[BIOINFO 2024]

HUMIPRO: A database of metagenomic profiles from over 30,000 gut, oral, and vaginal samples for the study of human microbiome

Geon Koh[1] and Insuk Lee[1,*]

[1]*Department of Biotechnology, College of Life Science & Biotechnology, Yonsei University, Seoul 03722, Republic of Korea*

**Corresponding author: insuklee@yonsei.ac.kr*

HUMIPRO is a large-scale database providing 30,528 taxonomic profiles derived from three key human body sites—gut, oral, and vaginal—collected from 43 countries. The database is designed to support microbiome research by offering high-quality whole metagenomic sequencing (WMS) data, accompanied by relevant metadata, to explore the relationships between microbial communities and various health conditions. HUMIPRO's pipeline is built on a two-track strategy to gather both published and unpublished WMS data, ensuring extensive sample collection. Compared to other existing taxonomic profile databases, HUMIPRO offers more comprehensive and diverse datasets, with a stronger emphasis on disease samples (covering 159 disease phenotypes) and more strict quality control. HUMIPRO includes samples from 233 datasets, providing users with metadata on age, health status, antibiotic use, and disease phenotypes, among others. The database is optimized for bioinformatics experts, allowing for direct access to processed matrices without the need for additional data preparation. HUMIPRO's extensive data collection makes it ideal for advanced analyses, such as differential abundance analysis, which can be used to identify microbial signatures associated with diseases. Additionally, the database's richness enables the development of predictive models using machine learning techniques to forecast disease states based on microbiome composition. Furthermore, the quality-controlled WMS data can be used for genome assembly, enabling the discovery of novel microbial species and further expanding microbiome research.