Pre- and post-sequencing recommendations for functional annotation of human fecal metagenomes

Michelle L. Treiber
University of California Davis

Diana H. Taft
University of California Davis

Ian Korf
University of California Davis

David A. Mills
University of California Davis

Danielle G. Lemay
USDA ARS Western Human Nutrition Research Center and University of California Davis Genome Center
https://orcid.org/0000-0003-3318-0485

Research article

Keywords: Metagenomes, metagenomics, fecal, stool, human, functional annotation, function, gene-centric analysis, bioinformatics, microbiome

Version of Record: A version of this preprint was published on February 24th, 2020. See the published version at https://doi.org/10.1186/s12859-020-3416-y.

Abstract

Background Shotgun metagenomes are often assembled prior to annotation of genes which biases the functional capacity of a community towards its most abundant members. For an unbiased assessment of community function, short reads need to be mapped directly to a gene or protein database. The ability to detect genes in short read sequences is dependent on pre- and post-sequencing decisions. The objective of the current study was to determine how library size selection, read length and format, protein database, e-value threshold, and sequencing depth impact gene-centric analysis of human fecal microbiomes when using DIAMOND, an alignment tool that is up to 20,000 times faster than BLASTX. Results Using metagenomes simulated from a database of experimentally verified protein sequences, we find that read length, e-value threshold, and the choice of protein database dramatically impact detection of a known target, with best performance achieved with longer reads, stricter e-value thresholds, and a custom database. Using publicly available metagenomes, we evaluated library size selection, paired end read strategy, and sequencing depth. Longer read lengths were acheivable by merging paired ends when the sequencing library was size-selected to enable overlaps. When paired ends could not be merged, a congruent strategy in which both ends are independently mapped was acceptable. Sequencing depths of 5 million merged reads minimized the error of abundance estimates of specific target genes, including an antimicrobial resistance gene. Conclusions Shotgun metagenomes of DNA extracted from human fecal samples sequenced using the Illumina platform should be size-selected to enable merging of paired end reads and should be sequenced in the PE150 format with a minimum sequencing depth of 5 million merge-able reads to enable detection of specific target genes. Expecting the merged reads to be 180-250bp in length, the appropriate e-value threshold for DIAMOND would then need to be more strict than the default. Accurate and interpretable results for specific hypotheses will be best obtained using small databases customized for the research question.

Background

Measurement of the human fecal microbiome can provide a snapshot of the functional capacity of the microbes of the distal gut. Previous studies have used methods in which a particular marker gene, e.g. 16S rRNA gene, is amplified and sequenced to estimate the taxonomy of microbes present [1]. A method, PICRUSt (phylogenetic investigation of communities by reconstruction of unobserved states), has been developed to predict the functional composition of metagenomes from 16S rRNA marker data and reference genomes [2]. However, in a comparison with full metagenomes, PICRUSt still missed a large percentage of genes known to be present and predicted many genes not found in adequately sequenced metagenomes [3]. The current golden standard remains to sequence all of the DNA present in the sample —the metagenome, also known as the "shotgun" metagenome—and to directly map DNA sequences to protein or gene family databases to infer functional capacity.

One common way of functionally annotating a shotgun metagenome is to first assemble the reads into longer fragments of DNA called contigs, then predict open reading frames in the contigs and map these ORFs to gene family databases. This is valuable because it can provide genomic context for individual

genes. However, when estimating the functional capacity of a community, metagenomics assembly biases in favor of the assembled contigs which have less fragmented gene sequences. This means that abundance estimates are biased in favor of the functions of the most abundant organisms. Assembly has been shown to be detrimental to obtaining true gene abundance estimates, as in the case of antibiotic resistance genes [4]. If the objective of the study is to assess the overall functional capacity of a community, it may not be necessary to put sequences in the context of their individual genomes.

Different bioinformatics strategies have been suggested to functionally annotate shotgun metagenomes (reviewed in [5–7]). Regardless of the bioinformatics strategy used, a fundamental step is the ability to detect genes-of-interest in short read sequencing data now commonly produced via Illumina platforms. At the time of this writing, common formats are single read 50bp or 100bp (SR50, SR100) and paired end 100bp and 150bp (PE100, PE150). Gene detection in short read sequences is impacted by decisions in both the pre- and post-sequencing phases. Pre-sequencing decisions include choices of size-selection during library preparation, read length and format (single or paired end), and sequencing depth. Post-sequencing decisions include selection of pre-processing steps, a reference database, and a mapping strategy. While some guidance has been provided on sequencing [8], recommendations are based largely on practices that best inform taxonomy, rather than function. Metagenome studies have generally been designed with microbial ecology (e.g. taxonomy) in mind as the foremost goal with functional annotation as an afterthought. As a result, pre-sequencing decisions for the estimation of taxonomy may be weighted towards creating non-overlapping paired ends of reads that maximize the ability to fully assemble the metagenome. However, tools that map reads to protein databases, such as BLASTX [9] and DIAMOND [10], are not able to leverage non-overlapping pairs of reads. A systematic assessment of the annotation of short reads demonstrated the general utility of homology-based mapping with the caution that read length, phylogeny, and database coverage impacts accuracy [11]. Since that important work which utilized BLAST [11], a protein alignment tool called DIAMOND was developed which is 20,000 times faster than BLASTX [10]. We therefore sought to identify appropriate e-value thresholds for functional annotation of shotgun metagenomics using DIAMOND for read lengths commonly available via the popular Illumina platform in the current study.

Another important limitation of prior work is that it has not been clear how pre- and post-sequencing decisions impact the ability to quantify abundances of genes with verified functionality. Databases of genes, gene families, or orthologous groups are often based on predictions or homology with few entries being verified for functionality. For example, the carbohydrate-active enzymes database (CAZy) requires only one member of each gene family to be biochemically characterized with the remaining members assigned based on amino acid sequence similarity [12]. It is possible for enzymes to be assigned to more than one family and for different enzymes in the same family to act on different substrates.

In the current study, we evaluated various pre- and post- sequencing choices. We first assembled a database of protein sequences for which the function has been experimentally verified and then used this database of "knowns" to generate simulated metagenomes consisting of sequence reads of proteins with verified functionality. Simulating metagenome sequence reads from this database of "knowns" enabled

us to accurately quantify how pre- and post- sequencing choices affected the number of true positives, false positives, true negatives, and false negatives. We then used these simulated metagenomes to investigate how read length, e-value threshold, and the choice of protein database impacted our ability to detect sequences of known abundance. Next, we used real human stool metagenomes from two different studies to investigate strategies for paired end reads when most reads don't overlap. Finally, we sub-sampled deeply sequenced fecal metagenomes to determine the minimum sequencing depth needed to quantify the abundance of the genes of specific enzymes.

Results

2.1 Evaluation of the effect of read length

We first constructed a database containing only protein sequences with experimentally verified functions (see Methods). Metagenomes were then simulated using sequences from this database with increasing proportions of a target enzyme, beta-galactosidase, with the non-beta-galactosidase reads being derived from other sequences in our known protein database. One hundred metagenomes were simulated with five different read lengths of 50 bp, 100 bp, 150 bp, 200 bp, and 250 bp for 500 total metagenomes. For each read length, the 100 simulated metagenomes were mapped to a custom beta-galactosidase database, the NCBI RefSeq database [13] or the SEED database [14] using DIAMOND with the "sensitive" flag and default e-value threshold. With a read length of 50 bp, only the custom beta galactosidase database, from which the beta-gal metagenomic reads were simulated, performed well with both high true positive rates and low false positive rates (**Figure 1**). True positive rates were expressed as the proportion of reads known to originate from the target that were correctly identified as the target. False positive rates would ordinarily be expressed as the proportion of reads not originating from the target as being incorrectly identified as the target. However, for metagenomes, such rates are as low as 0.0001, which seems small. However, when expressed as the number of reads per 10 million that were incorrectly identified as being the target (as in Figure 1), this is a more practical perspective. A false positive rate of 0.0001 translates to 1000 counts per 10 million, which is likely intolerable if the target were only 100 counts per 10 million because that target's accurate count of 100 per 10 million would become 1100 per 10 million. This is especially problematic for differential abundance testing. If the target's true abundance doubles with a false positive rate of 0, the fold-change is 2 (200/100), but if the target doubles in abundance with a false positive rate of 0.0001, the fold-change is only 1.09 (1200/1100).

The SEED and RefSeq databases required longer reads for more accurate matching of beta-galactosidase, preferably 200bp and above. Regardless of database choice, true positive rates increase with increasing read length (Figure 1A) but the number of false positives increase with increasing read length as well (Figure 1B). This results in increased sensitivity but decreased specificity with increasing read length (Additional File 1, Figure S1). However, overall accuracy is increased with increasing read length (Additional File 1, Figure S1).

2.2 Selection of appropriate e-value threshold for mapping reads

To evaluate an appropriate DIAMOND e-value threshold for mapping reads of different lengths to a custom database, error rates were calculated for different e-values. We simulated a metagenome of 2,500 beta-galactosidase reads per 100,000 total reads with the non-beta-galactosidase reads being derived from other sequences in our known protein database (see Methods). Metagenomes were simulated at all 5 read lengths: 50, 100, 150, 200, and 250bp. Reads were aligned to the beta-galactosidase database using e-value cutoffs of 1e-3 (default DIAMOND e-value threshold), 1e-10, 1e-25, 1e-50, and 1e-100 (Figure 2). Both true and false positives decrease with more stringent e-values (Figure 2A-B), resulting in lower sensitivity but higher specificity (Additional File 1, Figure S2). For every read length, the default e-value was most accurate for this unique situation where a custom database containing only the gene of interest is used to detect the gene of interest, beta-galactosidase, in the metagenome (Additional File 1, Figure S2).

While the choice to examine beta-galactosidase in our previous experiments was due to our lab's interest in this enzyme, we conducted another series of experiments with a protein sequence selected specifically to maximize the difficulty of distinguishing it from other proteins. Using our custom database of known proteins, we computed all pairwise percent identities of all sequences in the database and chose the protein with the highest percent identity with other proteins in the database. Homoserine O-acetyltransferase (HTA), which had 28 matches of mean 42% identity with other proteins in the database of known proteins, was selected for this experiment. Of the 28 sequences with high pairwise identity in the known protein database, only 14 are HTA and the other 14 proteins are not HTA. The 14 true HTA proteins were used to construct a custom database while all of the known proteins, including the 14 off-target sequences were used to create simulated metagenomes. Using simulated metagenomes containing 350 HTA reads per 100,000 total reads at all 5 read lengths, the metagenomes were aligned to a custom database containing the HTA protein sequences. To detect HTA, the best choice of e-value threshold was dependent on read length. For read lengths 100bp and higher, the best choice of e-value was not the default (Figure 2C-D). For 100-150bp, a threshold of 1-e10 yielded the highest accuracy (Additional File 1, Figure S3). For 200-250bp read lengths, 1e-25 was most accurate (Additional File 1, Figure S3). In general, the optimal e-value threshold increases with increasing read length.

2.3 Choice of protein database

Protein databases differ in their organization, level of resolution, and level of evidence for annotations. To understand how the choice of protein database impacts the accuracy of functional annotation, several databases—Carbohydrate Active Enzyme (CAZy), NCBI nonredundant RefSeq, SEED Subsystems and a customized database containing only beta-galactosidases—were used to annotate metagenomes that were simulated from a database of proteins of experimentally verified function (see Methods). CAZy is a database of sequence-based families of enzymes that assemble, modify and breakdown oligo- and polysaccharides [12]. The NCBI RefSeq database contains a comprehensive set of non-redundant protein sequences [13]. The SEED subsystems database contains collections of functionally related protein families [14]. The customized database contained only beta-galatosidase sequences with previously verified functionality.

One hundred metagenomes of 100,000 reads were simulated from our database of known proteins with increasing doses of beta-galactosidase sequences ranging from 25 to 2,500 sequences for each of the five different read lengths: 50 bp, 100 bp, 150 bp, 200 bp, and 250 bp. Each metagenome was annotated against each database to determine the total number of beta-galactosidase hits. At all read lengths, the custom beta-galactosidase database was closer to the "expected" dose-response than the other databases (**Figure 3**). At 50bp, only the custom database performed well. At higher read lengths, the CAZy database appeared to over-estimate and the RefSeq and SEED subsystems databases appeared to underestimate the number of betagalactosidase reads present.

To better understand why the CAZy database over-estimated beta-galactosidase abundance, we mapped the reads to CAZy families. For the beta-galactosidase dose experiment, some families showed an increasing dose while other families showed a constant number of hits across all doses (**Figure 4**). GH42 had the most obvious dose-response with a more moderate dose-response for GH2. Other GH families with potential beta-galactosidase members had an even response across the metagenomes. This phenomenon was true at all read lengths: 50bp, 100bp, 150bp, 200bp, and 250bp (Additional File 1, Figure S4). Some of the CAZy families which contain beta-galactosidase (GH1, GH2, GH16, GH35, GH42, and GH98) also contain other genes that are not beta-galactosidase. This suggests that there may be off-target hits when using the CAZy database.

2.4 Evaluation of strategies for paired end reads

When mapping paired end reads to a protein database, mapping tools such as DIAMOND are not able to leverage the joint nature of pair-end reads. One must either map the two reads independently or the reads can be merged and then mapped. The extent to which sequences of the two ends can be merged depends on the size selection of the DNA prior to sequencing and the length of the reads. In the current study, we analyzed stool metagenomes from two projects with different read formats and insert sizes. The adult stool metagenomes were sequenced in a 2x101bp format from DNA with an insert size range of 280-320 bp [15]. With the larger insert size, the average percent of read pairs that could be merged in the 30 largest metagenomes used in the present study was 34.6% and the average length of the merged reads was 112 bp. The infant stool metagenomes were sequenced in a 2x151bp format from DNA with an average insert size range 119-289 bp [16]. In these metagenomes, which were deliberately sequenced with a smaller insert size to improve overlaps, we found that the average percent of read pairs that could be merged was 87.9% and the average length of the merged read was 188 bp. Thus, with a custom size selection step during library preparation, it is possibly to vastly improve the number of overlapping reads and the resulting length of the merged reads.

We next evaluated whether an alternative strategy should be used when analyzing shotgun metagenomes in the paired end format with few overlapping paired reads. Other options besides merging reads included using only one of the read pairs (referred to here as the "R1 strategy") or independently mapping both reads and keeping only those hits in which both reads map to the same protein (referred to here as the "congruent strategy"). We used our custom database for beta-galactosidase, a well-known

enzyme present in both infant and adult stool, and mapped reads against this beta-gal database using the R1, congruent, or merged read strategy. With the infant metagenomes, in which most reads were mergeable and the resulting reads were not much longer than the single reads, both alternative strategies were highly correlated with the merged strategy, as expected. The correlation between merged and R1 strategies was r=0.97, p<2.2e-16 (**Figure 5A**) and the correlation between merged and congruent strategies was r=0.96, p<2.2e-16 (**Figure 5B**). Furthermore, the number of reads mapped using any of the strategies was similar, likely due to the high percent of overlapping reads (88%). With the adult metagenomes, in which only one third of the reads overlapped, the alternative strategies had a lower correlation with the merged strategy: the correlation between merged and R1 strategies was r= 0.91, p<4.9e-12 (**Figure 5C**) and the correlation between merged and congruent strategies was r= 0.95, p<1.1e-14 (**Figure 5D**). Roughly double the number of counts were obtained with the R1 strategy compared with the merged strategy. The fewest counts were obtained with the congruent strategy, although the correlation was higher. In summary, it is best to increase the number of overlapping reads during the library preparation phase; however, if few reads can be merged, the congruent strategy appears to be a viable alternative.

2.5 Evaluation of the effect of sequencing depth

The choice of sequencing depth would be expected to effect the likelihood of detecting all the proteins in a given population. To determine the minimum sequencing depth needed for metagenomic sequencing of stool metagenomes, we sub-sampled stool metagenomes that were deeply sequenced. The 33 infant metagenomes had a range of sequencing depths from 13.7 to 31.7 million reads per metagenome [16]. We also selected the 30 largest metagenomes from among 500 metagenomes in an adult stool metagenome project [15]. Each metagenome was randomly sub-sampled to produce metagenomes ranging from 100,000 to 10 million sequences each, with 10 sub-samples at each read depth. The standard deviation of the counts of the target gene, beta-galactosidase, across the 10 sub-samples was then calculated at each read depth. As expected, the greater the sequencing depth in infant stool metagenomes, the lower the standard deviation relative to quantitation from the full metagenome (**Figure 6**). For quantification of beta-galactosidase in infant metagenomes, it appears that even a few million merged reads is sufficient to accurately capture the relative abundance of beta-galactosidase. We repeated the experiment with adult stool metagenomes but because these were prepared with suboptimal insert sizes for merging, even the largest metagenomes had only 4 million merged reads, which would be insufficient for a sub-sampling experiment. We therefore applied the conguent strategy to all unmerged read pairs in the largest metagenomes with at least 20 million reads. The standard deviation of the count of beta galactosidases in the sub-sampled metagenomes relative to the full metagenomes reached a plateau near 10 million reads (Figure 6B).

Given that one of the infant metagenomes was very deeply sequenced (30 million reads with merging of nearly 90% of reads), we conducted further experiments to determine the read depth to classify other groups or enzymes of interest. For CAZy families GH29 or GH95 which contain fucosidases, 5 million reads appear to be sufficient for quantification (Additional File 1, Figure S5). This appears to be similarly

true for broader classifications of "animal carbohydrates", "plant cell wall carbohydrates", "mucin", and "sucrose/fructans" based on multiple CAZy families (**Figure 7A**).

Antimicrobial resistance genes in fecal metagenomes are of practical importance, but they are rare. We therefore asked what the sequencing depth would need to be to accurately quantitate an antimicrobial resistance gene, beta-lactamase, that was known to be both low abundance and differentially abundant in the infant metagenomes [16]. We sub-sampled six infant genomes—two known to be high in beta-lactamase, two with medium beta-lactamase and two with low beta-lactamase—and mapped against a custom database for beta-lactamases. The standard deviation of beta-lactamase counts, relative to the full metagenomes, seemed to be again minimized by 5 million merged reads (**Figure 7B**). In general, 5 million merged reads seem sufficient for functional assessment of infant stool metagenomes.

Discussion

As the cost of sequencing continues to decrease, more studies will use shotgun metagenomes to estimate the functional content of fecal microbiomes and scientists will be faced with practical choices in the pre- and post-sequencing phases. In the current study, we found that read length, e-value threshold, choice of protein database, strategy for paired end reads, and sequencing depth all impacted the ability to detect sequences of known abundance in shotgun metagenomes of human fecal microbiomes. Furthermore, decision-making around these parameters is intertwined as they affect each other.

Previous studies have observed a relationship between read length and the accuracy of gene detection [11, 17]. Consistent with their observations, we found that overall accuracy increases with read length. What is more interesting is translating this knowledge to determine which of the commonly available sequencing formats are therefore appropriate. SR50 should clearly never be used; SR100 should only be used with a small customized database, provided that quality trimming does not extensively trim the reads. PE100 should only be used for the same purpose as SR100 or it can be used more broadly if the sequencing library is size selected to enable overlapping paired ends, which are then merged to create merged reads longer than 100bp. When processing longer read formats, reads <100bp should simply be discarded. The Human Microbiome Project's WGS Read Processing protocol retaining all reads >60bp [18]; clearly, this threshold is too low for functional annotation of gene sequences. Given today's read format choices, the ideal choice would be PE150 with size-selection to enable overlapping reads. This can generate merged reads with a median length near 250bp (D. Lemay, unpublished observations).

We also investigated what to do with legacy data in which most paired end reads are not mergeable. For one of the largest sets of fecal metagenomes from healthy adults to date [15], we found that only one third of the reads overlapped and when they did, the mean read length was still rather short (just 112bp). In another dataset [16], merging reads was much more successful ((~88% overlap, mean 188bp, on PE150 data). Using these two datasets, we compared three read strategies: a "merged strategy" (using only merge-able reads), an "R1 strategy" (using only forward reads), and a "congruent strategy" (independentally mapping both reads and keeping only those hits in which both reads map to the same

protein). As expected, all three strategies are highly correlated with "merged" and "congruent" having the highest correlation. However, given that mergeable reads result in different absolute numbers of counts across the three strategies (R1 >> merged >> congruent), the counts from the three strategies should not be combined. A single strategy should be used for the whole data set. If the reads are mergeable, the merged strategy should be used, otherwise the congruent strategy is a reasonable alternative.

For the design of future experiments, should the sequence library be size-selected to enable overlapping pairs of reads? If functional annotation is a primary focus, then merged reads are highly desirable. Fewer hits were obtained with the "congruent" strategy, suggesting that the sensitivity would be lower if paired end reads cannot be merged. Rare genes, such as antimicrobial resistance genes, would then be more difficult to quantitate.

Knowing that appropriate e-value thresholds are dependent on the search tool used as well as read length and database size [17], we investigated what thresholds would be appropriate for use with DIAMOND at different read lengths. We found that two different protein sequences had two different optimal thresholds for accurate detection. This supports the opinion of other researchers that it may not be possible to identify a single cut-off for all proteins [7, 17]. However, given that we identified and tested a "worse case scenario", our data suggests that it is possible to identify reasonable ranges for mean read lengths. For example, using DIAMOND in "sensitive" mode, a threshold within the range of the default 1-e3 to 1-e10 would be appropriate for read lengths of 100-150bp when a small customized database is used. A threshold within the range of 1e-10 to 1e-25 would be appropriate for read lengths of 200-250bp. For a larger database, like the SEED, we observed higher false positives with increasing read length, suggesting that a threshold more strict than the default is needed for broad databases. Generally, if the research question in mind has low tolerance for false positives, thresholds towards the stricter end of the suggested ranges should be chosen.

As it is well-known that shotgun metagenomes will have reads of varying length, it may be useful to implement a binning strategy to apply e-value thresholds proportional to the read length. First, shorter reads (<100bp) should be removed. Then, if the read lengths vary dramatically within a sample, it may be useful to sort each sequence or merged pair of sequences into bins such that appropriate e-value thresholds can be applied to each bin.

Most metagenomic analyses to date are non-hypothesis driven, using a database of all available protein sequences. As more metagenomes become publicly available, one can imagine that these data will be revisited with specific hypotheses in mind. As sequencing becomes cheaper, it is also likely that shotgun metagenomes will be produced to address a specific research question on microbial functionality, rather than as a non-hypothesis-driven exploration. Bengtsson-Palme has suggested that a database that is specialized for the research question should be used and, if it does not exist, should be constructed based on genes of verified function described in the literature [7]. Our data supports that opinion. With the goal of detecting beta-galactosidase, the custom database for this enzyme was far superior to the other databases. NCBI and SEED tended to under-estimate the number of beta-galactosidase enyzmes in the

simulated metagenome, while the use of CAZy over-estimated them. The fact that beta-galactosidases exist in several distinct families in the CAZy database and that those families, in turn, include some sequences that are likely not beta-galactosidases further demonstrates the difficulty interpreting results for specific enzymes using the CAZy database. Criteria previously suggested for customization of databases include whether the sequences are experimentally verified, the quality of the data, and the functional and/or taxonomic coverage of the data [19]. If analyzing the content of a particular enzyme of interest, the database could be focused on experimentally verified sequences of that enzyme as we did for beta-galactosidase. By extension, if a particular pathway is the focus of the hypothesis, the protein sequences associated with the pathway should be curated for the database. Hypothesis-specific databases would be faster, more accurate, and highly interpretable.

Although it was expected that a custom database would perform best, this had not been previously demonstrated. Prior to performing the experiments, we did not know if the use of a custom database would increase the false positive rate. Without other proteins in the database, there could be a risk that non-target sequences would be more often matched to the target because the non-target sequence reads would not have a true match in the database. However, our experiments suggest this is not the case. The custom database containing only the target sequence provided superior performance.

Investigating the effect of sequencing depth (e.g. library size) on the ability to quantitate specific genes or functional categories in human fecal metagenomes, we found that 5 million merged reads or 10 million unmerged reads would be sufficient to quantitate even rare genes. In prior work, Nayfach and Pollard surprisingly found that reducing the sequencing depth of human gut metagenomes by 95% introduced < 2.5% variation in gene copy number estimates [20]. However, they also suggest that the effects differ for common versus rare genes. What is rare? Considering that there are, on average, around 500,000 unique microbial genes per individual's gut microbiome [21], many individual genes will be rare. We found that beta-galactosidase, the enyzme that digests lactose, has about 500 copies per 5 million merged reads in an adult metagenome and tens times as many in an infant metagenome. For beta-lactamase, an antimicrobial resistance gene, there were about around 100 copies per 5 million merged reads in a deeply sequenced infant metagenome. The concept of shallow sequencing—roughly 500,000 sequences per metagenome—has been put forward as an in-between alternative between 16S sequencing and current shotgun metagenomes [22]. With 500,000 unique microbial genes per individual's gut microbiome [21], shallow sequencing would sample each gene only once, on average, which would not be sufficient to assign presence/absence to most genes with confidence, nevermind comparative quantitation among samples. Answering hypotheses about specific genes or functions more narrow than say, "carbohydrate metabolism", will require a more deeply sequenced metagenome. We suggest a minimum of 5 million merged reads of at least 150bp in length.

Our recommendations on sequencing depth are limited to shotgun sequencing of human fecal samples. These samples are high biomass with nearly all DNA from microbial sources and low risk of substantial contamination. Furthermore, host DNA is easily removed [23, 24]. Others have indicated that problem of higher variation due to contamination in low biomass samples [25] which suggests that

recommendations based on human fecal samples may not extend to other samples. It is possible that recommendations may not even extend to stool samples of other animals. Zaheer et al reported that >50 million reads per sample would be needed to characterize the fecal resistome of cattle fecal samples [26]. Nearly all reads (>97%) were uncharacterizable in that study; the authors suggested that this may be due to the presence of feed-associated plants that were not in the database.

The experiments in the current study do not address how to use gene abundances to compare across metagenomes. Nayfach and Pollard review three methods: (1) gene relative abundance, which is relative amounts of genes found in a sample, (2) average genomic copy number, which is the expected number of copies of the gene per cell, and (3) gene absolute abundance, which cannot be estimated from sequence data alone [20]. The choice should probably be dependent upon the application. If the primary goal is to calculate the carbohydrate active enzyme profile of a fecal metagenome without regards to whether those enyzmes are derived from bacteria with a reference genome or even non-bacteria, such as archaea or yeasts, then the gene relative abundance may be expected to yield less bias than average genomic copy number. However, it is important that differential abundance testing is done using a count-adapted method, such as DESeq2 [27] and edgeR [28], which simultaneously account for differences in library sizes and the compositional nature of the data and have demonstrated applicability for metagenome data [29]. A comparison of 14 different methods demonstrated that these methods have the best performance to determine differnentially abundant genes in metagenomes [30].

Most analyses of metagenomes involve assembly of reads into contigs followed by functional annotation of the contigs [5–7]. However, it is known that this process results in draft assemblies of only the most abundant organisms [5]. Thus, estimates of functional capacity of assembled metagenomes are biased in favor of the functions of the most abundant members [7]. It has also been shown that the assembly process was detrimental to accurate estimation of antimicrobial resistance genes [4]. For these reasons, we recommend the direct mapping of reads without prior metagenomic assembly when the primary research goal is an unbiased quantification of the functional capacity of a microbial community.

Our study is intended to be agnostic to the analysis pipeline used, except for the use of DIAMOND as a mapping utility. An argument against using point-and-click computational computational pipelines is the inability to assign appropriate cut-offs for genes of interest [7]. An advantage of creating one's own pipeline or modifying one that is open source, such as SAMSA2 [31], is the ability for users to define e-value thresholds and use customized databases. SAMSA2, although built for metatranscriptomes, also works to assess gene abundances in metagenomes with the omission of the rRNA removal step, which is unnecessary for metagenomes. Therefore, a version of the SAMSA2 pipeline with that modification was used for the experiments in this paper. For the functional annotation of fecal metagenomes, it could be argued that a broad overview with mappings to the Clusters of Orthologous Genes (COG) database [32], KEGG Orthology [33], and/or SEED subsystems [14] databases is a reasonable first step followed by tests for specific hypotheses with databases customized for those hypotheses.

An alternative to the DIAMOND/BLAST approach would be a mapping strategy which uses hidden Markov models (HMMs). A profile HMM encodes the statistical probability of each amino acid in the sequence of a protein family used to train the model. This can be useful to identify distant homologs that may not be represented in a database. The Pfam database is a large collection of protein families, with each family represented by a multiple sequence alignment and HMM [34]. Pfam also collects entries into clans that are collections of proteins related by profile HMM. Several tools, such as HMM-GRASPx [35] and MetaCLADE [36], identify protein domains in metagenome and/or metatranscriptome sequences. These methods are likely most useful to annotate sequences when proteins are not well-represented in reference databases, such as in environmental sequence data [37], because protein domain information is better than no information. However, proteins with very different functions can share the same protein domain and there can be multiple protein domains in the same protein. For human fecal metagenomes, it is now possible to map 70-90% of reads to genes [38] using the integrated gene catalog (IGC) of the human gut metagenome [39]. Given that the average percentage of prokyarotic genomes that comprise gene-coding regions is 87%, mapability of reads from human gut microbiomes has approached the maximum achievable [39]. Thus, it is both possible and preferable to identify the complete protein rather than a protein domain to enable interpretation of the results. In theory, a profile HMM can be built using complete sequences from different taxa for a particular protein; whether such a strategy has higher performance than the DIAMOND/BLAST approach is beyond the scope of the current analysis.

Assignment of sequence reads from fecal metagenomes to broad functional categories gives the impression of stability [40, 41], but this impression is likely false when investigating function on a finer scale of individual genes. Publicly available human fecal metagenomes have been reanalyzed for functional content using alternate methods [42] or for meta-analyses across cohorts [38]. Legacy data should also be re-analyzed for specific hypotheses. The results of the current study suggests some guidelines for using legacy data to address hypotheses of interest: create a custom database of the gene or genes of interest, merge paired ends to create longer reads, and adjust the e-value threshold for the median read length or bin reads by read length for different e-value thresholds.

While many of our observations are unsurprising, they challenge the current paradigm in several ways. First, conventional sequence library preparation for metagenomes produces non-overlapping reads; our analysis suggests that this protocol should be altered to increase the percentage of overlapping reads when the primary research question involves microbial genes, rather than taxa. Second, if metagenomes with non-overlapping reads have already been sequenced, we demonstrate that a congruent mapping strategy is an appropriate alternative. Third, the default e-value thresholds for DIAMOND are not always appropriate for the analysis of metagenomes; this work provides appropriate e-values for different read lengths. Finally, the use of custom databases is non-standard in metagenomics analysis, and this work advocates for their use in hypothesis-driven analysis of metagenomes.

Conclusions

The accurate identification of sequences of known abundance in human fecal metagenomes was affected by read length, e-value threshold, choice of protein database, strategy for paired end reads, and sequencing depth. If the primary purpose of metagenomics analysis is to quantitate the functional capacity of microbial communities, then DNA extracted from human fecal samples sequenced using the Illumina platform should be size-selected to enable merging of paired end reads, and should be sequenced in the PE150 format or better with a minimum sequencing depth of 5 million merge-able reads, which will likely be closer to 10 million reads. Expecting the merged reads to be 180-250bp in length, the appropriate e-value threshold for DIAMOND would then be more strict than the default. Accurate and interpretable results for specific hypotheses will be best obtained using small databases customized for the research question.

Methods

5.1 Construction of a database of proteins with known function

In order to test various functional metagenomic techniques, a database containing only proteins with experimentally verified functions was constructed. The three sources for protein sequences were Swiss-Prot [43], New England BioLabs (NEB) Inc., and those previously experimentally verified in *Bifidobacterium* [16–24]. From Swiss-Prot, a fasta formatted database of reviewed bacterial sequences with experimentally verified functions was downloaded. Sequences of experimentally validated glycosidases were sent directly from the NEB techical team. We manually collected a list of *Bifidobacterium* enzymes that were biochemically confirmed for their function [16–24]. The protein database (~5000 protein sequences) was then blasted against itself to remove all sequences with greater than 50% identity using custom scripts, removing ~1000 protein sequences that were closely related to other sequences within the database. Lastly, we reverse translated the protein database into a nucleic acid database using the *E. Coli* codon table and the tool EMBOSS Backtranseq [53]. Note, because DIAMOND translates nucleic acid sequences into protein sequences, the choice of codon table did not make a difference in this scenario.

5.2 Construction of simulated metagenomes with increasing dosage of beta galactosidase

The database of proteins with known functions was split into two databases, one with only beta-galactosidase sequences and the other containing all other protein sequences. Sequences were randomly selected from the two databases to create 100 test databases with an increasing proportion of beta-galactosidase enzyme sequences (1/4,000 to 100/4,000). Next, *in silico* metagenomes were created using the next generation sequencing simulator software MetaSim [54]. Using MetaSim, 100 metagenomes of 100,000 reads with increasing dosages of beta-galactosidase (25/100,000 - 2,500/100,000) were simulated at 5 different read length of 50 bp, 100 bp, 150 bp, 200 bp, and 250 bp. Using the same reverse translated nucleic acid database, each sequence was extended by 20% (10% to each end) by randomly adding nucleotides based on actual nucleotide frequencies to represent more

realistic DNA fragments with flanking intergenic regions. Another 500 metagenomes were simulated from the elongated sequences of increasing beta-galactosidase proportions at the 5 different read lengths.

5.3 Fecal metagenome datasets

Two human fecal shotgun metagenomic datasets were analyzed in this paper. The dataset of infant fecal metagenomes included 33 samples [16]. The second dataset included 30 adult fecal metagenomes which were among the largest metagenomes from a cohort of 471 healthy, Western European adults of at least 18 years of age, as part of the 500 Functional Genomics project [15]. The infant metagenomes' sequence data are in DDBJ Center with SRA accession no. SRP133760, and the adult metagenomes' sequence data are in NCBI SRA with accession no. PRJNA319574.

5.4 Metagenomic sequence analysis

The SAMSA2 pipeline [55] was modified for metagenomics analysis with the following steps: 1) reads mapping to the human genome were first removed with BMTagger [23], 2) paired-end reads were merged using PEAR [56], 3) sequence adaptor contamination and low quality bases were removed using Trimmomatic [57] and 4) the quality reads were then annotated against a protein reference database using DIAMOND, a high-throughput squence aligner [10]. DIAMOND is highly sensitive and runs at a speed that is up to 20,000 times faster than BLASTX and up to 2,500 times faster with the "sensitive" option. The pipeline script, master_beta.galac.db_analysis_stoolmg.sh, is in

https://github.com/mltreiber/functional_metagenomics

To determine gene content, reads were mapped against three publicly available databases, NCBI RefSeq [58], SEED Subsystems [14], and CAZy [12] with only the best hit retained. Reads were also mapped against a custom-built database which was a fasta-formatted list of experimentally verified beta-galactosidase sequences.

5.5 Metrics of evaluation

The performance of sequence read mapping was evaluated using several metrics. Sequence reads originating from the target that were also classified as the target were considered true positives (TP). Sequence reads not originating from the target that were classified as the target were considered to be false positives (FP). Sensitivity refers to the ability to detect the target sequence while specificity refers to the ability to correctly identify sequences that are not the target; accuracy is the proportion of true results. Sensitivity, specificity, and accuracy were calculated as follows:

$$sensitivity = \frac{TP}{TP+FN} \quad specificity = \frac{TN}{FP+TN} \quad accuracy = \frac{TP+TN}{n}$$

False negatives (FN) were target sequence reads that were not correctly classified as the target. True negatives (TN) were non-target sequence reads that were correctly classified as non-target. The number of total reads was equal to n.

List Of Abbreviations

DIAMOND, a BLAST-like sequence aligner that uses translated sequence queries to greatly increase annotation speed over traditional BLAST.

MG-RAST, the MetaGenomics Rapid Annotation using Subsystems Technology server, a public analysis pipeline for handling metagenome and metatranscriptome datasets

PE, paired end read

SAMSA2, Simple Analysis of Metatranscriptomes through Sequence Annotation version 2, a standalone metatranscriptomics analysis pipeline

SEED, a protein database that seeks to group sequences into hierarchical categories, created by the Fellowship for Interpretation of Genomes (FIG group)

SR, single read

Declarations

Ethics approval and consent to participate

Not applicable. The current study used publicly available data.

Consent for publication

Not applicable.

Availability of data and material

All raw data supporting this study is publicly available. The infant fecal metagenomes analyzed in the current study are available in NCBI Sequence Read Archive (SRA) under accession number SRP133760. The adult fecal metagenomes used in the current study are available in NCBI SRA with accession number PRJNA319574.

The database of experimentally verified protein sequences developed for this study, as well as other databases used, is in the databases directory of https://github.com/mltreiber/functional_metagenomics.

The SAMSA2 analysis pipeline, modified for metagenomics, is called "master_beta.galac.db_analysis_stoolmg.sh", and is available in the scripts directory of https://github.com/mltreiber/functional_metagenomics.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was funded in part by the Inner Mongolia Mengniu Dairy (Group) Company Ltd.; National Institutes of Health awards F32HD093185 (D.H.T.), and R01AT008759 (D.A.M.); and U.S. Department of Agriculture 2032-51530-026-00D (DGL). The United States Department of Agriculture is an equal opportunity provider and employer. Funding bodies played no role in the design of the study or analysis or interpretation of data or in writing the manuscript.

Authors' contributions

DGL conceived of the study. MLT conducted all experiments. MLT, IK, and DGL contributed to code. MLT, DT, IK, DAM, and DGL interpreted data. MLT and DGL drafted the manuscript. DT and IK contributed to the manuscript. MLT, DT, IK, DAM, and DGL read and approved the final manuscript.

Acknowledgements

The authors thank Dr. Nina Kirmiz for the collection of *Bifidobacterium* enyzme sequences and Shannon E.K. Joslin for testing removal of human reads with BMTagger. The authors also thank Dr. Jinxin Liu, Dr. Lutz Froenicke, Dr. Matthew Settles, Dr. Zhengyao Xue, and Dr. Mary Kable for helpful discussions or comments.

References

1. Clarridge JE. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clinical Microbiology Reviews. 2004;17:840–62.
2. Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol. 2013;31:814–21.
3. Sun S, Jones RB, Fodor AA. Inference based PICRUSt accuracy varies across sample types and functional categories. bioRxiv. 2019;:655746. doi:10.1101/655746.
4. Bengtsson-Palme J, Boulund F, Fick J, Kristiansson E, Larsson DGJ. Shotgun metagenomics reveals a wide array of antibiotic resistance genes and mobile elements in a polluted lake in India. Front Microbiol. 2014;5:648. doi:10.3389/fmicb.2014.00648.

5. Prakash T, Taylor TD. Functional assignment of metagenomic data: challenges and applications. Brief Bioinform. 2012;13:711–27. doi:10.1093/bib/bbs033.

6. Sharpton TJ. An introduction to the analysis of shotgun metagenomic data. Frontiers in Plant Science. 2014;5 JUN.

7. Bengtsson-Palme J. Strategies for Taxonomic and Functional Annotation of Metagenomes. In: Metagenomics: Perspectives, Methods, and Applications. Elsevier Inc.; 2017. p. 55–79.

8. Rivers AR. A quick reference guide to metagenome sequencing at JGI. 2016.

9. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Research. 1997;25:3389–402.

10. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12:59–60. doi:10.1038/nmeth.3176.

11. Carr R, Borenstein E. Comparative analysis of functional metagenomic annotation and the mappability of short reads. PLoS One. 2014;9:e105776. doi:10.1371/journal.pone.0105776.

12. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 2014;42 Database issue:D490-5. doi:10.1093/nar/gkt1178.

13. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44:D733–45.

14. Overbeek R, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res. 2014;42:206–14. doi:10.1093/nar/gkt1226.

15. Schirmer M, Smeekens SP, Vlamakis H, Jaeger M, Oosting M, Franzosa EA, et al. Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity. Cell. 2016;167:1125-1136.e8.

16. Taft DH, Liu J, Maldonado-Gomez MX, Akre S, Huda MN, Ahmad SM, et al. Bifidobacterial Dominance of the Gut in Early Life and Acquisition of Antimicrobial Resistance. mSphere. 2018;3.

17. Nayfach S, Bradley PH, Wyman SK, Laurent TJ, Williams A, Eisen JA, et al. Automated and Accurate Estimation of Gene Family Abundance from Shotgun Metagenomes. PLoS Comput Biol. 2015;11.

18. Young S, Martin J, Kota K, Mitreva M. HMP WGS Read Processing. 2011. https://www.hmpdacc.org/hmp/doc/ReadProcessing_SOP.pdf.

19. Bengtsson-Palme J, Boulund F, Edström R, Feizi A, Johnning A, Jonsson VA, et al. Strategies to improve usability and preserve accuracy in biological sequence databases. Proteomics. 2016;16:2454–60.

20. Nayfach S, Pollard KS. Toward Accurate and Quantitative Comparative Metagenomics. Cell. 2016;166:1103–16. doi:10.1016/j.cell.2016.08.007.

21. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010;464:59–65.

22. Hillmann B, Al-Ghalith GA, Shields-Cutler RR, Zhu Q, Gohl DM, Beckman KB, et al. Evaluating the Information Content of Shallow Shotgun Metagenomics. mSystems. 2018;3.

23. Agarwala R, Morgulis A. BMTagger. 2014. ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger/.

24. Schmieder R, Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One. 2011;6.

25. Weiss S, Amir A, Hyde ER, Metcalf JL, Song SJ, Knight R. Tracking down the sources of experimental contamination in microbiome studies. Genome Biol. 2014;15:564. doi:10.1186/s13059-014-0564-2.

26. Zaheer R, Noyes N, Ortega Polo R, Cook SR, Marinier E, Van Domselaar G, et al. Impact of sequencing depth on the characterization of the microbiome and resistome. Sci Rep. 2018;8.

27. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550. doi:10.1186/s13059-014-0550-8.

28. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2009;26:139–40.

29. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol. 2014;10:e1003531. doi:10.1371/journal.pcbi.1003531.

30. Jonsson V, Österlund T, Nerman O, Kristiansson E. Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. BMC Genomics. 2016;17.

31. Westreich ST, Treiber ML, Mills DA, Korf I, Lemay DG. SAMSA2: A standalone metatranscriptome analysis pipeline. BMC Bioinformatics. 2018;19.

32. Tatusov RL, Koonin E V, Lipman DJ. A genomic perspective on protein families. Science. 1997;278:631–7. doi:10.1126/science.278.5338.631.

33. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44:D457-62. doi:10.1093/nar/gkv1070.

34. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47:D427–32. doi:10.1093/nar/gky995.

35. Zhong C, Edlund A, Yang Y, McLean JS, Yooseph S. Metagenome and Metatranscriptome Analyses Using Protein Family Profiles. PLoS Comput Biol. 2016;12:e1004991. doi:10.1371/journal.pcbi.1004991.

36. Ugarte A, Vicedomini R, Bernardes J, Carbone A. A multi-source domain annotation pipeline for quantitative metagenomic and metatranscriptomic functional profiling. Microbiome. 2018;6:149. doi:10.1186/s40168-018-0532-2.

37. Raes J, Foerstner KU, Bork P. Get the most out of your metagenome: computational analysis of environmental sequence data. Current Opinion in Microbiology. 2007;10:490–8.

38. Armour CR, Nayfach S, Pollard KS, Sharpton TJ. A Metagenomic Meta-analysis Reveals Functional Signatures of Health and Disease in the Human Gut Microbiome. mSystems. 2019;4.

39. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol. 2014;32:834–41. doi:10.1038/nbt.2942.

40. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, et al. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207–14.

41. Heintz-Buschart A, Wilmes P. Human Gut Microbiome: Function Matters. Trends Microbiol. 2018;26:563–74. doi:10.1016/j.tim.2017.11.002.

42. Bradley PH, Pollard KS. Proteobacteria explain significant functional variability in the human gut microbiome. Microbiome. 2017;5.

43. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. UniProtKB/Swiss-Prot. Methods Mol Biol. 2007;406:89–112. http://www.ncbi.nlm.nih.gov/pubmed/18287689. Accessed 21 Aug 2019.

44. Egan M, Motherway MOC, Ventura M, van Sinderen D. Metabolism of sialic acid by Bifidobacterium breve UCC2003. Appl Environ Microbiol. 2014;80:4414–26.

45. Ruiz-Moyano S, Totten SM, Garrido DA, Smilowitz JT, Bruce German J, Lebrilla CB, et al. Variation in consumption of human milk oligosaccharides by infant gut-associated strains of bifidobacterium breve. Appl Environ Microbiol. 2013;79:6040–9.

46. Kitaoka M, Tian J, Nishimoto M. Novel putative galactose operon involving lacto-N-biose phosphorylase in Bifidobacterium longum. Appl Environ Microbiol. 2005;71:3158–62.

47. Turroni F, Duranti S, Bottacini F, Guglielmetti S, Van Sinderen D, Ventura M. Bifidobacterium bifidum as an example of a specialized human gut commensal. Front Microbiol. 2014;5 AUG.

48. Turroni F, Bottacini F, Foroni E, Mulder I, Kim JH, Zomer A, et al. Genome analysis of Bifidobacterium bifidum PRL2010 reveals metabolic pathways for host-derived glycan foraging. Proc Natl Acad Sci U S A. 2010;107:19514–9.

49. Garrido D, Ruiz-Moyano S, Mills DA. Release and utilization of N-acetyl-D-glucosamine from human milk oligosaccharides by Bifidobacterium longum subsp. infantis. Anaerobe. 2012;18:430–5. doi:10.1016/j.anaerobe.2012.04.012.

50. Yoshida E, Sakurama H, Kiyohara M, Nakajima M, Kitaoka M, Ashida H, et al. Bifidobacterium longum subsp. infantis uses two different-galactosidases for selectively degrading type-1 and type-2 human milk oligosaccharides. Glycobiology. 2012;22:361–8.

51. Sela DA, Li Y, Lerno L, Wu S, Marcobal AM, German JB, et al. An infant-associated bacterial commensal utilizes breast milk sialyloligosaccharides. J Biol Chem. 2011;286:11909–18. doi:10.1074/jbc.M110.193359.

52. Sela DA, Garrido D, Lerno L, Wu S, Tan K, Eom HJ, et al. Bifidobacterium longum subsp. infantis ATCC 15697 α-fucosidases are active on fucosylated human milk oligosaccharides. Appl Environ Microbiol. 2012;78:795–803.

53. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 2019;47:W636–41. doi:10.1093/nar/gkz268.

54. Richter DC, Ott F, Auch AF, Schmid R, Huson DH. MetaSim-A Sequencing Simulator for Genomics and Metagenomics. PLoS One. 2008;3:e3373. doi:10.1371/journal.pone.0003373.

55. Westreich ST, Treiber ML, Mills DA, Korf I, Lemay DG. SAMSA2: A standalone metatranscriptome analysis pipeline. BMC Bioinformatics. 2018;19.

56. Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: A fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics. 2014;30:614–20.

57. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.

58. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44:D733–45.

Additional File Legends

File name: Additional_file_1.pdf

Title: Additional File 1

Description: This PDF file contains supplementary figures, Figures S1-S5.

Figures
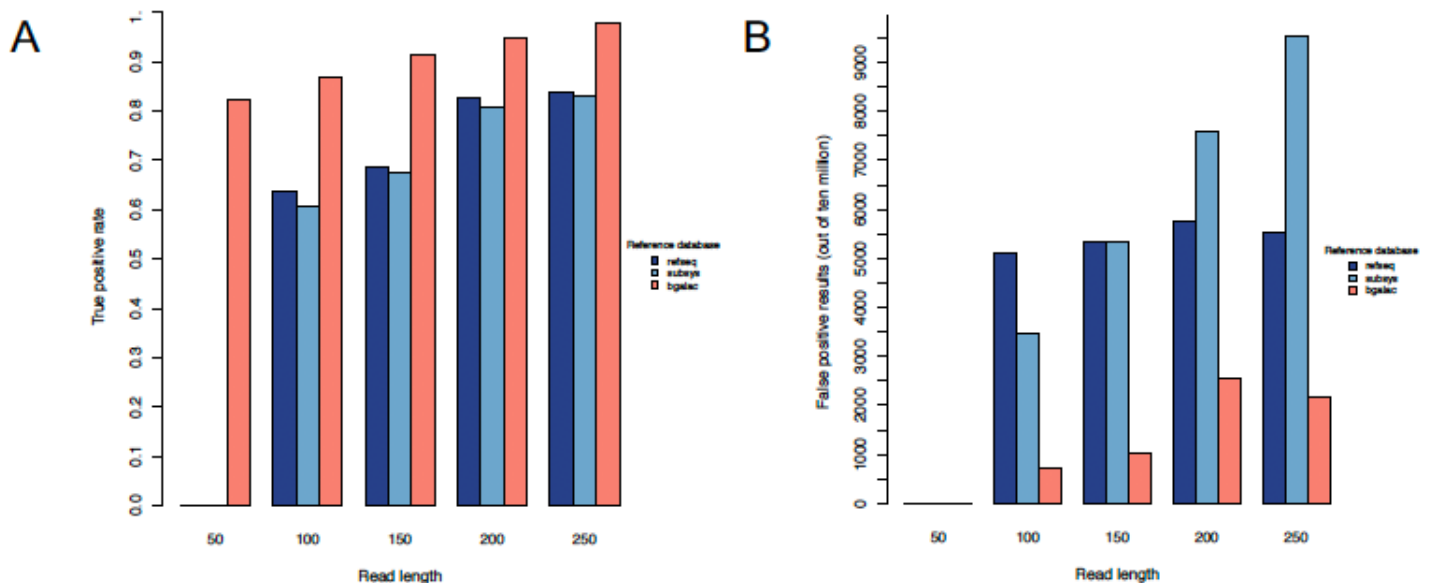
Figure 1

True (A) and false (B) positive rates, across different read lengths, of detecting the target beta-galactosidase sequence in simulated metagenomes using DIAMOND with "sensitive" flag and default e-

value and 3 databases: NCBI RefSeq (navy), SEED Subsystems (blue), and beta-galactosidase (salmon).
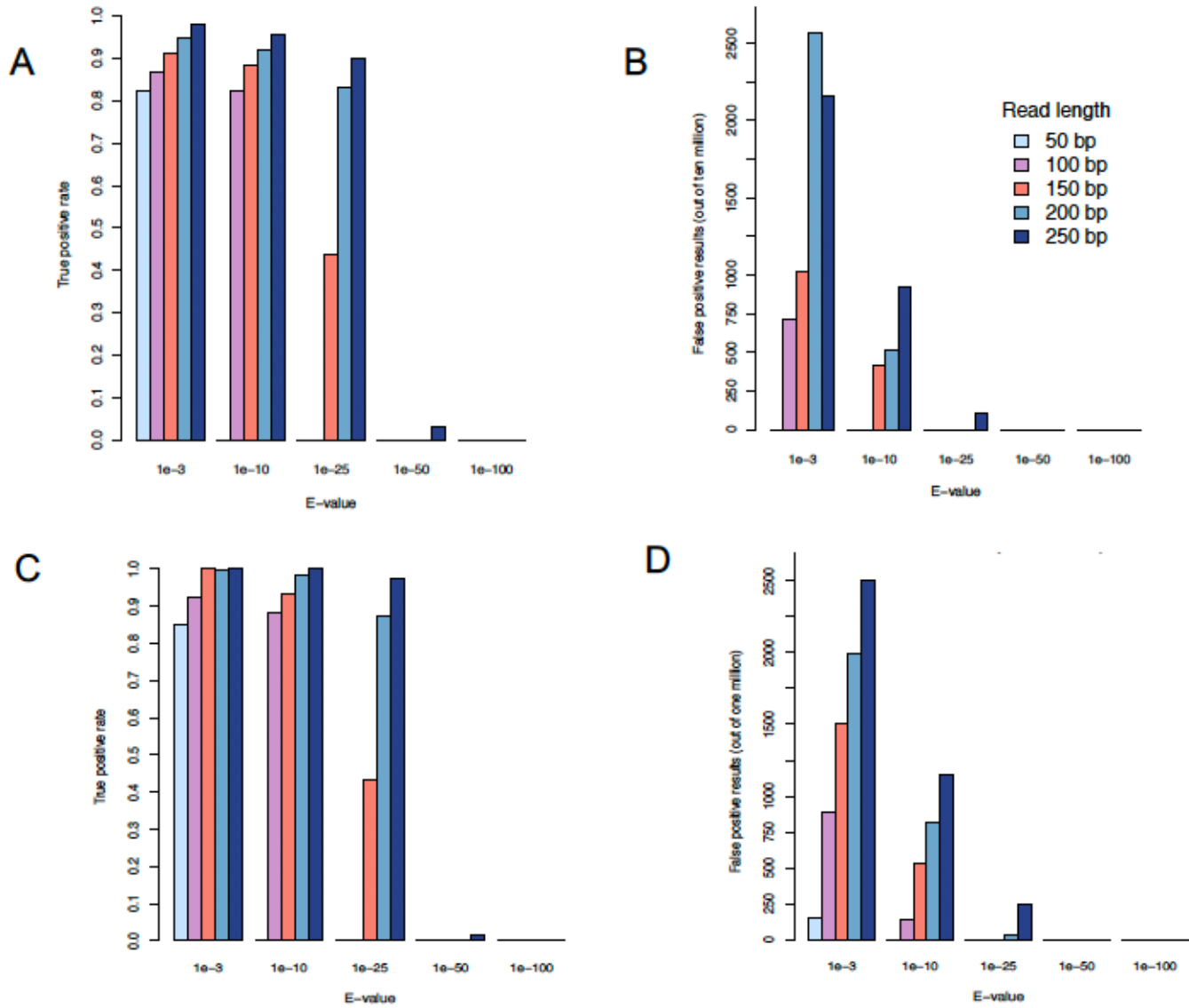
Figure 2

A) True positive rates and (B) false positive rate results from simulated metagenomes containing 2,500 beta-galactosidase sequences at varying e-value cutoffs and read lengths of 50 bp (light blue), 100 bp (purple), 150 bp (salmon), 200 bp (blue), and 250 bp (navy). (C) True positive rates and (D) false positive rate results from simulated metagenomes containing 350 Homoserine O-acetyltransferase at varying e-value cutoffs and read lengths of 50 bp (light blue), 100 bp (purple), 150 bp (salmon), 200 bp (blue), and 250 bp (navy).
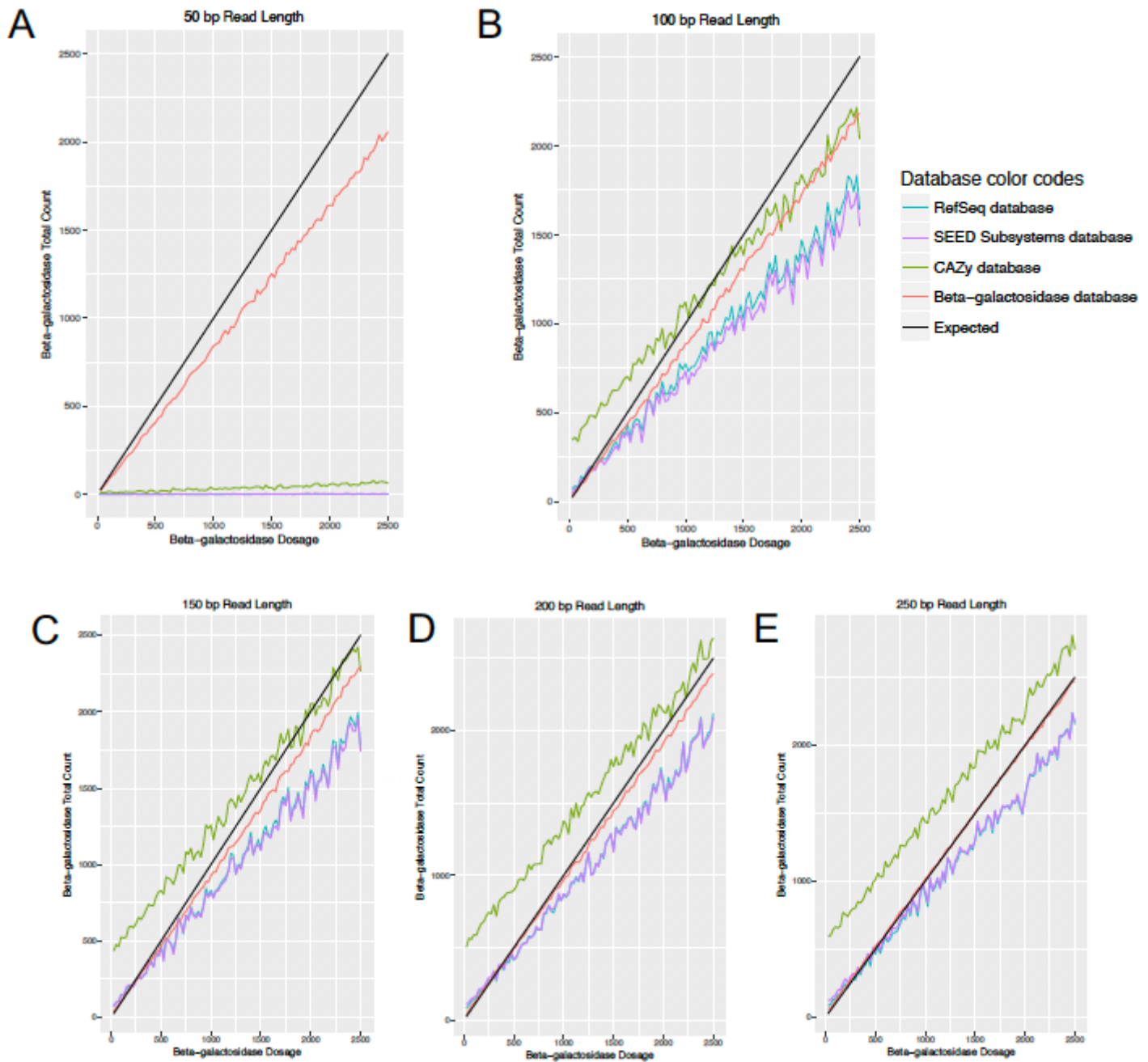
Figure 3

Comparison of total number of beta-galactosidase annotations across varying beta-gal dosages and read lengths of (A) 50 bp, (B) 100 bp, (C) 150 bp, (D) 200 bp, and (E) 250 bp between 4 databases: NCBI RefSeq (blue), SEED Subsystems (purple), CAZy (green), and Beta-galactosidase (red). The expected line is black.
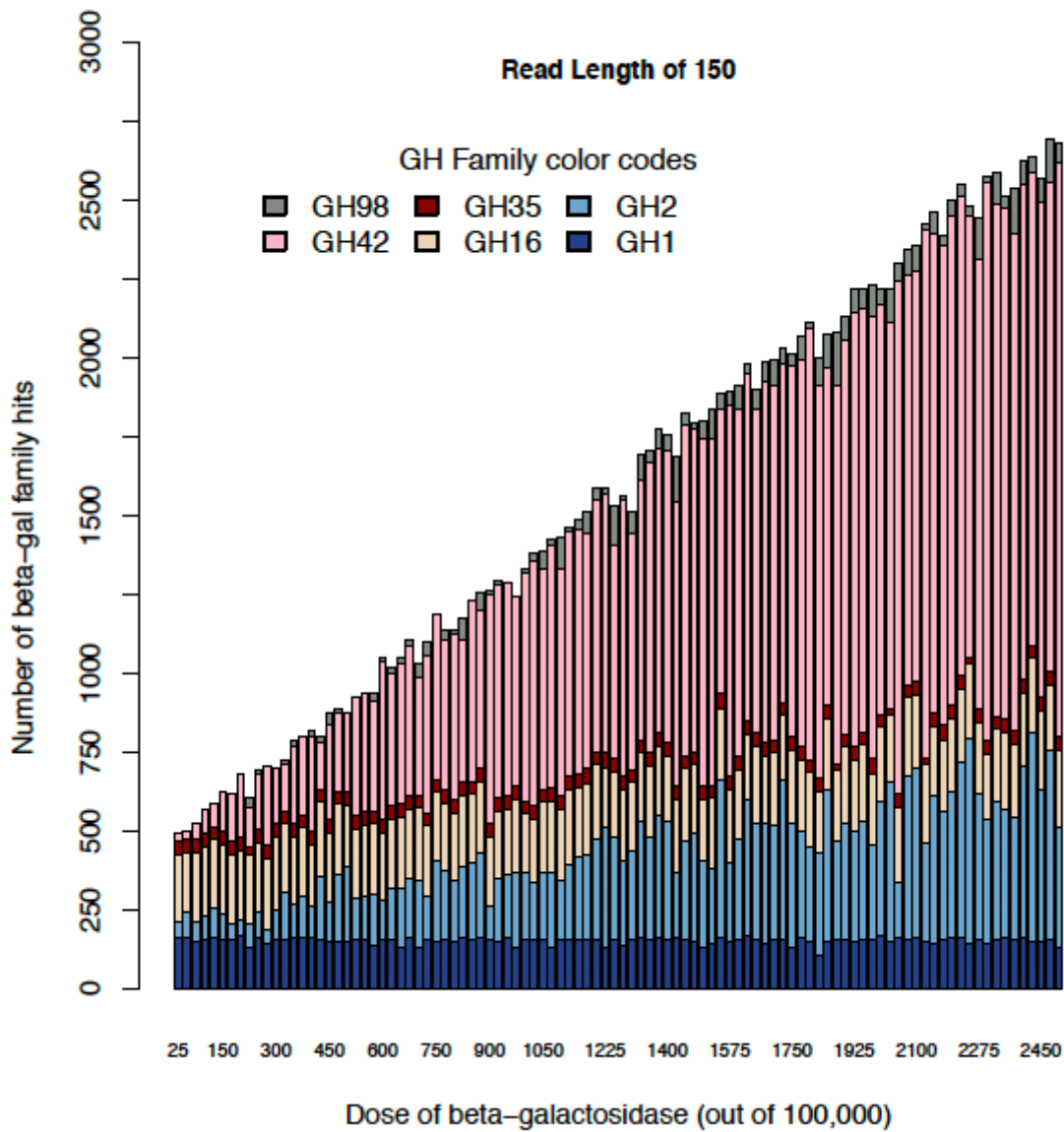
Figure 4

CAZy family hits with increasing beta-gal dosages at a read length of 150 bp. Annotations against 7 CAZy families that contain beta-galactosidase were included: GH1(navy), GH2 (blue), GH16 (tan), GH35 (red), GH42 (pink), and GH98 (grey).
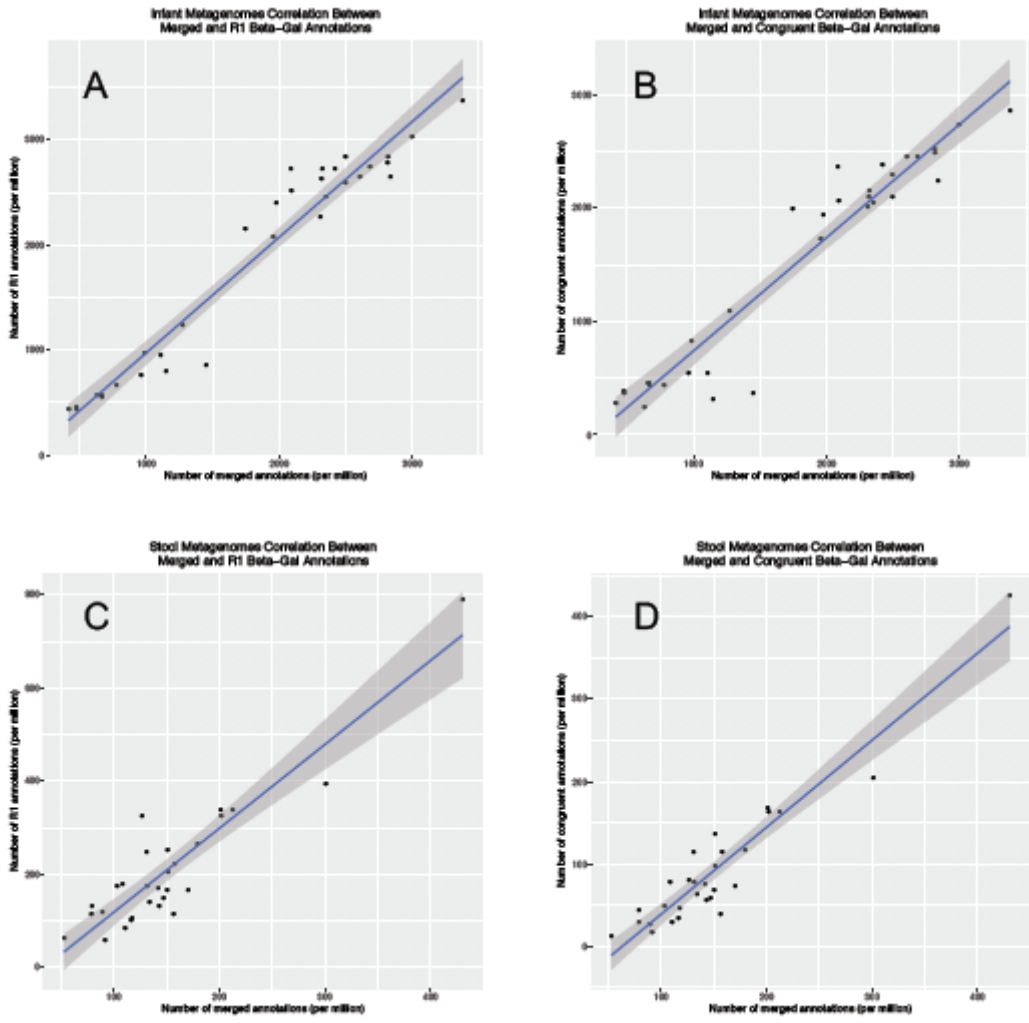
Figure 5

Correlations of the R1, congruent, and merged read annotations in (A-B) infant fecal metagenomes 2x150bp with nearly 90% merge-able reads and (C-D) adult fecal metagenomes 2x100bp format with only one third merge-able reads.
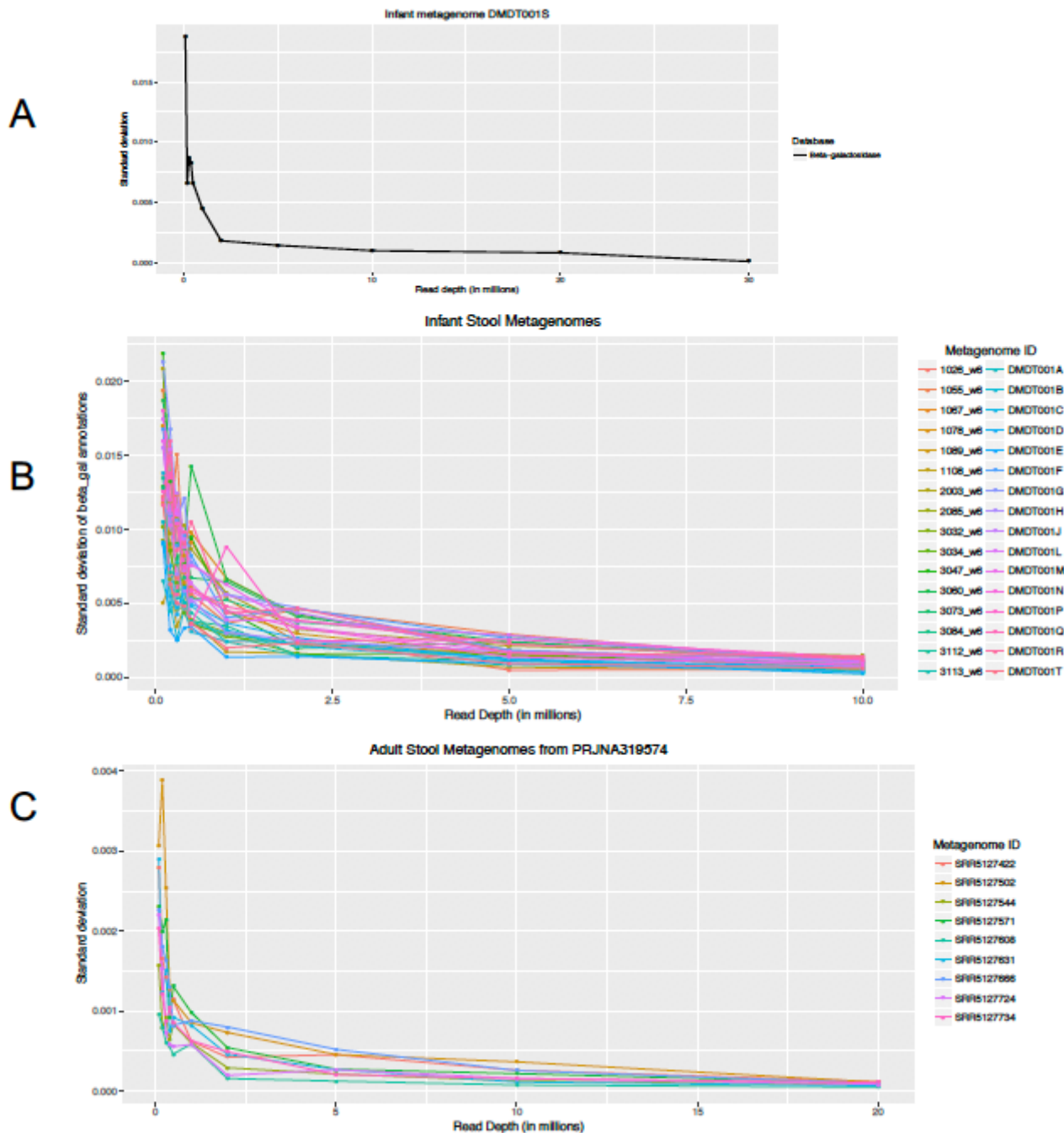
Figure 6

Effect of sub-sampled read depth on the standard deviation of target gene hits in the sub-sampled metagenome relative to the full metagenome. The target gene was beta-galactosidase, mapped using a custom database of beta-galactosidase sequences. The metagenomes tested were (A) an infant fecal metagenome of 30 million merged reads, (B) all infant fecal metagenomes of at least 10 million merged reads, and (C) adult fecal metagenomes with at least 20 million unmerged reads mapped using the congruent strategy.
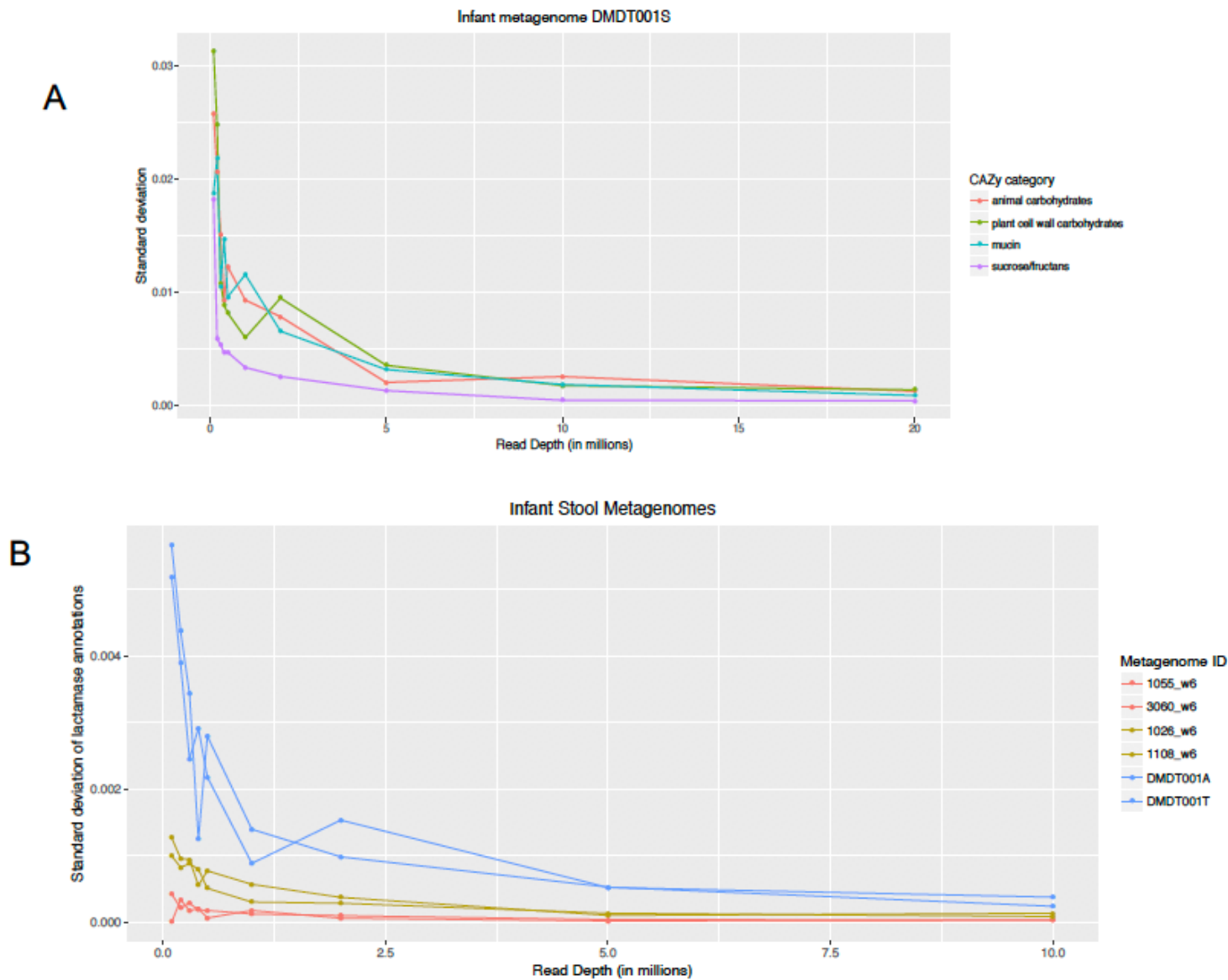
Figure 7

Effect of sub-sampled read depth on the standard deviation of target gene hits in the sub-sampled metagenome relative to the full metagenome. The target gene was (A) CAZy category, mapped using the CAZy database, or (B) beta-lactamase, using a custom database. The metagenomes tested were (A) an infant fecal metagenome of 30 million merged reads, and (B) six infant stool metagenomes representative of high (blue), medium (yellow), and low abundance (red) in beta-lactamase.