Human reference gut microbiome including 5,414 prokaryotic species, including newly assembled genomes from under-represented Asian metagenomes

Chan Yeong Kim[1†], Muyoung Lee[1†], Sunmo Yang[1], Kyungnam Kim[2], Dongeun Yong[2], Hye Ryun Kim[3], and Insuk Lee[1*]

[1] Department of Biotechnology, College of Life Science & Biotechnology, Yonsei University, Seoul 03722, Korea. [2] Department of Laboratory Medicine, Research Institute of Bacterial Resistance, College of Medicine, Yonsei University, Seoul 03722, Korea. [3] Division of Medical Oncology, Department of Internal Medicine, Yonsei Cancer Center, College of Medicine, Yonsei University, Seoul 03722, Korea

† These authors contributed equally to this work.

* Corresponding author: Insuk Lee

ABSTRACT

Metagenome sampling bias for geographical location and lifestyle is partially responsible for the incomplete catalog of reference genomes of gut microbial species. Here, we present a substantially expanded microbiome catalog, the Human Reference Gut Microbiome (HRGM). Incorporating newly assembled 29,082 genomes from 845 fecal samples collected from three under-represented Asian countries—Korea, India, and Japan—the HRGM contains 232,098 non-redundant genomes of 5,414 representative prokaryotic species, >103 million unique proteins, and >274 million single-nucleotide variants. This is an over 10% increase from the largest reference database. The newly assembled genomes were enriched for members of the *Bacteroidaceae* family, including species associated with high-fiber and seaweed-rich diet. Single-nucleotide variant density was positively associated with the speciation rate of gut commensals. Ultra-deep sequencing facilitated the assembly of genomes of low-abundance taxa, and deep sequencing (>20 million read pairs) was needed for the profiling of low-abundance taxa. Importantly, the HRGM greatly improved the taxonomic and functional classification of sequencing reads from fecal samples. Finally, mapping homologous sequences for human auto-antigens onto the HRGM genomes revealed the association of commensal bacteria with high cross-reactivity potential with inflammation. The HRGM (www.mbiomenet.org/HRGM/) will facilitate the identification and functional analysis of disease-associated gut microbiota.