MetaKSSD Enables Rapid and Online Metagenomic Taxonomic Profiling and Profile Searching

Huiguang Yi*, Qing Chang, Xiaoxin Lu

Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China.

*Correspondence: yihuiguang@caas.cn (H.Y.)

Abstract

Metagenomic taxonomic profiling, which involves mapping microbiome sample reads or *k*-mers to a pre-built reference taxonomic marker database (MarkerDB) to estimate taxon composition and abundances, is a fundamental step for microbiome-related studies. The explosive growth of reference genomes and metagenomic data presents significant scalability and efficiency challenges for existing metagenomic profilers. Additionally, many microbiome researchers prefer online analysis platforms like Galaxy and One Codex due to unfamiliarity with command line tools, but uploading large-scale data can be frustrating, especially with low bandwidth. To meet these challenges, we introduce MetaKSSD, a highly scalable method for MarkerDB construction and metagenomic taxonomic profiling. MetaKSSD can profile approximately 10 GB of metagenome data within seconds while utilizing only 0.5 GB of memory and its MarkerDB encompasses 85,202 species using only 0.17 GB of storage. Extensive benchmarking demonstrated that MetaKSSD's accuracy was comparable to the state-of-the-art metagenomic profilers. In microbiome-phenotype association study, MetaKSSD identified more phenotype-associated bacterial species than the predominant profiler MetaPhlAn4. Utilizing MetaKSSD, we profiled 382,016 metagenomic runs from NCBI SRA and developed efficient methods for searching similar profiles among them. Notably, MetaKSSD's client-server architecture performs local sketching of metagenomic data to a compact sketch, allowing swift transmission to remote server for real-time online metagenomic analyses and profile searching, thus facilitating use by non-expert users.

1. **Introduction**

Metagenomic taxonomic profiling, which involves estimating the taxon composition and abundances of a microbiome sample using whole genome sequencing (WGS) data, is a fundamental step in various microbiome-related studies, such as microbiome-phenotype association studies[1–6], microbiome-based disease prediction[7,8], examining microbiome changes with age[9,10], and microbiome co-occurrence and co-abundance studies[11–13]. Typically, a metagenomic taxonomic profiling method (also called a profiler) estimates taxon abundances through mapping reads or *k*-mers of samples to a pre-built reference taxonomic marker database (termed MarkerDB hereafter) [14,15].

MarkerDB is normally constructed from reference genomes. Different profilers have different methods for processing reference genomes, making MarkerDB from one profiler incompatible for another. Most profilers, such as MetaPhlAn[9,16] and mOTUs[17–19], do not support custom MarkerDB construction. However, a few profilers, like Bracken[20], do support it. The comprehensiveness of the MarkerDB is crucial, as a profiler can only detect taxa that exist in its MarkerDB. To maintain competitive performance, a profiler needs to update its MarkerDB periodically to cover taxonomic informative markers from as diverse genomes as possible, especially metagenome-assembled genomes (MAGs)[21–23]. However, low-quality, such as potentially misassembled genomes, should be excluded. The taxonomic annotation for the reference genomes is also important. There are two systems for taxonomic annotation: NCBI taxonomy[24] and Genome Taxonomy Database (GTDB) taxonomy. NCBI taxonomy is the traditional taxonomy system adopted by most profilers[25,26], whereas GTDB is newer system modified from NCBI taxonomy to provide rank-normalized, phylogenetically consistent taxonomy[27–30].

In the past decade, more than 20 metagenomic profilers have emerged, each with its strengths at the time of release[14,25,26]. However, as the number of reference genomes is growing exponentially[31], most profilers face challenges in MarkerDB scalability. Bracken2[20], a profiler that relies on the reads' classifier Kraken2[32,33], is perhaps an early victim of this trend, as its standard MarkerDB has grown to about 80GB[34]. Such a size poses difficulties for downloading in poor network conditions and running in environments with limited memory resources. The predominant profiler, MetaPhlAn4[9], has a MarkerDB of 20 GB, which has increased five times since its previous iteration[16], and will soon face similar issues. An exception is mOTUs[17–19], whose MarkerDB is only 3.5GB and grows slowly with iterations.

With the continually dropping of sequencing cost, large-scale metagenomic profiling is also a trend. Bracken2, one of the fastest profilers revealed in the second round of Critical Assessment of Metagenome Interpretation (CAMI2), requires about 0.66 hour to profile 100GB of data[25]. However, the data from a medium-scale microbiome sequencing project can easily surpass 5TB, requiring significant time for merely profiling. Furthermore, there are over 500,000 WGS metagenomic runs (total size estimated to be over 3PB) in the NCBI Short Read Archive (SRA)[35,36] as of this study. Metagenome profiling at this scale is necessary in the near future for

constructing large models[37] for microbiome data (Discussion). However, this task is far beyond the capacities of any existing profilers.

Metagenomic profilers have a very wide user base; however, a significant proportion of potential users, such as clinical or wet lab scientists, may not be accustomed to command line tools, which are the primary interface for most metagenomic profilers. Alternatively, they switch to online platform such as Galaxy[38] and One Codex[39] for metagenomic analysis. However, in addition to potential data security problems, uploading raw sequencing data can be time-consuming or frequently fail, especially with limited bandwidth.

To meet these challenges, we present MetaKSSD, a metagenomic profiler developed upon the genome sketching technique, K-mer Substring Space Decomposition (KSSD). KSSD is able to downsample a (meta)genome to a subset of *k*-mers (also called a sketch) with a reduction of thousands of times, while maintaining the ability to accurately estimate metagenome containment using sketches[40]. KSSD uniquely enables lossless sketch operations such as union, intersection, and subtraction, which are fundamental for various sketch-based analyses. Building upon these basic functionalities, MetaKSSD additionally introduces a feature that tracks *k*-mer counts within the sketch, and redefines the processes of MarkerDB construction and metagenomic taxonomic profiling as a series of sketch operations (Results).

We demonstrate here that MetaKSSD not only offers excellent MarkerDB scalability but also enables large-scale metagenomics profiling with ease. More importantly, the transmission of sample sketches over the Internet is swift, avoiding bandwidth bottlenecks. In the client-server mode, MetaKSSD realizes real-time online metagenomic profiling and instantaneous searches across hundreds of thousands of profiles for similar ones. This opens the door to a technique we call “sketch communications”, allowing users to search, compare, or share their sample sketches and profiles with others immediately (Discussion).

2 Results

2.1 Overview of MetaKSSD’s algorithm and main features

MetaKSSD redefines MarkerDB construction and metagenomic taxonomic profiling using sketch operations, including union, intersection, and subtraction.

- **MarkerDB Construction**: Reference genomes are first sketched and organized by species (Fig. 1a). Inspired by the 'pangenome' concept[41], a 'pangenome sketch' for each species is generated by uniting all sketches from that species. Species-specific markers are created by subtracting the union of pangenome sketches of other species. All species-specific markers are then consolidated into the MarkerDB (Fig. 1b, Methods).
- **Taxonomic Profiling**: For profiling, a sketch with tracked *k*-mer counts is created for the sample. The overlapping *k*-mer counts between this sketch and the species-specific markers in MarkerDB are used to compute summary statistics, which are then normalized to estimate the relative abundances for all species (Fig. 1c, Methods).

MetaKSSD operates on sketches, making its MarkerDB orders of magnitude smaller than using original reference genome[40]. For example, MetaKSSD's MarkerDB spans 85,202 species from 402,709 genomes (GTDBr214), far surpassing MetaPhlAn4 and mOTUs3 (Fig. 2a). It requires only 0.17GB storage for MarkerDB and 0.5GB memory for profiling, vastly outperforming MetaPhlAn3 in both metrics (Fig. 2b). Therefore, MetaKSSD achieves a MarkerDB scalability, defined by the ratio of species number to storage usage, of 501,628 species per GB, which is 53 times higher than its closest competitor mOTUs3 (9,429 species per GB).

MetaKSSD also enables rapid metagenomic profiling. For example, it profiled an 11 GB metagenome (ERR598952) in 11 seconds, 10 times faster than Bracken2, the fastest method from CAMI2[25]. Additional tests showed MetaKSSD being 4.6 to 17 times faster than Bracken2, which itself significantly faster than other state-of-the-art profilers (Fig. 2c). As will demonstrate later, MetaKSSD's remarkable computational efficiency allows unprecedentedly large-scale metagenomic profiling.

In addition to command-line mode, MetaKSSD supports a client-server architecture, enabling local sketching and server-based analysis (Code Availability). With a 4,096-fold dimensionality reduction, the typical metagenome sketch size is under 2MB, allowing swift online transmission. Tests showed the MetaKSSD client took 15 seconds to sketch a 2GB metagenome and 30 seconds for full analyses, significantly outperforming Galaxy[38] and One Codex[39] (Fig. 2d).

In principle, this significant size reduction should not compromise the profiling quality. MetaKSSD estimates species abundance using read coverages on species-specific *k*-mers sampled across genomes[40]. Each species in the 0.17GB MarkerDB occupies 1,820 bytes on average, equivalent to 455 *k*-mers, sufficient for most cases except highly strain-diverse samples (Discussion).

2.2 Benchmarking for MetaKSSD

Utilizing the Open-community Profiling Assessment tooL (OPAL)[42], we conducted a comprehensive evaluation of MetaKSSD against the top-performers identified in the CAMI2 taxonomic profiling challenges[25,26], focusing on their latest iterations, including MetaPhlAn3[16], MetaPhlAn4[9], mOTUs3[17] and Bracken2.8[20,32] (by speed). These profilers were first benchmarked on four synthetic metagenome datasets from CAMI2[25], representing diverse environmental contexts: Mouse gut (n=49 samples), Rhizosphere (n=21 samples), Marine (n=10 samples), and Strain_madness (n=100 samples) (Data Availability). Other profilers previously benchmarked in CAMI2 on these datasets were also included to provide a broader landscape for comparison. MetaKSSD was benchmarked across various parameter settings of K (the half-length of K-mer) and S (the minimum overlapped k-mer between a metagenome sketch and MarkerDB), while keeping L = 3 fixed (representing 4,096-fold dimensionality reduction). Since MetaKSSD uses GTDB taxonomy, whereas other profilers and the ground truth use NCBI taxonomy, MetaKSSD's results were converted to the NCBI taxonomy to ensure comparability (Methods).

The evaluation considered four OPAL metrics at the species level:

1. **Purity (Specificity)**: The percentage of correctly predicted species among all predicted species in a metagenome sample.
2. **Completeness (Sensitivity)**: The percentage of correctly predicted species among all species in a metagenome sample.
3. **L1 Norm Error**: The estimation accuracy of species abundance.
4. **Weighted Unifrac Error**: The estimation accuracy of abundance across all taxonomic ranks.

It showed no single profiler outperformed all others across all datasets, but the top-performers consistently emerged from the MetaKSSD, MetaPhlAn, or mOTUs families. The sole exception was Dudes[43], which achieved the highest purity in 'Strain_madness' dataset. Notably, MetaKSSD excelled in 'Mouse_gut' dataset, ranking highest in purity and completeness (e.g., MetaKSSD.L3K11S48 with purity: 0.87 ± 0.01, completeness: 0.976 ± 0.003). MetaKSSD also ranked among the top in taxa abundance estimation accuracy (L1 norm error: 0.448 ± 0.04, weighted unifrac error:1.24 ± 0.0789), comparable to MetaPhlAn4 (L1 norm error: 0.412 ± 0.0427, weighted unifrac error:1.19 ± 0.0795). It also achieved the highest accuracy in 'Rhizosphere' dataset and was a top-tier profiler in 'Marine' dataset. However, MetaKSSD showed only mediocre performance in the 'Strain_madness' dataset (Discussion) (Fig. 3).

MetaKSSD's performance on ‘Rhizosphere’, ‘Marine’, and ‘Strain_madness’ datasets may have been underestimated due to the anonymous nature of source genomes, preventing precise GTDB to NCBI taxonomy conversion (Methods). Substantial errors could arise from the differences of taxonomy systems, rather than genuine inaccuracies. In contrast, ‘Mouse_gut’ dataset provided genome IDs, enabling precise taxonomy conversion and improved performance.

Later-developed profilers generally outperformed older ones, as seen with MetaPhlAn4 and mOTUs3 versus MetaPhlAn2,3 and mOTUs1,2. These improvements might be due to the inclusion of CAMI2 source genomes in later-developed MarkerDBs, rather than methodological innovations, as noted in previous benchmarking studies[14,15]. To address this, we constructed five synthetic metagenome samples (termed 'New released' dataset) using genomes absent from all MarkerDBs (Tab. S1) for further benchmarking (Methods). On the 'New released' dataset, MetaKSSD outperformed other top profilers in almost all metrics (Fig. 3), with the sole exception being Bracken2.8, which achieved the highest completeness but the lowest purity. This may be due to Bracken2.8 using all taxonomically informative *k*-mers from whole reference genomes[33], while others used only quality-controlled marker genes (MetaPhlAn and mOTUs families) or sketches of markers (MetaKSSD).

While Fig. 3 depicts species-level evaluations, MetaKSSD achieved similar performance across other taxonomic ranks (Figs. S1-5). Overall, while no single metagenome profiling method is best for all situations, MetaKSSD emerges as one of the top-performing methods in most cases.

2.3 MetaKSSD improves microbiome-phenotype association study

The unprecedentedly comprehensive MarkerDB of MetaKSSD presumably allows for more refined metagenome profiling. However, this advantage has not fully materialized due to the GTDB to NCBI taxonomy conversion in the benchmarking experiments (Result Section 2.2). To explore its capabilities with the native GTDB taxonomy system, MetaKSSD was applied to a

microbiome-phenotype association study within the BGInature2012 cohort[44], and compared with Metaphlan4, the predominant metagenome profiling method for microbiome-phenotype association studies[1,45,46].

The BGInature2012 cohort was chosen for its well-characterized and fully open-access nature (Data Availability), unlike other cohorts that typically require an application and review process. The dataset includes stool microbiome WGS data and related metadata for 368 Chinese individuals. The metadata encompass gender, age, and 12 physiological phenotypes: Diabetic status, Body Mass Index (BMI), Fasting Blood Glucose (FBG), Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Fasting Serum Insulin (FINS), Fasting Serum C-peptide (FCP), Glycosylated Hemoglobin HbA1c (HbA1c), Triglycerides (TG), Total Cholesterol (TCHO), High-Density Lipoprotein (HDL), and Low-Density Lipoprotein (LDL). Based on the metagenome profiles generated by both MetaKSSD and MetaPhlAn4 for all the 368 samples, the species significantly associated with each of these phenotypes were identified (Methods).

Out of all the 12 phenotypes examined, MetaKSSD identified associations for 9 phenotypes, while MetaPhlAn4 identified associations for 7 phenotypes, missing findings in FBG and LDL. Neither method found associations in the three phenotypes—DBP, TCHO, and BMI—thus, they were excluded from our comparisons. Interestingly, associations emerged upon converting BMI to obesity status based on Chinese obesity criteria (BMI >= 28), therefore, BMI was replaced by obesity status, resulted in total 10 phenotypes in the comparisons. MetaKSSD consistently identified more associations than MetaPhlAn4 across the 10 phenotypes (Fig. 4a, b). In total, MetaKSSD identified 49 associated species, comprising 23 knowns and 26 unknowns (yet-to-be cultivated or named, detailed in Methods), whereas MetaPhlAn4 identified 28, consisting of 23 knowns and 5 unknowns (Tab. S2). Therefore, the increased number of associated species by MetaKSSD is mainly contributed by unknown species, a trend that persists across all phenotypes (Fig. 3b).

The analysis of phenotypic variations (Methods) showed that a greater number of species consistently explained a higher percentage of phenotypic variations across all the phenotypes and methods (Fig. 4b, c). Specifically, species identified by MetaKSSD consistently explained a higher percentage of phenotypic variations than those by Metaphlan4 across all phenotypes (Fig. 4c), affirming the effectiveness of the increased number of species identified by MetaKSSD.

Subsequently, the phenotypic variations solely explained by the unknown species (excluding the part explained by both of known and unknown species, designated as "unknown added") were calculated by subtracting the phenotypic variations explained by known species from the total phenotypic variations (Methods). The "unknown added" phenotypic variations constituted a substantial proportion of the total phenotypic variations if the unknown species present (Fig. 4b, c), affirming their effectiveness. Compared to MetaPhlAn4, MetaKSSD's unknown species explained significantly more phenotypic variations across most phenotypes, except for diabetic status, where there was a tie (Fig. 4c).

There are 8 common species detected by both profilers, accounting for only 16% (8 out of 49 species) and 29% (8 out of 28 species) of all species identified by MetaKSSD and MetaPhlAn4, respectively. However, extensive connections between the two profilers' results were observed at

higher taxonomic ranks (Fig. 4a, Tab. S3). Specifically, ninety percent of the species identified by MetaKSSD (44 out of 49 species) have a counterpart identified by MetaPhlAn4 for the same phenotype that is from the same class, while vice versa, 96% (27 out of 28 species) also hold true (Fig. 3d), indicating that the results of the two methods support each other. This observation is aligned with the fact that the GTDB taxonomy[27–30], used by MetaKSSD, is much more consistent at the class rank than at the species rank, compared to the NCBI taxonomy used by MetaPhlAn4.

In summary, these results indicate that MetaKSSD enhances microbiome-phenotype association studies by identifying a significantly larger number of effective phenotype-associated species, and a large proportion of them are unknown species that were overlooked by previous studies. Notably, findings in this study were extensively evidenced by other studies (Tab. S3). For example, both MetaKSSD and Metaphlan4 methods found *Roseburia intestinalis* negatively correlated with diabetes, coinciding with previous researches[47,48]. Similarly, the association detected by MetaKSSD between *Parabacteroides merdae* and fasting serum insulin was also found by Qiao et al.[49] These supporting studies affirms the reliability of our microbiome-phenotype association analysis process as a whole.

2.4 Large-scale metagenome profiling and profile searching with MetaKSSD

MetaKSSD was applied for the analyses of metagenomic data publicly available in NCBI SRA. Out of the 517,829 WGS metagenomic runs retrieved as of this study, 382,016 runs passed our quality control process and were successfully sketched and profiled by MetaKSSD, which spans 312, 664 samples,16,398 projects and 2,025 environments (detailed in Methods, see Tab. S4 for runs information and Data Availability for all sketches and profiles). Analyzing over all the profiling results showed that the number of samples versus the detected species (and genera) follows a long-tail distribution (Fig. S6). Specifically, thresholding the minimum MarkerDB overlapping *K*-mer at 18, 303,406 (97%) samples contain fewer than 1,000 species. However, 426 (0.14%) samples contain over 10,000 species, which are mostly from the environments of soil, rhizosphere or wetland, agrees with previous studies[50,51](Tab. S4).

Fifty-three commonly studied metagenomic environments were identified (encompassing at least 30 projects, detailed in Methods; see Tab. S5 for the list), which accounts for 49% of the total samples. In these environments, the 'bovine gut' exhibited the highest median species number (median m = 671), significantly exceeding those of the 'mouse gut' (m = 164) and 'human gut' (m = 145). Considering a larger proportion of 'bovine gut' *k*-mers are absent from the reference database, the extent of this exceeding is still underestimated. While 'wetland' included the sample with the largest species number (24,554 species for sample SAMN20864071), it had a relatively lower median species number (m=327.5). Although 'annelid' (m = 3) and 'coral' (m = 10) had the smallest median species numbers, there is likely a substantial proportion of hidden species in these environments not covered by the reference database (Fig. 5a). Similar rankings of environments were observed when using a threshold of S = 48 or at the genus level (Fig. S7). Furthermore, all detected species, after controlling for suspicious findings (Methods), were ranked by the number of environments in which they are present. The top species, as expected, are well-characterized species such as *Escherichia coli*, *Klebsiella pneumoniae*, and *Salmonella enterica* (Fig. S8). In total 17,742 environments-specific species were identified, spanning all the

fifty-three environments (Tab. S6), which may serve as an important reference for studies on prokaryote habitats[52]. These results provide a general characterization of the landscape of existing metagenome data, which has been difficult to achieve using conventional metagenomic profilers.

T-SNE (t-distributed Stochastic Neighbor Embedding) analysis on a subset of 151 well-characterized projects (spanning 30,592 samples and 17 different environments; see Tab. S7 for accessions, see Methods for enrollment criteria) showed clear clustering of samples by their originating environments, though sub-clusters were frequently observed, such as in mouse feces and air samples (brown and red dots in Fig. 5b, respectively). Further analysis showed that the sub-clusters within an environment were clearly grouped by projects (Fig. S9), which could be raised by either batch effects (i.e., technical causes) or the existence of uncharacterized project-specific sub-environments (i.e., biological causes). To investigate the main cause of sub-clustering, t-SNE analysis was applied to a subset of human stool samples related with lifestyles, which have even better characterization of originating environments (Tab. S8, initially adopted by the MetaPhlAn4 work[9]). It revealed a clear distinction between westernized (16,669 samples, 60 projects) and non-westernized (560 samples, 8 projects) lifestyles, with the latter forming a distinct cluster (Fig. 5c). This result demonstrated similar performance in discriminating westernized and non-westernized clusters compared to the MetaPhlAn4 result (as shown in Fig. S4c from the MetaPhlAn4 paper). Further dissection of non-westernized samples found that samples from seven different projects were mixed together, showing no significant project-specific batch effects. The only exception was PRJNA392180 (green dots in Fig. 5d), comprising 40 Hadza hunter-gatherer gut microbiome samples, which formed a distinct cluster. Previous study using a different dataset supported that Hadza hunter-gatherers indeed have a unique gut microbiome profile[53]. These clues jointly suggest that the sub-clusters discovered using MetaKSSD are more likely a result of the existence of biologically meaningful (often uncharacterized) sub-environments rather than batch effects.

To optimize computing efficiency, the abundance profiles were encoded as sparse vectors and saved in binary file format. Consequently, the entire abundance vector database, encompassing all 382,026 runs, used merely 400 MB of storage (Data Availability). Given an abundance vector, MetaKSSD can immediately identify the most similar abundance vectors from the database. This efficient search functionality is achieved using two optional distance (or similarity) measures: 1) normalized L1 norm and 2) cosine of angle, between the query and each abundance vector in the database (Methods). Both measures can effectively group samples by their originating environments using the Multiple Dimensional Scaling (MDS) method, as demonstrated with the test abundance vectors (Fig. S10, Tab. S9). This supports the use of these two measures for profile searching. The abundance vector searching functionality is particularly useful in a client-server mode (Code Availability), enabling users without bioinformatics skills to easily explore similar profiles in the database for their samples (Discussion).

3. Discussion

The performance of a metagenomic profiler is limited by the comprehensiveness of its MarkerDB[14,15,54]. However, preparing and updating MarkerDB are cumbersome and often significantly lags behind the release of exponentially growing reference genomes[14,31,55]. The

MetaKSSD method proposed in this study provides a swift and highly scalable approach for MarkerDB preparation, allowing for the construction of a more comprehensive MarkerDB while using significantly fewer computational resources. Benefiting from the comprehensiveness of its MarkerDB, MetaKSSD improves microbiome-phenotype association studies over the-state-of-the art method.
Compared to existing profilers, MetaKSSD enhances time and memory efficiency by orders of magnitude without compromising quantification accuracy. These advancements facilitate large-scale metagenome profiling and data-driven discoveries. For example, uncharacterized groups in the clustering of large-scale metagenome profiles suggest possible yet-to-be-discovered niches, such as unknow life-styles or health statuses, though further sample surveys and experimental validations are indispensable. Furthermore, a very recent research showcased that a large "language" model (LLM) trained on large-scale single-cell profiles can broadly benefit the field of single-cell biology[56]. With the exponential growth of metagenome data[55], a similar metagenomic LLM could be trained based on the large-scale MetaKSSD profiles in near future. In this analogy, a profile is a 'sentence' and species abundances are the 'words'. Such a metagenomic LLM could be used for species abundance enhancement, imputation, batch effect correction, sample annotation, and other applications, as demonstrated in the example of single-cell LLM[56].

Before profiling, MetaKSSD reduces raw data to a sketch typically smaller than 2 MB, allowing them to be swiftly transmitted over Internet regardless of the bandwidth bottleneck. This feature makes MetaKSSD (the client-server mode) the first real-time online metagenome analysis platform. This platform is designed to assist as many users as possible for metagene analysis, including those with no bioinformatics skills, e.g. aid physicians or lab medicine for the clinical usage[57]. Herein, metagenomic analysis can be done promptly by simply selecting and summitting FASTQ files in the processing window (see Code Availability for MetaKSSD clients). There is no technical barrier to extend this platform to support various types of omics data, such as genomic, transcriptomic, or epigenomic data, which would transform it into an omics sketches hub. Utilizing the established distance estimation functionalities for sketches[40] and profiles (Methods), this platform would allow users to search, compare and share sketches and profiles promptly with each other. However, addressing the related privacy and data security concerns is still part of our future plans.

MetaKSSD still faces some limitations. MetaKSSD estimates species abundance using the average of the 98th and 99th percentiles of MarkerDB overlapped *k*-mer counts (Methods). This approach aims to provide a simplified estimation of species abundance when a metagenome sample potentially contains two or more strains of a species, without involving complex statistical modeling. Though this measurement performed well for most of the testing datasets (typically containing fewer than three strains per species), it significantly underperformed compared to other state-of-the-art methods on the 'strain-madness' dataset, which represents a highly strain-diverse dataset (containing more than 20 strains per species)[25]. Fortunately, a species from real-world microbiome is typically dominated by a single strain[58–61], which justifies the application of MetaKSSD. MetaKSSD has also been tested on 16S amplicon sequencing (using 16-fold dimensionality-reduction, L1K9) and long-reads data for metagenome profiling but showcased inferior performances compared to using WGS short reads data (not shown). Further improvements of MetaKSSD on these metagenomic data types are feasible but not

deemed worthwhile, since amplicon metagenomic profiling is already computationally efficient and long-reads metagenomic data is primarily used for assembly rather than profiling[62].

4. Methods

4.1 MarkerDB construction

All 402,709 genomes from the Genome Taxonomy Database release 08-RS214 (GTDBr214) were downloaded utilizing the accessions provided in the metadata files "bac120_metadata_r214.tar.gz" and "ar53_metadata_r214.tar.gz", accessible at https://data.gtdb.ecogenomic.org/releases/release214/214.0/. These genomes were sketched by MetaKSSD (Code Availability) using the following command:

```
metakssd dist -L <L3K11.shuf> -o <L3K11_sketch> <all genomes Dir>
```

where "L3K11.shuf" is the parameter file defining the dimensionality reduction level (L) as 3 and the *k*-mer length as 22 (the dimensionality reduction rate is 16^L, resulting in 4096 when $L = 3$, and the *k*-mer length is twice the value of K, equating to 22 when $K = 11$). The resulting sketch of all the genomes was stored as "L3K11_sketch".

The genome names within the sketch were print line by line and direct the output to a file named "genome_name.txt" using the following command:

```
metakssd set -P <L3K11_sketch> > <genome_name.txt>
```

Then a grouping file named "group_name.txt" were prepared as follow: The "group_name.txt" file should contain the same lines as "genome_name.txt", where each line specifies the species name of the genome corresponding to the line in "genome_name.txt". Each line follows the format: "ID<TAB>species_name\n", where the ID is a non-negative integer uniquely labeling the species_name, and the species name of the genome can be found in the metadata files ".*_metadata_r214.tar.gz". For example, "1 Escherichia_coli\n" represents the species name "Escherichia coli" labeled with ID 1. The ID 0 is reserved for excluding genomes from the resulting sketch. Once "group_name.txt" is prepared, genomes can be grouped by their originating species using the following command:

```
metakssd set -g <group_name.txt> -o <L3K11_pan-sketch> <L3K11_sketch>
```

Here, "L3K11_pan-sketch" represents the consolidated ‘pangenome’ sketches for all species (Fig. 1b).

Subsequently, the union sketch of all species-specific marker, named "L3K11_union_sp-sketch", was obtained using this command:

```
metakssd set -q -o <L3K11_union_sp-sketch> <L3K11_pan-sketch>
```

Finally, the MarkerDB named “markerdb_L3K11” was generated by overlapping "L3K11_union_sp-sketch" with "L3K11_pan-sketch" using this command:

```
metakssd set -i <L3K11_union_sp-sketch> -o <markerdb_L3K11> <L3K11_pan-sketch>
```

4.2 Metagenome profiling and summary statistics for read coverage

For metagenome profiling, the microbiome sample, named “sample1.fastq” for example, is initially sketched with counts tracked using MetaKSSD (with ‘-A’ option, see Code Availability):

```
metakssd dist -L <L3K11.shuf> -A -o <sample1_sketch> <sample1.fastq>
```

Here, parameter file “L3K11.shuf” imposes a 4,096-fold dimensionality reduction for sketching.

Subsequently, a table detailing read coverage statistics for all identified species could be generated with this command:

```
metakssd composite -r <markerdb_L3K11> -q <sample1_sketch> > <species_coverage.tsv>
```

Where "markerdb_L3K11" is the pre-computed MarkerDB (Methods section 4.1).

The resulting table, “species_coverage.tsv”, comprises seven columns: sample name, species name, number of overlapping *k*-mer, mean read coverage, average of 98th and 99th percentile of read coverage, median read coverage and maximum read coverage, respectively. The fifth column (average of the 98th and 99th percentile of read coverage) was normalized to the relative abundances. This choice was made based on the performances of different summary statistics (Fig. S11), as well as the following considerations:

1. The microbiome may contain multiple strains within a single species, resulting in stacked read coverages for common *k*-mers. Therefore, using higher percentile read coverages provides a more accurate estimation of species-level read coverages.
2. However, the top 1% of *k*-mer read coverages are likely attributed to repetitive regions and are therefore excluded to avoid potential biases.

The normalization was conducted with this command:

```
perl possion.kssd2out.pl < species_coverage.tsv > <minimum overlapped k-mer > > <species relative abundance profile>
```

Here, the "minimum overlapped *k*-mer" is set as an integer, defaulting to 18, to control false discovery of species. For species in MarkerDB Script “possion.kssd2out.pl” normalized the summary statistics of reads coverage to 1.

4.3 MetaKSSD Benchmarking and Taxonomic system conversion

The benchmark datasets comprise four CAMI2 challenge datasets: ‘Mouse_gut’, ‘Rhizosphere’, ‘Marine’, and ‘Strain_madness’ (Data Availability) along with the ‘New_released’ dataset.
The ‘New_released’ dataset (n = 5 samples) was simulated using CAMISIM[63], similar to other CAMI2 challenge datasets (see https://github.com/CAMI-challenge/CAMISIM/). However, this dataset employed 162 newly released bacterial genomes as the source genomes (see Tab. S1 for accessions). These 162 genomes were selected as follows: Complete genomes that released after the marker database release date of the latest methods (March 2022 for mOTUs3) were chosen from the latest bacteria “assembly_summary.txt” file (ftp://ftp.ncbi.nih.gov/genomes/genbank/bacteria/assembly_summary.txt), and genomes present in the GTDBr214 database (used by MetaKSSD) were explicitly excluded, ensuring the genomes enrolled are not included in all the methods’ marker database. This step yielded 5193 genomes. To prevent overrepresentation of genomes from a few extensively studied species such as *Salmonella enterica* and *Escherichia coli*, only species harboring 2 to 15 genomes out of the 5193 were selected. Furthermore, to avoid conflicts between NCBI taxonomy and GTDB taxonomy, only genomes from species with consistent taxonomy between NCBI and GTDB were chosen. This selection process resulted in a total of 162 species, with one genome randomly selected from each.

MetaKSSD used GTDBr214 taxonomy while all other methods and the ground-truth profile used NCBI taxonomy. To make the MetaKSSD result comparable, MetaKSSD profiles were converted to NCBI taxonomy before the evaluation. For the ‘New_released’ dataset, the taxonomy conversion is simply a one-to-one mapping from GTDB to NCBI species of the source genomes, since it was simulated using species consistent genomes as mentioned above. For ‘Rhizosphere’, ‘Marine’, and ‘Strain_madness’ datasets, source genomes of which are anonymous and their GTDB taxonomy cannot be tracked, the GTDB to NCBI species mapping were created based on all GTDB genomes using the maximal supports rule: If a GTDB species consists of genomes from multiple NCBI species (the GTDB/NCBI taxonomy for all genomes is available at https://data.gtdb.ecogenomic.org/releases/release214/214.0/*metadata_r214.tar.gz), then the NCBI species with maximal support genomes of this GTDB species are mapped (see Tab. S10 for the mapping scheme). Such a conversion scheme can make mistakes if a source genome is from a minor NCBI species. As a consequence, the performance of MetaKSSD might be underestimated on these datasets. For the ‘Mouse_gut’ dataset (n=49 samples), which has all its source genomes’ accession IDs disclosed, we created a GTDB to NCBI species mapping scheme for each sample using the taxonomies information of source genomes of this sample as

above. The taxonomy of the MetaKSSD result for each sample was convert to NCBI taxonomy using its corresponding mapping scheme.

MetaKSSD together with the latest iterations of profilers that not yet benchmarked by the previous CAMI2 challenge experiment (including MetaPhlan3, MetaPhlan4, Bracken2.8, and mOUT3) were performed on all the five benchmark datasets. The performance metrics were evaluated using the OPAL framework[42] at the species-level. For each of the four CAMI2 datasets, the results of these new methods were presented together with the CAMI2 challenge results from previous methods (each dataset may have the results from different set of methods), which can be collected from CAMI2 GitHub repositories: https://cami-challenge.github.io/OPAL/cami_ii_mg/ and https://github.com/CAMI-challenge/second_challenge_evaluation/tree/master/profiling.

4.4 Microbiome-phenotype association study

All 368 stool microbiome data within the BGInature2012 cohort[44] were downloaded from NCBI SRA accession: SRA045646. Metaphlan4 used the default parameters and the MarkerDB vJan21 (the version used in its original paper[9]) for profiling these data. MetaKSSD sketched these data using a 22-mer at a 4,096-fold dimensionality-reduction rate (with the parameter file "L3K11.shuf", see Methods section 5.2). Due to additional spurious control measures applied later, the MarkerDB overlapping threshold S was not set, to retain as many species as possible in MetaKSSD profiling.

After profiling, species relative abundances from individual taxonomic profiles were compiled into a matrix, with rows representing species and columns representing individual/sample IDs. To control for spurious species, those with a prevalence of less than 0.1 were removed, retaining only species with non-zero abundance in more than 36 individuals.

All phenotypic measurements of the BGInature2012 cohort were obtained from the Supplementary Tables of its original paper[44] (link: https://static-content.springer.com/esm/art%3A10.1038%2Fnature11450/MediaObjects/41586_2012_BFnature11450_MOESM372_ESM.xls). Missing values in these measurements were imputed with the mean of the available data. One of these phenotypes, body mass index (BMI), was converted to obesity status (0 or 1) using Chinese obesity criteria (BMI >=28)[64,65].

Microbiome-phenotype associations were detected across all species-phenotype pairs using a linear model[66] with R [67], including age and gender as covariates:

```
Phenotype ~ Age + Gender + Species Relative Abundance.
```

P-values of the associations were subjected to Benjamini-Hochberg correction[68] to control for multiple testing. An association was considered significant at a false discovery rate (FDR) < 0.2, according to the usage conventions for Metaphlan4[6,9].

Spearman correlations of these phenotype-species pairs were calculated using the 'cor' function[69] in R. To calculate the proportion of variation in a phenotype explained by the

significantly associated species, a phenotype was regressed on all significantly associated species using a linear model[66] with R:

```
fit <- lm(Phenotype ~ sp1 + sp2 + ... + spn),
```

where sp_1, sp_2, ..., sp_n represent the significantly associated species. The value `summary(fit)$r.squared` was reported as the variation proportion explained. The variation explained by known species was calculated in the same way. Known species are defined as species not prefixed with "_SGB" in the name for Metaphlan4 results, and species not prefixed with the pattern "_sp[0-9]{9}" for MetaKSSD results; otherwise, they are considered known species, following the naming conventions in Metaphlan4[9] and GTDB[27,29,30].

4.5 Sketching of WGS Metagenomic data from SRA

All WGS metagenomic runs available as of Dec. 27, 2023, were obtained using the following search query executed at NCBI SRA:

```
((((((Public[Access]) AND ("0000"[Modification Date] : "2023/12/27"[Modification Date])) AND "wgs"[Strategy]) AND "metagenomic"[Source] AND "cluster_public"[prop] AND "filetype fastq"[Properties])) AND "filetype fastq"[Filter]) AND "cloud s3"[Filter]
```

The returned run information, comprising of 517,829 runs, was saved in a file "SraRunInfo.csv" for the downstream analysis.

Subsequently, runs using undesired technologies, such as 16S or other amplicon sequencing methods, nanopore, pacbio, or solid sequencing were excluded through the execution of the following command:

```
awk -F ',' '$5>2e8 && $5 > $8*1e6 && $13 == "WGS" && $14 ~ /PCR|RANDOM|unspecified/ && $15 == "METAGENOMIC" && $19 !~ /NANOPORE|PACBIO|SOLID/ ' SraRunInfo.csv |grep -iv 'dbgap\|16S' | csvcut -c1 > accession.list
```

It resulted in 420,708 runs.

Runs were sketched in parallel on AWS instances for efficiency:

```
nohup cat accession.list | parallel -j 16 " aws s3 cp s3://sra-pub-run-odp/sra/{}/{} $TMPDIR/ && kssd221 dist -L  <L3K11.shuf> -A -o <outdir> --pipecmd 'fastq-dump --skip-technical --split-spot -Z' $TMPDIR/{} && rm $TMPDIR/{} " >out.log 2>&1 &
```

A quality control step was then implemented: Sketches smaller than 20 Kb were first removed, as these were likely amplicon or virus data lacking explicit annotation information or encountered failures during the sketching process. The remaining sketches were profiled using

MetaKSSD (see Methods section 5.2), and runs with no prokaryotic species were excluded. This process resulted in 382,016 high-quality sketches (Data Availability).

4.6 Identifying commonly studied environments

Commonly studied metagenomic environments (Tab. S5) were identified as follows:
Projects were grouped by the "Scientific_name" (i.e., environment) from all runs metadata (Tab. S4). Initially, 66 environments encompassed at least 30 projects, but 13 were excluded from the final list due to ambiguous annotation or duplication: Homo_sapiens, Mus_musculus, bioreactor_metagenome, feces_metagenome, food_metagenome, freshwater_metagenome, gut_metagenome, human_metagenome, insect_metagenome, marine_metagenome, sediment_metagenome, synthetic_metagenome, and viral_metagenome.

4.7 Controlling for suspicious species before ranking by environments

To rank species by the number of environments in which they are present, suspicious findings were controlled as follow: a species was considered present in an environment only if it was detected in more than 10% of the projects from that environment, and a species was considered present in a project only if it was detected in more than 10% of the samples from that project.

4.8 t-SNE analysis

Projects with at least 30 samples and originating from environments (as indicated by the "Scientific_name" column in the metadata Tab. S4) that include at least two different projects with at least 30 samples were first selected. These projects were then manually checked to ensure that the project descriptions were consistent with their "Scientific_name" from the SRA metadata. Projects with insufficient descriptions or samples from environments inconsistent with the SRA metadata were excluded. Finally, 151 projects were selected, spanning 30,592 samples and 17 different environments (Tab. S7).

The MetaKSSD (L3K11S18) profiles were consolidated into a species-by-sample matrix for each environment. For samples with multiple run profiles, only one run was randomly selected. For each profile matrix, the t-SNE[70] method from the ‘manifold’ module in the Python ‘sklearn package’[71] was used for dimensionality reduction, with parameters perplexity=30, init="pca", metric="l1", and n_iter=80000. The dimensionality reduction results were visualized using the R package ‘ggplot2’[72].

4.9 Construction of Abundance Vector Database, Abundance Vector Searching and Clustering

Every metagenome sketch produced in the previous step was utilized to generate a species relative abundances profile, represented as a sparse vector and encoded in a binary file format (*.abv file), optimizing search efficiency. This can be achieved through the following command:

```
metakssd composite -r <markerdb_L3K11> -q <metagenome sketch> -b -o < markerdb_L3K11> / “abundance_Vec”
```

Subsequently, all *.abv files were indexed, and the abundance vector database was created within the same directory as "markerdb_L3K11" using the command:

```
metakssd composite -r <markerdb_L3K11> -i
```

Due to the parsing of abundance vector database is MarkerDB dependent, the abundance vector database “abundance_Vec” were put under the path of MarkerDB “markerdb_L3K11”.

To quantify the distance or similarity between two abundance vectors $\boldsymbol{a} = (a_1,a_2..a_n)^T$ and $\boldsymbol{b} =(b_1,b_2..b_n)^T$, where *n* represents the total number of species in the MarkerDB, the normalized L1 norm $||\boldsymbol{a} - \boldsymbol{b}||_1$ and the cosine of the angle between the vectors (also termed as Cosine similarity) were calculated as follow:

Normalized L1 norm distance:

$$l1nom\ distance\ = \frac{||\boldsymbol{a} - \boldsymbol{b}||_1}{2} = \frac{\sum_{i=1}^{n}|a_i - b_i|}{2}$$

Cosine similarity:

$$\cos(\theta) = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{(\sum_{i=1}^{n} a_i^2)(\sum_{i=1}^{n} b_i^2)}}$$

Cosine similarity can be easily converted to Cosine distance by:

$$Cosine\ distance = 1 - \cos(\theta) = 1 - \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{(\sum_{i=1}^{n} a_i^2)(\sum_{i=1}^{n} b_i^2)}}$$

Abundance vectors similar to an abundance vector "input.abv" could be retrieved from database as follow:

```
metakssd composite -r <markerdb_L3K11> -s<0 or 1> <input.abv>
```

Here, the options -s0 and -s1 enable searching based on L1 norm and cosine similarity, respectively.

For abundance vectors clustering, multidimensional scaling (MDS) method[73] was employed with the pairwise normalized L1 norm- or Cosine-distance as input.

5. Data Availability

- MarkerDB of MetaKSSD (L3K11): https://zenodo.org/records/11437234/files/markerdb.L3K11_gtdb_r214.tar.gz
- Abundance vector database of MetaKSSD (L3K11): https://zenodo.org/records/11437234/files/markerdb.abvdb231227.L3K11_gtdb_r214.tar.gz
- All metagenome sketches (L3K11, as of 20231227) are available through zenodo records:10609030,10614425,10614597 and 10676887.
- The MetaKSSD profiles for all sketched runs are available through zenodo record: 11345411
- Four CAMI2 benchmark datasets: 'Mouse_gut' dataset is available at https://frl.publisso.de/data/frl:6421672/dataset/; 'Rhizosphere' 'Marine', and 'Strain_madness' datasets are available at https://frl.publisso.de/data/frl:6425521/.
- CAMI2 results for the five benchmarking datasets are available at yhg926.github.io/KSSD2/OPAL
- Microbiome-phenotype association study

Whole genome sequencing (WGS) data of stool microbiome for all 368 Chinese individuals from BGInature2012 cohort is available under accession SRA045646. The related metadata is available from Supplementary Tables of the source paper[44] at https://static-content.springer.com/esm/art%3A10.1038%2Fnature11450/MediaObjects/41586_2012_BFnature11450_MOESM372_ESM.xls.

6. Code Availability

- MetaKSSD (Linux):

https://github.com/yhg926/MetaKSSD

- MetaKSSD Clients (Mac OS):

http://www.genomesketchub.com/download/MetaKSSD_Mac.zip

- MetaKSSD Clients (Windows OS):

http://www.genomesketchub.com/download/MetaKSSD_Windows.exe

References

1. Gacesa, R. *et al.* Environmental factors shaping the gut microbiome in a Dutch population. *Nature* **604**, 732–739 (2022).
2. Wang, J. & Jia, H. Metagenome-wide association studies: Fine-mining the microbiome. *Nature Reviews Microbiology* vol. 14 Preprint at https://doi.org/10.1038/nrmicro.2016.83 (2016).
3. Kurilshikov, A. *et al.* Large-scale association analyses identify host factors influencing human gut microbiome composition. *Nat Genet* **53**, (2021).
4. Gilbert, J. A. *et al.* Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature* vol. 535 Preprint at https://doi.org/10.1038/nature18850 (2016).
5. Kishikawa, T. *et al.* Metagenome-wide association study of gut microbiome revealed novel aetiology of rheumatoid arthritis in the Japanese population. *Ann Rheum Dis* **79**, (2020).
6. Manghi, P. *et al.* MetaPhlAn 4 profiling of unknown species-level genome bins improves the characterization of diet-associated microbiome changes in mice. *Cell Rep* **42**, (2023).
7. Zhu, J. *et al.* Statistical modeling of gut microbiota for personalized health status monitoring. *Microbiome* **11**, (2023).
8. Gupta, V. K. *et al.* A predictive index for health status using species-level gut microbiome profiling. *Nat Commun* **11**, (2020).
9. Blanco-Míguez, A. *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat Biotechnol* **41**, (2023).
10. Ghosh, T. S., Shanahan, F. & O'Toole, P. W. The gut microbiome as a modulator of healthy ageing. *Nature Reviews Gastroenterology and Hepatology* vol. 19 Preprint at https://doi.org/10.1038/s41575-022-00605-x (2022).
11. Faust, K. *et al*. Microbial co-occurrence relationships in the Human Microbiome. *PLoS Comput Biol* **8**, (2012).
12. Ma, B. *et al.* Earth microbial co-occurrence network reveals interconnection pattern across microbiomes. *Microbiome* **8**, (2020).
13. Chen, L. *et al.* Gut microbial co-abundance networks show specificity in inflammatory bowel disease and obesity. *Nat Commun* **11**, (2020).
14. Ye, S. H., Siddle, K. J., Park, D. J. & Sabeti, P. C. Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell* vol. 178 Preprint at https://doi.org/10.1016/j.cell.2019.07.010 (2019).
15. Sun, Z. *et al.* Challenges in benchmarking metagenomic profilers. *Nat Methods* **18**, (2021).
16. Beghini, F. *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with biobakery 3. *Elife* **10**, (2021).

17. Ruscheweyh, H. J. *et al.* Cultivation-independent genomes greatly expand taxonomic-profiling capabilities of mOTUs across various environments. *Microbiome* **10**, (2022).
18. Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic marker genes. *Nat Methods* **10**, (2013).
19. Milanese, A. *et al.* Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat Commun* **10**, (2019).
20. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: Estimating species abundance in metagenomics data. *PeerJ Comput Sci* **2017**, (2017).
21. Zeng, S. *et al.* A compendium of 32,277 metagenome-assembled genomes and over 80 million genes from the early-life human gut microbiome. *Nat Commun* **13**, (2022).
22. Sánchez-Navarro, R. *et al.* Long-Read Metagenome-Assembled Genomes Improve Identification of Novel Complete Biosynthetic Gene Clusters in a Complex Microbial Activated Sludge Ecosystem. *mSystems* **7**, (2022).
23. Hauptfeld, E. *et al.* Integrating taxonomic signals from MAGs and contigs improves read annotation and taxonomic profiling of metagenomes. *Nat Commun* **15**, 3373 (2024).
24. Schoch, C. L. *et al.* NCBI Taxonomy: A comprehensive update on curation, resources and tools. *Database* vol. 2020 Preprint at https://doi.org/10.1093/database/baaa062 (2020).
25. Meyer, F. *et al.* Critical Assessment of Metagenome Interpretation: the second round of challenges. *Nat Methods* **19**, (2022).
26. Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation - A benchmark of metagenomics software. *Nat Methods* **14**, (2017).
27. Rinke, C. *et al.* A standardized archaeal taxonomy for the Genome Taxonomy Database. *Nat Microbiol* **6**, (2021).
28. Parks, D. H. *et al.* A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat Biotechnol* **38**, (2020).
29. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* **36**, (2018).
30. Parks, D. H. *et al.* GTDB: An ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res* **50**, (2022).
31. National Center for Biotechnology Information. RefSeq Growth Statistics. *https://www.ncbi.nlm.nih.gov/refseq/statistics/* (2024).
32. Lu, J. *et al.* Metagenome analysis using the Kraken software suite. *Nat Protoc* **17**, (2022).
33. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol* **20**, (2019).
34. Derrick Wood, S. S. J. L. F. B. C. P. A. Z. D. B. M. S. and B. L. Kraken 2, KrakenUniq and Bracken indexes. https://benlangmead.github.io/aws-indexes/k2.
35. Leinonen, R., Sugawara, H. & Shumway, M. The sequence read archive. *Nucleic Acids Res* **39**, (2011).
36. Katz, K. *et al.* The Sequence Read Archive: A decade more of explosive growth. *Nucleic Acids Res* **50**, (2022).
37. Hao, M. *et al.* Large Scale Foundation Model on Single-cell Transcriptomics. *bioRxiv* (2023).
38. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res* **46**, (2018).
39. One Codex. *https://www.onecodex.com/platform/*.

40. Yi, H., Lin, Y., Lin, C. & Jin, W. Kssd: sequence dimensionality reduction by k-mer substring space sampling enables real-time large-scale datasets analysis. *Genome Biol* **22**, (2021).
41. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial 'pan-genome'. *Proc Natl Acad Sci U S A* **102**, (2005).
42. Meyer, F. *et al.* Assessing taxonomic metagenome profilers with OPAL. *Genome Biol* **20**, (2019).
43. Piro, V. C., Lindner, M. S. & Renard, B. Y. DUDes: A top-down taxonomic profiler for metagenomics. *Bioinformatics* **32**, (2016).
44. Wang, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, (2012).
45. Zhernakova, A. *et al.* Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science (1979)* **352**, (2016).
46. Huang, R. Y. *et al.* Metagenome-wide association study of the alterations in the intestinal microbiome composition of ankylosing spondylitis patients and the effect of traditional and herbal treatment. *J Med Microbiol* **69**, (2020).
47. Karlsson, F. H. *et al.* Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498**, (2013).
48. Al-Jameel, S. S. Association of diabetes and microbiota: An update. *Saudi Journal of Biological Sciences* vol. 28 Preprint at https://doi.org/10.1016/j.sjbs.2021.04.041 (2021).
49. Qiao, S. *et al.* Gut Parabacteroides merdae protects against cardiovascular damage by enhancing branched-chain amino acid catabolism. *Nat Metab* **4**, (2022).
50. Bahram, M. *et al.* Metagenomic assessment of the global diversity and distribution of bacteria and fungi. *Environ Microbiol* **23**, (2021).
51. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A* **95**, 6578–6583 (1998).
52. Mise, K. & Iwasaki, W. Environmental Atlas of Prokaryotes Enables Powerful and Intuitive Habitat-Based Analysis of Community Structures. *iScience* **23**, (2020).
53. Schnorr, S. L. *et al.* Gut microbiome of the Hadza hunter-gatherers. *Nat Commun* **5**, (2014).
54. Breitwieser, F. P., Lu, J. & Salzberg, S. L. A review of methods and databases for metagenomic classification and assembly. *Brief Bioinform* **20**, (2018).
55. Pavlopoulos, G. A. *et al.* Unraveling the functional dark matter through global metagenomics. *Nature* **622**, (2023).
56. Hao, M. *et al.* Large-scale foundation model on single-cell transcriptomics. *Nat Methods* (2024) doi:10.1038/s41592-024-02305-7.
57. Chiu, C. Y. & Miller, S. A. Clinical metagenomics. *Nature Reviews Genetics* vol. 20 Preprint at https://doi.org/10.1038/s41576-019-0113-7 (2019).
58. Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure & genetic diversity from metagenomes. *Genome Res* **27**, (2017).
59. Costea, P. I. *et al.* Subspecies in the global human gut microbiome. *Mol Syst Biol* **13**, (2017).
60. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* **493**, (2013).

61. Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, (2017).
62. Marx, V. Method of the year: long-read sequencing. *Nature Methods* vol. 20 Preprint at https://doi.org/10.1038/s41592-022-01730-w (2023).
63. Fritz, A. *et al.* CAMISIM: Simulating metagenomes and microbial communities. *Microbiome* **7**, (2019).
64. Zhou, B. F. Predictive values of body mass index and waist circumference for risk factors of certain related diseases in Chinese adults--study on optimal cut-off points of body mass index and waist circumference in Chinese adults. *Biomed Environ Sci* **15**, (2002).
65. Pan, X. F., Wang, L. & Pan, A. Epidemiology and determinants of obesity in China. *The Lancet Diabetes and Endocrinology* vol. 9 Preprint at https://doi.org/10.1016/S2213-8587(21)00045-0 (2021).
66. Wilkinson, G. N. & Rogers, C. E. SYMBOLIC DESCRIPTION OF FACTORIAL MODELS FOR ANALYSIS OF VARIANCE. *J Appl Stat* **22**, (1973).
67. Becker, R. A., Chambers, J. M. & Wilks, A. R. The New S Language. *Biometrics* **45**, (1989).
68. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Series B Stat Methodol* **57**, (1995).
69. Kendall, M. G. A New Measure of Rank Correlation. *Biometrika* **30**, (1938).
70. Van Der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research* **9**, (2008).
71. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, (2011).
72. Wilkinson, L. ggplot2: Elegant Graphics for Data Analysis by WICKHAM, H. *Biometrics* **67**, (2011).
73. Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29**, (1964).

Figure legends

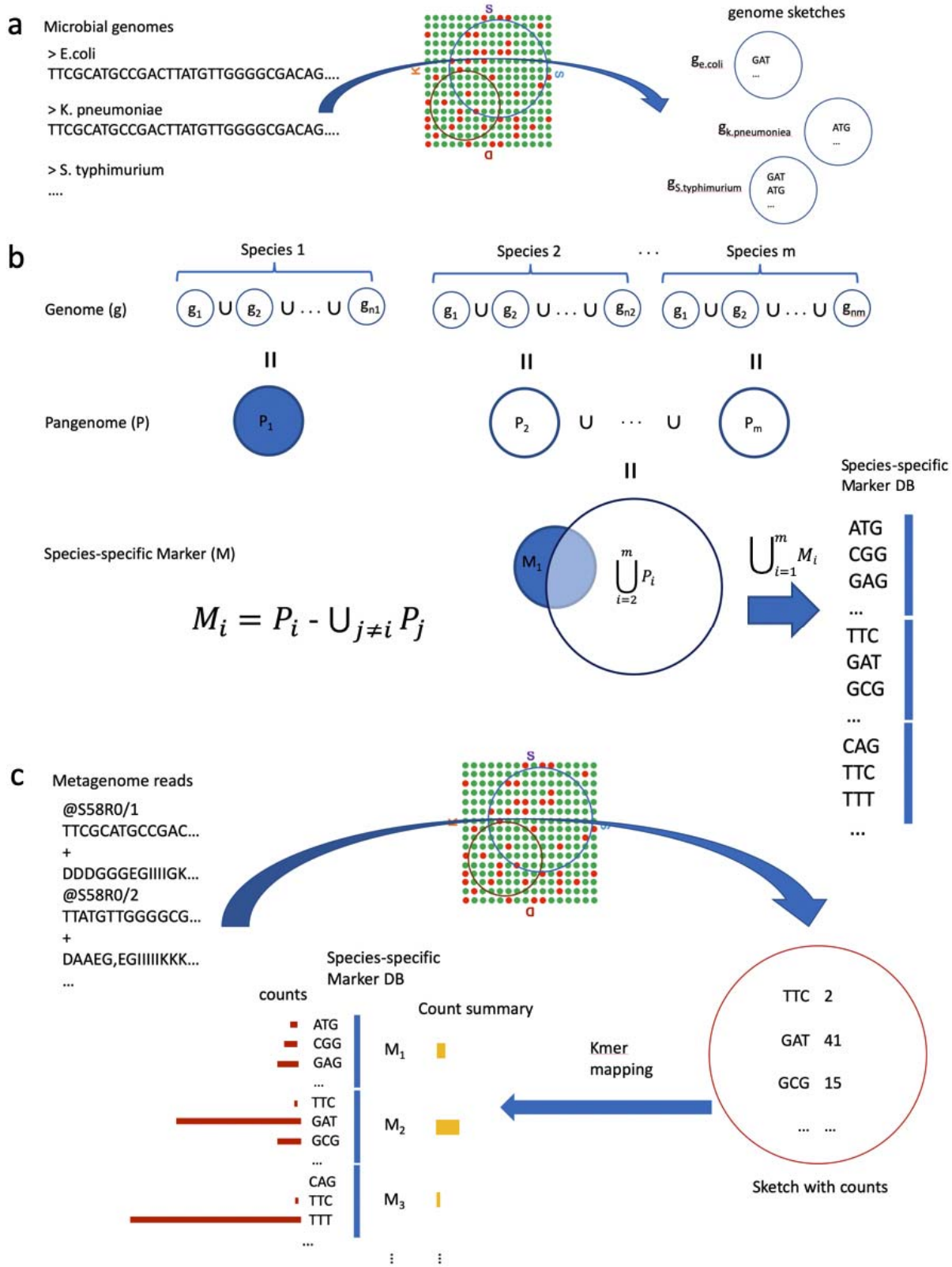

Fig. 1 MetaKSSD algorithm overview. (a) Reference genomes were sketched and organized by species. (b) MarkerDB construction: for each species *i* (*i=1..m*, where *m* is the total number of

species), the union sketch of all the sketches from species i was taken as its pangenome sketch P_i; Species-specific marker M_i was derived by subtracting from P_i the union of all other pangenome sketches; All species-specific markers were indexed to build the MarkerDB. (c) Metagenome taxonomic profiling: The sequencing data of a sample was first sketched with k-mer counts tracked. The sample sketch was then mapped to the MarkerDB, and the relative abundances for all species were estimated based on the statistical summaries of the overlapped k-mer counts.

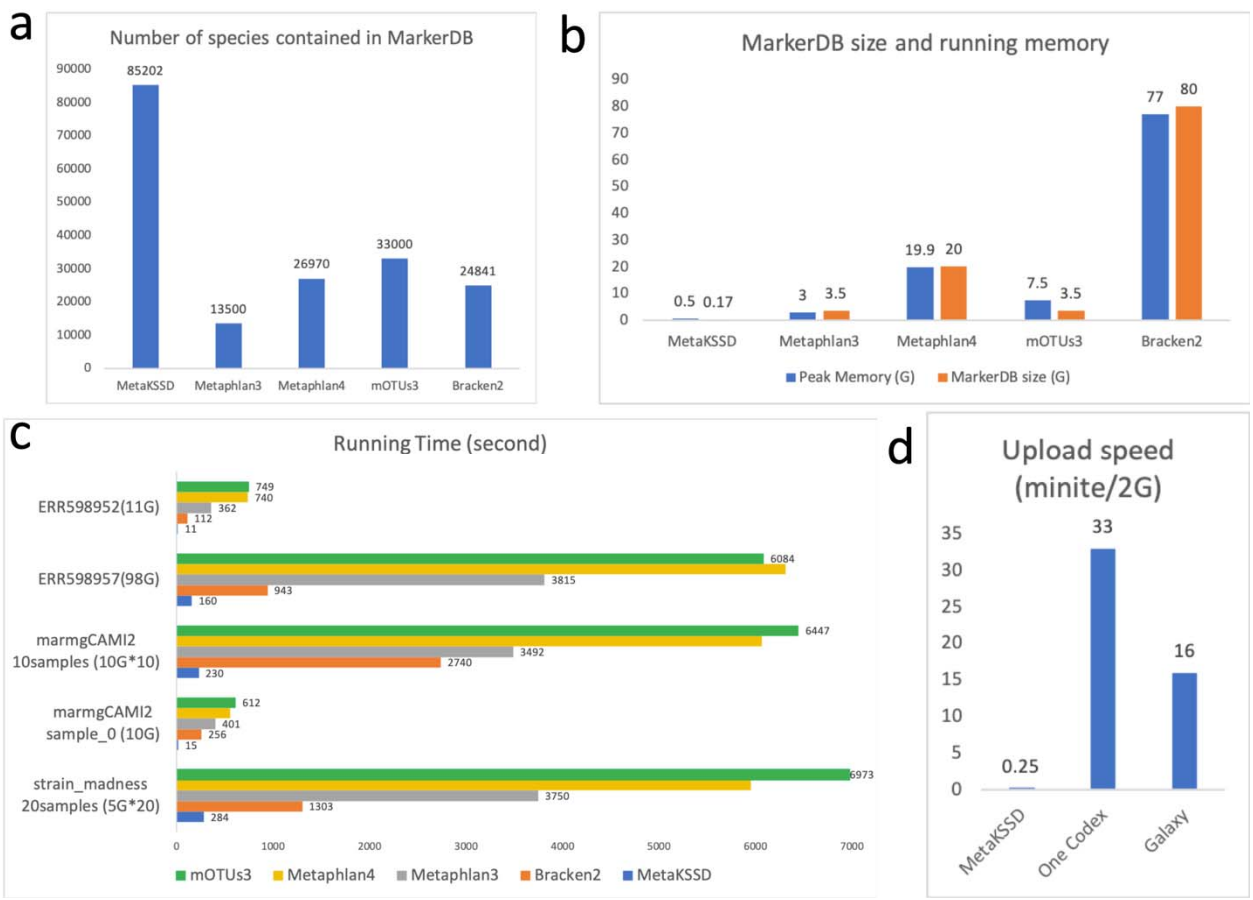

Fig. 2 MetaKSSD quantifies for more species while using less computational resources. Profiling speed testing (c) was conducted on an Amazon Web Services (AWS) c6a.24xlarge EC2 instance, featuring high I/O storage, utilizing 30 threads. Upload speed testing (d) was conducted on a MacBook Pro with 40Mbps bandwidth

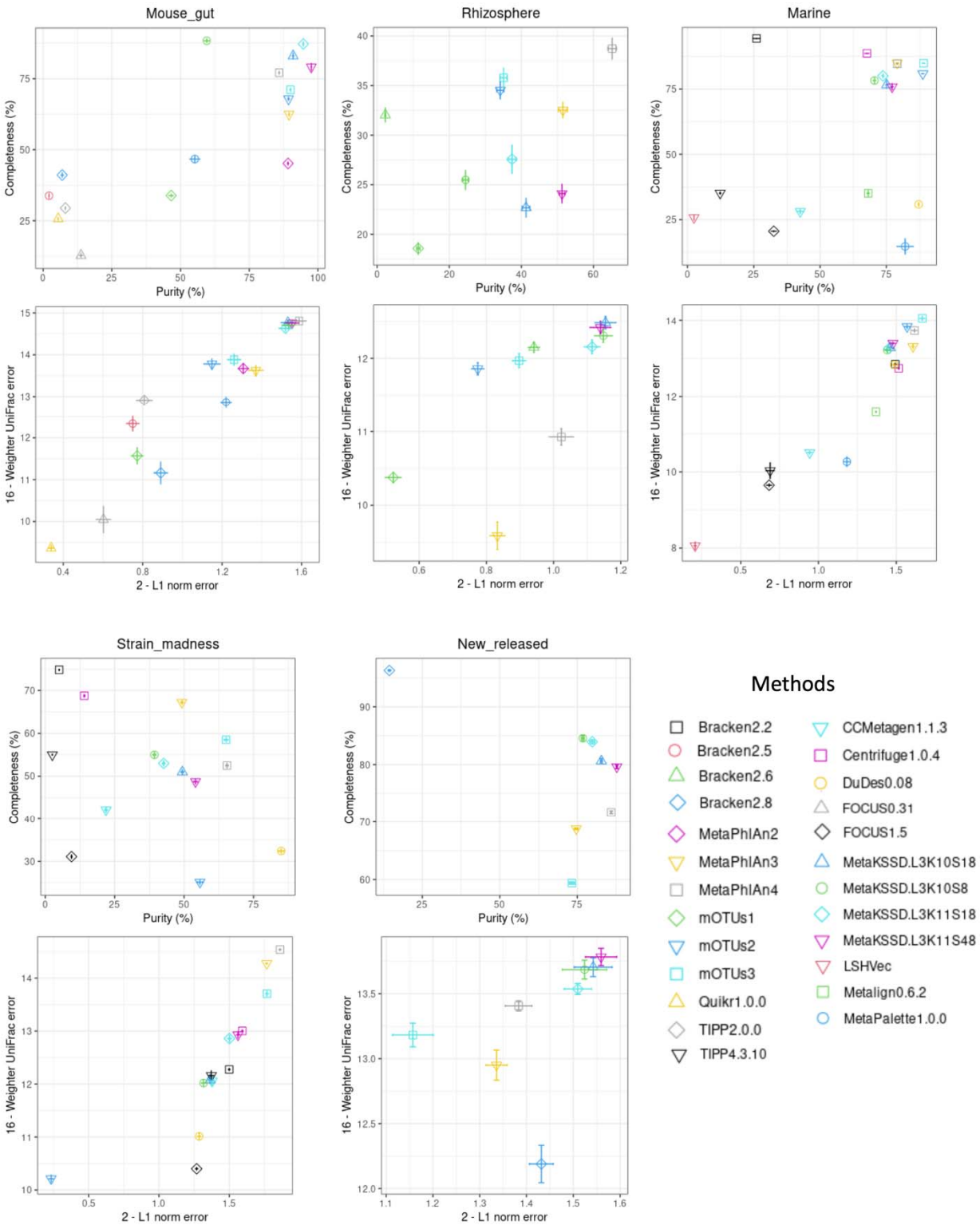

Figure 3. Performance Comparison of MetaKSSD and Other Methods.
MetaKSSD, along with different parameter settings and various versions of other metagenomic profiling methods, were benchmarked and compared across five benchmark datasets: mouse gut

(*n*=49 samples), Rhizosphere (*n*=21 samples), Marine (*n*=10 samples), Strain_madness (*n*=100 samples), and the 'New released' dataset (*n*=5 samples). Refer to the Data Availability and Methods sections for details on simulation. The performance metrics including Purity-Completeness (the top row of each dataset) and 2-L1 norm error-16- Weighted UniFrac error (the bottom row of each dataset) were evaluated at the species-level using the OPAL framework. Purity measures the ratio of correctly predicted taxa present to all predicted taxa at the selected taxonomic rank, ranging from 0 (worst) to 1 (best); Completeness measures the ratio of correctly predicted taxa present to all taxa present at the selected taxonomic rank, also ranging from 0 (worst) to 1 (best); L1 norm error calculates the sum of absolute differences between true and predicted abundances at the selected taxonomic rank, ranging from 0 (perfect reconstruction of relative abundances) to 2 (totally incorrect reconstruction); Weighted UniFrac error is a tree-based measure of similarity between true and predicted abundances at all taxonomic ranks, ranging from 0 (high similarity) to 16 (low similarity). The length of each cross represents the standard deviations of the corresponding metric for each data point.

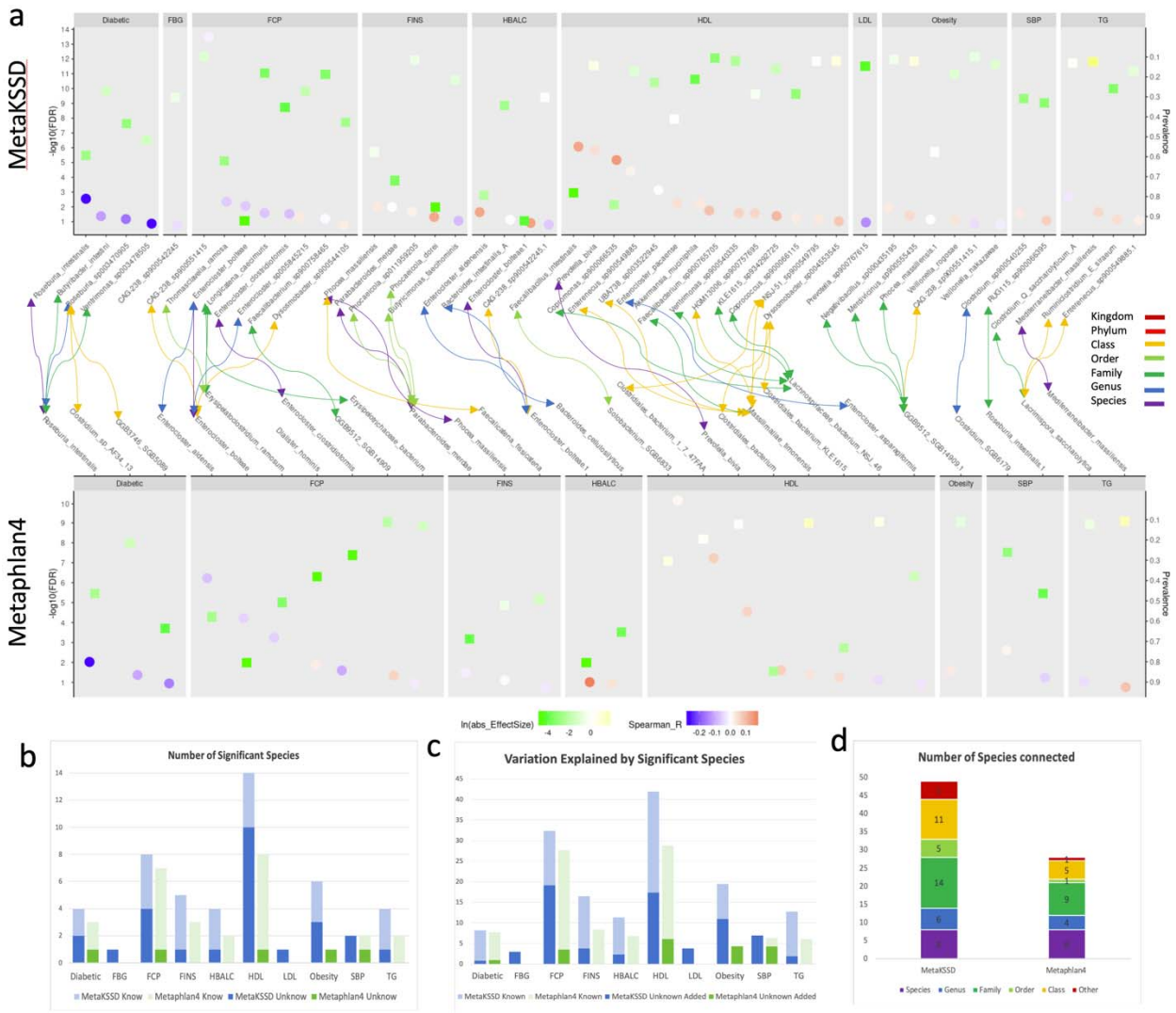

Figure 4. Comparison of MetaKSSD and Metaphlan4 for microbiome-phenotype association study.

All the 368 Chinese stool samples from the BGInature2012 cohort[1] were profiled utilizing both MetaKSSD and Metaphlan4 methods and bacteria species associated with 10 phenotypes were identified. These phenotypes include Diabetic status, fasting blood glucose (FBG), fasting serum C-peptide (FCP), fasting serum insulin (FINS), glycosylated hemoglobin HbAlc (HBALC), high-density lipoprotein (HDL), low-density lipoprotein (LDL), Obesity status, systolic blood pressure (SBP), and triglyceride (TG). (a) Negative log transformed False Discovery Rate (FDR, circle dots, left y-axis) and Prevalence (square dots, right y-axis) of the bacterial species (x-labels) identified by MetaKSSD and Metaphlan4 for each phenotype. A line connects a MetaKSSD species with its closest Metaphlan4 species if they belong to the same class, with the color of the line indicating their lowest common ancestry rank. Species name prefixed with pattern "sp[0-9]{9}" and “SGB” are unknown species for MetaKSSD and Metaphlan4, respectively. Otherwise is known species. Comparison of MetaKSSD and Metaphlan4 with regarding (b) the number of significant species identified, and (c) the percent of variations explained. (d) Summary of species connected in (a).

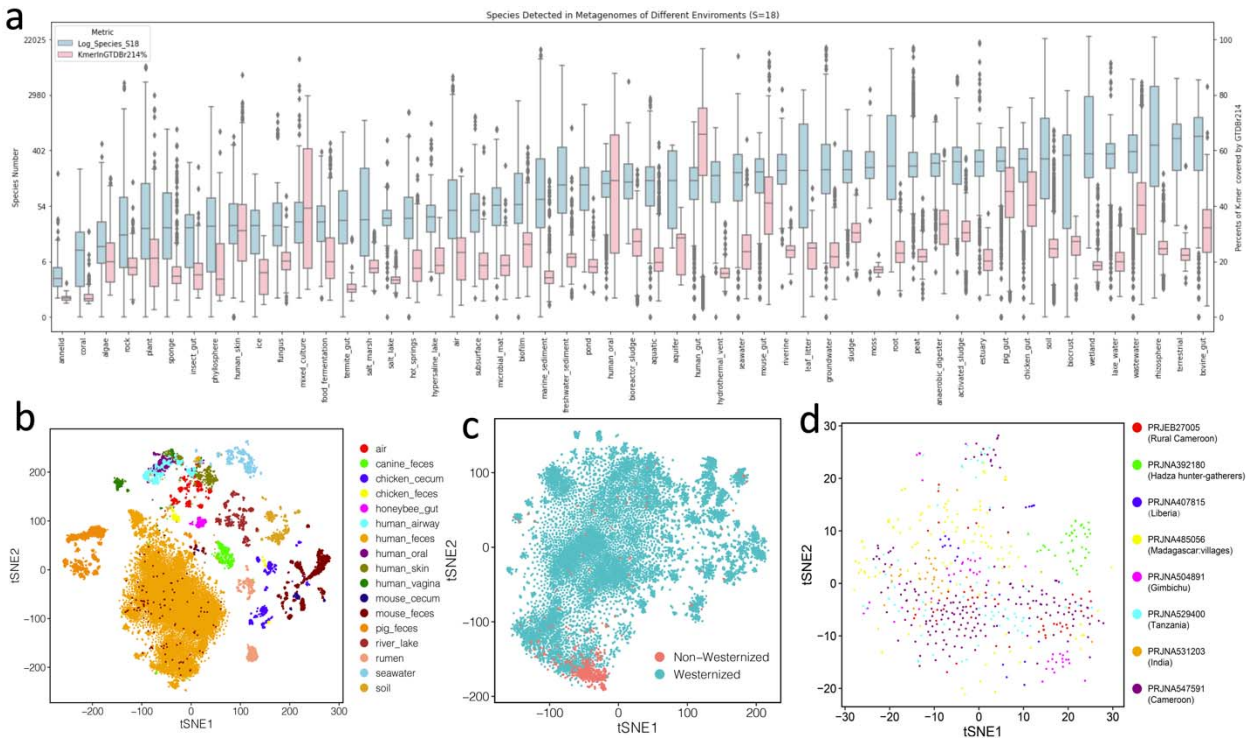

Figure 5. Large-scale metagenome profiling. (a) Species number (left y-axis) detected in different environments alongside *k*-mer percentages covered by GTDBr214 genomes (right y-axis). The MetaKSSD (L3K11, with MarkerDB constructed from GTDBr214) raw abundance profiles (Data availability) of runs from the 53 commonly studied metagenomic environments (Table S5) were selected. The species number for each run was determined from its raw profile by thresholding the MarkerDB overlapped *k*-mer number S = 18. The *k*-mer (*k*=22) percentage covered by GTDBr214 was estimated by the proportion of *k*-mers in a run sketch that overlapped with the union sketch of all GTDBr214 genomes. The species number and *k*-mer percentage of a sample is the average of the species numbers and *k*-mer percentages of all its runs. Environments were ranked by the sample median of species number. (b)-(d)T-SNE analysis of 151 well-characterized projects (Tab. S7), lifestyles related projects (Tab. S8) and non-Westernized samples, respectively.