Methods & Data availability

**Choice of Bacterial Genome WGS data:**

We downloaded high quality sequencing runs for publicly available metagenomes obtained from different parts of the human body from the NCBI SRA database. The data were selected according to rigorous filters in order to make comparison possible and limit irregularities between different sequenced runs from various studies. We selected runs obtained using Illumina sequencing instruments and whole genome sequencing. We selected for studies with a minimum threshold of 5Gb of total data and used experimental runs with over 100Mb. The SRA run selector interface was used to select for specific runs. Metagenomic data was screened by analyzing the papers that provided our data in order to identify steps to remove data contamination (from human or phiX genome). These rigorous steps allowed us to confirm the high quality and reliability of the selected data – without introducing preferential bias.

**Bioinformatics Workflow:**

The pipeline for bioinformatics analysis is as follows.

After raw datasets were obtained, paired short raw reads were assembled into long DNA contigs by SPAdes (version 3.14.1) using the *spades.py* *-meta* command.

Database for blastx alignment was built by DIAMOND (version 0.9.36) based on 2\_raw\_algs.tar file in EggNOG v5.0 bacterial database. Then the assembled DNA contigs from SPAdes (scaffolds.fasta in output files) were mapped to this EggNOG v5.0 bacterial database built by DIAMOND. The persistent index (PI) of each ortholog group (OG) was calculated by the formula:

PI value = # species in this OG / # species in total dataset

Highest PI was assigned to the region if it was aligned to two different OGs. If there was no OG alignment, the PI value for that region of DNA was defined to be 0 and that region was manually defined as a unique unknown gene. By now each position on the DNA scaffolds were uniquely assigned to an ortholog group (or a unique unknown gene) with a corresponding PI value.

The raw pair-ended reads were mapped back to the scaffolds in order to get the “coverage”, which was defined as the number of raw pair-ended reads that were successfully mapped back to the scaffolds. The process was done by Bowtie (version 2.3.4.1) under the default setting. Bedtools 2 (version 2.27.1) succeeded in getting the position-wise coverage under the function *genomeCoverageBed*.

Since different values of “coverage” were not comparable between each scaffold in a sample, the idea of TPM was borrowed to establish the uniform rule of coverage for comparison. Aligned with OG from EggNOG database or not, each segment of DNA scaffolds was treated as a “gene” in the form of TPM, especially when there is no OG alignment, the segment was regarded as a unique “gene”. TPM for each “gene” was calculated by first dividing the original coverage by the “gene length” – the length of the DNA segment (resulting in TPM0), grouping the same OG on different scaffolds together in this sample, and finally summing up all the TPM0 on each nucleotide to get final results of TPM. Hence, the TPM value for each “gene” was obtained.