SOFTWARE

Open Access

MetaPro: a scalable and reproducible data processing and analysis pipeline for metatranscriptomic investigation of microbial communities

Billy Taj[1†], Mobolaji Adeolu[1†], Xuejian Xiong[1], Jordan Ang[2], Nirvana Nursimulu[1,3] and John Parkinson[1,4,5*]

Abstract

Background Whole microbiome RNASeq (metatranscriptomics) has emerged as a powerful technology to functionally interrogate microbial communities. A key challenge is how best to process, analyze, and interpret these complex datasets. In a typical application, a single metatranscriptomic dataset may comprise from tens to hundreds of millions of sequence reads. These reads must first be processed and filtered for low quality and potential contaminants, before being annotated with taxonomic and functional labels and subsequently collated to generate global bacterial gene expression profiles.

Results Here, we present MetaPro, a flexible, massively scalable metatranscriptomic data analysis pipeline that is cross-platform compatible through its implementation within a Docker framework. MetaPro starts with raw sequence read input (single-end or paired-end reads) and processes them through a tiered series of filtering, assembly, and annotation steps. In addition to yielding a final list of bacterial genes and their relative expression, MetaPro delivers a taxonomic breakdown based on the consensus of complementary prediction algorithms, together with a focused breakdown of enzymes, readily visualized through the Cytoscape network visualization tool. We benchmark the performance of MetaPro against two current state-of-the-art pipelines and demonstrate improved performance and functionality.

Conclusions MetaPro represents an effective integrated solution for the processing and analysis of metatranscriptomic datasets. Its modular architecture allows new algorithms to be deployed as they are developed, ensuring its longevity. To aid user uptake of the pipeline, MetaPro, together with an established tutorial that has been developed for educational purposes, is made freely available at https://github.com/ParkinsonLab/MetaPro. The software is freely available under the GNU general public license v3.

Keywords Metatranscriptomics, Microbiome function, Taxonomic annotation, Sequence analysis pipeline

[†] Billy Taj and Mobolaji Adeolu contributed equally to the work and should be regarded as co-first authors.

*Correspondence:
John Parkinson
john.parkinson@utoronto.ca
Full list of author information is available at the end of the article

Introduction

Innovations in culture-independent microbiology, coupled with advances in high-throughput DNA sequencing, have profoundly transformed our understanding of the relationships between microbial communities and their environments [1–3]. In the context of human health, it is increasingly apparent that the composition of the intestinal microbiome has a significant impact on many diseases including type I diabetes, inflammatory bowel disease (IBD), obesity, and rheumatoid arthritis [4–10]. Due to technological and financial constraints, microbiome studies have historically relied on marker gene surveys (e.g., 16S rDNA sequences); a technology that focuses on community composition but provides limited insights regarding functional capacity [11, 12]. More recently, attention has been turning to the use of whole microbiome DNA and RNA sequencing (metagenomics and metatranscriptomics), which yield more meaningful mechanistic insights through broad analysis of microbiome gene content and gene expression [13–19]. These novel methods of analysis are enabled by the next-generation sequencing (NGS) platforms such as Illumina's HiSeq and NovaSeq platforms, capable of generating the millions of sequence reads required to report on the thousands of genes encoded and expressed by microbial communities [20]. A significant challenge is how best to process and interpret these rich datasets that can comprise upwards of hundreds of millions of sequence reads per sample.

While a need for fast and effective tools to automatically process metagenomic and metatranscriptomic pipelines has been identified, few tools are available, particularly for metatranscriptomic data. Web-based analysis platforms such as MG-RAST [21] and COMAN [22] have limited support for metatranscriptomic analyses, though the scope and customizability of possible analysis is narrow and the scale of analysis is limited by the availability of remote compute resources. Existing locally hosted metatranscriptomic pipelines such as SAMSA2 [23], IMP [24], and MetaTrans [25] offer significantly more options than the web-based analysis pipelines but are insufficiently parallelized, limiting their ability to scale to large (e.g., 100+GB) datasets. Further, since they require modifications of experimental protocols or intimate knowledge of computer operating systems to install and execute, they are less amenable to the non-expert. The HMP Unified Metabolic Analysis Network (HUMAnN3) is a fast and scalable platform that was primarily designed to analyse metagenomic datasets [26]. Its extension to analyze metatranscriptomic data comes with the expectation that paired (i.e., from the same sample) metagenomic data is available, a potential constraint due to sequencing costs.

Here, we present MetaPro, a flexible, portable, massively scalable end-to-end analysis pipeline for processing metatranscriptomic data. MetaPro is designed specifically to be easy to deploy and use. It is developed in Python3, and both the pipeline and associated tools are encapsulated in a Docker image, allowing for single-step installation and deployment on both local computers and scientific computing clusters [27, 28]. MetaPro also supports an auto-resume feature for subsequent runs of the same data through the pipeline or if the user wishes to re-run a specific stage of the pipeline. MetaPro is designed with the assumption that new bioinformatics tools will be created that will outperform existing tools currently utilized by MetaPro. Thus, to keep MetaPro relevant, the software architecture enables users to swap, remove, and insert new tools where applicable. To demonstrate improved performance and functionality of MetaPro, we benchmark the speed, resource utilization, and annotation capabilities of MetaPro and two state-of-the-art pipelines, HUMAnN3 and SAMSA2, against three complementary metatranscriptomic datasets. To promote user uptake, both for the application of metatranscriptomics to microbial communities, as well as the tool itself, MetaPro features a tutorial mode that takes the user through each step of the processing pipeline for educational purposes and to encourage adoption beyond the computer specialist.

Implementation

MetaPro—a flexible and scalable metatranscriptomics analysis pipeline

MetaPro was developed as a robust pipeline for the reliable analysis of metatranscriptomic datasets. Key design features include the flexibility to incorporate improved tools as they become available, a scalable architecture to facilitate analysis of hundreds of millions of sequence reads, and ease of use such that end-users are able to install MetaPro as a single software package. To enable these features, the MetaPro pipeline utilizes a modular architecture in which different tools are used at different stages of processing with standard inputs and outputs (Fig. 1A).

MetaPro accepts demultiplexed FASTQ-formatted sequences as the initial input file for analysis. Similar to other pipelines, MetaPro is primarily designed for processing the relatively short-read sequences (e.g., 100–150 bp) generated by Illumina platforms that allow for cost-effective profiling of gene expression in complex microbial communities. MetaPro accepts either single or paired-end sequence reads, and after trimming and filtering for adaptors, paired-end reads are merged and low-quality sequences, as well as duplicate reads are removed. Next, MetaPro filters for known host genomes, vectors, and both rRNA and tRNA sequences. For paired-end datasets, at any time one read of a pair (either forward or reverse) is filtered during adaptor or quality trimming, its matching read (assuming it passes the filter) is placed into a group of "singletons." For host, vector, and rRNA filtering, if either of a read pair matches a host, vector, or rRNA sequence, then by default its paired read would be assigned as host/vector/rRNA accordingly. This results in a set of reads of putative mRNA origin. Next, the remaining mRNA reads have their duplicates repopulated. Previous analyses have demonstrated improved annotation efficiency through assembling reads into longer contigs [35]. Here, we apply rnaSPAdes [36], a transcriptomic assembler within the SPAdes toolkit. Since contigs may contain multiple genes (i.e., as might occur with a polycistronic transcript), we subsequently apply MetaGeneMark [37] to split each contig into discrete "genes."

After the contig assembly step, two categories of sequences have been generated: (1) a set of discrete "genes" derived from contigs assembled from filtered paired-end and singleton reads and (2) a set of unassembled "singleton" reads representing reads (including merged and unmerged paired-end reads, as well as orphans that can arise during filtering of paired-end datasets). These putative mRNA sequences subsequently undergo three separate multistage processes to annotate: (1) gene identities, (2) taxonomic origin, and (3) enzymatic function. For gene annotations, MetaPro applies a tiered set of sequence similarity searches, starting with the fastest and least sensitive, BWA [16, 38], followed by pBLAT [39], then DIAMOND [40]. The former two tools rely on a non-redundant database of genome sequences, ChocoPhlAn [26], while DIAMOND utilizes the NCBI non-redundant (NR) protein database [41]. For enzyme annotation, MetaPro relies on an ensemble approach involving DETECT [42], PRIAM [43], and DIAMOND [40] searches against the UniProtKB/Swiss-Prot database [44]. Due to its greater precision, MetaPro incorporates all DETECT predictions while only incorporating the union of the predictions obtained from both the PRIAM and DIAMOND searches. Finally for taxonomic annotations, MetaPro uses taxonomic assignments of the genes identified through prior searches of ChocoPhlAn and NR protein databases, supplemented with predictions from Kaiju [45] and Centrifuge [46], two high-performance short read taxonomic classifiers. These latter classifiers use the NR protein and NCBI nucleotide databases [41], respectively. Predictions from all sources are combined within a consensus framework to derive a single taxonomic assignment to each contig/read using WEVOTE. [47]

Since we expect MetaPro will be typically deployed on cluster computing environments that feature limitations in memory and/or processing time requirements, MetaPro has been designed to protect against potential points of failure through implementation of intermediate, human-readable, output files that serve as checkpoints. To keep up with the current state of technology, MetaPro was written in Python3. The user is also able to insert their own databases. Further, since ease of installation and use often represent significant barriers to software adoption, MetaPro minimizes software dependencies through the use of the Docker software deployment infrastructure [27]. Docker simplifies the task of distributing pipelines by combining all the tools used by MetaPro in a single image resulting in uncomplicated, single-step installation, and deployment independent of computing architecture. Furthermore, MetaPro is fully compatible with Singularity [28], a secure containerization software based on Docker, that is frequently employed in cluster computing environments.

Results and discussion

Comparisons of MetaPro performance relative to other pipelines

Datasets for benchmarking

MetaPro's approach to metatranscriptomic analysis is different from those of SAMSA2 [23] and HUMAnN3 [26]. Though MetaPro and SAMSA2 merge paired-end reads, SAMSA2 discards unmerged reads while MetaPro considers them valid. HUMAnN3 does not merge reads at all. We illustrate the methodological differences using three complementary datasets (Table 1; Supplemental Table 1). These comprise 12 samples obtained from the cecum and colon of 5 germ-free, non-obese diabetic (NOD) mice inoculated with Altered Schaedler

(See figure on next page.)

Fig. 1 Overview of MetaPro workflow and performance relative to two state-or-the art pipelines. **A** Overview of MetaPro's workflow, including the tools and databases used at each step. The bar on the right indicates the relative time required for each phase of the pipeline. **B** Chord diagrams tracking the trajectory of sequence reads across each pipeline, in comparison to MetaPro. Each arch represents one category of reads, bands mapping between arcs indicate the proportion of reads assigned by each pipeline to the associated category. **C** Stacked barcharts depicting the number of reads annotated to specific taxa in NOD mouse samples, and kimchi samples by BWA alignments, MetaPro, HUMAnN3, and SAMSA2. The NOD mouse datasets were generated from gut samples from mice inoculated with a defined microbial consortium (Altered Schaedler Flora (ASF) [29]. In addition to the 8 taxa associated with ASF, reads were also assigned to *Parabacteroidesgoldsteinii*, a close relative of *Parabacteroides ASF519* (see legend). The kimchi datasets comprise five major taxa (see legend [30–34]). It should be noted that *Leuconostoc gasicomitatum* reported in the original publication is currently classified as a subspecies of *Leuconostoc gelidum*. For NOD sample SRR1828965, HUMAnN3 did not annotate any reads

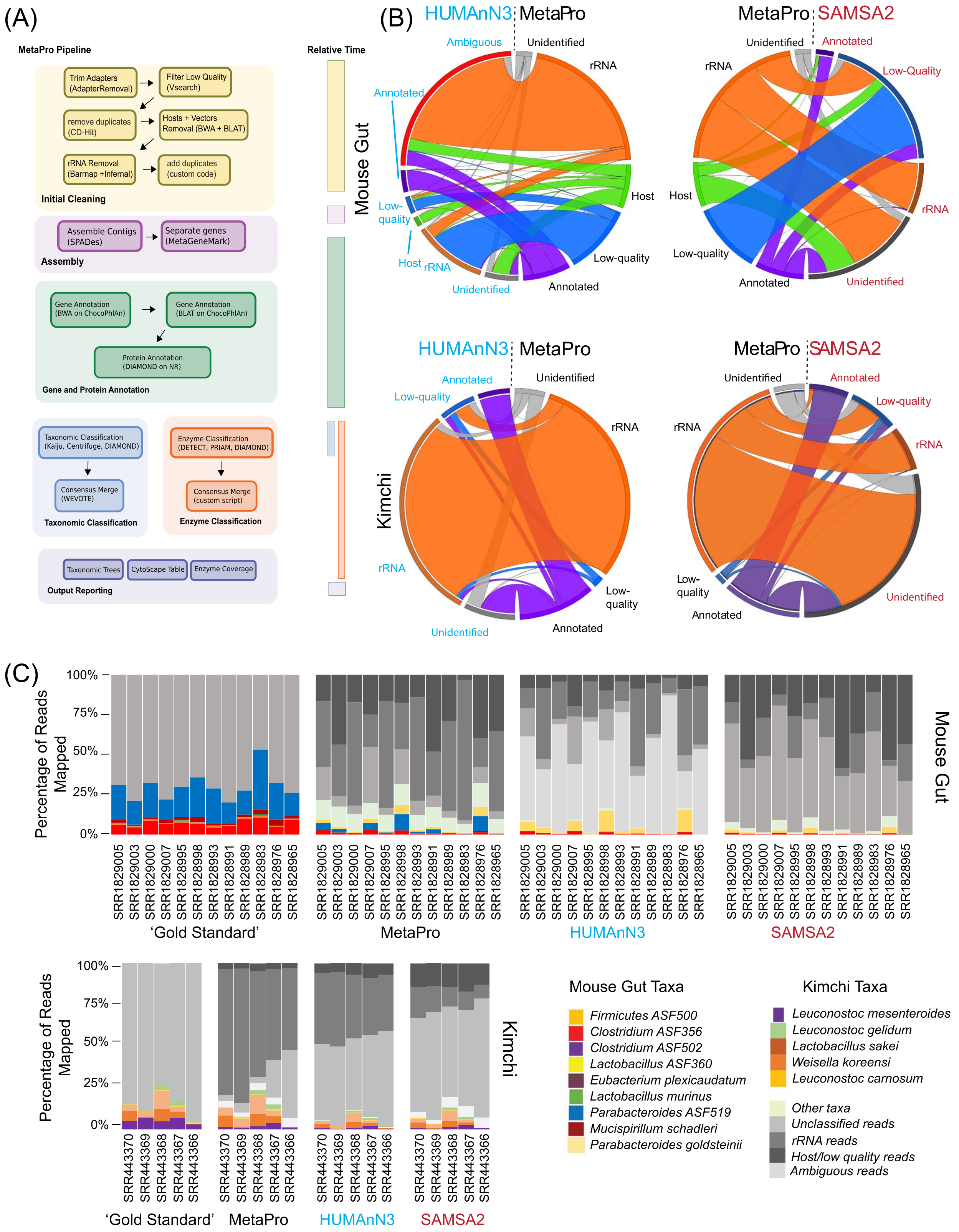

Fig. 1 (See legend on previous page.)

Table 1 Sequence read processing efficiency of MetaPro, HUMAnN3, and SAMSA2

Dataset	Pipeline	Mean % of total reads that are putative mRNA	Mean % of annotated putative reads	Mean annotated putative reads, as a % of total reads	Mean unique gene families[a]
NOD mouse	MetaPro	17.16	71.64	12.73	14,308
NOD mouse	HUMAnN3	79.95	2.80	2.24	2691
NOD mouse	SAMSA2	39.33	10.85	4.27	11,499
Kimchi	MetaPro	32.21	59.27	19.09	87,193
Kimchi	HUMAnN3	89.26	11.71	10.45	9166
Kimchi	SAMSA2	72.54	15.72	11.4	86,270
Oral Biofilm	MetaPro	97.48	96.96	94.52	95,707
Oral Biofilm	HUMAnN3	100	71.31	71.31	28,750
Oral Biofilm	SAMSA2	98.98	80.40	79.58	298,315

[a] Unlike MetaPro and SAMSA2, HUMAnN3 reports only the number of unique gene families

Flora (ASF) [29] bacteria; 5 samples obtained during a 29-day fermentation of kimchi [48] at days 7, 13, 18, 25, and 29; and 8 samples obtained during the maturation of an in vitro oral biofilm cultured from a complex human oral microbiome [49] at the 6, 9, 11, 13, 15, 17, 21, and 24-h time points. The ASF consortium represents a standardized collection of 8 bacteria used in studies of the mouse gut microbiome. The availability of their genome sequence [13] provides an effective gold standard to benchmark pipeline performance. Similarly, genome sequence data is available for the five species of lactic acid bacteria that, together, dominate over 97% of the kimchi fermentation datasets. Lastly, the human oral microbiome dataset represents putative mRNA reads from a highly complex microbial community encompassing over 700 bacterial taxa, many of which have yet to be cultured. A gold standard is not available for this dataset. However, the intention of analysis of the human oral microbiome dataset was to assess the performance of the three pipelines on a dataset exhibiting similar complexity to microbial communities typically associated, for example, with human health. The performance of each pipeline was assessed against each of these datasets to assess the following metrics: read filtering, gene annotation, taxonomic assignments, and enzyme function assignments.

Read filtering

Sequence processing pipelines require the filtering of input read sequences based on the sequence quality and potential contamination. For metatranscriptomics, sources of contamination include sequence adaptors, reads of host origin, and RNA reads derived from non-mRNA sources. Identification and removal of these reads are important to both reduce downstream processing time as well as to avoid potentially incorrect inferences of data generated by the pipeline. We therefore examined the performance of each pipeline to filter reads from each of the benchmark datasets (Table 1). Filtering of contaminant reads is particularly important in the context of host-associated metatranscriptomic datasets as sequences of host origin can represent a significant proportion of reads in a sample [15]. Both MetaPro and HUMAnN3 (which relies on a Bowtie2-based [50] sequence filtering tool termed Kneaddata [51]) provide the option to filter host-associated reads. Both pipelines allow for user-defined custom reference databases of host-associated reads to be used in this filtering step. In our performance testing, we used the same mouse and human host-associated sequence reference libraries in both tools to ensure comparability between the results. SAMSA2, on the other hand, does not allow for filtering contaminant reads. We also tracked the reads of each dataset through each pipeline to assess the relative performance of each pipeline's filtering protocols (Fig. 1B). For the mouse NOD dataset, while a minor fraction of reads annotated as host by MetaPro were similarly annotated by HUMAnN3, many others were annotated as either "unidentified" or "ambiguous" (this latter category reflects the challenge in accounting for read annotation associated with the processing of paired-end datasets by HUMAnN3—see below). SAMSA2 maps most reads annotated as host by MetaPro, to "unidentified," or were filtered for low-quality.

Due to the lack of a polyA tail, bacterial RNASeq datasets often feature a high abundance of non-mRNA reads such as those of rRNA or tRNA origin [15]. Identification and removal of these reads can significantly reduce the size of the putative mRNA dataset and increase the speed of downstream analyses. Furthermore, during the generation of many metatranscriptomic datasets rRNA depletion kits are typically employed to reduce these abundant sequence moieties. Such kits often perform better for specific taxa, resulting in taxonomic bias in the removal of rRNA sequences. It is therefore important to bioinformatically filter for any remaining rRNA to minimize their impact on taxonomic readouts. MetaPro filters rRNA and tRNA using a two-tiered approach consisting of the BAsic Rapid Ribosomal RNA Predictor (Barrnap) [52] and Infernal [53]. Barrnap is a fast RNA filtering tool utilizing hidden Markov models of sequence families, whereas Infernal is a slower, more sensitive RNA filtering tool utilizing both RNA sequence and secondary structure to identify sequence-divergent RNA homologs that conserve their secondary structure. Combining these tools results in sensitive filtering of rRNA and tRNA at an increased speed compared to using only Infernal. In contrast, HUMAnN3 uses the Bowtie2-based KneadData [51] for rRNA sequence filtering and SAMSA2 uses SortMeRNA [54], a fast sequence-based filtering tool, to filter rRNA sequences.

Due to the methodological differences discussed above, the three pipelines analyzed here each generated differing numbers of putative mRNA reads (Supplemental Table 1). When analyzing the NOD mouse dataset, MetaPro, SAMSA2, and HUMAnN3 predicted that an average of 24.0 ± 15.6%, 39.3 ± 11.5%, and 80.0 ± 8.3%, respectively, of the total reads in the sample as being of putative mRNA origin. The significantly higher proportion of predicted mRNA reads in these samples by HUMAnN3 likely represent insufficient filtering of low-quality or host-associated reads. When analyzing the kimchi fermentation dataset, MetaPro predicted that an average of 32.2 ± 12% of the total reads in the sample were putative mRNA while SAMSA2 and HUMAnN3 predicted 72.5 ± 3.9% and 89.3 ± 1.2%, respectively. Like the NOD mouse dataset, the high proportion of predicted mRNA reads by SAMSA2 and HUMAnN3 likely represent instances of insufficient filtering. Lastly, when analyzing the human oral biofilm dataset, which had low quality and putative rRNA reads prefiltered [49], all three pipelines predicted that nearly all of the sample consisted of putative mRNA reads. MetaPro, SAMSA2, and HUMAnN3 predicted 97.5 ± 1.1%, 99.0 ± 0.4%, and 100.0 ± 0.01% putative mRNA, respectively.

When comparing the relative performance of rRNA filtering between MetaPro and the other pipelines (Fig. 1B), we see in the mouse gut data that MetaPro identifies the same rRNA reads as SAMSA2; however, a substantial fraction of reads assigned as rRNA by MetaPro is assigned as either "unidentified" or "low-quality" by SAMSA2. For the HUMAnN3 pipeline, due to challenges in annotation accounting associated with paired read data (see below), the majority of reads annotated as rRNA by MetaPro have "ambiguous" assignments in the HUMAnN3 pipeline. Further, the majority of reads annotated by HUMAnN3 as rRNA were deemed "low quality" by MetaPro. While similar behavior is observed with the kimchi samples for the comparison with SAMSA2, MetaPro and HUMAnN3 exhibit a high degree of concordance in rRNA read assignments, highlighting the significant difference in how HUMAnN3 handles single-end and paired-end datasets.

Indeed many of the differences we observe across the three pipelines reflect the way in which each pipeline handles paired-end data. SAMSA2 separates the unmerged reads but only annotates the merged reads. This results in far fewer putative reads when compared to MetaPro or HUMAnN3. HUMAnN3, which uses the filtering tool—Kneaddata, treats the forward, and reverse reads as two concatenated sets, but with each read of a pair receiving the same sequence ID. Consequently as reads progress through the HUMAnN3 pipeline, if pairs of reads receive discordant annotations, since there is no unique ID for the pair—the assignment for that ID becomes ambiguous. For example, the same ID for a read defined as host can also correspond to a read defined as rRNA, while a read ID defined as rRNA can also appear as a putative gene. This results in HUMAnN3 annotating the most putative reads of the pipelines.

As noted earlier, MetaPro attempts to merge paired-end reads to reduce compute overhead. This data reduction process results in two broad categories of sequences: a set of discrete "genes" derived from contigs and a set of "singleton" reads representing unassembled merged and unmerged reads. In subsequent annotation efforts, we make the assumption that paired-end reads (merged or unmerged) derive from the same gene. Thus, during the annotation step, in the case where the two reads of a pair do not match to the same gene, we assign both reads to the gene which has the highest scoring match. We acknowledge that it is possible that the two reads of a pair could map to different genes (for example as might occur from a polycistronic transcript), and this possibility increases for libraries generated with larger insert sizes. At the same time, however, we note that at least for the NOD mouse datasets, only 0.2–7.8% of non-overlapping paired reads that assembled into contigs were annotated to different genes (Supplemental Table 2). This approach simplifies the accounting process and enables a level of transparency to the user, providing, for example, a map that documents the mapping of each read to "genes" within a contig. Reads do not appear more than once in this map, as each read assignment is based on the alignment score defined by BWA. Only the first, highest scoring, alignment match is used.

Assigning reads to genes

The ability to assign a short read to a gene is a critical step towards functional and taxonomic profiling of metagenomic and metatranscriptomic datasets. Here, we assessed the ability of each pipeline to assign reads of putative mRNA origin to individual transcripts. For the NOD mouse and Kimchi datasets, we were interested in comparing each tool's ability to map reads to genes associated with the taxa known to be present in these datasets. While MetaPro and SAMSA2 assign reads to distinct genes and proteins, HUMAnN3 reports only on the basis of gene families as defined by UniRef90 [55]. This is reflected in the reports of unique transcripts identified in the datasets (Table 1 and Supplemental Table 1), where MetaPro and SAMSA2 report similar results, except for the oral biofilm datasets where the larger number of unique transcripts reported by SAMSA2 is likely associated with less stringent criteria used in sequence similarity searches. For HUMAnN3, gene family abundances are split by taxa and reported in terms of reads per kilobase (RPK), which combines matches to members of a gene family normalized for the length of each member matched as well as for sequences that match to multiple reference genes. Thus, to compare to the gene-centric reporting of MetaPro and SAMSA2, we used intermediate files generated by the Bowtie2 and DIAMOND searches in the HUMAnN3 pipeline to map reads to individual genes and proteins. Reads mapping to genes from multiple taxa were allocated equally to each taxon (with the relative contribution of that read assigned to each taxon being a proportion of the number of taxa the read mapped to). Reads that were not subsequently aligned (and reported) to a UniRef90 gene family were removed.

From each pipelines mappings, we compared the number of putative mRNA reads annotated by each pipeline to genes and/or proteins associated with the 8 ASF (NOD mouse datasets) and 5 lactic acid bacteria (kimchi datasets) known to be present in each sample (Fig. 1C). To establish a gold standard for each sample, we used BWA [38] to perform sequence similarity searches of reads from each sample against a database comprising only the genomes of either the 8 ASF bacteria or the 5 lactic acid bacteria. The gold-standard was created using raw reads, prior to any filtering, allowing the performance of each pipelines filtering protocols to also be assessed in this analysis. While we note that sequence data for all 8 ASF and 5 lactic acid bacteria were present in databases used by each pipeline, we expect that the considerably greater taxonomic representation of sequences in these databases would result in false positive taxonomic assignments.

Across the 12 NOD mouse samples, our BWA-based benchmarking annotated a total of 8.6 million reads to genomes of the 8 ASF bacteria (Fig. 1C and Supplemental Table 3). In contrast to the number of reads annotated to be of putative mRNA origin, MetaPro was able to annotate more reads to a known gene/protein (4.25 million) compared to SAMSA2 (1.49 million) or HUMAnN3 (1.47 million) (Supplemental Tables 1 and 3). Furthermore, MetaPro also annotated a greater number and proportion of these reads to one of the 8 ASF bacteria (1.45 million reads, 34.1%), compared to both SAMSA2 (179,000 reads, 12%) and HUMAnN3 (177,000 reads, 12%). A majority of the reads annotated to ASF bacteria by SAMSA2 and HUMAnN3 were associated with *Clostridium* ASF356. However, *Parabacteroides* ASF519, the most prevalent species in the samples as defined by the gold standard (66.2% ± 6.5% of reads), was absent from the HUMAnN3 results and represented by only 7053 reads across all 12 samples annotated by SAMSA2. Instead, it appears that both pipelines erroneously assign many reads to transcripts derived from a closely related species, *Parabacteroides goldsteinii*. In contrast, MetaPro identified 581,849 reads aligning to *Parabacteroides ASF519* in all 12 samples. Furthermore, HUMAnN3 failed to assign any reads for sample SRR1828965 and was unable to assign reads to transcripts associated with *Clostridium* ASF502, *Eubacterium plexicaudatum* ASF492, *Firmicutes* ASF500, and *Lactobacillus* ASF360. MetaPro and SAMSA2 also exhibited poor annotation rates for these taxa, although we note that the version of ChocoPhlAn that MetaPro used in these analyses did not contain *Firmicutes* ASF500. Overall, through precision-recall analyses, we find that MetaPro exhibits greater recall than both HUManN3 and SAMSA2 for all 12 mouse samples, although we note that the limited number of annotations assigned by SAMSA result in a higher precision than MetaPro for several datasets (Fig. 2A).

To further track relationships of annotations across pipelines, we compared annotations derived from each pipeline against gold-standard assignments (Fig. 2B, Supplemental Fig. 1). In these analyses, we removed all rRNA reads detected by MetaPro. While many reads assigned by the gold standard to *P. ASF519* mapped to reads assigned to *P. ASF519* and its close relative, *P. goldsteinii*, by MetaPro, we also identified mappings to several other categories, notably "other organisms," that include assignments at different levels of taxonomy. Furthermore, we found a relatively small proportion of reads defined by the gold standard as *P. ASF519*, which are defined as "unidentified" by MetaPro. Notably, both HUMAnN3 and SAMSA2 map a greater proportion of gold standard-defined *P. ASF519* reads as "unidentified" (Supplemental Fig. 1A).

Across the five kimchi fermentation samples, BWA-based benchmarking annotated a total of 49.6 million reads to genomes of the 5 lactic acid bacteria (Fig. 1C and Supplemental Table 4). Similar to the NOD mouse samples, MetaPro was able to annotate more reads to a known gene/protein (19.6 million) compared to SAMSA2 (20.2 million) or HUMAnN3 (18.5 million) (Supplemental Tables 1 and 4). Furthermore, MetaPro also annotated more of these reads to one of the 5 lactic acid bacteria (22.2 million reads, 75%), compared to both SAMSA2 (11.3 million reads, 56%) and HUMAnN3 (15.2 million reads, 82%). Interestingly, sample SRR443366 proved challenging to identify 3 of the 5 lactic acid bacteria: *Weissella koreensis*, *Leuconostic carnosum*, and *Leuconostic gelidum*, relative to other samples. Few reads were annotated to these taxa by both MetaPro and SAMSA2, and no reads were annotated to these taxa by HUMAnN3. Precision-recall analyses again demonstrated a superior performance of MetaPro relative to the other two pipelines (Fig. 2A). Mapping of (non rRNA) read assignments across tools further revealed a higher degree of concordance between MetaPro and gold standard assignments relative to the other two tools (Fig. 2B and Supplemental Fig. 1B).

From these analyses, we found that MetaPro outperforms SAMSA2 and HUMAnN3 in terms of assigning putative mRNA reads to the appropriate taxon of origin. For HUMAnN3, the limited ability to assign reads to appropriate taxa may reflect the pipeline's reliance on paired metagenomic data and the use of MetaPhlAn3 [56] to identify taxa likely associated with the dataset being processed. This allows the use of smaller, more targeted, reference databases for sequence similarity searches, greatly reducing runtime. However, this strategy may be compromised if the taxa have not previously been well sampled, potentially explaining the inability of HUMAnN3 to annotate any reads for sample SRR1828965.

We note that HUMAnN3 was recently developed to supersede a previous version, HUMAnN2. Comparisons reveal these two versions generally yield similar results (Supplemental Fig. 2; Supplemental Tables 1, 3 and 4). However, for the NOD mouse datasets, HUMAnN3 assigned more reads to *Clostridium ASF356* over its predecessor, while HUMAnN2, unlike HUMAnN3, was able to assign at least some reads in sample SRR1828965. For the kimchi datasets, HUMAnN3 was able to assign more reads to *Leuconostoc gelidum*, and *Leuconostoc carnosum* and, unlike HUMAnN2, was able to assign reads in sample SRR443366. We suggest these differences in the ability to assign reads to the two datasets, and SRR1828965 and SRR443366 may be related to the different versions of MetaPhlAn used by each pipeline to generate targeted reference databases.

Overall, we found that MetaPro best reflected the gold standard assignments, although as for the other pipelines, MetaPro assigned many reads to sequences from other taxa not expected in the samples. While this highlights the need to improve gene annotation methods, we nonetheless note that such sequences represent potential homologs of genes encoded by the expected taxa and are thus likely to be functionally informative.

Inferring accurate taxonomic ranks

In the previous section, we benchmarked each pipeline's ability to map putative mRNA reads to transcripts of genes encoded by the genomes of bacterial taxa known to comprise the datasets. However, determining the taxonomic composition of a sample to elucidate taxa responsible for providing critical functions remains a challenge in metatranscriptomic analysis. MetaPro features an additional stage for taxonomic annotation to assign taxa more accurately to putative mRNA reads. MetaPro utilizes input from the short-read classifiers Kaiju and Centrifuge [46], together with the inferred taxonomy from gene annotations derived from the initial tiered set of sequence similarity searches, to determine the taxonomic composition of a sample. The taxonomic output from the short-read classifiers and the gene annotation stage are combined into a single consensus taxonomic classification using WEVOTE [47] in order to increase the precision of classification. WEVOTE considers the taxonomic predictions from the two short-read classifiers, as well as the gene annotations, using a simplified variant of the NCBI taxonomy tree structure and selects the highest confidence taxonomic classification based on the output from individual approaches. WEVOTE produces a consensus classification depending on the pattern of taxonomic classification produced by the three predictors. HUMAnN3 utilizes MetaPhlAn3 [57] for taxonomic profiling based on clade-specific marker genes in the sample whereas SAMSA2 solely utilizes the curated taxonomic annotations present in the RefSeq database to determine read classification and performing taxonomic classification after gene annotation.

To assess the performance of each pipeline, we compared the total number of reads that were correctly annotated to specific taxonomic ranks, from species to phylum, within the 8 ASF and 5 lactic acid bacteria for the NOD mouse and kimchi fermentation datasets (Fig. 2C and Supplemental Fig. 3). For the NOD mouse datasets [15], MetaPro can annotate an average of 71.6 ± 7.4% of reads across all samples, 59.7 ± 8.5% down to the species level, and 3.4 ± 0.5% to the genus level. The remaining 8.5 ± 3.5% of reads were annotated to the rank of family or higher. SAMSA2 annotated 8.1 ± 6.5% of each sample, with 0.2 ± 8.5% down to the species level, and 6.3 ± 5.7%

to the genus level, while HUMAnN3 annotated an average of 5.6 ± 6.7% reads to each sample only to the species level. HUMAnN3 defaults to reporting on species-level taxa.

For the Kimchi set, MetaPro can annotate an average of 85.8 ± 8.5% of putative reads across all samples, 76.3 ± 7.8% down to the species level, and 7.6 ± 2.3% down to the genus level. The remaining 1.9 ± 0.9% of the reads annotated to the rank of family or higher. SAMSA2 annotated an average of 13.4 ± 7.6% across all kimchi samples, with 5.2 ± 3.5% to the species level, and 6.59 ± 4.40% to the genus level. The remaining 1.6 ± 0.3% annotate to a taxon of family or higher. HUMAnN3 annotated an average of 11.7 ± 5.7% reads to the species level and only to the species level. Though we recognize that HUMAnN3 and MetaPro treat reads differently, we made a design choice to provide information at different taxonomic levels to better help the user understand the structure of their data. In contrast, HUMAnN3 relies on clade markers to improve annotation accuracy which can help, for example, troubleshoot potential errors that may have occurred during sample collection and processing.

In addition to the NOD mouse and kimchi fermentation datasets, we were also interested in examining the performance of the pipelines on classifying more complex datasets derived from human oral microbiomes (Fig. 2C and Supplemental Fig. 3). These datasets comprise 8 samples and provide more complex data to further assess the performance of the three tools, with the caveat that we do not know what taxa are present in these samples. MetaPro annotated an average of 93% ± 1% of the data, leaving 7% ± 1% unable to be identified SAMSA2 was able to annotate an average of 80% ± 1.9% of its putative reads, with 20% ± 1.9% of the reads unclassified to any known taxa. HUMAnN3 annotated an average of 63.8% ± 2.6% of its putative reads, leaving 36.1 ± 2.6% unmapped. Of MetaPro's putative reads, an average of 75.6 ± 3.2% were identified down to the species level, and 14% ± 2.8% of the reads were annotated to the genus taxa. SAMSA2 annotated an average of 65% ± 1.5% of its putative reads to a species, and 13% ± 0.5% of those reads to a genus. HUMAnN3 annotated all 64.8 ± 2.6% of its putative reads to a species.

Overall, we found that MetaPro's ensemble approach to taxonomic annotation consistently outperformed the approaches employed by SAMSA2 and HUMAnN3, although we did note a performance improvement for HUMAnN3 over HUMAnN2 (Supplemental Fig. 2).

Annotating enzymatic functions

The ability of each pipeline to accurately infer enzymatic functions of the datasets is essential for understanding the metabolic activity of a sample. Since annotations based on simple similarity searches can yield false positive rates of up to 50% [58], MetaPro relies on a robust approach that combines predictions from DETECT [42], with those from PRIAM [43] that are also confirmed by sequence similarity searches against the Swiss-Prot database [59]. Though these approaches are still based on sequence similarity, this combinatorial method outperforms the use of any one individual method and has been effectively applied in a number of settings [60–65]. Though the approaches MetaPro uses are also based on sequence similarity, they have additional scoring and qualitative factors. DETECT uses a global alignment score, while PRIAM uses a position-specific scoring matrix. These additional qualitative scores enhance their performance over the traditional sequence similarity methods.

HUMAnN3 reports MetaCyc pathway abundances, but to do this, the gene families are translated from their UniRef IDs into MetaCyc reactions, using a static relational map. This map also includes enzyme commission (EC) [66] numbers. HUMAnN3 then uses a pathway-to-reactions mapping, locating minimally satisfied paths using MinPath [67]. SAMSA2 uses a static relational mapping of the RefSeq [68] database for gene annotation, and the SEED subsystem [69] database for enzymatic function annotation. Both databases are searched using DIAMOND-based sequence similarity searches using a relatively permissive *e* value cutoff of 0.001.

MetaPro performs enzyme annotations by first translating genes identified through the BWA and pBLAT searches into proteins and adding these to the proteins identified through the DIAMOND-based searches. Enzyme predictions reported by MetaPro consist of all the DETECT results, together with the intersection of enzymes predicted by both PRIAM and DIAMOND using an *e* value cutoff of 1e − 5. MetaPro also includes a high-confidence mode that will only return enzymes (reported as ECs) that the pipeline detects with high stringency cutoffs for PRIAM (probability score greater than 0.5) and DIAMOND (*e* value cutoff of 1e − 10). MetaPro can sometimes predict multiple enzymes per protein. To resolve these instances, MetaPro compares the enzyme matches against a reference database of valid pairs of enzyme combinations compiled from the Swiss-Prot database. Pairs that lack experimental support (i.e., not previously identified in Swiss-Prot) are removed. In situations where more than two enzymes are predicted in a single protein, the enzymes with the two highest probability scores are used.

In all samples, we found that MetaPro identifies more unique ECs than HUMAnN3 and SAMSA2 (Fig. 3A, Supplemental Fig. 4; Supplemental Tables 5 & 6). For example, comparisons of predictions for the NOD mouse samples (Fig. 3A) reveals MetaPro identifies an average of 949 unique ECs/sample (622 defined as high-quality) compared to 465 and 170 for SAMSA2 and HUMAnN3, respectively. Of these, 350 were shared with SAMSA2 and 153 were shared with HUMAnN3. Both SAMSA2 (41 ECs) and HUMAnN3 (4 ECs) predicted combinations of ECs that have not been identified in the same protein in Swiss-Prot. The kimchi datasets show a similar performance (Fig. 3A), with MetaPro identifying 1982 unique ECs/sample (1238 defined as high quality) compared to 859 and 600 for SAMSA2 and HUMAnN3, respectively. Of these, 692 were shared with SAMSA2 and 512 were shared with HUMAnN3. For these datasets, SAMSA2 and HUMAnN3 predicted 75 and 32 EC combinations of ECs that have not been identified in the same protein in Swiss-Prot. Finally, for the oral biofilm samples (Fig. 3A), MetaPro identified an average of 1465 unique ECs/sample (1024 defined as high quality), compared to 858 and 1171 for SAMSA2 and HUMAnN3, respectively. Of these, 674 and 856 were shared with SAMSA2 and HUMAnN3, respectively. Furthermore, SAMSA2 and HUMAnN3 identified an average of 79 and 120 EC combinations not previously seen in Swiss-Prot annotated proteins.

In summary, MetaPro delivers superior performance over the other two tools with both a greater number of EC predictions and greater confidence of annotations (through the combined use of DETECT, PRIAM, and DIAMOND predictions) relative to the simple sequence-similarity-based approaches used by the other two tools. As a sidenote, we did find that HUMAnN3 showed improved performance over HUMAnN2, with the latter identifying fewer ECs in both the kimchi and oral biofilm datasets (588 and 986 unique ECs, respectively) while also predicting a higher number of combinations of ECs not supported by Swiss-Prot annotations (Supplemental Fig. 4 and Supplemental Table 7).

Result reporting

To allow for more intuitive exploration of processed data, MetaPro features several text and visual outputs. For context, HUMAnN3 produces three text files: (1) a gene families file that details the abundance of each gene family in the community (as measured by reads per kilobase; RPK), stratified to show the contribution from each species; (2) a pathway abundance file that details a normalized abundance of pathway components for each pathway, again stratified to show species contributions; and (3) a pathway coverage table, which provides a confidence score for the presence of a pathway in the community, as well as in individual species. SAMSA2 is typically used to compare between two datasets but does generate files that report: (1) DIAMOND search results, (2) a taxonomic summary of read pair counts, and (3) an enzyme classification summary of read pair counts. In contrast, MetaPro produces the following: (1) a histogram of read quality, together with a summary of read processing (reads filtered for quality, host contamination, rRNAs and tRNAs; putative mRNAs; annotated mRNAs; number of unique transcripts; and number of unique ECs); (2) a detailed gene-to-read mapping of every read the pipeline annotated; (3) a summary of every taxon identified in the sample, together with read pair counts; (4) a list of ECs identified, together with the gene or protein providing the annotation; (5) a table of genes with associated reads per kilobase of transcript, per million mapped reads (RPKM), with details on the contribution of taxa that with further details on the contribution of taxa that are individually responsible for at least 1% of putative reads; (6) a CytoScape-compatible annotation table that can be readily imported into Cytoscape and mapped on to KEGG defined pathways to illustrate the contribution of specific taxa to enzymes expressed in the selected pathway (Fig. 3B). This latter mapping requires the use of the KEGGmapper and enhancedGraphics app plugins for the Cytoscape platform and is described in more detail in the accompanying tutorial; and (7) a heatmap in png image format, showing the relative contribution of the 20 most prevalent taxa to enzyme expression grouped in Kyoto Encylcopedia of Genes and Genomes (KEGG)-defined superpathways (Fig. 3C).

The read summary table contains: (1) total number of reads in the sample; (2) the number of high-quality reads; reads where the adaptors removed, trimmed of low-quality reads, and then merged; (3) a percentage of (2), relative to the total reads in the sample (1); (4) the number of host reads that have been removed by MetaPro (including duplicates); (5) a percentage of column (4), relative to the total reads in the sample (1). (6) The number of vector reads removed by MetaPro (including duplicates). (7) A percentage of column (6), relative to the total reads of the sample (1). (8) The number of reads removed by MetaPro's rRNA filter. (9) A percentage of column (8), relative to the total number of reads (1). (10) The number of putative reads (mRNA) to be annotated. (11) A percentage of column (10) to the total reads in the sample (1). (12) The number of reads annotated by MetaPro's gene-annotation step. (BWA, pBLAT, and DIAMOND). (13) A percentage of the reads in column (12), relative to the number of putative reads (10). (14) The number of unique genes annotated by MetaPro's gene annotation step ((BWA, pBLAT, and DIAMOND). (15) The number of unique enzymes MetaPro detected in its enzyme annotation step, using the "high" stringency settings.

(16) The number of unique enzymes MetaPro detected in its enzyme annotation step, using the "low" stringency setting.

Computational overhead

Processing time represents an inherent tradeoff between the robustness of an analysis and the speed with which it can be completed. MetaPro, SAMSA2, and HUMAnN3 are each designed to perform specific analyses. MetaPro focuses on the annotation of samples of metatranscriptomic data to provide a detailed breakdown of taxonomic contributions. SAMSA2 was designed to perform comparative analyses between two datasets. HUMAnN3 was primarily designed primarily for the analysis of metagenomic data, reporting gene family abundances, but also accepts metatranscriptomic data. We compared the execution of the MetaPro, SAMSA2, and HUMAnN3 pipelines across all 25 datasets using one computing node equipped with 20 Intel Skylake 2-core-CPU, 202 GB RAM, running CentOS 7. Each pipeline differs significantly in their completion times (Table 2 & Supplemental Table 8). For the NOD mouse datasets, the mean completion time for each sample was 16,071 ± 5259, 2342 ± 321 and 1310 ± 267 s for MetaPro, SAMSA2, and HUMAnN3, respectively. Similarly, for the five kimchi fermentation datasets, mean completion time per sample was 53,733 ± 18,287, 5883 ± 1739, and 11,067 ± 903 s for MetaPro, SAMSA2, and HUMAnN3 respectively. However, for the 8 human oral biofilm datasets, we found the mean completion times were longer for SAMSA2 being 46,835 ± 2679, 53,939 ± 3191 and 14,282 ± 1126 s for MetaPro, SAMSA2, and HUMAnN3, respectively. With the exception of the oral biofilm datasets, MetaPro requires a longer execution time than the other pipelines, with, for example roughly 30% of runtime dedicated to the initial cleaning step, which includes rRNA removal (Fig. 1A). However, as indicated above, this increase in runtime stems from the preference to prioritize accuracy over speed. The majority of the execution time within MetaPro is accounted for by both the use of a large reference file (in the case of DIAMOND with the NR database), as well as the use of robust tools for enzymatic function inference (Fig. 1A) rather than static mappings used by other tools. For example, in assigning reads to genes, MetaPro uses the entire ChocoPhlAN database, instead of the subset used by HUMAnN3. Furthermore, MetaPro performs 3 times as many annotation phases as SAMSA2. For the oral biofilm dataset, it was interesting to note the longer execution time of SAMSA2. This appears related to the higher number of reported hits to the RefSeq database during read annotation. This likely resulted in a slow down due to the creation of the large output files associated with reporting the results of the searches.

Since the choice of database can impact runtime performance, MetaPro's ability to use customized databases offers the potential for the user to select smaller and more specialized databases that would result in significant performance speedups. While MetaPro's default use of large databases, and multiple passthroughs of data using different tools slows performance, it does ensure both enhanced coverage and improved accuracy of taxonomic and functional (i.e., EC annotations) assignments. Nonetheless, we have included a section on how to add user defined databases to our software documentation.

Easy installation of MetaPro through docker

MetaPro is packaged as a Docker [70] container for easy distribution. Every third-party software tool MetaPro uses is installed under the docker container, as well as all custom scripts we developed as part of the pipeline. The user will still need to obtain their own copy of MetaGeneMark's license key. MetaPro is compatible with Singularity [28], a specialized version of Docker meant for scientific computing environments. By using Docker and Singularity, MetaPro can interface with job schedulers

Table 2 Average runtime performance of MetaPro, HUMAnN3, and SAMSA2

Dataset	Pipeline	Pre-processing	St. dev	Data analysis	St. dev	Total time	St. dev
NOD mouse	HUMAnN3	339.50	101.00	970.92	211.83	1310.42	267.13
NOD mouse	SAMSA2	–-	–-	–-	–-	2341.75	320.92
NOD mouse	MetaPro	3869.12	3442.85	12,202.28	4696.80	16,071.40	5259.01
Kimchi	HUMAnN3	1960.80	133.92	9106.20	809.62	11,067.00	902.83
Kimchi	SAMSA2	–-	–-	–-	–-	5882.60	1738.68
Kimchi	MetaPro	24,853.02	3256.61	101,984.66	67,339.49	12,6837.68	70,477.12
Oral Biofilm	HUMAnN3	3481.00	203.84	10,800.63	967.63	14,281.63	1126.46
Oral Biofilm	SAMSA2	–-	–-	–-	–-	53,938.88	3190.63
Oral Biofilm	MetaPro	19,183.11	2479.90	41,996.21	3424.56	61,179.33	5191.15

like any other docker-ized software. The current iteration of MetaPro can only handle one data sample per run. It was not designed to work on multiple samples simultaneously.

Hardware requirements

MetaPro requires at least 707 GB of disc space to store the default reference libraries. MetaPro also currently requires up to 6 times the amount of disc space that the sample dataset occupies. This is to ensure there is room for the intermediate files the pipeline creates. As for RAM, there is no hard limit, but we recommend at least 128 GB of memory due to the requirements of DIAMOND. MetaPro was meant for large data processing nodes, with ample RAM and CPU cores.

MetaPro online tutorial

To introduce and guide researchers through the process of analyzing metatranscriptomic data, MetaPro features a dedicated tutorial mode. The tutorial has been designed for both computational novices as well as more seasoned bioinformaticians starting out in metatranscriptomic analysis. The interactive tutorial takes users through the tasks of filtering data, aligning reads to databases, and scanning identified genes through taxonomic and enzyme classification tools. The tutorial includes additional intermediate steps, such as reading the quality of the data, and interfacing the data with visualization software (CytoScape) to highlight the importance of each step and help the user understand how MetaPro parses the data. The tutorial is run within the MetaPro Docker container and was developed over several years of workshops and classes provided to both undergraduate and graduate students, as well as through the Canadian bioinformatics workshop series (https://www.bioinformatics.ca).

Conclusions

Increasingly, microbiome studies are shifting emphasis from identifying the composition of complex microbial communities to understanding how they function. Supporting this shift is the emergence of whole microbiome RNASeq (metatranscriptomics). However, despite the recognized value of these datasets, few dedicated tools are currently available for their analysis. To address this need, we developed MetaPro, a single package capable of processing and analyzing metatranscriptomic datasets. Our benchmarking analyses show that MetaPro delivers superior performance over existing pipelines in terms of gene, taxonomic, and enzyme annotations. Output is delivered as normalized gene expression profiles in terms of RPKM, together with a novel visualization framework based on the network visualization tool, Cytoscape, that allows the display of enzyme expression and the taxa responsible in the context of individual metabolic pathways. To assess the performance of MetaPro, we used samples derived from NOD mousececal samples, kimchi fermentation, and samples from a human oral biofilm dataset to compare against HUMAnN3 and SAMSA2. The differences in features are summarized in a table (Table 3).

MetaPro is readily installable with limited dependencies allowing deployment across multiple compute architectures, including cloud computing environments, allowing for massively parallel scale up required for the hundreds of millions of reads typically generated in a single experiment. MetaPro is designed to be flexible to account for the incorporation of improved algorithms as they are developed. To help users understand the various steps involved in metatranscriptomic analysis, we provide an established tutorial that has been developed with non-specialists in mind. Currently, the only major limitation to metatranscriptomic analysis using the MetaPro pipeline is speed of execution relative to existing tools. However, we consider speed to be secondary to accuracy. We are currently developing novel methods to enhance pipeline performance in terms of speed without compromising accuracy of the results.

Methods

MetaPro workflow

The MetaPro workflow involves a series of steps which first filters, then annotates the sequence data. In initial preprocessing steps are as folloes: MetaPro identifies and removes low-quality reads and adapters. Reads are subsequently merged and duplicates identified and removed. After merging, the reads are further filtered for reads of host origin and other potential contaminants. Next, reads associated with non-coding RNA moieties (e.g., rRNAs and tRNAs) are identified and filtered. The remaining, reads of putative mRNA origin are subsequently repopulated with any duplicate reads, before being assembled into contigs, which are used to define discrete genes.

Table 3 Comparison of tool characteristics

Feature	MetaPro	HUMAnN3	SAMSA2
Filter host using custom databases	Yes	No	No
Filter vectors	Yes	No	No
Filters rRNA	Yes	Yes	Yes
Uses multiple processors	Yes	Yes	No
Assembles contigs	Yes	No	No
Merges reads	Yes	No	Yes
Compatible with Docker	Yes	Yes	No

After pre-processing, MetaPro begins the annotation process by identifying which gene each read belongs to, using sequence aligners. Once complete, the pipeline attempts to assign the taxonomy of each read while also attempting to annotate enzymatic functions to each gene. Finally, MetaPro summarizes the details of its analysis into a series of output files (Fig. 1A).

Sequence preprocessing

MetaPro first identifies and removes segments of reads associated with adapters and low-quality bases at the 3' end using AdapterRemoval v2.1.7 [71]. The number of threads used is set to the maximum allowable number of cores the computing environment provides. This is automatically detected by MetaPro. The trimqualities flag is used, which trims the 5'/3' termini of reads with quality scores up to the AdapterRemoval default quality minimum of 2. Other settings rely on default values. Paired-end reads may become separated at this time; if one read is removed, while the other remains. In this instance, the orphaned reads are collected and labeled as singletons. If the data is paired-end, the reads are merged using VSEARCH v2.7.1 [72] with the flag -fastq_mergepairs. Merged reads are moved to the singletons collection. Once merging is complete, the paired-end reads and singletons are filtered for low-quality using VSEARCH with the flags –fastq_filter, –fastq_ascii, and –fastq_maxee. The ascii argument is determined by checking the reads for specific characters, to determine the ascii code offset (33 or 64). If the data is single-end, merging is skipped.

At this stage, MetaPro will have created two categories of data: paired-end reads that were not merged (pair_1 and pair_2) and paired-end reads that successfully merged (singletons). MetaPro next uses VSEARCH again to filter for low-quality reads from paired-end reads and singletons, with the –fastq_filter flag, using the default quality thresholds. Filtering low-quality reads out of paired-end data will result in mismatched reads whereby one read is high-quality, while the other is low quality. These orphaned reads are moved to the singletons by MetaPro using a custom script. Paired-end reads are strictly defined as read pairs featuring both forward and reverse reads. MetaPro then removes duplicate reads from the paired-end and singletons data using CD-HIT v4.6.8 [73], to reduce the data load for subsequent steps.

Next, MetaPro provides the user with an option for filtering contaminant sequences and sequences derived from any host organism that might be associated with the sample (for example if the sample was collected from a mouse or human intestinal sample). This is defined through a configuration option set by the user before launching the pipeline. Though MetaPro lets users choose their own host filter sequences, by default, MetaPro uses mouse and human genome sequences. When filtering paired-end reads, if either paired-end read matches to a host, vector, or rRNA sequence, then by default its paired read would be assigned as host/vector/rRNA accordingly. The user can override this feature to allow both reads of a pair to pass these filters, if at least one of the pair passes, through setting the “filter stringency configuration” option to “low”.

To identify sequence artifacts arising from library preparation, MetaPro utilizes the UniVec_Core dataset comprising known sequencing vectors, sequencing adapters, linkers, and PCR Primers derived from the NCBI UniVec_Core Database [41]. Host organism sequences to be filtered are provided as a single FASTA formatted file by the user and identified through sequence similarity searches using BWA 0.7.17 [38] and pBLAT 2.0 [39]. Due to pBLAT’s inability to support paired-end data, MetaPro further sorts the reads with internal logic. If either one of a pair of reads matches a sequence in the contaminants file, both reads are filtered from downstream analysis. This behavior can be toggled to a more permissive rule where pairs are retained if either one of the pair does not match a sequence in the contaminants file. Host and vector filtering is performed on paired-end reads, and singletons. During this step, paired-end reads may become separated, when either read of a pair (forward or reverse) matches with a host or vector. When this occurs, the orphaned read is moved to the singletons collection.

(See figure on next page.)

Fig. 2 Relative performance of taxonomic classifications assigned by each pipeline. **A** Precision-recall graphs of MetaPro, HUMAnN3, and SAMSA2’s gene annotation strengths against the gold standard hits. True positives are annotations of reads from the pipeline that agree with the gold standard. True negatives are reads that are unidentified by both the pipeline and the gold standard. False negatives are reads that the gold-standard identified but were not annotated by the pipeline. False positives are when the pipeline annotated a read where the gold standard did not. Due to the low number of true positives across all samples and all tools, precision and recall are low, and the plots do not resemble a typical precision-recall plot. **B** Chord diagrams mapping the relationship of reads annotated by the gold-standard and MetaPro for NOD mouse gut and kimchi datasets. Arcs represent categories of reads. Bands between arcs indicate the proportion of reads mapping between categories, for example while many of reads mapped to *Parabacteroides ASF519* are annotated as *Parabacteroidesgoldsteinii* in MetaPro, other reads are assigned by MetaPro to derive from a variety of other organisms. For the kimchi dataset, we observe a higher agreement between MetaPro and gold standard annotations. **C** Pie charts showing the breakdown of taxonomic assignments for each pipeline for a selection of samples. Each chart is constructed only from reads deemed to be of putative bacterial mRNA origin as defined by each pipeline. MetaPro annotates more putative reads to species than either HUMAnN3 or SAMSA2

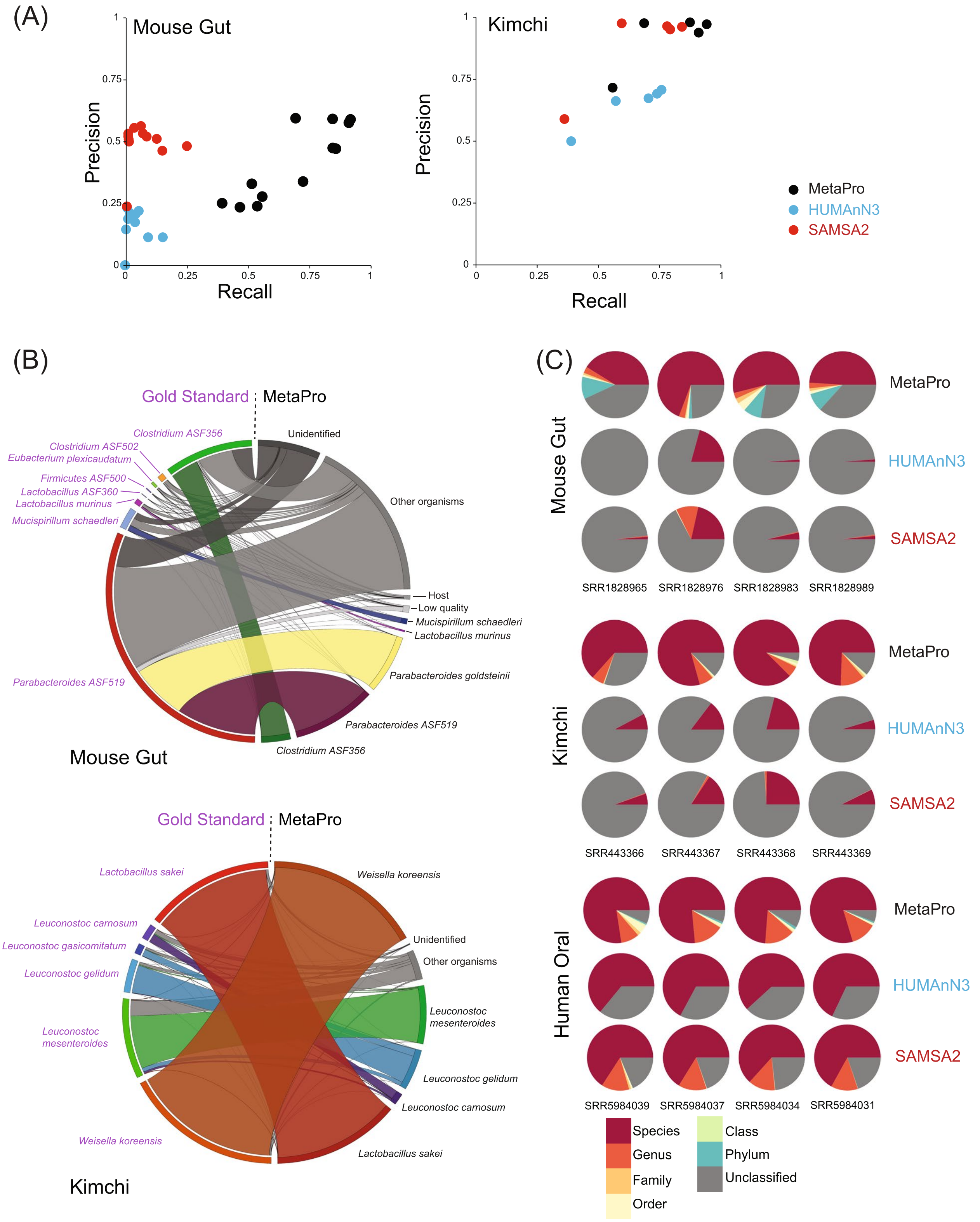

Fig. 2 (See legend on previous page.)

Having identified and removed low quality and contaminating reads, MetaPro next filters for reads of rRNA origin. In a typical RNA extract, 90% or more of reads can be of rRNA origin [74]. While rRNA depletion kits can significantly reduce the proportion of such reads, metatranscriptomic pipelines need to filter for any remaining reads of rRNA origin. rRNA filtering represents one of the most computationally expensive stages of the MetaPro pipeline. To accelerate this step, MetaPro first removes duplicate reads prior to filtering and repopulates the data after filtering. To further increase efficiency, rRNA filtering is first performed by Barrnap, a rRNA filtering tool based on nhmmer [75], and then by Infernal [53], which while more computationally intensive, provides greater sensitivity through the application of covariance models based both on sequence and secondary structure comparisons. To better exploit parallel computing environments, MetaPro splits read data into smaller chunks consisting of 50,000 reads. Each chunk is then processed by Barrnap in parallel, after which the pipeline segregates sequence reads into putative mRNA reads and other reads. The putative mRNA reads are then processed a second time by Infernal, and reads not identified as mRNA are removed. Singletons and paired-end reads are filtered for rRNA. Paired-end reads may become orphans from rRNA filtering. As previous, orphaned reads are moved into the singletons collection.

To improve speed and accuracy of the subsequent annotation steps, the pipeline attempts to assemble reads into longer contiguous sequences ("contigs") using the sequence assembler, rnaSPAdes [36]. Prior to contig assembly, duplicate reads that were removed by CD-HIT are added back in. This is to preserve the read depth of the data, used in assembly. However, only reads that have passed the various filtration steps are repopulated. Assembling reads into contigs reduces the amount of file reading and writing (I/O)-related tasks by shrinking the number of unique segments analyzed. Furthermore, by analyzing longer sequences, the subsequent alignments of the remaining data produce results of higher confidence. Rather than assembling data prior to filtering, we chose to assemble contigs last due to the possibility of assembling contigs with contaminated reads and hence minimize the occurrence of chimeras. We further ensure that assembled reads do not derive from host, rRNA, or tRNA. This was to ensure that we would not mislabel whole sections of reads that may not be a contaminant or annotate something that was contaminated. Preliminary analyses involving assembling the data first were found to have minimal impact on results.

As contigs may represent transcriptional units containing multiple genes, MetaPro applies MetaGeneMark v1 [37], a gene model predictor, to separate contigs into individual putative genes. MetaGeneMark defines sections within the contigs that are predicted to encode a gene. Reads are subsequently mapped to discrete genes through BWA sequence searches. The specific settings are a mismatch penalty (-B) of 40, a gap-open penalty (-O) of 60, a gap-extension penalty (-E) of 10, and a clipping penalty (-L) of 50. At the end of this stage, we have three categories of data: gene models derived from contigs, paired-end reads that did not align with these models, and singletons (filtered merged reads and filtered orphaned reads) that did not align with the models. A map of gene models to their mapped reads is created to track all reads as they travel through the pipeline (contig map).

Sequence preprocessing design rationale

MetaPro was made with several novel design choices that we will explain further. Firstly, we chose to merge the sequence reads prior to host/vector/rRNA filtering, and contig assembly to minimize the occurrence of chimeras, and to optimize the processing times of MetaPro. Though we acknowledge that this may result in lower quality downstream steps, we have found this has minimal impact on the filtering. Based on an initial set of 3.1 million reads generated from the cecal contents of

(See figure on next page.)

Fig. 3 Enzyme annotation performance comparison and example outputs. **A** Stacked barcharts indicate the number of enzymes, as defined through enzyme classification (EC) assignments, annotated by each pipeline for the three sets of datasets: NOD mouse gut, kimchi, and human oral biofilm. In addition to displaying ECs' unique or shared between MetaPro and the other two tools, also shown are ECs, predicted by HUMAnN3 and SAMSA2 to occur in combination with another EC, in the same transcript, with no supporting evidence that such a combination has been previously observed (as defined through Swiss-Prot annotations). Further, for HUMAnN3, we show the number of EC assignments that occur in combinations of greater than 2 ECs. **B** A Cytoscape network representation of the tricarboxylic acid (TCA) cycle, together with the breakdown of EC expression by taxon. MetaPro generates a Cytoscape compatible annotation file which can be used to map gene expression data onto KEGG defined pathways using the KEGGscape and enhancedGraphics plugin applications. Here, each node represents an individual enzyme with its size indicating the overall expression of that enzyme in the dataset (as defined by RPKM). Colored sectors indicate the contribution of each taxon to expression of that enzyme. To simplify the display, taxa were manually merged into 8 taxonomic groupings. Here, we see that members of *Klebsiella* and *Citrobacter* are the main contributors to the TCA cycle. These outputs were generated from a sample in the human oral biofilm dataset (SRR5984039). **C** A summary overview showing the contribution of major taxa to KEGG-defined superpathways. After annotation of enzymes, MetaPro generates a heatmap in PNG format showing taxa responsible for expression (calculated as the sum of RPKMs assigned to each EC annotated for each taxon). Only taxa associated with at least 1% of total reads are shown (see the "Methods" section)

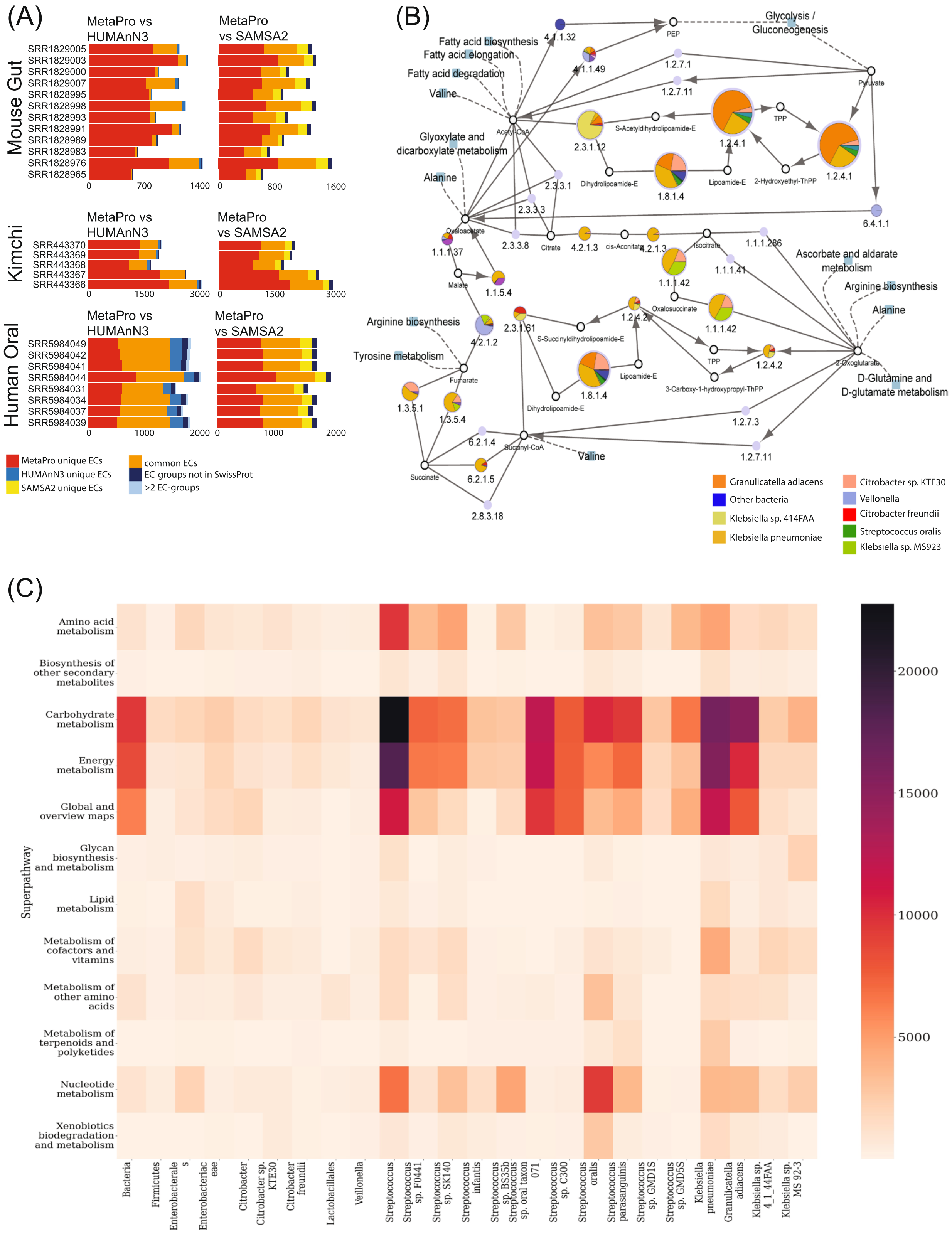

Fig. 3 (See legend on previous page.)

a non-obese diabetic (NOD) mouse (SRR1828998), we found that this new order identified only 13,179 additional host reads, representing 0.42% of the dataset.

Secondly, we have chosen to assemble the sequence reads after filtering to avoid the inclusion of contaminant reads in the assembly. We tested an alternate data workflow (Scenario B: (1) filtering for adaptors and low quality, (2) assembly of reads into contigs, (3) filtering of host and vector contaminants, and (4) filtering of rRNA) against the current workflow (Scenario A: (1) filtering for adaptors and low quality, (2) removal of sequence duplicates, (3) filtering of host and vector contaminants, (4) filtering of rRNA, and (5) adding back sequence duplicates) on the NOD mouse dataset SRR1828998 of 3.1 million paired reads. In this test, scenarios A and B identified 893,895 and 556,362 reads as rRNA/tRNA, respectively. Of these, 552,574 reads were common to both scenarios, indicating that Scenario A missed 3788 rRNA/tRNA reads picked up by Scenario B, while Scenario B missed 341,321 rRNA/tRNA reads picked up by Scenario A. This analysis reveals that the assembly of rRNA/tRNA reads into contigs can result in their misidentification in downstream analyses. Next, we examined the presence of polycistronic reads. We scanned the assembled paired-end reads against the MetaGeneMark genes and tallied the number of paired-end reads that aligned to different genes. We only considered paired-end reads that had successful alignments on both pairs. Though they do exist, we have found that for the NOD mouse dataset, only 0.2–7.84% of all assembled paired-end reads exhibited discordant alignments (Supplemental Table 2). Nevertheless, it is appreciated that polycistronic transcripts have the potential to contribute to the co-expression of neighboring (and likely functionally related) genes. While outside the scope of the current study, we expect future iterations of MetaPro will consider the impact of polycistronic transcripts.

Gene annotation

During the annotation step, all data are processed in the same way. The forward and reverse unassembled paired-end data are annotated together. Unassembled singletons are annotated separately, as are contigs. Prior to assembly, each piece of data is split into smaller chunks in an effort to distribute the workload across all CPUs on the system. Each chunk of data is annotated separately using BWA 0.7.17 [38] against the ChocoPhlAn [26] database from HUMAnN3. Once the annotations for BWA have finished for all data, MetaPro interprets the reports, and filters the reads that have been annotated based on that report using a post-processing program we designed. In this post-processing step, the BWA report (samfile) is parsed, and files containing unannotated reads will be created based on the samfile. To determine whether a read has been annotated, the CIGAR string is used with a 90% alignment threshold chosen as a cutoff by deafult. While this setting can be altered by the user, our prior experience suggests that 90% offers a good compromise between stringency and flexibility to account for sequencing errors and strain differences. In addition to unannotated reads, MetaPro also exports a gene-to-read map for every gene and read that was annotated. The reads not aligned by BWA are then sent through pBLAT.

MetaPro will filter the reads annotated by pBLAT [39] against ChocoPhlan again to retrieve unidentified reads, which are subsequently annotated through DIAMOND [40] comparisons to the non-redundant protein database (NR). When DIAMOND is finished on all remaining reads, MetaPro will filter the annotated reads, resulting in unidentified reads that failed to be annotated by either BWA, pBLAT, or DIAMOND. Each step also creates a separate gene-to-read map. pBLAT and DIAMOND share the same cutoff schemes, based on three values extracted from the m8 report: percent-identity, alignment length, and bitscore. The alignment length is used in conjunction with the sequence length, to calculate an alignment percentage. The criteria for pBLAT and DIAMOND to declare a read's match acceptable is as follows: (1) if either the percent-identity score falls below the threshold of 85%), (2) the alignment length percentage falls below the threshold of 65%, or (3) the bitscore is below the threshold of 60, the match is rejected. All thresholds are capable of being overridden by the user in the configuration settings.

We make the assumption that the forward read and the reverse read represent the same piece of datum (they derive from a single transcript). We also assume that there is a single best match for this datum (determined through quality scores derived from the annotation tool). Therefore, it is possible for the tools to declare multiple acceptable matches on a single read in a report, due to paired-end reads sharing the same sequence read identification, as well as multiple hits from the alignment tools. To resolve these differences, MetaPro will use different metrics depending on the tool. If there are multiple hits in BWA, MetaPro will consider the alignment score of each match. The highest score becomes the match for the read. In events of a tie, the first match in the report will be used. A read match's alignment score must exceed the previous best to be used. In pBLAT and DIAMOND, the bitscore will be used in the same way.

Enzyme annotation

After annotating putative mRNA transcripts to genes, then converting those genes to proteins, MetaPro will attempt to annotate both the translated genes and the proteins identified by DIAMOND searches of NR, to enzymatic functions, as defined by Enzyme Classification (EC) identifiers. This involves a combination of the enzyme profile tools, DETECT [42] and PRIAM [43], together with DIAMOND searches against the Swiss-Prot database [59]. The results are combined by prioritizing the annotation from DETECT and, in cases where enzymatic function could not be assigned using DETECT, enzymatic functions that are consistently assigned by both DIAMOND and PRIAM. This process may assign multiple ECs per transcript. Proteins with multiple EC assignments are further filtered to include only Ecs that have been observed to co-occur in the Swiss-Prot database. MetaPro exports two EC reports; low-quality and high-quality. Both EC reports will contain all results from DETECT together with the intersection of predictions from PRIAM and DIAMOND, with an *e* value cutoff of $< 1e-5$ for the low-quality report, and an *e* value cutoff of $< 1e-10$, together with a PRIAM probability score of ≥ 0.5 for the high-quality report.

Taxonomic annotation

The MetaPro pipeline uses a consensus of three strategies to assign reads to taxonomic classifications. Firstly, the taxonomic identification numbers ("taxid") for each gene/protein for which putative mRNA reads were previously annotated are retrieved based on accession numbers derived from the NCBI accession2taxid database [41]. Secondly, all putative mRNA sequences are processed by the short-read classifiers Kaiju [45] and Centrifuge [46]. Taxonomic classifications from each source (NCBI lookup, Kaiju and Centrifuge) are then merged, using the classification consensus tool, WEVOTE [47]. Since we expect NCBI lookup-based annotations to be of higher quality, MetaPro assigns greater weight to the NCBI lookup assignments (60% NCBI lookup, 20% Centrifuge, and 20% Kaiju). This effectively results in the use of Kaiju and Centrifuge to cover potential gaps in annotation.

Output and visualizations

Read metrics and data quality

MetaPro compiles various read metrics within a text file. This file reports the total number of sequence reads from the raw input, the remaining number of high-quality reads after filtering for quality and adaptors, the number of reads associated with the host (if applicable), the amount of rRNA and tRNA reads removed from the rRNA filtration step, the number of putative mRNA reads, the number of reads annotated to a gene or protein, the total number of unique transcripts found, and the number of unique enzymes detected in the data. MetaPro also creates a histogram of the number reads and their quality scores, before and after they have been filtered for low-quality reads. Finally, the pipeline reports on the N50 and L50 read statistics of the contigs formed during the assembly stage.

Gene expression

MetaPro produces a collection of genes, along with their constituent reads, summarized in a gene/protein-to-read map. This gene/protein map labels each gene (ChocoPhlAn) or protein (NR), along with all the read IDs of the input data that annotate to that gene/protein. MetaPro additionally provides a table of genes/proteins, associated EC, and RPKM, as well as another table of genes/proteins, with their full taxonomy. MetaPro also produces a contig-to-read map that shows the reads associated with each contig generated by rnaSPAdes.

Enzyme annotation

To add context to the enzyme annotation results, MetaPro exports a table file that is compatible with Cytoscape [76] import functions. The table provides a list of Ecs, and the expression (as measured by RPKM) of the various genes and proteins that identify to that EC. EC expression is further broken down by taxa, selected on the basis of minimal abundance. Specifically for each taxon, starting at the rank of species, if that taxon is not associated with at least 1% of total reads (default cutoff), then it is merged with other taxa that share the genus that also do not exceed the 1% criteria. This process is repeated to define groups of taxa at the level of family, order, class, and phyla that represent at least 1% of total reads. Together with the installation of two Cytoscape apps (KEGGscape and enhanceGraphics), KEGG-defined metabolic pathways can be imported as KGML files and enzymes in the pathways annotated with the MetaPro-defined EC file. In typical applications, this can result in the representation of each enzyme as a pie chart or annular ring, in which the size of the pie chart/ring represents total expression (as defined by RPKM values) of that EC in the dataset and the segments indicate the taxa responsible for contributing to the expression of that enzyme. In addition, MetaPro also generates a table of ECs, grouped together by super-pathways (EC_coverage.csv) and a table of super-pathways, and their constituent ECs, expressed as RPKM for the formation of the EC super-pathway heatmap image.

Taxonomic annotation

MetaPro reports the taxonomic findings of the sample as a table of taxa and read pair counts. In brief, annotated genes and proteins are assigned to and grouped on the basis of taxonomic assignment. Reads assigned to each gene/protein that belongs to a specific taxon are summed to provide to produce a read pair count for each taxa.

Computing considerations

Since cluster computing environments can impose limitations on processing time available to the user, MetaPro was developed to improve data throughput through parallelization and ensure the completion of the pipeline through the implementation of an auto-resume feature. To address the former, for the rRNA filtration and gene annotation steps, we deploy a technique based on MapReduce [77] in which datasets are divided into smaller chunks of 50,000 sequences that are processed on parallel threads. Other steps were found not to benefit from splitting the dataset and are simply run as serial processes on the global dataset. Since the processing time required by MetaPro is typically unknown before runtime, to ensure that the processing of a dataset is completed, particularly in computing environments where allocations may be time limited, MetaPro features a multi-level auto-resume feature. This is enabled through a robust bookmarking system that keeps track of the state of processing prior to termination due to time constraints. Subsequent restarts allow MetaPro to resume processing on the remaining sharded data (i.e., for the rRNA filtering and gene annotation steps). For other steps, MetaPro will restart from the beginning of the stage that failed.

Performance scaling

MetaPro's ability to scale in performance based on available resources comes in the form of coordinating the flow of parallel job launches of the third-party tools. For example, the gene annotation steps are all launched sequentially, based on the available hardware on the system. MetaPro monitors resource usage and controls the traffic of job launches until there a free processor and enough free RAM to ensure a safe execution. The current iteration of this mechanism has user-defined interval check times. Subsequent iterations will have automatically calibrated delays. Currently, only rRNA removal and gene annotation identified as our largest bottlenecks use this controller.

Interfacing with a job scheduler

Singularity and Docker are software that manage software environments (system configurations) for deploying onto a single computer. This software is typically used in conjunction with Kubernetes (a container orchestration software) in datacenters with multiple machines clustered together. MetaPro is set up through Singularity to capitalize on this ease of deployment. Singularity has two modes: interactive mode (where users deploy the environment and interact with it directly) or execution mode (where users can access components within the environment through automated means, without human intervention). To interface with a job scheduler (an automated method), users would use Singularity's execution mode to access MetaPro as a single command. MetaPro can only be used through the command-line-interface.

Initial set-up

To run MetaPro using the default libraries, a minimum of 707 GB of disc space is required. This is to store the reference databases and all associated index files for the various tools that are used. To run MetaPro with a sample dataset, a minimum disc space of 6 times the size of your data is needed. This space is to allow MetaPro to split and store interim data for faster processing times.

MetaPro features configurable parameters for many features that impact processing runtime. In a master configuration file, users define the locations of the library and reference file paths required by the various tools employed by the pipeline. The configuration file also includes a series of settings to define stringency levels of filters and displays such as the cutoff for defining taxa to be reported in the enzyme annotation output file, the CIGAR match length cutoff for BWA gene annotation, and the identity, length, and score cutoffs for accepting pBLAT and DIAMOND gene annotations. The configuration file lets users control the efficiency of the parallelization, by adjusting the size of the data chunks processing in the rRNA filtration and gene annotation steps. The user can also control the number of concurrent processes running at each step, and the memory allocated for rRNA filtering, and gene annotation. Finally, the configuration file lets users control the state of the interim files; interim files can be retained intact, compressed, or removed to save disk space.

Database expansion

Datasets and reference database sizes have exponentially expanded and are expected to continue this trajectory as the field grows. MetaPro's scalability was designed to keep up with this pace by basing its performance on the available computing resources, while offering the most comprehensive coverage in annotation. However, the sizes of databases and state-of-the-art computing still necessitate computation times on the range of days, if not weeks. We are currently investigating methods to pre-filter databases to reduce MetaPro's workload on any given sample. The strategy is to create curated database subsets for MetaPro using a preliminary scan of a taxonomic classification tool, similar to the design of HUMAnN3.

Read accounting for HUMAnN3 associated with paired-end data

While SAMSA2 and MetaPro provide transparency in terms of read accounting, mapping the assignment of individual reads in paired-end data remains a challenge for HUMAnN3 processed datasets as duplicate read IDs (representing the forward and reverse reads) can occur in multiple categories. To identify the fate of individual reads we therefore performed the following steps. First, we collected read IDs for each step from Kneaddata—*rRNA*, *host*, *putative mRNA*, *post-trimmed reads* (for quality and adaptor), and *post-repeat-trimmed* reads (for repeats), as well as the *read-taxon map* generated through the bowtie comparisons performed by HUMAnN3. From these ID mappings, we then defined (1) rRNA reads as the set of *rRNA* read IDs subtracting out the read IDs common to *host* or *putative mRNA*; (2) putative mRNA reads as the set *of putative mRNA* read IDs subtracting out the read IDs common to *rRNA* or *host*; (3) host reads as the set of *host* read IDs subtracting out the read IDs common to *rRNA* or *putative mRNA*; (4) trimmed reads as the set of raw (input) read IDs subtracting out the set of *post-trimmed read* IDs; (5) repeated reads as the set of *post-trimmed* read IDs subtracting out the set of *post-repeat-trimmed* read IDs; (6) annotated reads as the set of read IDs associated with the rRNA reads defined by (1) that overlap with read IDs present in the *read taxon* map; (7) unclassified reads as putative mRNA reads subtracting out read IDs common to annotated reads; and (8) ambiguous reads as read IDs that are found in two locations (host and rRNA, rRNA, and putative mRNA, and host and putative mRNA).

Availability and requirements

Project name: MetaPro.

Project home page: https://github.com/ParkinsonLab/MetaPro

Operating system(s): Platform independent.

Programming language: Python3.

Other requirements: MetaPro is deployed within Docker software deployment infrastructure [27].

License: GNU general public license v3 / The user will need to obtain their own copy of MetaGeneMark's license key.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40168-023-01562-6.

Additional file 1: Fig. S1. Read relationships of HUMAnN3 and SAMSA2 against The Gold standard. Chord diagrams showing the relationships of the gold-standard annotated reads of the NOD mouse gut (A) and kimchi datasets (B) as they are processed by the HUMAnN3 and SAMSA2 pipelines. Each arc of the diagram is a category of reads. Each band joining 2 arcs represents the proportion of reads that map between categories. Many of the unidentified mouse gut reads in HUMAnN3 were identified as *Parabacteroides ASF519*. Similar to MetaPro, SAMSA2 identified a portion of the *Parabacteroides ASF519* gold-standard reads to be *Parabacteroides goldsteinii*. In the kimchi set, HUMAnN3 has more branching chords than MetaPro. SAMSA2 has a larger proportion of reads in *Other organisms* compared to MetaPro. In an ideal scenario, there would be a 1:1 relationship between the Gold-standard and the pipelines.

Additional file 2: Fig. S2. Gene annotation performance of MetaPro, HUMAnN3, and HUMAnN2. Stacked barcharts depicting the number of reads annotated to specific taxa in (A) NOD mouse samples, and (B) Kimchi samples by BWA alignments, MetaPro, HUMAnN3, and HUMAnN2. The NOD mouse datasets were generated from gut samples from mice inoculated with a defined microbial consortium (Altered Schaedler Flora (ASF); [29]). In addition to the 8 taxa associated with ASF, reads were also assigned to *Parabacteroides goldsteinii,* a close relative of *Parabacteroides ASF519* (see legend). The kimchi datasets comprise five major taxa (see legend; [30–34]). It should be noted that *Leuconostoc gasicomitatum* reported in the original publication is currently classified as a subspecies of *Leuconostoc gelidum*. For NOD sample SRR1828965, HUMAnN3 did not annotate any reads; for kimchi sample SRR443366, HUMAnN2 did not annotate any reads.

Additional file 3: Fig. 3. Taxonomic classification performance of MetaPro, HUMAnN3, and HUMAnN2. For the NOD mouse (A) and Kimchi (B) datasets, each pie chart shows a breakdown of taxonomic assignments at the different taxonomic levels indicated, that are closest to the last common ancestor of the expected bacteria within the sample. For the human oral datasets (C), given the lack of gold standard assignments, each pie chart represents the relative abundance of reads assigned to different taxonomic levels. Unclassified reads represent annotated reads with no assigned taxon. The graphs indicate a substantial improvement in HUMAnN3's annotation abilities over HUMAnN2.

Additional file 4: Fig. S4. Enzyme annotation performance of MetaPro and HUMAnN2. Stacked barcharts indicate the number of enzymes, as defined through enzyme classification (EC) assignments, annotated by each pipeline for the three sets of datasets: (A) NOD mouse, (B) kimchi, and (C) human oral biofilm data. In addition to displaying ECs unique or shared between MetaPro and HUMAnN2, also shown are ECs, predicted by HUMAnN2 to occur in combination with another EC, in the same transcript, with no supporting evidence that such a combination has been previously observed (as defined through Swiss-Prot annotations). Further, for HUMAnN2, we show the number of EC assignments that occur in combinations of three or more ECs.

Additional file 5: Table S1. Summary of Sequence Read Processing for Three Metatranscriptomic Datasets (NOD Mouse gut; Kimchi and Human Oral Biofilm) Processed by HUMAnN3, HUMAnN2, MetaPro and SAMSA2. This table reports the processing results from the four pipelines on samples from three different datasets. HUMAnN2 and HUMAnN3's preprocessing tool concatenates paired reads into 1 single file and treats them as 2 separate reads. The NOD mouse samples are paired-end data, while the kimchi and human oral biofilm represent single-end sequence datasets. Unlike MetaPro and SAMSA2, HUMAnN3 and HUMAnN2 do not report transcripts but instead group proteins identified in their pipelines into gene families that are reported in the final column.

Additional file 6: Table S2. Polycistronic read statistics for NOD mouse. This table shows the tally of non-overlapping paired-end reads that were assembled into contigs by MetaPro through rnaSPADes and subsequently annotated into discrete genes by MetaGeneMark. This table also shows the prevalence of polycistronic reads that exist within the data. BWA was used to align the assembled paired-end reads against the genes to identify discordant alignments between the forward and reverse-end read of a pair with the same ID. This table has seven columns: 1) the sample ID. 2) the sample description. 3) the total number of alignments is the number of alignments of a read to a gene that BWA reported. 4) The total number of pairs is the number of IDs that BWA aligned, be it forward, reverse, or both paired-end reads. 5) The paired-end disagreements column are the number of times a forward-end and reverse-end read had different alignments for each NOD mouse sample. 6) The paired-end agreements column shows the number of times a forward-read and reverse-end read aligned to the same gene. 7) The percentage of paired-end disagreements, relative to the total number of paired-end reads in the sample. The percentage of disagreements (polycistronic reads) are at-best 0.23%, and at-worst 7.8% of assembled, non-overlapped paired-end reads in the NOD mouse samples.

Additional file 7: Table S3. Read annotation statistics for NOD mouse datasets from MetaPro, HUMAnN3, HUMAnN2, SAMSA2 compared with the gold standard. This table shows the number of reads in each NOD mouse sample each pipeline assigned to the 8 ASF bacteria: *Clostridium ASF356, Clostridium ASF502, Eubacterium plexicaudatum, Firimicutes ASF500, Lactobacillus ASF360, Lactobacillus murinus, Mucisprillim schaedleri, and Parabacteroides ASF519*. Due to the similarity *between P. ASF519*, and *P. goldsteinii*, the pipelines will sometimes annotate to *P. goldsteinii* rather than to *P.ASF519*. *P. goldsteinii* was also a dominant species found within the samples outside of the 8 ASF bacteria. The expected results were produced by annotating the reads with a reference containing only the 8 ASF, using BWA.

Additional file 8: Table S4. Read annotation statistics for kimchi fermentation datasets from MetaPro, HUMAnN3, HUMAnN2, SAMSA2, compared with the gold standard. This table shows the number of reads in each kimchi sample, annotated to the expected 5 lactic acid bacteria (LAB) from each pipeline: *Leuconostic mesenteroides*, *Lactobacillus sakei*, *Weissella koreensis*, *Leuconostoc carnosum*, and *Leuconostoc gelidum*. The expected results were obtained by annotating the kimchi datasets against a database containing only the reference gene sequences for the 5 LAB, using BWA.

Additional file 9: Table S5. Comparisons of Enzyme annotations between MetaPro and HUMAnN3 for NOD mouse, kimchi, and human oral biofilm datasets. This table compares the ECs of MetaPro against HUMAnN3 on the NOD mouse, kimchi, and human oral biofilm datasets. The HUMAnN3 ECs were filtered for EC co-occurrence pairs that were not found in Swiss-Prot, and multiple unique ECs that annotated to the same gene family. The resulting HUMAnN3 ECs were contrasted against MetaPro's EC, yielding a common set of ECs found in both tools, ECs found only by MetaPro, and ECs found only by HUMAnN3. The same comparison is shown for MetaPro's high-quality EC predictions.

Additional file 10: Table S6. Comparisons of Enzyme annotations between MetaPro and SAMSA2 for NOD mouse, kimchi, and human oral biofilm datasets. This table compares the ECs of MetaPro against SAMSA2 on the NOD mouse, kimchi, and human oral biofilm datasets. The SAMSA2 ECs were filtered for EC co-occurrence pairs that were not found in Swiss-Prot, and multiple unique ECs that annotated to the same gene family. The resulting SAMSA2 ECs were contrasted against MetaPro's EC, yielding a common set of ECs found in both tools, ECs found only by MetaPro, and ECs found only by SAMSA2. The same comparison is shown for MetaPro's high-qaulity EC predictions.

Additional file 11: Table S7. Comparisons of Enzyme annotations between MetaPro and HUMAnN2 for NOD mouse, kimchi, and human oral biofilm datasets. This table compares the ECs of MetaPro against HUMAnN2 on the NOD mouse, kimchi, and human oral biofilm datasets. The HUMAnN2 ECs were filtered for EC co-occurrence pairs that were not found in Swiss-Prot, and multiple unique ECs that annotated to the same gene family. The resulting HUMAnN2 ECs were contrasted against MetaPro's EC, yielding a common set of ECs found in both tools, ECs found only by MetaPro, and ECs found only by HUMAnN2. The same comparison is shown for MetaPro's high-qaulity EC predictions.

Additional file 12: Table S8. Computational performance statistics of MetaPro, HUMAnN3, and SAMSA2. This table reports the amount of processing time required for each run of the three pipelines. MetaPro additionally exports the timing data of each stage independently. HUMAnN3's pre-processing step is a separate stage using a separate tool called KneadData. SAMSA2 cleans the data in the pipeline, but it is integrated and does not export timing data.

Acknowledgements

We would like to thank Ana Popovic, Emil Jurga, James St. Pierre, and Donny Chan for their enthusiastic help in pipeline testing and providing valuable feedback on pipeline implementation. We would also like to thank Ben Jackel for his advice on analyzing output generated by HUMAnN2 and HUMAnN3.

Authors' contributions

BT, MA and XX developed the pipeline. NN, JA and JP helped in the design of the pipeline. BT, MA and JP wrote the main manuscript. JP conceived the project. The author(s) read and approved the final manuscript.

Funding

This work was supported through funding from the Natural Sciences and Engineering Research Council (RGPIN-2019-06852), the Canadian Institutes for Health Research (MRT-168043), the University of Toronto's Medicine by Design initiative which receives funding from the Canada First Research Excellence Fund, the Alberta Livestock and Meat Agency, the Ontario Ministry of Agriculture, Food and Rural Affairs and the Canadian Poultry Research Council. NN was supported by a student Restracomp fellowship given by the Hospital for Sick Children. Computing resources were provided by the SciNet HPC Consortium. SciNet is funded by the Canada Foundation for Innovation under the auspices of Compute Canada, the Government of Ontario, Ontario Research Fund—Research Excellence, and the University of Toronto. The funders had no role in the study design, data collection, and analysis; decision to publish; or preparation of the manuscript.

Availability of data and materials

The software is freely available under the GNU public license V3 and can be accessed through https://github.com/ParkinsonLab/MetaPro.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Program in Molecular Medicine, The Hospital for Sick Children, Toronto, ON M5G 0A4, Canada. [2]Department of Chemical and Physical Sciences, University of Toronto, Mississauga, ON L5L 1C6, Canada. [3]Department of Computer Science, University of Toronto, Toronto, ON M5S 3G4, Canada. [4]Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 3G4, Canada. [5]Department of Biochemistry, University of Toronto, Toronto, ON M5S 3G4, Canada.

Received: 24 December 2022 Accepted: 28 April 2023
Published online: 27 June 2023

References

1. Lee W-J, Hase K. Gut microbiota–generated metabolites in animal health and disease. Nat Chem Biol. 2014;10:416.
2. Mendes R, Garbeva P, Raaijmakers JM. The rhizosphere microbiome: significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms. FEMS Microbiol Rev. 2013;37:634–63.
3. Huttenhower C, et al. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207.
4. Kuhn KA, Pedraza I, Demoruelle MK. Mucosal immune responses to microbiota in the development of autoimmune disease. Rheum Dis Clin North Am. 2014;40:711–25. https://doi.org/10.1016/j.rdc.2014.07.013.
5. Frank DN, et al. Disease phenotype and genotype are associated with shifts in intestinal-associated microbiota in inflammatory bowel diseases. Inflamm Bowel Dis. 2011;17:179–84. https://doi.org/10.1002/ibd.21339.

6. Li, E. *et al.* Inflammatory bowel diseases phenotype, C. difficile and NOD2 genotype are associated with shifts in human ileum associated microbial composition. *PLoS One* 7, e26284, doi:https://doi.org/10.1371/journal.pone.0026284 (2012).
7. Markle JG, Frank DN, Adeli K, von Bergen M, Danska JS. Microbiome manipulation modifies sex-specific risk for autoimmunity. Gut Microbes. 2014;5:485–93. https://doi.org/10.4161/gmic.29795.
8. Alkanani AK, et al. Alterations in intestinal microbiota correlate with susceptibility to type 1 diabetes. Diabetes. 2015;64:3510–20. https://doi.org/10.2337/db14-1847.
9. Hara N, et al. The role of the intestinal microbiota in type 1 diabetes. Clin Immunol. 2013;146:112–9. https://doi.org/10.1016/j.clim.2012.12.001.
10. Zhang, Y. *et al.* Identification of candidate adherent-invasive E. coli signature transcripts by genomic/transcriptomic analysis. *PLoS One* 10, e0130902, doi:https://doi.org/10.1371/journal.pone.0130902 (2015).
11. Langille MG, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol. 2013;31:814.
12. Aßhauer KP, Wemheuer B, Daniel R, Meinicke P. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. Bioinformatics. 2015;31:2882–4.
13. Xiong, X. *et al.* Generation and analysis of a mouse intestinal metatranscriptome through Illumina based RNA-sequencing. *PLoS One* 7, e36009, doi:https://doi.org/10.1371/journal.pone.0036009 (2012).
14. Weckx S, et al. Metatranscriptome analysis for insight into whole-ecosystem gene expression during spontaneous wheat and spelt sourdough fermentations. Appl Environ Microbiol. 2011;77:618–26. https://doi.org/10.1128/AEM.02028-10.
15. Jiang Y, Xiong X, Danska J, Parkinson J. Metatranscriptomic analysis of diverse microbial communities reveals core metabolic pathways and microbiome-specific functionality. Microbiome. 2016;4:2. https://doi.org/10.1186/s40168-015-0146-x.
16. Qin J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010;464:59–65. https://doi.org/10.1038/nature08821.
17. Human Microbiome Project, C. Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214, doi:https://doi.org/10.1038/nature11234 (2012).
18. Smith MI, et al. Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. Science. 2013;339:548–54. https://doi.org/10.1126/science.1229000.
19. Backhed F, et al. Dynamics and stabilization of the human gut microbiome during the first year of life. Cell Host Microbe. 2015;17:690–703. https://doi.org/10.1016/j.chom.2015.04.004.
20. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57–63. https://doi.org/10.1038/nrg2484.
21. Meyer F, et al. The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. 2008;9:386.
22. Ni Y, Li J, Panagiotou G. COMAN: a web server for comprehensive metatranscriptomics analysis. BMC Genomics. 2016;17:622.
23. Westreich ST, Treiber ML, Mills DA, Korf I, Lemay DG. SAMSA2: a standalone metatranscriptome analysis pipeline. BMC Bioinformatics. 2018;19:175.
24. Narayanasamy S, et al. IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. Genome Biol. 2016;17:260.
25. Martinez X, et al. MetaTrans: an open-source pipeline for metatranscriptomics. Sci Rep. 2016;6:26447.
26. Franzosa EA, et al. Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods. 2018;15:962.
27. Boettiger C. An introduction to Docker for reproducible research. ACM SIGOPS Oper Syst Rev. 2015;49:71–9.
28. Kurtzer GM, Sochat V, Bauer MW. Singularity: scientific containers for mobility of compute. PLoS One. 2017;12: e0177459.
29. Orcutt R, Gianni F, Judge R. Development of an "Altered Schaedler Flora" for NCI gnotobiotic rodents. Microecol Ther. 1987;17:59.
30. Kim J, Chun J, Han HU. Leuconostoc kimchii sp. nov., a new species from kimchi. Int J Syst Evol Microbiol. 2000;50:1915–19.
31. Kim N, Park K-R, Park I-S, Cho Y-J, Bae YM. Application of a taste evaluation system to the monitoring of Kimchi fermentation. Biosens Bioelectron. 2005;20:2283–91.
32. Lee J-S, et al. Analysis of kimchi microflora using denaturing gradient gel electrophoresis. Int J Food Microbiol. 2005;102:143–50.
33. Park JA, et al. Change of microbial communities in kimchi fermentation at low temperature. Korean J Microbiol. 2003;39:45–50.
34. Park J-M, et al. Identification of the lactic acid bacteria in kimchi according to initial and over-ripened fermentation using PCR and 16S rRNA gene sequence analysis. Food Sci Biotechnol. 2010;19:541–6.
35. Celaj A, Markle J, Danska J, Parkinson J. Comparison of assembly algorithms for improving rate of metatranscriptomic functional annotation. Microbiome. 2014;2:39. https://doi.org/10.1186/2049-2618-2-39.
36. Bushmanova E, Antipov D, Lapidus A, Przhibelskiy AD. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. GigaScience. 2019;8:giz100.
37. Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. Nucleic Acids Res. 2010;38:e132–e132.
38. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26:589–95.
39. Wang M, Kong L. pblat: a multithread blat algorithm speeding up aligning sequences to genomes. BMC Bioinformatics. 2019;20:28.
40. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12:59.
41. Coordinators NR. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2018;46:D8.
42. Nursimulu N, Xu LL, Wasmuth JD, Krukov I, Parkinson J. Improved enzyme annotation with EC-specific cutoffs using DETECT v2. Bioinformatics. 2018;34:3393–5.
43. Claudel-Renard C, Chevalet C, Faraut T, Kahn D. Enzyme-specific profiles for genome annotation: PRIAM. Nucleic Acids Res. 2003;31:6633–9.
44. Consortium, U. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2018;47:D506–15.
45. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun. 2016;7:11257.
46. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 2016;26:1721–9.
47. Metwally AA, Dai Y, Finn PW, Perkins DL. WEVOTE: weighted voting taxonomic identification method of microbial sequences. PLoS One. 2016;11:e0163527.
48. Jung JY, et al. Metatranscriptomic analysis of lactic acid bacterial gene expression during kimchi fermentation. Int J Food Microbiol. 2013;163:171–9.
49. Edlund A, et al. Uncovering complex microbiome activities via metatranscriptomics during 24 hours of oral biofilm assembly and maturation. Microbiome. 2018;6:217.
50. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9. https://doi.org/10.1038/nmeth.1923.
51. Shi AM, Lauren. KneadData. 2014. https://huttenhower.sph.harvard.edu/kneaddata/.
52. Seemann T. Barrnap: basic rapid ribosomal RNA predictor. 2013. https://github.com/tseemann/barrnap.
53. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29:2933–35.
54. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. 2012;28:3211–7. https://doi.org/10.1093/bioinformatics/bts611.
55. Suzek BE, et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2015;31:926–32. https://doi.org/10.1093/bioinformatics/btu739.
56. Segata N, et al. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods. 2012;9:811–4. https://doi.org/10.1038/nmeth.2066.
57. Truong DT, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods. 2015;12:902.
58. Hung SS, Wasmuth J, Sanford C, Parkinson J. DETECT–a density estimation tool for enzyme classification and its application to Plasmodium falciparum. Bioinformatics. 2010;26:1690–8. https://doi.org/10.1093/bioinformatics/btq266.
59. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 2000;28:45–8.
60. Blazejewski T. et al. Systems-based analysis of the sarcocystis neurona genome identifies pathways that contribute to a heteroxenous life cycle. mBio. 2015;6:e02445–02414. https://doi.org/10.1128/mBio.02445-14.

61. Cotton JA, et al. The genome of Onchocerca volvulus, agent of river blindness. Nat Microbiol. 2016;2:16216. https://doi.org/10.1038/nmicrobiol.2016.216.
62. Xiong X, et al. Perilipin-2 modulates dietary fat-induced microbial global gene expression profiles in the mouse intestine. Microbiome. 2017;5:117.
63. Swapna LS, Molinaro AM, Lindsay-Mosher N, Pearson BJ, Parkinson J. Comparative transcriptomic analyses and single-cell RNA sequencing of the freshwater planarian Schmidtea mediterranea identify major cell types and pathway conservation. Genome Biol. 2018;19:124. https://doi.org/10.1186/s13059-018-1498-x.
64. Coghlan A, et al. Comparative genomics of the major parasitic worms. Nat Genet. 2019;51:163–74. https://doi.org/10.1038/s41588-018-0262-1.
65. Curran DM, et al. Modeling the metabolic interplay between a parasitic worm and its bacterial endosymbiont allows the identification of novel drug targets. eLife. 2020;9:e51850. https://doi.org/10.7554/eLife.51850.
66. Webb, E. C. Enzyme nomenclature 1992. Recommendations of the nomenclature committee of the international union of biochemistry and molecular biology on the nomenclature and classification of enzymes. San Diego: Academic Press; 1992. pp. 863.
67. Ye Y, Doak TG. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. PLOS Comput Biol. 2009;5:e1000465. https://doi.org/10.1371/journal.pcbi.1000465.
68. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2006;35:D61–5.
69. Overbeek R, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 2005;33:5691–702. https://doi.org/10.1093/nar/gki866.
70. Merkel D. Docker: lightweight linux containers for consistent development and deployment. Linux J. 2014;2014:2.
71. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes. 2016;9:88.
72. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4: e2584.
73. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150–2. https://doi.org/10.1093/bioinformatics/bts565.
74. Stewart FJ, Ottesen EA, DeLong EF. Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. ISME J. 2010;4:896.
75. Wheeler TJ, Eddy SR. nhmmer: DNA homology search with profile HMMs. Bioinformatics. 2013;29:2487–9.
76. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2010;27:431–32.
77. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. Commun ACM. 2008;51:107–13. https://doi.org/10.1145/1327452.1327492.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.