Duplex-specific nuclease efficiently removes rRNA for prokaryotic RNA-seq

Hana Yi[1], Yong-Joon Cho[2], Sungho Won[3], Jong-Eun Lee[4], Hyung Jin Yu[4], Sujin Kim[4], Gary P. Schroth[5], Shujun Luo[5] and Jongsik Chun[1,2,6,*]

[1]Institute of Molecular Biology and Genetics, [2]School of Biological Sciences & Institute of Bioinformatics (BIOMAX), Seoul National University, Seoul, [3]Department of Statistics, Chung-Ang University, Seoul, [4]DNA Link, Inc., Sungsan-dong 591-4, Mapo-gu, Seoul, Republic of Korea, [5]Illumina, Inc., 25861 Industrial Blvd., Hayward, CA, USA and [6]Chunlab, Inc., Seoul National University, Seoul, Republic of Korea

Received April 12, 2011; Revised June 11, 2011; Accepted July 14, 2011

ABSTRACT

Next-generation sequencing has great potential for application in bacterial transcriptomics. However, unlike eukaryotes, bacteria have no clear mechanism to select mRNAs over rRNAs; therefore, rRNA removal is a critical step in sequencing-based transcriptomics. Duplex-specific nuclease (DSN) is an enzyme that, at high temperatures, degrades duplex DNA in preference to single-stranded DNA. DSN treatment has been successfully used to normalize the relative transcript abundance in mRNA-enriched cDNA libraries from eukaryotic organisms. In this study, we demonstrate the utility of this method to remove rRNA from prokaryotic total RNA. We evaluated the efficacy of DSN to remove rRNA by comparing it with the conventional subtractive hybridization (Hyb) method. Illumina deep sequencing was performed to obtain transcriptomes from *Escherichia coli* grown under four growth conditions. The results clearly showed that our DSN treatment was more efficient at removing rRNA than the Hyb method was, while preserving the original relative abundance of mRNA species in bacterial cells. Therefore, we propose that, for bacterial mRNA-seq experiments, DSN treatment should be preferred to Hyb-based methods.

INTRODUCTION

RNA-seq is a novel method for elucidating the transcriptome of cells. This method uses high-throughput next-generation sequencing technology and has revolutionized the way in which gene expression profiles are examined (1). The RNA molecules present in prokaryotic cells are mostly rRNA species, whereas mRNA constitutes only 1–5% of total RNA. Therefore, efficient enrichment of mRNA is a critical step for successful mRNA-seq experiments. Because mRNA molecules extracted from bacterial cells mostly lack poly-A tails, the methods developed so far have focused on removing non-mRNAs rather than selecting mRNAs. Several techniques (2–6) have been applied to deplete rRNA from the total bacterial RNA population, but the efficiency and robustness of these methods have not been objectively compared. Recently, He *et al.* (7) compared the two most popular rRNA removal methods, namely subtractive hybridization (Hyb) and exonuclease digestion, using Illumina-based RNA-seq of synthetic microbial metatranscriptomes. Their results suggested that the Hyb method introduced less bias in the relative proportion of the mRNA population compared to exonuclease digestion. However, no study has been conducted to evaluate these rRNA removal methods in mRNA-seq based on pure cultured strains.

Zhulidov *et al.* (8) introduced a simple cDNA normalization method based on duplex-specific nuclease (DSN) aimed at enhancing the detection of rare transcripts in eukaryotic cDNA libraries by decreasing the prevalence of highly abundant transcripts. This DSN method includes the denaturation of cDNA, its subsequent reassociation and enzymatic degradation of the double-stranded (ds) DNA fraction using DSN isolated from the Kamchatka crab (9). Because the Hyb rate for each transcript is proportional to the square of its concentration (10), abundant transcripts form ds DNA more effectively during the reassociation step and are subjected to DSN-mediated degradation. DSN has a strong preference for cleaving dsDNA, and there is no significant cleavage of single-stranded (ss) DNA under the directed working conditions of the enzyme (9).

The DSN method has been successfully applied to normalize transcripts in cDNA libraries from various eukaryotes (11). The treatment was usually performed on cDNA libraries enriched with mRNAs using either mRNA-specific poly(A) tail selection in the RNA state (12) or

[*] To whom correspondence should be addressed. Tel: +82 2 880 8153; Fax: +82 2 874 8153; Email: jchun@snu.ac.kr

an oligo(dT) primer approach for reverse transcription from total RNA (13,14). However, the application of the DSN method for the purpose of rRNA removal from total RNA has not been reported in prokaryotic or eukaryotic transcriptome studies. Here, the use of DSN normalization as an rRNA removal method was evaluated and compared to the conventional subtractive Hyb method. Illumina deep sequencing of the transcriptomes of *Escherichia coli* grown under four conditions demonstrated that the DSN method is suitable for rRNA removal while preserving the original relative abundance of each mRNA transcript.

MATERIALS AND METHODS

Bacterial cultures

Escherichia coli K-12/MG1655 was grown in LB medium (Difco) at 37°C with continuous shaking. The freshly grown cells were inoculated into two culture flasks containing LB medium and incubated under anaerobic or aerobic conditions. At exponential phase ($OD_{0.6}$), the aerobic culture was evenly distributed into three sterile culture flasks under aseptic conditions. Three aliquots were then subjected to three different conditions as follows. The first aliquot was subjected to instant RNA extraction at exponential phase, and the second was extracted at stationary phase ($OD_{2.0}$). The third aliquot was subjected to heat shock stress by incubating the culture at 42°C for 30 min. The anaerobic cells were also grown to exponential phase ($OD_{0.6}$), and the cells were subjected to instant RNA extraction.

RNA extraction, rRNA removal and sequencing library construction

Total RNA was independently extracted from the four culture conditions (aerobic exponential, aerobic stationary, aerobic heat shock and anaerobic exponential) using the hot phenol method with additional purification using an RNeasy Mini kit (Qiagen) following the manufacturer's instructions. The quantity and quality of the RNA were evaluated before and after the rRNA removal processes using RNA electropherograms (Agilent 2100 Bioanalyzer) and the RNA integrity number (RIN) (15). Total RNAs from cultures treated under the four conditions were aliquoted into three portions. The first aliquot of total RNA (200 ng) was used to generate a sequencing library using an mRNA-seq library prep kit (Illumina) without any other treatment. The RNA was directly subjected to fragmentation without the mRNA purification step (poly-A selection). The second aliquot was subjected to a subtractive Hyb-based rRNA removal process using the MICROBExpress Bacterial mRNA Enrichment Kit (Ambion). The resultant RNA (100 ng) was used for sequencing library construction using the mRNA-seq library prep kit, omitting the poly-A selection step. The last aliquot (200 ng) was used to generate a sequencing library using an mRNA-seq library prep kit with some modifications. The RNA was directly subjected to fragmentation without the mRNA purification step (poly-A selection). The first- and second-strand cDNA was synthesized from the fragmented RNA using random hexamer primers. End repair, A-tailing, adaptor ligation, cDNA template purification and enrichment of the purified cDNA templates using PCR were then performed. The resulting sample libraries were subjected to DSN treatment using the Trimmer-Direct cDNA Normalization Kit (Evrogen) as follows. The sample library mixed with Hyb buffer was denatured at 98°C for 2 min and incubated at 68°C for 5 h. DSN buffer and 2 µl of the DSN enzyme were added to the mixture and incubated at 68°C for 25 min followed by the addition of stop solution. After purification of the DSN-treated library using SPRI beads, the library was enriched by PCR using PE1.0 and PE2.0 primers. The library construction was completed by final purification of the PCR product using SPRI beads. Because the commercial Hyb and DSN kits already employed highly optimized reagents and conditions for *E. coli* RNA, we adopted the procedures suggested by the respective manufacturers. The summary of the experimental procedures of the control, DSN and Hyb methods is given in Supplementary Figure S1.

Sequencing and alignment of the transcriptome

RNA deep sequencing was performed using two runs of the Illumina Genome Analyzer IIx to generate single-ended 36-bp reads. The genome sequence and functional annotation information of this strain were obtained from the NCBI database (accession number NC_000913.2). Quality-filtered reads were aligned to the reference genome sequence using CLC Genomics Workbench 4.0 (CLC bio). Mapping was based on the minimal length of 32 bp with an allowance of up to two mismatches. The relative transcript abundance was measured in reads per kilobase of exon per million mapped sequence reads (16) (RPKM) using the following formula:

$$\mathrm{RPKM} = \frac{10^9 \times (\text{number of mapped reads of an mRNA})}{(\text{total number of mapped reads in a sample}) \times (\text{sum of the exons in base pairs})}$$

Determination of detection threshold

The library-size normalization was performed by dividing the raw read count of each mRNA by the number of total mapped reads in each Illumina lane and then multiplying by the average total mapped read numbers of four control samples. The detection threshold was determined by calculating the number of reads of an mRNA that was significantly different from zero read considering the experimental errors. For this calculation, duplicate control RNA data were analyzed under the assumption that controls 1 and 2 should have an identical true expression level, but measurement errors may cause different observed expression levels. The sample standard deviation was used to calculate the confidence interval that distinguished the observed expression level of an mRNA from an undetected mRNA. The detailed calculations are as follows.

It was assumed that $x_{i1}(x_{i2})$ was an observed expression level for control 1 (control 2) and that μ_i was an unknown true expression level for both x_{i1} and x_{i2}. If x_{i1} was 0, it was removed from the analysis, and it was assumed that x_{i1} was larger than 0. For the measurement error, the following log-normal distributions for x_{i1} and x_{i2} were assumed, respectively:

$$\log_{10} x_{i1} \sim N\left(\mu_i, \sigma^2_{\mu_i}\right)$$
$$\log_{10} x_{i2} \sim N\left(\mu_i, \sigma^2_{\mu_i}\right),$$

where x_{i1} and x_{i2} are independent. The variance in x_{i1} and x_{i2} depends on μ_i because the proportion of the measurement error got smaller for higher μ_i, and results in Supplementary Figure S3a also confirmed that the inverse of their variance was proportional to the true expression level. Because the mean parameters for x_{i1} and x_{i2} were assumed to be equal, the following could be obtained:

$$\log_{10} x_{i1} - \log_{10} x_{i2} \sim N\left(0, \sigma^2_{\mu_i}\right) \Rightarrow \frac{1}{\sqrt{2}} \log_{10} \frac{x_{i1}}{x_{i2}} \sim N\left(0, \sigma^2_{\mu_i}\right).$$

As a result, the confidence interval for x_{i1} given μ_i could be empirically calculated using $\frac{1}{\sqrt{2}} \log_{10} \frac{x_{i1}}{x_{i2}}$. To minimize the dependence of the variance on μ_i, only the observed intensities for which $\frac{1}{\sqrt{2}} \log_{10} x_{i1} x_{i2}$ was <1 were considered. If the variance was denoted as $\hat{\sigma}^2_{<1}$, the normality of the observed expression level resulted in the confidence interval $(-\alpha, 3.09\sqrt{\hat{\sigma}^2_{<1}})$ at a 0.001 significance level.

Statistical analyses

Statistical analyses were performed using the RPKM values of mRNAs detected from all experimental conditions after detection threshold filtering using library size-normalized data. The statistical significance (*P*-value) of differences in rRNA removal efficiency was obtained using the likelihood ratio test based on a generalized linear model based on Poisson regression. The robustness of mRNA relative abundance conservation was analyzed using general linear regression and Lowess nonlinear regression models. Hierarchical clustering was performed using the unweighted pair group method with arithmetic mean (UPGMA) clustering algorithm using Pearson's product-moment correlation coefficient. All statistical analyses were performed using the R package, version 2.11.0 (www.r-project.org).

RESULTS

Illumina deep sequencing

Four RNA samples prepared from *E. coli* K12/MG1655 cells grown under four conditions (aerobic exponential, aerobic stationary, aerobic heat shock and anaerobic exponential) were aliquoted and subjected to three different protocols. The first aliquot (control) was processed without any rRNA removal treatment; the second aliquot was treated using DSN normalization (DSN; Trimmer-Direct cDNA Normalization Kit, Evrogen); and the third aliquot was treated using subtractive Hyb (MICROBExpress Bacterial mRNA Enrichment kit, Ambion). The RNA quality measured using RNA electropherograms showed that the extracted total RNA was of good quality, with an average RNA integrity number (RIN) of 9.2 (Supplementary Figure S2). The disappearance of rRNA peaks after the rRNA removal process using Hyb was also visualized in the electropherograms.

DNA sequencing was performed using two eight-lane flow cells of the Illumina Genome Analyzer IIx to generate single-ended 36-bp reads. The first run contained eight lanes of untreated control samples, which consisted of duplicate lanes of each condition to allow an increased number of mRNA reads for samples. The second run contained four lanes of DSN-treated samples and four Hyb-treated samples. The number of quality-filtered reads for each sample ranged from 32 to 39 million, and >99.0% of the reads (average 99.4%) were mapped to the reference genome sequence, indicating good sequencing quality and negligible contamination (Table 1). The Illumina reads of duplicate control sample lanes were combined for further analyses.

The sequence coverage is defined as the proportion of mRNAs that have one or more mapped reads with respect to all annotated genes (4493 genes in the reference *E. coli* genome). The average coverage of the control samples was 98.6%, which is not significantly different from those treated with DSN (99.1%) or Hyb (99.2%), suggesting that the sequencing depths in this study were sufficient (Supplementary Table S1). The resultant expression profile of *E. coli* is shown in Supplementary Table S2, and the top 10 highly expressed genes in each growth and rRNA treatment conditions are summarized in Supplementary Table S3.

The composition of major RNA types, namely rRNA, tRNA, mRNA and other RNA (miscRNA), was similar to that of prokaryotes (7).

rRNA removal efficiency

Illumina deep sequencing of the total RNA in *E. coli* revealed that rRNA was indeed the major component in all four growth conditions, ranging from 93.1% to 94.5%, whereas a substantially smaller proportion (2.9–3.8%) was identified as mRNA (Table 1). After the rRNA removal treatments, the proportion of rRNA in the cDNA libraries was reduced to 13.3–38.6% using DSN and 60.4–79.5% using Hyb (Figure 1). The mapped mRNAs increased 17.3-fold using DSN and 6.5-fold using Hyb compared to the untreated controls. The difference in the efficiency of rRNA removal treatments was statistically significant ($P = 0.00007$). The rRNA removal efficiency of DSN was 2.5 times higher than Hyb. However, the ratio of unmapped reads was higher in samples that underwent Hyb (0.86%) than those treated with DSN (0.39%) or than the control (0.50%). The efficiency of rRNA removal in DSN method increased in the order of rRNA

Table 1. Sequencing and alignment statistics

Description	Run ID	Total reads	Mapped reads	mRNA	rRNA	tRNA	misc_RNA	intergenic	Unmapped reads
Control									
Exponential	Ex-C (Ex-C1, Ex-C2)	73 501 029	73 146 186	2 752 700	69 216 309	10 917	116 628	1 049 632	354 843
Stationary	St-C (St-C1, St-C2)	65 848 252	65 554 024	2 482 828	62 197 077	1389	195 905	676 825	294 228
Heat shock	He-C (He-C1, He-C2)	76 958 662	76 537 899	2 904 834	71 652 945	6490	183 421	1 790 209	420 763
Anaerobic	An-C (An-C1, An-C2)	73 860 611	73 469 539	2 120 611	69 437 582	1435	274 246	1 635 665	391 072
DSN									
Exponential	Ex-D	32 461 464	32 346 699	18 516 330	9 593 620	55 732	425 762	3 755 255	114 765
Stationary	St-D	34 429 138	34 274 833	24 816 321	4 566 206	11 204	747 676	4 133 426	154 305
Heat shock	He-D	33 583 557	33 470 900	21 926 911	7 125 455	38 092	604 116	3 776 326	112 657
Anaerobic	An-D	32 309 941	32 179 277	16 188 510	12 459 843	7358	727 996	2 795 570	130 664
Hyb									
Exponential	Ex-H	39 013 862	38 765 857	9 461 600	25 217 660	29 688	253 429	3 803 480	248 005
Stationary	St-H	35 007 172	34 706 767	7 897 907	22 624 587	5341	406 012	3 772 920	300 405
Heat shock	He-H	39 419 310	39 030 251	12 374 014	23 788 881	32 205	620 168	2 214 983	389 059
Anaerobic	An-H	37 470 516	37 106 501	5 635 773	29 788 740	7606	519 716	1 154 666	364 015

The statistics of untreated control samples were obtained from data combined from two lanes.

size (23 S > 16 S > 5 S), but the ratio of 5 S:16 S:23 S was pretty conserved in DSN (0.1:35:65) compared to control (0.02:32:68), while the ratio was significantly shifted in Hyb (0.4:9:91).

RNA-seq detection threshold

To improve the accuracy of the RNA-seq analysis, the mRNA detection threshold was determined to exclude Illumina reads in which the expression level was severely affected by measurement errors. It was assumed that the biological duplicates, namely controls 1 and 2 under the same conditions, had unknown identical means and that the measurement errors caused the differences in the Illumina read counts. To determine the detection threshold value, read count 1 was used as the one-tailed upper limit of minimum read at the 0.001 significance level. Because the variance of the measurement decreased with the level of the unknown true expression level (Supplementary Figure S3a), the area in which the variance showed normal distribution was determined first (Supplementary Figure S3b) to minimize the dependence of the variance on expression values. For the selected area, the variance was estimated, and the detection threshold of an mRNA was determined to be six Illumina reads in our library size-normalized data at the 0.001 significance level. Consequently, 94.58% of mRNAs that had more than six mapped reads in all samples that passed the filtering. The 234 failed ORFs (5.42%) were omitted from further statistical analyses. By removing the insignificant low read-count genes, the linearity of regression between the two duplicate controls was improved (Supplementary Figure S3c and d).

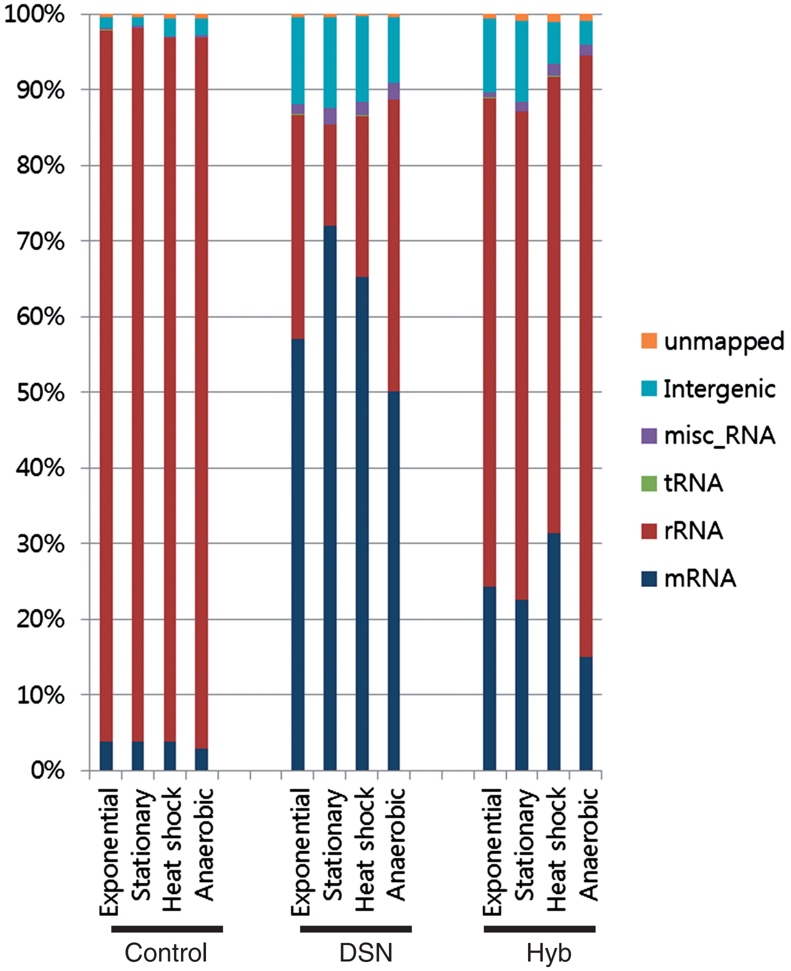
Robustness of mRNA relative abundance

The correlation of mRNA expression patterns between control samples and the two rRNA removal treatments is summarized in Figure 2a and Supplementary Figure S4. The average slope of the linear regression line between the controls and DSN-treated samples was 0.99 ($r = 0.99$, Pearson's correlation coefficient), whereas the corresponding value between the controls and Hyb-treated samples was significantly lower (0.75 on average, $r = 0.93$). Similarly, the Lowess fit of DSN-treated samples converged to linear regression, whereas Hyb-treated samples showed a departure from linearity, with a skewed shape in the area of the low to middle 'expressers'. Hierarchical clustering analysis of all samples using the Pearson's product-moment correlation coefficient indicated that the mRNA expression profiles of the untreated controls were more similar to that of DSN-treated samples than that of Hyb-treated samples (Figure 2b).

Robustness of mRNA profiles

The fold-change in expression levels of each mRNA was calculated by dividing the RPKM of stationary (St), heat shock (He) or anaerobic (An) samples by the RPKM of the exponential (Ex) sample and these fold-change values were represented as St/Ex, He/Ex or An/Ex, respectively. The fold difference between the control and the two rRNA removal treatments was calculated by subtracting the log-scale fold value of the corresponding control sample from that of DSN- or Hyb-treated sample. The regression analyses (Supplementary Figure S5) and boxplots (Figure 3) indicated a smaller fold difference in samples treated with DSN than Hyb. The slopes of the linear regression lines of DSN- and Hyb-treated samples were 0.96 ($r = 0.98$) and 0.74 ($r = 0.90$), respectively. The Lowess fit analysis of Hyb-treated samples also showed a departure from linearity, whereas the fit of DSN-treated samples converged to linear regression. The boxplots (Figure 3) demonstrated no fold differences in DSN-treated samples, with average values close to 0, whereas the corresponding values of Hyb-treated samples ranged between 0.03 and 0.10, implying significant expression level differences caused by the Hyb treatment.

The transcripts for which relative abundance was severely biased using Hyb treatment were identified



	Control	DSN	Hyb
rRNA (%)	93.9 ± 0.6	25.6 ± 10.9	67.3 ± 8.4
mRNA (%)	3.5 ± 0.5	61.1 ± 9.6	23.3 ± 6.7

Figure 1. rRNA removal efficiency of DSN and Hyb methods. The proportion of Illumina reads assigned to mRNA or rRNA indicates that DSN is more effective at rRNA removal than Hyb.

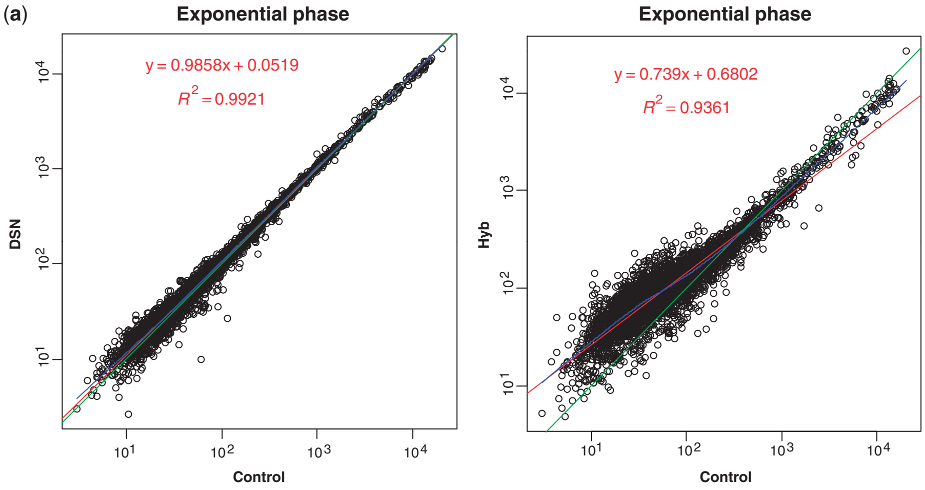


Figure 2. Comparison of robustness of mRNA relative abundance using DSN and Hyb. (**a**) Correlation of relative gene expression between the control and rRNA removal treatments. The exponential phase libraries (Ex) are shown as an example. Left, control (Ex-C) versus DSN (Ex-D); right, control (Ex-C) versus Hyb (Ex-H). The points in the plot indicate the RPKM value of each individual mRNA transcript. Red indicates the linear regression line, and blue indicates non-linear regression (Lowess) fit. Green is a straight line with a slope of 1. (**b**) Hierarchical clustering of all samples tested in this study using UPGMA based on Pearson's correlation. The regression plot and dendrogram represent the superior conservation of mRNA relative abundance in DSN treatment compared to Hyb treatment.

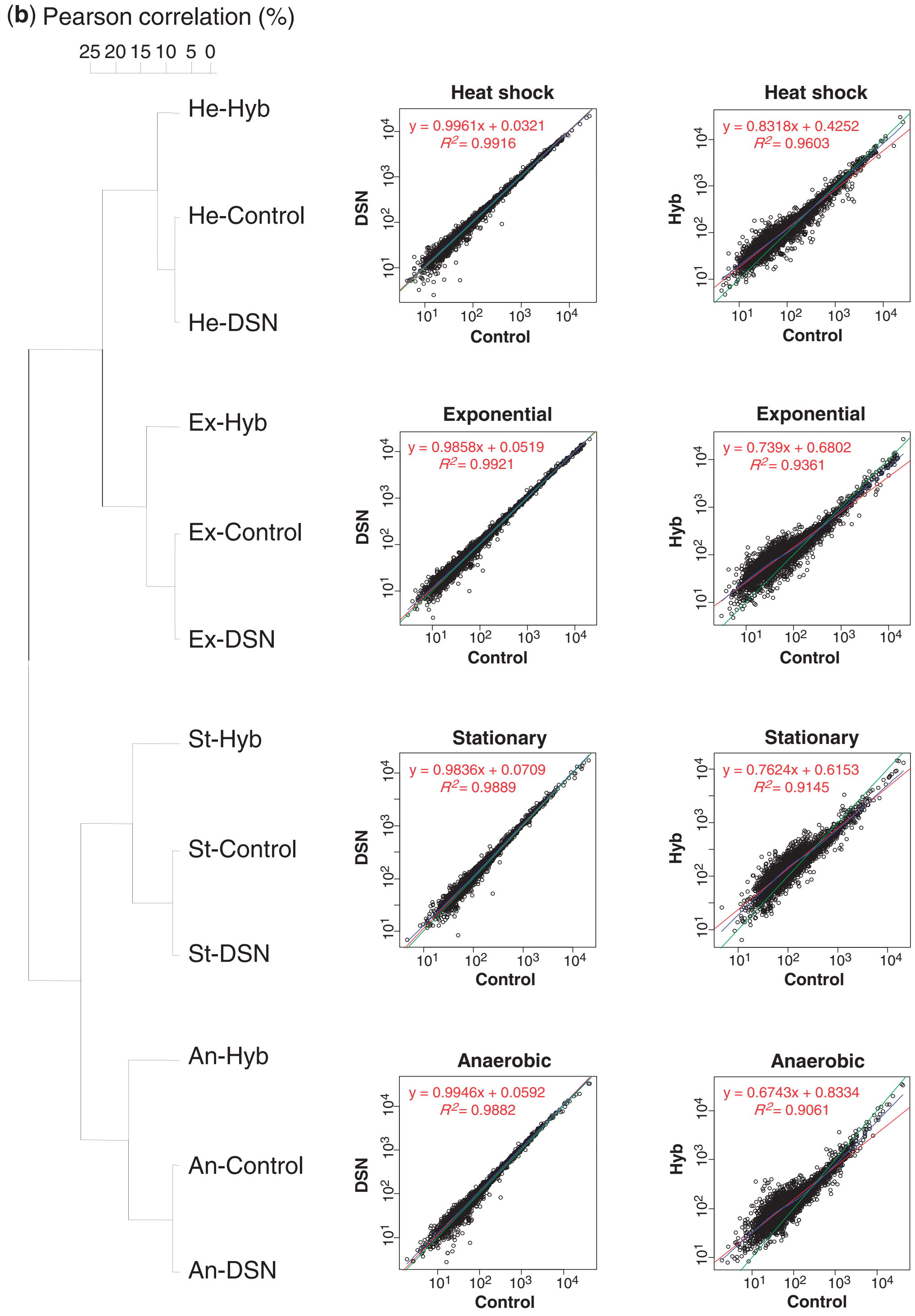


Figure 2. Continued.

using a threshold of 0.5 at Δlog (fold value) because the majority of transcripts of DSN-treated samples were within the threshold (Figure 3). A total of 127 ORFs were identified from Hyb-treated samples (Supplementary Table S4), and their functional categories are provided in Supplementary Figure S6. No correlation between the functional categories of the transcripts and bias was detected. No selective loss or gain of mRNAs depending on GC content was either observed in this data set.

DISCUSSION

Given the prevalence of rRNA in the total transcriptome of prokaryotes, it is essential to enrich mRNA prior to sequencing-based genome-wide gene expression studies.

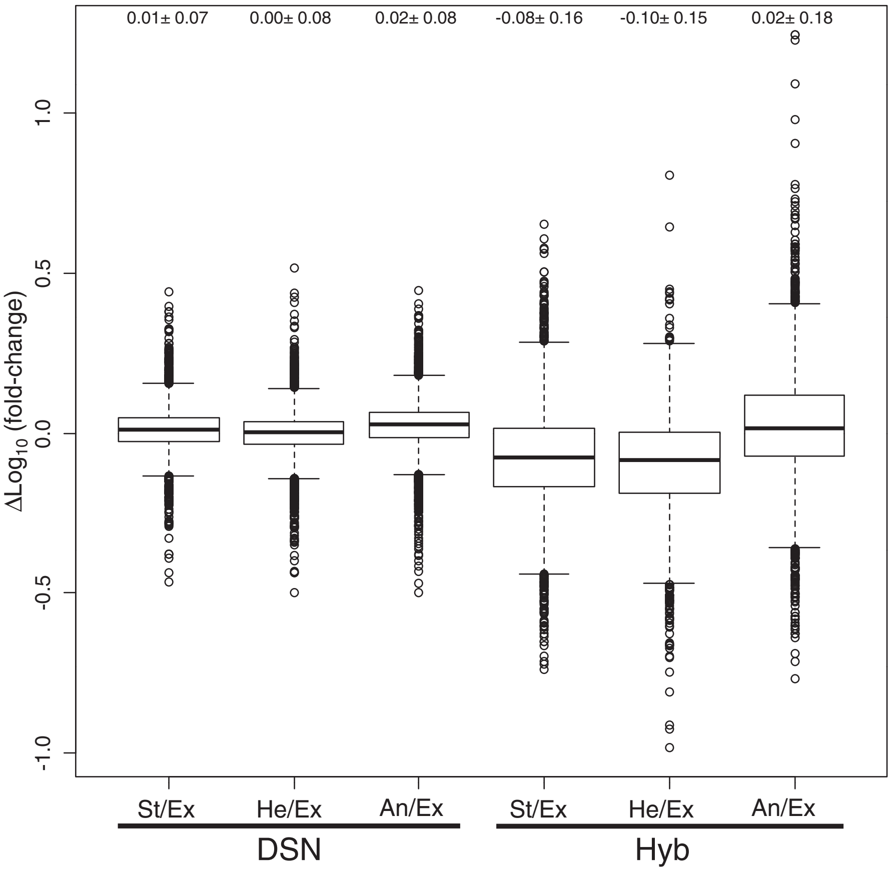


Figure 3. Boxplots showing fold difference profiles. The fold-change of each mRNA was calculated by dividing the RPKM of St, He or An samples by the RPKM of the Ex sample and is represented as St/Ex, He/Ex or An/Ex, respectively. The fold difference profile caused by rRNA removal treatment was calculated by subtracting the log-scale fold-change of the corresponding control sample from the DSN- or Hyb-treated samples. The average and standard deviation are shown at the top of the plots. No fold difference was observed using DSN, with the average values close to 0, whereas the corresponding values of Hyb ranged between 0.10 and 0.03, implying a significant expression level difference caused by the Hyb treatment.

However, it is equally important to preserve the overall gene expression patterns while enriching the mRNA. In this study, we evaluated two commercially available methods for performing such a task using *E. coli*, the most widely used model bacterium.

In our study, both the Hyb and DSN methods removed a substantial proportion of rRNA species, with DSN being 2.5-fold more efficient than Hyb. Our results suggest that researchers can reduce the sequencing cost of RNA-seq by 2.5-fold by using DSN method, compared to the commercial Hyb kit. The rRNA removal efficiency of Hyb method obtained in this study was comparable with the previous studies. It has been known that the rRNA removal efficiency of Hyb method varies widely for community RNA samples (17,18), as well as for single-species analyses (19). For example, the amount of rRNA remained after the commercial Hyb treatment ranged from 43.6% to 98.6% depending on microbial species (7). This is because the Hyb method is based on Hyb between rRNA and oligonucleotides that target conserved regions of bacterial rRNA; therefore, the removal efficiency is largely dependent on the selected oligonucleotide sequences. In contrast, the DSN method does not depend on particular rRNA sequences; therefore, in theory, it can be used for any organism, including archaea. However, the variation of rRNA removal efficiency between samples was also observed in DSN treatment, but the reason is unclear at this stage.

The regression analyses demonstrated the lower robustness of the Hyb method compared to DSN, especially in the low- to middle-expression range (approximate RPKM < 300). Many transcripts in this range seemed to be expressed at higher levels than those in the untreated controls. The depletion of high expressers may result in the relative overrepresentation of lower expressers. A number of severely biased transcripts were found in samples that underwent Hyb treatment, although no correlation between the functional categories of the transcripts and bias was found. Undesired binding between mRNAs and the capture oligonucleotides may cause this phenomenon. Indeed, in the case of ribosomal proteins (*rps, rpl* and *rpm*) that have obvious homology with rRNAs, the average RPKM value was reduced to 76% of the untreated control after Hyb treatment, whereas it was conserved at 105% in DSN-treated samples. An increased proportion of unmapped reads after Hyb treatment compared to the control or DSN treatment was also noted, although the reason for this is unclear.

Because of the function of the DSN enzyme, which preferentially degrades highly abundant transcripts over transcripts of low abundance, we expected that abundant transcripts could be affected by DSN treatment. Because the absolute levels of gene expression in bacteria vary over as much as six orders of magnitude (20–22), the concentration of highly abundant mRNA classes could be affected by DSN degradation. However, as far as we surveyed, even the most abundant mRNA typically comprise 1–7% of mRNAs (6,16,23), and as much as 10% in severe cases (24). Thus, the amount of rRNA in a cell significantly outnumbers even the most abundant mRNA transcripts. In fact, the 10 most expressed genes in each condition showed a little bit declined expression level in DSN (7.9% on average) compared to untreated control as shown in Supplementary Table S3, while the amount of decline was much more severe in Hyb (20.3%). Moreover, the log scale regression analysis in this study proved that the amount of decline observed in the abundant mRNAs did not hamper the overall robustness.

The typical ratio of rRNA:non-rRNA revealed by RNA-seq is 95–99% in a wide taxonomic range of pure cultured bacteria and archaea (7). Thus, our DSN method is, at least, applicable to a broad range of prokaryotes in laboratory culturing condition. In addition, it has been known that the ratio of rRNA compared to mRNA is higher in resting cells than actively growing cells (25,26). Thus, we think that the overwhelming abundance of rRNA compared to non-rRNA would be the case for even slowly growing cells or metabolically inactive cells, though experimental evidence will be required for DSN applicability for these conditions in future. Because the rRNA ratios reported in a metatranscriptomic study are also as high as 74–97% (6), it is fair to say that DSN method would be applicable to mixed population samples.

Several rRNA removal methods are known, but all of these methods have some limitations. The methods based on Hyb between rRNA and DNA targeting rRNA may generate bias depending on the taxonomy of bacteria. These methods include the RNase H digestion method based on reverse transcription with rRNA-specific primers (4) and the Hyb method evaluated in this study. Though recently developed subtractive Hyb method has solved the bias by generating customized oligonucleotide probes targeting sample specific rRNAs (6), it is not still free from non-specific binding of rRNA probe to mRNAs. The size selection method using gel electrophoresis (5) has an apparent limitation because some mRNAs can co-migrate with rRNA and are subsequently omitted. The poly(A) tail addition methods (3,27), which use preferential poly(A) adenylation of mRNA in crude RNA, also have the possible limitation of uneven poly(A) adenylation efficiency among mRNAs, which may generate expression-level bias. Because poly(A) tail addition method have not been tested in pure cultures using RNA-seq, further evaluation is required. In the case of 5′-phosphate-dependent exonuclease digestion of rRNA with 5′ monophosphate, it has already been demonstrated that this method compromises the relative proportion of the mRNA population more severely than the Hyb method (7). This change in the relative proportion may occur because exonuclease can also eliminate 5′-monophosphorylated mRNA species, which are produced during mRNA processing by endoribonucleases (28,29) and RppH (30).

The only drawback of DSN over the Hyb method is that it requires more experimental steps and, therefore, a longer time to prepare cDNA libraries for deep sequencing. However, because the rRNA removal step is performed after cDNA library construction, the possibility of unintentional RNA degradation is greatly reduced compared to other methods involving pre-treatments. In addition, the amount of total input RNA used for deep sequencing library construction was much less in the DSN method (200 ng) than in the conventional mRNA-seq accompanying poly(A) selection (1–10 µg). Considering the loss of RNAs during the mRNA enrichment step using other pre-treatments, the amount of total RNA to be extracted is even smaller using the DSN method.

Another advantage of the DSN method is that it works well even with partially degraded total RNA, whereas the Hyb method requires intact rRNA for the successful binding of the oligonucleotides to targeted conserved sites. Indeed, a well-supported positive correlation between RIN values, which represent the degree of rRNA integrity, and rRNA removal efficiency (r = 0.88) was observed in our Hyb experiments, as well as in a previous report (7); however, this type of correlation was not observed in our DNS treatments (r = 0.27). Although the number of samples (four total RNAs) and the range of RNA degradation (RIN value 8.8–10.0) were not strong enough for statistical analyses, the lower importance of the RNA fragmentation status for effective DNS treatment compared to the Hyb method was clear.

Because the conventional RNA-seq of eukaryotic organisms relies on poly(A)$^+$ capture in the first step of sequencing library preparation, the outcome generally contains only information on poly(A)-tailed RNAs. If the DSN method used in this study is applied to eukaryotic RNA, it probably will provide information on both non-poly(A) RNA sequences and poly(A) RNA. Further study of the feasibility of this method for eukaryotic RNA-seq without poly(A) selection is therefore needed.

In this study, we performed deep sequencing of total RNA of *E. coli*. Although *E. coli* is a well studied, widely used model organism, its precise genome-wide expression profile has not been documented previously. To our knowledge, this is the first report on the intact genome-wide RNA profile of *E. coli* without any treatment and selection. This information clearly demonstrates the overall abundance of different RNA species in a bacterial cell.

DSN-based normalization showed a higher efficiency of rRNA removal than the Hyb method, while preserving the relative abundance of mRNA. The thermodynamic principle of this technique allows its application to any kind of eukaryotic or prokaryotic organism. Therefore, DSN-based mRNA enrichment can be readily used in bacterial mRNA-seq experiments.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Drs Kangseok Lee and You-Hee Cho for their helpful comments and suggestions on this manuscript.

FUNDING

Priority Research Centers Program (grant number 2010-0094020); National Research Foundation grant (2011-0016498) through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology, Republic of Korea. Funding for open access charge: National Science Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.
2. Pang,X., Zhou,D., Song,Y., Pei,D., Wang,J., Guo,Z. and Yang,R. (2004) Bacterial mRNA purification by magnetic capture-hybridization method. *Microbiol. Immunol.*, **48**, 91–96.
3. Frias-Lopez,J., Shi,Y., Tyson,G.W., Coleman,M.L., Schuster,S.C., Chisholm,S.W. and Delong,E.F. (2008) Microbial community gene expression in ocean surface waters. *Proc. Natl. Acad. Sci. USA*, **105**, 3805–3810.
4. Dunman,P.M., Murphy,E., Haney,S., Palacios,D., Tucker-Kellogg,G., Wu,S., Brown,E.L., Zagursky,R.J., Shlaes,D. and Projan,S.J. (2001) Transcription profiling-based identification of *Staphylococcus aureus* genes regulated by the *agr* and/or *sarA* loci. *J. Bacteriol.*, **183**, 7341–7353.
5. McGrath,K.C., Thomas-Hall,S.R., Cheng,C.T., Leo,L., Alexa,A., Schmidt,S. and Schenk,P.M. (2008) Isolation and analysis of mRNA from environmental microbial communities. *J. Microbiol. Methods*, **75**, 172–176.
6. Stewart,F.J., Ottesen,E.A. and DeLong,E.F. (2010) Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. *ISME J.*, **4**, 896–907.
7. He,S., Wurtzel,O., Singh,K., Froula,J.L., Yilmaz,S., Tringe,S.G., Wang,Z., Chen,F., Lindquist,E.A., Sorek,R. *et al.* (2010) Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nat. Methods*, **7**, 807–812.
8. Zhulidov,P.A., Bogdanova,E.A., Shcheglov,A.S., Vagner,L.L., Khaspekov,G.L., Kozhemyako,V.B., Matz,M.V., Meleshkevitch,E., Moroz,L.L., Lukyanov,S.A. *et al.* (2004) Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acids Res.*, **32**, e37.
9. Shagin,D.A., Rebrikov,D.V., Kozhemyako,V.B., Altshuler,I.M., Shcheglov,A.S., Zhulidov,P.A., Bogdanova,E.A., Staroverov,D.B., Rasskazov,V.A. and Lukyanov,S. (2002) A novel method for SNP detection using a new duplex-specific nuclease from crab hepatopancreas. *Genome Res.*, **12**, 1935–1942.
10. Young,B.D. and Anderson,M. (1985) In Hames,B.D. and Higgins,S.J. (eds), *Nucleic Acids Hybridisation, a Practical Approach*. IRL Press, Oxford, Washington DC, pp. 47–71.
11. Bogdanova,E.A., Shagin,D.A. and Lukyanov,S.A. (2008) Normalization of full-length enriched cDNA. *Mol. Biosyst.*, **4**, 205–212.
12. Simon,A., Glockner,G., Felder,M., Melkonian,M. and Becker,B. (2006) EST analysis of the scaly green flagellate *Mesostigma viride* (Streptophyta): implications for the evolution of green plants (Viridiplantae). *BMC Plant Biol.*, **6**, 2.
13. Zhu,Y.Y., Machleder,E.M., Chenchik,A., Li,R. and Siebert,P.D. (2001) Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques*, **30**, 892–897.
14. Danley,P.D., Mullen,S.P., Liu,F., Nene,V., Quackenbush,J. and Shaw,K.L. (2007) A cricket Gene Index: a genomic resource for studying neurobiology, speciation, and molecular evolution. *BMC Genomics*, **8**, 109.
15. Schroeder,A., Mueller,O., Stocker,S., Salowsky,R., Leiber,M., Gassmann,M., Lightfoot,S., Menzel,W., Granzow,M. and Ragg,T. (2006) The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol. Biol.*, **7**, 3.
16. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
17. Poretsky,R.S., Bano,N., Buchan,A., LeCleir,G., Kleikemper,J., Pickering,M., Pate,W.M., Moran,M.A. and Hollibaugh,J.T. (2005) Analysis of microbial gene transcripts in environmental samples. *Appl. Environ. Microbiol.*, **71**, 4121–4126.
18. Hewson,I., Poretsky,R.S., Dyhrman,S.T., Zielinski,B., White,A.E., Tripp,H.J., Montoya,J.P. and Zehr,J.P. (2009) Microbial community gene expression within colonies of the diazotroph, *Trichodesmium*, from the Southwest Pacific Ocean. *ISME J.*, **3**, 1286–1300.
19. Yoder-Himes,D.R., Chain,P.S., Zhu,Y., Wurtzel,O., Rubin,E.M., Tiedje,J.M. and Sorek,R. (2009) Mapping the *Burkholderia cenocepacia* niche response via high-throughput sequencing. *Proc. Natl. Acad. Sci. USA*, **106**, 3976–3981.
20. Taniguchi,Y., Choi,P.J., Li,G.W., Chen,H., Babu,M., Hearn,J., Emili,A. and Xie,X.S. (2010) Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, **329**, 533–538.
21. Wilhelm,B.T., Marguerat,S., Watt,S., Schubert,F., Wood,V., Goodhead,I., Penkett,C.J., Rogers,J. and Bahler,J. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, **453**, 1239–1243.
22. Zaslaver,A., Bren,A., Ronen,M., Itzkovitz,S., Kikoin,I., Shavit,S., Liebermeister,W., Surette,M.G. and Alon,U. (2006) A comprehensive library of fluorescent transcriptional reporters for *Escherichia coli*. *Nat. Methods*, **3**, 623–628.
23. Gilbert,J.A., Field,D., Huang,Y., Edwards,R., Li,W., Gilna,P. and Joint,I. (2008) Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS One*, **3**, e3042.
24. Kim,J.B., Porreca,G.J., Song,L., Greenway,S.C., Gorham,J.M., Church,G.M., Seidman,C.E. and Seidman,J.G. (2007) Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science*, **316**, 1481–1484.
25. Johnson,L.F., Abelson,H.T., Green,H. and Penman,S. (1974) Changes in RNA in relation to growth of fibroblast .1. Amounts of messenger-RNA, ribosomal-RNA, and tertiary RNA in resting and growing cells. *Cell*, **1**, 95–100.
26. ter Kuile,B.H. and Bonilla,Y. (1999) Influence of growth conditions on RNA levels in relation to activity of core metabolic enzymes in the parasitic protists *Trypanosoma brucei* and *Trichomonas vaginalis*. *Microbiology*, **145**, 755–765.
27. Wendisch,V.F., Zimmer,D.P., Khodursky,A., Peter,B., Cozzarelli,N. and Kustu,S. (2001) Isolation of *Escherichia coli* mRNA and comparison of expression using mRNA and total RNA on DNA microarrays. *Anal. Biochem.*, **290**, 205–213.
28. Matsunaga,J., Dyer,M., Simons,E.L. and Simons,R.W. (1996) Expression and regulation of the *rnc* and *pdxJ* operons of *Escherichia coli*. *Mol. Microbiol.*, **22**, 977–989.
29. Sim,S.H., Yeom,J.H., Shin,C., Song,W.S., Shin,E., Kim,H.M., Cha,C.J., Han,S.H., Ha,N.C., Kim,S.W. *et al.* (2010) *Escherichia coli* ribonuclease III activity is downregulated by osmotic stress: consequences for the degradation of *bdm* mRNA in biofilm formation. *Mol. Microbiol.*, **75**, 413–425.
30. Deana,A., Celesnik,H. and Belasco,J.G. (2008) The bacterial enzyme RppH triggers messenger RNA degradation by 5′ pyrophosphate removal. *Nature*, **451**, 355–358.