microorganisms

MDPI

Article

A Pipeline for Constructing Reference Genomes for Large Cohort-Specific Metagenome Compression

Linqi Wang [1], Renpeng Ding [2], Shixu He [2], Qinyu Wang [1] and Yan Zhou [1,2,*]

[1] State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, Shanghai 200438, China; linqiwang21@m.fudan.edu.cn (L.W.); qinyuwang22@m.fudan.edu.cn (Q.W.)
[2] MGI Tech, Shenzhen 518083, China; dingrenpeng@genomics.cn (R.D.); heshixu@genomics.cn (S.H.)
* Correspondence: zhouy@fudan.edu.cn

Abstract: Metagenomic data compression is very important as metagenomic projects are facing the challenges of larger data volumes per sample and more samples nowadays. Reference-based compression is a promising method to obtain a high compression ratio. However, existing microbial reference genome databases are not suitable to be directly used as references for compression due to their large size and redundancy, and different metagenomic cohorts often have various microbial compositions. We present a novel pipeline that generated simplified and tailored reference genomes for large metagenomic cohorts, enabling the reference-based compression of metagenomic data. We constructed customized reference genomes, ranging from 2.4 to 3.9 GB, for 29 real metagenomic datasets and evaluated their compression performance. Reference-based compression achieved an impressive compression ratio of over 20 for human whole-genome data and up to 33.8 for all samples, demonstrating a remarkable 4.5 times improvement than the standard Gzip compression. Our method provides new insights into reference-based metagenomic data compression and has a broad application potential for faster and cheaper data transfer, storage, and analysis.

Keywords: metagenomics; sequence data; reference-based compression

Citation: Wang, L.; Ding, R.; He, S.; Wang, Q.; Zhou, Y. A Pipeline for Constructing Reference Genomes for Large Cohort-Specific Metagenome Compression. *Microorganisms* **2023**, *11*, 2560. https://doi.org/10.3390/microorganisms11102560

Academic Editor: Juan M. Gonzalez

Received: 5 August 2023
Revised: 16 September 2023
Accepted: 18 September 2023
Published: 14 October 2023

1. Introduction

Metagenomics is one of the most important methods of microbiome research, which uses genomic strategies to investigate the genetic composition and functional patterns of all microorganisms present in specific environmental samples [1]. Over the years, several large-scale collaborative microbiome projects have been initiated, including the well-known Human Microbiome Project (HMP) and the Earth Microbiome Project (EMP) [2,3], the Metagenomics of the Human Intestinal Tract (MetaHIT) for gut microbiota [4], and the TARA Oceans Project for marine microorganisms [5]. These projects usually use non-targeted shotgun sequencing strategies to obtain higher taxonomic resolution and more detailed functional information about the genome, resulting in the generation of massive sequencing data. As of October 2022, the Human Microbiome Project Data Portal (https://portal.hmpdacc.org/) (accessed on 1 October 2022) contains 31,596 samples from 18 studies, with a 48.54 TB file volume of data. The expansion of project size and the growth of data volume poses a challenge in terms of computation, as there may not be enough resources to store and process data. This underscores the importance and urgency of developing methods to compress metagenomic data efficiently.

Several data compression strategies have been developed and can be divided into reference-free and reference-based methods. Reference-free methods compress raw sequencing data based on their natural statistics. Redundant DNA sequences are identified as far as possible and then processed using general compression methods such as Gzip and Bzip2 [6–9]. Reference-based methods exploit similarities between reads and a reference genome, which usually maps the target sequence to the reference genome and stores the

information needed to reconstruct the sequence: position in the reference genome and differences [10–14]. Due to the high similarity among genomes of homologous species, reference-based compression usually achieves high compression ratios, but this approach strongly relies on high-quality reference genomes.

There are many challenges in the reference-based compression of metagenomic sequencing data. Unlike classical genomic samples, metagenomic samples consist of many different organisms, resulting in a lack of universal reference genomes among samples. The microbial composition of cohorts may vary considerably as different studies have various technical approaches to sample collection and sequencing. In addition, microbial reference genomes are often stored in integrated microbial databases, such as the NCBI reference sequence database (RefSeq) [15], iMicrobe [16], EBI Metagenomics [17], IMG/M [18], and MG-RAST [19]. Nevertheless, these databases have extremely large data volumes and a high redundancy of genomes, which inevitably consume significant computational resources and time costs when used as reference genomes for compression.

In this work, we propose a scheme for the construction of lightweight and cohort-specific reference genomes. By overcoming the limitations of existing approaches and utilizing cohort-specific reference genomes, we achieve significant improvements in compression ratios and storage efficiency. Our results on diverse habitats demonstrate the effectiveness and applicability of our pipeline for enhancing the performance of reference-based compression tools and reducing storage costs.

2. Materials and Methods

2.1. Cohort Description

Studies from January 2015 to October 2022 with paired-end shotgun metagenomic sequencing on human gut, mouth, skin, vaginal, soil, marine, freshwater, and wastewater samples were searched on the National Center for Biotechnology Information (NCBI). In total, 29 studies with a sample size greater than 100 and data sizes greater than 100 GB of published metagenomic datasets were included (Table 1). Two hundred samples were randomly selected from each dataset (all were included if sample size < 200). The details of the total 5669 samples of all datasets are reported in Supplementary Table S1. Metagenomic datasets were locally downloaded from the European Nucleotide Archive (ENA) via the Aspera ascp command line client (v3.9.1). Sequencing depth distributions within these datasets are shown in Figure S1.

Table 1. Summary of the metagenomic datasets used in this study.

Habitats	BioProject	Sample Size	Data Size	Sample Used in This Study	References
Human gut	PRJEB11419	39,038	1003.18 Gb	200	[20]
	PRJEB12449	882	371.38 Gb	200	[21]
	PRJEB14847	372	1.18 Tb	200	[22]
	PRJEB19857	350	520.69 Gb	200	[23]
	PRJEB24041	633	480.72 Gb	200	[24]
	PRJEB32631	1679	1.34 Tb	200	[25]
	PRJEB34871	1197	1.13 Tb	200	[26]
	PRJEB39223	2196	2.97 Tb	200	[27]
	PRJEB39610	644	974.51 Gb	200	[28]
	PRJNA613947	348	871.69 Gb	200	[29]
Human mouth	PRJNA745160	888	657.43 Gb	200	[30]
	PRJNA647796	108	219.63 Gb	108	[31]
	PRJEB6997	530	2.01 Tb	200	[32]

Table 1. *Cont.*

Habitats	BioProject	Sample Size	Data Size	Sample Used in This Study	References
Human skin	PRJNA46333	8774	1.87 Tb	200	[33]
	PRJNA604820	516	269.29 Gb	200	[34]
	PRJNA763232	289	210.76 Gb	200	[35]
Human vagina	PRJNA797778	542	649.33 Gb	200	[36]
Soil	PRJEB42019	7557	726.36 Gb	200	[37]
	PRJEB24121	290	180.83 Gb	200	[38]
	PRJEB44414	195	239.80 Gb	195	[39]
Marine	PRJNA656268	1942	1.16 Tb	200	[40]
	PRJNA681031	305	268.79 Gb	200	[41]
	PRJEB38290	308	1.11 Tb	200	[42]
Freshwater	PRJNA287840	729	154.65 Gb	200	[43]
	PRJNA662092	184	265.07 Gb	184	[44]
	PRJNA418866	790	335.47 Gb	116	[45]
Wastewater	PRJNA746354	3188	447.20 Gb	200	[46]
	PRJNA801677	2105	452.38 Gb	200	[47]
	PRJEB31650	567	689.20 Gb	200	[48]

2.2. Data Pre-Processing and Taxonomic Profiling

Metagenomic sequencing reads were trimmed using fastp (v0.23.1) with a minimum N base number of 0 (--n_base_limit 0) and minimum read length of 60 (--length_required 60) for trimming. Taxonomic profiles were obtained using the default parameters of MetaPhlAn3 [49], which uses a database of clade-specific markers to quantify bacteria constituents at the species and higher taxonomic levels.

2.3. Data Analysis

To evaluate the heterogeneity of different datasets, we performed a CLR transformation of the species relative abundance data, followed by principal component analysis (PCA) using the scikit-learn module in Python. For each sample, the proportion of species that contributed more than 80% to abundance (abundant species) was first calculated. Next, the richness of the abundant species of various datasets was evaluated using sampling-unit-based rarefaction and extrapolation curves, which were generated using the first (species richness, q = 0) Hill number. The rarefaction/extrapolations were computed using the R package iNEXT (v2.0.12) [50] as the mean of 100 replicate bootstrapping runs to estimate 95% confidence intervals.

2.4. Construction of the Basic Reference Database

We chose MetaPhlAn3 as the basis for the construction of our basic reference database, as its marker gene database is the largest and most comprehensive available. Compared with other tools based on marker genes or 16S rRNA, MetaPhlAn3 can detect more microbial species, including some rare or newly discovered ones [49]. We constructed the MetaPhlAan3_db_25k database by selecting the best genomes for each species from NCBI based on the species list of MetaPhlAn3. For each species, we chose representative genomes or reference genomes if available, or otherwise the highest quality assembled genomes. We also downloaded some external genomes, which belong to different clades but also contain MetaPhlAn3 marker genes. These genomes may have acquired marker genes from different species due to horizontal gene transfer or other mechanisms. We believe that these genomes are valuable because they can increase the species diversity and coverage of the basic reference database. For genomes in the form of chromosomes, scaffolds, or contigs, we removed the gaps in the genomes and concatenated all segments into one complete sequence for subsequent alignment and compression. For the MetaPhlAan3_db_25k database, we collected a total of 25,435 microbial genomes, covering various taxa such as bacteria,

archaea, and eukaryotes. Similarly, we also constructed the Env_db_6k database, which is a database containing 6149 reference genomes of environmentally relevant microorganisms. Compared with the MetaPhlAan3_db_25k database, the Env_db_6k database contains more environmental microbial species, but it also has some overlaps and crossovers. The two databases can complement each other in reflecting the diversity and variability of the microbial community.

2.5. Evaluation of Short-Read Sequence Aligners

Ref1000 and a subset of 10 samples from PRJEB11419 were used to evaluate the performance of aligners during indexing and alignment. Ref1000 is a collection of 1000 bacterial genomes randomly selected from the MetaPhlAn3_db_25k database. Details of Ref1000 and the sample dataset can be found in Tables S2 and S3, respectively. We evaluate the time and memory consumption of short-read aligners during indexing using one thread. The BWA-index was run with default parameters because it does not provide a multi-threaded mode.

2.6. Construction of Cohort-Specific Reference Genomes and Compression

We randomly selected 50 samples from each dataset and built the cohort-specific reference genomes consisting of the top 1000 genome sequences with the highest alignment rates based on the Met-aPhlAn3_db_25k and Env_db_6k databases. The rest of the samples in each dataset were used for compression tests. The sequencing data of each sample were compressed using Genozip (v13.0.20) under both reference-free and reference-based modes, followed by the calculation of the compression ratio (size of the original FASTQ file/size of the compressed file). The following parameters were used to obtain the best compression: "--best=No_REF" for reference-free mode; "—best" and "—pair" for reference-based mode.

3. Results

3.1. Microbiota Composition Analysis of Samples from Various Cohorts

We collected 5669 metagenomic samples from 29 public datasets and divided them into human and environmental samples based on whether they were collected from human body parts (gut/mouth/skin/vagina) or environmental materials (soil/marine/freshwater/wastewater). The detailed information of these datasets is shown in Table 1. All shotgun sequencing data of human samples were processed with MetaPhlAn3 for quantitative species-level taxonomic profiling, and then the microbiota composition of samples from various cohorts was evaluated (Figure 1). Compositional PCA for beta diversity showed a high dispersion of the data (Figure 1A). For gut samples, PCA demonstrated the separation of samples from different cohorts (PC1: 22.03%, PC2: 4.43%). Samples of the majority of cohorts tended to cluster to the bottom left or right, indicating inter-cohort similarity in microbiota composition. An exception was PRJEB14847, samples which were far away from other cohorts and clustered into two centers, suggesting the existence of within-study clusters. The microbiota composition of mouth samples showed a high dispersion, with samples from different cohorts clustered separately into centers. The distribution of skin samples was highly concentrated and had the least variation across all habitats (PC1: 14.05%, PC2: 5.41%), whereas within-cohort vaginal samples showed a higher degree of variation, but the PCA ultimately only explains a small amount of the total variance (PC1: 16.54, PC2: 13.95%). Overall, the separation of samples across cohorts is greater than the separation within cohorts.

Abundant species are those that comprise the top 80% of the total abundance in a sample when ranked by decreasing relative abundance. The proportion of abundant species of different cohorts was consistently low, with a median of 8.15% to 21.57% (Figure 1B). This result has been previously observed in other studies, where a small number of species contributed most of the abundance in the entire microbial community [51,52]. As a function of sampling effort, the rarefaction/extrapolation (R/E) curve provides the expected number of observed/predicted abundant species, with the number of abundant species in most cohorts reaching saturation at 50 samples (Figure 1C).

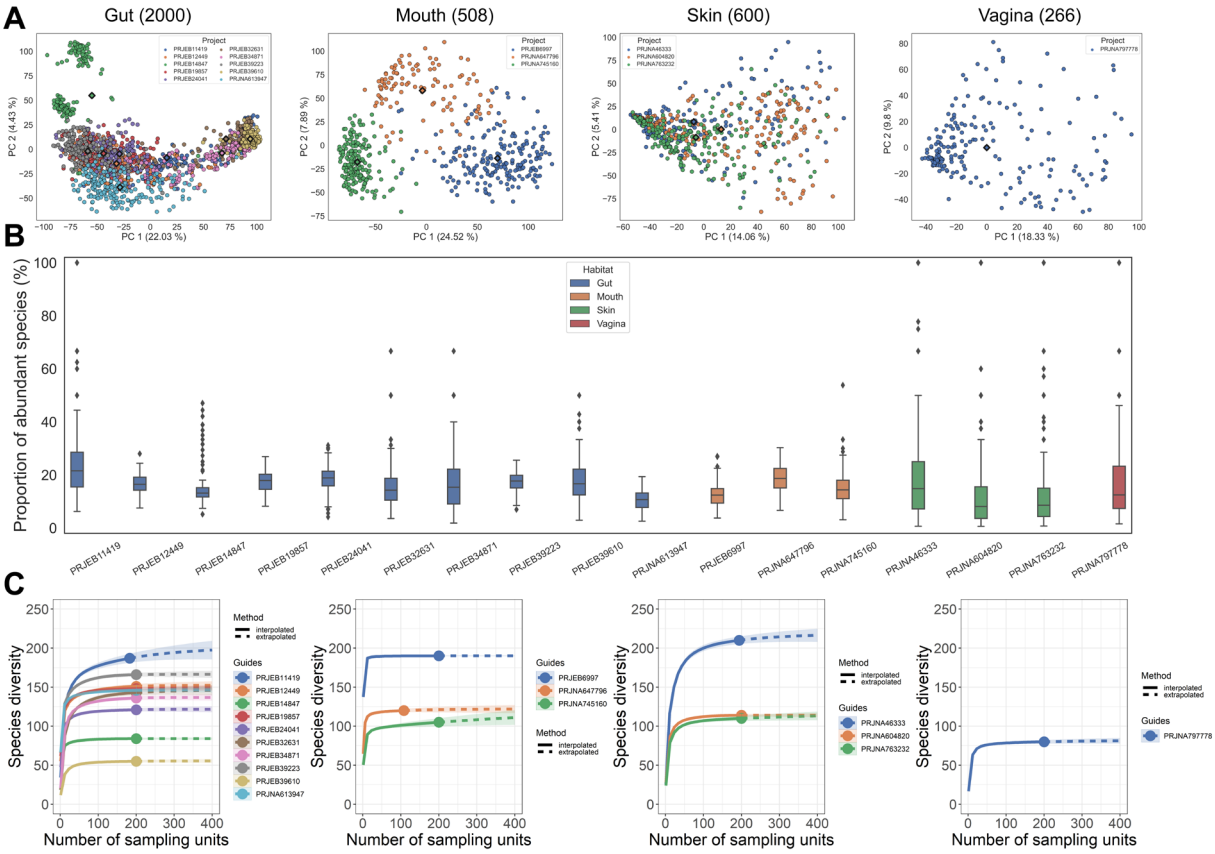

Figure 1. Microbiota composition analysis of human samples from various cohorts. (**A**) PCA demonstrates the clustering of samples from different cohorts based on the CLR-transformed species relative abundance data of various cohorts. The diamonds indicate cluster centroids. Number in parentheses represents the sample size. (**B**) Boxplot illustrates the proportion of abundant species in each sample. Abundance species are those that made up the top 80% of the total abundance in a sample. Each dot represents a sample from that cohort. Outliers are shown as dots beyond the whiskers of the boxplot, which reflect samples that deviate significantly from the normal range of abundant species proportions. Number in parentheses represents the sample size. (**C**) Sampling-unit-based rarefaction and extrapolation (R/E) curves of abundant species of each cohort. The dots indicate the actual number of specimen records and separate the interpolated (solid line) from extrapolated (dashed line) regions of each curve. The shaded areas represent 95% confidence intervals (based on a bootstrap method with 100 replications). Number in parentheses represents the sample size.

3.2. Workflow of Constructing Reference Genomes for Specific Cohorts

The first step of the analysis workflow is the sequencing quality control to obtain high-quality reads, followed by the alignment of short reads to the genomes in the basic reference database (Figure 2). Next, reads that do not meet the post-filtering criteria are discarded: low-quality reads (MAPQ < 5), reads with non-perfect matches (insertion, deletion, skipped region, soft clipping, and hard clipping), and reads with more than three mismatches. For reads that mapped to multiple positions in the reference genome, we only kept records of the best alignment. Reference genomes were ranked based on the mapping rate (number of reads mapped to each reference genome/number of total reads), and reference genomes for each cohort were constructed depending on the number of genomes specified by users. In brief, users can easily construct cohort-specific reference genomes of any size by simply providing shotgun sequencing data.

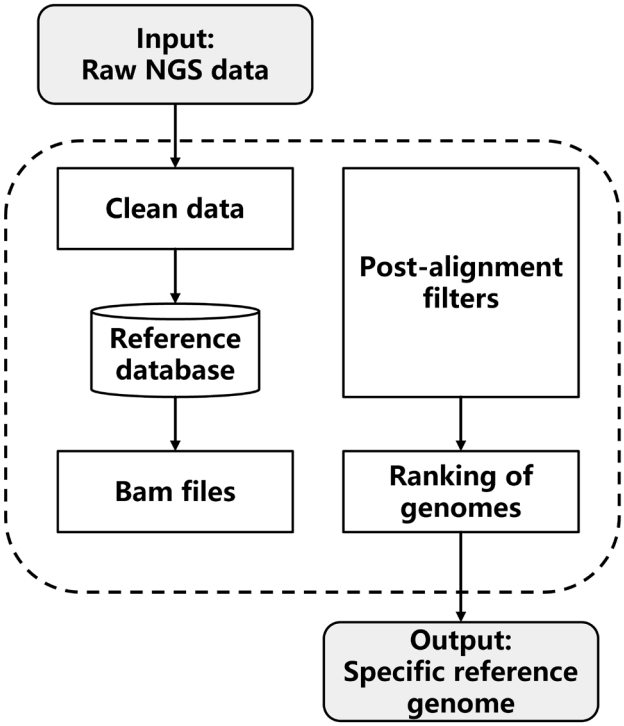

Figure 2. Workflow of reference genome construction. The next-generation sequencing (NGS) reads of input samples were trimmed and then mapped to the reference database. Post-alignment filters were used to remove low-quality reads, characterized by sequencing errors, small variants (indels and SNPs), and multiple mismatches. Genomes in the reference database were ranked based on the matching rate, and the final specific reference genome was constructed according to the output numbers of genomes specified by users.

3.3. Evaluation of Short-Read Sequence Aligners

The construction of cohort-specific reference genomes relies on the alignment of short sequencing reads to the basic reference database, which necessitates the inclusion of as many microbial reference sequences as feasible in the database. Large and comprehensive reference databases require more computational resources, especially memory. Therefore, we evaluated the performance of three commonly used aligners, Bowtie2 (v2.3.5.1) [53], BWA (v0.7.17) [54], and Minimap2 (v2.24) [55], during indexing and alignment. Minimap2 shows a significant speed advantage as it is about 92 times faster than Bowtie2 and 35 times faster than BWA, but at the cost of high memory usage (Table 2). BWA has the lowest memory consumption among all aligners, making it superior for indexing large reference databases.

Table 2. Indexing time and peak memory of indexing for aligners using one thread.

Aligner	Ref1000 (3.8 GB)	
	Time (s)	Mem (GB)
Bowtie2	15,081	10
BWA	5821	6
Minimap2	164	25

Next, we investigated how the aligners scaled with the number of threads by running them with one, four, and eight threads as multithreading is the standard use case (Figure 3). The alignment time nearly halves for the aligners when doubling the number of threads, suggesting that the tools make efficient use of the resources. Minimap2 runs the fastest of all the aligners, but this advantage diminishes as the number of threads rises. In terms of

memory usage, Bowtie2 has the lowest peak memory across all experiments. Followed by this is BWA, whose memory usage is slightly higher than Bowtie2 when using one thread but increases fast as the number of threads grows. BWA and Bowtie2 are both BWT-based aligners, whose indices may be shared by multiple tasks to reduce memory consumption. Minimap2's memory usage is relatively high compared to memory-efficient tools Bowtie2 and BWA, which reach a peak memory usage of 15 GB when running 10 tasks simultaneously. In summary, BWA and Bowtie2 may save more memory when working with large reference genome databases, while Minimap2 is faster when using small reference databases. A large reference database could be divided into sub-databases when memory resources are restricted, and our pipeline will automatically integrate the results after mapping them individually.

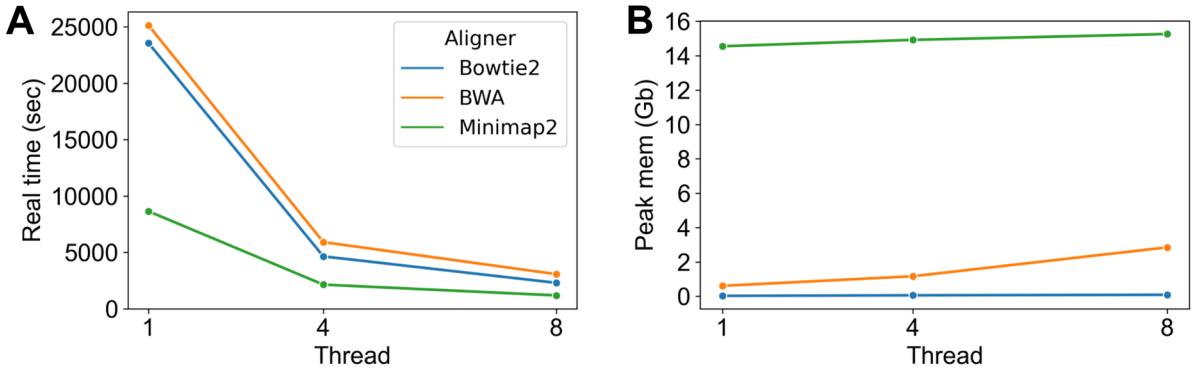

Figure 3. Real time and peak memory in multithread mode. (**A**) The real time (in seconds) of three aligners (Bowtie2, BWA and Minimap2) with different number of threads (1, 4 and 8). (**B**) The peak memory (in GB) of three aligners with different number of threads.

3.4. Construction of Specific Reference Genomes for Each Dataset

Comprehensive microbial reference databases are essential for the construction of specific reference genomes; thus, we constructed two microbial reference databases, MetaPhlAn3_db_25k and Env_db_6k, applicable to human and environmental samples. MetaPhlAn3_db_25k is a microbial reference database comprising 25,435 bacterial, archaeal, and eukaryotic reference genomes. As a complement to MetaPhlAn3_db_25k, the Env_db_6k database contains 6149 reference genomes of microorganisms associated with environmental metagenomes (mainly soil and water). We designated MetaPhlAn3_db_25k as the basic reference database for human-associated samples and randomly selected 50 samples from each dataset for the construction of specific reference genomes, as 50 samples might reflect the abundance species composition of the whole cohort. To identify the optimal number of genome sequences, we investigated how the mapping rate varied with the number of genomes (Figure S2). The mapping rate of most cohorts saturated when the number of genomes reached 1000; hence, we output the top 1000 genomes as the specific reference genome for each dataset. The size of the reference genome constructed for each dataset ranged from 2.4 to 3.8 GB, which is comparable to the size of the human genome. For environment-associated samples, both the MetaPhlAn3_db_25k and Env_db_6k databases were used as the basic reference databases to construct the reference genomes of 2.8–3.9 GB in size for each dataset.

3.5. Performance of Reference Genome-Based Compression of Each Dataset

We evaluated the performance of reference-based compression using Genozip, a lossless compression tool designed for genomic data with both reference-based and reference-free modes [13]. Genozip was selected for its superior performance in compressing genomic sequences compared with other methods. Moreover, it supports various input and output formats, such as FASTA, FASTQ, SAM, BAM, CRAM, etc., and provides a series of extendable downstream analysis tools for high flexibility and convenience. In general, reference-based compression yielded a higher compression ratio than the general compres-

sion tool Gzip and the reference-free mode (Figure 4, Table S4). The average compression ratio of reference-based compression for the total of 2457 human samples is 9.9, which is 2.5 times and 1.6 times better than that of Gzip and reference-free mode, respectively. When compressed using cohort-specific reference genomes, the compression ratios for all samples exceeded four, and half of the samples (1401/2457) reached eight. Over 10% (290/2457) of samples obtained a score of 15, with the best result being 29.4. The average compression ratio of each cohort ranged from 5.6 to 18.1, with an average improvement of 74–264% over Gzip (Table S4). The low compression ratios of some datasets were potentially due to the relatively shallow sequencing depth, such as PRJEB24041 and PRJEB12449 from the human gut and PRJNA46333 from skin (Figure S1). The average compression ratio for 1693 environmental samples is 8.2, which is 1.9 times better than that of Gzip and 1.1 times higher than the reference-free method (Table S4). Approximately 7% (120/1693) of the environmental samples were able to obtain a compression ratio of 15, with the best result being 33.8. Samples with high compression ratios were mostly from the wastewater datasets PRJNA80167 and PRJNA746354. The average compression ratios for each dataset ranged from 5.5 to 17.2, with an average improvement of 49–183% compared to Gzip.

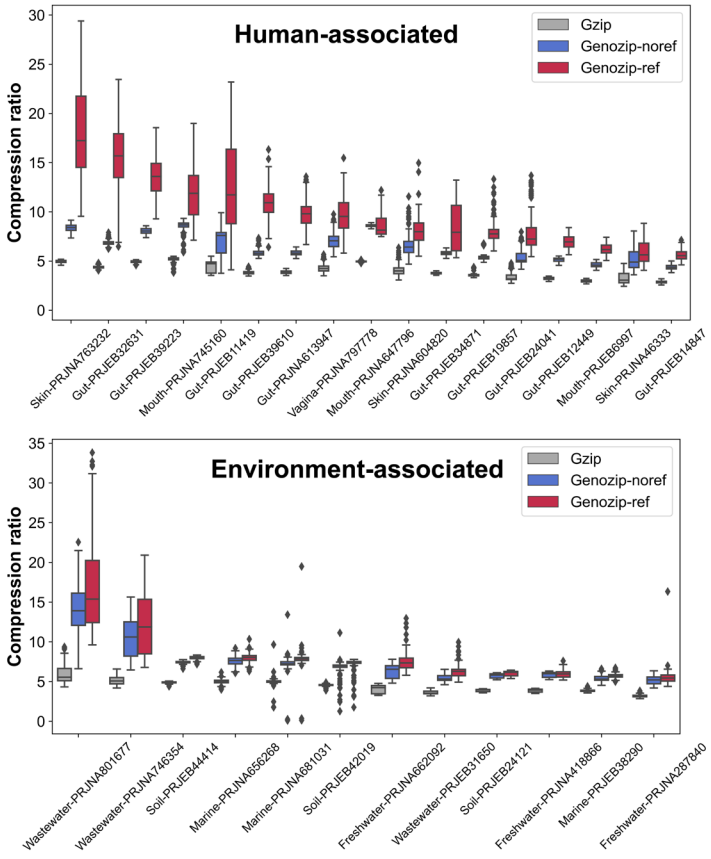

Figure 4. Compression ratio of samples from 29 cohorts under different compression tools and modes. The gray, blue, and red boxes represent the compression ratios of samples compressed with Gzip, Genozip without reference, and Genozip with reference, respectively. Compression ratio is defined as the ratio of the original file size to the compressed file. Outliers are shown as dots beyond the whiskers of the boxplot.

4. Discussion

Novel reference-based compression tools for genomic data have been developed in recent years, but their acceptable reference genome size is usually limited, yet sensitive to the choice of reference genomes [56–58]. To streamline current reference databases and enhance the compression ratio of metagenomes, we introduce a pipeline for constructing a

lightweight and cohort-specific reference genome using a small sample size, which can be used for compressing large cohort samples.

We collect genome sequences with high mapping rates in the reference database to create a cohort-specific reference genome, as the mapping rate has a positive correlation with the compression ratio for samples within the same cohort. Nevertheless, it should be noted that high mapping rates may not necessarily result in high compression rates when comparing across cohorts. This result is predictable as the reference genome only impacts the compression of nucleotide sequences, while the actual compression also includes header lines and quality scores. Quality scores occupy a significant chunk of the storage space and are more difficult to compress compared with other components of FASTQ files [59]. The coverage and depth of each contig can be shown by the extension tools of Genozip, which is a valuable feature that may simplify our pipeline for the construction of reference genomes. Unfortunately, we were unable to test this feature using either MetaPhlAn3_db_25k or Env_db_6k reference databases due to the limits of the free edition of Genozip regarding the size of the reference genome. In addition to the mapping rate, other factors such as the location, host information, and collection method of the samples may also affect the quality and compression efficiency of the constructed reference sequences. However, some metadata (e.g., host information and collection methods) are difficult to access due to inconsistencies in the metadata formats of different public platforms, which poses challenges for further analysis and comparison. We encourage future researchers to provide richer metadata information when uploading data to facilitate subsequent analysis and comparison.

Our findings indicate that environmental samples have lower species richness and less compression improvement than human samples when using reference genomes, as illustrated in Figure S3. This highlights the diversity and complexity of environmental microbiome communities, and the lack of adequate and comprehensive reference genomes for less-studied microbial species. To enhance the quality of reference genomes and improve habitat specificity, minimizing genomic redundancy and developing habitat-specific reference databases have proven effective. Notably, several well-curated reference genome datasets of the human microbiome have been developed in recent years, including those from low-abundance and unculturable bacteria, enriching the current reference catalog for the human microbiome [60–62]. Efforts have also been made to develop habitat-specific reference databases for environmental metagenomes. For example, Choi et al. developed RefSoil, a curated reference database of soil microbial genomes, and extended it to include plasmids of soil microorganisms as the RefSoil+ database [63,64]. Klemetsen et al. constructed the marine prokaryotic genome databases MarRef and MarDB using non-redundant genome and metagenome datasets obtained from ENA and NCBI [65]. These habitat-specific databases have facilitated the classification and analysis of metagenomic samples, and as these specific databases continue to be expanded and improved on, reference genomes constructed using our pipeline will become more targeted.

In summary, our work offers fresh ideas for reference-based compression and the construction of cohort-specific reference genomes. With approximately 2–4 GB of reference genomes, we achieved impressive compression ratios of 5.5–18.1 (1.5–3.6 times better than Gzip) across 29 datasets, with significant savings in storage space and cost. This is especially beneficial for large-scale genomic projects that generate and store massive amounts of data. Our method reduces the I/O time for large files and improves the efficiency and scalability of genomic data processing and analysis pipelines. Furthermore, our approach is adaptable to improvements in storage systems, which can enhance data access speed. This, in turn, facilitates rapid data retrieval and sharing among researchers and clinicians.

While our method has demonstrated impressive advantages in reference-based compression, it is important to acknowledge its limitations and consider future directions for improvement. Although the size of the microbial reference genome is already approaching that of the human genome, there is still potential for further compression by removing redundant regions based on their coverage degree. It is also important to test our method on third-generation sequencing data, which may have different characteristics and require-

ments for compression compared to NGS short reads. A significant challenge is to improve the availability and quality of reference genomes in the database, especially for less-studied microbial species, which may affect the compression efficiency for samples collected from environmental sources. As genome databases continue to evolve and improve, we anticipate that the quality and specificity of cohort-specific databases will enhance, further strengthening the effectiveness and applicability of our method.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/microorganisms11102560/s1, Figure S1: Kernel distribution estimation plot of sequence depths within each dataset used in this study; Figure S2: Variation in the mapping rates of various datasets with the number of genomes; Figure S3: Scatter plot of species richness and compression ratio for 29 datasets; Table S1: Metadata of metagenomic samples used in this study; Table S2: Summary of 1000 bacterial reference genomes (Ref1000); Table S3: Metadata of samples used to evaluate the performance of short-read aligners; Table S4: Summary of reference genomes and compression ratio for each dataset.

Author Contributions: Conceptualization, Y.Z., L.W., R.D. and S.H.; investigation, L.W., R.D. and S.H.; writing—original draft preparation, L.W., R.D., Q.W. and Y.Z. All authors discussed the results and implications and commented on the manuscript at all stages. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Open-source software was used for analysis, including the Aspera ascp command line client (v3.9.1) for data downloading, fastp (v0.23.1) and MetaPhlAn3 for data pre-processing, Python and R package iNEXT for data analysis, BWA for data evaluation, and Genozip (v13.0.20) for data compression test. Our pipeline can be freely download at https://github.com/wanglinqi123/MetaRef (accessed on 20 March 2022). We provide the Accession number of microbial genomes for MetaPhlAn3_db_25k and Env_db_6k databases on Github, and users can build these two base databases by themselves using scripts. The download address is https://github.com/wanglinqi123/MetaRef/tree/main/BasicRefDB (accessed on 20 March 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Handelsman, J.; Rondon, M.R.; Brady, S.F.; Clardy, J.; Goodman, R.M. Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chem. Biol.* **1998**, *5*, R245–R249. [CrossRef] [PubMed]
2. Gevers, D.; Knight, R.; Petrosino, J.F.; Huang, K.; McGuire, A.L.; Birren, B.W.; Nelson, K.E.; White, O.; Methe, B.A.; Huttenhower, C. The Human Microbiome Project: A community resource for the healthy human microbiome. *PLoS Biol.* **2012**, *10*, e1001377. [CrossRef] [PubMed]
3. Gilbert, J.A.; Jansson, J.K.; Knight, R. The Earth Microbiome project: Successes and aspirations. *BMC Biol.* **2014**, *12*, 69. [CrossRef] [PubMed]
4. Ehrlich, S.D.; Consortium, M. MetaHIT: The European Union Project on metagenomics of the human intestinal tract. In *Metagenomics of the Human Body*; Nelson, K.E., Ed.; Springer: New York, NY, USA, 2011; pp. 307–316, ISBN 978-1-4419-7089-3.
5. Sunagawa, S.; Acinas, S.G.; Bork, P.; Bowler, C.; Tara Oceans, C.; Eveillard, D.; Gorsky, G.; Guidi, L.; Iudicone, D.; Karsenti, E.; et al. Tara Oceans: Towards global ocean ecosystems biology. *Nat. Rev. Microbiol.* **2020**, *18*, 428–445. [CrossRef]
6. Bonfield, J.K.; Mahoney, M.V. Compression of FASTQ and SAM format sequencing data. *PLoS ONE* **2013**, *8*, e59190. [CrossRef]
7. Hach, F.; Numanagic, I.; Alkan, C.; Sahinalp, S.C. SCALCE: Boosting sequence compression algorithms using locally consistent encoding. *Bioinformatics* **2012**, *28*, 3051–3057. [CrossRef]
8. Selva, J.J.; Chen, X. SRComp: Short read sequence compression using burstsort and Elias omega coding. *PLoS ONE* **2013**, *8*, e81414. [CrossRef]
9. Janin, L.; Schulz-Trieglaff, O.; Cox, A.J. BEETL-fastq: A searchable compressed archive for DNA reads. *Bioinformatics* **2014**, *30*, 2796–2801. [CrossRef]
10. Fritz, M.H.-Y.; Leinonen, R.; Cochrane, G.; Birney, E. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.* **2011**, *21*, 734–740. [CrossRef]
11. Huang, Z.A.; Wen, Z.; Deng, Q.; Chu, Y.; Sun, Y.; Zhu, Z. LW-FQZip 2: A parallelized reference-based compression of FASTQ files. *BMC Bioinform.* **2017**, *18*, 179. [CrossRef]
12. Jones, D.C.; Ruzzo, W.L.; Peng, X.; Katze, M.G. Compression of next-generation sequencing reads aided by highly efficient de novo assembly. *Nucleic Acids Res.* **2012**, *40*, e171. [CrossRef] [PubMed]

13. Lan, D.; Tobler, R.; Souilmi, Y.; Llamas, B. Genozip—A Universal Extensible Genomic Data Compressor. *Bioinformatics* **2021**, *37*, 2225–2230. [CrossRef] [PubMed]
14. Hach, F.; Numanagic, I.; Sahinalp, S.C. DeeZ: Reference-based compression by local assembly. *Nat. Methods* **2014**, *11*, 1082–1084. [CrossRef] [PubMed]
15. O'Leary, N.A.; Wright, M.W.; Brister, J.R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **2015**, *44*, D733–D745. [CrossRef]
16. Youens-Clark, K.; Bomhoff, M.; Ponsero, A.J.; Wood-Charlson, E.M.; Lynch, J.; Choi, I.; Hartman, J.H.; Hurwitz, B.L. iMicrobe: Tools and data-dreaiven discovery platform for the microbiome sciences. *GigaScience* **2019**, *8*, giz083. [CrossRef]
17. Mitchell, A.; Bucchini, F.; Cochrane, G.; Denise, H.; ten Hoopen, P.; Fraser, M.; Pesseat, S.; Potter, S.; Scheremetjew, M.; Sterk, P.; et al. EBI metagenomics in 2016–An expanding and evolving resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res.* **2016**, *44*, D595–D603. [CrossRef]
18. Chen, I.A.; Markowitz, V.M.; Chu, K.; Palaniappan, K.; Szeto, E.; Pillay, M.; Ratner, A.; Huang, J.; Andersen, E.; Huntemann, M.; et al. IMG/M: Integrated genome and metagenome comparative data analysis system. *Nucleic Acids Res.* **2017**, *45*, D507–D516. [CrossRef]
19. Meyer, F.; Paarmann, D.; D'Souza, M.; Olson, R.; Glass, E.M.; Kubal, M.; Paczian, T.; Rodriguez, A.; Stevens, R.; Wilke, A.; et al. The metagenomics RAST server—A public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinform.* **2008**, *9*, 386. [CrossRef]
20. McDonald, D.; Hyde, E.; Debelius, J.W.; Morton, J.T.; Gonzalez, A.; Ackermann, G.; Aksenov, A.A.; Behsaz, B.; Brennan, C.; Chen, Y.; et al. American Gut: An Open Platform for Citizen Science Microbiome Research. *mSystems* **2018**, *3*, e00031-18. [CrossRef]
21. Vogtmann, E.; Hua, X.; Zeller, G.; Sunagawa, S.; Voigt, A.Y.; Hercog, R.; Goedert, J.J.; Shi, J.; Bork, P.; Sinha, R. Colorectal Cancer and the Human Gut Microbiome: Reproducibility with Whole-Genome Shotgun Sequencing. *PLoS ONE* **2016**, *11*, e0155362. [CrossRef]
22. Costea, P.I.; Zeller, G.; Sunagawa, S.; Pelletier, E.; Alberti, A.; Levenez, F.; Tramontano, M.; Driessen, M.; Hercog, R.; Jung, F.E.; et al. Towards standards for human fecal sample processing in metagenomic studies. *Nat. Biotechnol.* **2017**, *35*, 1069–1076. [CrossRef] [PubMed]
23. Pre-BreedYield Consortium EMBL Nucleotide Sequence Database (Project PRJEB19857). Available online: https://www.ebi.ac.uk/ena/browser/view/PRJEB19857 (accessed on 1 August 2022).
24. Korpela, K.; Costea, P.; Coelho, L.P.; Kandels-Lewis, S.; Willemsen, G.; Boomsma, D.I.; Segata, N.; Bork, P. Selective maternal seeding and environment shape the human gut microbiome. *Genome Res.* **2018**, *28*, 561–568. [CrossRef] [PubMed]
25. Shao, Y.; Forster, S.C.; Tsaliki, E.; Vervier, K.; Strang, A.; Simpson, N.; Kumar, N.; Stares, M.D.; Rodger, A.; Brocklehurst, P.; et al. Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature* **2019**, *574*, 117–121. [CrossRef]
26. Auguet, O.T.; Niehus, R.; Gweon, H.S.; Berkley, J.A.; Waichungo, J.; Njim, T.; Edgeworth, J.D.; Batra, R.; Chau, K.; Swann, J.; et al. Population-level faecal metagenomic profiling as a tool to predict antimicrobial resistance in Enterobacterales isolates causing invasive infections: An exploratory study across Cambodia, Kenya, and the UK. *E Clin. Med.* **2021**, *36*, 100910. [CrossRef] [PubMed]
27. Asnicar, F.; Berry, S.E.; Valdes, A.M.; Nguyen, L.H.; Piccinno, G.; Drew, D.A.; Leeming, E.; Gibson, R.; Le Roy, C.; Khatib, H.A.; et al. Microbiome connections with host metabolism and habitual diet from 1098 deeply phenotyped individuals. *Nat. Med.* **2021**, *27*, 321–332. [CrossRef] [PubMed]
28. Masi, A.C.; Embleton, N.D.; Lamb, C.A.; Young, G.; Granger, C.L.; Najera, J.; Smith, D.P.; Hoffman, K.L.; Petrosino, J.F.; Bode, L.; et al. Human milk oligosaccharide DSLNT and gut microbiome in preterm infants predicts necrotising enterocolitis. *Gut* **2021**, *70*, 2273–2282. [CrossRef]
29. Valles-Colomer, M.; Blanco-Míguez, A.; Manghi, P.; Asnicar, F.; Dubois, L.; Golzato, D.; Armanini, F.; Cumbo, F.; Huang, K.D.; Manara, S.; et al. The person-to-person transmission landscape of the gut and oral microbiomes. *Nature* **2023**, *614*, 125–135. [CrossRef]
30. Pettigrew, M.M.; Kwon, J.; Gent, J.F.; Kong, Y.; Wade, M.; Williams, D.J.; Creech, C.B.; Evans, S.; Pan, Q.; Walter, E.B.; et al. Comparison of the Respiratory Resistomes and Microbiota in Children Receiving Short versus Standard Course Treatment for Community-Acquired Pneumonia. *mBio* **2022**, *13*, e0019522. [CrossRef]
31. Pre-BreedYield Consortium EMBL Nucleotide Sequence Database (Project PRJNA647796). Available online: https://www.ebi.ac.uk/ena/browser/view/PRJNA647796 (accessed on 1 August 2022).
32. Zhang, X.; Zhang, D.; Jia, H.; Feng, Q.; Wang, D.; Liang, D.; Wu, X.; Li, J.; Tang, L.; Li, Y.; et al. The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nat. Med.* **2015**, *21*, 895–905. [CrossRef]
33. Oh, J.; Byrd, A.L.; Deming, C.; Conlan, S.; Kong, H.H.; Segre, J.A. Biogeography and individuality shape function in the human skin metagenome. *Nature* **2014**, *514*, 59–64. [CrossRef]
34. Jo, J.H.; Harkins, C.P.; Schwardt, N.H.; Portillo, J.A.; Zimmerman, M.D.; Carter, C.L.; Hossen, M.A.; Peer, C.J.; Polley, E.C.; Dartois, V.; et al. Alterations of human skin microbiome and expansion of antimicrobial resistance after systemic antibiotics. *Sci. Transl. Med.* **2021**, *13*, eabd8077. [CrossRef] [PubMed]
35. Pre-BreedYield Consortium EMBL Nucleotide Sequence Database (Project PRJNA763232). Available online: https://www.ebi.ac.uk/ena/browser/view/PRJNA763232 (accessed on 1 August 2022).

36. France, M.T.; Fu, L.; Rutt, L.; Yang, H.; Humphrys, M.S.; Narina, S.; Gajer, P.M.; Ma, B.; Forney, L.J.; Ravel, J. Insight into the ecology of vaginal bacteria through integrative analyses of metagenomic and metatranscriptomic data. *Genome Biol.* **2022**, *23*, 66. [CrossRef] [PubMed]
37. Pre-BreedYield Consortium EMBL Nucleotide Sequence Database (Project PRJEB42019). Available online: https://www.ebi.ac.uk/ena/browser/view/PRJEB42019 (accessed on 1 August 2022).
38. Pre-BreedYield Consortium EMBL Nucleotide Sequence Database (Project PRJEB24121). Available online: https://www.ebi.ac.uk/ena/browser/view/PRJEB24121 (accessed on 1 August 2022).
39. Bahram, M.; Espenberg, M.; Pärn, J.; Lehtovirta-Morley, L.; Anslan, S.; Kasak, K.; Kõljalg, U.; Liira, J.; Maddison, M.; Moora, M.; et al. Structure and function of the soil microbiome underlying N_2O emissions from global wetlands. *Nat. Commun.* **2022**, *13*, 1430. [CrossRef] [PubMed]
40. Larkin, A.A.; Garcia, C.A.; Garcia, N.; Brock, M.L.; Lee, J.A.; Ustick, L.J.; Barbero, L.; Carter, B.R.; Sonnerup, R.E.; Talley, L.D.; et al. High spatial resolution global ocean metagenomes from Bio-GO-SHIP repeat hydrography transects. *Sci. Data* **2021**, *8*, 107. [CrossRef]
41. Pre-BreedYield Consortium EMBL Nucleotide Sequence Database (Project PRJNA681031). Available online: https://www.ebi.ac.uk/ena/browser/view/PRJNA681031 (accessed on 1 August 2022).
42. Schultz, D. Mechanisms of Polysaccharide-Degradation in Particle-Associated Microbial Communities. Available online: https://nbn-resolving.org/urn:nbn:de:gbv:9-opus-59014 (accessed on 1 August 2022).
43. Rossum, T.V.; Uyaguari-Diaz, M.I.; Vlok, M.; Peabody, M.A.; Tian, A.; Cronin, K.I.; Chan, M.; Croxen, M.A.; Hsiao, W.W.L.; Isaac-Renton, J.; et al. Spatiotemporal dynamics of river viruses, bacteria and microeukaryotes. *bioRxiv* **2018**, 259861. [CrossRef]
44. Pérez-Carrascal, O.M.; Tromas, N.; Terrat, Y.; Moreno, E.; Giani, A.; Corrêa Braga Marques, L.; Fortin, N.; Shapiro, B.J. Single-colony sequencing reveals microbe-by-microbiome phylosymbiosis between the cyanobacterium Microcystis and its associated bacteria. *Microbiome* **2021**, *9*, 194. [CrossRef]
45. Bai, Y.; Wang, Q.; Liao, K.; Jian, Z.; Zhao, C.; Qu, J. Fungal Community as a Bioindicator to Reflect Anthropogenic Activities in a River Ecosystem. *Front. Microbiol.* **2018**, *9*, 3152. [CrossRef]
46. Pre-BreedYield Consortium EMBL Nucleotide Sequence Database (Project PRJNA746354). Available online: https://www.ebi.ac.uk/ena/browser/view/PRJNA746354 (accessed on 1 August 2022).
47. Pre-BreedYield Consortium EMBL Nucleotide Sequence Database (Project PRJNA801677). Available online: https://www.ebi.ac.uk/ena/browser/view/PRJNA801677 (accessed on 1 August 2022).
48. Poulsen, C.S.; Ekstrom, C.T.; Aarestrup, F.M.; Pamp, S.J. Library Preparation and Sequencing Platform Introduce Bias in Metagenomic-Based Characterizations of Microbiomes. *Microbiol. Spectr.* **2022**, *10*, e0009022. [CrossRef]
49. Beghini, F.; McIver, L.J.; Blanco-Miguez, A.; Dubois, L.; Asnicar, F.; Maharjan, S.; Mailyan, A.; Manghi, P.; Scholz, M.; Thomas, A.M.; et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife* **2021**, *10*, e65088. [CrossRef]
50. Hsieh, T.C.; Ma, K.H.; Chao, A. iNEXT: An R package for rarefaction and extrapolation of species diversity (H ill numbers). *Methods Ecol. Evol.* **2016**, *7*, 1451–1456. [CrossRef]
51. Avolio, M.L.; Forrestel, E.J.; Chang, C.C.; La Pierre, K.J.; Burghardt, K.T.; Smith, M.D. Demystifying dominant species. *New Phytol.* **2019**, *223*, 1106–1126. [CrossRef] [PubMed]
52. Loftus, M.; Hassouneh, S.A.; Yooseph, S. Bacterial associations in the healthy human gut microbiome across populations. *Sci. Rep.* **2021**, *11*, 2828. [CrossRef] [PubMed]
53. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef]
54. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* **2013**, arXiv:1303.3997. [CrossRef]
55. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **2018**, *34*, 3094–3100. [CrossRef] [PubMed]
56. Li, P.; Jiang, X.; Wang, S.; Kim, J.; Xiong, H.; Ohno-Machado, L. HUGO: Hierarchical mUlti-reference Genome compression for aligned reads. *J. Am. Med. Inform. Assoc.* **2014**, *21*, 363–373. [CrossRef]
57. Deorowicz, S.; Danek, A. GTShark: Genotype compression in large projects. *Bioinformatics* **2019**, *35*, 4791–4793. [CrossRef]
58. Dufort y Álvarez, G.; Seroussi, G.; Smircich, P.; Sotelo-Silveira, J.; Ochoa, I.; Martín, Á. RENANO: A REference-based compressor for NANOpore FASTQ files. *Bioinformatics* **2021**, *37*, 4862–4864. [CrossRef]
59. Ochoa, I.; Asnani, H.; Bharadia, D.; Chowdhury, M.; Weissman, T.; Yona, G. QualComp: A new lossy compressor for quality scores based on rate distortion theory. *BMC Bioinform.* **2013**, *14*, 187. [CrossRef]
60. Almeida, A.; Nayfach, S.; Boland, M.; Strozzi, F.; Beracochea, M.; Shi, Z.J.; Pollard, K.S.; Sakharova, E.; Parks, D.H.; Hugenholtz, P.; et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **2021**, *39*, 105–114. [CrossRef]
61. Zhu, J.; Tian, L.; Chen, P.; Han, M.; Song, L.; Tong, X.; Sun, X.; Yang, F.; Lin, Z.; Liu, X.; et al. Over 50,000 Metagenomically Assembled Draft Genomes for the Human Oral Microbiome Reveal New Taxa. *Genom. Proteom. Bioinform.* **2022**, *20*, 246–259. [CrossRef] [PubMed]
62. Zou, Y.; Xue, W.; Luo, G.; Deng, Z.; Qin, P.; Guo, R.; Sun, H.; Xia, Y.; Liang, S.; Dai, Y.; et al. 1520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat. Biotechnol.* **2019**, *37*, 179–185. [CrossRef] [PubMed]
63. Choi, J.; Yang, F.; Stepanauskas, R.; Cardenas, E.; Garoutte, A.; Williams, R.; Flater, J.; Tiedje, J.M.; Hofmockel, K.S.; Gelder, B.; et al. Strategies to improve reference databases for soil microbiomes. *ISME J.* **2017**, *11*, 829–834. [CrossRef] [PubMed]

64. Dunivin, T.K.; Choi, J.; Howe, A.; Shade, A. RefSoil+: A Reference Database for Genes and Traits of Soil Plasmids. *mSystems* **2019**, *4*, e00349-18. [CrossRef] [PubMed]
65. Klemetsen, T.; Raknes, I.A.; Fu, J.; Agafonov, A.; Balasundaram, S.V.; Tartari, G.; Robertsen, E.; Willassen, N.P. The MAR databases: Development and implementation of databases specific for marine metagenomics. *Nucleic Acids Res.* **2017**, *46*, D692–D699. [CrossRef]