Validation of two ribosomal RNA removal methods for microbial metatranscriptomics

Shaomei He[1,2,5], Omri Wurtzel[3,5], Kanwar Singh[1], Jeff L Froula[1], Suzan Yilmaz[1], Susannah G Tringe[1], Zhong Wang[1], Feng Chen[1], Erika A Lindquist[1], Rotem Sorek[3] & Philip Hugenholtz[1,2,4]

[1]Department of Energy Joint Genome Institute, Walnut Creek, California, USA. [2]Energy Biosciences Institute, University of California-Berkeley, Berkeley, California, USA. [3]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel. [4]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, Australia. [5]These authors contributed equally to this work. Correspondence should be addressed to P.H. (p.hugenholtz@uq.edu.au).

The predominance of rRNAs in the transcriptome is a major technical challenge in sequence-based analysis of cDNAs from microbial isolates and communities. Several approaches have been applied to deplete rRNAs from (meta)transcriptomes, but no systematic investigation of potential biases introduced by any of these approaches has been reported. Here we validated the effectiveness and fidelity of the two most commonly used approaches, subtractive hybridization and exonuclease digestion as well as combinations of these treatments, on two synthetic five-microorganism metatranscriptomes using massively parallel sequencing. We found that the effectiveness of rRNA removal was a function of community composition and RNA integrity for these treatments. Subtractive hybridization alone introduced the least bias in relative transcript abundance, whereas exonuclease and in particular combined treatments greatly compromised mRNA abundance fidelity. Illumina sequencing itself also can compromise quantitative data analysis by introducing a G+C bias between runs.

Rapid technological advances in ultra-high-throughput sequencing are making *de novo* sequencing of transcriptomes (RNA-seq) a viable alternative to microarray analysis of microbial isolates and communities[1]. A major technical challenge for *de novo* transcriptome sequencing is the low relative abundance of mRNAs in total cellular RNA (1–5%; ref. 2), the bulk of which is rRNAs and tRNAs[3]. Unlike eukaryotic mRNAs, which can be selectively synthesized into cDNA by virtue of their poly(A) tails[4], bacterial and archaeal cDNAs are predominantly rRNA

sequences[5,6]. Therefore, prokaryotic rRNAs are often removed before sequencing to improve mRNA detection sensitivity. Different methods have been used to eliminate prokaryotic rRNA, including subtractive hybridization with rRNA-specific probes[7,8], digestion with exonuclease that preferentially acts on rRNA, poly(A) tail addition to discriminate against rRNA[9,10], reverse transcription with rRNA-specific primers followed by RNase H digestion to degrade rRNA:DNA hybrids[11], and gel electrophoresis size separation and extraction of non-rRNA bands[12].

Among these methods, subtractive hybridization and exonuclease digestion have become the most popular owing to the availability of commercial kits from Ambion (MICROBExpress Bacterial mRNA Enrichment kit) and Epicentre (mRNA-ONLY Prokaryotic mRNA Isolation kit). The former kit uses a subtractive hybridization with capture oligonucleotides specific to 16S and 23S rRNAs. It has been applied to both bacterial isolates and environmental samples, in one or two rounds[6,13–18]. The Epicentre kit uses exonuclease to preferentially degrade processed RNAs with 5′ monophosphate (the majority of which are believed to be rRNAs)[19,20]. In some instances, these methods have been used in combination to improve rRNA removal[21–23]. There is no consensus, however, on the best approach, and existing data are anecdotal. Here we validated the performance of these kits with two synthetic five-member microbial communities using Illumina sequencing and found that rRNA removal efficiencies were community- and RNA integrity–dependent and that only subtractive hybridization adequately preserved relative transcript abundance for quantitative analyses.

RESULTS

Experimental design

We constructed two five-member synthetic microbial communities by pooling equimolar amounts of total RNAs extracted independently from microbial isolates with sequenced genomes that span a wide phylogenetic, (G+C)-content and genome-size range. Two of the five species were common to both communities (**Table 1**). We tested five rRNA depletion methods involving subtractive hybridization (Hyb), exonuclease digestion (Exo) and combined treatments in two experiments using the two synthetic communities. In experiment 1, we tested Hyb, Exo and Hyb followed by Exo (Hyb + Exo) using synthetic community 1. Then we conducted experiment 2 using synthetic community 2 to confirm the results of experiment 1 and expanded the depletion methods tested to also include two rounds of Hyb (2Hyb) as well as Exo followed by Hyb (Exo

+ Hyb) (**Table 2**). As a control, we used total RNA without rRNA removal. We initially assessed rRNA removal using RNA electropherograms (Agilent 2100 Bioanalyzer) (**Supplementary Figs. 1** and **2** and **Supplementary Note 1**). We then sequenced technical replicates (after RNA pooling) distributed within and between four runs (flowcells; runs 1 and 2 in experiment 1, and runs 3 and 4 in experiment 2) on an Illumina Genome Analyzer II sequencer to evaluate intra- and interrun variation (**Table 2**). We generated ~10–17 million 76-base-pair single-end reads for each sample, and for all but one sample we mapped 99% of reads to a reference, indicating good read quality and negligible contamination (**Supplementary Tables 1** and **2** and **Supplementary Note 2**).

Technical reproducibility

We first evaluated technical reproducibility by comparing the relative transcript abundance of technical replicates calculated as reads per gene normalized by total mapped mRNA reads. The four intrarun technical replicates in three independent runs (controls and Hyb in run 1; **Table 2**) were highly reproducible (Pearson's product-moment correlation coefficient, $r = 0.997 \pm 0.001$; **Fig. 1a**) suggesting that both sample preparation–associated variation and intrarun sequencing variation were minor. In addition, the sole interrun technical replicate between runs 1 and 2 (Exo 1 i and Exo 1 ii; **Table 2**) was also highly reproducible (**Fig. 1b**), suggesting negligible technical variation between these two runs. By contrast, the interrun technical replicates between runs 3 and 4 were less reproducible owing to multiple distinct regressions that correspond to individual organisms (**Fig. 1c**). Linear regression slopes for each organism positively correlated with their genomic G+C content (**Fig. 1c**). This indicates that run 3 systematically underrepresented (G+C)-rich sequences relative to run 4, which we believe was due to variability in run quality (**Table 2**). Two simple normalization strategies based on the total mRNA counts from the source organism and gene G+C content (**Supplementary Note 3**) improved the overall correlation (**Fig. 1d**), confirming that the run-to-run variation was indeed largely associated with G+C content. Therefore, to accurately assess rRNA depletion efficiency and mRNA fidelity, we restricted comparisons of treatment-control pairs to the same run in experiment 2 (**Table 2**). We took the interrun average after we performed the intrarun comparison.

Efficiency of rRNA depletion

As expected, the relative rRNA content of the controls was 95–97% of total RNA, typical of bacteria and archaea, with each community member being approximately equally represented in the controls (**Fig. 2a,b**). After the various treatments, the observed rRNA content decreased by as little as 3.6% (Exo) and up to 19.9% (Hyb + Exo). This decrease in rRNA percentage in total RNA reflected the redistribution of rRNA relative to non-rRNA owing to the treatment. For experiment 1, the rank order of rRNA removal efficiencies was Exo < Hyb < Hyb + Exo; for experiment 2, it was Hyb ≈ 2Hyb < Exo < Exo + Hyb < Hyb + Exo ('≈' indicates no statistically significant difference, $P > 0.05$). The difference in rank order of removal efficiencies between the two experiments was due to community composition differences (see below), although combined treatments resulted in the greatest removal in both experiments (**Fig. 2**).

At first glance, the applied rRNA depletion methods did not appear to be very effective, but the observed postdepletion rRNA fraction did not accurately reflect the amount of rRNA that was actually removed, with observed rRNA removal not exceeding 20% until actual removal was > 80% (**Fig. 2c** and **Supplementary Note 4**). For each community member we also calculated actual rRNA removal based on the relative proportion of rRNA and non-rRNA for individual organisms (**Supplementary Tables 3** and **4** and **Fig. 2d**). *Halorhabdus utahensis* exhibited the lowest percentage removal for all five treatments in experiment 2. We expected this for the Hyb treatments because archaea are not targeted by the Ambion Hyb kit (**Table 1**). Low removal of *Halorhabdus* rRNAs by exonuclease, however, was unexpected (**Fig. 2b,d**). We suspect that some of the processed 5′ ends of archaeal rRNAs may have exposed hydroxyl groups rather than a monophosphate because adapters requiring an exposed 5′-monophosphate ligated poorly to archaeal *Sulfolobus solfataricus* rRNAs[24]. This indicates that it will be difficult to deplete rRNAs in microbial communities dominated by archaea using the tested methods.

Percentage removal of *Streptomyces* sp. str. LCC rRNA was markedly lower than average for all three of the applied treatments in experiment 1 (**Fig. 2d**). The most poorly performing treatment for this organism was Hyb, with only 15% actual and 0.2% observed rRNA removal. Efficient removal of *Catenulispora acidiphila* rRNA, another member of the phylum Actinobacteria, ruled out the possibility that actinobacterial rRNA is generally poorly targeted by Hyb (**Fig. 2d**). Rather, we noticed that the *Streptomyces* RNA had the lowest RNA integrity number[25] (RIN of 5.5; **Table 1**). Intact rRNA is important for the success of the Hyb treatment

because degradation increases the proportion of rRNA fragments without the conserved Hyb target sites[18]. This likely explains the reduced effectiveness of the Hyb treatment despite the presence of the rRNA target sites in this organism. Lower RNA integrity may also explain the apparently reduced effectiveness of the Exo treatment on *Streptomyces* rRNA because partially degraded mRNAs can be targeted by exonuclease skewing the mRNA:rRNA ratio used to calculate actual rRNA removal (Online Methods) Lower integrity of *Desulfovibrio vulgaris* RNA (RIN 7.4) may similarly explain lower rRNA removal for this organism in some of the treatments (**Fig. 2d**). Indeed, we observed a high positive correlation between rRNA removal and RIN values for most treatments (**Supplementary Fig. 3**) suggesting that RIN can be used as an indicator of likely rRNA removal efficiency for these methods.

Improvement in mRNA detection sensitivity

We calculated the fold enrichment in mRNAs by comparing the proportion of total mapped mRNA reads between the control and each treatment (**Fig. 3a**), which produced results ranging from 1.9- to 5.7-fold, largely consistent with the observed rRNA removal. Although mRNA fold enrichment is an intuitive way to express the effect of rRNA removal treatments, a more informative metric may be improvement in mRNA detection sensitivity calculated as the percentage increase in the number of detected mRNAs in a given treatment relative to the number of detected mRNAs in the corresponding control.

Defining the detection threshold as 30 mapped reads per gene (after normalization for dataset size; Online Methods), we calculated a 50% to 230% increase in mRNA detection sensitivity (**Fig. 3b**). This shows that even a modest depletion of rRNA (3.6% observed removal) resulted in more efficient use of sequence data (50% increase in mRNA detection sensitivity) justifying the application of rRNA removal treatments.

Two rounds of subtractive hybridization marginally increased the detection sensitivity improvement from 159% for one-round Hyb to 172% for two rounds ($P = 0.004$). When testing the effect of the combination of treatments, the order of applying the rRNA removal kits produced statistically different results ($P = 0.0003$): Hyb followed by Exo gave greater improvement in mRNA detection sensitivity than Exo followed by Hyb (**Fig. 3**). We hypothesize that using exonuclease as the first treatment produced fragmented rRNAs and reduced the number of Hyb rRNA target sites, resulting in lower mRNA enrichment.

Fidelity of mRNA relative abundance

Another metric by which we assessed the rRNA removal methods was fidelity: how well relative transcript abundance was preserved after a given treatment relative to the untreated control, assuming an accurate representation of relative transcript abundance in the control. We calculated Bray Curtis (BC) similarities between samples by pairwise comparisons of relative transcript abundances of individual mRNAs above the detection threshold in the control. For each experiment, we inferred a dendrogram from the matrix of sample pair similarities using unweighted pair group method with arithmetic mean (UPGMA) (**Fig. 4**). To further evaluate fidelity, we also counted the number of mRNAs exhibiting greater than twofold change between the treatment and corresponding control and presented it as a percentage of the total detected mRNAs (**Fig. 4**).

The rank order of treatments providing better transcript fidelity relative to the control was the same in both experiments: Hyb (≈2 Hyb) > Exo > Hyb + Exo (≈Exo + Hyb) and also was largely consistent for individual organisms (**Supplementary Tables 5** and **6**). Subtractive hybridization alone caused the smallest amount of transcript skewing with only 2–3% of transcripts exhibiting greater than twofold change in expression relative to the control, and one- and two-round Hyb treatments were not statistically different ($P = 0.81$). Exo treatments produced a higher degree of skewing than Hyb treatments alone ($P < 0.05$), probably owing to exonuclease action on partially degraded mRNAs. For example, genes from Clusters of Orthologous Groups of proteins (COG) functional category J were significantly underrepresented in Exo treatments relative to the control (Fisher's exact test $P = 1.65^{-13}$ corrected for multiple-hypothesis testing; Online Methods). This category comprises all of the ribosomal protein families that are known to have high turnover rates and mRNA pools with short half lives that presumably are partially degraded. Exonuclease alone produced drastically different skewing in both communities (30% and 5% of genes exhibiting greater than twofold change in communities 1 and 2 respectively) indicating that mRNA degradation was also dependent on community composition, and that the mRNAs of members of community 1 were on the whole more sensitive to the exonuclease treatment than members of community 2 (**Supplementary Tables 5** and **6**).

Hyb and Exo applied together produced the greatest skew in mRNA fidelity (>50% of detectable transcripts showing greater than twofold change from control) regardless of the order

in which they were applied (**Fig. 4**). Somewhat unexpectedly, a consistent mRNA skew was introduced by the two tested combinations in experiment 2 (BC similarity = 90.6 ± 1.1%) despite both having a lower correlation to Exo treatment alone (BC similarity = 67.2 ± 1.8%) and Hyb treatment alone (BC similarity = 66.6 ± 3.0%). This suggests that the interaction between the two kits has a more complex effect on mRNA fidelity than a simple addition of the effect that each kit has alone and yet was also reproducible and independent of the order in which the kits were applied.

DISCUSSION

Recently, two studies showed that RNA-seq has high technical reproducibility using Illumina sequencing on a bacterial isolate[18] and Genome Sequencer FLX pyrosequencing (Roche) on bacterial isolates and an environmental sample[26]. We, too, found that technical replicates of Illumina sequencing were highly reproducible with the notable exception of one interrun comparison. We discovered a consistent G+C content bias that would have been missed without data from multiple reference organisms with a broad G+C content range. We speculate that this bias is directly linked to Illumina run quality, with lower-quality runs having greater underrepresentation of high G+C content templates. A G+C content bias has previously been reported in Illumina sequences of mouse transcripts[27], lending some support to this hypothesis.

Observed rRNA removal for the synthetic communities was within the 5 to 58% range calculated for published sequence-based metatranscriptomic analyses of natural communities using the same depletion methods[13,14, 17, 21-23]. However, the different removal treatments varied greatly in preserving relative transcript abundance (mRNA fidelity), providing a second key determinant for method selection.

The Ambion kit (Hyb) has been previously used to analyze ocean water[13,14] and soil samples[17]. In the latter case, two rounds of rRNA depletion had been used presumably on the assumption that this would result in improved mRNA enrichment. We did not find that two rounds of Hyb produced a significant improvement in rRNA depletion. This suggests that rRNAs with target sites were efficiently removed in the first round and additional rounds only introduced extra unnecessary handling. Remaining rRNAs either had mismatches to the target sites or were fragments lacking a target site. We observed both instances. First, the rRNAs of the sole archaeal member of community 2, which are incompatible with the capture oligos, had the

lowest depletion levels, which substantially compromised overall mRNA enrichment in community 2, although 26–32% of rRNAs were still removed despite target mismatches. Second, RNA integrity (indicated by RIN) was positively correlated with Hyb removal (**Supplementary Fig. 2**) confirming that more fragmented rRNAs had lower removal efficiencies, consistent with a similar observation for a *Burkholderia cenocepacia* isolate[18].

Unlike with the Hyb treatment, for which potential limitations are well recognized, variability in exonuclease treatment is less well understood. Differential RNA degradation by the processive 5′ to 3′ exonuclease used in the Epicentre kit (Exo) is based on the presence of 5′-monophosphates on mature rRNAs (susceptible) and 5′-triphosphates on intact mRNAs (resistant). However, our results demonstrated that some degradation of mRNAs must be occurring owing to a consistent drop in mRNA fidelity for both experiments. This is most likely due to the exonuclease targeting partially degraded mRNAs, supported by the observation that highly expressed genes with short half-lives were preferentially lost after Exo treatment. Conceivably, Exo treatment could provide a useful snapshot of stable full-length gene expression.

We believe that the findings from these two simple synthetic communities will be applicable to more complex natural communities. We therefore suggest the use of one round of Hyb for standard processing of microbial community RNAs with the caveat that archaeal-dominated communities will likely have low levels of mRNA enrichment. To overcome this limitation, sample-specific subtractive hybridization strategies are a promising complement or alternative approach to generic capture beads[26], particularly as our results predict that hybridization should not compromise transcript fidelity. Combined treatments should be avoided as they severely impact mRNA fidelity, and Exo treatment alone should not be used unless there is a specific biological interest in studying longer-lived transcripts. Workup of total RNAs should include a determination of the RIN, and we suggest only sequencing RNAs with RIN > 7.0 to maximize the potential effectiveness of the Hyb treatment. Finally, comparisons of samples between Illumina runs should be viewed cautiously given the potential for G+C content bias. These recommendations should help guide researchers in making informed decisions on how to enrich mRNA for *de novo* sequencing, particularly for microbial communities in which many more variables may be at play.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Note: Supplementary information is available on the Nature Methods *website.*

AUTHOR CONTRIBUTIONS

S.H., K.S., S.G.T., F.C., E.A.L. and P.H. planned the experiments, S.H., K.S. and S.Y. executed the experiments, S.H., O.W., J.L.F., Z.W., R.S. and P.H. performed the data analysis, S.H. and P.H wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

REFERENCE

1. Sorek, R. & Cossart, P. Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nat. Rev. Genet.* **11**, 9–16 (2010).

2. Neidhardt, F.C. & Umbarger, H.E. Chemical composition of *Escherichia coli*. in *Escherichia coli and Salmonella: Cellular and Molecular Biology* 2nd edn., vol. 1 (eds. Neidhardt, F.C. *et al.*) 13–17 (ASM Press, Washington, D.C., 1996).

3. Karpinets, T.V., Greenwood, D.J., Sams, C.E. & Ammons, J.T. RNA: protein ratio of the unicellular organism as a characteristic of phosphorous and nitrogen stoichiometry and of the cellular requirement of ribosomes for protein synthesis. *BMC Biol.* **4**, 30 (2006).

4. Zhao, J., Hyman, L. & Moore, C. Formation of mRNA 3′ ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol. Mol. Biol. Rev.* **63**, 405–445 (1999).

5. Urich, T. *et al.* Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS ONE* **3**, e2527 (2008).

6. Mao, C., Evans, C., Jensen, R. & Sobral, B. Identification of new genes in *Sinorhizobium meliloti* using the Genome Sequencer FLX system. *BMC Microbiol.* **8**, 72 (2008).

7. Su, C. & Sordillo, L.M. A simple method to enrich mRNA from total prokaryotic RNA. *Mol. Biotechnol.* **10**, 83–85 (1998).

8. Pang, X. *et al.* Bacterial mRNA purification by magnetic capture-hybridization method. *Microbiol. Immunol.* **48**, 91–96 (2004).

9. Frias-Lopez, J. *et al.* Microbial community gene expression in ocean surface waters. *Proc. Natl. Acad. Sci. USA* **105**, 3805–3810 (2008).

10. Shi, Y., Tyson, G.W. & DeLong, E.F. Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature* **459**, 266–269 (2009).

11. Dunman, P.M. *et al.* Transcription profiling-based identification of *Staphylococcus aureus* genes regulated by the agr and/or sarA loci. *J. Bacteriol.* **183**, 7341–7353 (2001).

12. McGrath, K.C. *et al.* Isolation and analysis of mRNA from environmental microbial communities. *J. Microbiol. Methods* **75**, 172–176 (2008).

13. Gilbert, J.A. *et al.* Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS ONE* **3**, e3042 (2008).

14. Gilbert, J.A. *et al.* Potential for phosphonoacetate utilization by marine bacteria in temperate coastal waters. *Environ. Microbiol.* **11**, 111–125 (2009).

15. He, S. *et al.* Metatranscriptomic array analysis of 'Candidatus Accumulibacter phosphatis'-enriched enhanced biological phosphorus removal sludge. *Environ. Microbiol.* **12**, 1205–1217 (2010).

16. Hu, Z., Zhang, A.X., Storz, G., Gottesman, S. & Leppla, S.H. An antibody-based microarray assay for small RNA detection. *Nucleic Acids Res.* **34**, e52 (2006).

17. Shrestha, P.M., Kube, M., Reinhardt, R. & Liesack, W. Transcriptional activity of paddy soil bacterial communities. *Environ. Microbiol.* **11**, 960–970 (2009).

18. Yoder-Himes, D.R. *et al.* Mapping the *Burkholderia cenocepacia* niche response via high-throughput sequencing. *Proc. Natl. Acad. Sci. USA* **106**, 3976–3981 (2009).

19. Burgmann, H. *et al.* Transcriptional response of *Silicibacter pomeroyi* DSS-3 to dimethylsulfoniopropionate (DMSP). *Environ. Microbiol.* **9**, 2742–2755 (2007).

20. Garbeva, P. & de Boer, W. Inter-specific interactions between carbon-limited soil bacteria affect behavior and gene expression. *Microb. Ecol.* **58**, 36–46 (2009).

21. Hewson, I. *et al.* Microbial community gene expression within colonies of the diazotroph, *Trichodesmium*, from the Southwest Pacific Ocean. *ISME J.* **3**, 1286–1300 (2009).

22. Poretsky, R.S. *et al.* Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. *Environ. Microbiol.* **11**, 1358–1375 (2009).

23. Hewson, I. *et al.* In situ transcriptomic analysis of the globally important keystone N_2-fixing taxon *Crocosphaera watsonii. ISME J.* **3**, 618–631 (2009).

24. Wurtzel, O. *et al.* A single-base resolution map of an archaeal transcriptome. *Genome Res.* **20**, 133–141 (2010).

25. Schroeder, A. *et al.* The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol. Biol.* **7**, 3 (2006).

26. Stewart, F.J., Ottesen, E.A. & DeLong, E.F. Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. *ISME J.* **4**, 896–907 (2010).

27. Rosenkranz, R., Borodina, T., Lehrach, H. & Himmelbauer, H. Characterizing the mouse ES cell transcriptome with Illumina sequencing. *Genomics* **92**, 187–194 (2008).

FIGURES
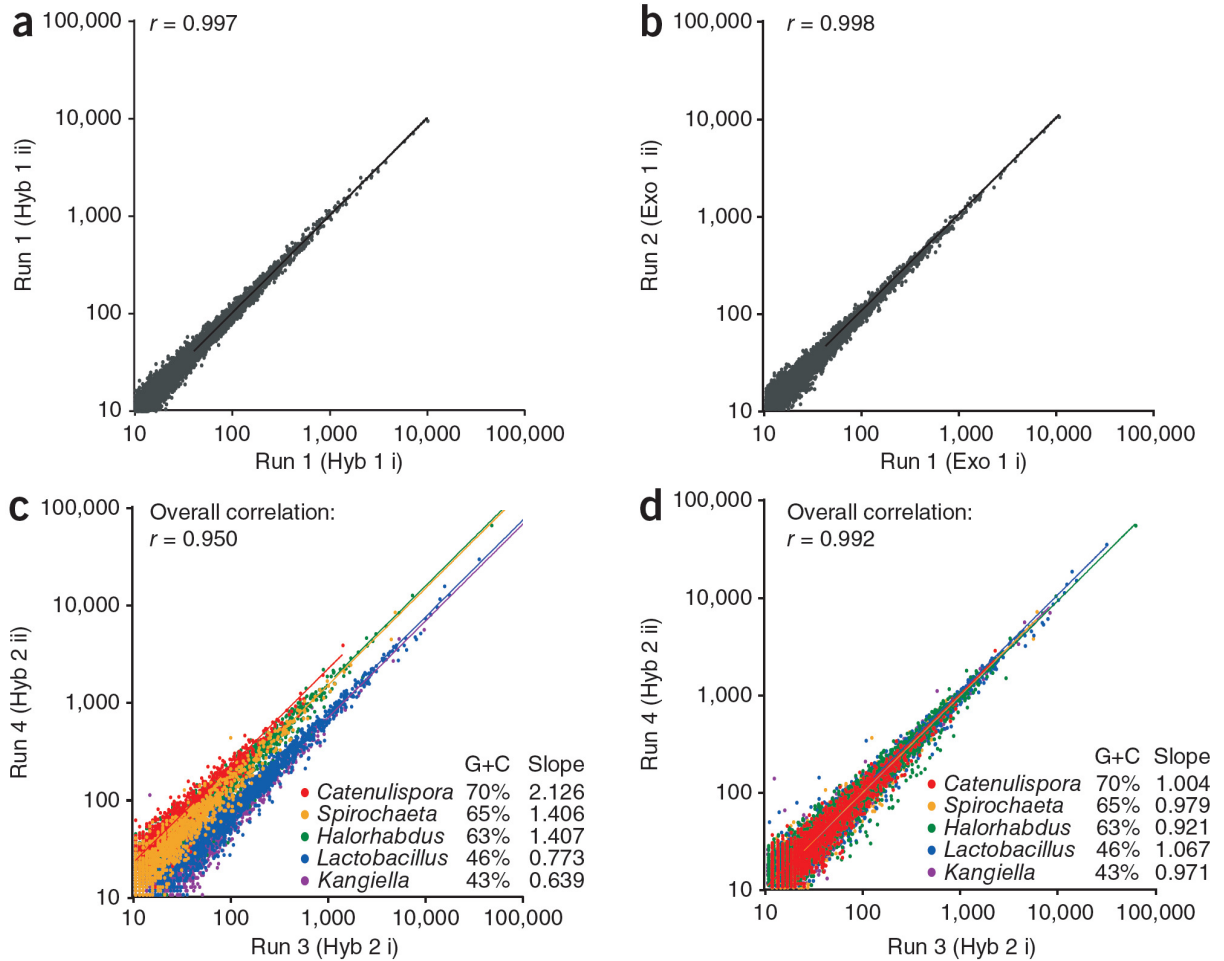
Figure 1 | Technical reproducibility. (**a–d**) In the correlation plots, each point indicates the abundance of an individual mRNA transcript in two technical replicates. Analysis of technical replicates (i and ii) of Hyb in Illumina run 1 (**a**), Exo between runs 1 and 2 (**b**), Hyb between runs 3 and 4 (**c**; color-coded by source organisms), and Hyb between runs 3 and 4 after normalizing for G+C content by organism (**d**). Pearson's product moment correlation coefficient, *r,* for all data points regardless of source organism is shown. Slopes from linear regression of data points for each organism are indicated in **c** and **d**.
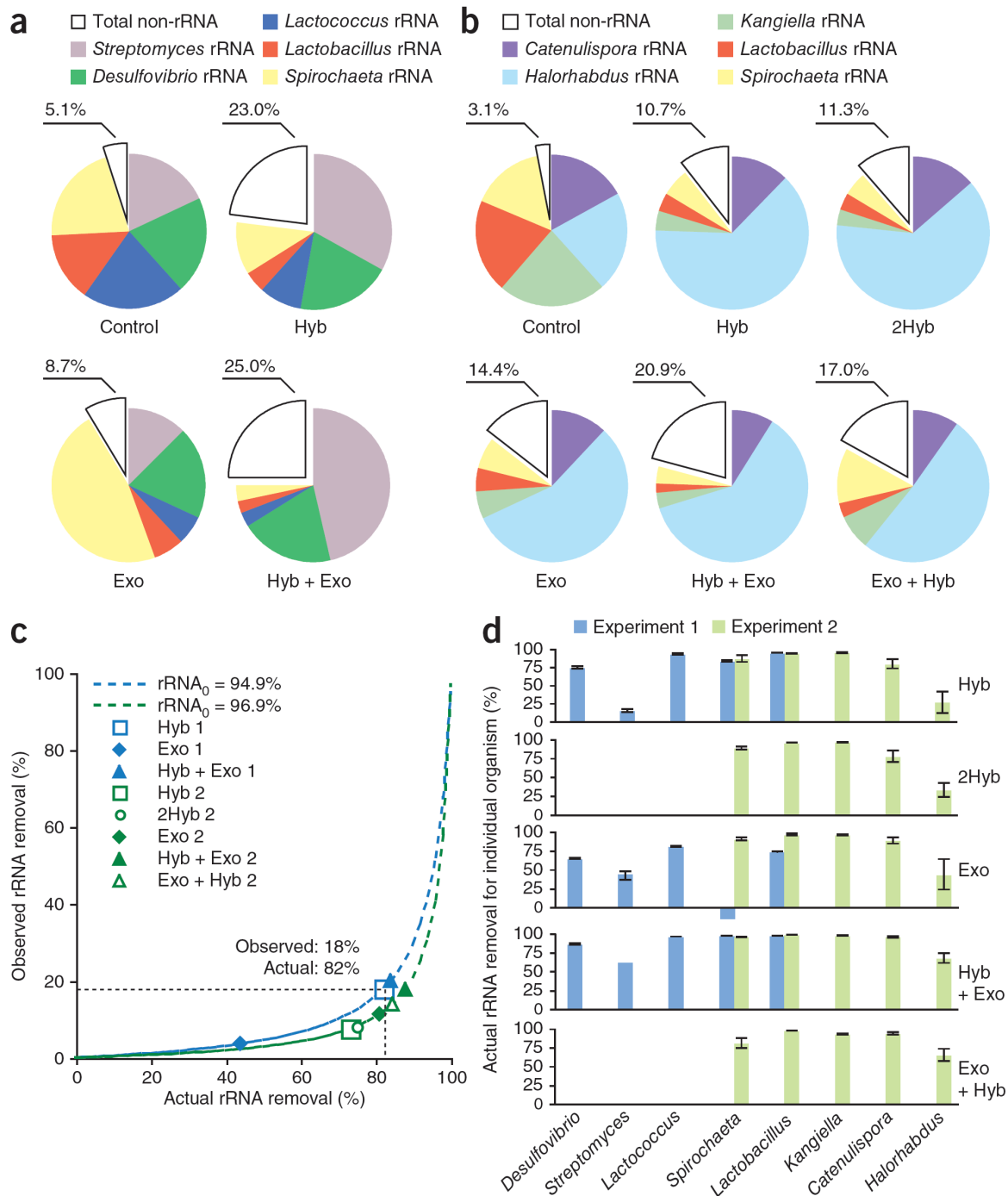
Figure 2 | Effectiveness of bulk rRNA removal. (**a,b**) Distribution of reads between rRNAs (divided by community member) and total non-rRNAs for each treatment in experiments 1 (**a**) and 2 (**b**). (**c**) Observed and actual rRNA percentage removals for the three and five treatments in experiments 1 and 2. Dashed lines are simulations of observed and actual rRNA percentage removals, when starting rRNA (in controls, $rRNA_0$) accounted for 94.9% (community 1) and 96.9% (community 2) of total RNA. (**d**) Actual rRNA percentage removal for each organism. Error bars, s.d. There was no net rRNA removal for *Spirochaeta* by Exo in experiment 1, indicated by an arbitrary negative value.
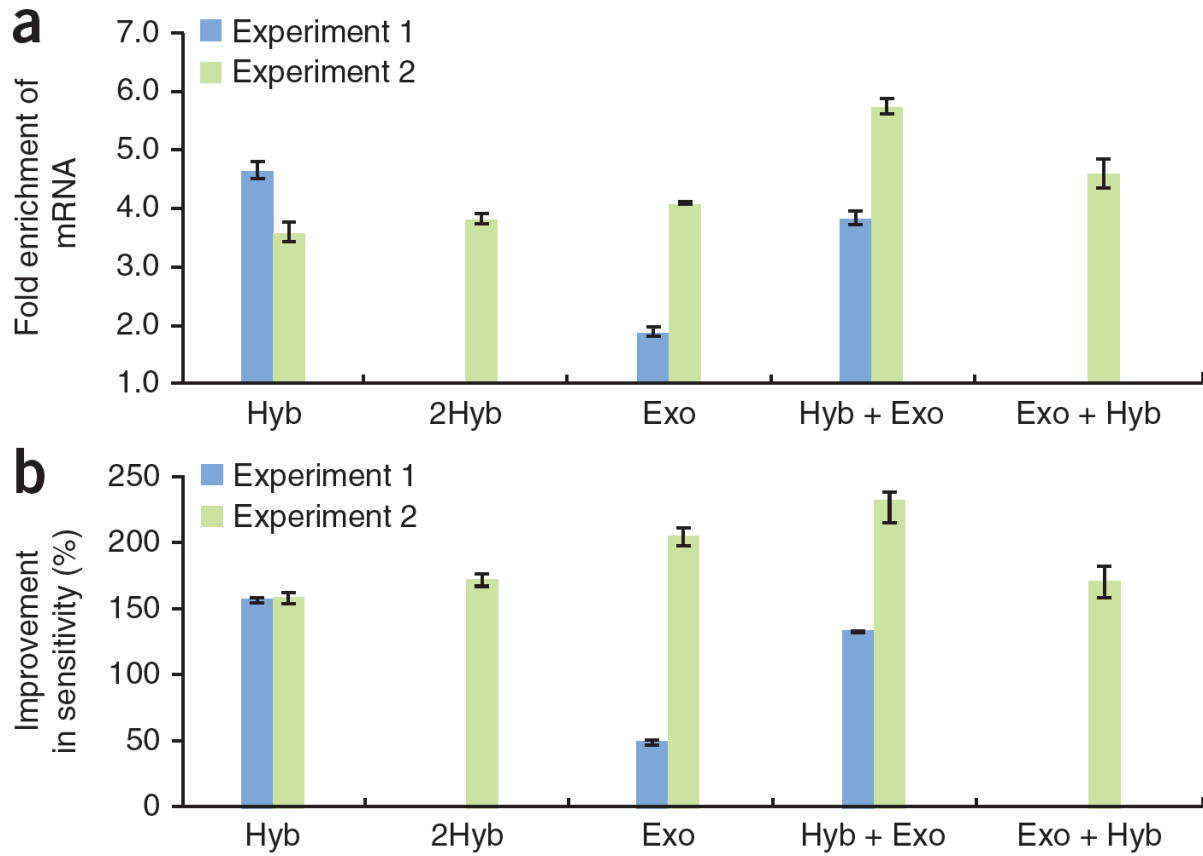
Figure 3 | Enrichment of mRNA in the synthetic communities. (**a**) Fold enrichment of total mRNA abundance. (**b**) Percentage improvement in mRNA detection sensitivity.
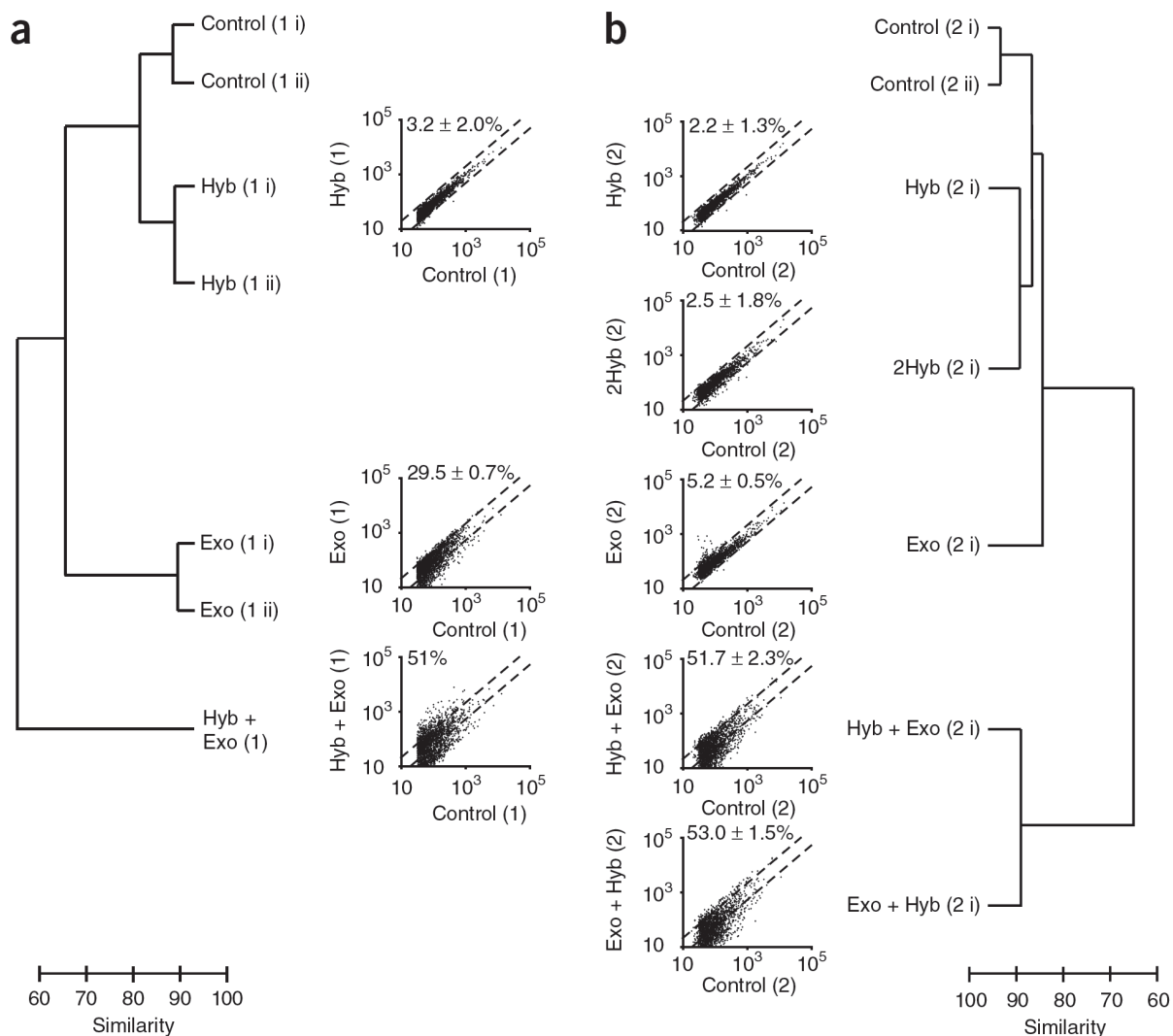
Figure 4 | Fidelity of mRNA relative abundance. (**a**,**b**) Analysis of all seven samples in experiment 1 (**a**) and the seven samples from run 3 in experiment 2 (**b**). Bray Curtis similarities between samples are indicated by a dendrogram showing increasing loss of mRNA fidelity with distance from controls. Increasing loss of fidelity between treatments (y axes) and corresponding controls (x axes) is also visually shown using scatter plots. The average percentage and s.d. of mRNAs in treatments exhibiting greater than twofold difference from respective controls (indicated by diagonal dashed lines) is shown in each scatter plot.

TABLES

Table 1 | Details of microbial isolates used in the two synthetic communities

Organism	Genome size (Mbp)	G+C content (%)	Phylum	Match Hyb target sites[a]	23S/16S[b]	RIN[b]	Community
Desulfovibrio vulgaris strain Hildenborough	3.7	63	Proteobacteria	Yes	0.9	7.4	1
Streptomyces sp. strain LCC	8–10[c]	71	Actinobacteria	Yes	1.4	5.5	1
Lactococcus lactis subspecies *lactis* IL1403	2.53	35	Firmicutes	Yes	1.6	9.8	1
Spirochaeta aurantia subspecies *aurantia* M1	4.3	65	Spirochaeta	Yes	2.1	9.8	1 and 2
Lactobacillus brevis ATCC[d] 367	2.3	46	Firmicutes	Yes	1.4	9.9	1 and 2
Kangiella koreensis DSM[d] 16069	2.9	43	Proteobacteria	Yes	1.4	10	2
Catenulispora acidiphila DSM[d] 44928	10.5	70	Actinobacteria	Yes	1.3	8.6	2
Halorhabdus utahensis DSM[d] 12940	3.1	63	Euryarchaeota	No	1.9	10	2

[a]Information about 16S and 23S rRNAs matching to the capture oligos in the MICROBExpress mRNA Enrichment Kit was obtained from Ambion. [b]The ratio of 23S/16S rRNAs and RNA integrity number (RIN) were determined using the Agilent 2100 Bioanalyzer. RIN values range from 1 (most degraded) to 10 (most intact). [c]Draft assembly; genome size was estimated to be 8–10 mega–base pairs (Mbp), which is in the range of the size of genomes of this organism's close relatives, such as *S. coelicolor* A3(2) (9,054,847 bp), *S. griseus* subspecies *griseus* NITE Biological Resource Center (NBRC) 13350 (8,545,929 bp) and *S. avermitilis* MA-468 (9,119,895 bp). [d]ATCC: American Type Culture Collection. DSM: Deutsche Sammlung von Mikrorganismen.

Table 2 | Ribosomal RNA depletion methods tested and run (flowcell) layout

	Experiment 1 Community 1		Experiment 2 Community 2		Published (meta)transcriptomic studies using rRNA depletion methods listed in run 4[a]
	Run 1	Run 2	Run 3	Run 4	
Lane 1	Control (i)[b]		Control (i)	Control (i)	
Lane 2	Control (ii)		Control (ii)	Control (ii)	
Lane 3	Hyb[c] (i)		Hyb (i)	Hyb (ii)	Ocean water[13,14], bioreactor[15] and bacterial isolates[6, 16, 18]
Lane 4	Hyb (ii)		2Hyb (i)	2Hyb (ii)	Soil[17]
Lane 5	Exo (i)		Exo (i)	Exo (ii)[d]	Bacterial isolates[19, 20]
Lane 6	Exo (ii)[d]		Hyb+Exo (i)	Hyb+Exo (ii)	
Lane 7	Hyb+Exo	Exo (ii)	Exo+Hyb (i)	Exo+Hyb (ii)	Ocean water[21–23]
Cluster pass rate[e]	84 ± 3%	88%	53 ± 6%	71 ± 3%	

[a]For completeness, there are two other metatranscriptomic studies not listed in the table that used addition of poly(A) tails which preferentially ligate to mRNAs providing the basis for mRNA enrichment[9,10].
[b]l(i) and (ii) indicate technical replicate. [c]Hyb, Subtractive hybridization (MICROBExpress Bacterial mRNA Enrichment Kit, Ambion); 2Hyb = two rounds of Hyb, Exo = Exonuclease digestion (mRNA-ONLY Prokaryotic mRNA Isolation Kit, Epicentre); Hyb+Exo = Hyb followed by Exo; Exo+Hyb = Exo followed by Hyb. [d]Lanes were excluded from analysis owing to technical failure. [e]Cluster pass rate is defined as the number of clusters that passed the Illumina chastity filter (³0.6) divided by total number of identified clusters. We use average cluster pass rate from lanes included in each run to indicate run quality.

ONLINE METHODS

Microbial isolates. One archaeal and seven bacterial isolates were used in this study (**Table 1**). *Lactococcus lactis* was grown in M17 medium (Difco) supplemented with 0.5% glucose, *Lactobacillus brevis* was grown in ATCC (American Type Culture Collection) medium 416, and both were incubated overnight in 50 ml shake flasks at 30 °C. *Streptomyces* sp. str. LCC was initially maintained on oatmeal agar (DSMZ (German Collection of Microorganisms and Cell Cultures) medium 84) at 30 °C, and axenic colonies from plates were transferred into yeast malt extract broth and grown at 30 °C. Other microbial isolates were grown using DSMZ recommended media and conditions (http://www.dsmz.de/). All isolates had complete reference genomes with the exception of *Streptomyces* sp. str. LCC for which a draft genome comprising 5,199 contigs is available (http://img.jgi.doe.gov/er).

RNA extraction and synthetic metatranscriptome construction. Total RNA was extracted independently from each microbial isolate using the RiboPure Bacteria Kit (Ambion) with DNase I digestion, following the manufacturer's instructions. Note that 5S rRNAs and tRNAs are largely excluded by this kit based on their size (<200 base paris (bp)). A modification was made when extracting RNA from *Halorhabdus* (a halophile that grows only in high salt) by applying twice the recommended volume of RNAWIZ buffer to cope with the high salt residues in the cell pellet. The quantity and quality of RNA was evaluated using Agilent 2100 Bioanalyzer electropherograms. Two synthetic metatranscriptomes were then constructed by pooling equimolar amounts of total RNAs from the two sets of five organisms, as described in **Table 1**. The pooled total RNA was aliquoted to form the starting material for the control (total RNA without rRNA removal) and various treatments (rRNA removal procedures), as well as their corresponding technical replicates (**Table 2**).

rRNA removal treatments. The MICROBExpress Bacterial mRNA Enrichment Kit (Ambion) and mRNA-ONLY Prokaryotic mRNA Isolation Kit (Epicentre), were applied in various combinations to synthetic communities 1 and 2 (**Table 2**) according to manufacturers' instructions. For each treatment, a technical replicate (following RNA pooling) was included, except for Hyb + Exo in experiment 1. For both experiments, ~8.5 μg of pooled total RNA was used as the starting material for each treatment evaluated. rRNA removal following each treatment was visually assessed by Agilent 2100 Bioanalyzer RNA electropherograms (**Supplementary Fig. 1**).

RNA fragmentation and cDNA synthesis. The pooled total RNAs from controls (100 ng) and mRNA-enriched RNAs by treatments (~5–60 ng) were chemically fragmented to ~150–250 bp by incubation with 1× fragmentation solution (Ambion) at 70 °C for 5 min. Double-stranded cDNAs were synthesized from fragmented RNAs using SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen), with the first strand primed by random hexamers and the second strand synthesized by nick translation.

Illumina library construction and sequencing. The Illumina genomic sample prep kit (Illumina) was used for Illumina library preparation, following manufacturer's instructions. Briefly, fragmented cDNA was end-repaired to generate blunt ends. An adenine was added to the 3′ end of the blunt-ended cDNA fragments, allowing the subsequent ligation of Illumina adaptors. A 200–500 bp size selection of ligation products was conducted by gel electrophoresis purification using the MiniElute Gel Extraction Kit (Qiagen) according to manufacturer's instructions, except that the gel was dissolved at 37 °C, instead of the recommended 55 °C, to minimize denaturation of (A+T)-rich fragments[28]. A 12-cycle PCR with adaptor primers was applied to comparable amounts of adaptor-modified cDNA fragments. All sequencing was performed using the Illumina Genome Analyzer II platform to generate single-end 76 bp reads with the exception of sample Exo (1-ii), which was paired-end data. Technical replicates were distributed within and between four Illumina runs (flowcells) (**Table 2**).

Alignment and relative transcript abundance evaluation. Quality-filtered reads were aligned to the reference genomes using blastn with a word-size of 11. The alignment required a minimal length of 42 bp allowing up to six mismatches. The average length of the alignment for samples in experiment 2 was 72.5 ± 3.0 bp. This indicates that the vast majority of sequences were fully mapped to their genome of origin. The number of reads mapped to each gene (including mRNA, rRNA, tRNA and intergenic regions; **Supplementary Table 2**) was counted and normalized by the library size to account for different numbers of reads generated for each sample. This normalization was performed by dividing the raw read count per transcript by the number of total mapped reads in each Illumina lane and then multiplying by the average number of total mapped reads in the controls in each experiment. An mRNA was defined as 'detected' if the library size-normalized read count (relative transcript abundance) for a given mRNA was ≥30. This detection threshold was selected as it was found to be the minimum number of reads needed to pass a Chi-

square statistical test for differential expression ($P < 0.05$) based on empirical simulations (corrected for multiple hypothesis testing; data not shown).

G+C content–associated interrun variation normalization. Technical reproducibility was evaluated using the Pearson's product-moment correlation coefficient (r) of relative transcript abundances of individual mRNAs between two technical replicates. As interrun variation associated with G+C content was observed, we tried two simple normalization strategies. The first strategy (by organism) was performed by normalizing the relative transcript abundances of individual mRNAs by the total mRNA read count from its source organism in each run. As the taxonomic classification of mRNAs in real metatranscriptomic datasets may not be available, we also tried a second strategy (by gene), which normalized the transcript abundance by the G+C content of individual genes, described below.

For each gene, $N_A' = N_A \times (\%G+C)^a$, in which N_A is raw read count and N_A' is G+C content–normalized read count for a given gene in run A; $\%G+C$ is G+C content of that gene; and a is the correction factor between runs A and B, determined by the best fit that minimizes the sum of difference between N_A' and N_B (read count from its corresponding technical replicate in run B) for all mRNAs with technical replicates in runs A and B.

rRNA percent removal calculation. The observed and actual rRNA percentage removal was calculated as:

$$\text{Observed rRNA percentage removal} = r_0 - r_1$$
$$\text{Actual rRNA percentage removal} = 1 - (r_1 / r_0) \times ((1 - r_0) / (1 - r_1))$$

r_0 and r_1 are the percentages of rRNA reads among all mapped reads in the control and treatment, respectively.

To accurately assess the rRNA depletion efficiency (mRNA fold enrichment, detection sensitivity improvement and fidelity, as described below), we restricted comparisons of treatment-control pairs to the same run in experiment 2, because of the observed G+C content bias between runs 3 and 4. The average and s.d. was derived from all available treatment-control comparisons in experiment 1 and all available intrarun comparisons from runs 3 and 4 in experiment 2.

mRNA enrichment and detection sensitivity increase calculation. The fold enrichment of total mRNA abundance and the percentage improvement in mRNA detection sensitivity were calculated as:

$$\text{mRNA fold enrichment} = m_1 / m_0$$

$$\text{Percentage improvement in mRNA detection sensitivity} = ((M_1 - M_0) / M_0) \times 100$$

m_0 and m_1 are the percentage of total mRNA reads among all mapped reads in the control and treatment respectively; and M_0 and M_1 are the number of mRNAs above the detection threshold (30 reads per gene after normalization by library size) in the control and treatment, respectively.

mRNA fidelity evaluation. mRNA fidelity is a measure of the preservation of relative transcript abundance between a treatment and corresponding control. As low-abundance transcripts may become detectable after an mRNA enrichment treatment, these transcripts may appear to be skewed because of statistically insignificant read counts in the control. Therefore, we used a filter to only include mRNAs that are above the 30-read detection threshold in a majority of the controls to assess mRNA fidelity. Individual mRNA read counts were further normalized by total read counts from all mRNAs included in the analysis. BC similarities between samples were calculated using Primer 6 (Version 6.1.9) (PRIMER-E Ltd) to generate similarity matrices based on pairwise comparisons of relative transcript abundance for individual mRNAs included. A BC similarity-based hierarchical clustering of samples was performed using the UPGMA clustering algorithm. In addition, normalized read counts for individual mRNAs detected in the control were plotted against their normalized read counts in each treatment. The number of mRNAs exhibiting greater than twofold difference between a treatment and corresponding control was counted and represented as a proportion of total detected mRNAs in the corresponding control.

Underrepresented mRNA functional category identification. A gene expression index (GEI) for all mRNAs was calculated by dividing the number of reads in the mRNA per kilobase, followed by normalization according to the total number of mRNA reads in the sample. Differential expression between samples was determined by performing a chi-square test of the GEI of all mRNAs in different conditions, followed by a Bonferonni correction multiple-hypothesis testing with a threshold of $P < 0.05$. Available functional classifications of genes was downloaded from US National Center for Biotechnology Information (NCBI) records (http://www.ncbi.nlm.nih.gov/COG/), and enrichment for underrepresented mRNAs in different

COG functional categories was calculated with a single-sided Fisher's exact test followed by a Bonferonni correction for multiple-hypothesis testing.

Other statistical analyses. Student's *t*-tests were performed to obtain statistical significance (*P* value) when comparing two variables. The Pearson's product-moment correlation coefficient (*r*) was used to determine the correlation between actual rRNA percent removal and RIN.

REFERENCE

28. Quail, M.A. *et al.* A large genome center's improvements to the Illumina sequencing system. *Nat. Methods* **5**, 1005–1010 (2008).

SUPPLEMENTARY INFORMATIOIN

Supplementary Figure 1	Agilent 2100 Bioanalyzer electropherograms of RNA
Supplementary Figure 2	Normalized read coverage plots of 16S and 23S rRNAs
Supplementary Figure 3	Correlation between actual rRNA percent removal and RIN.
Supplementary Table 1	Sample description, sequencing and read mapping statistics
Supplementary Table 2	Read distribution (%) among different RNA species
Supplementary Table 3	Read distribution (%) among different RNA species for each organism in Experiment 1
Supplementary Table 4	Read distribution (%) among different RNA species for each organism in Experiment 2
Supplementary Table 5	Percentage of mRNAs exhibiting >2-fold change in Experiment 1
Supplementary Table 6	Percentage of mRNAs exhibiting >2-fold change in Experiment 2
Supplementary Note 1	RNA electropherograms
Supplementary Note 2	Illumina sequencing and read quality
Supplementary Note 3	Normalization for G+C content-associated interrun variation
Supplementary Note 4	Difference between actual and observed rRNA removal

SUPPLEMENTARY FIGURES
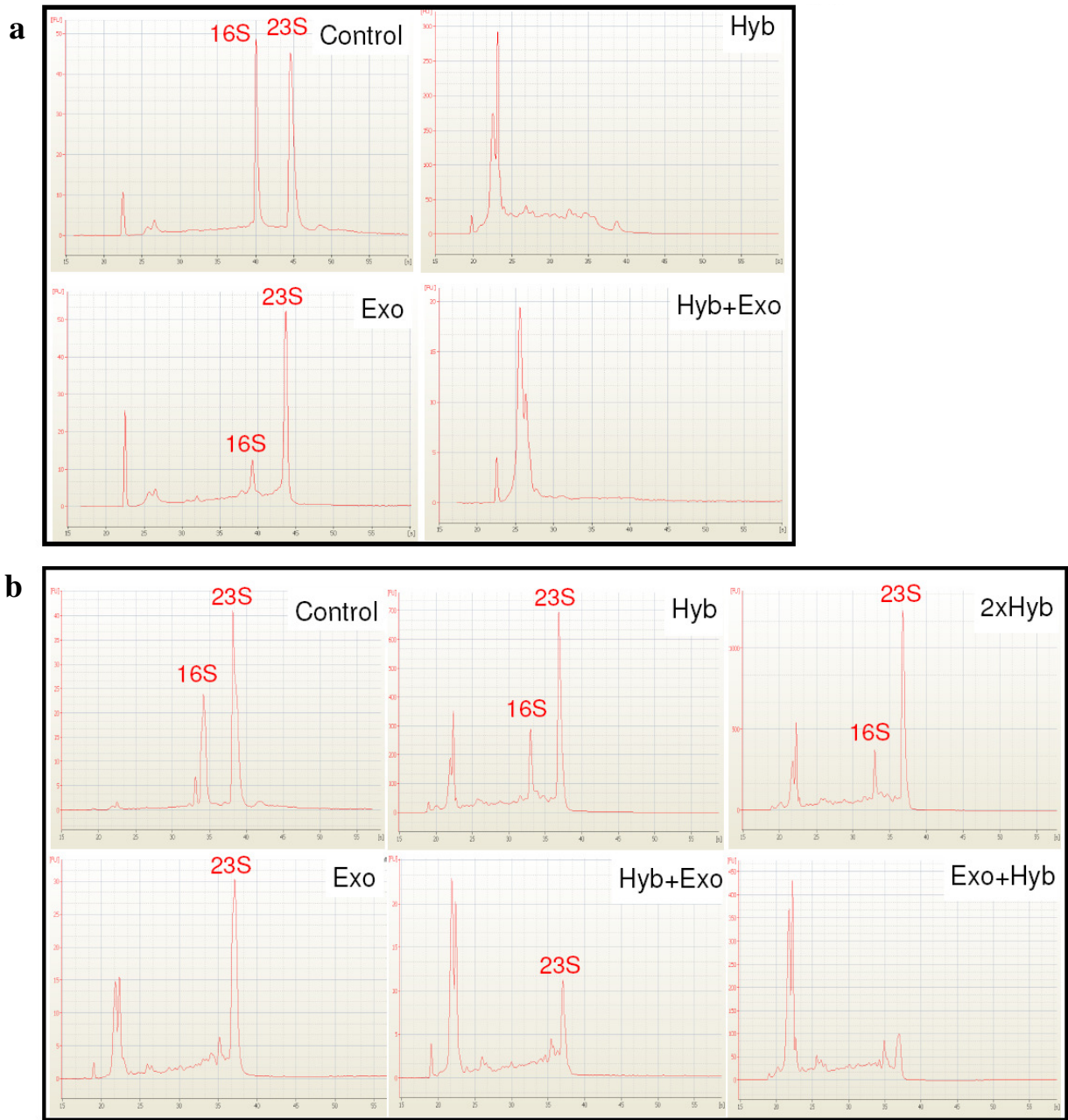
Supplementary Figure 1. Agilent 2100 Bioanalyzer electropherograms of RNA. (**a**) and (**b**) are electropherograms in Experiments 1 and 2 respectively before and after different rRNA removal procedures.

Desulfovibrio vulgaris normalized coverage plot for 16S rRNA

Highest exonuclease processivity

Likely Hyb target sites
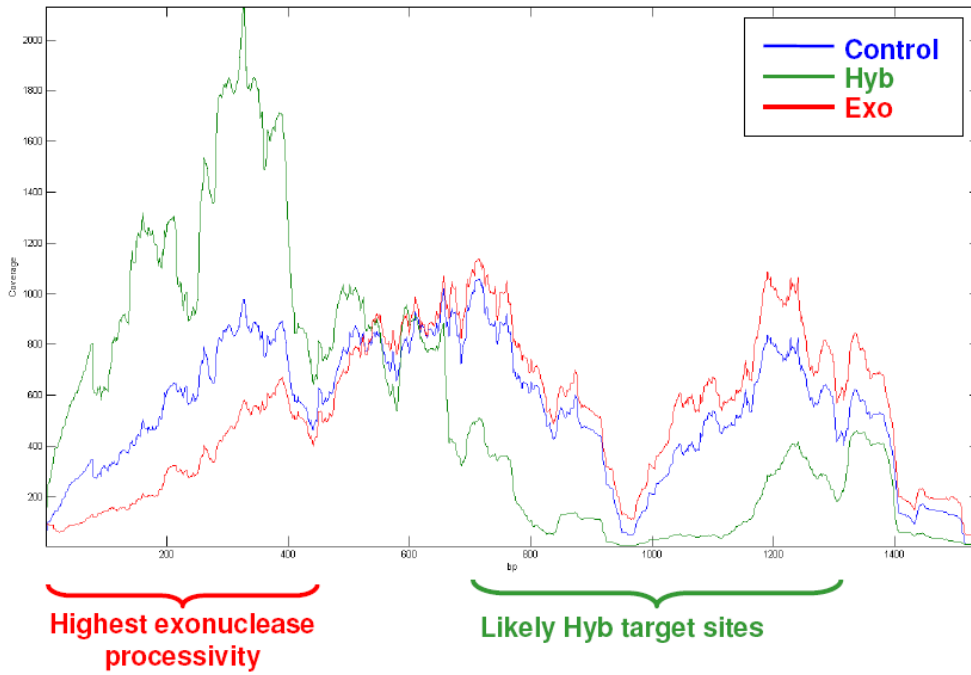
Desulfovibrio vulgaris normalized coverage plot for 23S rRNA

Highest exonuclease processivity

Likely Hyb target sites
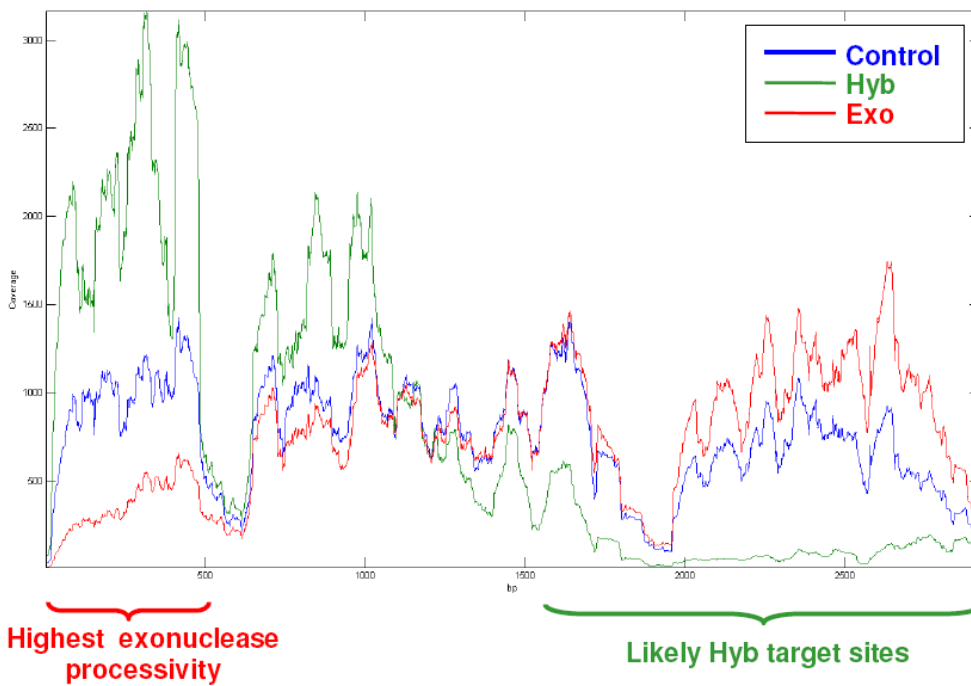
Supplementary Figure 2. Normalized read coverage plots of 16S and 23S rRNAs from *Desulfovibrio* showing inferred zones of highest exonuclease activity and Hyb target sites.
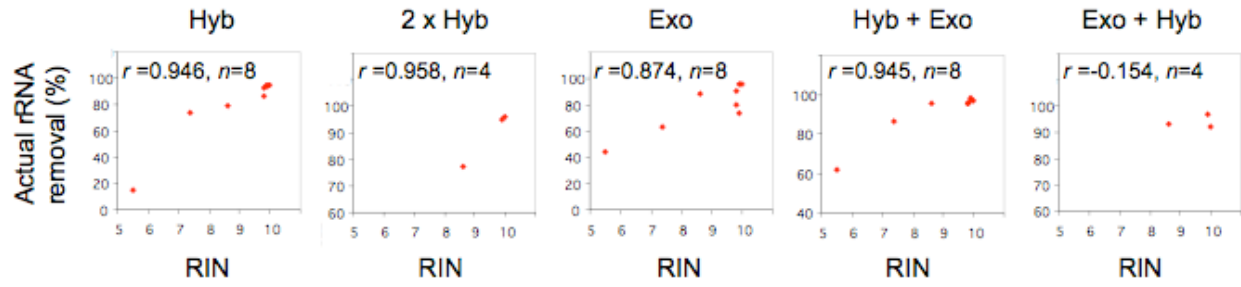
Supplementary Figure 3. Correlation between actual rRNA percent removal and RIN (RNA integrity number). Pearson's product-moment correlation coefficient (*r*) is indicated for each treatment, *n* is number of comparisons included. Note that *Halorhabdus* was not included as its rRNAs are not targeted by the Hyb kit.

SUPPLEMENTARY TABLES

Supplementary Table 1. Sample description, sequencing and read mapping statistics

Description	Experiment, Replicate	Run ID	Total reads	Mapped reads	% mapped
Control	1, i	1	10,735,075	10,698,419	99.66%
	1, ii	1	10,124,865	10,089,716	99.65%
Hyb	1, i	1	10,280,467	10,194,078	99.16%
	1, ii	1	10,779,165	10,694,839	99.22%
Exo	1, i	1	12,195,994	12,159,888	99.70%
	1, ii	2	32,941,724	32,584,366	98.92%
Hyb+Exo	1	1	12,060,728	11,938,847	98.99%
Control	2, i	3	14,779,022	14,726,899	99.65%
		4	16,091,913	16,018,756	99.55%
	2, ii	3	15,580,348	15,508,317	99.54%
		4	17,483,263	17,387,147	99.45%
Hyb	2, i	3	14,764,917	14,724,160	99.72%
	2, ii	4	16,795,730	16,717,188	99.53%
2xHyb	2, i	3	14,095,012	13,974,915	99.15%
	2, ii	4	16,925,206	16,880,270	99.73%
Exo	2, i	3	17,455,434	15,251,159	87.37% [a]
Hyb+Exo	2, i	3	15,227,972	15,174,780	99.65%
	2, ii	4	17,240,660	17,173,333	99.61%
Exo+Hyb	2, i	3	14,172,493	14,111,503	99.57%
	2, ii	4	16,651,811	16,584,746	99.60%

[a] the low read mapping result for this sample may have been due to lower than average fragment length in the library preparation.

Supplementary Table 2. Read distribution (%) among different RNA species.

Sample[a]	Run ID	mRNA	rRNA	tRNA	Misc RNA	Intergenic
Control (1, i)	1	3.95	95.05	0.04	0.45	0.51
Control (1, ii)	1	4.13	94.79	0.07	0.48	0.53
Hyb (1, i)	1	18.40	77.35	0.07	2.06	2.11
Hyb (1, ii)	1	18.99	76.58	0.06	2.13	2.24
Exo (1, i)	1	7.32	91.45	0.03	0.36	0.84
Exo (1, ii)	2	7.80	91.14	0.03	0.22	0.81
Hyb+Exo (1)	1	15.41	74.97	0.16	3.97	5.50
Control (2, i)	3	2.33	96.92	0.01	0.08	0.64
Control (2, ii)	3	2.36	96.90	0.02	0.08	0.64
Control (2, i)	4	2.37	96.90	0.02	0.09	0.62
Control (2, ii)	4	2.37	96.90	0.02	0.08	0.62
Hyb (2, i)	3	8.72	88.91	0.03	0.25	2.09
Hyb (2, ii)	4	8.14	89.76	0.07	0.22	1.81
2xHyb (2, i)	3	8.74	88.89	0.06	0.20	2.10
2xHyb (2, ii)	4	9.19	88.53	0.04	0.23	2.01
Exo (2, i)	3	9.53	85.65	0.76	0.51	3.54
Hyb+Exo (2, i)	3	13.15	79.56	0.09	1.41	5.79
Hyb+Exo (2, ii)	4	13.82	78.63	0.53	1.31	5.71
Exo+Hyb (2, i)	3	10.24	83.52	0.05	1.03	5.17
Exo+Hyb (2, ii)	4	11.36	82.43	0.14	1.00	5.08

[a](1) and (2) indicate experiments 1 and 2 respectively, and (i) and (ii) indicate technical replicates.

Supplementary Table 3. Read distribution (%) among different RNA species for each organism in Experiment 1

Organism	RNA species	Control (i) Run 1	Control (ii) Run 1	Hyb (i) Run 1	Hyb (ii) Run 1	Exo (i) Run 1	Exo (ii) Run 2	Hyb+Exo Run 1
Lactococcus	mRNA	6.97	6.76	43.36	44.9	28.37	29.91	38.61
	rRNA	91.05	91.09	44.6	42.6	66.7	66.34	30.06
	tRNA	0.08	0.15	0.15	0.12	0.12	0.08	0.43
	misc_RNA	1.33	1.43	8.1	8.1	2.24	1.15	18.54
	Intergenic	0.57	0.57	3.81	4.28	2.57	2.52	12.37
	Total	100	100	100	100	100	100	100
Desulfovibrio	mRNA	4.41	4.68	15.94	15.04	11.85	12.35	21.37
	rRNA	94.63	94.27	80.98	82.08	86.28	85.9	70.48
	tRNA	0.02	0.03	0.06	0.05	0.05	0.05	0.27
	misc_RNA	0.34	0.37	0.94	0.87	0.48	0.35	2.13
	Intergenic	0.6	0.65	2.09	1.97	1.35	1.35	5.75
	Total	100	100	100	100	100	100	100
Streptomyces	non-rRNA	1.20	1.96	1.43	1.39	2.04	2.23	3.09
	rRNA	98.80	98.04	98.57	98.61	97.96	97.77	96.91
	Total	100	100	100	100	100	100	100
Lactobacillus	mRNA	4.11	4.13	41.04	42.16	13.85	14.52	25.15
	rRNA	95.02	94.91	51.53	50.37	83.56	83.27	37.6
	tRNA	0.05	0.08	0.24	0.22	0.06	0.04	0.27
	misc_RNA	0.29	0.31	1.98	1.73	0.52	0.29	20.55
	Intergenic	0.53	0.57	5.21	5.53	2.01	1.88	16.42
	Total	100	100	100	100	100	100	100
Spirochaeta	mRNA	2.27	2.45	13.72	12.94	1.63	1.65	30.16
	rRNA	96.89	96.64	83.1	83.97	98.04	98.09	48.36
	tRNA	0.03	0.06	0.07	0.06	0.01	0.01	0.25
	misc_RNA	0.08	0.09	0.4	0.43	0.03	0.02	1.56
	Intergenic	0.73	0.76	2.71	2.59	0.29	0.23	19.67
	Total	100	100	100	100	100	100	100

Supplementary Table 4. Read distribution (%) among different RNA species for each organism in Experiment 2

Organism	RNA species	Control (i) Run 3	Control (ii) Run 3	Control (i) Run 4	Control (ii) Run 4	Hyb (i) Run 3	Hyb (ii) Run 4	2xHyb (i) Run 3	2xHyb (ii) Run 4	Exo (i) Run 3	Hyb+Exo (i) Run 3	Hyb+Exo (ii) Run 4	Exo+Hyb (i) Run 3	Exo+hyb (ii) Run 4
Catenulispora	mRNA	0.82	0.85	1.20	1.19	3.74	6.42	3.14	5.85	5.39	15.82	21.19	11.38	14.61
	rRNA	98.72	98.69	98.26	98.29	94.30	91.21	94.88	91.90	90.51	76.50	70.36	83.09	79.20
	tRNA	0.01	0.01	0.01	0.01	0.02	0.04	0.03	0.02	0.67	0.08	0.47	0.05	0.12
	misc_RNA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Intergenic	0.46	0.45	0.52	0.50	1.94	2.32	1.94	2.23	3.43	7.60	7.98	5.48	6.07
	Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Halorhabdus	mRNA	2.37	2.27	2.83	2.67	3.05	4.01	3.53	3.95	3.37	6.13	7.84	5.32	7.44
	rRNA	97.16	97.28	96.66	96.84	96.39	95.37	95.82	95.33	95.85	92.15	90.16	92.92	90.47
	tRNA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.01
	misc_RNA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Intergenic	0.47	0.45	0.51	0.49	0.56	0.61	0.65	0.71	0.78	1.72	2.00	1.76	2.08
	Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Kangiella	mRNA	1.71	1.74	1.48	1.53	23.32	18.91	23.66	26.87	17.07	15.64	11.57	7.49	6.63
	rRNA	97.63	97.63	97.89	97.86	67.52	71.87	66.10	62.99	68.97	53.82	58.89	77.06	76.80
	tRNA	0.01	0.01	0.01	0.01	0.05	0.16	0.08	0.08	1.01	0.15	1.10	0.06	0.22
	misc_RNA	0.02	0.02	0.03	0.02	0.07	0.12	0.07	0.11	0.34	2.10	2.16	0.83	0.97
	Intergenic	0.63	0.61	0.59	0.57	9.04	8.94	10.09	9.95	12.61	28.29	26.28	14.56	15.37
	Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Lactobacillus	mRNA	5.07	5.22	4.63	4.81	43.64	40.32	45.83	49.10	39.96	51.37	46.35	44.03	42.38
	rRNA	93.69	93.51	94.18	93.98	46.83	50.68	44.81	41.06	43.03	20.69	22.98	31.73	32.62
	tRNA	0.03	0.04	0.05	0.06	0.23	0.53	0.38	0.32	3.03	0.37	2.68	0.20	0.69
	misc_RNA	0.32	0.30	0.35	0.32	2.21	2.52	1.92	2.45	3.43	11.65	12.73	8.26	9.78
	Intergenic	0.89	0.92	0.79	0.83	7.09	5.96	7.06	7.07	10.55	15.92	15.26	15.78	14.53
	Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Spirochaeta	mRNA	1.00	0.98	1.46	1.34	7.62	12.41	10.25	11.89	8.16	20.03	19.08	4.39	7.44
	rRNA	98.20	98.20	97.70	97.81	89.00	83.37	85.97	83.96	84.09	67.79	68.37	92.19	87.88
	tRNA	0.02	0.03	0.03	0.04	0.09	0.26	0.20	0.14	2.84	0.60	2.86	0.09	0.33
	misc_RNA	0.05	0.04	0.06	0.06	0.46	0.56	0.56	0.73	0.94	1.32	1.65	0.36	0.52
	Intergenic	0.75	0.76	0.74	0.75	2.82	3.39	3.03	3.28	3.96	10.26	8.04	2.96	3.83
	Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Supplementary Table 5. Percentage of mRNAs exhibiting >2-fold change in Experiment 1

Treatment	*Desulfovibrio*	*Lactococcus*	*Spirochaeta*	*Lactobacillus*	Community 1
Hyb	0.8 ± 0.2	0.3 ± 0.3	0.7 ± 0.5	0.4 ± 0.5	**3.2 ± 2.0**
Exo	7.4 ± 0.6	40.4 ± 1.5	20.6 ± 1.0	39.3 ± 1.2	**29.5 ± 0.7**
Hyb+Exo	18.7 ± 0	69.9 ± 0	40.4 ± 0	59.8 ± 0	**51.1 ± 0**

Note: *Streptomyces* is not included due to its draft genome containing a large number of partial genes.

Supplementary Table 6. Percentage of mRNAs exhibiting >2-fold change in Experiment 2

Treatment	*Spirochaeta*	*Lactobacillus*	*Kangiella*	*Catenulispora*	*Halorhabdus*	Community 2
Hyb	3.3 ± 2.3	1.4 ± 0.5	0.7 ± 0.3	0.5 ± 0.7	2.1 ± 0.8	**2.2 ± 1.3**
2xHyb	3.2 ± 0.8	2.3 ± 1.8	1.9 ± 1.7	3.0 ± 3.6	2.4 ± 2.0	**2.5 ± 1.8**
Exo	6.0 ± 2.8	10.7 ± 0.4	0.8 ± 0.8	0.4 ± 0.5	1.4 ± 0.5	**5.2 ± 0.5**
Hyb+Exo	37.2 ± 2.8	53.7 ± 1.7	52.8 ± 2.0	2.1 ± 0.6	19.6 ± 1.1	**51.7 ± 2.3**
Exo+Hyb	52.2 ± 0.8	59.4 ± 0.7	59.2 ± 1.8	4.2 ± 1.7	32.8 ± 0.8	**53.0 ± 1.5**

SUPPLEMENTARY NOTES

Supplementary Note 1. RNA electropherograms

Prior to sequencing, rRNA removal by the different treatments was initially assessed using RNA electropherograms (Agilent 2100 Bioanalyzer) (**Supplementary Fig. 1**). Technical replicates of each treatment exhibited very similar profiles (data not shown). For Experiment 1, removal of 16S and 23S rRNAs appeared to be efficient for Hyb and Hyb + Exo as individual rRNA peaks could not be resolved following these treatments. This is consistent with the presence of Hyb target sites for all organisms in Community 1 (**Table 1**). By contrast, 16S and 23S peaks remained in Experiment 2 even after two rounds of Hyb, presumably because the Hyb probes are not compatible with the rRNA of one member of Community 2, the archaeon *Halorhabdus* (**Table 1**), as later confirmed by rRNA removal data. In both experiments, applying Exo alone did not completely eliminate 16S and 23S peaks and there appeared to be preferential elimination of 16S rRNAs, i.e. the remaining 23S peak was higher than the 16S peak (**Supplementary Fig. 1**), probably due to the decreased exonuclease processivity on the longer molecules (**Supplementary Fig. 2**).

Supplementary Note 2. Illumina sequencing and read quality

Illumina sequencing was used in this study because of the availability of reference genomes for mapping short reads and because it provides significantly higher coverage than pyrosequencing for the same cost. This enables a higher detection sensitivity of mRNAs, especially in the control which is mostly comprised of rRNAs, for testing statistical significance of the difference between control and treatments. Approximately 10-17 million 76-bp single-end Illumina reads were generated for each sample (**Supplementary Table 1**). Two samples, Exo (ii) (Run 1) and Exo (ii) (Run 4), were excluded due to technical failure and Exo (ii) in Experiment 1 was re-run (Run 2) producing ~33 million 76 bp paired-end reads (**Supplementary Table 1**). With the exception of Exo (i) (Run 3), ~99% of reads for each sample mapped to a reference genome (**Supplementary Table 1**), indicating both sound read quality and negligible contamination.

Supplementary Note 3. Normalization for GC content-associated inter-run variation

We tried two simple normalization strategies: by organism and by gene. When normalizing the mRNA abundance from both runs by the total mRNA read counts from its source organism, the five slopes converged to a value of 1

(**Fig. 1d**), and the correlation for all the genes increased from 0.950 to 0.992. This indicates that inter-run variation is reduced by separating organisms with distinct GC contents. When normalizing the transcript abundance by the GC content of individual genes on one run (See **Online Methods** for details), the overall correlation improved (*r* between Runs 3 and 4 improved from 0.970 ± 0.014 to 0.995 ± 0.007), confirming that the run to run variation is indeed largely associated with GC content. Therefore, to accurately assess rRNA depletion efficiency and mRNA fidelity, we restricted comparisons of treatment-control pairs to the same run in Experiment 2. The inter-run average was taken after the intra-run comparison was performed.

Supplementary Note 4. Difference between actual and observed rRNA removal

The observed post-depletion rRNA fraction does not accurately reflect the amount of rRNA that is actually removed. To illustrate this, let us assume a hypothetical total RNA sample containing 95 rRNA and 5 mRNA molecules (i.e. 95% observed rRNA). Removal of 80% of the rRNA leaves 19 rRNA molecules and the original 5 mRNAs (i.e. 79% observed rRNA). This demonstrates that the observed rRNA depletion is only 16% despite 80% actual removal. **Fig. 2c** describes the relation between actual and observed rRNA removal, and shows that observed rRNA removal does not exceed 20% until actual removal is >80%. The higher the starting percentage of rRNA, the larger the difference is between the observed and actual percent removals. For example, an 18% observed removal in Experiment 1 (94.9% initial rRNA) and Experiment 2 (96.9% initial rRNA) corresponds to 82% and 88% actual removal, respectively (**Fig. 2c**).

LEGAL DISCLAIMER