Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics

Title: Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics

Contributors: Frank J. Stewart*[1], Elizabeth A. Ottesen*[1], and Edward F. DeLong♦[1]

Affiliations and footnotes:

* These authors contributed equally to this work.

♦ Corresponding author**.** Mailing address: Division of Biological Engineering & Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Parsons Laboratory 48-427, 15 Vassar Street, Cambridge, MA 02139. Phone: (617) 253-5271. Email: delong@MIT.edu

[1]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Parsons Laboratory 48, 15 Vassar Street, Cambridge, MA 02139.

Running title: Ribosomal RNA extraction for microbial metatranscriptomics

NOTE: all revisions/additions to this manuscript have been highlighted in green

Abstract

Metatranscriptomes generated by pyrosequencing hold significant potential for describing functional processes in complex microbial communities. Meeting this potential requires protocols that maximize mRNA recovery by reducing the relative abundance of ribosomal RNA, as well as systematic comparisons to identify methodological artifacts and test for reproducibility across datasets. Here, we implement a protocol for subtractive hybridization of bacterial rRNA (16S and 23S) that uses sample-specific probes and is applicable across diverse environmental samples. To test this method, rRNA-subtracted and unsubtracted transcriptomes were sequenced (454 FLX technology) from bacterioplankton communities at two depths in the oligotrophic open ocean, yielding 10 datasets representing ~350 Mbp. Subtractive hybridization reduced bacterial rRNA transcript abundance by 40 to 58%, increasing recovery of non-rRNA sequences up to fourfold (from 12-20% of total sequences to 40-49%). In testing this method, we established criteria for detecting sequences replicated artificially via pyrosequencing errors and identified such replicates as a significant component (6 to 39%) of total pyrosequencing reads. Following replicate removal, statistical comparisons of reference genes (identified via BLASTX to NCBI-nr) between technical replicates and between rRNA-subtracted and unsubtracted samples showed low levels of differential transcript abundance (< 0.2% of reference genes). However, gene overlap between datasets was remarkably low, with no two datasets (including duplicate runs from the same pyrosequencing library template) sharing greater than 17% of unique reference genes. These results indicate that pyrosequencing captures a small subset of total mRNA diversity and underscores the importance of reliable rRNA subtraction procedures to enhance sequencing coverage across the functional transcript pool.

Keywords: functional genomics / gene expression / ribosomal RNA / Roche 454 pyrosequencing / RNA amplification / marine bacterioplankton

Introduction

Metatranscriptomic analysis using pyrosequencing is dramatically improving our understanding of gene expression in natural microbial communities (DeLong, 2009; Poretsky *et al*., 2009; Shi *et al.,* 2009). In these analyses, cDNA is synthesized from total RNA extracted from an environmental sample and used directly for massively parallel shotgun sequencing. Metatranscriptomes from functionally diverse habitats, including seawater and soil, can be sequenced using pyrosequencing methodologies (GS 20 or GS FLX systems, Roche 454 Life Sciences), yielding tens to hundreds of thousands of sequence fragments from the RNA pool (Frias-Lopez *et al.,* 2008; Poretsky *et al.,* 2009; Urich *et al.,* 2008). Such analyses provide detailed information on the taxonomic and functional diversity in the transcriptionally active community, as they simultaneously characterize both the ribosomal and messenger RNA components of the transcript pool. However, to advance the generic application and utility of pyrosequencing-based transcriptomics to microbial ecology studies, it is necessary to develop quality assurance and methodological troubleshooting techniques that both enhance current protocols and minimize bias in the interpretation of the read data.

It is not yet clear to what extent pyrosequencing methods capture the full breadth of expressed functional genes in microbial community transcriptomes. Indeed, pyrosequencing likely fails to capture many functionally important transcripts that occur at low frequencies. This is particularly true of datasets dominated by ribosomal RNA sequences (e.g., prokaryotic 5S, 16S and 23S rRNA, eukaryotic 18S and 28S). In the first study of a marine microbial

metatranscriptome sequenced using pyrosequencing (Roche GS 20 system), rRNA reads represented 53% of total sequences (Frias-Lopez *et al.,* 2008). A similar study using the same technology reported rRNA contamination at 37% of total sequences (Poretsky *et al.,* 2009), following the application of two commercial kits designed to enrich for mRNA (mRNA-ONLY™ Prokaryotic mRNA isolation kit, Epicentre; MICROBExpress™ Bacterial mRNA enrichment kit, Ambion). Subsequent analyses of marine transcriptomes using the Roche GS FLX platform, which yields average read lengths (~200-250 bp) over twice those produced via the GS 20 technology, have shown rRNA abundances that often exceed 90% of total reads (DeLong lab, unpublished data; Hewson *et al*., 2009b; Hewson, pers. comm.). The increase in rRNA abundance observed with FLX-based technology relative to GS 20 is likely due to differences in the pyrosequencing protocols themselves. For example, the FLX protocol includes a bead binding purification step that selects for longer transcripts, relative to GS20, which may increase the relative rRNA representation. As read lengths increase – the Roche GS FLX Titanium reagents now facilitate average read lengths greater than 400 bp – achieving adequate sequencing depth of non-rRNA reads, primarily mRNA, requires effective subtractive procedures to minimize rRNA abundance. Ribosomal RNA subtraction becomes particularly relevant as metatranscriptomics moves from a purely descriptive phase to one in which the method is applied experimentally to track low frequency changes in gene expression, e.g., in response to environmental perturbations (Delong, 2009).

Here, we introduce a sample-specific method for the subtraction of rRNA from total RNA. The method employs subtractive hybridization using antisense rRNA probes transcribed *in vitro* from PCR products amplified from coupled DNA samples, thereby ensuring the specificity of the probe mix. We tested this method by sequencing rRNA-subtracted and

unsubtracted transcriptomes of open ocean bacterioplankton communities at two depths in the North Pacific Subtropical Gyre (Station ALOHA, North Pacific), demonstrating a substantial reduction in the targeted RNA fraction (bacterial 16S and 23S rRNA). Our method can be generally extended to any microbial community sample, and can be easily expanded to target other rRNA transcripts, including those of archaea or eukaryotes.

The increasing use of pyrosequencing-based metagenomics and transcriptomics would also benefit from explicit quality control tests to determine the quantitative reproducibility of the data generated and to ensure reasonable extrapolations of read abundance to *in situ* transcript abundance. Gomez-Alvarez *et al.* (2009) showed that metagenomic datasets generated using pyrosequencing contain significant numbers (11-35% of total reads) of sequencing artifacts in the form of replicate sequences. These artificial replicates, putatively generated during the emulsion PCR stage of pyrosequencing (Briggs *et al.,* 2007; Gomez-Alvarez *et al.,* 2009), can erroneously increase the apparent abundance of transcripts from which these sequences derive. In order to address these and other concerns related to the reproducibility of expression profiles, we performed a set of technically replicated pyrosequencing-based transcriptomic analyses. These analyses determined criteria for replicate removal, underscored the potential for pyrosequencing artifacts (replicate reads) to confound interpretation of transcript diversity and abundance, and highlighted the potentially limited extent to which standard sequencing depths reveal the diversity of total transcript pools.

Materials and Methods

Sample Collection

Planktonic microorganisms were sampled from the photic zone (25 m and 75 m) at Station ALOHA (22° 45'N, 158° 00'W) as part of the Hawaii Ocean Time-series (HOT) program (HOT-186 cruise, casts S2C27 and S2C30) in October 2006, as described in Shi *et al.* (2009). Replicate seawater samples (1.8-2.0 L) for RNA extraction were prefiltered through 1.6 um GF/A filters (47 mm dia., Whatman) and collected onto 0.22 um Durapore filters (25 mm dia., Millipore) using a peristaltic pump. Filters were immediately transferred to microcentrifuge tubes containing 300 ul RNAlater® (Ambion) and frozen at -80C. Less than 20 min elapsed between sample collection (arrival on deck) and fixation in RNAlater®. Samples for DNA extraction were collected from the same water sample used for RNA collection as in Frias-Lopez *et al*. (2008). For each sample, seawater (220 L) was prefiltered through a 1.6 um GF/A filter (125 mm dia., Whatman) onto a 0.22 um Steripak-GP20 filter (Millipore). The filter units were filled with lysis buffer (50 mM Tris•HCl, 40 mM EDTA, and 0.75 M sucrose), capped, and frozen at -80°C until extraction.

RNA and DNA isolation

Total RNA was extracted from filters using a modification of the *mir*Vana™ miRNA Isolation kit (Ambion) as described previously (Shi *et al.*, 2009). Briefly, samples were thawed on ice, and the RNAlater® surrounding each filter was removed by pipetting and discarded. Filters were immersed in Lysis/Binding buffer (Ambion) and vortexed to lyse attached cells. Total RNA was then extracted from the lysate according to the manufacturer's protocol, incubated (37 ºC for 30 min) with TURBO DNA-free™ to remove genomic DNA, and purified and concentrated using the RNeasy MinElute Cleanup kit (Qiagen). Genomic DNA was extracted from Steripak filters as described previously (Frias-Lopex *et al*., 2008).

rRNA subtraction

Subtractive hybridization using sample-specific biotinylated rRNA probes was used to remove bacterial 16S and 23S rRNA from total RNA samples (Figure 1 and 2). The probe synthesis reaction was adapted from an *in situ* hybridization method by DeLong *et al.* (1999) and combined with a subtractive hybridization protocol similar to that of Su and Sordillo (1998). Ribonucleotide probes targeting bacterial 16S and 23S rRNA genes were generated from community DNA samples collected in tandem with each total RNA sample. Templates for probe generation were first prepared by PCR using universal primers flanking nearly the full length of the bacterial 16S gene and ~ 85% of the 23S rRNA gene, with reverse primers modified to contain the T7 RNA polymerase promoter sequence (Table 1; DeLong *et al.,* 1999). PCR reactions (50 ul each) included 100 ng template DNA, 1 ul of Herculase® II Fusion DNA Polymerase (Stratagene), 1X Herculase reaction buffer, 10 mM dNTP, and 10 uM each of forward and reverse primers. Reaction conditions were as follows: 2min at 92°C; 35 cycles of 20s at 95°C, 20s at 39°C (23S reactions) or 55°C (16S reactions), 75s (16S) or 90s (23S) at 72°C; 3min at 72°C. Resulting products were visualized via gel electrophoresis and purified via the QIAquick PCR purification kit (Qiagen). The above rRNA probe generation step can also potentially be modified by generating amplicons via reverse transcription (RT)-PCR, using RNA (rather than DNA) as starting template. We deliberately chose to amplify from the DNA pool, however, as RT-PCR is less efficient at longer sequence lengths, and we sought to maximize probe coverage along the length of all rRNA genes.

Biotinylated antisense rRNA probes were generated by *in vitro* transcription (IVT) with T7 RNA polymerase using T7 promoter-containing 16S and 23S amplicons as templates. IVT

was conducted using the *MEGAscript*® High Yield Transcription kit (Ambion), with the following modifications. Probes for 16S and 23S rRNA were generated separately in 20 ul reactions, each containing: 1X buffer, T7 RNA polymerase, SUPERase•In™ RNase inhibitor (10 U), ATP (7.5 mM), GTP (7.5 mM), CTP (5.625 mM), UTP (5.625 mM), biotin-11-CTP (1.875 mM, Roche), biotin-16-UTP (1.875 mM, Roche), and 16S/23S DNA template (250-500 ng). Reactions were run at 37°C for 4-5 h, then DNAse digested with TURBO™ DNAse (Ambion) for 15 min at 37°C. Products were purified using the MEGAclear™ kit (Ambion). Assuming the template amplicons used for IVT were ~50% GC, the biotin labeling density in the resulting 16S/23S probes was ~1 in 8 nucleotides.

Biotinylated rRNA probes were hybridized to complementary rRNA molecules in the total RNA sample. Hybridization reactions (50 ul), each containing formamide (20%), 1X SSC buffer (0.15 M sodium chloride, 0.015 M sodium citrate), SUPERase•In™ RNase inhibitor (20 U), template RNA (25 m ~600 ng; 75 m ~ 200 ng), and equal amounts of 16S and 23S rRNA probes at a final template-to-probe ratio of 1:2 (per probe), were denatured at 70°C for 5 min and incubated at room temperature (RT) for 3 min. Biotinylated double-stranded rRNA was then removed from the sample by hybridization (10 min at RT) to streptavidin-coated magnetic beads (New England Biolabs; 50 ul aliquot, washed 3X in 1X SSC), followed by separation on a magnetic rack (2 min) and removal of the rRNA-subtracted supernatant via pipet. An additional 50 uL 1X SSC was applied to the beads for washing, separated as above, and pooled with the original supernatant. The pooled products were purified via the RNeasy MinElute Cleanup kit (Qiagen). Subtraction efficiency was evaluated by monitoring the removal of 16S and 23S peaks from total RNA profiles using a 2100 Bioanalyzer and the RNA 6000 Pico chip kit (Agilent; Figure 2).

Minor updates to this protocol were implemented following characterization of the Station ALOHA (HOT 186) samples. The fully optimized and updated protocol is included as a pdf file in the Supplemental Online Materials. In brief, 1) an initial wash with 0.1N NaOH was incorporated into the streptavidin bead preparation steps to ensure complete removal of RNases, 2) the denaturation/hybridization step was changed from 5 min at 70°C followed by 3 min at RT to 5 min at 70°C followed by a step-down procedure with 1 min each at 5°C intervals from 65°C to 25°C, and 3) probes for Archaeal and Eukaryotic large and small subunit rRNA were incorporated into the protocol (see Supplementary text and Table S3 for primer design and sequences), and additional streptavidin-coupled beads were used to ensure complete removal of these additional probes.

RNA amplification and cDNA synthesis

rRNA-subtracted and unsubtracted total RNA (~35-80 ng) was amplified using the MessageAmp™ II-Bacteria kit (Ambion) as described previously (Frias-Lopez *et al.,* 2008; Shi *et al.,* 2009). Briefly, total RNA was polyadenylated using *Escherichia coli* poly(A) polymerase. Polyadenylated RNA was converted to double-stranded cDNA via reverse transcription primed with an oligo(dT) primer containing a promoter sequence for T7 RNA polymerase and a recognition site for the restriction enzyme BpmI (T7-BpmI-$(dT)_{16}$VN, Table 1). cDNA was then transcribed *in vitro* at 37°C (25m for 7hr, 75m for 14 hr), yielding large quantities (10-100 ug) of single-stranded antisense RNA. Amplified RNA (~5-10 ug aliquot) was then converted to double-stranded cDNA using the SuperScript® III First-Strand Synthesis System (Invitrogen) with priming via random hexamers for first-strand synthesis, and the SuperScript™ Double-Stranded cDNA synthesis kit (Invitrogen) for second-strand synthesis. cDNA was then purified

with the QIAquick PCR purification kit (Qiagen), digested with BpmI for 2-3 hrs at 37°C to remove poly(A) tails, and used directly for pyrosequencing

Pyrosequencing and technical replicates

Prior to sequencing, poly(A)-removed cDNA was purified via the AMPure® kit (Agencourt®). Purified cDNA was used for the generation of single-stranded DNA libraries and emulsion PCR according to established protocols (454 Life Sciences, Roche). Clonally amplified library fragments were then sequenced on a Genome Sequencer FLX System (Roche).

To produce a technically replicated transcriptomic analysis, equal aliquots of the HOT-186 75 m total RNA sample were separated and used independently for rRNA subtraction, RNA amplification, and pyrosequencing – see samples 75 m A and B, unsubtracted and rRNA-subtracted (Table 2). All sample processing parameters were kept consistent across replicates.

Data analysis

Ribosomal RNA-derived reads were identified using BLASTN to compare all reads against a rRNA database composed of both prokaryotic and eukaryotic small and large subunit rRNA nucleotide sequences (5S, 16S, 18S, 23S and 28S rRNA) from available microbial genomes and sequences in the ARB SILVA LSU and SSU databases (http://www.arb-silva.de). Reads producing alignments with bit scores greater than 50 were identified as rRNA sequences and removed from pyrosequencing datasets.

Non-rRNA sequences were checked for replicate sequences using the open-source program CD-HIT (Li and Godzik, 2006) according to the protocol of Gomez-Alvarez *et al.* (2009). Replicates were defined as sequences sharing greater than 99% nucleotide identity, with

an allowable length difference of 1 bp, and a requirement that the first 3 bp of the replicate sequences be identical. This cutoff was chosen by comparison of multiple pyrosequencing runs prepared from a single pyrosequencing library (sample 75 m A, rRNA-subtracted), as a conservative measure that significantly reduced the size of sequence clusters appearing at high frequency in one sequencing run but at low frequency in repeat sequencing runs (Table S1; Figure S2). Additional details on criteria for identifying replicates can be found in the Supplementary Online Material.

Non-rRNA sequence reads were compared to the National Center for Biotechnology Information non-redundant protein database (NCBI-nr, as of February 1, 2009) using BLASTX. Top BLASTX hits with e-values less than 1×10^{-5} were used for nr-protein designations. Reads with multiple, equal hits were assigned to the reference protein (hit) with the highest number of previously assigned reads. The total number of reads assigned to each reference protein was tracked through all analyses for consistent assignment between pyrosequencing runs. Statistically significant differences (two-tailed $P<0.05$) in the expression levels (abundances) of nr-designated transcripts were determined in pairwise comparisons between datasets using the method of Audic and Claverie (1997; AC test), which accounts for variation in database size and assumes a Poisson distribution for the number of transcripts representing a given gene. P-values were adjusted using a false discovery rate (FDR) correction to account for potential false positives due to multiple comparisons, as in Benjamini and Hochberg (1995).

Nucleotide sequence data generated during this study will be deposited in public databases prior to publication and can be made available to reviewers upon request.

Results and Discussion

Pyrosequencing read statistics

A total of 1,786,949 sequence reads representing ~350 Mbp over 10 pyrosequencing runs were generated from rRNA-subtracted and unsubtracted samples from depths of 25 and 75 m in the open ocean photic zone (Table 2). Reads with significant BLASTN hits (bit score > 50) to either prokaryotic or eukaryotic rRNA sequences represented 80 to 88% of total reads in unsubtracted samples. Of these, bacterial rRNA represented 74-83% of total reads (Table 2), with archaeal and eukaryotic rRNA representing 0.9-1.4% and 3.5-4.8%, respectively. The high percentages of rRNA reads observed here are consistent with experimental evidence suggesting that rRNA accounts for ~80-90% of total RNA in a typical bacterium (Wendisch *et al.*, 2001). Indeed, upon transitioning to the 454 GS FLX sequencing system, the percentage of rRNA observed in unsubtracted cDNA datasets derived from marine bacterioplankton has averaged 88% in our lab (range: 74-97%; n = 20 FLX cDNA datasets), underscoring the necessity for an effective rRNA subtraction approach when using our linear amplification protocol.

rRNA subtraction

Subtractive hybridization to sample-specific rRNA probes lowered bacterial rRNA abundance by 40-58% relative to unsubtracted samples, reducing bacterial rRNA to 35-46% and total rRNA to 52-61% of pyrosequencing reads (Table 2, Figure 2, Figure S3). In response, the fraction of non-rRNA reads in each dataset increased up to fourfold, raising the proportion of reads with significant BLASTX hits to NCBI-nr proteins from 3.1-4.9% in unsubtracted samples to 7.3-20.4% in subtracted samples (Table 2). Bacterial 16S rRNA showed a greater proportional decrease in abundance than 23S rRNA following rRNA subtraction (Table 2), which may be due to a combination of broader coverage across diverse bacterial phyla by the

16S primers relative to the 23S primers, and to differences in the extent to which 16S and 23S primer sets span the full length of the target molecules (~95% of 16S rRNA, ~85% of 23S rRNA). Since our probe sets targeted only bacterial rRNA, the relative percentages of archaeal and eukaryotic rRNA reads increased approximately three to fivefold (to 3.5-4.8% and 10.7-12.2% of total reads, respectively; Table 2). However, among those reads not identified as bacterial rRNA, the percentage represented by archaeal and eukaryotic rRNA reads did not differ substantially between unsubtracted (25-29%; mean: 23.8%) and subtracted (20-29%; mean: 26.6%) samples, suggesting little non-specific probe binding but emphasizing the need to develop additional probe sets to target these rRNA fractions.

The subtraction of rRNA occurred non-uniformly along the length of rRNA transcripts. For example, among the rRNA reads remaining in the subtracted sample, the proportion mapping to the central region of the 23S rRNA (~bp 1000-2500) decreased relative to the unsubtracted sample, while the proportion mapping to the terminal region (~ bp 2500-2900) increased substantially (by ~50%; Figure S1). This pattern may be caused in part by the exclusion of the terminal 400 bp of the ~2900 bp 23S rRNA gene by our probe set (Table 1), although a clear, but as of yet unexplained, bias toward 3' 23S rRNA reads was also observed in the unsubtracted sample. A shift in the relative abundances of reads representing varying regions of the 16S rRNA was also apparent following subtraction, with a noticeable proportional increase in reads mapping to the 5' 16S rRNA region (Figure S1). These patterns indicate differential subtraction efficiencies along the rRNA transcripts, and therefore the potential that these molecules had been fragmented in the pre-subtracted samples.

After developing and testing the rRNA subtraction protocol described above, we have applied our method to a diverse range of samples and further optimized it to include primer sets

315 targeting both Archaeal and Eukaryotic rRNAs (see Table S3). Using this revised protocol, rRNA abundance in microbial community cDNAs has averaged 36% of total reads (range: 28-55%) across 10 different bacterioplankton samples from three distinct oceanic regions (Table 3; see Figure S3 for a representative total RNA profile following Domain-specific rRNA subtraction). We have also applied the method to a pure monoculture of actively growing 320 *Dokdonia* sp. using 16S and 23S rRNA probe sets that perfectly match this target species. For this pure culture, our method successfully reduced rRNA to an average of 8% of the total cDNA reads (range: 3-11%; n = 3; see Table S3 for primers).

In addition to the method presented here, several commercial rRNA reduction protocols are also now available. One commonly used commercial product, the MICROBExpress™ 325 Bacterial mRNA enrichment kit (Ambion), employs a subtractive hybridization to proprietary oligonucleotide probes, followed by rRNA removal via bead-immobilized capture oligonucleotides (in contrast to sample-specific, near full-length probes and biotin-streptavidin capture in our protocol). While the MICROBExpress™ kit has been shown to be compatible for rRNA subtraction from a variety of bacteria, the efficiency of rRNA removal using this method 330 can vary widely for community RNA samples (e.g., Poretsky *et al.,* 2005; McGrath *et al.,* 2008; Hewson *et al.,* 2009a,b), as well as for single-species analyses (e.g., Yoder-Himes et *al*., 2009). Indeed, oligonucleotide capture probes used in this method are predicted to be sensitive to target sequence variability known to be present in microbial community rRNAs, and the manufacturers explicitly state that the commercial kit is only partially compatible, or even incompatible, with a 335 variety of microorganisms, including all *Archaea* (http://www.ambion.com/techlib/misc/microbe.html). Another commercial rRNA removal strategy, the mRNA-ONLY™ kit (Epicentre), uses specific exonucleases to selectively digest

rRNA, but not mRNA. A potential complication of the exonuclease method in the mRNA-ONLY™ protocol is its potential to catalyze secondary, non-5' monophosphate RNAse activity that may degrade specific mRNA transcripts in addition to the rRNA (Epicentre website). In a recent analysis of a microbial community associated with a *Trichodesmium* bloom, the mRNA-ONLY™ kit , used in conjunction with the MICROBExpress™ kit, was unable to reduce rRNA abundance below 94% of total FLX-based reads (Hewson *et al.*, 2009b; Hewson, pers. comm.), again suggesting potentially high variability in commercially available protocols. Recognizing the potential limitations of these protocols is important in selecting or developing rRNA subtraction methods for microbial transcriptome analyses.

Other alternatives to commercially available subtraction techniques have also been explored. Noting mixed results with the MICROBExpress™ method, McGrath *et al.* (2008) proposed physical removal of rRNA bands from RNA samples by agarose gel electrophoresis and extraction. While effective, this technique requires large starting concentrations of RNA and may bias functional gene expression profiles by eliminating mRNA transcripts that co-migrate with the rRNA fraction. Gilbert *et al.* (2008) reported low rRNA representation in a metatranscriptomic survey following amplification of the cDNA via multiple displacement amplification (MDA) using the GenomiPHI™ V2 kit (GE Healthcare). However, this kit is optimized for amplification of genomic DNA, and has not been extensively tested for reproducibility and bias when used for amplification of short cDNA fragments. Indeed, numerous studies have shown that even for single template genomic DNA, MDA using ϕ29 polymerase amplifies unevenly across different loci, suggesting that the reliability and utility of MDA for quantitative analyses is uncertain at best (Bergen *et al.*, 2005; Ballantyne *et al.*, 2007). Recently, Armour *et al.* (2009) proposed a novel rRNA depletion method based on the use of a

computationally defined subset of hexameric primers that exclude target rRNA sequences during cDNA synthesis from total RNA. While potentially effective in low-complexity samples or single-species monocultures, this method, as for most published rRNA subtraction procedures, is not easily extrapolated to complex metatranscriptomic samples with diverse rRNA pools. A definitive comparison of rRNA removal protocols from previously published metatranscriptomic studies is not feasible, since prior studies did not include unsubtracted controls for comparison, utilized distinct cDNA synthesis and downstream pyrosequencing preparation protocols, and analyzed microbial communities with differing compositions.

The protocol we describe here offers some potential advantages over the existing rRNA-subtraction procedures noted above. The method can be easily tailored to synthesize sample- or taxon-specific probes targeting either specific strains or a broad array of Archaeal, Bacterial, and Eukaryotic rRNAs. While we chose to target only Bacterial rRNA for the initial development of this method, the protocol can be expanded to include both Archaeal and Eukaryotic rRNA probe sets generated using the broad-specificity primers listed in Table S3. The method can also be applied to less complex, non-environmental samples (e.g., experimental cultures or consortia). If necessary for such samples, taxon-specific primers can be used in place of universal primers for probe generation, as demonstrated for a *Dokdonia* culture analyzed in our lab (Table 3 and S3). When followed by a linear RNA amplification step, this subtractive protocol can be used with relatively small amounts of starting material; in our hands, subtraction of rRNA from as little as 20 ng total RNA has yielded amounts of mRNA-enriched template sufficient for amplification and pyrosequencing. As the protocol implemented here biases the composition of any rRNA reads remaining after subtraction (as does any rRNA subtraction procedure), our method is specifically designed for maximizing coverage of the functional RNA pool, potentially

identifying unique or interesting transcripts that can inform or suggest more targeted gene-specific studies to follow.

Replicate reads

Sequencing artifacts in the form of replicated sequences are a common source of error in pyrosequencing datasets (Briggs *et al.,* 2007; Gomez-Alvarez *et al.,* 2009). Careful identification and removal of such sequences, which can account for more than 30% of pyrosequencing reads (Gomez-Alvarez *et al.,* 2009), is therefore essential for accurately extrapolating observed read abundance to *in situ* DNA or cDNA abundance. Replicates are hypothesized to originate during the emulsion PCR step of pyrosequencing. Incomplete emulsion can result in the attachment of a single PCR product to multiple beads (Briggs *et al.*,, 2007; Gomez-Alvarez *et al.*,, 2009), which upon sequencing results in an increase in the number of observed occurrences of that sequence in the dataset. In an idealized case, sequences generated from identical, replicated template molecules would share start and stop sites, resulting in identical sequence along their full length. However, sequencing errors and quality trimming result in the generation of non-identical sequences of different lengths from identical template molecules. The challenge therefore is to differentiate imperfectly sequenced artificial replicates from legitimate sequences derived from multiple, similar DNA fragments.

Gomez-Alvarez *et al.,* (2009) calculated that, for the genomic DNA of an idealized microbial community, the probability of multiple reads starting at the same position in a randomly-sheared metagenome analysis is extremely low (~1 x 10^{-10}). These authors therefore identify artificial replicates as sequences (of potentially varying lengths) sharing greater than 90% nucleotide identity and having identical beginning sequences (over the first 3 bp). This

definition is likely robust for microbial community DNA, but is perhaps overly conservative for microbial community transcriptomes. Unlike random genomic DNA fragments, transcripts have clearly defined start and stop sites and routinely occur in multiple copies per cell. As a result, criteria for identifying replicates from metatranscriptomes must, to the greatest extent possible, distinguish artifacts from legitimate re-sampling of multiple transcripts originating from the same gene.

We analyzed pairs of sequencing runs (independent emulsion PCR and sequencing reactions generated from the same adaptor-ligated template library) to establish criteria for replicate removal from our transcriptomic libraries, identifying replicates as sequences differing by no more than 1 bp in length, sharing 99% nucleotide identity, and having identical start sites (first 3 bp) (see Supplementary Online Text, Table S1, Figure S2). Using these criteria, re-sampled transcripts were rare (0.3-3.4% of sequences) compared to artificial replicate transcripts (6-39% of sequences). Though clusters of replicate reads were identified in all samples examined, the percentage of replicated reads to total reads (replicate frequency) varied greatly among the 10 different pyrosequencing runs (Table 2). Substantial variation occurred even between multiple runs generated via independent emulsion PCR and sequencing reactions from the same template library. Notably, replicate frequency varied between runs from 9 to 38% and 14 to 36% in the rRNA-subtracted and unsubtracted 75 m A samples, respectively, and from 11 to 39% between runs of the unsubtracted 75 m B sample (Table 2). In each of these comparisons, replicate frequency was negatively correlated with the number of total reads per run, suggesting a link between artifact generation and the efficiency of the emulsion PCR and pyrosequencing steps. This immediately suggested important criteria for quality control and assessment of individual sequencing runs.

The presence of these replicates can bias the apparent diversity of the transcript pool. For example, prior to removal of replicate sequences the slopes of rarefaction curves describing the number of unique non-redundant (nr) peptide reference genes identified per dataset as a function of sequencing depth differed notably between repeated runs (1 and 2) of the subtracted 75 m A samples (Table 2; Figure 3). Removal of replicate sequences eliminated this discrepancy, yielding highly similar rarefaction curves. Together, these data indicate that artificial replicates are a common and potentially biasing component of pyrosequencing-based transcriptomic datasets. However, effective criteria for replicate removal may vary among samples of differing taxonomic complexity and functional state. Though not yet routine in pyrosequencing-based analyses, systematic comparisons of technically repeated pyrosequencing runs may help delineate sample-specific criteria for replicate removal.

Statistical comparison of transcript abundances in technical replicates

Validation of rRNA subtraction and computational replicate removal methods required statistical determination of changes in transcript abundance with a high degree of taxonomic resolution. Statistical comparisons of metagenomic and transcriptomic profiles have largely focused on gene clusters and functional groupings rather than individual genes (Rodriguez-Brito *et al.,* 2006; Frias-Lopez *et al*., 2008; Huson *et al.,* 2009; Poretsky *et al.,* 2009). However, this approach potentially lacks the resolution to detect changes in expression of specific transcripts (e.g., due to the non-specific binding of rRNA probe to an mRNA transcript during subtractive hybridization). Here, to assess the impact of rRNA subtraction and replicate removal on expression profiles, non-rRNA transcripts were mapped to specific protein sequences in the NCBI-nr database, and the relative abundances of transcripts matching single reference genes

were compared between datasets (Table 4, Figures 4 and 5). Dalevi *et al.* (2008) showed that a similar mapping approach accurately represented the functional and taxonomic characteristics of ~100 bp DNA fragments; we expect an even greater degree of assignment accuracy given our longer average read length (> 200 bp). A statistical test from the expressed sequence tag literature was then applied to identify differentially represented reference genes (AC test; Audic and Claverie 1997), along with a false-discovery rate (FDR) minimizing test (Benjamini and Hochberg, 1995) to correct for the large number of comparisons ($P<0.05$). We used this test to examine the reproducibility of transcriptional profiles generated by pyrosequencing, and the effect of rRNA subtraction on the abundance on non-rRNA transcripts.

Removal of replicate sequences was key to the generation of reproducible transcriptional profiles (see starred datasets in Table 4). For example, the variation between rarefaction curves generated from the raw sequencing data from the rRNA-subtracted 75 m A sample (Figure 3) was also reflected in reference gene abundances; 144 genes were identified as 'differentially expressed' between the two sequencing runs (Table 4). Removal of replicate sequences reduced this to just 30 references with significantly different abundances. The majority of these differences involved genes represented by very few (or zero) reads in one of the two datasets. Replicate removal may therefore have the most significant effect on apparent differences between low abundance transcripts, for which even small numbers of artificial replicates may have a disproportionately large effect on apparent expression level.

A thorough analysis of the reproducibility of metatranscriptomic profiles must also take into account experimental variation, as the generation of transcriptomic libraries from extracted total RNA requires extensive processing steps, including linear amplification in our procedure, that might introduce variability into the observed transcriptional profile. In order to address this

issue, we subdivided the total RNA sample from 75 m to generate four samples (unsubtracted and rRNA-subtracted samples A and B) which were processed in parallel to examine the reproducibility of resultant transcriptomic datasets. For the unsubtracted 75 m sample, no nr reference genes were differentially expressed between replicates A and B (Table 4). For the rRNA-subtracted 75 m sample, following removal of replicate reads, only 0.03% (7 of 25,174) of total identified references were differentially expressed (Table 4; Table S1). Of those genes represented by greater than 0.1% of the sequence reads in the subtracted 75 m A dataset (n = 42), only three varied in expression between replicates (Table 5, Figure 4). As observed in comparisons of replicate runs from a single template library, the majority of the significant differences between the full technical replicates involved relatively low abundance reference genes having significantly higher representation in one of the two datasets (Table 5).

Statistical comparison of rRNA-subtracted and unsubtracted samples

Having established the reproducibility of transcriptional profiles generated during metatranscriptomic analyses, we examined bias potentially introduced during the removal of rRNA by subtractive hybridization. Subtractive protocols have the potential to alter functional gene expression profiles if removal of non-rRNA transcripts occurs due to non-specific probe binding. In the 25 m sample, only 0.03% of total nr reference genes (13 of 48,090 total, which includes 2378 shared between datasets + 45,712 unique to either dataset) showed significantly different abundances between subtracted and unsubtracted samples (Table 4). Of these, five were represented at abundances greater than 0.1% of total reads in the unsubtracted dataset (Figure 5). In the 75 m sample, only 0.01% (3 of 32,340) were differentially represented. These low levels of variation are within the range observed between replicate pyrosequencing runs

derived from independent emulsion PCR reactions using template from the same adapter-ligated library (0-0.2% of genes differentially expressed for comparisons 25 m S1 vs. S2, 75 m AS1 vs. AS2, AU1 vs. AU2, BU1 vs. BU2, Table 4). This suggests that these differences are due to stochastic variation introduced during pyrosequencing, and that the subtractive hybridization protocol does not significantly alter the apparent expression profile generated from transcript libraries.

Conclusions

A distinguishing characteristic of community DNA and RNA sequencing efforts is the high complexity of the resulting sequence data. For example, in this study, pyrosequencing of a single sample of seawater collected at 25 m generated 266,859 unique non-rRNA sequences (including runs from both rRNA-subtracted and unsubtracted libraries), 117,809 (44%) of which had significant hits to 48,090 unique nr reference proteins. An initial, half-plate run derived from the RNA-subtracted portion of the 25 m sample yielded 21,011 nr references, of which only 37% were recaptured in a second full-plate run, which contained nearly twice as many sequences. Sequencing depth can therefore clearly limit the analytical capability of metatranscriptomic analyses. Deeper sequencing not only increases the likelihood of sampling novel transcripts but also facilitates statistically significant comparisons of transcripts appearing across multiple datasets (e.g., experimental treatments). This is particularly relevant for low-abundance transcripts that constitute the majority of diversity in the mRNA pool.

Though changes in the transcriptional profiles of highly expressed genes yield important insight into microbial communities (e.g., Frias-Lopez *et al.*, 2008; Hewson *et al.*, 2009b), the relatively shallow extent to which pyrosequencing captures gene-level diversity among

functional transcripts increases the need to optimize message recovery. Ribosomal RNA represented 80 - 88% of sequences recovered from the ocean samples analyzed in this study. Given the mRNA transcript diversity encountered here, the presence of rRNA at such high proportions hinders the detection of potentially tens to hundreds of thousands of unique functional transcripts via a standard 454 pyrosequencing run. Even incremental reductions in the rRNA pool can therefore substantially increase our knowledge of genes expressed at low frequency. When analyzing expression at such resolution, it becomes increasingly important to distinguish sequencing artifacts from genuine variation in transcript abundance. This analysis, along with other recent studies (e.g., Gomez-Alvarez *et al.*, 2009), highlights the need for meaningful criteria for identifying and removing artificially replicated sequences that confound statistical comparisons of expression. Comparisons between technically replicated libraries constitute an effective method for establishing such criteria and should become more commonplace in pyrosequencing-based analyses. Upon removal of sequencing artifacts, transcriptomic pyrosequencing datasets appear highly reproducible and, in conjunction with rRNA-subtraction methods that maximize message recovery, can provide new insights into the diversity and dynamics of less abundant transcripts. This is particularly relevant as microbial metatranscriptomics is increasingly used to monitor community responses to experimentally-induced perturbations, some of which may elicit subtle, but important, functional changes in non-dominant community members.

Acknowledgements

We thank Yanmei Shi and the captain and crew of the *R/V Kilo Moana* for their help in collecting samples for this study, Jay McCarren and Asuncion Martinez for help with DNA

545 extractions, Rachel Barry for her tireless work in preparing samples for pyrosequencing, Tracy Teal for providing scripts for use in identifying replicate pyrosequencing reads, Hiro Kimura for generously providing data on rRNA subtraction efficiency in *Dokdonia*, and Adrian Sharma for providing data on rRNA subtraction efficiency in marine bacterioplankton samples from Bermuda. This work was supported by a gift from the Agouron Institute, and grants from the
550 Gordon and Betty Moore Foundation (EFD), the Office of Science (BER) U. S. Department of Energy, and NSF Science and Technology Center Award EF0424599. This work is a contribution of the Center for Microbial Oceanography: Research and Education (C-MORE)

References

Armour CD, Castle JC, Chen R, Babak T, Loerch P, Jackson S, Shah JK, Dey J, Rohl CA, Johnson JM, Raymond CK. (2009). Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nat Methods* **6**: 647-649.

Audic S, Claverie JM. (1997). The significance of digital gene expression profiles. *Genome Res* **7**: 986-995.

Benjamini Y, Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc B* **57**: 289-300.

Ballantyne KN, van Oorschot RAH, Muharam I, van Daal A, Mitchell RJ. (2007). Decreasing amplification bias associated with multiple displacement amplification and short tandem repeat genotyping. *Anal Biochem* **368**: 222-229.

Bergen AW, Qi Y, Haque KA, Welch RA, Chanock SJ. (2005). Effects of DNA mass on multiple displacement whole genome amplification and genotyping performance. *BMC Biotechnol* **5**: 24.

Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prüfer K, Meyer M, Krause J, Ronan MT, Lachmann M, Pääbo S. (2009). Patterns of damage in genomic DNA sequences from a Neandertal. *Proc Natl Acad Sci USA* **104**: 14616-14621.

Dalevi D, Ivanova NN, Mavromatis K, Hooper SD, Szeto E, Hugenholtz P, Kyrpides NC, and Markowitz VM. (2008). Annotation of metagenome short reads using proxygenes. *Bioinformatics* **24**: i7-i13.

DeLong EF. (2009). The microbial ocean from genomes to biomes. *Nature* **459**: 200-206.

DeLong EF, Taylor LT, Marsh TL, Preston CM. (1999). Visualization and enumeration of marine planktonic archaea and bacteria by using polyribonucleotide probes and fluorescent in situ hybridization. *Appl Environ Microbiol* **65**: 5554-5563.

Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW, DeLong EF. (2008). Microbial community gene expression in ocean surface waters, *Proc Natl Acad Sci USA* **105**: 3805-3810.

Gilbert JA, Field D, Huang Y, Edwards R, Li W, Gilna P, Joint I. (2008). Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS One* **3**:e3042.

Gomez-Alvarez V, Teal TK, Schmidt, TM. (2009). Systematic artifacts in metagenomes from complex microbial communities. *ISME J* **1-4**.

Hewson I, Poretsky RS, Beinart RA, White AE, Shi T, Bench SR, Moisander PH, Paerl RW, Trip HJ, Montoya JP, Moran MA, Zehr JP. (2009a). *In situ* transcriptomic analysis of the globally important keystone N_2-fixing taxon *Crocosphaera watsonii*. *ISME J* **3**:618-631.

Hewson I, Poretsky RS, Dyhrman ST, Zielinski B, White AE, Tripp HJ, Montoya JP, Zehr JP. (2009b). Microbial community gene expression within colonies of the diazotroph, *Trichodesmium*, from the Southwest Pacific Ocean. *ISME J* **1-15**

Hunt DE, Klepac-Ceraj V, Acinas SG, Gautier C, Bertilsson S, Polz MF. (2006). Evaluation of 23S rRNA PCR primers for use in phylogenetic studies of bacterial diversity. Appl. *Environ Microbiol* **72**: 2221-2225.

Huson DH, Richter DC, Mitra S, Auch AF, Schuster SC. (2009). Methods for comparative metagenomics. *BMC Bioinformatics* **10**(Suppl 1): S12.

Li W, Godzik A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**:1658-1659.

McGrath KC, Thomas-Hall SR, Cheng CT, Leo L, Alexa A, Schmidt S, Schenk PM. (2008). Isolation and analysis of mRNA from environmental microbial communities. *J Microbiol Meth* **75**: 172-176.

Poretsky RS, Bano N, Buchan A, LeCleir G, Kleikemper J, Pickering M, Pate WM, Moran MA, Hollibaugh JT. (2005). Analysis of microbial gene transcripts in environmental samples. *Appl Environ Microbiol* **71**: 4121-4126.

Poretsky RS, Hewson I, Sun S, Allen AE, Zehr JP, Moran MA. (2009). Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. *Environ Microbiol* **11**: 1359-1375.

Rodriguez-Brito B, Rohwer F, Edwards RA. (2006). An application of statistics to comparative metagenomics. *BMC Bioinformatics* **7**: 162.

Shi Y, Tyson GW, DeLong EF. (2009). Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature* **459**: 266–269.

Su C, Sordillo LM. (1998). A simple method to enrich mRNA from total prokaryotic RNA. *Mol Biotechnol* **10**: 83-85.

Urich T, Lamzen A, Qi J, Huson DH, Schleper C, Schuster SC. (2008). Simultaneous assessment of soil microbial community structure and function through analysis of the metatranscriptome. *PLoS One* **3**:e2527.

Wendisch VF, Zimmer DP, Khodursky A, Peter B, Cozzarelli, Kustu S. (2001). Isolation of *Escherichia coli* mRNA and comparison using mRNA and total RNA on DNA microarrays. *Anal Biochem* **290**: 205-213.

Yoder-Himes DR, Chain PSG, Zhu Y, Wurtzel O, Rubin EM, Tiedje JM, Sorek R. (2009). Mapping the *Burkholderia cenocepacia* niche response via high-throughput sequencing. *Proc Natl Acad Sci USA* **106**: 3976-3981.

Figure Legends

Figure 1. Flowchart of the metatranscriptomic sample processing pipeline, illustrating steps for the sample-specific subtraction of bacterial ribosomal RNA (16S and 23S) from total RNA in an environmental sample.

Figure 2. Size distribution of total RNA in unsubtracted and rRNA-subtracted portions of the HOT-186 25 m sample.

Figure 3. Rarefaction curve for HOT-186 75 m sample AS. The number of unique nr reference genes identified via BLASTX (at e-values $\leq 1 \times 10^{-5}$) is shown as a function of sequencing depth. Runs 1 and 2 represent multiple pyrosequencing runs from the same adapter-ligated sample library, before and after removal of replicate sequences.

Figure 4. Relative abundance of NCBI-nr reference genes in rRNA-subtracted pyrosequencing 75 m A datasets. Reference genes representing > 0.1% of the 75 m A library are shown in descending order. Their abundance in a replicate library (75 m B rRNA-subtracted) is shown in red. Reference genes with significantly different abundances are labeled with a FDR-corrected p-value.

Figure 5. Relative abundance of NCBI-nr reference genes in HOT-186 25 m pyrosequencing datasets with and without rRNA subtraction. Reference genes representing > 0.1% of the unsubtracted library are shown in descending order. Their abundance in the library generated

from post-subtraction RNA is shown in red. Reference genes with significantly different abundances are labeled with a FDR-corrected p-value.

Figure 1.

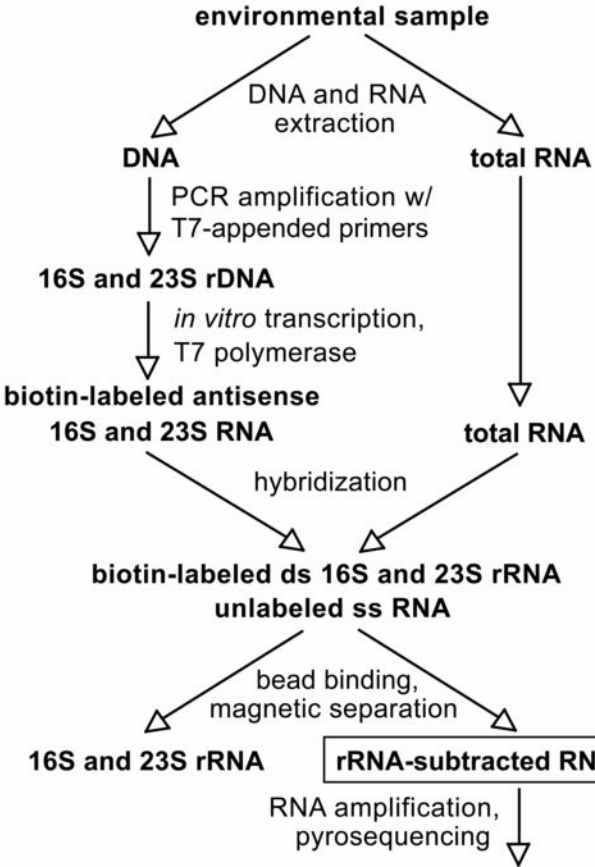
environmental sample
DNA and RNA extraction
DNA
total RNA
PCR amplification w/ T7-appended primers
16S and 23S rDNA
in vitro transcription, T7 polymerase
biotin-labeled antisense 16S and 23S RNA
total RNA
hybridization
biotin-labeled ds 16S and 23S rRNA
unlabeled ss RNA
bead binding, magnetic separation
16S and 23S rRNA
rRNA-subtracted RNA
RNA amplification, pyrosequencing

Figure 2.

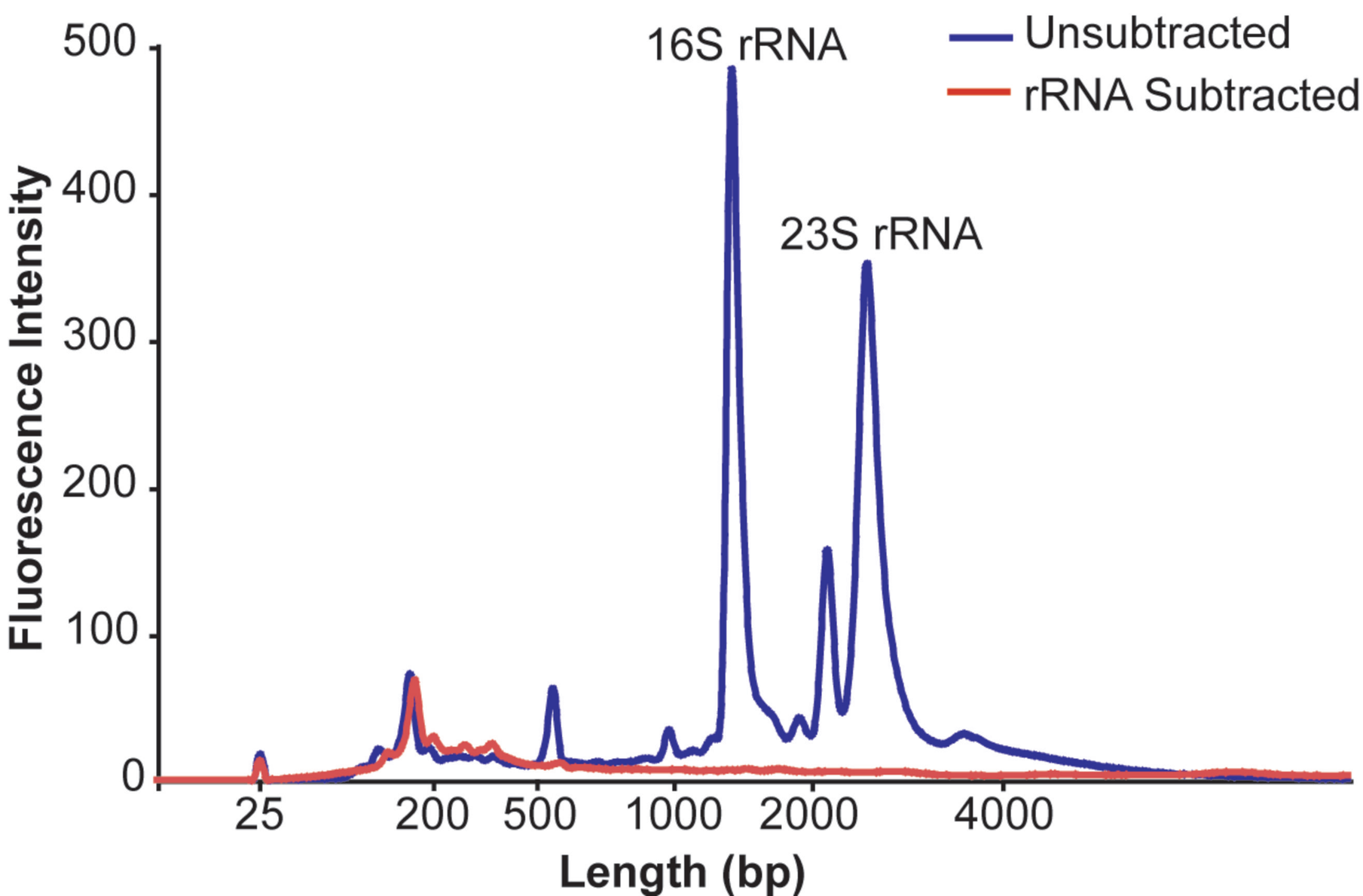
500
400
300
200
100
0
Fluorescence Intensity
16S rRNA
23S rRNA
Unsubtracted
rRNA Subtracted
25
200
500
1000
2000
4000
Length (bp)

Figure 3.

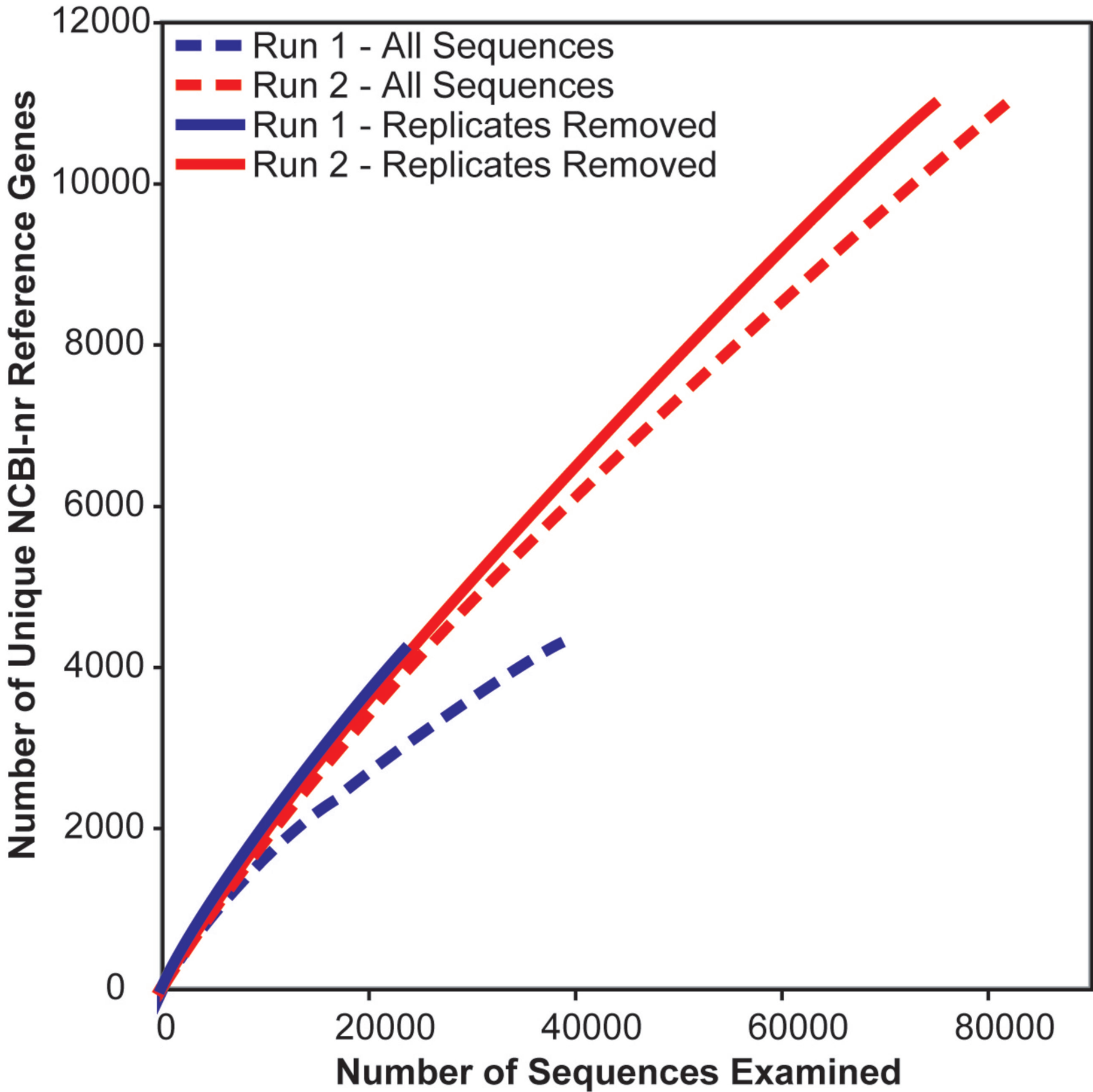
Run 1 - All Sequences
Run 2 - All Sequences
Run 1 - Replicates Removed
Run 2 - Replicates Removed
Number of Unique NCBI-nr Reference Genes
12000
10000
8000
6000
4000
2000
0
0
20000
40000
60000
80000
Number of Sequences Examined

Figure 4.

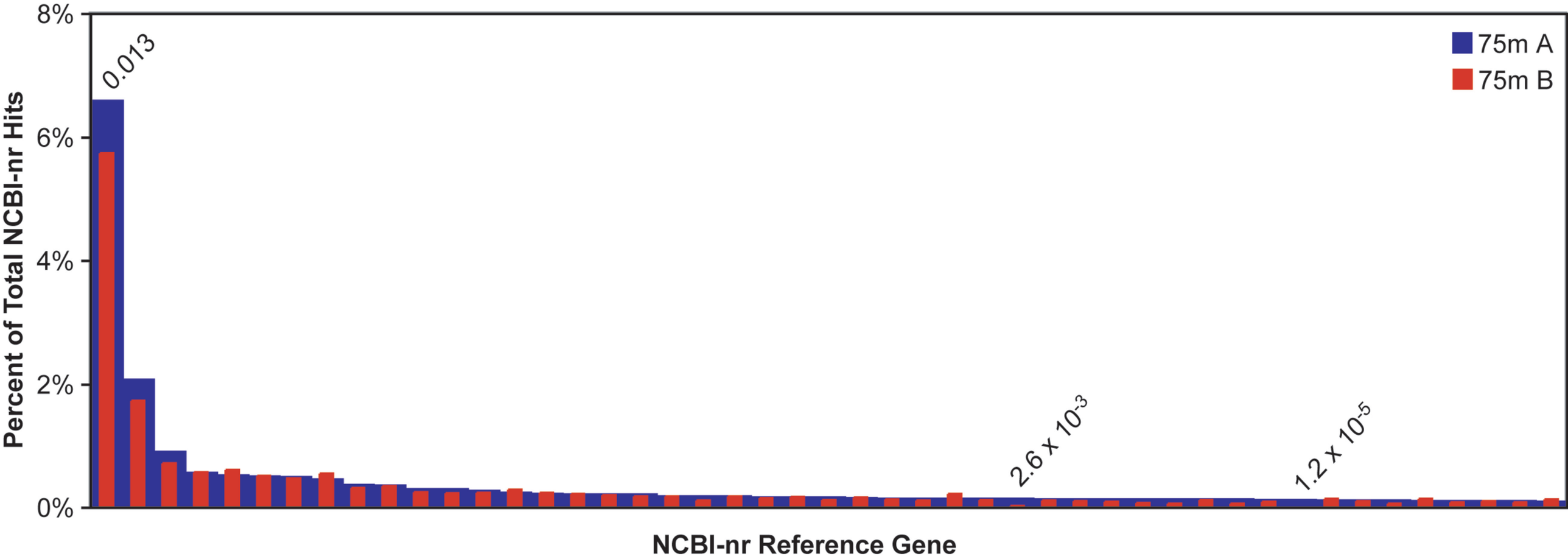
8%
6%
4%
2%
0%
Percent of Total NCBI-nr Hits
0.013
2.6 x 10-3
1.2 x 10-5
75m A
75m B
NCBI-nr Reference Gene

Figure 5.

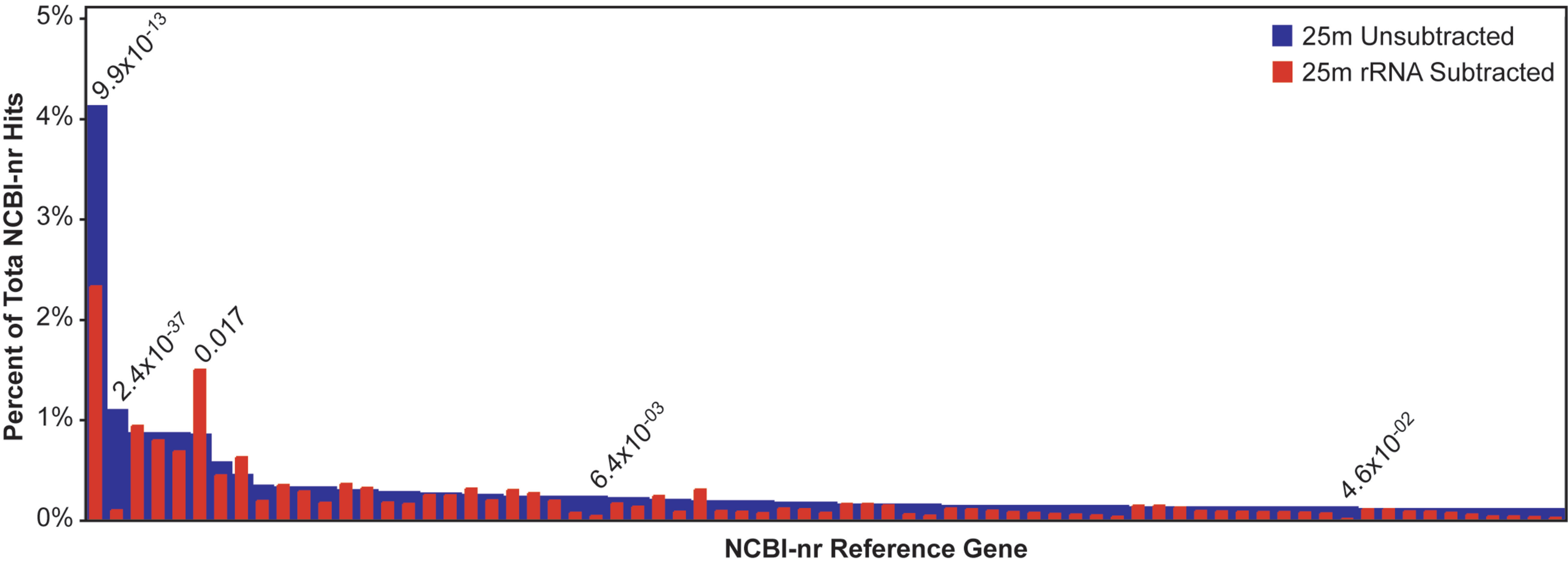
5%
4%
3%
2%
1%
0%
Percent of Tota NCBI-nr Hits
9.9x10-13
2.4x10-37
0.017
6.4x10-03
4.6x10-02
25m Unsubtracted
25m rRNA Subtracted
NCBI-nr Reference Gene

Table 1. Primers used for PCR to generate bacterial 16S and 23S rRNA probes and for reverse transcription of polyadenylated cDNA

locus	primer	sequence (5' to 3')
16S	*27F*	AGAGTTTGATCCTGGCTCAG
	1492R_T7[1]	GCCAGTGAATTG**TAATACGACTCACTATA**GGGACGGCTACCTTGTTACGACTT
23S[2]	*189F*	GAASTGAAACATCTHAGTA
	2490R_T7[1]	GCCAGTGAATTG**TAATACGACTCACTATA**GGGCGACATCGAGGTGCCAAA
poly(A)[3]	T7-BpmI-$(dT)_{16}$VN	GCCAGTGAATTG**TAATACGACTCACTATA**GGGGCGACTGGAGTTTTTTTTTTTT TTTTVN

[1] see DeLong *et al.* (1999) for design of primers appended with T7 promoters (underlined above)
[2] 23S primers are based on those of Hunt *et al.* (2006)
[3] targets molecules containing poly(A) residues; used for reverse transcription prior to RNA amplification

Table 2. Read numbers and statistics

				% rRNA reads[4]							% non-rRNA reads[5]		
					Bacteria		Archaea		Eukaryota				
sample	S/U[1]	run[2]	reads[3]	all	LSU	SSU	LSU	SSU	LSU	SSU	all	unique[6]	nr hits[7]
25m	U	1	138 269	88.1	64.9	18.3	0.8	0.1	3.2	0.8	11.9	10.3	4.7
	S	1	195 031	51.5	29.3	5.4	4.4	0.1	11.4	0.8	48.5	45.5	20.4
		2	366 790	51.7	29.5	5.3	4.6	0.2	11.4	0.8	48.3	44.6	19.5
75m A	U	1	63 091	80.3	61.5	13.9	1.3	0.1	3.0	0.5	19.7	12.5	3.1
		2	199 807	82.7	64.3	13.6	1.1	0.1	3.0	0.6	17.3	14.8	3.3
75m A	S	1	99 275	61.2	43.8	1.4	3.7	0.3	10.0	1.9	38.8	23.9	7.3
		2	206 823	60.2	44.7	1.2	3.2	0.3	8.7	2.0	39.8	36.3	9.3
75m B	U	1	40 732	80.5	60.9	13.4	1.3	0.1	4.2	0.6	19.5	11.8	3.6
		2	225 507	81.9	63.7	12.4	1.2	0.1	3.7	0.8	18.1	16.1	4.9
75m B	S	1	251 624	54.5	37.0	1.2	3.4	0.3	10.0	2.4	45.5	40.0	11.0

[1] S = bacterial rRNA subtracted via hybridization, U = rRNA unsubtracted from sample,
[2] independent emulsion PCR reactions and pyrosequencing runs generated from the same library
[3] total number of sequence reads per run
[4] percentage of total pyrosequencing reads with significant (bit score > 50) BLASTN hits to prokaryotic (Bacteria, Archaea) or eukaryotic small (SSU: 16S, 18S) and large (LSU: 5S, 23S, 28S) subunit rRNA
[5] non-rRNA reads, as a percentage of total pyrosequencing reads
[6] non-rRNA reads without replicates; see *Methods* for replicate criteria
[7] non-replicate, non-rRNA reads with significant (e-value ≤ $1x10^{-5}$) BLASTX hits to proteins in the NCBI non-redundant database (nr)

Table 3. rRNA abundance in metatranscriptomic profiles of diverse samples following subtractive hybridization

sample	primers[1]	total reads	rRNA reads	% rRNA
Pure culture				
Dokdonia sp., rep 1	Dok	630 260	65 339	10.4
Dokdonia sp., rep 2	Dok	195 278	4 859	2.5
Dokdonia sp., rep 3	Dok	91 437	10 784	11.8
Bacterioplankton				
Bermuda, tropical				
20m	B	511 525	146 530	28.6
50m	B	365 838	87 240	23.8
100m	B	519 951	143 907	27.7
OMZ, experimental incubation				
OMZ t0	BAE	27 300	9 805	35.9
OMZ t1	BAE	105 274	58 240	55.3
OMZ t2	BAE	64 463	29 590	45.9
Monterey Bay[2]				
10m sample WCR3	BAE	248 016	82 932	33.4
10m sample WCR5	BAE	238 635	90 767	38.0
10m sample WCR6	BAE	235 339	82 501	35.1
10m sample BAC16	BAE	102 024	40 833	40.0

[1] probe: B = Bacterial 16S and 23S (primers in Table 1); A = Archaeal 16S and 23S (Table S3); E = Eukaryotic 18S and 28S (Table S3); Dok = *Dokdonia* strain-specific 16S and 23S (Table S3)
[2] Archaeal 16S probe excluded – PCR yielded multiple bands

Table 4. Dataset (DS) comparisons – non-rRNA sequences mapped to non-redundant (nr) NCBI reference sequences

DS compared[1]		total refs[2]		refs unique to DS[3]		% refs	refs w/ sig. diff.	% reads in sig. diff. refs[6]	
DS1	DS2	DS1	DS2	DS1	DS2	shared[4]	abundance[5]	DS1	DS2
25m S1	25m S2	21 011	33 097	13 261	25 347	16.7	0	0	0
25m U	25m S	4 110	46 358	1 732	43 980	4.9	13	7.0	4.0
75m AS1*	75m AS2*	4 278	11 040	2 978	9 740	9.3	144	27.0	4.1
75m AS1	75m AS2	4 231	11 011	2 939	9 719	9.3	30	11.0	7.5
75m AU1	75m AU2	1 275	4 193	975	3 893	5.8	6	2.7	0
75m BU1	75m BU2	1 086	6 794	747	6 455	4.5	2	0.8	0
75m AS*	75m BS*	14 018	14 860	10 434	11 276	14.2	75	15.0	8.5
75m AS	75m BS	13 950	14 790	10 384	11 224	14.2	7	7.2	5.8
75m AU*	75m BU*	5 213	7 586	3 955	6 328	10.9	14	3.0	3.5
75m AU	75m BU	5 168	7 541	3 918	6 291	10.9	0	0	0
75m U	75m S	11 459	25 174	7 166	20 881	13.3	3	1.5	1.9
25m all	75m all	48 090	32 340	36 341	20 591	17.1	306	18.0	22.0

[1] as listed in Table 2, where * represents dataset comparisons without removal of replicate sequences and samples without a specified run number (i.e. 25m S) represent comprehensive datasets of all runs associated with that sample (i.e. 25m S1 and 25m S2 combined)
[2] total number of reference genes identified via BLASTX of non-rRNA reads against the NCBI non-redundant (nr) database (e-value ≤ $1x10^{-5}$)
[3] reference genes present in only one dataset
[4] distinct nr-reference genes shared between datasets, as a percentage of total distinct reference genes identified via BLASTX of the two datasets under comparison
[5] reference genes differing significantly in abundance (reads per reference) between datasets ($P < 0.05$)
[6] percentage of total reads matching (via BLASTX) reference genes that differ significantly in abundance

Table 5. NCBI-nr reference genes differing significantly in abundance between technical replicates A and B of the rRNA-subtracted HOT-186 75 m sample

nr reference genes	percentage of BLASTX hits[1]		P-value[2]	FDR[3]
	75m A	75m B		
EDZ60346: proteorhodopsin [*Candidatus Pelagibacter* sp. HTCC7211]	0.125%	0.004%	4.6E-10	1.2E-05
ZP_01223243: flagellar protein [*marine gamma proteobacterium* HTCC2207]	0.144%	0.025%	3.1E-07	2.6E-03
ZP_01612947: hypothetical protein ATW7_13848 [*Alteromonadales bacterium* TW-7]	0.091%	0.004%	2.2E-07	2.7E-03
ZP_01048944: RNA polymerase sigma-70 factor [*Cellulophaga* sp. MED134]	0.072%	ND	6.1E-07	3.9E-03
YP_001090510: ammonium transporter [*Prochlorococcus marinus* str. MIT 9301]	6.591%	5.743%	2.6E-06	1.3E-02
YP_001483709 bacteriochlorophyll synthase [*Prochlorococcus marinus* str. MIT 9215]	0.061%	ND	5.3E-06	2.2E-02
YP_002126505: CN5-related N-acetyltransferase [*Alteromonas macleodii* 'Deep ecotype']	0.068%	0.004%	1.2E-05	4.4E-02

[1] percentage of the total number of significant hits to nr (e-value ≤ $1x10^{-5}$) via BLASTX of non-rRNA reads
[2] P-value as calculated in Audic and Claverie (1997) for pairwise tests of differential abundance
[3] P-values following an FDR-correction for multiple tests (Benjamini and Hochberg 1995)