Applied and Environmental Microbiology, Mar. 2008, p. 1453–1463 Vol. 74, No. 5
0099-2240/08/$08.00+0 doi:10.1128/AEM.02181-07

Metagenomics: Read Length Matters▽†

K. Eric Wommack,[1] Jaysheel Bhavsar,[1] and Jacques Ravel[2]*

Delaware Biotechnology Institute, University of Delaware, 15 Innovation Way, Newark, Delaware 19711,[1] and Institute for Genome Sciences, Department of Microbiology and Immunology, University of Maryland School of Medicine, 20 Penn Street, Baltimore, Maryland 21201[2]

Received 24 September 2007/Accepted 3 January 2008

Obtaining an unbiased view of the phylogenetic composition and functional diversity within a microbial community is one central objective of metagenomic analysis. New technologies, such as 454 pyrosequencing, have dramatically reduced sequencing costs, to a level where metagenomic analysis may become a viable alternative to more-focused assessments of the phylogenetic (e.g., 16S rRNA genes) and functional diversity of microbial communities. To determine whether the short (~100 to 200 bp) sequence reads obtained from pyrosequencing are appropriate for the phylogenetic and functional characterization of microbial communities, the results of BLAST and COG analyses were compared for long (~750 bp) and randomly derived short reads from each of two microbial and one virioplankton metagenome libraries. Overall, BLASTX searches against the GenBank nr database found far fewer homologs within the short-sequence libraries. This was especially pronounced for a Chesapeake Bay virioplankton metagenome library. Increasing the short-read sampling depth or the length of derived short reads (up to 400 bp) did not completely resolve the discrepancy in BLASTX homolog detection. Only in cases where the long-read sequence had a close homolog (low BLAST E-score) did the derived short-read sequence also find a significant homolog. Thus, more-distant homologs of microbial and viral genes are not detected by short-read sequences. Among COG hits, derived short reads sampled at a depth of two short reads per long read missed up to 72% of the COG hits found using long reads. Noting the current limitation in computational approaches for the analysis of short sequences, the use of short-read-length libraries does not appear to be an appropriate tool for the metagenomic characterization of microbial communities.

Sequence polymorphism analysis of discrete genes within environmental samples has revolutionized our view of the diversity and the composition of microbial communities. Since it was first proposed as a universal phylogenetic marker of life on earth (29), sequence analysis of the small-subunit rRNA gene has become the gold standard for the assessment of microbial diversity within environmental samples. Today, a plethora of techniques for the assessment of microbial species richness and evenness are based on sequence polymorphism within this single gene (see reference 9 for a review). More recently, the conceptual approaches developed for the analysis of 16S rRNA gene microbial diversity have been applied to functional genes involved in chemical transformations critical to the carbon (e.g., RuBisCo [6]), nitrogen (e.g., NifH [31]), and sulfur (sulfite reductase [18]) cycles. These analytical approaches have revealed a significant diversity of microorganisms that are capable of mediating the chemical transformations that maintain global biogeochemical nutrient cycles. Despite the extraordinary view of microbial taxonomic and functional diversity that single-gene approaches provide, these approaches have two significant limitations, namely, the inability to provide a picture of the broader genomic context of a given gene of interest and the requirement of prior sequence information necessary for the design of oligonucleotide PCR primers and probes.

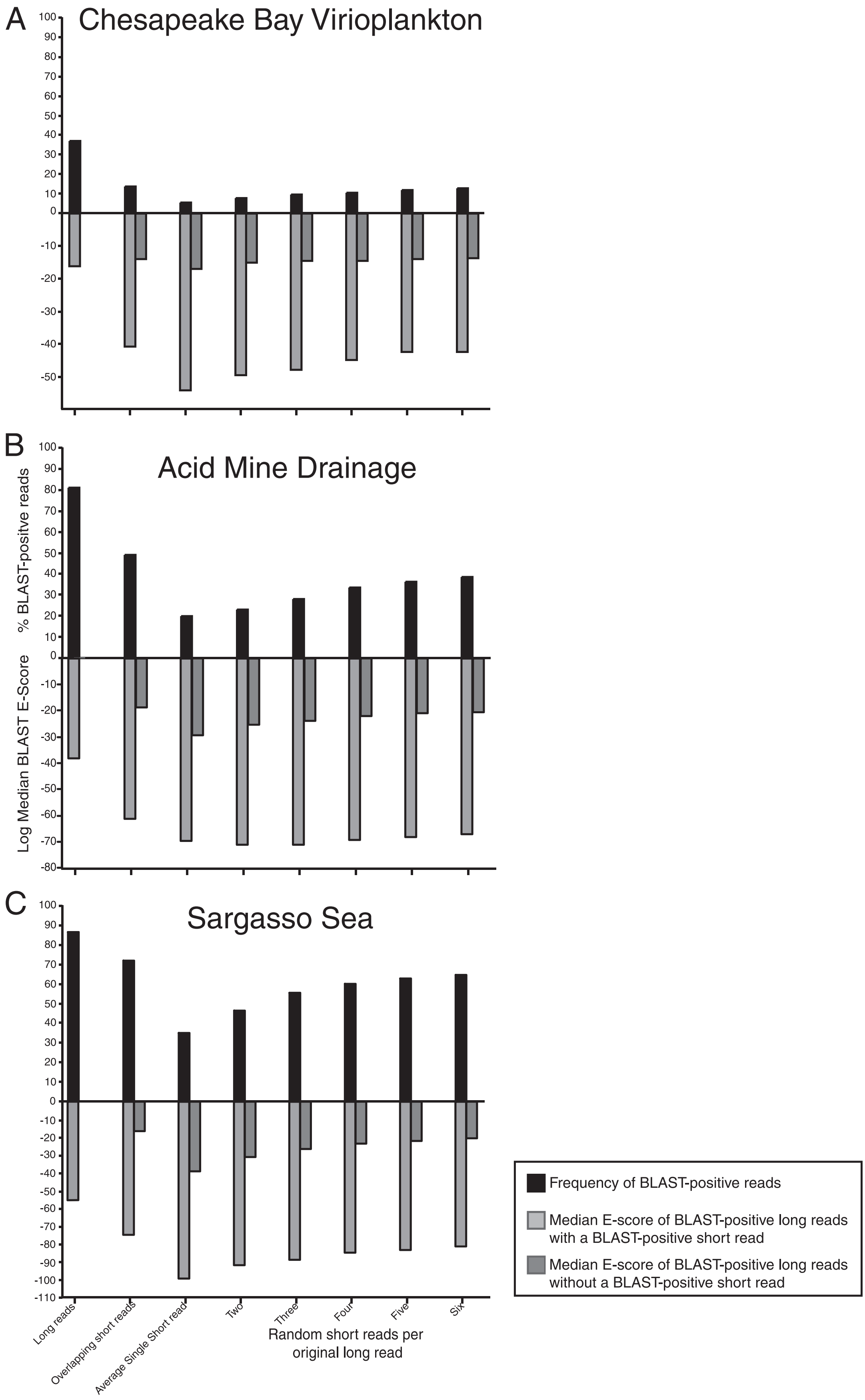
The ideal technique for the assessment of microbial diversity would circumvent the need for selective PCR amplification and provide sequence information of sufficient length to discern connections between the taxonomy and physiology of every microbe within the community. At present, high-throughput sequencing applied to whole-microbial-community DNA, a collective suite of techniques known as microbial metagenomics, is the only approach capable of nearing this holistic view of the taxonomic and functional diversity within extant microbial communities. To date, microbial metagenomic investigations of marine ecosystems have revealed the enormous diversity of potentially photoheterotrophic prokaryotes in the Sargasso Sea (27) and the pervasiveness of cyanophages within the euphotic zone of the pelagic ocean (8). Sequencing of small-insert shotgun libraries of microbial community DNA from low-pH acid mine drainage (AMD) environments (26) and the symbiotic microbial flora of a gutless marine oligochaete (30) have enabled the nearly complete assembly of microbial genomes without the necessity of cultivation. Although significantly smaller in overall scale, shotgun sequencing of viral DNA within environmental samples has revealed that communities of double-stranded DNA viruses are very diverse (3, 5) and contain an extraordinary amount of novel sequence (2).

To date, most metagenomic investigations have adapted whole-genome shotgun sequencing approaches to the cloning and sequencing of microbial community DNA collected from environmental samples. In this approach, small-insert DNA

* Corresponding author. Mailing address: Institute for Genome Sciences, Department of Microbiology and Immunology, University of Maryland School of Medicine, 20 Penn Street, Baltimore, MD 21201. Phone: (410) 706-5674. Fax: (410) 706-1482. E-mail: jravel@som.umaryland.edu.

† Supplemental material for this article may be found at http://aem.asm.org/.

▽ Published ahead of print on 11 January 2008.


A
Chesapeake Bay Virioplankton
100
90
80
70
60
50
40
30
20
10
0
-10
-20
-30
-40
-50
B
Acid Mine Drainage
100
90
80
70
60
50
40
30
20
10
0
-10
-20
-30
-40
-50
-60
-70
-80
% BLAST-positive reads
Log Median BLAST E-Score
C
Sargasso Sea
100
90
80
70
60
50
40
30
20
10
0
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
Long reads
Overlapping short reads
Average Single Short read
Two
Three
Four
Five
Six
Random short reads per original long read
■ Frequency of BLAST-positive reads
▨ Median E-score of BLAST-positive long reads with a BLAST-positive short read
▨ Median E-score of BLAST-positive long reads without a BLAST-positive short read

clone libraries are analyzed using Sanger dideoxy chain terminator sequencing (22) to yield DNA sequence libraries consisting of sequence reads ranging from ca. 600 to 900 bp in length with accuracies exceeding 99.97%. The limitations of this approach are the overall cost of sequencing and the potential biases introduced in constructing clone libraries. Recently, a novel sequencing-by-synthesis technology, 454 pyrosequencing, was introduced which dramatically lowers the per-base pair cost of sequencing and circumvents the need for clone library construction (17). Prior metagenome investigations utilizing 454 pyrosequencing obtained ca. 25 to 40 Mb of DNA sequence data per analytical run at a single-read accuracy of ~98% (1, 10). While the benefits of this technology are substantial and have already proven it to be a useful advance for whole-genome sequencing approaches (13), its limitations are lower individual read accuracy and significantly shorter read lengths (100 to 200 bp) than for Sanger dideoxy sequencing.

Nevertheless, the lower cost, avoidance of potential cloning biases, and large amount of sequence generated in a single run would appear to make 454 pyrosequencing the ultimate approach for metagenomic analysis of microbial communities. To date, this approach has been applied to analyses of microbial communities within low-pH AMD environments (10), the mouse (25), and virioplankton communities of four oceanic provinces (1). In the particular case of metagenome analysis of marine virioplankton communities by 454 pyrosequencing, only a small fraction of sequence reads (>10%) showed significant BLAST homology to sequences within organismal (GenBank nt/nr) and environmental (GenBank env_nt/env_nr) sequence databases combined. In contrast, the frequency of BLAST homologous sequences within Sanger-based metagenome sequence libraries of double-stranded DNA viral communities is typically around 60%, with half of these sequences showing homology to only environmental sequences (2, 11). The substantially lower BLAST homolog frequency within virioplankton metagenome sequence libraries obtained through 454 pyrosequencing prompted this analysis to determine whether low BLAST homolog frequency is a general property of short-read microbial metagenome sequence data. In addition, metagenomic sequence read data sets (long or short) are highly fragmented and cannot be assembled efficiently with the current bioinformatics tools available. These poor-quality assemblies have led scientists to perform read-level analysis of these data sets.

MATERIALS AND METHODS

Data sets. (i) Chesapeake Bay virioplankton metagenomic data. Virioplankton DNA was extracted from a viral concentrate obtained from 50 liters of Chesapeake Bay water. A metagenomic library (insert size, 1 to 2 kb) was constructed in vector pSMART as previously reported (for an example, see reference 14). A total of 6,407 Sanger sequencing reads were obtained.

(ii) AMD metagenomic data. A multi-Fasta file containing the trimmed Sanger sequence of 125,183 reads from the AMD project (26) was provided to us by Jill Banfield.

(iii) Sargasso Sea metagenomic data. The Sanger random shotgun sequencing data (1,667,992 reads) (27) was obtained from the Venter Institute in the form of a multi-Fasta file. Sequences homologous to sequences of *Burkholderia* spp. or *Shewanella* spp. were removed from the Sargasso data set, as the overrepresentation of these sequences is believed to be the result of shipboard contamination of the original metagenome sample (7, 16).

A total of 1,000 reads (>600 bp) were randomly selected from each metagenome sequence library. This subset of 1,000 sequences was considered the long-read reference library for all subsequent comparisons. The average read lengths for these data sets were 681 bp, 812 bp, and 877 bp for the Chesapeake Bay, AMD, and Sargasso Sea metagenomes, respectively.

Short-read data simulations from long-read reference libraries. (i) Tiled-and-overlapping short reads. A set of 100-bp fragments tiled over the entire length of the long read (>80 bp) and overlapping by 20 bp was generated for each of the 1,000 long reads. In this case, there were averages of 7.6, 9.6, and 10.4 short reads for each long-read sequence in the Chesapeake Bay, AMD, and Sargasso libraries, respectively (Fig. 1). This data set represented the ideal case of oversampling that is often generated with short-read sequencing methodologies.

(ii) Multiple random short reads. One-hundred-bp fragments were randomly and independently selected up to six times from each long read. Six data sets were constructed, each containing from one to six random short reads per long-read reference sequence. Increasing the number of short reads sampled per long read simulated the real-world situation of oversampling, albeit with less bias than the first set. In addition, the set containing only one short read per long read was independently replicated six times to simulate the lack of oversampling that would be obtained from an environmental sample containing a diverse range of microorganisms or viruses.

(iii) Increased lengths of short reads. To simulate the effect of longer read length, the first 150, 200, 250, 300, 350, and 400 bp of each of the 1,000 long-read sequences were sampled and used in six independent analyses.

BLAST analysis. Each data set (long reads and short reads) was analyzed with NCBI BLASTX (version 2.2.13) against the GenBank nr database (February 8, 2006 release). The output was collected in XML format and analyzed as described below. An E-score (expect value) cutoff of 10^{-3} or 10^{-5} was applied, and the top five hits were collected (options: blastall -p blastx -v 5 -b 5 -e 1e-3 -m 7).

(i) BLASTX output analysis. For each sequence, the top BLASTX hit with an E-score of $<10^{-3}$ (all short reads, 100 to 400 bp) or E-score of $<10^{-5}$ (long reads) was collected, generating two files that were further analyzed independently. The hit rate (total hit/total query reads) was calculated for each cutoff. Information on the subject sequence length, hit description and GenBank accession number, hit length, and hit E-score was collected for each positive BLAST result.

(ii) Comparative analysis of BLASTX results between long-read and short-read data sets. BLASTX results were compared for each long-read reference sequence and its corresponding short read(s) to estimate the level of concordance between the two data sets. When both sequences hit the same GenBank nr sequence (i.e., same GenBank accession number), the protein homolog of the short read was considered to be the same as that of the long read. When the BLASTX hit for the short read was different from the long read, the results were examined to decide if both hits were similar. For the tiled-and-overlapping and the multiple-short-read data sets, the BLASTX results for each set of short reads and their originating long read were compared. If one or more of the multiple short reads had a BLASTX hit similar to that of the long read, this was scored as an agreement between the two data sets. The total number of positive agreements was expressed as a percentage of the number of original long reads (e.g.,

FIG. 1. Long- and short-read BLAST homolog hit frequency for increasing levels of short-read sampling. (A) Chesapeake Bay virioplankton metagenome sequence data (average long-read length, 681 bp). (B) AMD microbial metagenome sequence data (average long-read length, 812 bp). (C) Sargasso Sea microbial metagenome sequence data (average long-read length, 877 bp). Black bars show frequency of significant BLASTX hits for each sequence category. Significance levels were E-scores of $<10^{-5}$ and E-scores of $<10^{-3}$ for long and short reads, respectively. Light gray bars show median E-scores of BLAST-positive long-read sequences for which at least one derived short read also had a BLAST hit. Dark gray bars show median E-scores of the long-read hits for which no derived short reads were BLAST positive. For the overlapping-short-read experiment, the average numbers of short reads per long read were 7.6, 9.6, and 10.4 for the Chesapeake Bay, acid mine, and Sargasso libraries, respectively. The data for average single short reads are the means of the results of six independent experiments.

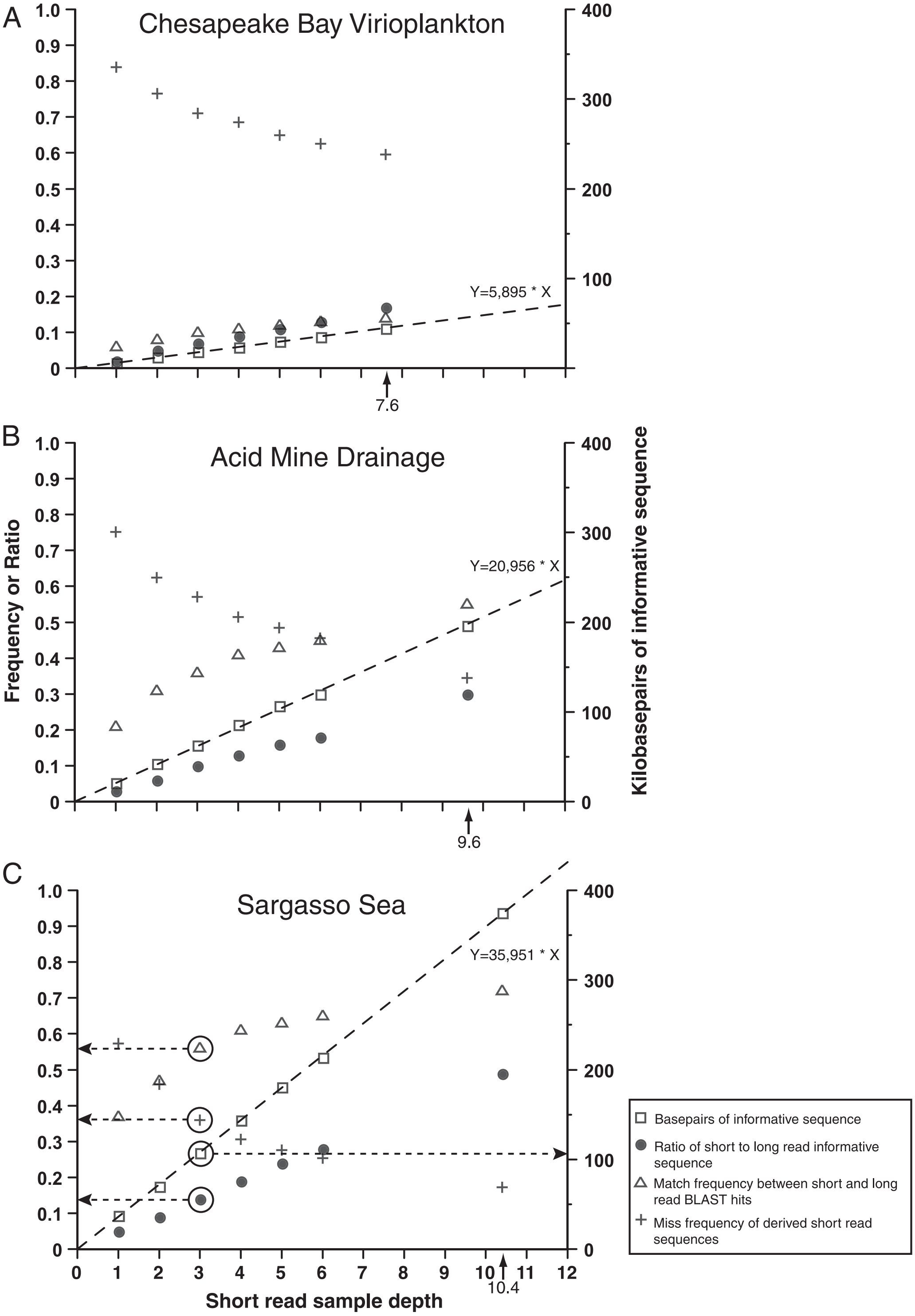


FIG. 2. Comparison of long- and short-read BLAST homolog hit frequency and amount of informative sequence for increasing levels of short-read sampling. (A) Chesapeake Bay virioplankton metagenome sequence data. (B) AMD microbial metagenome sequence data. (C) Sargasso Sea microbial metagenome sequence data. Open squares show base pairs of informative sequence. Significance levels for BLAST homologs were E-scores of $<10^{-5}$ and E-scores of $<10^{-3}$ for long and short reads, respectively. Closed circles show ratios of short- to long-read informative sequence. Open triangles show match frequencies between BLAST homologs of derived short reads and the original long-read sequence. Plus signs show miss frequencies of derived short-read sequences. As shown for the Sargasso Sea library (C; circled data points) at a sample depth of three short reads per long read, 56% of the derived short reads matched the original long-read BLAST hit; 36% of the short reads missed the BLAST homolog found by the original long read; and short reads yielded 107.5 kb of informative sequences, which was 14% of the amount contained in the original long-read Sargasso Sea data set. Numbers noted on the ordinate axis are the average number of derived short reads needed to completely cover a long-read sample with a 20-bp overlap between sequences (tiled-and-overlapping short reads). Regression line represents the rate of informative sequence acquisition with increasing levels of short-read sampling. The data for average single short reads are the means of the results of six independent experiments.

if 103 positives were scored, the percent hit rate for the short-read data set was reported as 10.3% [103/1,000 long reads]). Alternatively, when the short reads did not have a BLASTX hit but the originating long read did, this was expressed as a percentage of the total number of BLASTX hits for the long-read reference library (e.g., out of the 373 BLASTX hits to long-read sequences in the Chesapeake Bay virioplankton data set, 250 were not found in the short-read data set, yielding a 71% [250/373] miss rate for short-read sequences).

(iii) Estimate of the quality of homologs missed by short reads. To determine whether the tendency for short- and long-read sequences to find the same BLASTX homolog was randomly distributed among the population of BLASTX-positive long reads, the median E-score of the originating long-read sequences was determined for the two instances described above (i.e., agreement between short and long reads and no hit for the short read but a hit for the originating long read) (Fig. 1 and 2).

(iv) Estimate of the amount of informative sequence. The total number of base pairs of information was estimated by summing the lengths of all reads within a data set that had a significant BLASTX hit to the GenBank nr database. For the tiled-and-overlapping short-read data set and the data sets which consist of multiple short reads per long read, the total number of BLASTX-positive sequences was used, irrespective of whether multiple short reads hit the same GenBank entry.

(v) Functional annotation. A functional distribution analysis of the BLASTX hits and misses was generated by comparing the protein sequences of the hits from each long-read data set to the database of Clusters of Orthologous Groups of proteins (COG) version 2, which consists of 138,458 proteins forming 4,873 COGs from 66 unicellular genomes (24). For each long-read data set, we collected from GenBank the protein sequence of each top BLASTX hit. These proteins were compared to the COG database using NCBI BLASTP (options: blastall -p blastp -v 5 -b 5 -e 1e-3 -m 7). The output allowed the assignment of each protein to a functional category. The results were reported as the percentage of the total proteins for each category. To analyze the functional annotation of the proteins for which the long reads had homology but which the short reads missed, these missed proteins were analyzed as described above and the functional distribution of these proteins was compared to that of all the long-read hits.

RESULTS

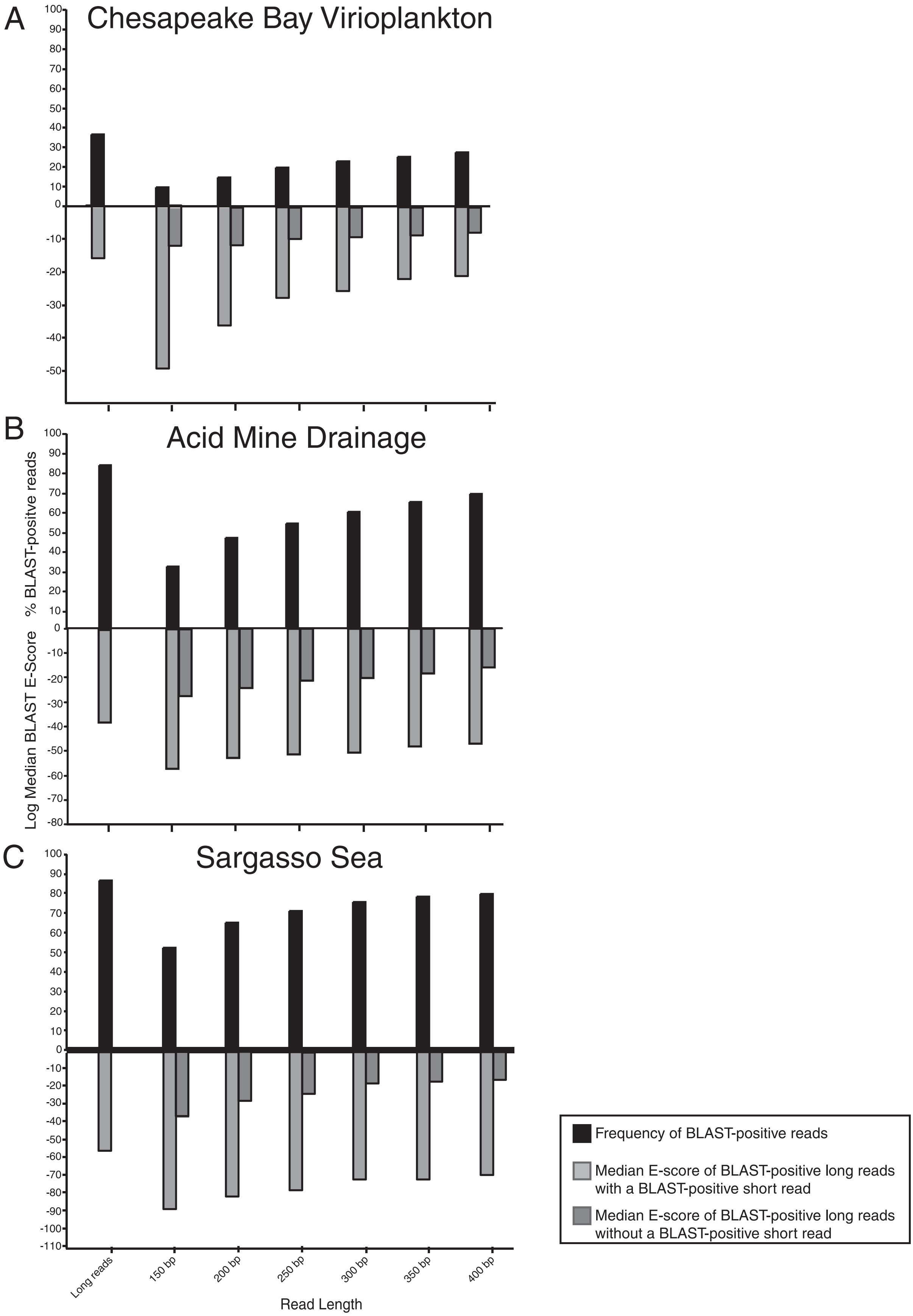
Several treatments were applied to a randomly selected, 1,000-sequence subset of each of three long-read metagenome sequence libraries. Each treatment was designed to simulate various scenarios for sampling a microbial or viral metagenome using high-throughput short-read sequencing approaches. To account for the lower BLAST expect value that occurs with shorter regions of homology, short reads were considered BLAST-positive at an E-score of $<10^{-3}$, while a significance cutoff of an E-score of $<10^{-5}$ was used for long reads. The high E-score of a $<10^{-3}$ significance cutoff has been used in previous analyses of short-read metagenome sequence data (10, 21).

Viral assemblages are genetically divergent. For the subset of 1,000 long-read sequences of each metagenome, the frequencies of BLASTX homologs were 87, 83, and 37% for the Sargasso and AMD microbial and Chesapeake Bay virioplankton metagenomes, respectively. These values are reflective of those seen in the larger sequence data set, validating the sampling approach used in this study. The quality of BLASTX alignments, as determined by the log converted median E-score of BLASTX-positive sequences, was highest for the Sargasso Sea sequences (log median E-score of −56), lowest for the Chesapeake Bay virioplankton (−16.6), and intermediate for the AMD subsample (−39) (Fig. 1 and 3). Again, these values were similar to those of the larger data set and are an indication that bacterial sequences from the Sargasso Sea are genetically closer to known bacterial sequences within the GenBank nr database than are bacterial sequences within the AMD metagenome. Together, the dramatically higher log median E-score and low BLASTX hit frequency for the Chesapeake Bay virioplankton sequences confirm that known viral genes within GenBank poorly represent the genetic diversity present within autochthonous communities of marine viruses.

Most long-read BLAST homologs are missed by a single, randomly selected short read. An in silico experiment in which a single 100-bp fragment was randomly selected from each long read and then subjected to BLASTX homology searching was repeated six times. The averages ± standard deviations of BLASTX hit frequency for short reads across each of these experiments were 35% ± 1.8%, 20 ± 3.4% and 6 ± 0.7% for the Sargasso and AMD microbial and Chesapeake Bay virioplankton metagenomes, respectively (Fig. 1). In all cases, the originating long-read sequence was also BLASTX positive, and the average log converted median E-scores of the BLASTX alignments for these long reads were −101 ± 4.4, −70 ± 3.5, and −55 ± 7.5 for the three libraries, respectively. These log converted median E-scores were substantially lower than the overall values for the long-read sample libraries, indicating that short-read sequences preferentially find highly similar sequences within the subject database.

Alternatively, for the majority of BLASTX-positive long-read sequences, the derived short reads did not find a BLAST homolog. The miss frequency for short-read experiments (Fig. 2 and 4) was the ratio of BLASTX-positive long-read sequences with no corresponding BLASTX-positive derived short reads to the total number of BLASTX-positive long-read sequences. Overall, the short-read sequence libraries missed 60% ± 0.02%, 76% ± 0.04%, and 85% ± 0.02% of the BLAST homologs found by long-read sequences in the Sargasso, AMD, and Chesapeake Bay data sets, respectively (Fig. 2). The corresponding average log median E-scores for the missed long-read sequences were only slightly higher (−41 ± 1, −30 ± 1.7, and −15 ± 0.3) than the overall values for each of the long-read subsample libraries (Fig. 1). Thus, short reads tended to miss more-distant genetic homologs that were found using long-read sequences.

All levels of short-read sampling miss BLAST homologs found with long reads. To examine the effect of increasing the short-read sampling depth, a series of in silico experiments were performed in which the number of randomly sampled short reads was increased stepwise from two to six per long-read sequence. An additional experiment was also performed in which short reads were tiled with a 20-bp overlap across each long-read sequence. This case reflects an idealized oversampling of genetic sequences within a metagenome. For the Sargasso data, the BLASTX hit rate across the two- to six-short-read series increased from 46 to 65%, while the miss rate (i.e., instances where short reads did not find a BLASTX homolog but the originating long read did) dropped from 47 to 25% (Fig. 1C and 2C). The largest percent rise in the hit rate and fall in the miss rate occurred at between one and two random short sequences per long-read sequence. Each step increase in the short-read sampling rate resulted in ever smaller improvements in BLAST homology searches. The idealized case of tiled-and-overlapping short reads for each long-read sequence had the highest hit rate (72%) (Fig. 1C), yet 17% of the BLASTX homologs found with the long sequences were missed by the short reads (Fig. 2C). The miss frequency of


A
Chesapeake Bay Virioplankton
B
Acid Mine Drainage
C
Sargasso Sea
% BLAST-positive reads
Log Median BLAST E-Score
100
90
80
70
60
50
40
30
20
10
0
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
Long reads
150 bp
200 bp
250 bp
300 bp
350 bp
400 bp
Read Length
Frequency of BLAST-positive reads
Median E-score of BLAST-positive long reads with a BLAST-positive short read
Median E-score of BLAST-positive long reads without a BLAST-positive short read

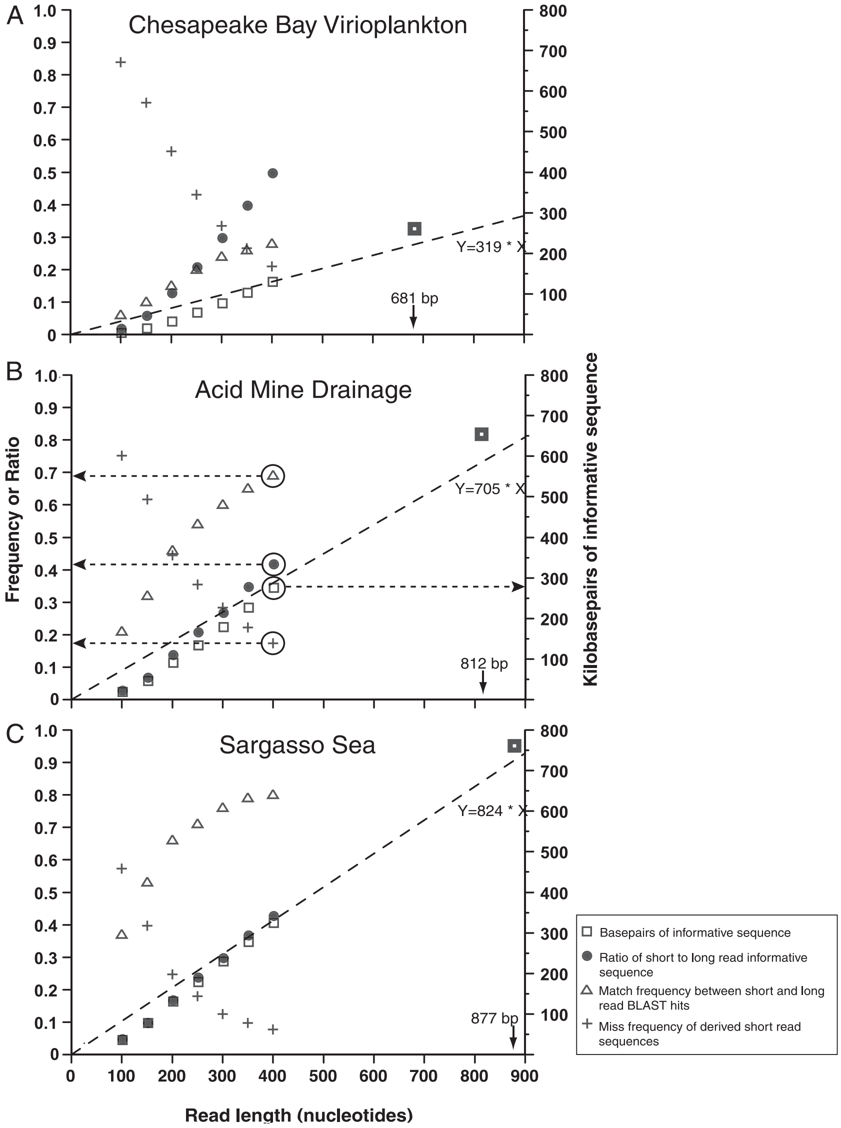


FIG. 4. Comparison of long- and short-read BLAST homolog hit frequency and amount of informative sequence for increasing length of short-read sequences. (A) Chesapeake Bay virioplankton metagenome sequence data. (B) AMD microbial metagenome sequence data. (C) Sargasso Sea microbial metagenome sequence data. Open squares show base pairs of informative sequence. Bold open squares represent the original long-read data sets. Significance levels for BLAST homologs were E-scores of $<10^{-5}$ and E-scores of $<10^{-3}$ for long and short reads, respectively. Closed circles show ratios of short- to long-read informative sequence. Open triangles show match frequencies between BLAST homologs of derived short reads and the original long-read sequence. Plus signs show miss frequencies of derived short-read sequences. As shown for the AMD library (B), at a short-read length of 400 bp, 69% of the derived short reads matched the original long-read BLAST hit; 17% of the short reads missed the BLAST homolog found by the original long read; and short reads yielded 277.6 kb of informative sequences, which was 42% of the amount contained in the original long-read acid mine data set. Numbers noted on the ordinate axis are average read lengths for the original long-read data set. Regression line represents the rate of informative sequence acquisition with increasing sequence length.

short reads in the tiled-and-overlapping experiment means that none of the derived short reads from a given sequence had significant ($E \leq 10^{-3}$) homology to a sequence in the subject BLAST database. The results of this experiment illustrate that the homology-finding capability of short reads is limited ultimately by their length, not by the overall sequence coverage of the short-read library.

A similar trend of increases in the hit rate and decreases in

FIG. 3. Long- and short-read BLAST homolog hit frequency for random derived read lengths between 150 and 400 bp. (A) Chesapeake Bay virioplankton metagenome sequence data (average long-read length, 681 bp). (B) AMD microbial metagenome sequence data (average long-read length, 812 bp) (C) Sargasso Sea microbial metagenome sequence data (average long-read length, 877 bp). Black bars show frequency of significant BLASTX hits for each sequence category. Significance levels were E-scores of $<10^{-5}$ and E-scores of $<10^{-3}$ for long and short reads, respectively. Light gray bars show median E-scores of BLAST-positive long-read sequences for which at least one derived short read also had a BLAST hit. Dark gray bars show median E-scores of the long-read hits for which no derived short reads were BLAST positive

the miss rate with increasing short-read sampling levels was also seen in the AMD and Chesapeake Bay data sets (Fig. 1A and B and 2A and B). However, as with the single-read experiment, the magnitude of the miss rate at all levels of short-read sampling was highest for the Chesapeake Bay virioplankton sequences. In the tiled-and-overlapping experiment, nearly 60% of the hits found with long-read Chesapeake Bay virioplankton sequences were missed by the short reads (Fig. 2A).

Short reads tend to miss genetically distant sequences. Despite increasing levels of sampling, derived short reads consistently failed to detect more-distant BLASTX homologs (higher E-score) of long-read sequences. Here, two cases are considered: (i) the median E-score of long-read BLASTX-positive sequences for which at least one derived short read also had a BLAST hit (match); and (ii) the median E-score of long-read hits for which no derived short reads were BLASTX positive (miss). Across all three metagenomes, as the level of derived short-read sampling increased, so did the long-read median E-scores for both matches and misses (Fig. 1). The increases in the long-read median E-scores for match and miss cases were greatest for the Sargasso Sea data set. From a sampling depth of one to a sampling depth of six short reads per long read, the median E-score increased by ca. 16 logs for both matches and misses. These increases in the median E-score rose to 26 and 23 logs for matches and misses, respectively, across the one-short-read to tiled-and-overlapping short-read sampling gradient. While the idealized short-read oversampling experiment improved BLAST homolog discovery, many significant hits were missed, as the median E-score for the 149 missed Sargasso Sea long reads in the tiled-and-overlapping experiment was 2×10^{-18} (Fig. 1C). In contrast, increased levels of short-read sampling within the AMD and Chesapeake Bay libraries led to only slight increases in the median E-score. For the one-short-read to tiled-and-overlapping short-read sampling gradient, the match and miss median E-scores increased by ~10 logs in the AMD library, with a dramatic jump of 8 logs in the match score between the six-short-read and the tiled-and-overlapping sampling (Fig. 1b). This same sampling gradient in the Chesapeake Bay library produced a 9-log increase in the match median E-score but only a 3-log increase in the miss median E-score (Fig. 1A). This relative lack of change in the miss median E-score for the Chesapeake Bay library reflects the fact that, for a proportion of BLASTX-positive long-read sequences, no level of short-read sampling is capable of detecting sequence homologs.

Long-read metagenome libraries yield more-informative sequence. Each BLAST sequence homolog within a metagenome library can be viewed as a component piece of information necessary for describing the phylogenetic and functional diversity within a microbial community. The amount of information contained within each long-read and derived short-read metagenome library was quantified by summing the total sequence length of BLASTX-positive sequences (Fig. 2). The significantly higher BLASTX hit rates and longer read lengths of the AMD and Sargasso 1,000-read-subsample libraries resulted in 676 and 760 kb of informative sequence, respectively (Fig. 4B and C). In contrast, less than half this amount of information (254 kb) was derived from the Chesapeake Bay virioplankton subsample library (Fig. 4A). The rate at which informative sequence was gained with increasing levels of short-read sampling differed sharply between the libraries and reflected the poor representation of viral sequences within GenBank. For the Chesapeake Bay virioplankton library, only 5.9 kb of informative sequence was gained for each step increase in the number of 100-bp short-read samples from each long read (Fig. 2A). In contrast, 36 and 20.5 kb of informative sequence per short-read sampling were gained for the Sargasso and AMD libraries, respectively (Fig. 2B and C). The bacterial and viral libraries also differed when the short-read match frequency and the ratio of long- to short-read informative sequence were compared. For the Chesapeake Bay virioplankton library, the match frequency and short- to long-read informative sequence ratio were very similar (Fig. 2A); however, for the AMD and Sargasso libraries, the match rate was more than double the informative sequence ratio. In the case of the Sargasso library, the relatively high short-read match rate of 65% in the six-short-read sampling experiment yielded only 28% of the informative sequence found through long reads (Fig. 2C).

Short reads (≤400 bp) miss a significant amount of the BLAST homologs found with long reads. The current average read length obtained through pyrosequencing (17) technology is ca. 100 bp; however, improvements to the technology have increased the average read length to ca. 200 bp. Read lengths of up to 500 bp have been achieved in the laboratory, but a commercially available instrument capable of producing such reads is at least two generations in the future (12). To test the efficacy of future improvements in pyrosequencing technology, simulations were run in which the sampling rate (i.e., the number of derived short reads from each long read) was held to 1, while the length of the random derived short reads was increased from 150 to 400 bp. For all short-read lengths, the frequency of BLAST-positive sequences was less than that for long reads (Fig. 3). In the case of the 400-bp short reads, the short-read miss rates were 21%, 17%, and 8% for the Chesapeake Bay virioplankton, AMD, and Sargasso libraries, respectively (Fig. 4). Similar to the results of the short-read sampling rate experiments, none of the short-read-length scenarios tested were capable of detecting all of the BLASTX homologs found with the original long-read sequences. Increasing the short-read length improved the match frequency with long reads which had lower quality (higher E score) BLAST homology (Fig. 3).

Longer short reads (150 to 400 bp) are better than higher sampling depth with 100-bp reads. It is interesting to note that the recent improvements in pyrosequencing technology (i.e., ~200 bp) improve the BLASTX miss rate to a level nearly equal to that found in the tiled-and-overlapping experiment (Fig. 2 and 4). This effect is most pronounced for the Chesapeake Bay virioplankton library, where the miss rates were 43% for the 250-bp reads and ca. 60% for the tiled-and-overlapping short-read sampling experiment (Fig. 2A and 4A). Thus, high levels of 100-bp sampling within a viral or bacterial community will not yield as much information as lower coverage using 250-bp sequences. The improvements in homolog discovery by using increasingly longer sequences were most pronounced for the Chesapeake Bay virioplankton library, where the miss frequency of the 400-bp sequences was fivefold lower than that of the 100-bp reads. For the Sargasso and

AMD libraries, the improvements were more modest, at 2.3- and 3.4-fold, respectively (Fig. 4). Dramatic differences in the acquisition of informative sequence with increasing short-read length were readily apparent between the microbial and viral metagenome sequence data. For the AMD and Sargasso libraries, ~35 kb and ~41 kb, respectively, of informative sequence were added for each 50-bp increase in short-read length, whereas less than half this rate was seen for the Chesapeake Bay virioplankton sequences (Fig. 4). Here again, the slower rate is a result of the overall low BLAST homology frequency for environmental viral sequences.

Short reads miss whole gene function classes within a metagenome library. To determine whether biases occurred among the types of protein groups detected with short-read sequences, the distribution of COGs among BLAST-positive long-read sequences was compared to the match frequency of short-read sequences. Among the three libraries, the Chesapeake Bay virioplankton contained the highest frequency of BLAST-positive sequences for which no COG could be assigned (see Fig. S1 to S3 in the supplemental material). For the AMD and Chesapeake Bay libraries, none of the 100-bp sampling levels was capable of detecting all COGs detected with long-read sequences (see Fig. S1 and S2 in the supplemental material). With the exception of those for replication, recombination, and repair (35 sequences) and nucleotide transport and metabolism (11 sequences), most COGs were rare within the original set of long-read Chesapeake Bay virioplankton sequences. For these two COGs, no level of sampling captured all of the BLAST homologs found with long-read virioplankton sequences. At a level of two random 100-bp reads per long read, nearly all COGs detected in the long-read Sargasso library were also detected at least once by derived short reads. Nevertheless, (with the exception of two rare COGs, for cell motility and cytoskeleton), at the highest sampling rate (six short reads per long read), the per-COG short-read match frequencies were ca. 75% or less for the Sargasso metagenome data (see Fig. S3 in the supplemental material). This match frequency is similar to that seen for the entire data set at a derived short-read sampling rate of 6 (65%) (Fig. 2) and indicates that, for this microbial community, short reads would have missed all protein groups with similar frequency.

Increasing short-read length improves detection of gene function classes. Similar to the global analyses of BLAST homolog frequency, longer single short reads (150 to 400 bp) performed better at detecting COGs than high levels of sampling with 100-bp reads. For all three libraries, the entire set of COGs detected with long reads were also detected with single short reads of 150 bp (AMD and Sargasso) or 300 bp (Chesapeake Bay). Within the Chesapeake Bay data set, 74 to 91% of the BLAST homologs within the two most common COGs were detected with a single short read of 300 bp. Not surprisingly, the highest match frequencies across COGs were seen for the 400-bp derived short reads from the Sargasso Sea sequences, yet only 5 of the 22 COGs were matched with 100% frequency by the 400-bp derived sequences (see Fig. S3 in the supplemental material). For the AMD library, short-read match frequencies of 100% to a given COG (cell motility) only occurred at the 400-bp short-read length (see Fig. S2 in the supplemental material).

DISCUSSION

Based on prior information (21), the underlying assumption of these analyses was that the original long-read sequences would find significantly more BLASTX homologs than short reads. Overall in these experiments, this assumption was upheld, as all BLASTX homologs found with short reads were also found with long reads. Thus, in these in silico analyses, short reads did not discover gene homologs that were not found using the originating long-read sequences.

Determining the connections between the composition (structure) and physiology (function) of microbial communities is a central objective in synecological studies of microorganisms. Currently, the application of high-throughput sequencing approaches to the characterization of genomic DNA from autochthonous microorganisms provides the most-encompassing and least-biased means of approaching this objective (15). It is also true that, recognizing the complexity of most microbial communities, increasing levels of sequence information obviously should lead to greater resolution and accuracy in determining the phylogenetic diversity and functional capabilities within a microbial community. Thus, it is not surprising that the low cost and high sequence yield of new, short-read sequencing-by-synthesis technologies make them appear to be an attractive option for microbial metagenomic investigations. However, sampling depth is not the only limitation to these investigations, as effective characterization of microbial diversity and function from metagenome data ultimately relies on sequence homology to genes of known or putative function and taxonomic origin. In total, these data simulation experiments decisively show that, in comparison to established Sanger sequencing technology, short reads have a significantly reduced ability to detect genetic homology to known proteins using BLAST.

This effect was most notable for metagenome sequence data from natural viral communities. Across the few Sanger-read viral community metagenome shotgun sequence libraries reported to date (2–5, 11), the translated BLAST hit rate against the GenBank nr database has been around 35%. This homolog frequency is similar to that commonly reported for whole-phage genomes and is a reflection of the relatively poor representation of phage genetic diversity within databases of known proteins (19). In contrast, the BLAST homolog frequencies for short-read libraries of virioplankton from four marine provinces were between 1 and ~11% against known sequences (1). Derived-short-read data simulation experiments from the Chesapeake Bay virioplankton metagenome library found similarly low levels of translated BLAST homolog detection. Thus, regardless of the environment sampled, the frequency of homolog detection using translated BLAST analysis will be low for short-read sequences from viral communities.

An obvious counterargument to the reduced ability of short reads to find homologous known proteins is the larger total amount of sequence data obtained from a single run of 454 pyrosequencing (ca. 30 to 40 Mbp). Thus, greater overall sequencing depth should compensate for the lower homolog detection frequency and ultimately result in a more complete and comprehensive characterization of the microbial community. Two aspects of the results of these data simulation experiments do not support this assumption. First, in the subset of cases where derived short reads had BLAST homologs, the originating long read tended to have a high-quality (low E-score) BLAST hit. Short reads derived from long reads with more-distant BLAST homology (high E-score) typically did not find a BLAST homolog, as evidenced by the higher median E-score of long reads without a matching BLAST-positive short read. This trend occurred across all levels of short-read sampling. Thus, short reads tend to detect only well-known or conserved proteins that are easily detected with longer reads at a higher overall confidence. Second, the analysis of derived short reads according to COG functional category showed that, for frequent COGs, even high levels of short-read sampling did not detect all of the homologs detected with long reads. This same trend also held for many of the less-frequent COGs detected within the long-read data sets. In no case did a short read detect a COG or BLAST homolog that was not detected by the originating long read. A recent report characterizing prokaryote assemblages within two AMD environments from short-read metagenome sequences confirms the poor ability of short reads to describe the functional characteristics of a microbial community (10). In the case of the library for the "Red" AMD environment, no protein homologs were detected for 9 of the 21 functional groups examined. In contrast, a long-read library from another AMD environment (26) detected all but three functional groups. While it is possible that the "Red" environment contained such a high density of novel genes that no homologs to a critical functional group such as "nucleosides and nucleotides" were detected, the more-parsimonious explanation is that the short reads did not provide sufficient sensitivity to detect known protein homologs by translated BLAST analysis.

Overall, the performance of the simulated short-read data was similar to reported 454 pyrosequencing metagenome data from other environments with regard to both BLAST homolog frequency and COG discovery. The only report comparing 454 and Sanger metagenome sequence libraries obtained from the same sample (cecal contents of mice) found notable differences in COG detection frequency between short and long reads and a substantially lower assignment of short reads to homologs within the nr or COG database (25). In addition, this experiment showed even, lower short-read BLAST homolog and COG detection ability than was predicted from our data. The collection of short-read data simulation experiments reported here represents unachievable levels of short-read sequencing accuracy, as no attempt was made to replicate the error rate or homopolymer failure rate of 454 sequencing. In reality, the individual-read insertion-deletion error rates of 454 sequences (ca. 1.7%) are significantly higher than that known for Sanger sequencing (17). Because accurate assembly of metagenome sequence data from a diverse microbial community can be difficult, most studies have relied on BLAST analysis of single reads for the functional and taxonomic characterization of microbial diversity. In the case of 454 pyrosequencing, such a high insertion-deletion error rate would produce frame shifts within the translations used for BLASTX analysis and further lessen the sensitivity and accuracy of short reads for the characterization of microbial diversity.

Besides the assumption that higher sequence yields will return more information despite the lower hit frequency, the dramatically lower per-bp cost, elimination of clone library construction, and speed of 454 pyrosequencing are often cited as a justification for the use of this technology in metagenomics. To test whether the lower per-bp cost actually leads to a lower overall cost of information, we determined the amount of Sanger sequencing reads necessary to match the overall amount of informative viral metagenome sequence data in the short-read virioplankton libraries reported by Angly et al. (1). Assuming an average read length of 750 bp at a cost of $1.25 per lane (template preparation and sequencing) plus $1,500 for clone library construction, the Sanger libraries would have cost ca. $7,100, $5,500, and $3,400 for the Gulf of Mexico, coastal British Columbia, and Sargasso libraries, respectively. Each of these cost estimates is comparable to or below the cost of a single 454 pyrosequencing run (ca. $9,000). Only in the case of the Arctic library, which comprised two 454 runs, would the cost of a Sanger library have been substantially higher (~$18,000 versus $38,500). This was largely due to the higher number of prophage-like sequences within the Arctic library and the fact that these virioplankton assemblages were estimated to contain a relatively low number of viral genotypes (532 genotypes; ~30 to ~200 times fewer) in comparison to the numbers in the other three environments (1). While this per-lane cost estimate is lower than that of commercial sequencing services, it is ~40% higher than the per-lane charges of nonprofit sequencing centers, such as the Joint Technology Center of the J. Craig Venter Institute.

Several studies utilizing 454 pyrosequencing have posited the elimination of cloning bias as an important advantage of this technology for metagenomic analysis of microbial communities (1, 10, 25). However, no studies have empirically determined the level of analytical bias attributable to the cloning step required for Sanger sequencing of a microbial metagenome. The closest approximation of such an experiment was a comparison of Sanger and 454 technologies for whole-genome sequencing of six marine bacterial strains (13). In that study, the G+C content and size of each bacterial genome were similar to those commonly seen within oceanic bacterioplankton communities. With 6- to 10-fold coverage, the physical gaps from shotgun cloning and Sanger sequencing comprised between 0 and 5.7% (mean, 1.4%) of the genomes. By comparison, the gaps from 454-only sequencing at 12- to 36-fold coverage comprised between 0.2 and 23% (mean, 12%) of the genomes. In all cases, 454-only sequencing missed a significantly larger proportion of the genome. The most egregious example was the *Erythrobacter litoralis* (HTCC 2594) genome, where >243 kb was missing from the 454 genome data compared to 632 bp for Sanger sequencing.

A recent analysis of cloning-bias data from 79 bacterial whole-genome sequencing projects revealed that the overall number of "unclonable" genes was small (0.6%). Biases in cloning efficiency across the genome projects were found to be largely gene dependent and related to both the expression level and copy number of the insert DNA (23). The influence of these factors on clonability can be minimized through the use of low-copy-number, transcription-terminated plasmid cloning vectors (e.g., pSMART; Lucigen Corp. [20]). In the case of the Chesapeake Bay virioplankton metagenome library, a transcription-terminated vector was used to allow for the cloning of potentially toxic bacteriophage genes (2). While cloning biases are well known from whole-genome investigations and a number of technical approaches have been devised to overcome them, it is clear that the cloning-independent nature of 454 pyrosequencing is probably not a panacean rationale for the use of this technology in metagenomics. On the contrary, library construction can be viewed as a significant advantage of Sanger sequencing approaches, as metagenome clone libraries can be an important resource for subsequent investigations of gene function or diversity.

The overwhelming recommendation provided by these short-read data simulation experiments is that longer-read Sanger sequencing is superior to short reads for the characterization of the functional and taxonomic composition of viral and prokaryotic communities. Indeed, the best methodological approach to this objective remains paired-end Sanger sequencing of small-insert (2 to 3 kb) clone libraries. Longer, high-quality, paired-end reads allow for more-accurate open reading frame detection and significantly greater sensitivity in the detection of sequence homology by BLAST. Nevertheless, short-read sequencing technologies do have significant utility for gap-closure in whole-genome sequencing projects (13) and for rapid sequencing of large inserts within bacterial artificial chromosome and fosmid vectors (28). In these cases, the lack of a cloning step and significant time savings of short-read sequencing technologies offer a significant advantage over Sanger sequencing.

ACKNOWLEDGMENTS

We are grateful for the support of National Science Foundation Microbial Observatories grant MCB-0132070 to K.E.W.

We are indebted to the Delaware EPSCoR program for support consisting of the computational time needed for BLAST analyses. This investigation would not have been possible without the Chesapeake Bay and acid-mine drainage metagenome sequence data sets graciously provided by Shellie R. Bench and Jill Banfield, respectively.

REFERENCES

1. **Angly, F. E., B. Felts, M. Breitbart, P. Salamon, R. A. Edwards, C. Carlson, A. M. Chan, M. Haynes, S. Kelley, H. Liu, J. M. Mahaffy, J. E. Mueller, J. Nulton, R. Olson, R. Parsons, S. Rayhawk, C. A. Suttle, and F. Rohwer.** 2006. The marine viromes of four oceanic regions. PLoS Biol. **4:**e368.
2. **Bench, S. R., T. E. Hanson, K. E. Williamson, D. Ghosh, M. Radosovich, K. Wang, and K. E. Wommack.** 2007. Metagenomic characterization of Chesapeake Bay virioplankton. Appl. Environ. Microbiol. **73:**7629–7641.
3. **Breitbart, M., B. Felts, S. Kelley, J. M. Mahaffy, J. Nulton, P. Salamon, and F. Rohwer.** 2004. Diversity and population structure of a near-shore marine-sediment viral community. Proc. Biol. Sci. **271:**565–574.
4. **Breitbart, M., I. Hewson, B. Felts, J. M. Mahaffy, J. Nulton, P. Salamon, and F. Rohwer.** 2003. Metagenomic analyses of an uncultured viral community from human feces. J. Bacteriol. **185:**6220–6223.
5. **Breitbart, M., P. Salamon, B. Andresen, J. M. Mahaffy, A. M. Segall, D. Mead, F. Azam, and F. Rohwer.** 2002. Genomic analysis of uncultured marine viral communities. Proc. Natl. Acad. Sci. USA **99:**14250–14255.
6. **Chen, F., K. Wang, J. J. Kan, D. S. Bachoon, J. R. Lu, S. Lau, and L. Campbell.** 2004. Phylogenetic diversity of *Synechococcus* in the Chesapeake Bay revealed by ribulose-1,5-bisphosphate carboxylase-oxygenase (RuBisCO) large subunit gene (rbcL) sequences. Aqua. Microb. Ecol. **36:**153–164.
7. **DeLong, E. F.** 2005. Microbial community genomics in the ocean. Nat. Rev. Microbiol. **3:**459–469.
8. **DeLong, E. F., C. M. Preston, T. Mincer, V. Rich, S. J. Hallam, N. U. Frigaard, A. Martinez, M. B. Sullivan, R. Edwards, B. R. Brito, S. W. Chisholm, and D. M. Karl.** 2006. Community genomics among stratified microbial assemblages in the ocean's interior. Science **311:**496–503.
9. **DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen.** 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. **72:**5069–5072.
10. **Edwards, R. A., B. Rodriguez-Brito, L. Wegley, M. Haynes, M. Breitbart, D. M. Peterson, M. O. Saar, S. Alexander, E. C. Alexander, Jr., and F. Rohwer.** 2006. Using pyrosequencing to shed light on deep mine microbial ecology under extreme hydrogeologic conditions. BMC Genomics **7:**57.
11. **Edwards, R. A., and F. Rohwer.** 2005. Viral metagenomics. Nat. Rev. Microbiol. **3:**504–510.
12. **Egholm, M.** 2006. What is happening in next generation sequencing?, p. 15. Genomes Med. Environ. Conf., Hilton Head, SC, 16 to 18 October 2006.
13. **Goldberg, S. M., J. Johnson, D. Busam, T. Feldblyum, S. Ferriera, R. Friedman, A. Halpern, H. Khouri, S. A. Kravitz, F. M. Lauro, K. Li, Y. H. Rogers, R. Strausberg, G. Sutton, L. Tallon, T. Thomas, E. Venter, M. Frazier, and J. C. Venter.** 2006. A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proc. Natl. Acad. Sci. USA **103:**11240–11245.
14. **Govind, R., J. A. Fralick, and R. D. Rolfe.** 2006. Genomic organization and molecular characterization of *Clostridium difficile* bacteriophage PhiCD119. J. Bacteriol. **188:**2568–2577.
15. **Handelsman, J.** 2004. Metagenomics: application of genomics to uncultured microorganisms. Microbiol. Mol. Biol. Rev. **68:**669–685.
16. **Mahenthiralingam, E., and P. Drevinek.** 2007. Comparative genomics of Burkholderia species, p. 53–81. *In* T. Coenye and P. Vandamme (ed.), Burkholderia: molecular microbiology and genomics. Horizon Scientific Press, New York, NY.
17. **Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg.** 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature **437:**376–380.
18. **Minz, D., J. L. Flax, S. J. Green, G. Muyzer, Y. Cohen, M. Wagner, B. E. Rittmann, and D. A. Stahl.** 1999. Diversity of sulfate-reducing bacteria in oxic and anoxic regions of a microbial mat characterized by comparative analysis of dissimilatory sulfite reductase genes. Appl. Environ. Microbiol. **65:**4666–4671.
19. **Pedulla, M. L., M. E. Ford, J. M. Houtz, T. Karthikeyan, C. Wadsworth, J. A. Lewis, D. Jacobs-Sera, J. Falbo, J. Gross, N. R. Pannunzio, W. Brucker, V. Kumar, J. Kandasamy, L. Keenan, S. Bardarov, J. Kriakov, J. G. Lawrence, W. R. Jacobs, Jr., R. W. Hendrix, and G. F. Hatfull.** 2003. Origins of highly mosaic mycobacteriophage genomes. Cell **113:**171–182.
20. **Riesenfeld, C. S., P. D. Schloss, and J. Handelsman.** 2004. Metagenomics: genomic analysis of microbial communities. Annu. Rev. Genet. **38:**525–552.
21. **Rohwer, F., R. Edwards, M. Breitbart, S. Kelley, P. Salamon, J. Nulton, B. Felts, J. Mahaffy, J. Mueller, and C. Carlson.** 2006. Metagenomic analysis reveals biogeography within marine virus communities and that sequences from cyanophages and ssDNA viruses are common, session TS-B19. ASLO Summer Meet. 2006, Victoria, BC, Canada, 4 to 9 June 2006.
22. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:**5463–5467.
23. **Sorek, R., Y. Zhu, C. J. Creevey, M. P. Francino, P. Bork, and E. M. Rubin.** 2007. Genome-wide experimental determination of barriers to horizontal gene transfer. Science **318:**1449–1452.
24. **Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale.** 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics **4:**41.
25. **Turnbaugh, P. J., R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon.** 2006. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature **444:**1027–1031.
26. **Tyson, G. W., J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, and J. F. Banfield.** 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature **428:**37–43.
27. **Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers, and H. O. Smith.** 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science **304:**66–74.
28. **Wicker, T., E. Schlagenhauf, A. Graner, T. J. Close, B. Keller, and N. Stein.** 2006. 454 sequencing put to the test using the complex genome of barley. BMC Genomics **7:**275.
29. **Woese, C. R., and G. E. Fox.** 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. USA **74:**5088–5090.
30. **Woyke, T., H. Teeling, N. N. Ivanova, M. Huntemann, M. Richter, F. O. Gloeckner, D. Boffelli, I. J. Anderson, K. W. Barry, H. J. Shapiro, E. Szeto, N. C. Kyrpides, M. Mussmann, R. Amann, C. Bergin, C. Ruehland, E. M. Rubin, and N. Dubilier.** 2006. Symbiosis insights through metagenomic analysis of a microbial consortium. Nature **443:**950–955.
31. **Zehr, J. P., M. T. Mellon, and S. Zani.** 1998. New nitrogen-fixing microorganisms detected in oligotrophic oceans by amplification of nitrogenase (*nifH*) genes. Appl. Environ. Microbiol. **64:**3444–3450.