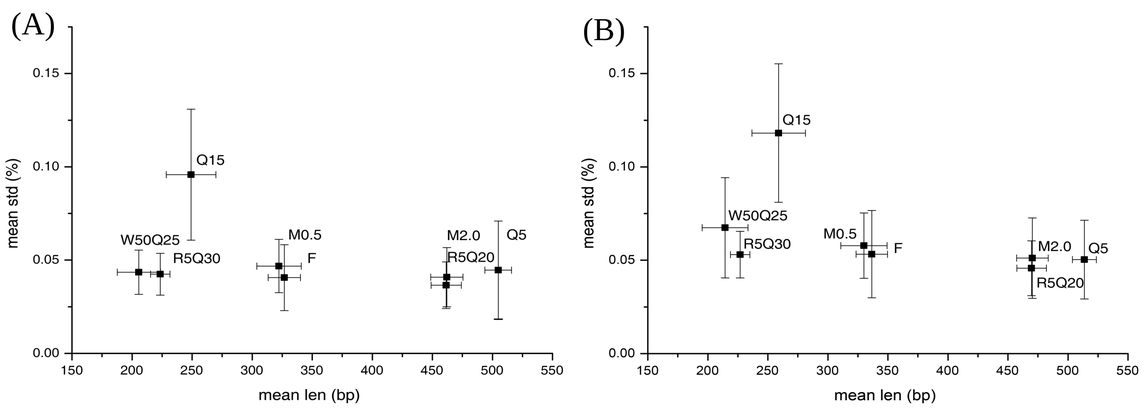


Fig. 3. Difference in filtering rate between species after preprocessing . Average of standard deviation of filtering rates among species (y axis) for six sequence sets plotted with average length of remaining reads (x axis) after shortest 10% (A), 15% (B) of reads are filtered. Error bar represent standard deviation of read length and standard deviation among six sequence sets.

CONCLUSION

Read length distribution are highly different among species. Especially, *Clostridium beijerinckii* show shorter read lengths than other species at V13 forward region as species A of figure 1.

Read number bias is highly dependent on trimming method. R5Q20 trimming method show lowest average bias and this method always show lower bias in all of 6 regions.

Read number bias was also affected by filtering rate. Read number bias was proportionate with filtering rate in most cases by various trimming methods and sequencing regions.

First of all, we strongly recommend check read length distribution before/after preprocessing. But, basically we recommend that trimming read at the last high-quality region (R5Q20) and filtering low-quality reads using more lenient thresholds to mitigate read number bias.