MIDAS2: Metagenomic Intra-species Diversity Analysis System

Chunyu Zhao, Boris Dimitrov, Miriam Goldman, Stephen Nayfach, Katherine S. Pollard

Abstract

Summary

The Metagenomic Intra-Species Diversity Analysis System (MIDAS) is a scalable metagenomic pipeline that identifies single nucleotide variants (SNVs) and gene copy number variants (CNVs) in microbial populations. Here, we present MIDAS2, which addresses the computational challenges presented by increasingly large reference genome databases, while adding functionality for building custom databases and leveraging paired-end reads to improve SNV accuracy. This fast and scalable reengineering of the MIDAS pipeline enables thousands of metagenomic samples to be efficiently genotyped.

Availability and Implementation

The source code is available at https://github.com/czbiohub/MIDAS2. The documentation is available at https://midas2.readthedocs.io/en/latest/.

Supplementary Information

Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Metagenotyping, the identification of intraspecific genetic variants in metagenomic data, is a powerful approach to characterizing population genetic diversity in microbiomes. Most pipelines identify variants based on alignment of reads to reference databases of microbial genomes and/or gene sequences (Supplementary Figure S1). While comprehensive reference databases can reveal strain-level relationships which would be otherwise overlooked (Beghini et al. 2021), alignment to large databases is computationally intensive. Furthermore, the divergence of reference genomes from strains in the metagenomic sample affects sensitivity and precision (Bush et al. 2020; Olm et al. 2021), and existing metagenotyping tools do not automatically adapt database files based on information in the metagenome. In this paper, we introduce MIDAS2 (Supplementary Figure S2), a major update to MIDAS

(Nayfach et al. 2016) (Supplementary Table S1) that addresses these challenges through (1) a new database infrastructure geared to run on AWS Batch and S3 that achieves elastic scaling for constructing database files from large collections of genomes; and (2) a fast and scalable implementation of the SNV calling pipeline that enables metagenotyping in thousands of samples with improved accuracy achieved through utilization of paired-end reads and databases customized to the species present in the samples. As the only tool that integrates all steps of the metagenotyping process, from database customization to alignment and variant calling, MIDAS2 helps to promote reproducible research.

2 IMPLEMENTATION

2.1 MIDAS Reference Database Updates

We generated MIDAS Reference Databases (MIDAS DBs), comprised of species pangenomes, marker genes, and representative genomes, from two public microbial genome collections: UHGG v.1 (Almeida et al. 2021) (4,644 species / 286,997 genomes) and GTDB v202 (Parks et al. 2022) (47,893 species / 258,405 genomes). This is a significant increase in database content compared to MIDAS DB v1.2 (5,952 species / 31,007 genomes) and other tools (Supplementary Table S2). We implemented a new infrastructure that dramatically simplifies building a new MIDAS DB for other genome collections by using a table-of-contents file assigning genomes to species and denoting the representative genome for each species (Supplemental Fig S3). MIDAS DBs can be built locally, which enables customized selection of representative genomes, a key component of accurate SNV calling (Bush et al. 2020).

2.2. MIDAS2 Analysis Algorithm Updates

Metagenotyping SNVs across large numbers of samples is computationally intensive. First, alignment and pileup are applied to each species in each sample (single-sample step). Then these pileup results must be scanned for each genomic site to compute population SNVs (across-samples step). Metagenotyping methods such as MIDAS and metaSNV v2 parallelize these computations over species, capping the maximum number of processors (CPUs) at the number of species being genotyped. The SNV module of MIDAS2 achieves better CPU utilization by splitting

genomic sites into multiple chunks per species. We execute parallelization over chunks in a way that does not destroy cache coherence to the point where computation stalls on input or output (I/O; Supplementary Notes).

3 RESULTS

3.1 Performance Improvements

We compared the running time and memory utilization of the single-sample and across-samples SNV modules of MIDAS and MIDAS2, using the same database (MIDAS DB v1.2) and 211 samples from an inflammatory bowel disease cohort (NCBI accession: PRJNA400072). The single-sample SNV module of MIDAS2 is slightly faster than MIDAS (Supplementary Figure S4), with database customization and Bowtie2 alignment taking up to 75% of run time (Supplementary Figure S5). The across-samples SNV module benefited more from parallelization, running 2.33 times faster in MIDAS2 with 48 CPUs (Figure 1A) and scaling linearly (Supplementary Figure S4).

3.1 Reference Genomes and Post-alignment Filters Determine Metagenotype Accuracy

MIDAS2, inStrain and metaSNV v2 were applied to three aliquots of a standardized bacterial community (Olm et al. 2021), and SNVs were compared between aliquots which should have identical metagenotypes (Supplementary Note). metaSNV v2 has the fewest false positives by only using uniquely aligned reads, but it only genotyped five of the eight species in the community (Supplementary Table S5). InStrain and MIDAS2 correctly detected all eight species. When both are run with a genome database containing only the reference genomes of the strains in the community, MIDAS2 has fewer false positives (Figure 1B). However, the false positive rate of MIDAS2 is higher when using the MIDAS DB v1.2, in which these species' reference genomes are diverged from the sample. Thus, high-quality reference genomes and post-alignment filters that balance false positives against false negatives are crucial for metagenotyping.

3.3 Scalability of MIDAS2 versus metaSNVv2

Since metaSNV v2 was previously shown to be efficient enough to metagenotype thousands of samples (Van Rossum et al. 2021), we assessed the scalability of MIDAS2 compared to metaSNV v2 on 1,097 samples from the PREDICT study (NCBI accession: PRJEB39223), using MIDAS DB UHGG with both tools (Supplementary

Note). Despite the same species selection criteria, MIDAS2 metagenotyped many more species (44 versus 14 for metaSNV v2) (Supplementary Note). MIDAS2 used more memory (21.21 GB versus 4 GB peak RAM utilization) and ran slightly longer (average 106 versus 84 minutes per species) to achieve this. We conclude that MIDAS2 can metagenotype thousands of samples with reasonable computational costs, providing a more sensitive alternative to metaSNV v2.

3.4 MIDAS2 Detects Metagenomes with Multiple Strains

For each of the 44 species from PREDICT with MIDAS2 metagenotypes, we quantified evidence of a single dominant strain versus mixtures of multiple strains with an existing method (Nayfach and Pollard 2015; Garud et al. 2019; Nayfach et al. 2016). While most species showed evidence of distinct lineages across samples (Supplementary Figure S8), single samples often had a single dominant strain (Figure 1C). However, samples with strain mixtures were common for several species, including *Bacteroides_B dorei* (62%) and *Faecalibacterium prausnitzii_G* (49%) (Supplementary Figures S9 and S10).

ACKNOWLEDGMENTS

This work was funded by the Chan Zuckerberg Biohub, Gladstone Institutes, and NSF grant #1563159.

Conflicts of Interest: none declared.

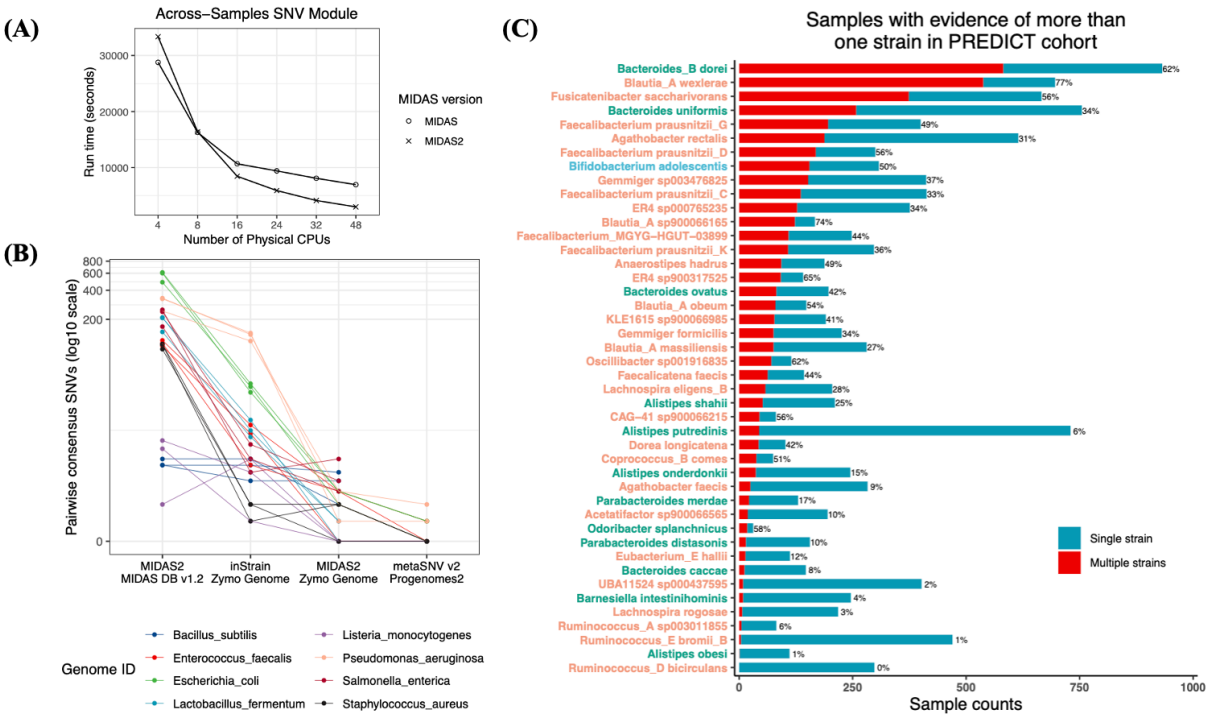


Figure 1. Speed, accuracy, and application of MIDAS2. A: The SNV module of MIDAS2 was re-engineered to parallelize within species, making it increasingly faster than MIDAS as more CPUs are deployed. This analysis was performed with 211 metagenomic samples (NCBI accession: PRJNA400072) and population SNVs. **B:** Benchmarking with three aliquots of a standardized microbial community shows that reference genome selection and post-alignment filters determine metagenotype accuracy. The aliquots are identical, so all SNVs are false positives. This analysis uses consensus SNVs. Population SNVs show the same pattern (Supplemental Figure S4). **C:** The distribution of samples with evidence of a strain mixture versus one dominant strain for 44 species metagenotyped by MIDAS2 in the PREDICT cohort (NCBI accession: PRJEB39223).

Figure 1 Speed, accuracy, and application of MIDAS2

References

Almeida, A. *et al.* (2021). A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology*, *39*(1), 105–114. https://doi.org/10.1038/s41587-020-0603-3

Andreas Heger, K. J. (n.d.). *pysam*. Retrieved June 13, 2022, from https://github.com/pysam-developers/pysam

Beghini, F., McIver. *et al.* (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife*, *10*. https://doi.org/10.7554/eLife.65088

Bush, S. J. *et al.* (2020). Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism-calling pipelines. *GigaScience*, *9*(2). https://doi.org/10.1093/gigascience/giaa007

Clarke, E. L. *et al.* (2019). Sunbeam: an extensible pipeline for analyzing metagenomic sequencing experiments. *Microbiome*, *7*(1), 46. https://doi.org/10.1186/s40168-019-0658-x

Commichaux, S. *et al.* (2021). A critical assessment of gene catalogs for metagenomic analysis. *Bioinformatics* . https://doi.org/10.1093/bioinformatics/btab216

Daillère, R. *et al.* (2016). Enterococcus hirae and Barnesiella intestinihominis Facilitate Cyclophosphamide-Induced Therapeutic Immunomodulatory Effects. *Immunity*, *45*(4), 931–943. https://doi.org/10.1016/j.immuni.2016.09.009

Danecek, P. *et al.* (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

Garud, N. R. *et al.* (2019). Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLoS Biology*, *17*(1), e3000102. https://doi.org/10.1371/journal.pbio.3000102

Jain, C. *et al.* (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications*, *9*(1), 5114. https://doi.org/10.1038/s41467-018-07641-9

Jiang, H. *et al.* (2015). Altered fecal microbiota composition in patients with major depressive disorder. *Brain, Behavior, and Immunity*, *48*, 186–194. https://doi.org/10.1016/j.bbi.2015.03.016

Mende, D. R. *et al.* (2020). proGenomes2: an improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes. *Nucleic Acids Research*, *48*(D1), D621–D625. https://doi.org/10.1093/nar/gkz1002

Nayfach, S., & Pollard, K. S. (2015). Population genetic analyses of metagenomes reveal extensive strain-level variation in prevalent human-associated bacteria. In *bioRxiv* (p. 031757). https://doi.org/10.1101/031757

Nayfach, S., & Pollard, K. S. (2016). Toward Accurate and Quantitative Comparative Metagenomics. *Cell*, *166*(5), 1103–1116. https://doi.org/10.1016/j.cell.2016.08.007

Nayfach, S. *et al.* (2016). An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Research*, *26*(11), 1612–1625. https://doi.org/10.1101/gr.201863.115

Olm, M. R. *et al.* (2021). inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nature Biotechnology*, *39*(6), 727–736. https://doi.org/10.1038/s41587-020-00797-0

Paradis, E. *et al.* (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* , *20*(2), 289–290. https://doi.org/10.1093/bioinformatics/btg412

Parks, D. H. *et al.* (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043–1055. https://doi.org/10.1101/gr.186072.114

Parks, D. H. *et al.* (2022). GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically

consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*, *50*(D1), D785–D794. https://doi.org/10.1093/nar/gkab776

Robinson, J. T. *et al.* (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26. https://doi.org/10.1038/nbt.1754

Rognes, T. *et al.* (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, *4*, e2584. https://doi.org/10.7717/peerj.2584

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* , *30*(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Truong, D. T. *et al.* (2017). Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Research*, *27*(4), 626–638. https://doi.org/10.1101/gr.216242.116

Van Rossum, T. *et al.* (2021). metaSNV v2: detection of SNVs and subspecies in prokaryotic metagenomes. *Bioinformatics* . https://doi.org/10.1093/bioinformatics/btab789

Supplementary Note

MIDAS2: Strain-level Reads-to-Table Metagenotyping Pipeline

MIDAS2 performs strain-level metagenomics analysis through a SNV module and a CNV module. Each module includes two sequential steps: single-sample analysis and across-samples integration of single-sample results to identify population variants (Supplementary Figure S2). The SNV module is based on alignment of metagenomic reads to one representative genome per species (rep-genomes), and the CNV module is based on alignment of reads to pan-genomes; both use an estimate of which species are present in the sample to customize the sample-specific reference database.

Given a collection of shotgun metagenomics samples, both the SNV and CNV module start with identifying the list of species that are sufficiently abundant for metagenotyping, based on profiling 15 universal, single copy genes (SCGs). By default, all species with median SCG coverage > 2X are metagenotyped. If the purpose is to genotype low abundance species, users are responsible for adjusting the cutoff (e.g., median SCGs coverage > 0). Species to be metagenotyped are then used to construct a sample-specific rep-genome database that only contains the representative genomes for these species (SNV module) and/or a pan-genome database that only contains the pan-genomes for these species (CNV module). This customization of a pre-computed MIDAS DB to the sample improves metagenotype accuracy by reducing "cross-mapping", whereby metagenomic sequencing reads align erroneously to conserved regions in the genomes of closely related species.

The SNV module starts with single-sample pileup, by aligning reads to a sample-specific rep-genome database. The alignment results for all the species in the rep-genome database are reported, such as *fraction_covered* *mean_coverage*, and *mapped_reads*. MIDAS2 purposely holds off on any filtering or species selection with the single-sample pileup results until across-samples SNV analyses are performed. Upon the completion of the single-sample SNV analyses for all the samples, MIDAS2 will compute the population SNVs across all the samples (see Population SNV Computation section).

It is worth noting that the MIDAS2 single-sample SNV module is not designed for consensus SNV calling, and by design MIDAS2 does not apply any species or variants calling filters at the single-sample level. However, in

recognition of the need for single-sample consensus SNV calling, we added an *--advanced* mode to the single-sample SNV analysis in MIDAS2 to report per species major allele and minor allele for all the genomic sites covered by at least two reads, upon which any variant calling filter would be applied by the users. We highly recommend users set the parameter *--ignore_ambiguous* to ignore ambiguous alleles, where genomic sites recruit tied read counts.

The CNV module starts with single-sample alignment of reads to the sample-specific pan-genome database. As in the SNV module, this only includes the pan-genomes of species being genotyped (i.e., sufficiently abundant species based on median SCGs coverage). The alignment summaries for all the species in the customized pan-genome database are reported (e.g. *mean_coverage*, *mapped_reads*), and these are used to determine the scope of across-samples CNV analysis. For example, only species that pass vertical pan-gene read depth > 1 (by default) are included in the across-samples CNV computation. For each gene in each species, coverage is normalized by the mean coverage of all the SCGs to estimate copy number per cell (Nayfach and Pollard 2016). A brief comparison of the CNV module with other methods can be found in Supplementary Table S3.

Features Distinguishing Different Reference-based Metagenotyping Pipelines

All read mapping based metagenotyping pipelines share an underlying structure, yet they differ from one another in the availability of genome databases, the customization of reference genomes to the species present in the sample, read aligners, pileup tools, post-alignment filters, and the contents of input and output files (Supplementary Figure S1).

Database. Whole-genomes are used as the reference for most pipelines (MIDAS2, metaSNV v2 (Van Rossum et al. 2021), InStrain (Olm et al. 2021)), while species-specific marker genes are used by StrainPhlan (Truong et al. 2017). MIDAS2 is the only pipeline providing precomputed reference genome databases: namely MIDAS DB UHGG and MIDAS DB GTDB. MIDAS2 builds a database of representative genomes (rep-genomes) customized to only include species that are sufficiently abundant in the sample, which is determined via coverage of 15 SCGs. In contrast, metaSNV v2 and inStrain leave database selection and customization to their users.

Alignment and pileup. MIDAS2 and inStrain provide post-alignment filters based on the read alignment quality (e.g., alignment sequence similarity and uniqueness), while metaSNV v2 leaves this step to the user. MIDAS2 performs read pileup directly from counting reads covering each nucleotide (*count_coverage*), while metaSNV v2 and inStrain both use samtools *mpileup* (Danecek et al. 2021) or pysam (Andreas Heger et. al. 2021) for the pileup task. See Supplementary Table S12 for the resulting differences in variant calling results. One possible reason for differences in metagenotypes is that samtools does not use soft-clipped reads. This may be desirable for certains applications. Yet since all microbiome pipelines apply post-alignment filtering to the raw alignment file before the pileup and variant calls, an unfiltered read pileup seems to be appropriate, so this is what we implemented in MIDAS2.

Genotypes. For single-sample SNV calling, inStrain only reports non-reference variants (when the major allele is different from the reference allele), which can cause missing allele information for downstream across-samples population SNV calling. On the other hand, metaSNV v2 does not report any single-sample pileup results, and only reports the across-samples non-reference variants. Adding samples to an existing analysis, metaSNV v2 therefore has to be run again from the beginning. MIDAS2 reports single-sample pileup results for the all genomic sites covered by at least 2 reads. Therefore, MIDAS2 can avoid repeating single-sample pileup when more samples are added to the analysis. MIDAS2 reports population SNVs of all genomic sites meeting the user-defined types (e.g, *snp_type*).

In summary, MIDAS2 is the only reads-to-table metagenotyping pipeline. By integrating and automating all steps of the metagenotyping process, MIDAS2 helps to promote reproducible research.

Population SNV Computation

Population SNV analysis (across-samples step) restricts attention to "sufficiently well" covered species in "sufficiently many" samples. To be specific, a given <*species, sample*> pair will only be kept if it has more than 40% horizontal genome coverage (*fraction_covered*) and 5X vertical genome coverage (*mean_coverage*). Furthermore, only "sufficiently prevalent" species with "sufficiently many" samples (by default *sample_counts* >

2) are eligible for across-samples population SNV analyses. Therefore, different species may have different lists of relevant samples. For each genomic site, a sample is "relevant" if the corresponding site depth falls between the user-defined range, otherwise it is ignored for the pooled-SNV compute. Therefore, different genomic sites from the same species may have different panels of "relevant samples". And genomic site prevalence can be computed as the ratio of the number of relevant samples for the given site over the total number of relevant samples for the given species.

There are three main steps to compute and report population SNV in MIDAS2. First, for each site, MIDAS2 determines the set of alleles present across all relevant samples. Specifically, for each allele (A, C, G, T), *merge_snps* subcommand (1) tallies the sample counts (*sc*) of relevant samples containing corresponding allele, and (2) sums up the read counts (*rc*) of the corresponding allele across all the relevant samples. Second, population major and minor alleles for a single site can be computed based on the accumulated read counts or sample counts across all relevant samples. The population major allele refers to the most abundant/prevalent allele, and the population minor allele refers to the second most prevalent/abundant allele. Third, MIDAS2 collects and reports the sample-by-site matrix of the corresponding (1) site depth and (2) allele frequency of the population minor allele for all the relevant samples

Chunkified Pileup Implementation

MIDAS parallelized pileup work by species, and because some species required more work than others, this resulted in a few CPU cores working longer while the rest sat idle waiting. So MIDAS2 subdivided the work for those large species into smaller units that we called chunks, by splitting the species' genomes into segments. We size the chunks so they take reasonable time (default *chunk_size* = 1000000). If chunks are too big, we will have CPU cores sitting idle while others finish running. If chunks are too small, the per-chunk setup costs will be high. The chunkified pileup algorithm can sweep through a large data set coherently, ensuring data locality so the active chunks fit into available filesystem cache, and it therefore achieves better utilization.

To compute the per-chunk population SNV, all the pileup results of corresponding sites across all the samples need

to be read into memory. Therefore at any given moment, up to total CPU cores * total number of sites per chunk * total number of the relevant samples pileup results will be read into RAM. Therefore, for a highly prevalent species, MIDAS2 dynamically (*--robust_chunk*) adjusts to a smaller chunk size. There is a trade-off between the chunk size, the running time and the memory usage. Users can customize the chunk size according to their computing environment. For example, constant RAM usage can be achieved despite an increasing number of samples with dynamic chunk sizes. When all chunks from the same species finish processing, chunk-level pileup results will be merged into species-level pileup results. Below is the pseudocode for the chunkified population SNV implementation.

```
def process(species_id, chunk_id):
    if chunk_id == -1:
        collect_chunks()
    chunk_worker()
def chunk_worker():
    for sample in list_of_relevant_samples:
        accumulate()
    call_population_snps()
    write_population_snps()
multiprocessing_map(process, list_of_species, num_cores)
```

This implementation makes reference-based population SNV analysis across thousands of samples possible.

Consensus and Population SNVs for a Standardized Microbial Community Reveal that Reference Genome Selection and Post-alignment Filters Determine Metagenotyping Accuracy

We compared the consensus sequences reported by three metagenotyping pipelines using three metagenomes from aliquots of the same ZymoBIOMICS Microbial Community Standard product (catalog no. D6300, BioProject PRJNA648136). No genetic differences were expected between these samples. For each species, SNVs were called by comparing all pairs of metagenomes; and we computed the pairwise consensus SNVs and population SNVs to assess the false positive rates of each method as in (Olm et al. 2021).

Analysis. Three metagenotyping pipelines that use alignment to reference databases were compared: MIDAS2,

inStrain, and metaSNV v2. We ran inStrain with the Zymo reference genomes as in (Olm et al. 2021), metaSNV v2 with proGenomes 2 as in (Mende et al. 2020), and MIDAS2 with MIDAS DB v1.2 for the purpose of reproducibility with results in (Olm et al. 2021). For the post-alignment filters, we used the default or recommended filter parameters by each tool (Supplementary Table S4). To evaluate the role of a well-matched reference database, we also ran MIDAS2 with the Zymo reference genomes, while keeping everything else unchanged. This allowed for a direct comparison of the post-alignment filters deployed in inStrain and MIDAS2, as well as a direct assessment of the reference database between MIDAS DB v1.2 and Zymo reference genomes. All the tools report horizontal genome coverage, upon which species presence is called. The same set of variant calling filters as inStrain were applied to all the tools. Last, pairwise consensus-based SNVs (consensus SNVs) and population SNVs (Olm et al. 2021) were computed. Simply put, consensus SNVs refer to two major alleles that are different between a pair of samples, whereas population SNVs refer to sites that differ (neither major or minor allele match) between a pair of samples.

Database. For the eight bacterial Zymo reference genomes, we found the corresponding species in MIDAS DB v1.2 and proGenomes2 based on the highest Average Nucleotide Identity (ANI) (Jain et al. 2018). It is worth mentioning that for Lactobacillus fermentum and Salmonella enterica, there are two pairs of highly closely related strains in the proGenomes2: 98.8% and 98.5% (Supplementary Table S6). We also reported the genome quality of all the representative genomes in all three databases using checkM (Parks et al. 2015) (Supplementary Table S7).

Results. The three aliquot samples originated from the same microbial population, therefore any observed SNVs could be a product of (1) read misalignment: some regions of the reference genome erroneously recruits reads originating from other genome regions or even other species; (2) failure of the post-alignment filters to recognize the misalignments as such; or (3) microdiversity created in the lab.

MIDAS2 reported more accurate consensus sequences compared to inStrain when both tools were run using the Zymo reference genomes matched to the strains in the metagenomes (Figure 1B). However, the accuracy of MIDAS2 was much inferior when the Zymo reference database was replaced by MIDAS DB v1.2, in which the representative genomes are less similar to the community and also suffer from fragmented assemblies. Population

SVNs followed the same pattern (Supplementary Figure S6). This shows that the reference genomes are the major determinant of accuracy for consensus SNVs, explaining a previous report of poor performance for MIDAS (Olm et al. 2021). As suggested by Olm et al., a good representative genome has the following characteristics: high quality contiguous sequence and a high degree of shared gene content with the taxa it is meant to represent.

On the other hand, with its stringent post-alignment filters (e.g., recruiting only uniquely aligned reads), metaSNV v2 only reported metagenotypes for five of the eight species in the mock community. Three species did not pass its default 40% horizontal coverage threshold (Supplementary Table S8). A few reasons can explain this failure of metaSNV v2. First, there is not an appropriate representative genome of the species *Bacillus subtilis* in the proGenomes 2; the closest genome has average nucleotide identity (ANI) = 92.54% (txid 224308). Second, there are two very closely related genomes in the proGenome 2 database for the other two species: *Lactobacillus fermentum* (ANI 98.83%) and *Salmonella enterica* (ANI 98.53%), which led to extremely low numbers of uniquely alignable reads for these two species. Since metaSNV v2 requires uniquely mapped reads, these two species have low horizontal genome coverage (0.04% and 30% of the whole genome, respectively) and hence many fewer sites metagenotyped. Third, metaSNV v2 only reports non-reference variants in the output. Therefore significantly fewer genomic sites were metagenotyped and further compared between a pair of samples by metaSNV v2. As a result, metaSNV v2 ran faster than the other two tools and reported fewer false positive SNVs (Supplementary Table S5). For the 1097 samples from the PREDICT study (PRJEB39223), the analysis-ready BAM file of MIDAS2 is 1.53 times larger than metaSNV v2, again due to the different post-alignment filters. Thus, metaSNV v2 generates few false positive SNVs and runs efficiently, but it has lower sensitivity than MIDAS2 and inStrain.

To summarize, we demonstrated building a high quality genome database and utilizing post-alignment filters that balance false positive versus false negative SNVs determine metagenotype accuracy.

MIDAS2 Reference Database Target Layout and Construction

MIDAS Reference Database (MIDAS DB) refers to a set of custom files needed to run MIDAS2. There are three

components in a MIDAS DB: rep-genome database, SCG marker database, and pan-genome database (Supplemental Figure S3). In each part, data is organized by species. For a given species, the rep-genome database contains all representative genomes of that species annotated using Prokka (Seemann 2014), and the SCG marker database is built on identified homologs of 15 universal SCGs from the representative genomes. The pan-genome database refers to the set of non-redundant genes within all genomes from that species (Commichaux et al. 2021), clustered at 99% sequence identity using vsearch (Rognes et al. 2016).

The original release of MIDAS provided a default bacterial reference database (MIDAS DB v1.2), constructed from a collection of 5952 bacterial species clusters representing 31,007 high-quality bacterial genomes. With the rapid growth of the number of sequenced microbial genomes, particularly with the addition of metagenome-assembled genomes (MAGs) from diverse habitats, it is necessary to update the MIDAS DB.

For MIDAS2, we took advantage of several published collections of prokaryotic genome collections (UHGG (Almeida et al. 2021), GTDB (Parks et al. 2022)) in which genomes were already clustered into species groups. We developed a new database infrastructure that is geared to run on AWS Batch and S3, achieving elastic scaling for database construction for each of these large genome collections. Specifically, MIDAS DB construction can be executed in AWS using hundreds of instances, depositing built products in S3. For example, the species pan-genome for all 47,894 species of GTDB r202 was built in roughly one week at a cost of $80K, using 100 r5d.24xlarge instances. The new database infrastructure reads in a table-of-contents (TOC) file containing genome-to-species assignments for all the genomes and a choice of representative genome for each species cluster. Six-digit numeric species ids are randomly assigned and stored in the corresponding metadata file (*metadata.tsv*). MIDAS2 users can also build a new MIDAS DB locally for a small collection of representative genomes of interest.

UHGG Separable/Inseparable Species

Many species have a closely related species with pairwise average nucleotide identity (ANI) near the species boundary (e.g., 95% ANI). For example, in the UHGG v1 database (Almeida et al. 2021), there are 981 species

with closest pairwise ANI higher than 92%. We explored how effective SCGs are in separating the species in the UHGG database. For a given SCG in a given species, we compute all 31-mers and compare them to 31-mers present in all species. We define the marker as separable if it has >100 unique 31-mers. We then define a species as separable if more than 50% of its 15 SCGs are separable. We found that there are 3,956 separable species and 689 inseparable species (Supplementary Tables S13 and S14). The majority (76%) of inseparable species are from the *Collinsella* genus. The representative genomes of the 3,956 separable species were used as the reference database for the PREDICT analysis in this study.

Bioinformatics Processing of Publicly Downloaded Samples

Metagenomic sequencing reads from an inflammatory bowel disease cohort (NCBI accession: PRJNA400072) were pre-processed using Sunbeam (Clarke et al. 2019), and 211 samples with more than 5 million reads were used for benchmarking. The benchmark work was done on a m5.16xlarge or m5.24xlarge EC2 instance, depending on the CPU counts needed.

Metagenomic sequencing reads from the PREDICT cohort (NCBI accession: PRJEB39223) were pre-processed using Sunbeam (Clarke et al. 2019), and 1,097 samples with more than 5 million reads were used in our analysis. The benchmark of MIDAS2 and metaSNV v2 was done on a r5-24xlarge instance.

The bioinformatics details of the ZymoBIOMICS benchmarking experiment are shown in Supplementary Table S4. The parameters used in the PREDICT analysis are shown in Supplementary Table S10.

Parameters for paired-end read alignment in MIDAS2

During Bowtie2 read alignment, the default maximum fragment length is 500 base pairs (bp). With this setting, any read pairs sequenced from DNA fragments longer than 500 bp are labeled as not properly aligned (*is_proper_pair* = FALSE), and the reported template length may also be incorrect (equal to zero) (Supplementary Figure S12). Therefore, to properly use the paired-end option of the single-sample SNV module, it is crucial to set the -*X* (maximum fragment length) to an appropriate value for the metagenomes being analyzed (default =5,000

bp), even if longer maximum fragment length will make the alignment slower.

Quasiphasable Species Model

We applied the model from Garud, Good et al. (Garud et al. 2019) to the PREDICT study to determine whether each species in each metagenomic sample was quasi-phasable (QP) or not (i.e., one dominant strain versus colonization by multiple bacterial lineages). The model uses synonymous sites in genes of the core genome of a given species. If the fraction of these sites with intermediate allele frequencies is high, this is taken as evidence for a strain mixture.

Bioinformatics. We use similar sample and site filters for population SNVs as in (Garud et al. 2019). Specifically, we filter the per species *snps_info.tsv* and *snps_freqs.tsv* files as follows:

- Minimal per sample median site depth of bi-allelic SNVs from protein coding sequences ($\bar{D}$) is 20
- Only include 4-fold degenerate synonymous sites
- Sample site depths must be between .3 * $\bar{D}$ and 3 * $\bar{D}$
- Minimal site depth is 10
- Minimal site prevalence is 5

This produces a filtered SNVs allele frequency file.

QP calculation. For each species, we estimated whether each sample is QP as follows:

1) For each non-intermediate site, we defined population major allele direction ($\bar{f}_{l_{dir}}$) as 1 if the majority of the allele frequencies are higher than 0.8 and 0 if not.

$$\bar{f}_{l_{dir}} = 1 \text{ if } (\sum_{n}^{N} f_n \geq 0.8 \geq \sum_{n}^{N} f_n \leq 0.2) \text{ else } \bar{f}_{l_{dir}} = 0$$

where n is the sample index, N is the number of relevant samples for the given site.

2) For each sample n, we computed the dominant haplotype of each non-intermediate site ($f_{n,l_{dir}}$) as 1 if the corresponding allele frequency is higher than 0.8 and 0 if not

$$f_{n,l_{dir}} = 1 \text{ if } f_{n,l} \geq 0.8 \text{ else } f_{n,l_{dir}} = 0$$

where l is the site index.

3) For each non-intermediate site, the population major allele frequency ($\bar{f}_l$) is computed as:

$$\bar{f}_l = \frac{\sum_n^N f_{n,l_{dir}} == \bar{f}_{l_{dir}}}{N}$$

4) For each sample, we estimated N_D as the average genetic distance between sample *n* and the alleles present in the remainder of the samples ($\bar{f}_{l_{dir}}$) as:

$$N_D = \sum_l^L \bar{f}_l \text{ if } f_{n,l_{dir}} == \bar{f}_{l_{dir}} \text{ else } 1 - \bar{f}_l$$

where, l is site index, L is total number of non-intermediate sites; n is the sample index.

5) For each sample, we computed the number of intermediate alleles over all sites ($N_<$) :

$$N_< = \sum_l^L 1 \text{ if } 0.2 < f_{n,l} < 0.8 \text{ else } 0$$

6) For each sample, if $\frac{N_<}{N_D} < 0.1$ then we say the sample is QP, because there is evidence of one dominant strain with sufficiently high coverage and sufficiently low rates of intermediate alleles.

PCoA and Manhattan distance

Manhattan distance was calculated based on the filtered SNV site-by-sample allele frequency matrix to evaluate the dissimilarity between samples. Principal Coordinate Analysis (PcoA) was calculated based on the Manhattan distance matrix using the ape package (Paradis et al. 2004).

Supplementary Figures

Fig S1: MIDAS2 is the only reads-to-table whole-genome based metagenotyping pipeline.

Fig S2: MIDAS2 performs strain-level metagenomics analysis through two modules.

Fig S3: MIDAS2 Reference Database (MIDAS DB) target layout and major construction steps.

Fig S4: MIDAS2 achieved better CPU utilization compared to MIDAS.

Fig S5: Database building and alignment on average take 75 % of run time in the MIDAS2 single-sample SNV module.

Fig S6: ZymoBIOMICS benchmark shows reference genome selection and post-alignment filters determine metagenotyping accuracy.

Fig S7: Compute performance comparison between MIDAS and MIDAS2 for single-sample CNV module.

Fig S8: PCoA plot of two species with genetically distinct lineages across samples, yet a single dominant strain within most samples.

Fig S9: Many samples are colonized with more than one strain of Bacteroides_B doeri.

Fig S10: Many samples are colonized with more than one strain of Faecalibacterium prausnitzii_G.

Fig S11: Importance of setting proper -*X* for paired-ends reads alignment

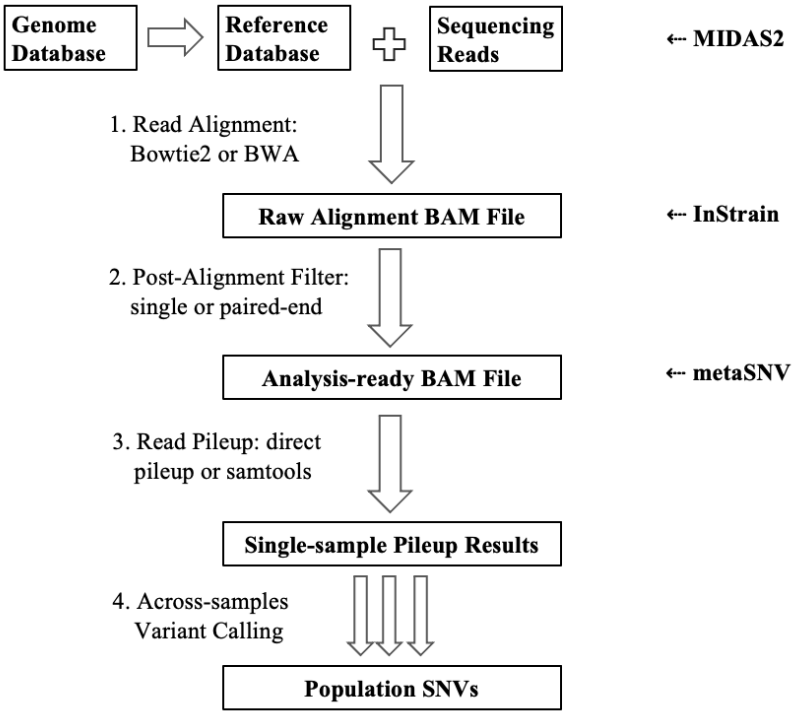


Fig S1. **MIDAS2 is the only reads-to-table whole-genome based metagenotyping pipeline**. Reference-based metagenotyping pipelines share an underlying structure, yet they differ from one another in whether or not they include code for generating and customizing the genome database to which metagenomic reads are aligned, the choice of reference genomes in the database, the reads alignment algorithm, pileup tools, post-alignment filters, as well as the input and output file contents. MIDAS2 starts from metagenomic sequencing reads and then proceeds through alignment, post-alignment filtering, pileup, and variant calling. It includes pre-computed reference databases, code for building a database, and code for customizing the database to the species present in the metagenome. The other tools leave database and alignment to the user. InStrain (Olm et al. 2021) starts from a BAM file and performs post-alignment filtering as a first step, whereas metaSNV v2 (Van Rossum et al. 2021) starts from a BAM file that has already been filtered. By integrating and automating all steps of the metagenotyping process, MIDAS2 helps to promote reproducible research.

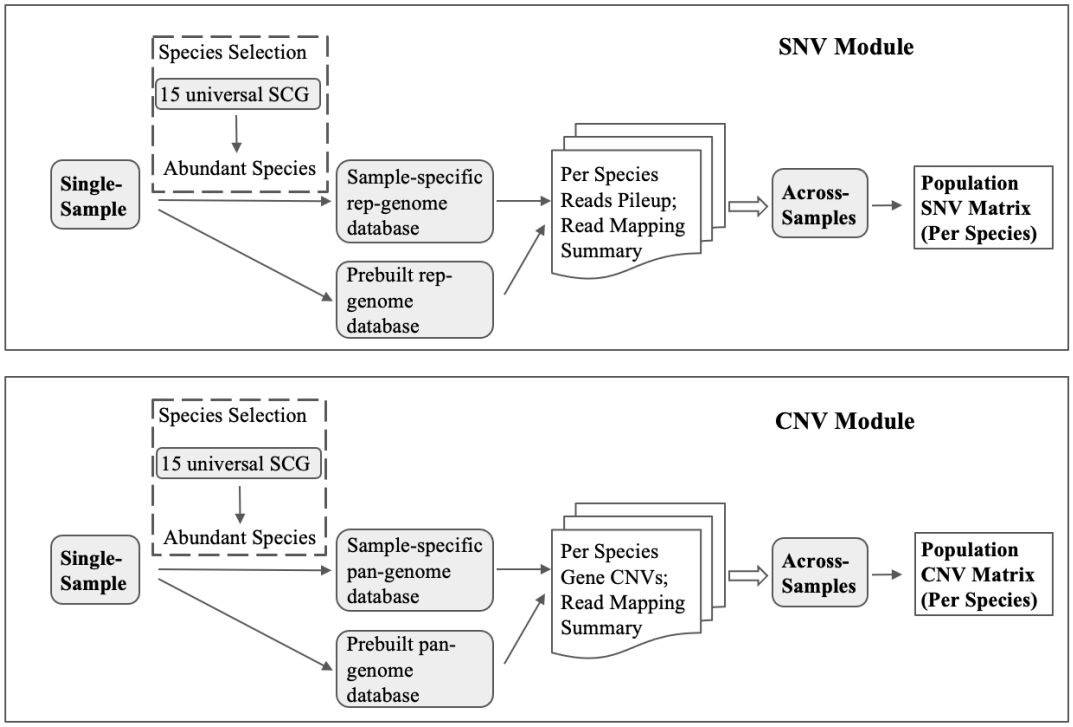


Fig S2. **MIDAS2 performs strain-level metagenomics analysis through two modules:** population single nucleotide variants analysis (SNV module) and pan-genome gene copy number variants analysis (CNV module). Each module includes two sequential steps: single-sample analysis and across-samples analysis. The single-sample step for both modules uses an estimate of which species are present in the sample to customize the sample-specific bowtie2 database. Across-sample analysis reports both the read alignment and genotyping summary, and multiple per-species sample-by-site SNV or sample-by-gene CNV matrices.

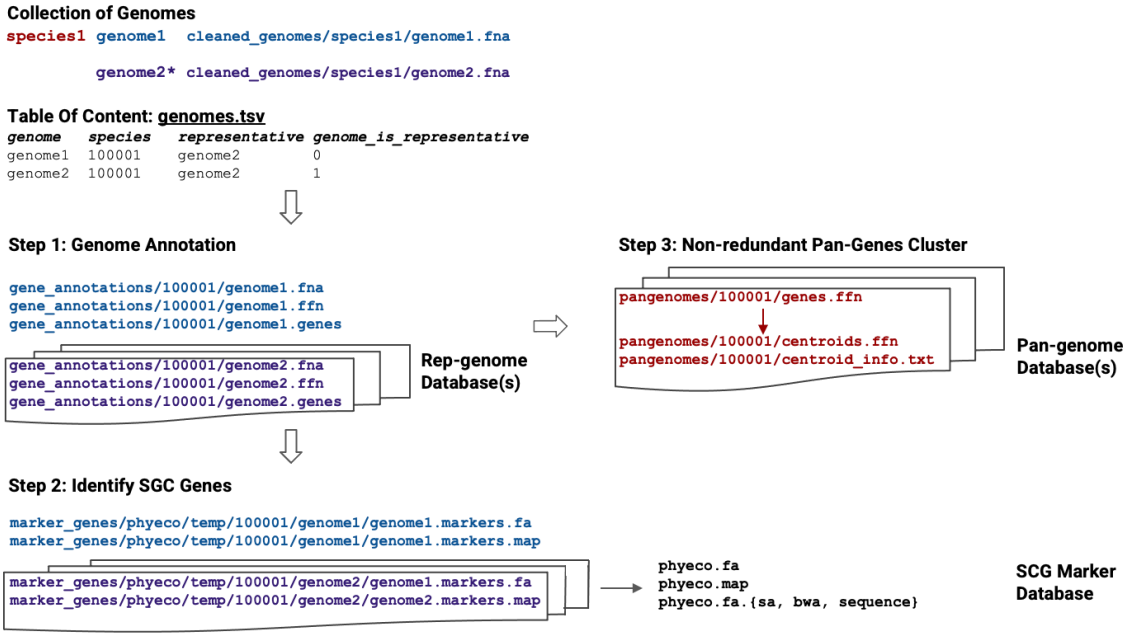


Fig S3. **MIDAS2 Reference Database (MIDAS DB) target layout and major construction steps.** There are three components in a MIDAS DB: rep-genome database, Single Copy Gene (SCG) marker database, and pan-genome database.

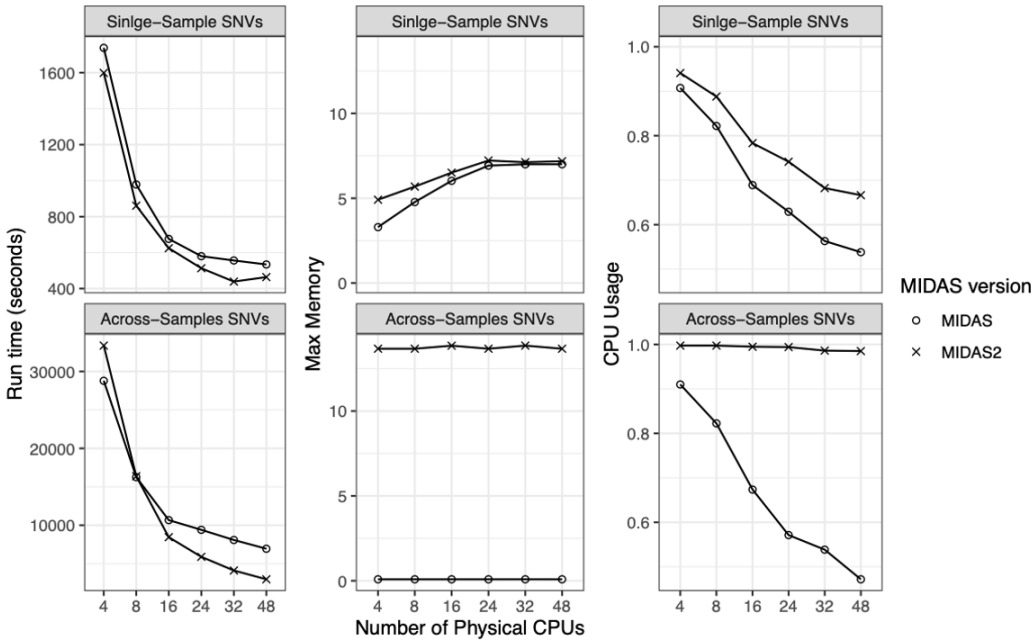


Fig S4. MIDAS2 achieves better CPU utilization compared to MIDAS. The single-sample SNV module of MIDAS2 is on average up to 1.27 times faster than MIDAS (number of CPUs = 32). This improvement is limited because 75% of the single-sample SNV analysis is spent on building sample-specific genome database, read alignment and post-alignment filtering (Supplementary Figure S6), which are outside the chunk-parallelization we implemented. The across-samples SNV module benefits more from chunk-parallelization. It is up to 2.33 times faster than MIDAS (number of CPUs = 48). CPU usage is defined as CPU percentage / number of cores. The chunk-parallelization leads to the across-samples module scaling almost linearly with the number of physical cores (tested up to 48 cores).

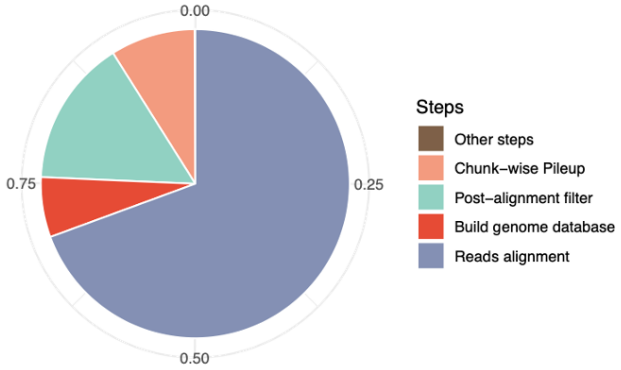


Fig S5. Database building and alignment on average take 75 % of run time in the MIDAS2 single-sample SNV module. We analyzed the distribution single-sample SNV module run times for 211 metagenomic samples (PRJNA400072) with 8 physical CPUs and using paired-end post-alignment filtering. Single-sample pileup is a multi-step task: build sample-specific genome database, perform read alignment, apply post-alignment filter, and do pileup.

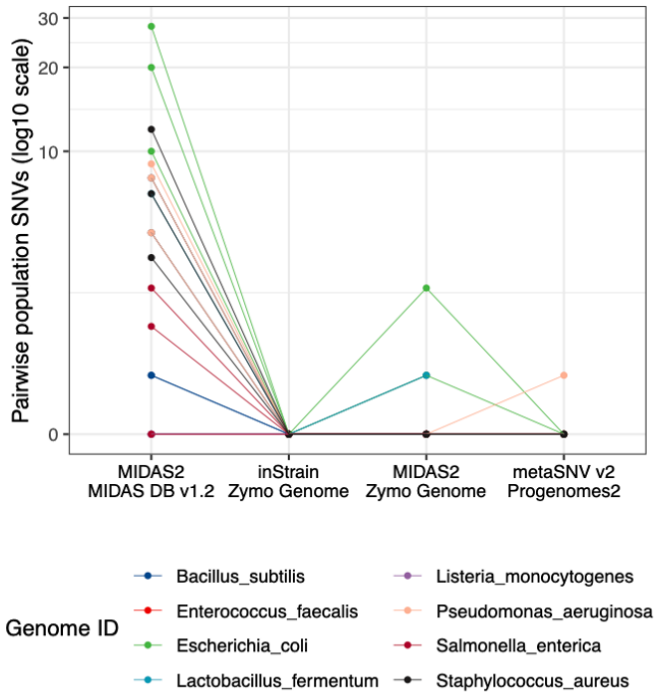


Fig S6. **ZymoBIOMICS benchmark shows reference genome selection and post-alignment filters determine metagenotype accuracy**. MIDAS2 with Zymo reference genome reported more accurate consensus SNVs compared to inStrain, yet the accuracy of MIDAS2 was inferior when the Zymo genomes were replaced by a more diverse database MIDAS DB v1.2 (genomes for many more species, genomes of the eight on-target species diverged from those in the metagenome). Fewer genomic sites were metagenotyped by metaSNV v2 due to its strict alignment uniqueness filters, resulting in the lowest false positive rates. However, only five out of the eight species were metagenotyped by metaSNV v2, compared to all eight for inStrain and MIDAS2 (Supplemental Table S9). Across methods, population SNVs were more accurate than consensus SNVs. These results show that metagenotype precision and recall depends upon building a good reference genome database and utilizing appropriate post-alignment filters. More details see Supplemental Notes.

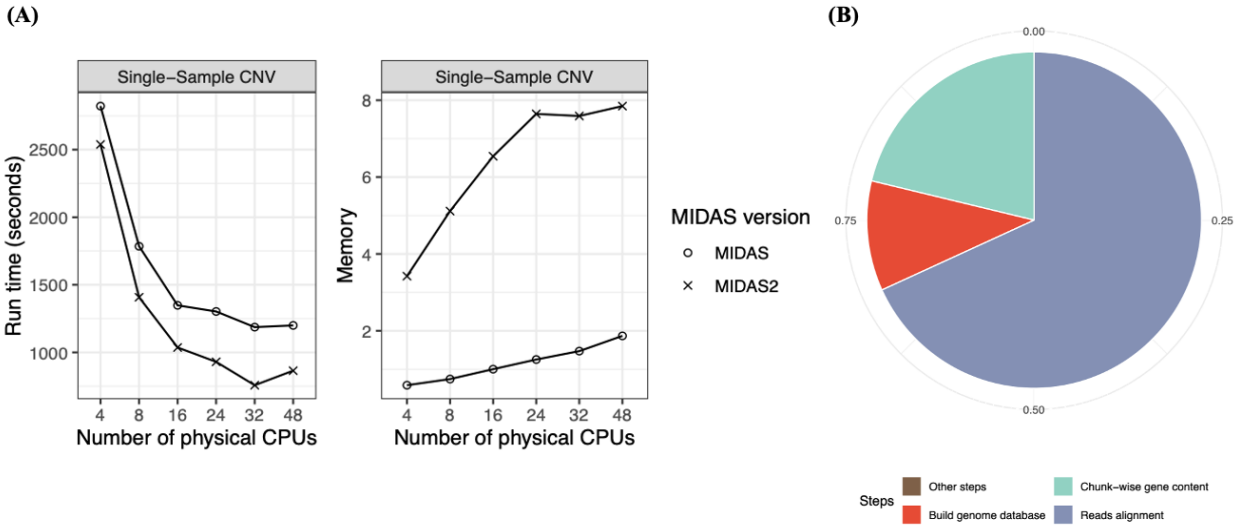


Fig S7. Compute performance comparison between MIDAS and MIDAS2 for single-sample CNV module. This benchmark was performed using 211 samples (PRJNA400072). **A:** The single-sample CNV module of MIDAS2 is on average up to 1.57 times faster than that of MIDAS (number of cores = 32). **B:** This minor improvement is because 80% of the single-sample CNV module is spent on building a sample-specific genome database plus read alignment.

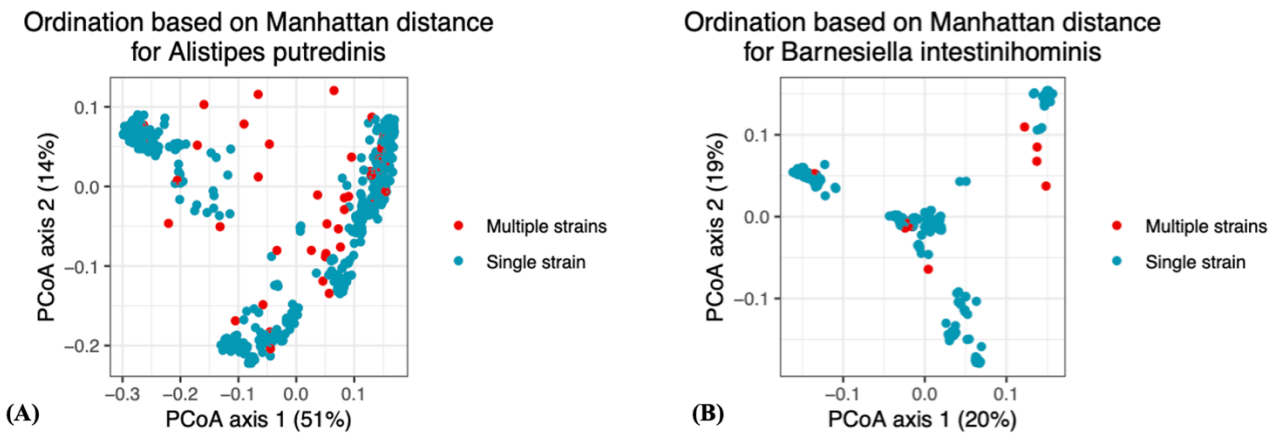


Fig S8. PCoA plot of two species with genetically distinct lineages across samples, yet a single dominant strain within most samples. PCoA was performed based on the pairwise Manhattan distance between samples computed from the population SNV minor allele frequency matrix of one single species. Each dot in the PCoA plot represents one sample, and the distance between a pair of dots represents the genetic similarity of that species in the two metagenomes. **A:** *Alistipes putredinis,* a species which is overly abundant in depression (Jiang, H. et al. 2015) patients. There appear to be two or three distinct clusters of samples, each comprised primarily of samples with a single strain. **B:** *Barnesiella intestinihominis* (Daillère, R. et al. 2016), a species associated with anti-cancer effects. There are four distinct clusters of samples.

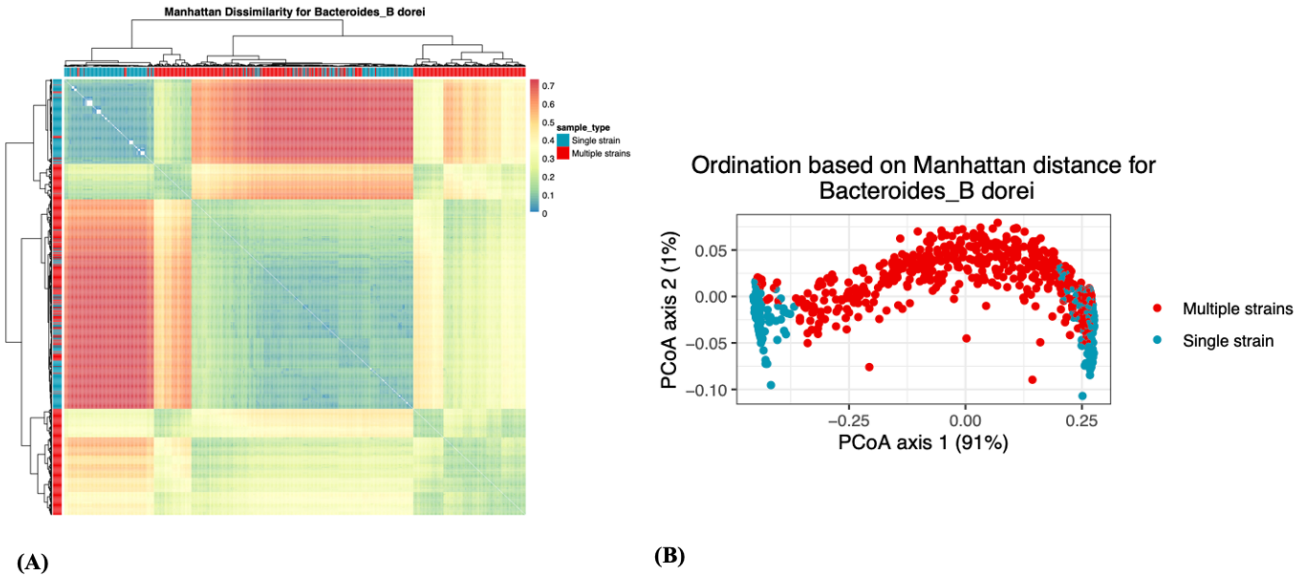
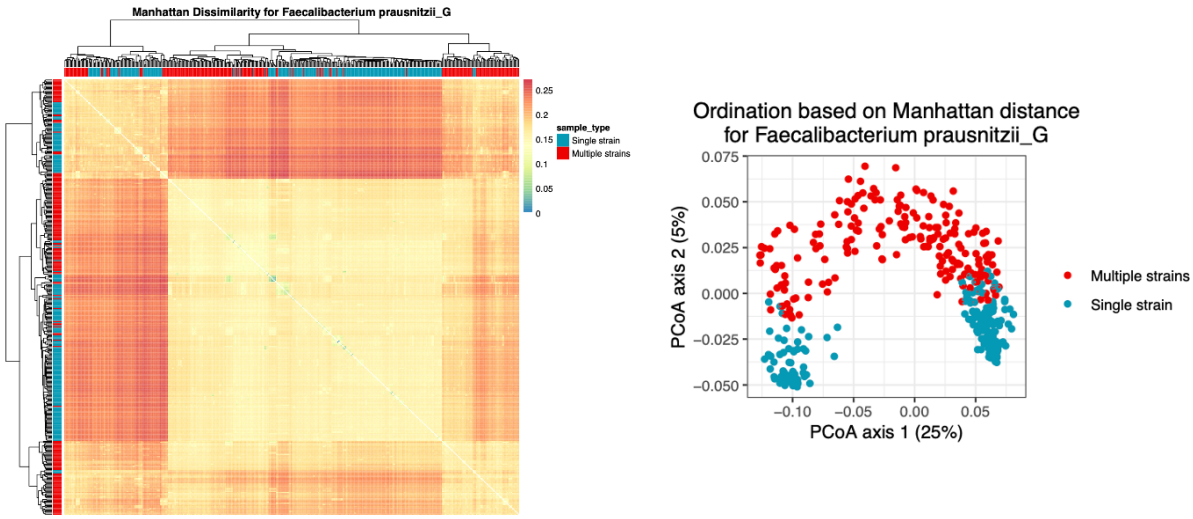
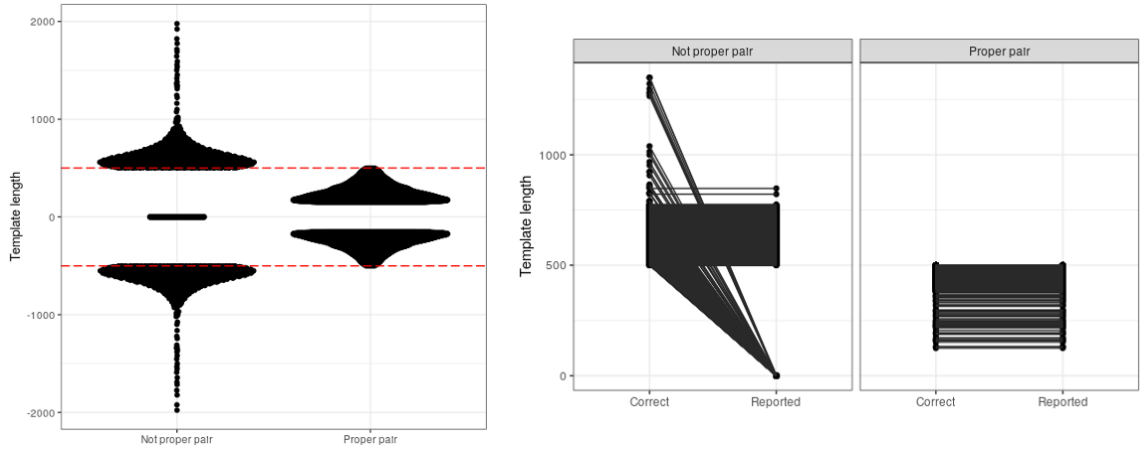


Fig S9. Many samples contain more than one strain of *Bacteroides_B dorei*. 62% of the 932 PREDICT samples with *B. dorei* have allele frequency spectra consistent with two or more distinct lineages colonizing the host. Samples with one dominant strain of *B. dorei* form two genetically distinct clusters, and the rest of the samples are intermediate between these, consistent with colonization by strains from both clusters. **A:** Heatmap of pairwise Manhattan distances between PREDICT samples. **B:** PCoA plot based on the Manhattan distances.



(A) **(B)**

Fig S10. Many samples contain more than one strain of *Faecalibacterium prausnitzii_G*. 49% of the 401 PREDICT samples with *F. prausnitzii_G* have allele frequency spectra consistent with two or more distinct lineages colonizing the host. Samples with one dominant strain of *F. prauznitzii_G* form two genetically distinct clusters, and the rest of the samples are intermediate between these, consistent with colonization by strains from both clusters. **A:** Heatmap of pairwise Manhattan distances between PREDICT samples. **B:** PCoA plot based on the Manhattan distances.



(A) **(B)**

Fig S11. **Importance of setting proper *-X* for paired-end read alignment**. Simulated paired-end reads with average fragment length 600 bps were aligned to the representative genome with default maximum fragment length of Bowtie2 (500 bps). **A:** We observed that read pairs sequenced from DNA fragments longer than 500 bps are often labeled as "Not proper pair" (*is_proper_pair* = FALSE). **B:** We also observed that the template length reported by Bowtie2 for read pairs originating from longer fragments could be falsely reported as 0. Therefore it is crucial to set a long enough *-X* to ensure reads from longer fragments will be properly aligned, even though it will make alignment slower.

Supplementary Tables

Table S1: MIDAS2 software updates

Table S2: Comparison of prokaryote species pan-genome databases.

Table S3: Comparison of gene profiling tools for metagenomes.

Table S4: ZymoBIOMICS benchmarking experiment.

Table S5: Consensus SNVs and population SNVs results for the ZymoBIOMICS benchmark.

Table S6: Corresponding species in MIDAS DB v1.2 and proGenomes 2 for 8 bacterial ZymoBIOMICS genomes.

Table S7: Genome quality of Zymo genomes and corresponding representative genomes in MIDAS DB v1.2 and proGenomes 2.

Table S8: Comparison of average horizontal genome coverage in a single sample in Zymo benchmark experiment.

Table S9: Comparison of average fraction of bases compared in a pair of samples in the Zymo benchmark experiment.

Table S10: Bioinformatics Details of experimental design for PREDICT analysis.

Table S11: Comparison of compute resources for PREDICT benchmarking using a r5-24xlarge instance.

Table S12: Discrepancy of variant calling results with samtools and MIDAS2 at one genomic site.

Table S13: 3956 separable UHGG species list.

Table S14: 689 inseparable UHGG species list.