[BIOINFO 2024]

Evaluating Metagenomics Fosmid Libraries Using Vector-Adjacent Sequences

Suvin Baek[1,2], Wonjae Seong[1], Eugene Rha[1], Kil Koang Kwon[1], Seung-Goo Lee[1], Haseong Kim[1,2*]

[1] *Synthetic Biology Research Center, Korea Research Institute of Bioscience and Biotechnology*
[2] *Graduate School of Engineering Biology, Korea Advanced Institute of Science and Technology*
**Corresponding author: haseong@kribb.re.kr*

Recently, metagenomics is increasingly used with artificial intelligence models to explore valuable genetic resources, analyze microbial community composition, and trace the transmission pathways of pathogenic viruses. Metagenomics fosmid libraries are notable for allowing the long-term storage of diverse genetic resources in physical form, enabling the functional screening and discovery of novel genetic resources. In this study, we propose a quantitative metric to evaluate the quality of metagenome libraries, particularly those created by inserting metagenomes into vectors. Conventional metrics used to evaluate sample metagenomics libraries include alpha diversity, colony counting, and sequence count; however, these metrics do not adequately capture changes in library quality that can arise during the construction of fosmid libraries. In this study, we utilize long-read sequencing technology to calculate the diversity of the library based on sequences adjacent to the fosmid vector, allowing for the correction of inflated library size estimates caused by redundant sequences generated during the library construction process. The calculated diversity metrics are represented as rarefaction curves, enabling the evaluation of sequence diversity across samples and determining whether sequencing depth is sufficient. The approach offers a quantitative metric for assessing library quality and is anticipated to provide a critical decision-making framework for determining whether to proceed with experiments that involve additional sequencing costs.