Supplementary Note

Evaluate the Precision-Recall of different metagenomic profilers. It is well known that the precision-recall analysis is largely impacted by the reference database used by different profilers[1]. To avoid the bias introduced by the database differences, a uniform database should be used for benchmarking studies. However, it is not allowed to customize reference database for most DNA-to-Marker methods (e.g., MetaPhlAn2 and mOTUs2), rendering the comparison between DNA-to-DNA and DNA-to-Marker methods based on the precision-recall analysis very challenging. To resolve this issue, in this work we selected genomes to generate synthetic communities based on the intersection between the reference databases, which enables us to evaluate the performance of different profilers in an unbiased manner.

We evaluated the precision and recall of four representative metagenomic profilers using the same 25 simulated communities as used in main text **Fig.3a-b**. First, we calculated the precision, recall, and F1 score (i.e., the harmonic mean value of precision and recall) using the raw profiling results of each profiler without additional abundance thresholding (see main text, **Fig.3c-e**). We found that: (1) DNA-to-DNA methods have a much lower raw precision (0.63~0.71) than DNA-to-Marker methods (~0.89) (**Fig.3c**); (2) There is only a marginal difference in the raw recall between DNA-to-DNA methods and DNA-to-Marker methods (**Fig.3d**); (3) DNA-to-DNA methods display a relatively lower raw F1 score than DNA-to-Marker methods (**Fig.3e**). These results suggest that both DNA-to-DNA and DNA-to-Marker methods can identify true positives with high accuracy when the reference database covers the species in the sample, but DNA-to-DNA methods identify more false positives in their default profiling results.

Then, we calculated the precision, recall, and F1 score using a set of commonly used abundance thresholds (e.g., 0.0001, 0.001, etc, see **Fig.3f-h**). With increasing abundance threshold, we observed that: (1) The improvement of precision is particularly effective for DNA-to-DNA methods (which suffer from the high false-positive rate issue), while DNA-to-Marker methods (not subject to the high false-positive rate issue) maintain a quite stable and high precision with a range of abundance thresholds (**Fig.3f**); (2) The recall tends to be quite stable when the abundance threshold is less than 0.0005 (which is the minimum species abundance for the simulated communities), but it declines significantly as the abundance threshold keeps increasing, especially when the threshold is over 0.001, for both DNA-to-DNA and DNA-to-Marker methods (**Fig.3g**); (3) F1 score shows a clear trade-off between precision and recall (**Fig.3h**). This pattern is particularly clear for DAN-to-DNA methods. In our calculations, the "sweet point" of the abundance threshold is ~0.0005 (which is the minimum species abundance for the simulated communities), where F1 score reaches its maximum. We emphasize that this "sweet point" is dataset/sample dependent. In reality, without any knowledge of the ground truth, it would be rather challenging (if not impossible) for end users of any metagenomic profiler to choose such an optimal abundance threshold.

Finally, we calculated the Area Under the Precision-Recall Curve (AUPRC) with the abundance threshold systematically tuned from 0 to 1. We found that DNA-to-DNA methods have slightly higher AUPRC than DNA-to-Marker methods (**Fig.S3**). Note that in a previous benchmark study[2], the authors reported a much bigger AUPRC difference between DNA-to-DNA and DNA-to-Marker methods (see their Fig.3A). This is because in that figure, rather than using a uniform reference database, profiler-specific default reference databases were used. As mentioned earlier, the difference in the reference databases of different profilers will largely affect the precision-recall analysis. In another figure (Fig.3B), the authors tried to use a uniform reference database --- RefSeq complete genomes (RefSeq CG), but unfortunately, they cannot apply it to DNA-to-Marker methods (which do not allow users to customize the reference database). Hence, they didn't compare the AUPRC of DNA-to-DNA and DNA-to-Marker methods in a unbiased manner. By contrast, our precision-recall analysis, based on genomes selected from the intersection among the different reference databases of MetaPhlAn2, mOTUs2, and Bracken/Kraken2, naturally avoids the bias of reference databases. Hence, our AUPRC comparison between DNA-to-DNA and DNA-to-Marker methods is unbiased.

Note that, although AUPRC can avoid tricky selection of abundance threshold, it is considered to be potentially biased toward low-precision and high-recall profilers, thus cannot reflect the low precision of DNA-to-DNA methods using their default profiling results. In other words, AUPRC strongly favors profilers (e.g., DNA-to-DNA methods) with high recall at the expense of low precision.

We emphasize that the calculations of precision, recall, F1 score, and AUPRC for each sample are not affected by using sequence or taxonomic abundance as the ground truth. This is simply because those metrics only concern the difference of presence/absence patterns in the ground truth and predicted abundance profiles, and by definition the ground-truth sequence abundance and taxonomic abundance profiles of our simulated microbiome samples share exactly the same presence/absence pattern.

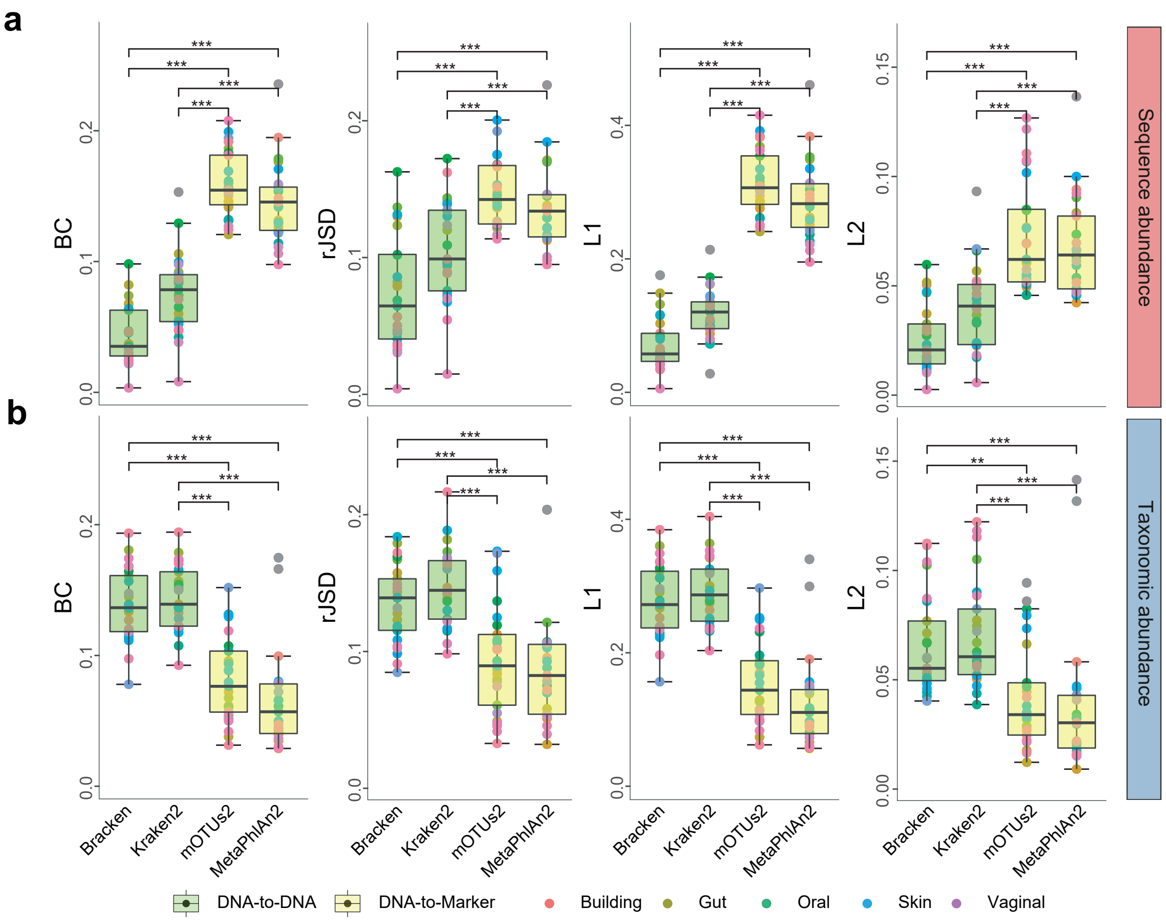

Figure S1. Differential benchmarking results based on modified dissimilarity/distance measures using two types of relative abundance as the ground truth. **a**, sequence abundance and **b**, taxonomic abundance. The boxplots indicate the dissimilarities (based on modified L1, modified L2, modified root Jensen-Shannon divergence (modified rJSD) and modified Bray-Curtis (modified BC) between the ground-truth profiles and the profiles predicted by different metagenomics profilers (Bracken, Kraken2, mOTUs2, and MetaPhlAn2) at the species level. The modified distances were calculated based on the true positives in the resulting profiles as compared to the ground truth. For each metagenomic profiler, we performed the dissimilarity/distance calculations based on n = 25 simulated microbial communities from five representative environmental habitats (gut, oral, skin, vagina and building) separately. Significance levels: p-value<0.05 (*), <0.01 (**), <0.001 (***), NS (non-significance); Two-sided Wilcoxon signed-rank test, exact p-values are provided in the source data.. The lower and upper hinges correspond to the first and third quartiles, and the center refers to the median value. The upper (or lower) whisker extends from the hinge to the largest (smallest) value no further (at most) than 1.5 * IQR from the hinge. Data beyond the end of the whiskers are plotted individually.

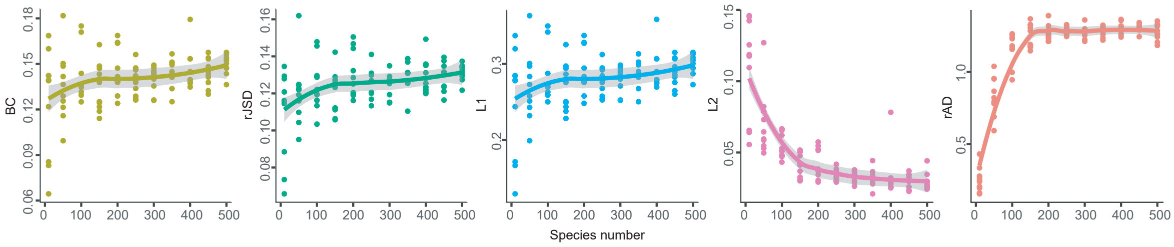

Figure S2. Dissimilarity/distance between sequence abundance and taxonomic abundance with varied species numbers. For each species number, we simulated 10 repeats of profiles. The dissimilarity/distance was then measured by different measures: Bray-Curtis (BC, yellow), rJSD (green), L1 (blue), L2 (purple), and rAD (red). rAD between these types of abundance profiles positively correlated with the species richness when < 200 microbial species presented in a community, yet saturated after the number of species reaching 200. L1, BC and rJSD can also reveal the difference between the two abundance types yet they were not affected by the species-level richness. L2 distance between the two abundance types dramatically dropped with the increase in the species-level richness. In the complex community with the number of species over 200, L2 distance metric almost lost the discriminatory power of these two abundance profiles.

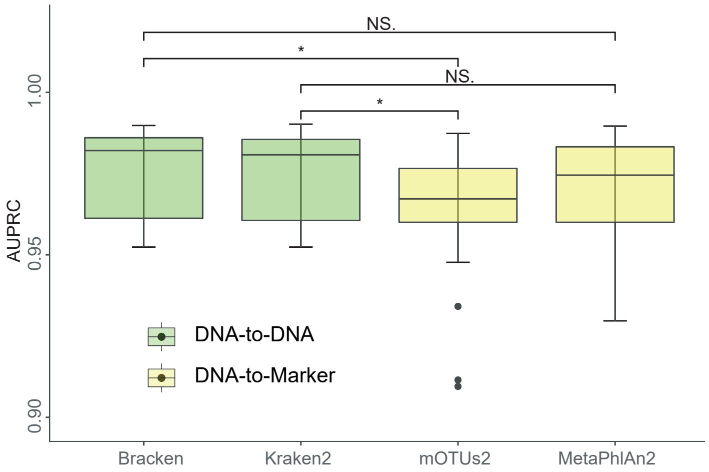

Figure S3. The AUPRC of four representative metagenomic profilers on 25 simulated communities. The boxplot indicates the AUPRC based on the default profiling results (without any abundance thresholding) of four profilers using either sequence abundance (green) or taxonomic abundance (yellow) as the ground truth. n = 25 simulated datasets. Significance levels: p-value<0.05 (*), <0.01 (**), <0.001 (***), NS (non-significance); two-sided Wilcoxon signed-rank test, p-values are 0.166, 0.019, 0.196, and 0.01 (from top to bottom) in the figure. The lower and upper hinges correspond to the first and third quartiles, and the center refers to the median value. The upper (or lower) whisker extends from the hinge to the largest (smallest) value no further (at most) than 1.5*IQR from the hinge. Data beyond the end of the whiskers are plotted individually.

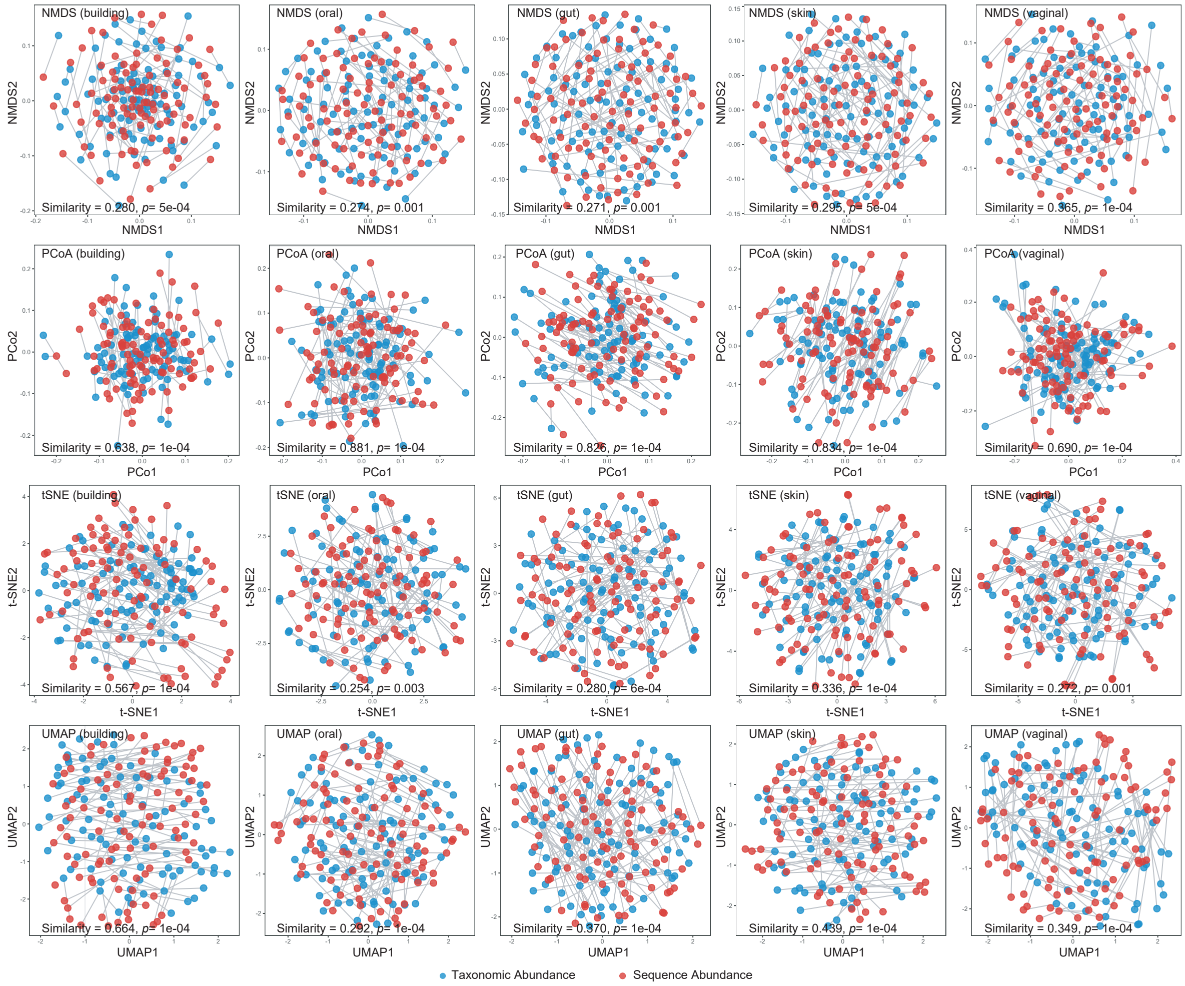

Figure S4. Ordination analyses of simulated profiles based on BC. Scatter plots of NMDS, PCoA, t-SNE and UMAP illustrate the dissimilarities between the sequence abundance (red dots) and taxonomic abundance (blue dots), which are the ground truth of the simulated profiles of 100 build environment, 100 oral, 100 skin, and 100 vaginal samples. Bray-Curtis (BC) distance was used to for the ordination analyses. The two abundance types from the same profile were connected using grey lines to show the difference of beta diversity.

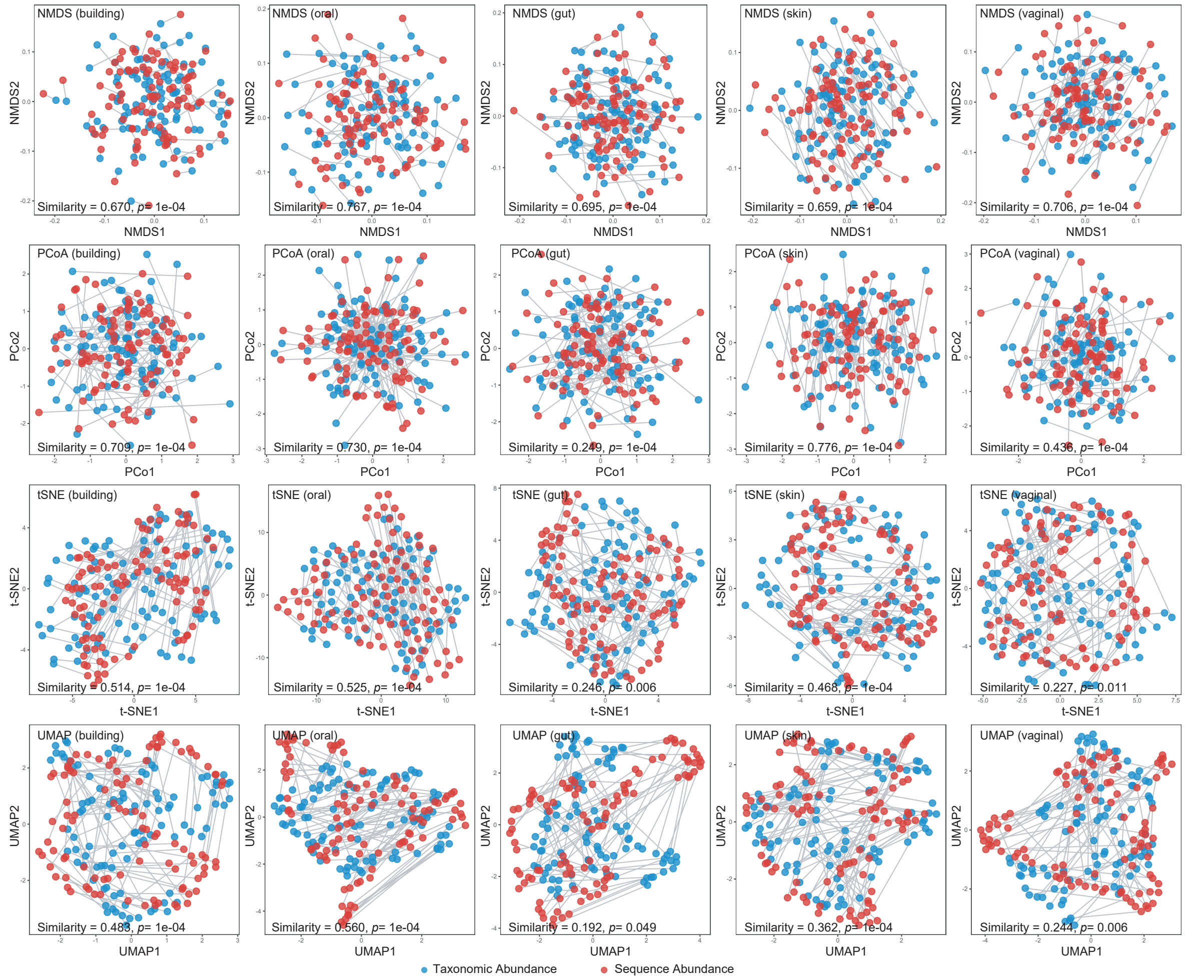

Figure S5. Ordination analyses of simulated profiles based on rAD. Scatter plots of NMDS, PCoA, t-SNE and UMAP illustrate the dissimilarities between the sequence abundance (red dots) and taxonomic abundance (blue dots), which are the ground truth of the simulated profiles of 100 build environment, 100 oral, 100 skin, and 100 vaginal samples. Robust Aitchison distance (rAD) was used to for the ordination analyses. The two abundance types from the same profile were connected using grey lines to show the difference of beta diversity.

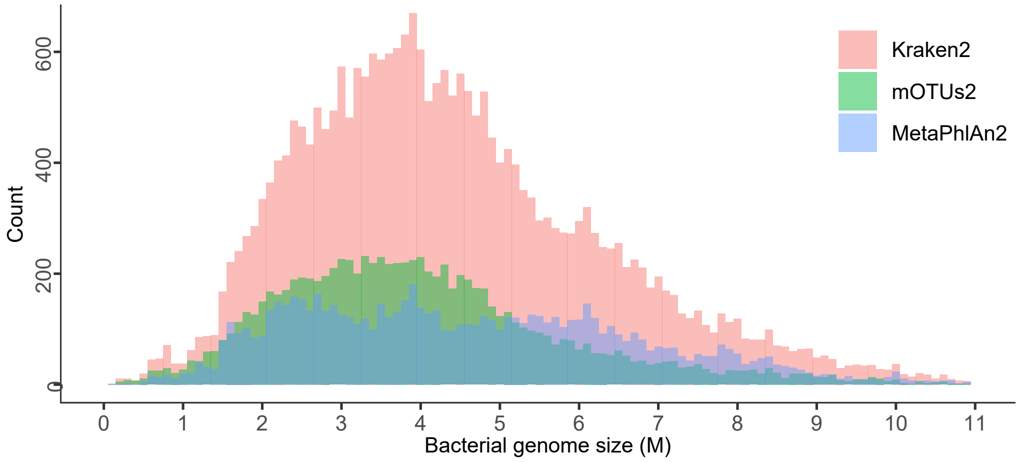

Figure S6. Genome size distribution of bacteria in the reference databases used in Kraken2, MetaPhlAn2, and mOTUs2.

Supplementary Reference

1. Breitwieser, F.P., Lu, J. & Salzberg, S.L. A review of methods and databases for metagenomic classification and assembly. *Briefings in bioinformatics* **20**, 1125-1136 (2019).

2. Ye, S.H., Siddle, K.J., Park, D.J. & Sabeti, P.C. Benchmarking metagenomics tools for taxonomic classification. *Cell* **178**, 779-794 (2019).