Fast, lightweight, and accurate metagenomic functional profiling using FracMinHash sketches

Mahmudur Rahman Hera[1,‡], Shaopeng Liu[2,‡], Wei Wei[2], Judith S. Rodriguez[2], Chunyu Ma[2], and David Koslicki[1,2,3]

[1]School of Electrical Engineering and Computer Science, Pennsylvania State University, USA
[2]Huck Institutes of the Life Sciences, Pennsylvania State University, USA
[3]Department of Biology, Pennsylvania State University, USA
[‡]*The authors contributed equally to the work*

Abstract

Motivation Functional profiling of metagenomic samples is essential to decipher the functional capabilities of these microbial communities. Traditional and more widely used functional profilers in the context of metagenomics rely on aligning reads against a known reference database. However, aligning sequencing reads against a large and fast-growing database is computationally expensive. In general, k-mer-based sketching techniques have been successfully used in metagenomics to address this bottleneck, notably in taxonomic profiling. In this work, we describe leveraging FracMinHash (implemented in `sourmash`, a publicly available software), a k-mer-sketching algorithm, to obtain functional profiles of metagenome samples. We show how pieces of the `sourmash` software (and the resulting FracMinHash sketches) can be put together in a pipeline to functionally profile a metagenomic sample.

Results We report that the functional profiles obtained using this pipeline demonstrate superior completeness and purity compared to the profiles obtained using other alignment-based methods when applied to simulated metagenomic data. At the same time, we also report that our functional profiling pipeline is 42-51x faster in CPU time, 10-15.8x faster in running time, and consumes up to 20% less memory. Coupled with the KEGG database, this method not only replicates fundamental biological insights but also highlights novel signals from the Human Microbiome Project datasets.

Reproducibility This fast and lightweight metagenomic functional profiler is freely available and can be accessed here: https://github.com/KoslickiLab/funprofiler. All scripts of the analyses we present in this manuscript can be found on GitHub.

1 Introduction

Metagenomic profiling, the process of evaluating and identifying the genomic information present within a given environmental sample, offers valuable insights into the genetic diversity, functional capacity, and ecological contributions of microorganisms. Specifically, metagenomic functional profiling (a.k.a. functional annotation), the computational process that involves identifying and quantifying functional components within metagenomic data, is essential to understanding microbial functionality, phenotype-genotype association, host interaction, disease progress, etc. [13, 14, 41, 56, 44].

A related and very important concept often explored in biology, and in gene studies in particular, is orthologous genes. If two genes in two different species are evolutionary descendants of the same gene in the least common ancestor of these species, then these two genes are called orthologs [60]. The study of orthologous genes contributes crucially to comparative genomics and evolutionary studies, and often these studies make use of the well-known fact that orthologous genes often carry out equivalent or identical biological functions across species boundaries [18, 36, 61]. For example, researchers have successfully identified functions of genes in newly sequenced genomes using orthology in model organisms [60, 36, 10]. Previously, it was common practice to annotate the function of human genes using the orthologs in mice, as it is easier to experiment on the latter [23]. Largely speaking, orthology traces genes to the same ancestral gene, and therefore, orthologous relationships are considered to be the most accurate way to capture differences and similarities in the composition of genomes from different species [16]. Naturally, computational efforts have explored the area, and constructed databases of orthologous genes, such as AspGD [4], MBGD [67], OrthoDB [39], and KEGG [31, 30].

Coming back to the study of genes in metagenomes: it is important to note that numerous computational techniques have been developed to identify genes that are present in an environmental sample. In a broader sense, these techniques fall into two categories: (a) gene prediction tools – which attempt to predict genes *ab initio* using only the sequences [48, 59, 77], and (b) gene classification/annotation tools – which aim to annotate a metagenome sample against a database of known genes. The latter category is more relevant for functional profiling since we need to have prior knowledge of the functions of known genes. The majority of the tools in this category rely on alignment-based approaches. For example, `eggNOG-Mapper` performs alignments with orthologs utilizing profile HMM models for query assignment [25]. `MG-RAST` [35] uses `BLAST` [74] to search against the M5nr database [70]. `BlastKOALA` [32], `GhostKOALA` [32], and `KofamSCAN` [3], which utilize the BLASTP [74], GHOSTX [63], and hmmsearch [29] algorithms respectively, are developed to search against Kyoto Encyclopedia of Genes and Genomes (KEGG) database [31]. Perhaps the fastest (and most widely used) tool in the context is `DIAMOND` [8], which utilizes double indexing on both query and reference to enable fast alignment. `DIAMOND`'s novel algorithm can match the accuracy of `BLAST` while having much better efficiency [8].

Despite the popular use of these tools, the primary use of alignment-based algorithms makes these a poor practical choice in terms of scalability. With the ever-growing volume of sequencing data, alignment-based methods, despite their historical success, will eventually be overshadowed by more scalable techniques. Computational biologists, therefore, continue to turn to sketching-based methods, which are often faster and more lightweight; and theoretical guarantees of the sketching algorithms ensure their high accuracy. These more popular alignment-based tools also lack the use of orthology relationships of the genes. Using well-studied orthologous groups has the potential to make functional profiling pipelines more efficient, both in time and memory.

In this paper, we present a sketching-based pipeline to functionally profile a metagenome in terms of orthologous gene groups. We used FracMinHash as our sketching technique to develop the pipeline – recent investigations have shown the successful use of FracMinHash (used in the software package `sourmash` [7]) in metagenomics analysis [47]. We used KEGG as the database of orthologous genes. The orthologous groups in KEGG are called KOs. Our pipeline can take FracMinHash sketches of a given metagenome and the KOs, and progressively discovers what KOs are present in the metagenome using the algorithm `sourmash gather` [52]. The pipeline can also annotate the relative abundances of the KOs. It is fast and lightweight because of using FracMinHash sketches, and is accurate when the sequencing depth is moderately high. We investigated the performance and resource usage of the pipeline against popular gene annotation tools `DIAMOND` and KofamScan using simulations. We also used the pipeline to functionally profile Human Gut Microbiota. Furthermore, the KO groups are very well studied in the KEGG database in the

form of pathways – which allowed us to make insightful deductions about these microbiota. Finally, to demonstrate the versatile usefulness of the results, we showed how two KO profiles can be consolidated into a single metric using the pathways' hierarchy. The pipeline is freely available and can be accessed here: https://github.com/KoslickiLab/funprofiler.

2 Background

2.1 KEGG and gene orthology

Selecting an appropriate reference database is crucial in metagenomic functional profiling. Orthologous genes, which share a common evolutionary ancestry, are commonly employed for inferring functions [12]. Numerous efforts have been undertaken to construct databases for gene orthology, such as NCBI COG [17], AGNOSTOS-DB [69], OrthoDB [39], and Ensembl Compara [21]. The KEGG database is a comprehensive and integrated resource for biological interpretation by providing gene annotations and mapping genes to manually created pathway maps [31, 30]. These KO groups serve as a foundation for functional annotation in metagenomics and can be further connected into pathway maps to illustrate the hierarchical interconnections of biological functions within living organisms.

2.2 FracMinHash and sketching-based method

With the large number of reference genomes and exponential surge in metagenomic data production, there is an imperative need for the creation of computational models that are both scalable and robust, ensuring precision in analysis. K-mer-based algorithms, particularly sketching-based methods, are gaining increasing popularity for metagenomic profiling implementations in this regard. A k-mer is a sequence of k consecutive nucleotides, extracted from a longer sequence. Algorithms designed to work with k-mers split the entire sample into k-mers, and analyze the number of shared/dissimilar k-mers among multiple samples. The number of all distinct k-mers in a sequencing sample can often be huge, so sketching-based methods take a fingerprint of the k-mers (called a *sketch*) and work with these much smaller sets, ensuring less consumption of computational resources. The most popular sketching method for many years has been MinHash, introduced in the context of document comparisons [6]. Mash [50] was developed to apply MinHash to genomic data and has been very widely used. However, recent studies have shown that the error in comparing two MinHash sketches depends on the sketch size [53], and the size needs to grow quadratically to compensate for the error [50, Fig S1]. More recent works have shown that when sets of very dissimilar sizes are compared, MinHash sketches perform relatively poorly [42, 37]. Researchers have proposed many adjustments to MinHash to tackle this issue; using a variable sketch size (instead of MinHash's fixed-size scheme), the recently introduced FracMinHash sketch [27, 20] is one such example. Simply speaking, a FracMinHash sketch retains a fraction of the k-mers in the original sample. Formally, given a perfect hash function $h : \Omega \to [0, H]$ for some $H \in \mathbb{R}$ and a *scale factor* s where $0 \leq s \leq 1$, a FracMinHash sketch of a set A is defined as follows:

$$\mathbf{FRAC}_s(A) = \{\, h(a) \mid a \in A \text{ and } h(a) \leq Hs \}\,. \tag{1}$$

The scale factor s is a tunable parameter that can modify the size of the sketch. For a fixed s, if the set A grows larger, the sketch $\mathbf{FRAC}_s(A)$ grows proportionally. This sketching technique was first introduced in the software package `sourmash` [7, 52], and has recently been theoretically analyzed to justify its use [20]. `sourmash` has successfully used FracMinHash to obtain genome-wide comparisons of biological sequences. In the second round of Critical Assessment of Metagenomic Interpretation (CAMI) challenges [57, 47], the tool `sourmash gather` [26], exhibited the highest completeness and purity in taxonomic profiling from metagenomic data for multiple datasets [47] across the genus and the species levels.

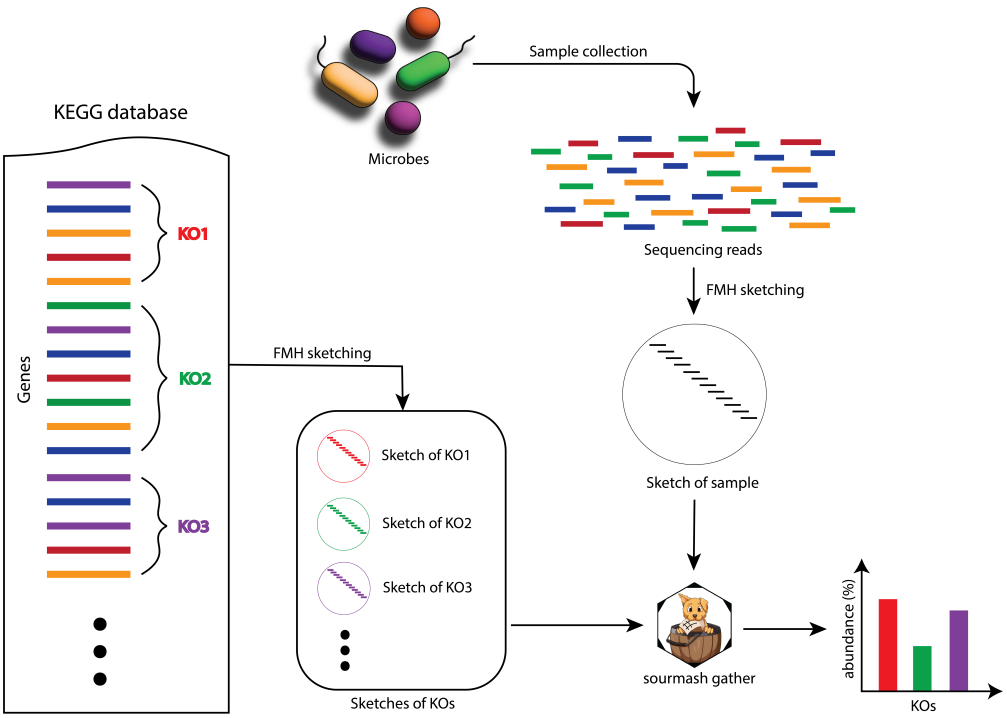


Figure 1: An overview of the pipeline used to identify genes/KOs present in a metagenome. The pipeline requires a list of KOs to calculate the relative abundance of these KOs in the sample. FracMinHash sketches of the KOs and the metagenome sample are computed using the command `sourmash sketch`, and then `sourmash gather` is invoked to identify the KOs using these sketches.

3 Results

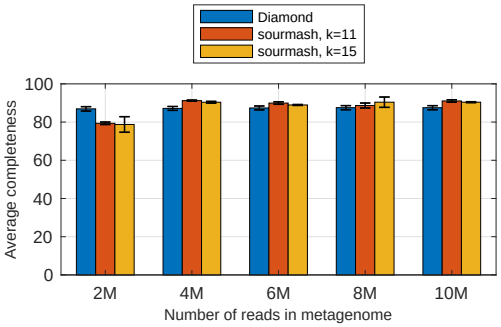
3.1 `sourmash` accurately identifies KOs using less computational resources

Our pipeline (shown in Figure 1) to identify KOs in a metagenome uses the KEGG database coupled with the `sourmash gather` command in the `sourmash` software package. The pipeline is detailed in the Methods section. We start presenting the results by showing the performance of this pipeline in identifying KOs in simulated metagenomes. Details of these simulations are elaborated in the Methods section.

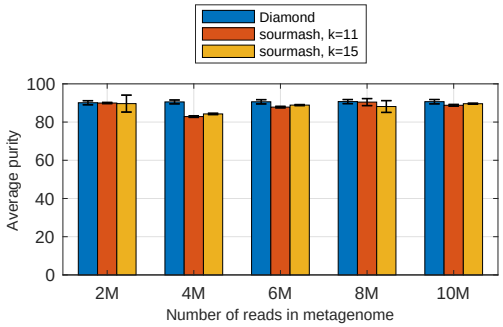
We made comprehensive comparisons with the tools `KofamScan` and `DIAMOND`. We compared with `DIAMOND` because this is the fastest and most widely used protein alignment tool, being more than 20,000 times faster than BLASTX [8]. We also compared our pipeline with `KofamScan`, primarily because this is a KO-identifying tool developed specifically for the KEGG database, and our use of the KEGG database naturally warrants this comparison. `KofamScan` uses `hmmsearch` to directly assign KO IDs to reads in a metagenome, whereas `DIAMOND` assigns gene id to a sequencing read. We summarized the outputs generated by `DIAMOND` using gene-to-KO mapping available in the KEGG database.

From our simulation experiments, we found that `KofamScan` fails to scale to metagenomes with millions of reads (taking more than seven days to complete on a simulated metagenome with 1M reads) – making it an impractical choice for this task. Nevertheless, because `KofamScan` was developed so closely with the KEGG database, we present the comparison in this manuscript.

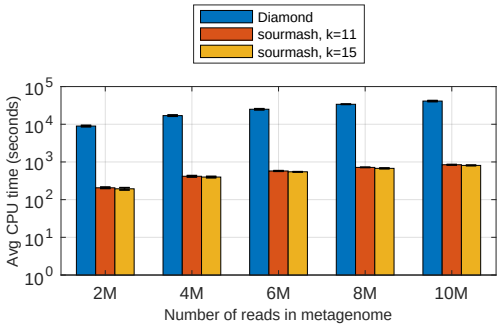
The results for metagenomes with millions of reads are shown in Figure 2 (a and b). We see that `sourmash gather` slightly outperforms `DIAMOND` for these settings in terms of both completeness and purity as the sample is sequenced in more depth. We used two different k-mer sizes when running `sourmash`. In these experiments, we used a single active thread to run the `sourmash gather` program, and 64 threads to run `DIAMOND` to generate these results. The computational resources (total CPU time and memory) to generate these results are shown in Figure 2 (c and d). Clearly, `sourmash gather` outputs the list of KOs faster than `DIAMOND`. This is expected, given the computation-intensive alignments performed by `DIAMOND`



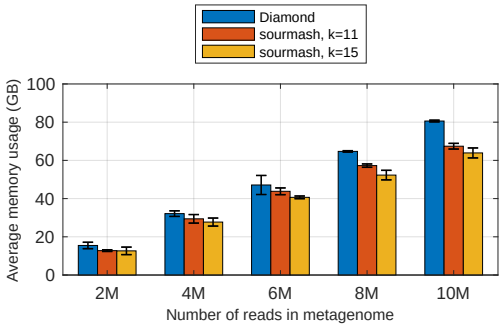
(a) Completeness scores in identifying KOs.



(b) Purity scores in identifying KOs.



(c) Total CPU time required to identify KOs.



(d) Memory usage to identify KOs.

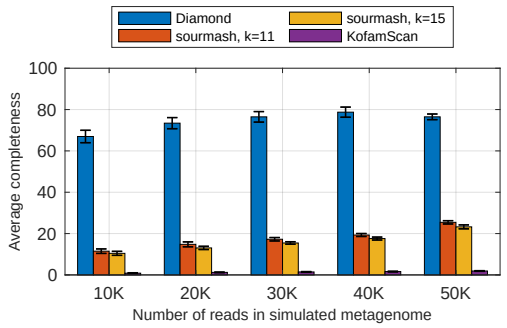
Figure 2: Average performances and computational resources of various tools in identifying KOs in simulated metagenomes. The simulated metagenomes were annotated to the truly present genes and KOs using brute force. Then, the outputs of the tools were matched against this ground truth. For a particular setting, ten different seed values were used to simulate a metagenome – the error bars indicate one standard deviation over these seeds. We found that `sourmash` shows high purity and completeness, just as `DIAMOND` – but `sourmash` consumes less computational resources(13.7-20.7% less memory, and 42-51x less CPU time).

to identify KOs. On the other hand, the use of lightweight sketches allows `sourmash` to avoid alignment altogether, and identify the list of all present KOs more accurately, using fewer computational resources.

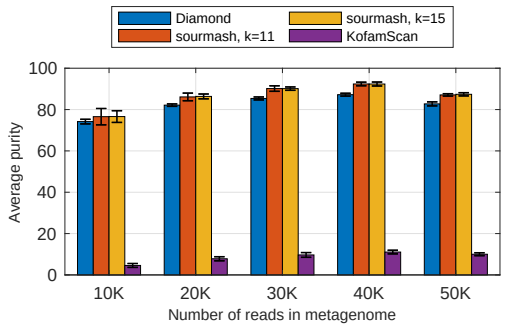
We also investigated the performance of `DIAMOND`, `KofamScan`, and our pipeline for much smaller-sized metagenomes (10-50K reads). These metagenomes were simulated from the same 1000 randomly selected bacterial genomes, and hence barely cover the diversified genes present in these genomes (with a low coverage). Therefore, these low-coverage samples would hardly be representative of real-world data. Nevertheless, we include these results to highlight the computationally intensive performance of `KofamScan`.

Our experiments revealed that when applied to low-coverage samples, `sourmash` has relatively high purity compared to the other two tools (shown in Figure 3b), i.e. almost all KOs that it identifies are truly in the sample. The completeness (shown in Figure 3b) is not so convincing compared to `DIAMOND`, i.e. `sourmash` cannot recover all KOs that are present when the sequencing is too shallow. Nonetheless, it does recover the most highly abundant KOs and is much faster and more lightweight compared to the other two tools (see running time and memory consumption in Figures 3c and 3d). We also found that `KofamScan` has exceptionally high resource requirements, and yet did not show promising performance.

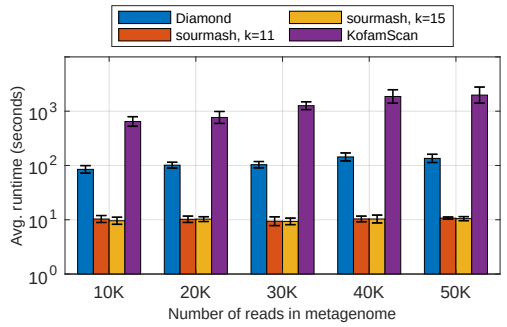
We conclude this section with the following remarks: `sourmash` clearly is the better choice when high-coverage samples are available. If the sequencing data does not have high coverage, then `sourmash` can recover the most abundant orthologous genes with high precision and may miss some gene groups whose presence is not so dominant. In all cases, `sourmash` runs faster, and consumes less memory.



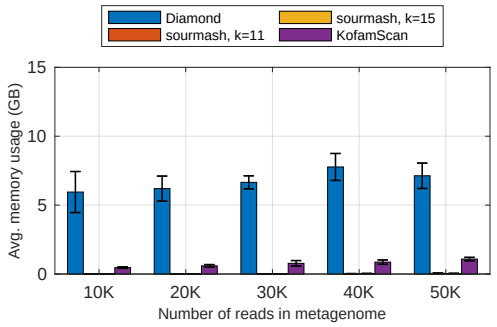
(a) Completeness scores in identifying KOs.



(b) Purity scores in identifying KOs.



(c) Running time required to identify KOs.



(d) Memory usage to identify KOs.

Figure 3: Average performances and computational resources of various tools in identifying KOs in simulated metagenomes. The simulated metagenomes were annotated to the truly present genes and KOs using brute force. Then, the outputs of the tools were matched against this ground truth. For a particular setting, ten different seed values were used to simulate a metagenome – the error bars indicate one standard deviation over these seeds.

3.2 `sourmash` reveals functional insights of human metagenomic data

After demonstrating the accuracy and improved efficiency of functional profile via FracMinHash sketches in simulated data, we subsequently employ this approach on actual human metagenomic data from the Human Microbiome Project (HMP)[66]. We use the KEGG (Kyoto Encyclopedia of Genes and Genomes) database[31], with KEGG Orthology (KOs) specifically, to profile metagenomic functions. The analysis encompasses a total of 1747 filtered samples, spanning 6 ethnic groups, 2 genders, and 3 study groups.

Functional profiles reveal the presence of housekeeping KOs and pathways consistently observed across the majority of samples, irrespective of conditions (Fig. 4). In addition to the core housekeeping KOs present in most samples, numerous other KOs (functional units) appear in just one or two conditions. The top hits of most observed KOs include K01872 for Aminoacyl-tRNA biosynthesis, K03701 for excinuclease ABC subunit A, and K02004 for (putative) ABC transport system permease protein. Most of the top hits are recognized as vital cellular functions, thus constituting a compilation of housekeeping functions. More universal functional insights of human microbiota can be obtained from the right tail of Figure 4b. We further checked the functional pathways at KEGG level 2 and level 3 and found that genes in Biosynthesis of amino acids (Ko01230), Carbon metabolism (Ko01200), and Amino sugar and nucleotide sugar metabolism (Ko00520) are abundant in most samples. These observations are consistent with the previous functional analysis in human gut microbiomes [79, 40, 65, 51], and suggest that these, along with Purine metabolism (Ko00230), Amino sugar and nucleotide sugar metabolism (Ko00520), and Glycolysis / Gluconeogenesis

(Ko00010) etc., are the core functions of the gut microbiota.

Next, we analyze the distinct functions among different conditions (Type 2 Diabetes, T2D; Healthy, HHS; and Inflammatory Bowel Disease, IBD). We conducted a LEfSe analyses [58] to unveil the key functional units/pathways that underlie the distinctions between the condition T2D vs. HHS and IBD vs. HHS.

Overall, our analysis identified a total of 17,225 KOs and 476 pathways across all samples. When applying the default parameters in LEfSe (log10 LDA > 2 and adjusted p-value < 0.05) to detect differentially abundant functional units, we identified 206 KOs and 71 pathways in the IBD group; 316 KOs and 83 pathways in the T2D group compared to the healthy control (HHS group). Figure 5 displays the top 10 results from the comparison between T2D and HSS. The most notable finding, map01100, indicates a strong link to Carbohydrate metabolism, a well-established association in the context of type 2 diabetes [64]. Likewise, we've detected functions related to glycan degradation (map00511) and sphingolipids (map00600), both of which have substantial supporting evidence for their involvement in disease status or development [11, 68, 76, 28, 73, 34, 1, 72, 24]. These observations substantiate the reliability of these functional profiles, thus signifying that the utilization of FracMinHash in conjunction with the KEGG database may serve as a fast and accurate hypothesis-generating methodology within the realm of functional analysis.

Digging into the KO-level analysis, the most significantly differentially abundant hit is NF-kappa-B-activating protein (K25931), which may play a key role in the pathogenesis of vascular complications of diabetes [62]. We also found starch-binding outer membrane protein (K21572), hydrophobic/amphiphilic exporter-1 (K03296), and carboxyl-terminal processing protease (K03797) etc. All of these discoveries may provide valuable insights for further investigating the functional roles of the gut microbiome.

3.3 `sourmash`-generated functional profiles can be used to compute Functional UniFrac distance

These FracMinHash-generated profiles can easily be coupled with other tools in further downstream functional analysis. In this section, we demonstrate this using two dimension reduction techniques: Principle Coordinate Analysis (PCoA) on pairwise Euclidean distances between profiles, and multidimensional scaling (MDS) using a recently developed method FunUniFrac [38].

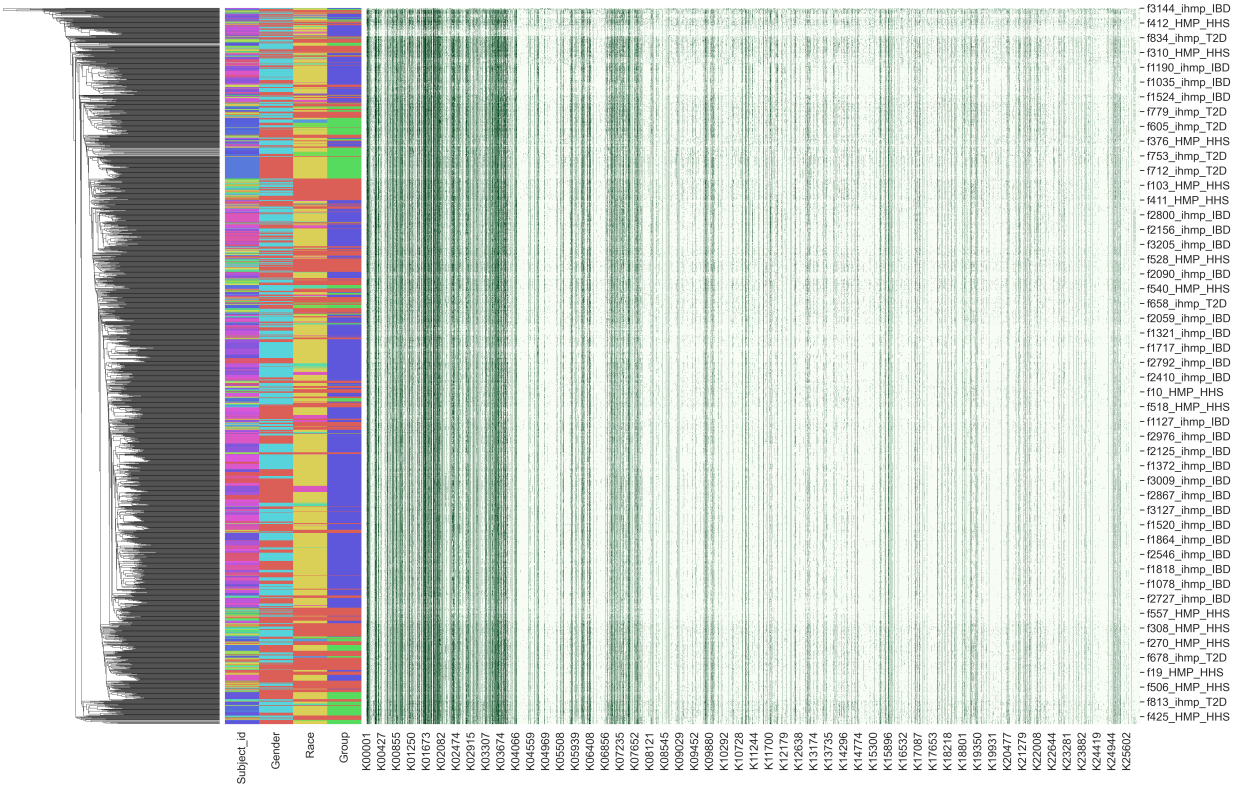
FunUniFrac stands for functional UniFrac. It is a dissimilarity metric for metagenomic functional profiles derived from the traditional UniFrac metric [45]. The computation of such UniFrac-like metrics requires an underlying tree structure that reflects some sort of intrinsic hierarchy or distance between the components (in our case, the KOs) in the samples. FunUniFrac uses the "KEGG KO tree" (see Subsection 4.2.2) from the KEGG database with branch lengths assigned based on pairwise distances between KOs obtained using `sourmash sketch`. This underlying KEGG tree allows the dissimilarity between two functional profiles to be quantified and potentially the differentially abundant protein functions to be identified.

For this demonstration, we used the same HMP dataset as used in Section 3.2 which consists of samples from healthy individuals (HHS) as well as individuals with type 2 diabetes (T2D) and inflammatory bowel disease (IBD). Using the functional profiles as input, we computed the pairwise FunUniFrac distances for T2D vs. HHS and performed MDS on the resulting pairwise distance matrices for visualization. We also clustered the functional profiles based on their pairwise Euclidean distances and performed a PCoA analysis. The comparisons of the results of these two clustering methods are presented in Figure 6. From this figure, we can observe their similarity in the segregation of the samples from these two environments, which agrees with prior discoveries of a connection between gut microbiome and T2D [2, 33, 54, 19, 5]. This demonstrates the versatility of these FracMinHash-generated functional profiles as inputs for downstream analyses in producing biologically meaningful results.

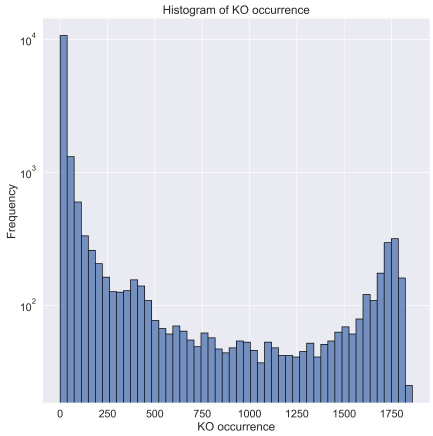
4 Methods

4.1 Overview of the pipeline

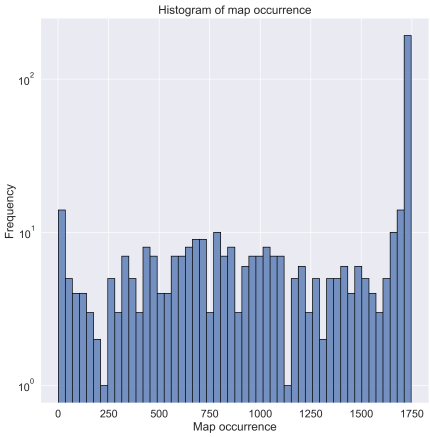
We begin with an overview of our KO identification pipeline, which differs from traditional gene identification tools (which mostly use alignment) by primarily utilizing `sourmash gather`, which breaks genes and reads into k-mers and approximately solves a set cover problem [26], producing a list of genes 'covering' observed k-mers in the metagenome. The problem with using `sourmash gather` with a list of genes is that genes are



(a) Functional landscape of the gut microbiota.

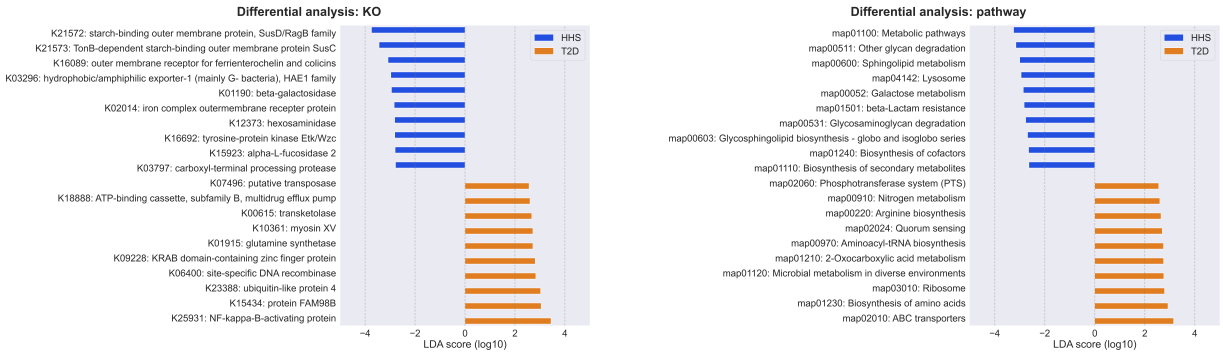


(b) Frequency of KOs detected in all samples



(c) Frequency of KEGG pathways detected in all samples

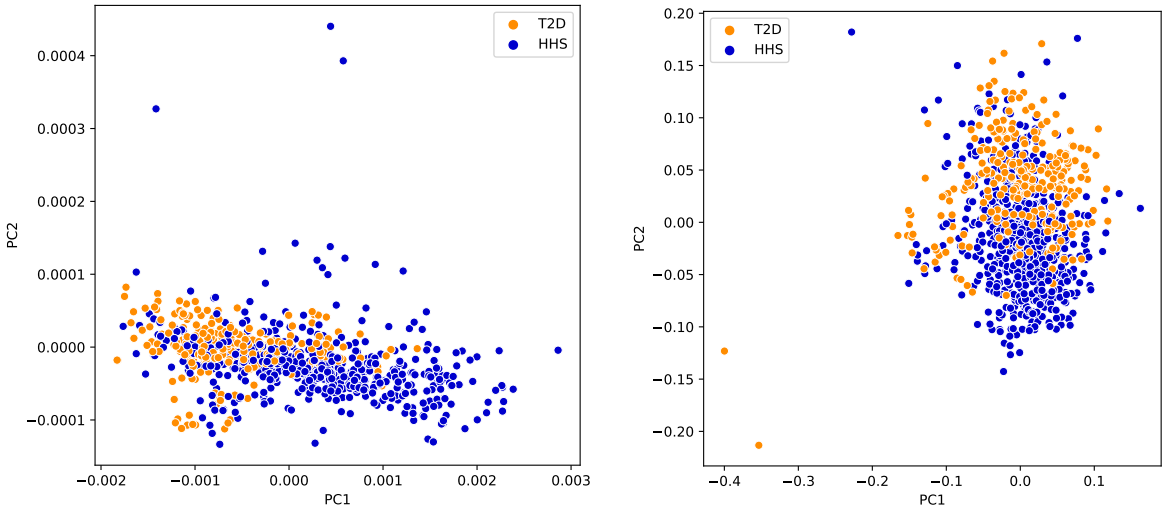
Figure 4: Functional landscape of the gut microbiota. (a) Heatmap of KO occurrence in all gut metagenomic samples. Samples are clustered by KO profile and the left color bar indicates the metadata of the sample: subject, gender, race, and study group. The green color indicates the presence of a given KO and a vertical green "line" suggests that a KO shows up in most of the samples. (b) Frequency distribution of KO occurrence: the right side contains KOs identified in most of the samples, i.e. "housekeeping". (c) Frequency distribution of KEGG pathway occurrence.



(a) T2D vs HHS: top differential KOs

(b) T2D vs HHS: top differential pathways

Figure 5: Differential analysis for T2D vs HSS. (a) Top 10 differential KOs in T2D samples compared to healthy samples. (b) Top 10 differential KEGG pathways in T2D samples compared to healthy samples.



(a) T2D vs HHS: PCoA of clustering based on pairwise Euclidean distances of functional profiles, with an outlier removed for better visualization.

(b) T2D vs HHS: MDS of clustering based on pairwise FunUniFrac distances.

Figure 6: Functional profiles recapitualate previously ovserved patterns: a) PCoA plot of functional profiles of gut microbiomes based on Euclidean distances for healthy and type 2 diabetes patients. b) MDS plot of samples clustered based on FunUniFrac distances computed using the same functional profiles; showing correspondence in observable patterns.

tinier compared to the sequences `sourmash gather` is designed to work on and that the number of genes in a known database can be very large. In the KEGG database, we found about 8M genes in the bacterial genomes. This large number can make the approximation algorithm in `sourmash gather` very slow and impractical to use. In initial investigations, we found that for even very small metagenomes (about 30 genomes, 150K reads in the metagenomes, and approximately 100K genes), it was computationally infeasible (15+ hours), and therefore, we abandoned this direction. Indeed, `sourmash` is intended to sketch many large sequence sets, not massive amounts of small sequences. Therefore, instead of finding genes, we focus on finding the orthologous gene *groups* in a metagenome. The KEGG database groups genes into orthologous groups called

'KEGG Orthologs' (referred to as KOs henceforth). The number of KOs (a total of only 25K) is much smaller than the number of genes, and the number of k-mers in a KO is much larger than that of a single gene. Considering these factors, we designed our pipeline to invoke `sourmash gather` with a list of all KOs in the KEGG database, and then to output a list of KOs that 'cover' all observed k-mers in a given metagenome.

4.2 Datasets

4.2.1 Simulated data

We used BBMap [9] to simulate a metagenome from 1000 randomly selected genomes from all 4498 bacterial genomes present in the KEGG database. The genomes were obtained from the KEGG database, as of 1 June, 2023. Genomes used in the simulations were selected uniformly, with each genome being equally likely to be included in the simulation. Tiny sequencing error was introduced in all simulated metagenomes, with insertion, deletion, and substitution rates of 0.001 (0.1%) each. BBMap was configured to store the chromosome coordinates of the sequencing reads when generating the metagenome file. We also collected the start and end positions of all 7956912 genes in these bacterial genomes. These start and end coordinates of the genes allowed us to construct a ground truth, i.e. which genes are actually present in the simulated metagenome. Using the gene-to-KO mapping available in the KEGG database, we then computed the ground truth at the KO level: which KOs are present in the simulated metagenome. This ground truth allowed us to measure the performance (completeness and purity) of the tools, presented in Figures 2 and 3.

We generated simulated metagenomes with the number of reads varying from 10K to 50K, and 2M to 10M. For each of these settings, we simulated 20 metagenomes with different seeds.

4.2.2 Reference data

The KEGG data, including KEGG Orthology (KO), gene sequences, and protein sequences), were downloaded from the KEGG FTP server, as of 1 June 2023. We used the KO identifiers that belong to the KEGG brite hierarchy **Ko00001** which contains 25,413 KOs. After filtering out the KOs without protein sequences, we were left with a total of 25,347 KOs used for the downstream analysis. The hierarchical structure of **Ko00001** was used as "KEGG KO tree".

We employed `sourmash` to establish FracMinHash sketches from amino acid sequences from those 25,347 KOs. These FracMinHash sketches were formulated using the command `sourmash sketch` with parameters `-p protein, k=7, k=11, k=15, abund, scaled=1000` (that is to say, we use k-mers for amino acid sequences and generate FracMinHash sketches using a scaling factor of 1000 for three different k values: 7, 11, and 15. Besides, k-mer abundances are tracked for quantification purposes). Subsequently, a sequence bloom tree (SBT) was constructed from the sketched KO database to accelerate the profiling process, though this is not required. These reference data are freely available on Zenodo.

4.2.3 Real metagenomes

All human metagenomic samples were sourced from the HMP data portal [66], using specific search terms: "feces" for body sites, "fastq" for format, and "WGS raw seq set" for type. On January 2023, a total of 3550 files were retrieved, tagged under the manifest id "160bdc491e". However, for quality assurance, additional filters were applied: (1) 185 files from the "MOMS-PI" project remained private; (2) 6 files were eliminated due to broken links; (3) 691 files exhibited unmatched MD5 values compared to metadata; (4) 4 files were corrupt; and (5) 148 files were identified as assembled scaffolds. Consequently, a final selection of 2516 files (a total of 4.6TB) was used. In subsequent analysis, we found numerous samples with very low sequencing depth. Therefore, we excluded files with sizes less than 200MB and files with less than 1000 KOs identified. After all these filtered steps, we have 1747 high-quality data remaining for the downstream analysis, including 547 healthy samples, 274 type 2 diabetes samples, and 926 samples related to inflammatory bowel disease.

4.3 Functional profiles of HMP data

We used command `sourmash sketch` with parameters `-p protein, k=11, abund, scaled=1000` for FracMinHash sketches of the HMP data (though it's suggested for the reference data to have multiple k values to

fit various purposes, one proper k size is sufficient for functional profiles). Next, functional profiles of them were generated using `sourmash gather` with parameters `-k 11 --protein --threshold-bp 500` ($k = 11$, which corresponds to 33 in DNA sequences, is a reasonable value for protein sequence comparison). It's worth mentioning that `sourmash gather` provides comprehensive information, among which, we utilized "f_unique_weighted" since this represents the relative abundance of each KO.

4.4 Differential analysis

LEfSe (Linear discriminant analysis Effect Size) was performed to determine enrichment in functional profiles of pairs of conditions. We adhered to the default settings for this analysis. Features (KO or pathway) with LDA score (log10) higher than 2 and adjusted p-value < 0.05 were recognized as significant. We used the Seaborn package in Python for plotting. For simplicity, we only showed partial results in the manuscript.

5 Discussion

5.1 Conclusions

In this manuscript, we combined the FracMinHash sketching technique along with the KEGG orthology to construct a robust pipeline for generating functional profiles from metagenomic data. Our findings demonstrate that the functional profiles derived through this pipeline exhibit advantages in terms of both completeness and purity when compared to `DIAMOND`, a widely used alignment-based tool for functional profiling. Moreover, our findings highlight the versatility of k-mer sketches, not just in taxonomic profiles but also in large-scale functional profiling, delivering enhanced scalability and effectiveness. It is worth noting that the FracMinHash may not be the ideal choice for the analysis of many small reference sequences in practice (e.g. millions of genes in the KEGG database) due to its top-down approximation algorithm. Opting for KEGG orthology, a family of genes sharing similar functions, is more appropriate in this situation.

By using this pipeline, we conducted a comprehensive functional annotation analysis on all available human gut metagenomic data from the Human Microbiome Project[66], comprising a total of 2.5K samples. We successfully generated sample-specific functional signatures and performed differential analyses across different conditions. The top signatures we identified align with existing literature. Moreover, the KO-based functional profiles expanded our insights by introducing additional disease-related functional units. To further illustrate the benefits of using the KEGG hierarchical structure on these KO profiles, we showcased their integration with the recently developed FunUniFrac method, highlighting how they can facilitate hierarchy-based sample comparisons. We believe this pipeline holds significant value in the realm of metagenomic functional profiles, especially when scalability and hypothesis-generating analyses are paramount.

5.2 Going beyond functional profiles

Gene Ontology (GO) enrichment analysis is widely used for understanding biological processes in experimental datasets. KEGG is such a comprehensive database that offers structured modules and pathways that delineate functional elements, associations with diseases, and interactions with drugs. It serves as an invaluable resource for exploration and hypothesis generation in diverse research endeavors. The potential of this direction can be fully exploited through the integration of functional profiles with meticulously curated biomedical knowledge graphs (e.g. Hetionet [22], RTX-KG2 [71]). This integration represents an exciting avenue for advancing our comprehension of intricate biological systems: metagenomic functional profiles provide insights into the functional potential of microbial communities; KEGG's rich resource of hierarchical modules and pathways offers a structured framework to interpret and contextualize these functions; and knowledge graphs, which link genes, proteins, diseases, drugs, and other biological entities, can facilitate the discovery of underlying relationships between different entities. By connecting functional profiles with biomedical knowledge graphs, researchers can explore the functional intersections of diseases [55, 49], identify potential therapeutic targets [75, 78, 46], and predict novel associations between microbial functions and human health [15, 43]. Pursuing this integrated approach can not only aid in hypothesis generation but also pave the way for innovative research.

References

[1] D. Ağagündüz, M. A. Icer, O. Yesildemir, T. Koçak, E. Kocyigit, and R. Capasso. The roles of dietary lipids and lipidomics in gut-brain axis in type 2 diabetes mellitus. *Journal of translational medicine*, 21(1):240, 2023.

[2] K. H. Allin, T. Nielsen, and O. Pedersen. Mechanisms in endocrinology: Gut microbiota in patients with type 2 diabetes mellitus. *European journal of endocrinology*, 172(4):R167–R177, 2015.

[3] T. Aramaki, R. Blanc-Mathieu, H. Endo, K. Ohkubo, M. Kanehisa, S. Goto, and H. Ogata. Kofamkoala: Kegg ortholog assignment based on profile hmm and adaptive score threshold. *Bioinformatics*, 36(7):2251–2252, 2020.

[4] M. B. Arnaud, G. C. Cerqueira, D. O. Inglis, M. S. Skrzypek, J. Binkley, M. C. Chibucos, J. Crabtree, C. Howarth, J. Orvis, P. Shah, et al. The aspergillus genome database (aspgd): recent developments in comprehensive multispecies curation, comparative genomics and community resources. *Nucleic acids research*, 40(D1):D653–D659, 2012.

[5] O. Aydin, M. Nieuwdorp, and V. Gerdes. The gut microbiome as a target for the treatment of type 2 diabetes. *Current Diabetes Reports*, 18(8):55, 2018.

[6] A. Z. Broder. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pages 21–29. IEEE, 1997.

[7] C. T. Brown and L. Irber. sourmash: a library for minhash sketching of dna. *Journal of open source software*, 1(5):27, 2016.

[8] B. Buchfink, K. Reuter, and H.-G. Drost. Sensitive protein alignments at tree-of-life scale using diamond. *Nature methods*, 18:366–368, 4 2021.

[9] B. Bushnell. Bbmap: a fast, accurate, splice-aware aligner. Technical report, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2014.

[10] K. Dolinski and D. Botstein. Orthology and functional conservation in eukaryotes. *Annu. Rev. Genet.*, 41:465–507, 2007.

[11] B. Ezzamouri, D. Rosario, G. Bidkhori, S. Lee, M. Uhlen, and S. Shoaie. Metabolic modelling of the human gut microbiome in type 2 diabetes patients in response to metformin treatment. *NPJ Systems Biology and Applications*, 9(1):2, 2023.

[12] G. Fang, N. Bhardwaj, R. Robilotto, and M. B. Gerstein. Getting started in gene orthology and functional analysis. *PLoS computational biology*, 6(3):e1000703, 2010.

[13] E. A. Franzosa, T. Hsu, A. Sirota-Madi, A. Shafquat, G. Abu-Ali, X. C. Morgan, and C. Huttenhower. Sequencing and beyond: integrating molecular'omics' for microbial community profiling. *Nature Reviews Microbiology*, 13(6):360–372, 2015.

[14] E. A. Franzosa, L. J. McIver, G. Rahnavard, L. R. Thompson, M. Schirmer, G. Weingart, K. S. Lipson, R. Knight, J. G. Caporaso, N. Segata, et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nature methods*, 15(11):962–968, 2018.

[15] C. Fu, X. Jiang, T. He, and X. Jiang. Mdepressionkg: a knowledge graph for metabolism-depression associations. In *Proceedings of the 2nd International Symposium on Artificial Intelligence for Medicine Sciences*, pages 63–68, 2021.

[16] T. Gabaldón and E. V. Koonin. Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics*, 14(5):360–366, 2013.

[17] M. Y. Galperin, K. S. Makarova, Y. I. Wolf, and E. V. Koonin. Expanded microbial genome coverage and improved protein family annotation in the cog database. *Nucleic acids research*, 43(D1):D261–D269, 2015.

[18] D. Graur and W.-H. Li. *Molecular evolution*. Sinauer Associates, Sunderland, MA, 1997.

[19] A. V. Hartstra, K. E. C. Bouter, F. Bäckhed, and M. Nieuwdorp. Insights into the role of the microbiome in obesity and type 2 diabetes. *Diabetes Care*, 38(1):159–165, 2014.

[20] M. R. Hera, N. T. Pierce-Ward, and D. Koslicki. Deriving confidence intervals for mutation rates across a wide range of evolutionary distances using fracminhash. *Genome Research*, 6 2023.

[21] J. Herrero, M. Muffato, K. Beal, S. Fitzgerald, L. Gordon, M. Pignatelli, A. J. Vilella, S. M. Searle, R. Amode, S. Brent, et al. Ensembl comparative genomics resources. *Database*, 2016:bav096, 2016.

[22] D. S. Himmelstein, A. Lizee, C. Hessler, L. Brueggeman, S. L. Chen, D. Hadley, A. Green, P. Khankhanian, and S. E. Baranzini. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, 6:e26726, 2017.

[23] H. L. Hinds, C. T. Ashley, J. S. Sutcliffe, D. L. Nelson, S. T. Warren, D. E. Housman, and M. Schalling. Tissue specific expression of fmr–1 provides evidence for a functional role in fragile x syndrome. *Nature genetics*, 3(1):36–43, 1993.

[24] T. Hla and A. J. Dannenberg. Sphingolipid signaling in metabolic disorders. *Cell metabolism*, 16(4):420–434, 2012.

[25] J. Huerta-Cepas, K. Forslund, L. P. Coelho, D. Szklarczyk, L. J. Jensen, C. von Mering, and P. Bork. Fast genome-wide functional annotation through orthology assignment by eggnog-mapper. *Molecular biology and evolution*, 34:2115–2122, 8 2017.

[26] L. Irber, P. T. Brooks, T. Reiter, N. T. Pierce-Ward, M. R. Hera, D. Koslicki, and C. T. Brown. Lightweight compositional analysis of metagenomes with fracminhash and minimum metagenome covers. *bioRxiv*, pages 2022–01, 2022.

[27] L. C. Irber Jr. *Decentralizing indices for genomic data*. University of California, Davis, 2020.

[28] R. Jia, M. Huang, L. Qian, X. Yan, Q. Lv, H. Ye, L. Ye, X. Wu, W. Chen, Y. Chen, et al. The depletion of carbohydrate metabolic genes in the gut microbiome contributes to the transition from central obesity to type 2 diabetes. *Frontiers in Endocrinology*, 12:747646, 2021.

[29] L. S. Johnson, S. R. Eddy, and E. Portugaly. Hidden markov model speed heuristic and iterative hmm search procedure. *BMC bioinformatics*, 11:1–8, 2010.

[30] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, 45(D1):D353–D361, 2017.

[31] M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[32] M. Kanehisa, Y. Sato, and K. Morishima. Blastkoala and ghostkoala: Kegg tools for functional characterization of genome and metagenome sequences. *Journal of molecular biology*, 428(4):726–731, 2016.

[33] F. H. Karlsson, V. Tremaroli, I. Nookaew, G. Bergström, C. J. Behre, B. Fagerberg, J. Nielsen, and F. Bäckhed. Gut metagenome in european women with normal, impaired and diabetic glucose control. *Nature*, 498(7452):99–103, 2013.

[34] B. D. Kayser, E. Prifti, M. Lhomme, E. Belda, M.-C. Dao, J. Aron-Wisnewsky, A. Kontush, J.-D. Zucker, S. W. Rizkalla, et al. Elevated serum ceramides are linked with obesity-associated gut dysbiosis and impaired glucose metabolism. *Metabolomics*, 15:1–13, 2019.

[35] K. P. Keegan, E. M. Glass, and F. Meyer. Mg-rast, a metagenomics service for analysis of microbial community structure and function. *Microbial environmental genomics (MEG)*, pages 207–233, 2016.

[36] E. V. Koonin. Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, 39:309–338, 2005.

[37] D. Koslicki and H. Zabeti. Improving minhash via the containment index with applications to metagenomic analysis. *Applied Mathematics and Computation*, 354:206–215, 2019.

[38] KoslickiLab. FunUniFrac: A repository for fununifrac. `https://github.com/KoslickiLab/FunUniFrac`.

[39] E. V. Kriventseva, D. Kuznetsov, F. Tegenfeldt, M. Manni, R. Dias, F. A. Simão, and E. M. Zdobnov. Orthodb v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic acids research*, 47(D1):D807–D811, 2019.

[40] R. Lin, W. Liu, M. Piao, and H. Zhu. A review of the relationship between the gut microbiota and amino acid metabolism. *Amino acids*, 49:2083–2090, 2017.

[41] N. Liu, H. Li, M. G. Chevrette, L. Zhang, L. Cao, H. Zhou, X. Zhou, Z. Zhou, P. B. Pope, C. R. Currie, et al. Functional metagenomics reveals abundant polysaccharide-degrading gene clusters and cellobiose utilization pathways within gut microbiota of a wood-feeding higher termite. *The ISME Journal*, 13(1):104–117, 2019.

[42] S. Liu and D. Koslicki. Cmash: fast, multi-resolution estimation of k-mer-based jaccard and containment indices. *Bioinformatics*, 38(Supplement_1):i28–i35, 2022.

[43] T. Liu, G. Lan, K. A. Feenstra, Z. Huang, and J. Heringa. Towards a knowledge graph for pre-/probiotics and microbiota–gut–brain axis diseases. *Scientific Reports*, 12(1):18977, 2022.

[44] Y. Liu, Z. Li, Y. Qi, X. Wen, and L. Zhang. Metagenomic analysis reveals a changing microbiome associated with the depth of invasion of oral squamous cell carcinoma. *Frontiers in Microbiology*, 13:795777, 2022.

[45] C. Lozupone and R. Knight. Unifrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, 71(12):8228–8235, 2005.

[46] C. Ma, Z. Zhou, H. Liu, and D. Koslicki. Kgml-xdtd: a knowledge graph–based machine learning framework for drug treatment prediction and mechanism description. *GigaScience*, 12:giad057, 2023.

[47] F. Meyer, A. Fritz, Z.-L. Deng, D. Koslicki, T. R. Lesker, A. Gurevich, G. Robertson, M. Alser, D. Antipov, F. Beghini, et al. Critical assessment of metagenome interpretation: the second round of challenges. *Nature methods*, 19(4):429–440, 2022.

[48] H. Noguchi, J. Park, and T. Takagi. Metagene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic acids research*, 34(19):5623–5630, 2006.

[49] S. C. P. Nunes. *Predicting gene-disease associations with knowledge graph embeddings over multiple ontologies*. PhD thesis, Universidade de Lisboa (Portugal), 2021.

[50] B. D. Ondov, T. J. Treangen, P. Melsted, A. B. Mallonee, N. H. Bergman, S. Koren, and A. M. Phillippy. Mash: fast genome and metagenome distance estimation using minhash. *Genome biology*, 17(1):1–14, 2016.

[51] A. Pant, T. K. Maiti, D. Mahajan, and B. Das. Human gut microbiota and drug metabolism. *Microbial Ecology*, 86(1):97–111, 2023.

[52] N. T. Pierce, L. Irber, T. Reiter, P. Brooks, and C. T. Brown. Large-scale sequence comparisons with sourmash. *F1000Research*, 8, 2019.

[53] M. Pătraşcu and M. Thorup. The power of simple tabulation hashing. *Journal of the ACM (JACM)*, 59(3):1–50, 2012.

[54] J. Qin, Y. Li, Z. Cai, S. Li, J. Zhu, F. Zhang, S. Liang, W. Zhang, Y. Guan, D. Shen, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60, 2012.

[55] A. Renaux, C. Terwagne, M. Cochez, I. Tiddi, A. Nowé, and T. Lenaerts. A knowledge graph approach to predict and interpret disease-causing gene interactions. *BMC bioinformatics*, 24(1):324, 2023.

[56] M. Schirmer, A. Garner, H. Vlamakis, and R. J. Xavier. Microbial genes and pathways in inflammatory bowel disease. *Nature Reviews Microbiology*, 17(8):497–511, 2019.

[57] A. Sczyrba, P. Hofmann, P. Belmann, D. Koslicki, S. Janssen, J. Dröge, I. Gregor, S. Majda, J. Fiedler, E. Dahms, et al. Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nature methods*, 14(11):1063–1071, 2017.

[58] N. Segata, J. Izard, L. Waldron, D. Gevers, L. Miropolsky, W. S. Garrett, and C. Huttenhower. Metagenomic biomarker discovery and explanation. *Genome biology*, 12:1–18, 2011.

[59] R. Silva, K. Padovani, F. Góes, and R. Alves. generfinder: gene finding in distinct metagenomic data complexities. *BMC bioinformatics*, 22(1):1–17, 2021.

[60] E. L. Sonnhammer and E. V. Koonin. Orthology, paralogy and proposed classification for paralog subtypes. *TRENDS in Genetics*, 18(12):619–620, 2002.

[61] R. A. Studer and M. Robinson-Rechavi. How confident can we be that orthologs are similar, but paralogs differ? *Trends in Genetics*, 25(5):210–216, 2009.

[62] S. V. Suryavanshi and Y. A. Kulkarni. Nf-$\kappa\beta$: a potential target in the management of vascular complications of diabetes. *Frontiers in pharmacology*, 8:798, 2017.

[63] S. Suzuki, M. Kakuta, T. Ishida, and Y. Akiyama. Ghostx: an improved sequence homology search algorithm using a query suffix array and a database suffix array. *PloS one*, 9(8):e103833, 2014.

[64] T. Takeuchi, T. Kubota, Y. Nakanishi, H. Tsugawa, W. Suda, A. T.-J. Kwon, J. Yazaki, K. Ikeda, S. Nemoto, Y. Mochizuki, et al. Gut microbial carbohydrate metabolism contributes to insulin resistance. *Nature*, pages 1–7, 2023.

[65] A. Tanca, M. Abbondio, A. Palomba, C. Fraumene, V. Manghina, F. Cucca, E. Fiorillo, and S. Uzzau. Potential and active functions in the gut microbiota of a healthy human cohort. *Microbiome*, 5:1–15, 2017.

[66] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon. The human microbiome project. *Nature*, 449(7164):804–810, 2007.

[67] I. Uchiyama. Mbgd: microbial genome database for comparative analysis. *Nucleic acids research*, 31(1):58–62, 2003.

[68] S. Upadhyaya and G. Banerjee. Type 2 diabetes and gut microbiome: at the intersection of known and unknown. *Gut microbes*, 6(2):85–92, 2015.

[69] C. Vanni, M. S. Schechter, T. O. Delmont, A. M. Eren, M. Steinegger, F. O. Glöckner, and A. Fernandez-Guerra. Agnostos-db: a resource to unlock the uncharted regions of the coding sequence space. *bioRxiv)*, 2021.

[70] A. Wilke, T. Harrison, J. Wilkening, D. Field, E. M. Glass, N. Kyrpides, K. Mavrommatis, and F. Meyer. The m5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. *BMC bioinformatics*, 13:1–5, 2012.

[71] E. Wood, A. K. Glen, L. G. Kvarfordt, F. Womack, L. Acevedo, T. S. Yoon, C. Ma, V. Flores, M. Sinha, Y. Chodpathumwan, et al. Rtx-kg2: a system for building a semantically standardized knowledge graph for translational biomedicine. *BMC bioinformatics*, 23(1):400, 2022.

[72] L. Wu, Z.-H. Lin, X.-J. Lu, X. Hu, H.-J. Zhong, D.-J. Lin, T. Liu, J.-T. Xu, W.-Y. Lin, Q.-P. Wu, et al. Washed microbiota transplantation improves patients with overweight by the gut microbiota and sphingolipid metabolism. *Biomedicines*, 11(9):2415, 2023.

[73] M. Yamaguchi and K. Yamamoto. Mucin glycans and their degradation by gut microbiota. *Glycoconjugate Journal*, pages 1–20, 2023.

[74] J. Ye, S. McGinnis, and T. L. Madden. Blast: improvements for better sequence analysis. *Nucleic acids research*, 34(suppl_2):W6–W9, 2006.

[75] J. Youn, N. Rai, and I. Tagkopoulos. Knowledge integration and decision support for accelerated discovery of antibiotic resistance genes. *Nature Communications*, 13(1):2360, 2022.

[76] F. Zhang, M. Wang, J. Yang, Q. Xu, C. Liang, B. Chen, J. Zhang, Y. Yang, H. Wang, Y. Shang, et al. Response of gut microbiota in type 2 diabetes to hypoglycemic agents. *Endocrine*, 66:485–493, 2019.

[77] W. Zhu, A. Lomsadze, and M. Borodovsky. Ab initio gene identification in metagenomic sequences. *Nucleic acids research*, 38(12):e132–e132, 2010.

[78] Y. Zhu, C. Che, B. Jin, N. Zhang, C. Su, and F. Wang. Knowledge-driven drug repurposing using a comprehensive drug knowledge graph. *Health Informatics Journal*, 26(4):2737–2750, 2020.

[79] Y. Zou, W. Xue, G. Luo, Z. Deng, P. Qin, R. Guo, H. Sun, Y. Xia, S. Liang, Y. Dai, et al. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nature biotechnology*, 37(2):179–185, 2019.