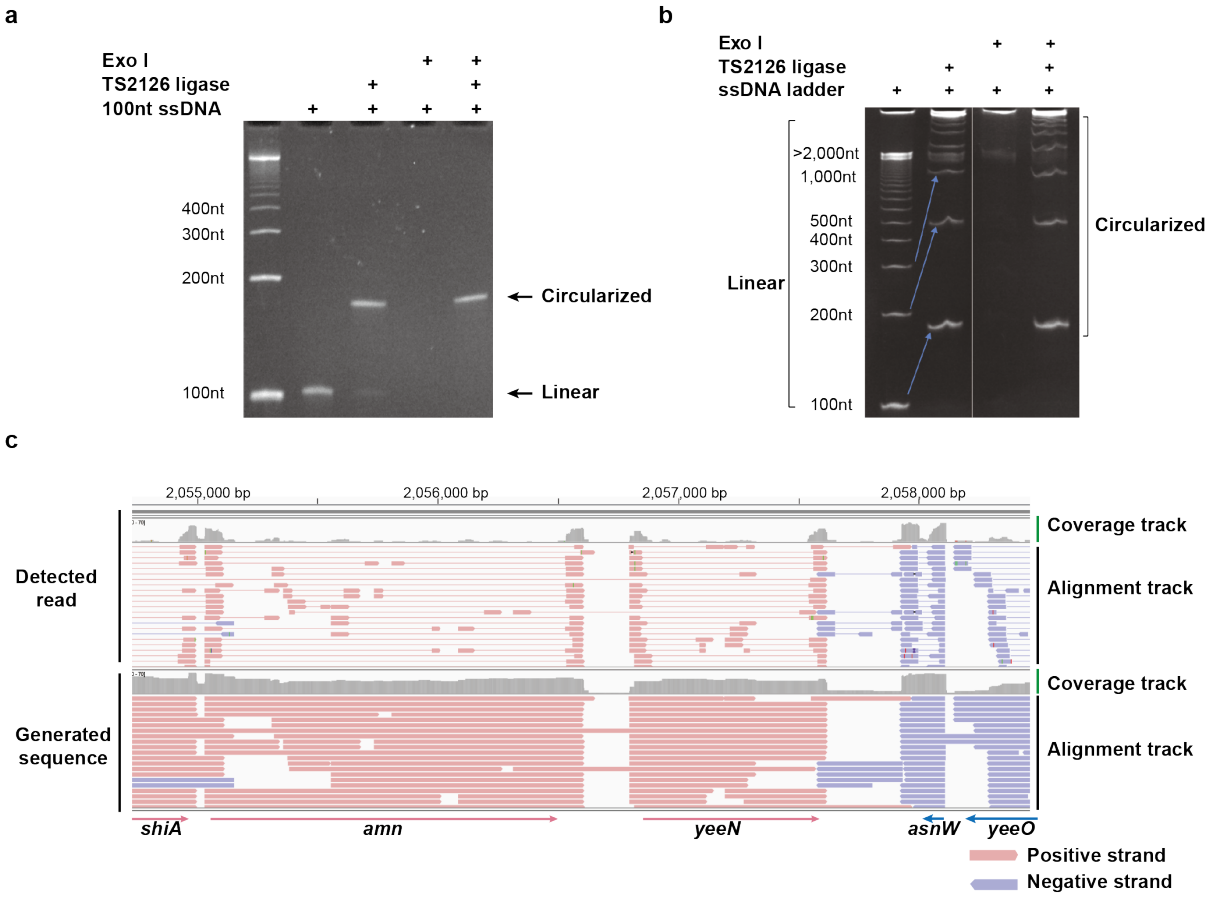
Supplementary Information for

Full-length RNA profiling reveals pervasive bidirectional transcription terminators in bacteria

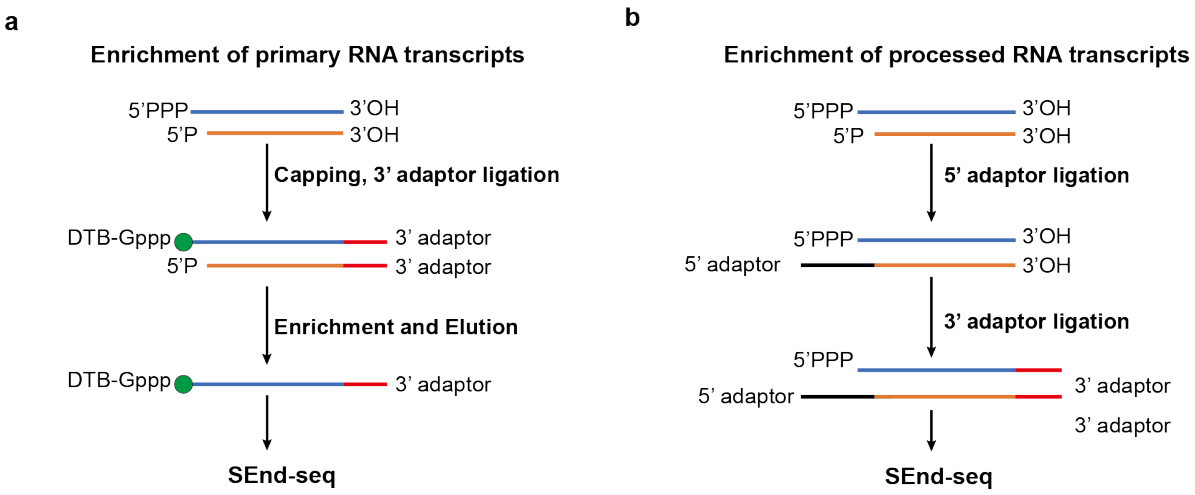
Xiangwu Ju, Dayi Li, Shixin Liu*
*e-mail: shixinliu@rockefeller.edu

Table of Contents

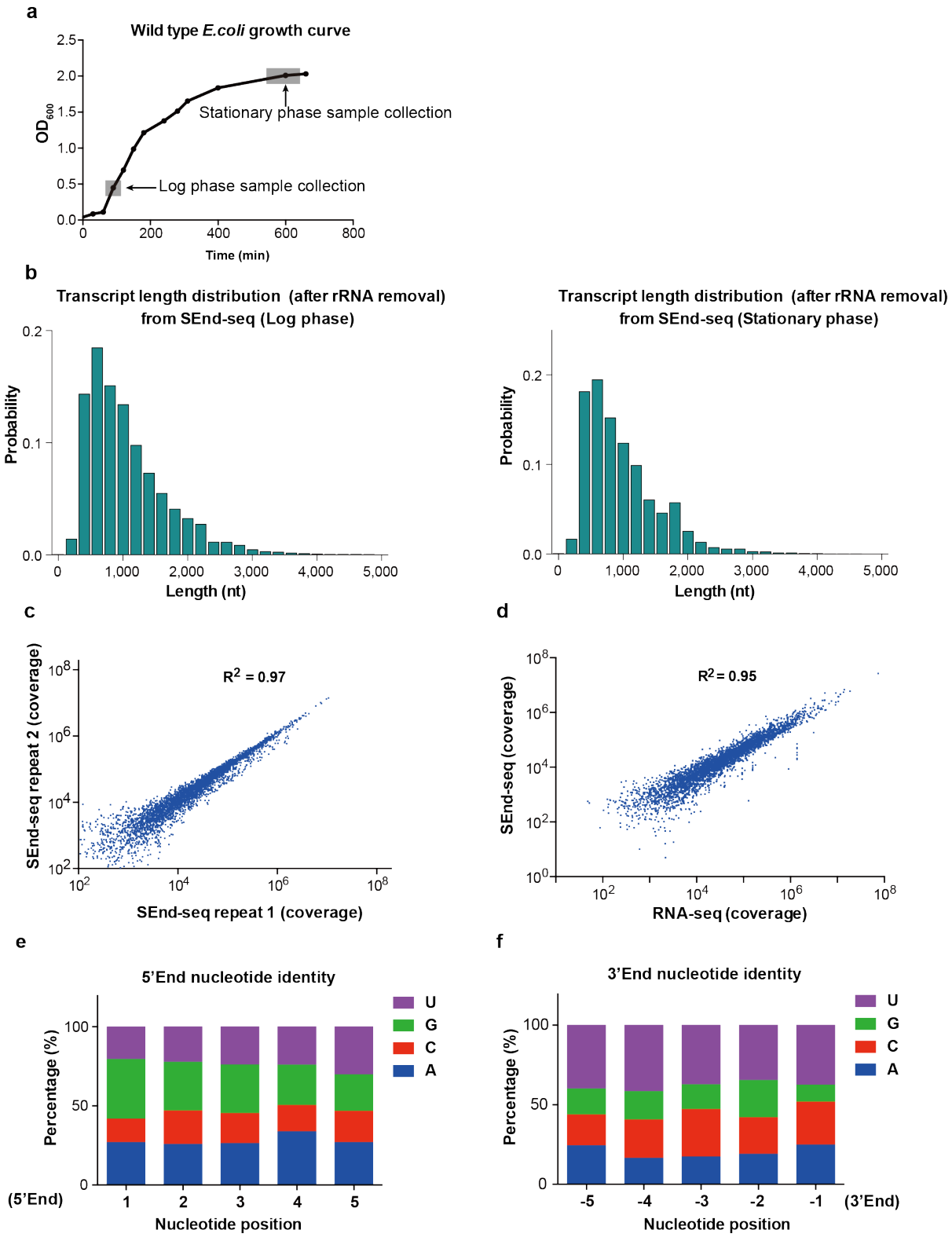
Supplementary Table 1. Genomic position of TSS identified by SEnd-seq (separate Excel file)
Supplementary Table 2. Genomic position and predicted secondary structure of TTS identified by SEnd-seq (separate Excel file)
Supplementary Table 3. Transcription units (TU) defined by SEnd-seq (separate Excel file)
Supplementary Table 4. Genomic position and structural analysis of overlapping bidirectional TTS and the *E. coli* genes involved (separate Excel file)
Supplementary Table 5. Oligonucleotides used in this study (attached here)



Supplementary Figure 1 | Additional information on the SEnd-seq workflow. a, A 100-nt ssDNA was circularized by the TS2126 ligase and visualized on a 10% urea-PAGE gel. Circularized DNA is resistant to exonuclease (Exo I) treatment, while linear DNA is efficiently digested by Exo I. The lack of concatemeric products suggests predominant intramolecular ligation (circularization). **b**, Gel showing the circularization efficiency of a ssDNA ladder ranging from 100 nt to >2,000 nt. **c**, A representative IGV data track illustrating how to generate full-length RNA sequences from raw SEnd-seq reads such as the one shown in Fig. 1b. Data are representative of three independent experiments.

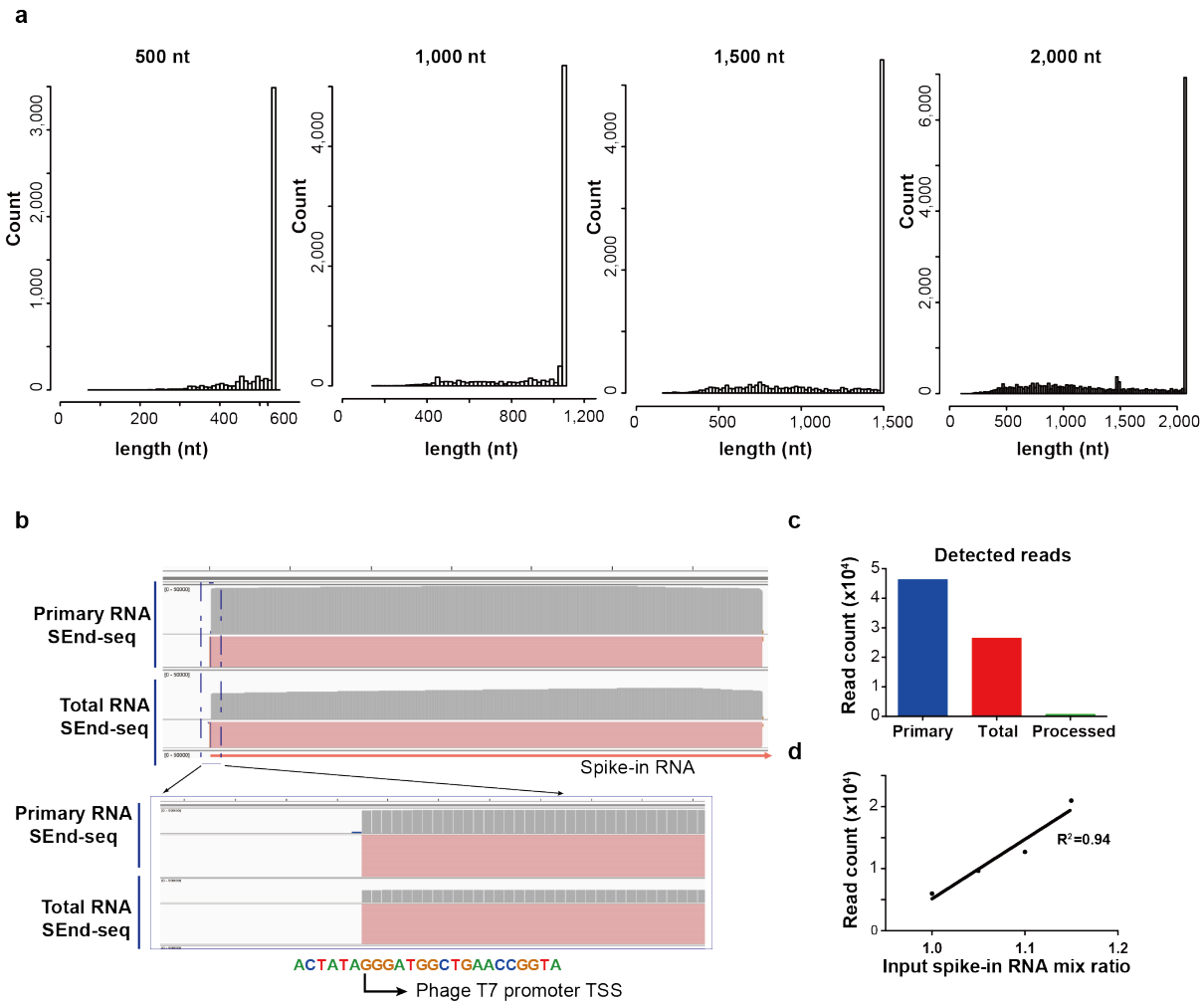


Supplementary Figure 2 | Strategies for enriching primary and processed transcripts. a, 5'-triphosphorylated primary transcripts are exclusively capped with desthiobiotin and then isolated by streptavidin beads[1]. To check the efficiency of processed RNA removal, we examined the abundance of the ribosomal RNA (rRNA)—which are predominantly 5' monophosphorylated—in the primary RNA dataset. Only less than 10% of the reads were mapped to rRNA as compared to 80% in the total RNA dataset. **b**, Alternatively, 5'-monophosphorylated processed RNA are exclusively ligated to a 5' adaptor. The rest of the workflow is the same as that shown in Fig. 1a.

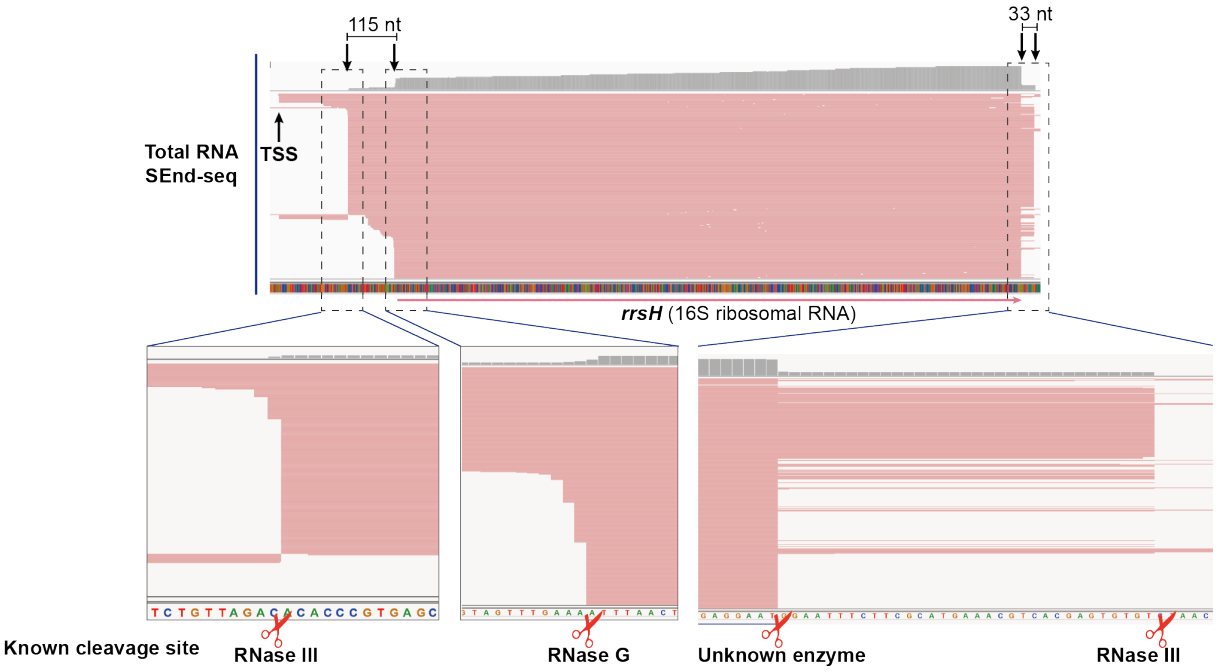


Supplementary Figure 3 | Analysis of the transcriptome datasets yielded by SEnd-seq. a, *E. coli* growth curve indicating the time points at which cells were collected for log-phase and

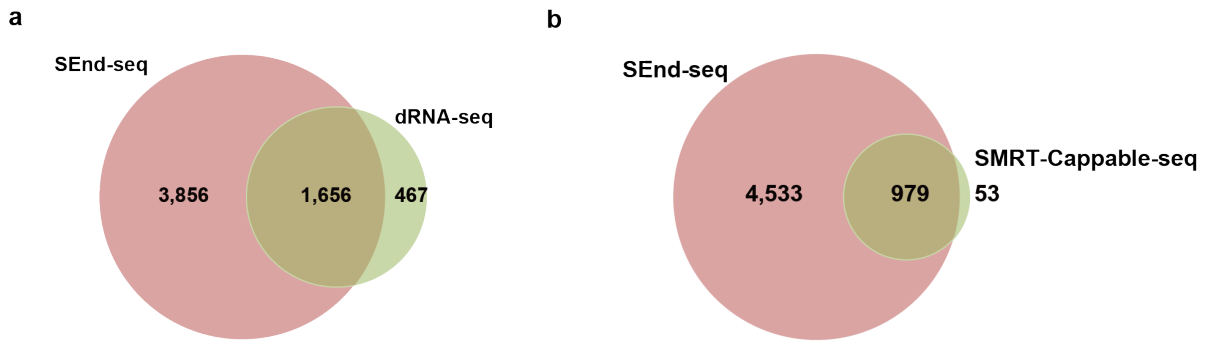
stationary-phase samples. **b**, Distribution of the transcript length (ribosomal RNA removed) recovered from SEnd-seq for log-phase (left) and stationary-phase (right) samples. **c**, Correlation for the coverage to each individual gene between two independent SEnd-seq replicates. **d**, Correlation for individual gene coverage between SEnd-seq and standard RNA-seq datasets. Pearson correlation coefficients are shown. **e**,**f**, Distribution of the nucleotide identity at the 5' end (**e**) and 3' end (**f**) of transcripts recovered from primary RNA SEnd-seq. The bias for A and G at the 5' end is consistent with their known enrichment at TSS[2]. The enrichment for U near the 3' end is also expected given the U-tract in intrinsic terminators. Data are representative of three independent experiments.



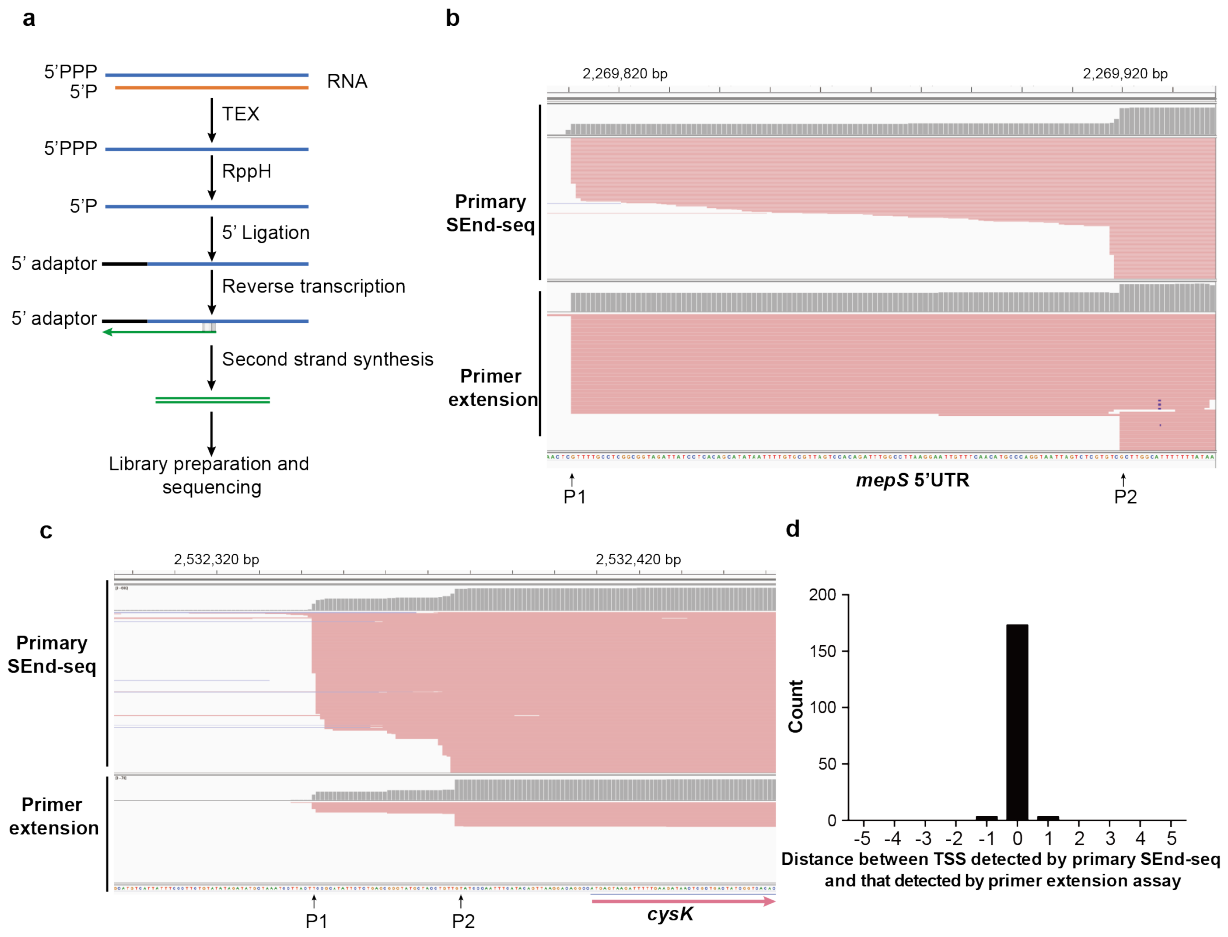
Supplementary Figure 4 | Evaluation of the performance of SEnd-seq with spike-in RNA. a, Distribution of the spike-in RNA length recovered after SEnd-seq analysis for four different input species. **b**, SEnd-seq data track for a 680-nt spike-in RNA. The anticipated TSS controlled by a phage T7 promoter is shown at the bottom. **c**, Read count of the spike-in RNA in the total, primary and processed RNA datasets. The low count of the spike-in RNA—which are 5' triphosphorylated—in the Processed column suggests successful depletion of primary RNA from the processed RNA sample. **d**, Correlation between the read count for different spike-in RNA species and their input amount.



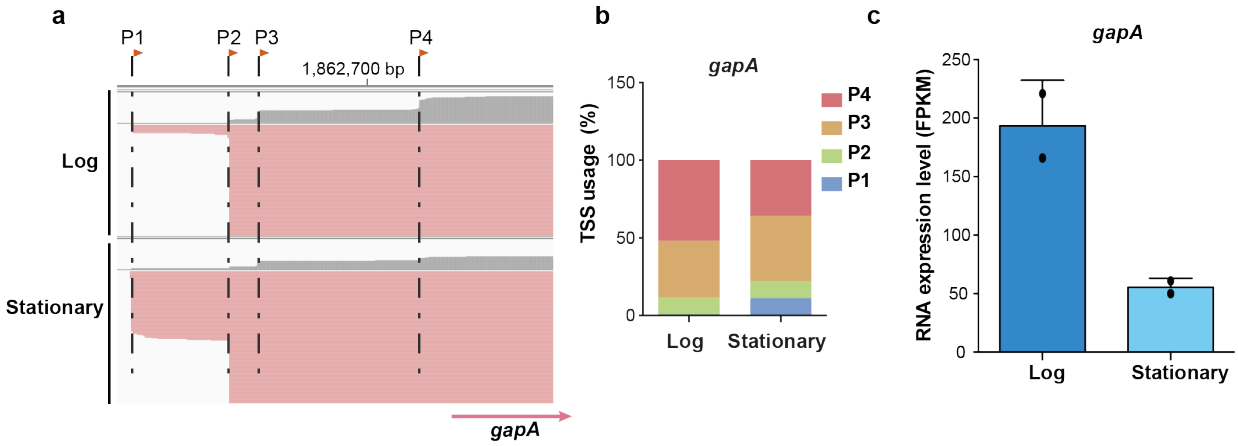
Supplementary Figure 5 | Demonstration of single-nucleotide resolution afforded by SEnd-seq. SEnd-seq detects the major intermediates in the maturation pathway of the 16S rRNA. The 16S rRNA precursor is trimmed by RNase III and subsequently by RNase G on the 5' end (their cleavage sites are separated by 115 nt); its 3' end is trimmed by RNase III and another unknown enzyme (their cleavage sites are separated by 33 nt)[3].



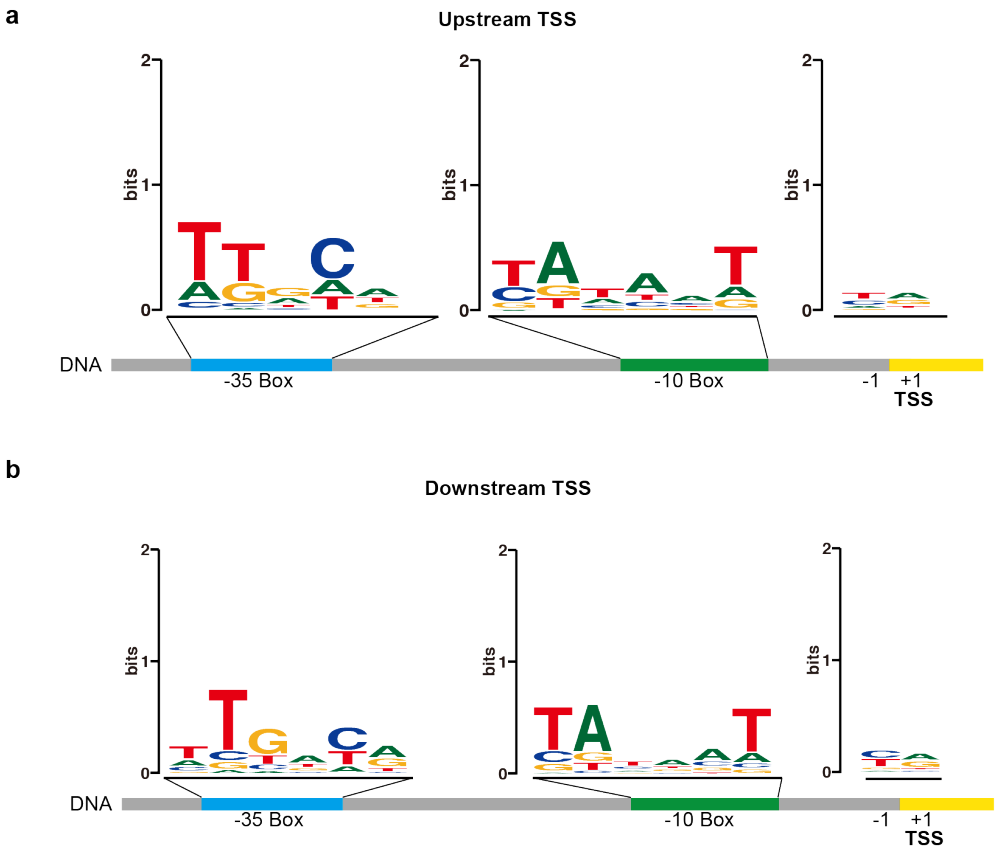
Supplementary Figure 6 | Comparison between SEnd-seq and other RNA 5'-end mapping methods. a, Venn diagram comparing TSS identified by SEnd-seq and by dRNA-seq[4]. **b**, Venn diagram comparing TSS identified by SEnd-seq and by SMRT-Cappable-seq[5]. Datasets obtained with log-phase samples were used for the comparison.



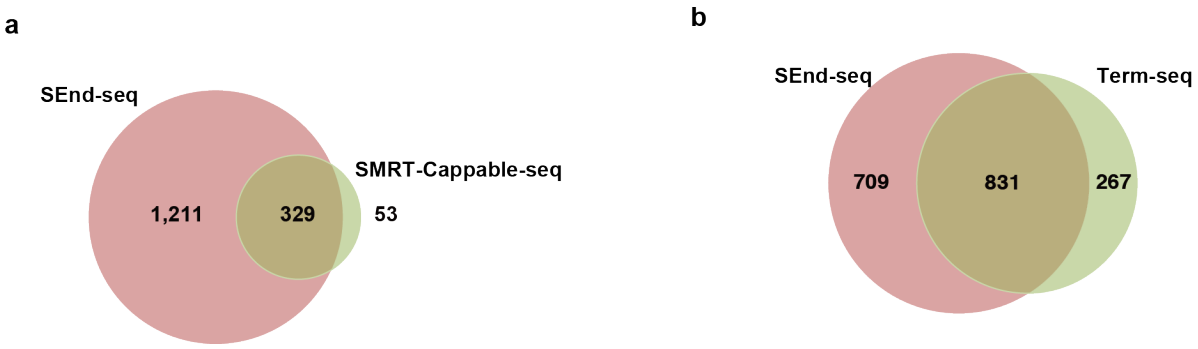
Supplementary Figure 7 | Validation of TSS by primer-extension assays. a, Workflow of the primer-extension assay. See Methods for details. **b**,**c**, Examples of primary SEnd-seq and primer extension results showing consistent RNA 5' ends. Data are representative of two independent experiments. **d**, Distribution of the distance between the TSS position from primary SEnd-seq and that for the same TSS obtained from the primer-extension assay. Out of the 180 TSS tested—38 of which are newly identified sites—the vast majority exhibit an exact match between the two assays.



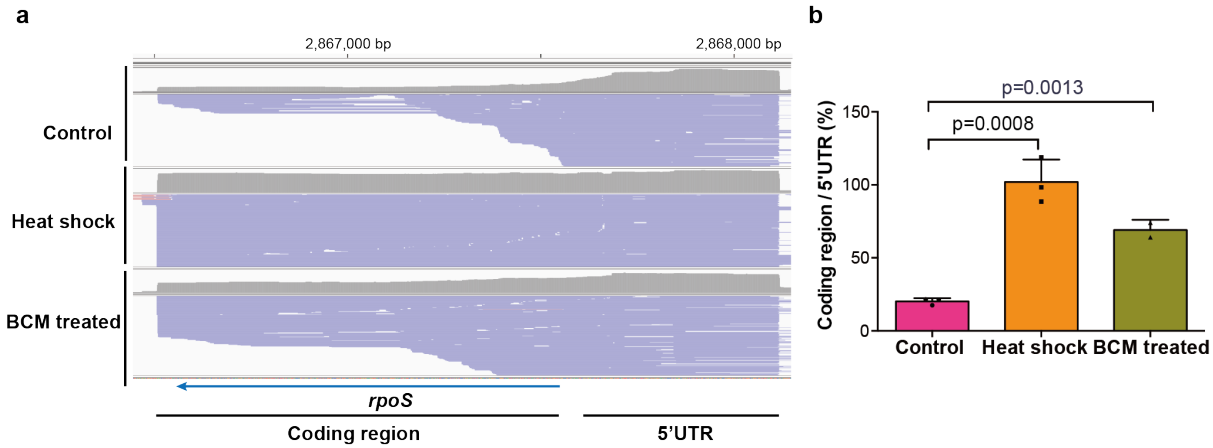
Supplementary Figure 8 | Growth-condition-dependent TSS usage revealed by SEnd-seq. a, SEnd-seq data track for *gapA* showing the usage of alternative TSS at different growth stages. Data are representative of two independent experiments. **b**, Quantification of the fraction of *gapA* transcripts starting at a given TSS. **c**, Expression level of the *gapA* gene in the log versus stationary phase reported by standard RNA-seq. Data are mean ± s.d. from two independent experiments.



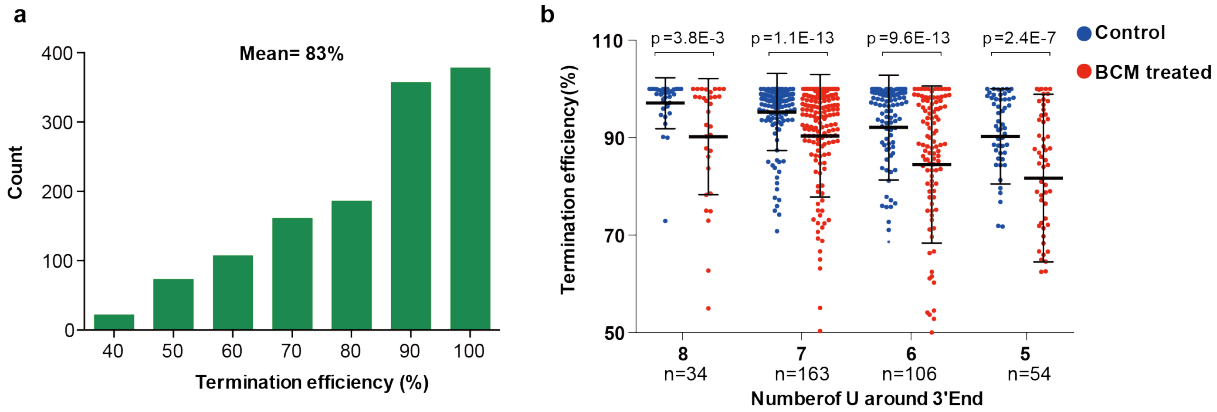
Supplementary Figure 9 | Motif analysis of the sequences around the upstream TSS (a) and downstream TSS (b) for genes that employ multiple start sites.



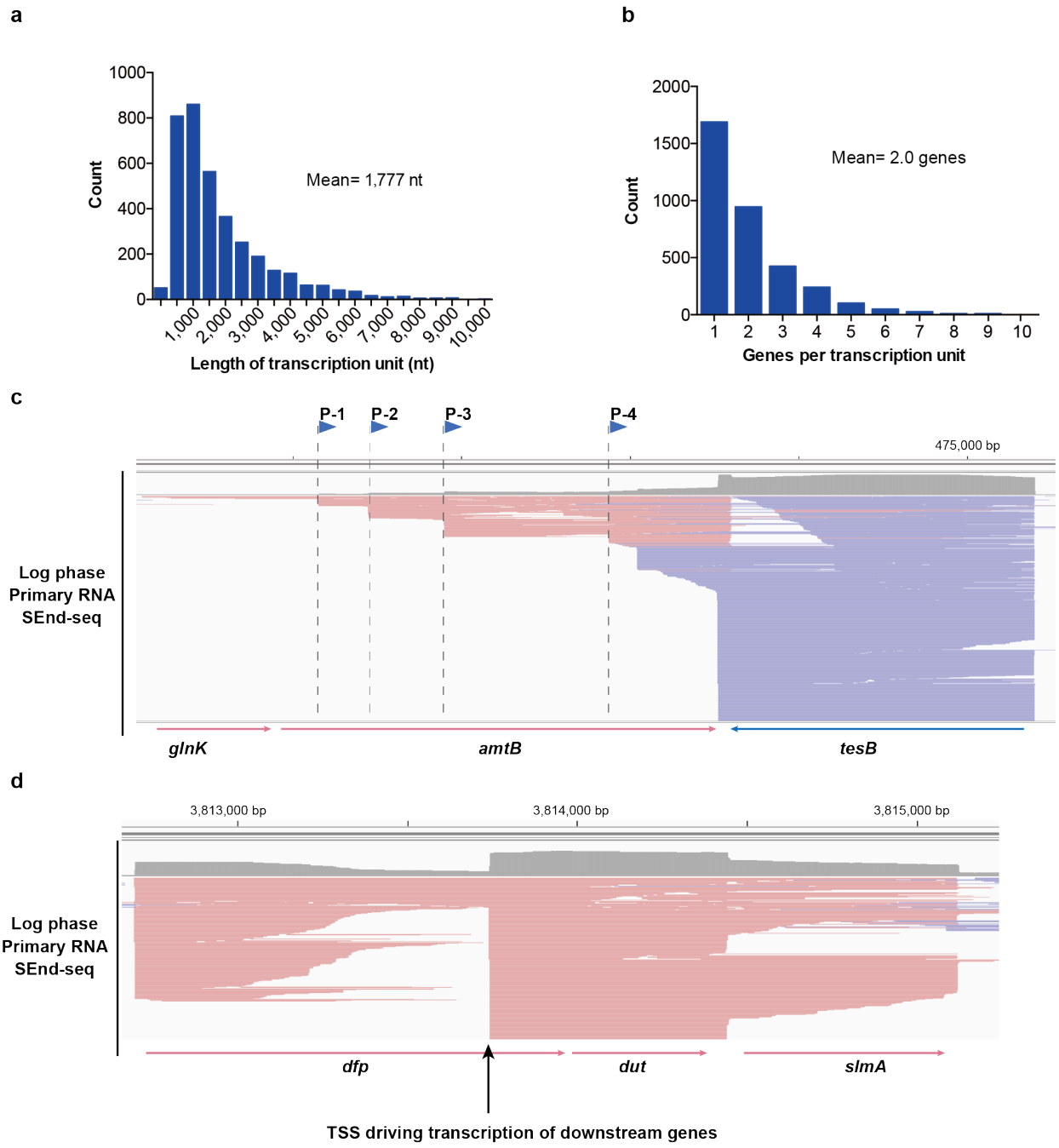
Supplementary Figure 10 | Comparison between TTS detected by SEnd-seq and other RNA 3'-end mapping methods. a, Venn diagram comparing TTS identified by SEnd-seq and by SMRT-Cappable-seq[5]. **b**, Venn diagram comparing TTS identified by SEnd-seq and by Term-seq[6].



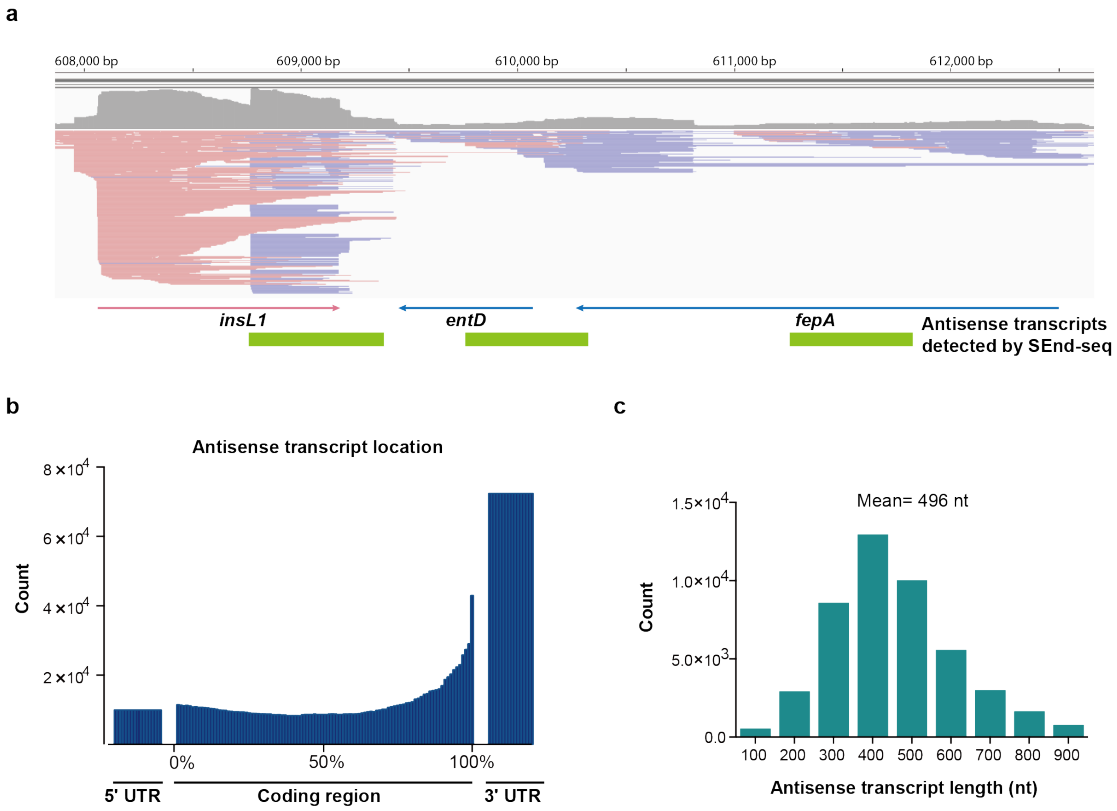
Supplementary Figure 11 | SEnd-seq detects premature termination events. a, SEnd-seq data track for the *rpoS* gene, which encodes a general stress sigma factor. Consistent with previous results[7], our data showed that most *rpoS* transcripts prematurely terminate in the 5' UTR during normal log phase growth. In contrast, full-length transcription is greatly stimulated under heat shock conditions. Inhibition of the Rho activity by BCM also suppresses premature termination. **b**, Number of *rpoS* reads mapped to the coding region over that mapped to the 5' UTR under normal, heat shock, and BCM-treated conditions. Data are mean ± s.d. from three independent experiments. *P* values were determined by two-sided unpaired Student's *t*-tests.



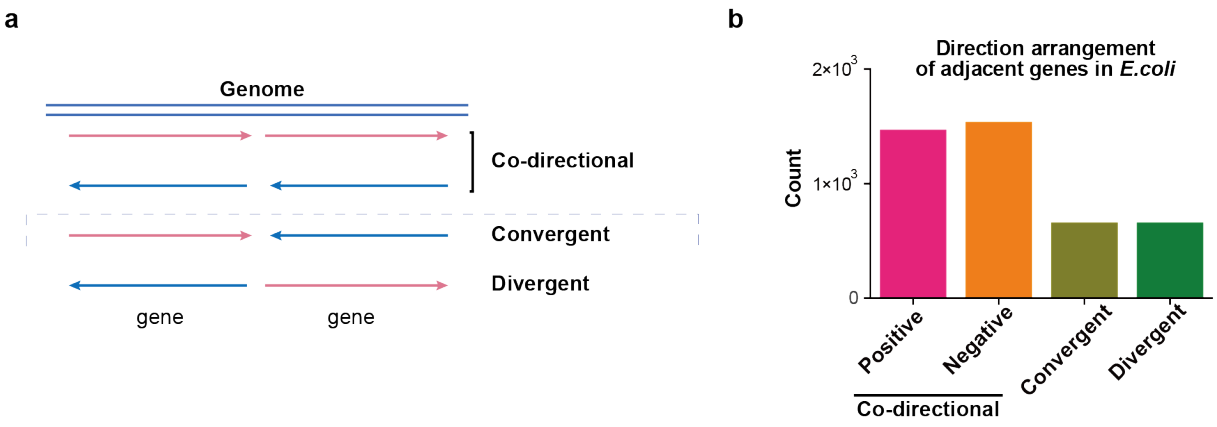
Supplementary Figure 12 | Additional characterization of TTS identified by SEnd-seq. a, Distribution of termination efficiencies for all TTS identified by SEnd-seq. Log-phase samples were used for the analysis. Data are representative of two independent experiments. **b**, Termination efficiency as a function of the number of uracil residues found at the 3' flank of the hairpin in the absence or presence of the Rho inhibitor BCM. *n* denotes the number of terminators analyzed in each category. Data are mean ± s.d. *P* values were determined by two-sided unpaired Student's *t*-tests.



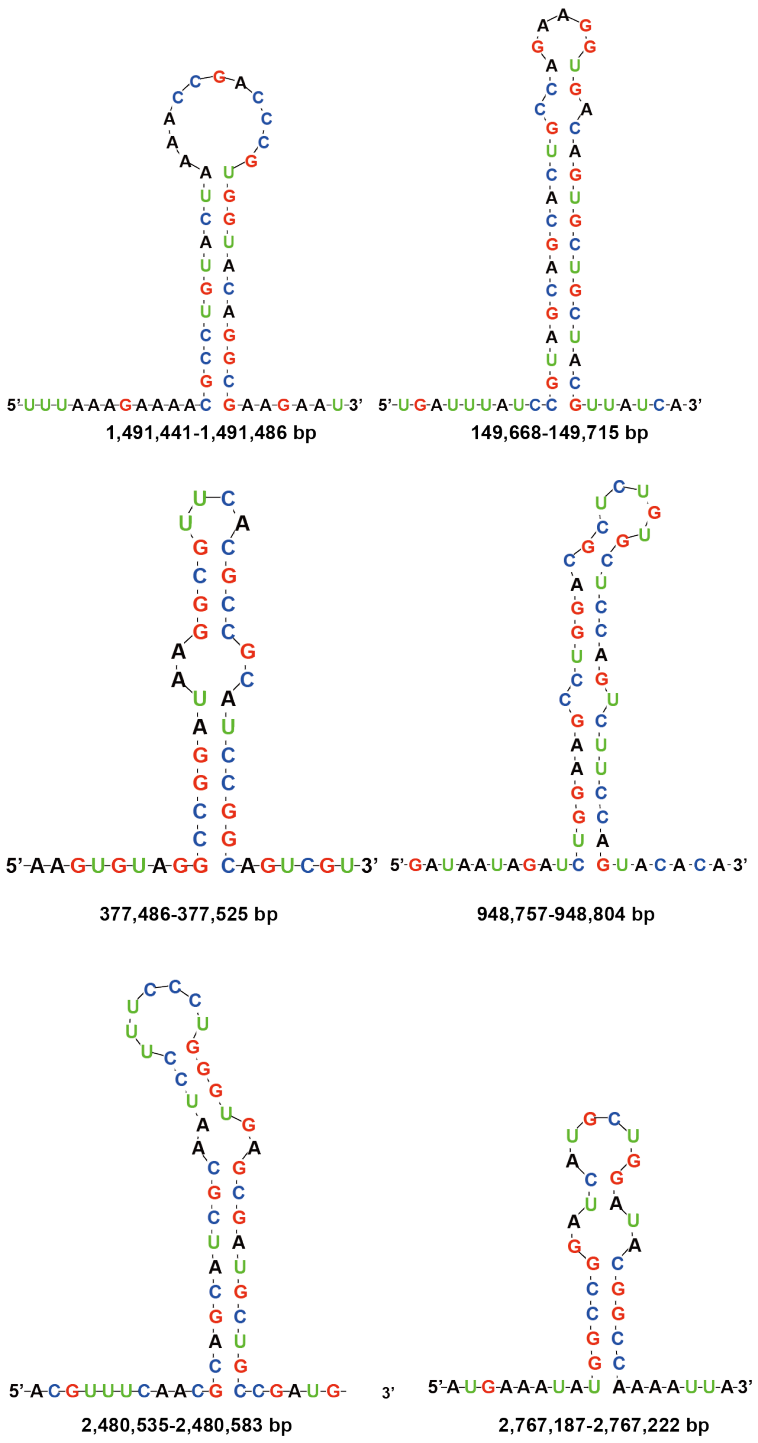
Supplementary Figure 13 | Analysis of transcription units (TU) defined by SEnd-seq. a, Length distribution of TU bound by a unique TSS and TTS. **b**, Distribution of the number of genes covered by each TU. **c**, An example of TU with TSS located inside the coding region of an annotated gene (*amtB*). **d**, An example of TU with an intragenic TSS that drives transcription of downstream genes (*dut* and *slmA*). Data are representative of three independent experiments.



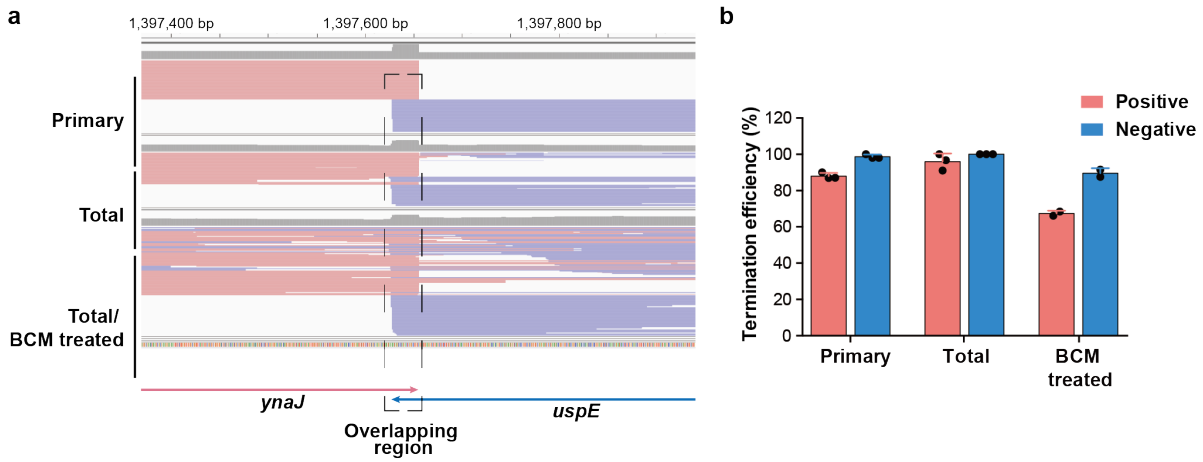
Supplementary Figure 14 | Antisense transcripts detected by SEnd-seq. a, SEnd-seq data track showing examples of antisense transcripts (blue reads for the *insL1* gene, red reads for *entD* and *fepA*). **b**, Distribution of the position of detected antisense transcripts with respect to annotated genes. **c**, Length distribution of detected antisense transcripts. The scarcity of transcripts shorter than 100 nt is due to the procedure of sample preparation, which removes most of short RNA species including tRNA.



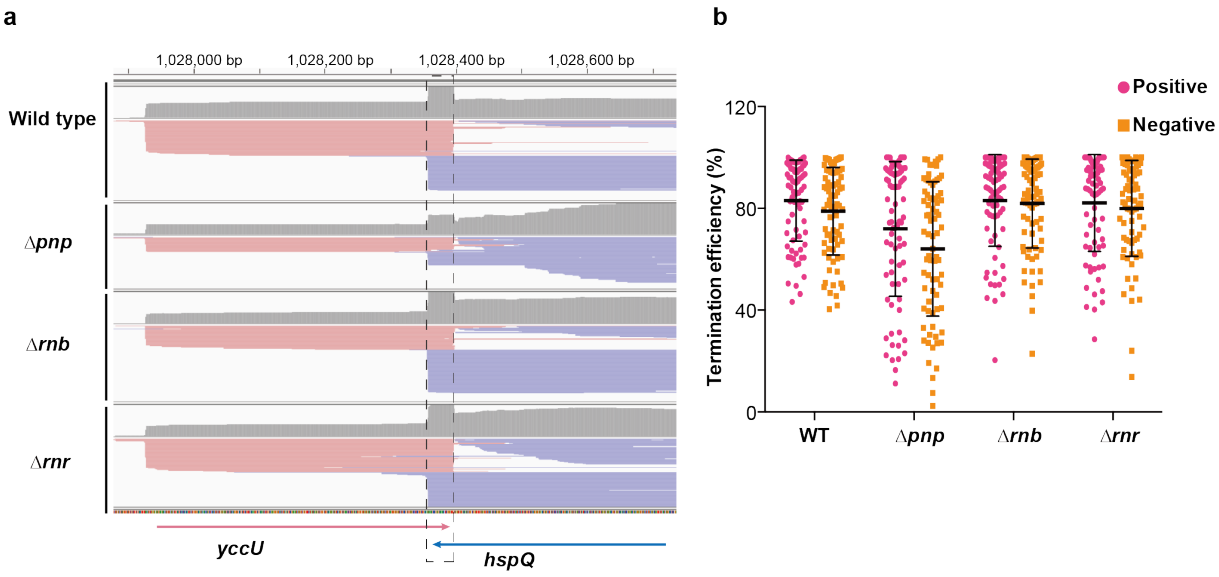
Supplementary Figure 15 | Direction arrangement of adjacent genes in the *E. coli* genome. a, Possible scenarios for the arrangement of adjacent genes. **b**, Statistics for the above scenarios in the *E. coli* genome.



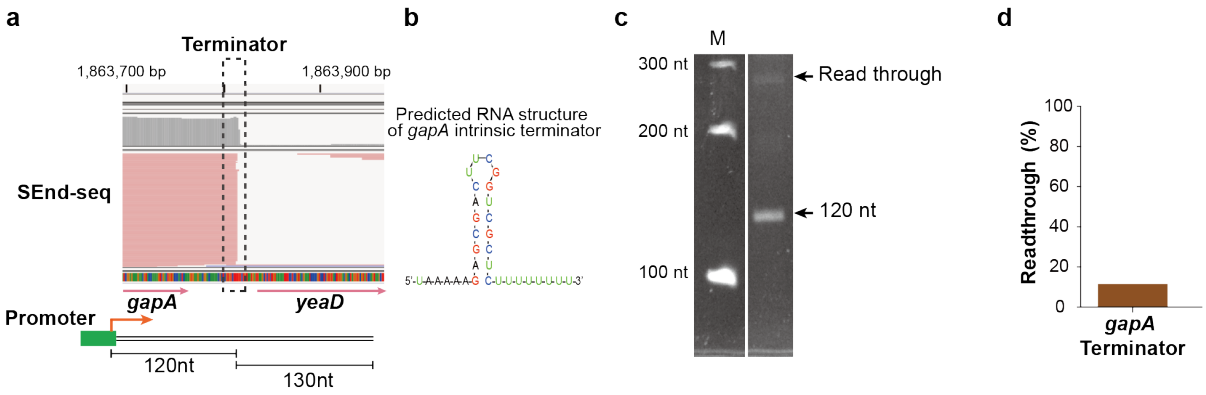
Supplementary Figure 16 | Examples of predicted stem-loop structure formed at the overlapping bidirectional TTS.



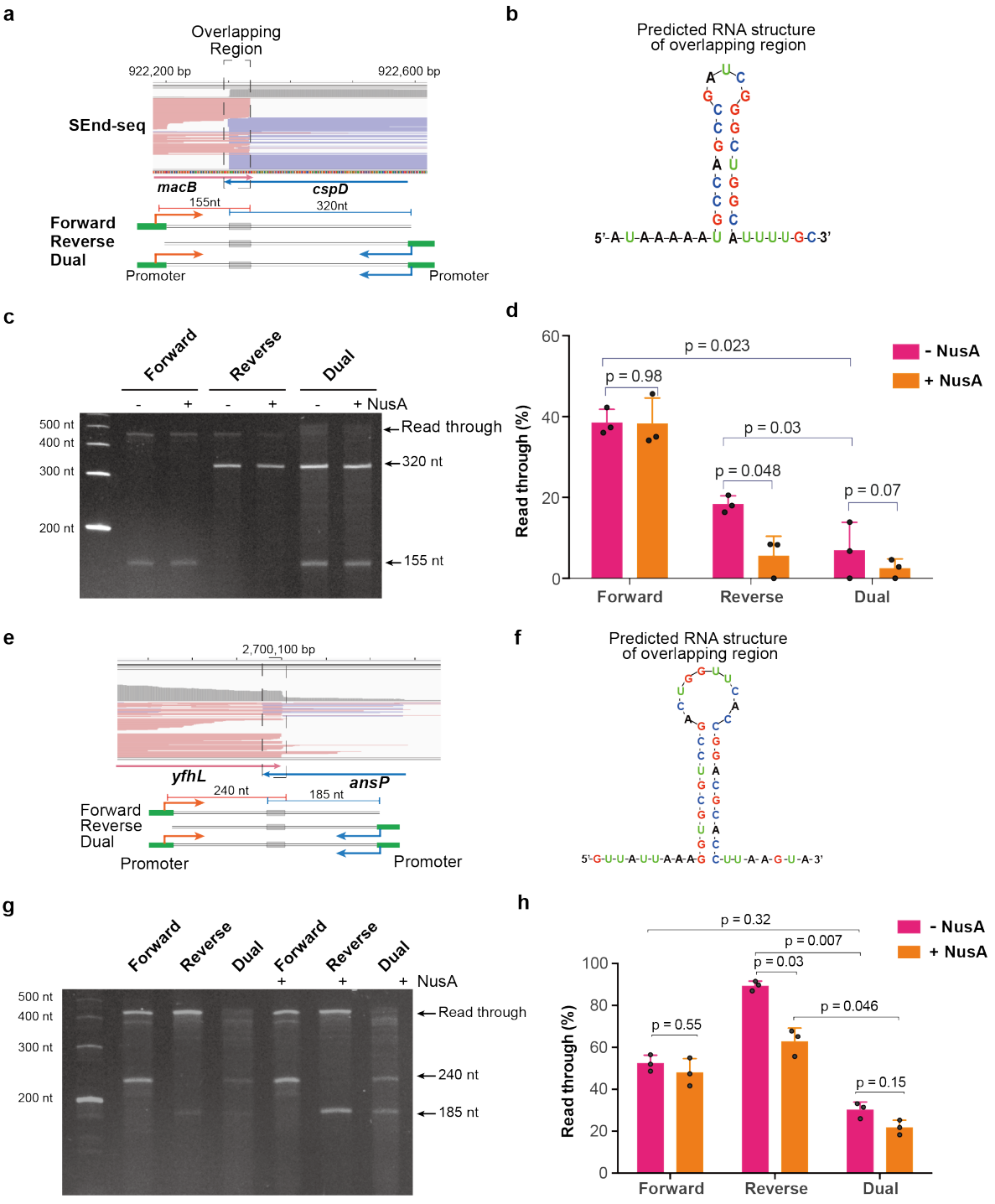
Supplementary Figure 17 | Effect of Rho inhibition on the termination efficiency of overlapping bidirectional TTS. a, SEnd-seq data track around the bidirectional TTS between the *ynaJ-uspE* convergent gene pair using primary RNA, total RNA, or total RNA treated with the Rho inhibitor BCM. Data are representative of two independent experiments. **b**, Average termination efficiency of bidirectional TTS featuring a highly expressed gene pair (*n* = 78) in the positive or negative direction for the different samples described in (**a**). Data are mean ± s.d. from two or three independent experiments.



Supplementary Figure 18 | Comparison of the termination efficiency of overlapping bidirectional TTS among different exonuclease knockout strains. a, SEnd-seq data track around the bidirectional TTS between the *yccU-hspQ* convergent gene pair for the wildtype, *pnp*-, *rnb*-, and *rnr*-knockout strains. Data are representative of two independent experiments. **b**, Termination efficiency of overlapping bidirectional TTS that feature a highly expressed gene pair (n = 78) in either orientation for the different strains described in (**a**). Data are mean ± s.d.

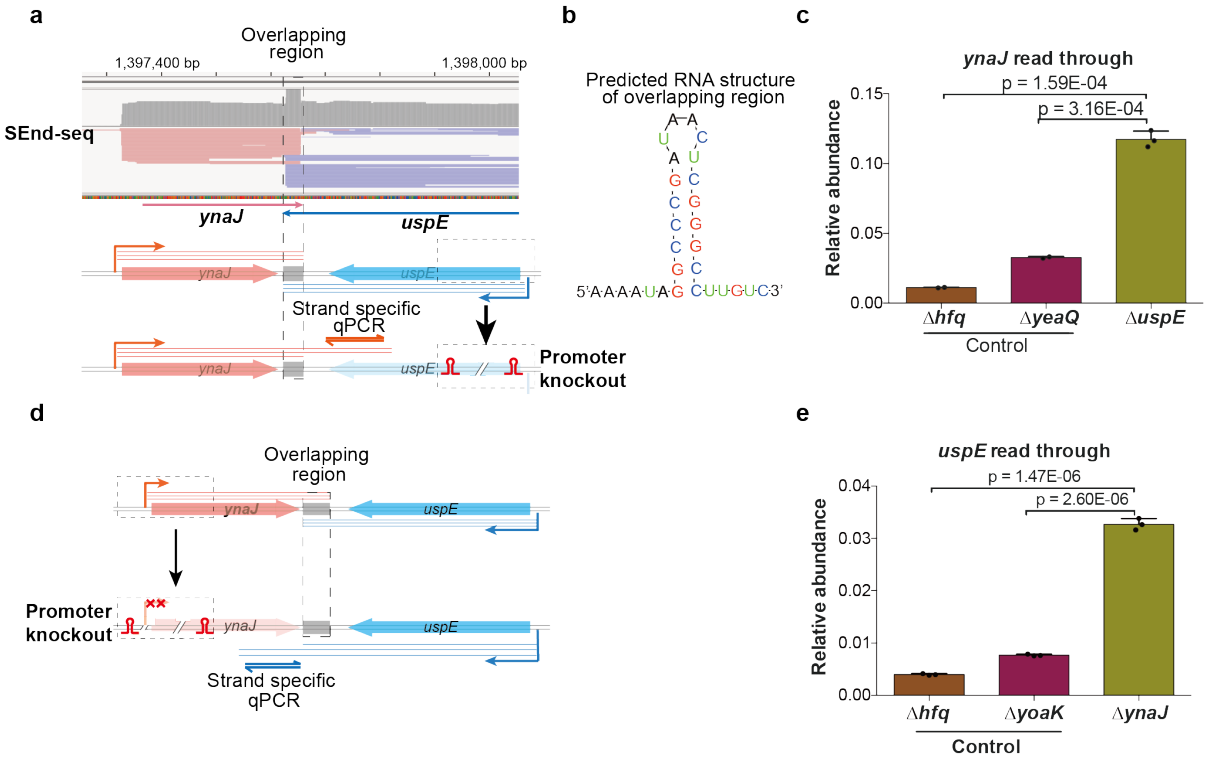


Supplementary Figure 19 | Example of a strong intrinsic terminator that causes efficient transcription termination in vitro. a, A DNA template harboring a T7A2 promoter and the intrinsic terminator downstream of the *gapA* gene was constructed for in vitro transcription. SEnd-seq reveals a high termination efficiency for the *gapA* terminator in vivo. **b**, Predicted secondary structure of the *gapA* terminator showing features of a strong intrinsic terminator (a GC-rich stem-loop and an 8-nt 3' U-tract). **c**, Gel showing the transcription products using the template described in (**a**). Data are representative of three independent experiments. **d**, Quantification of the fraction of readthrough transcripts.

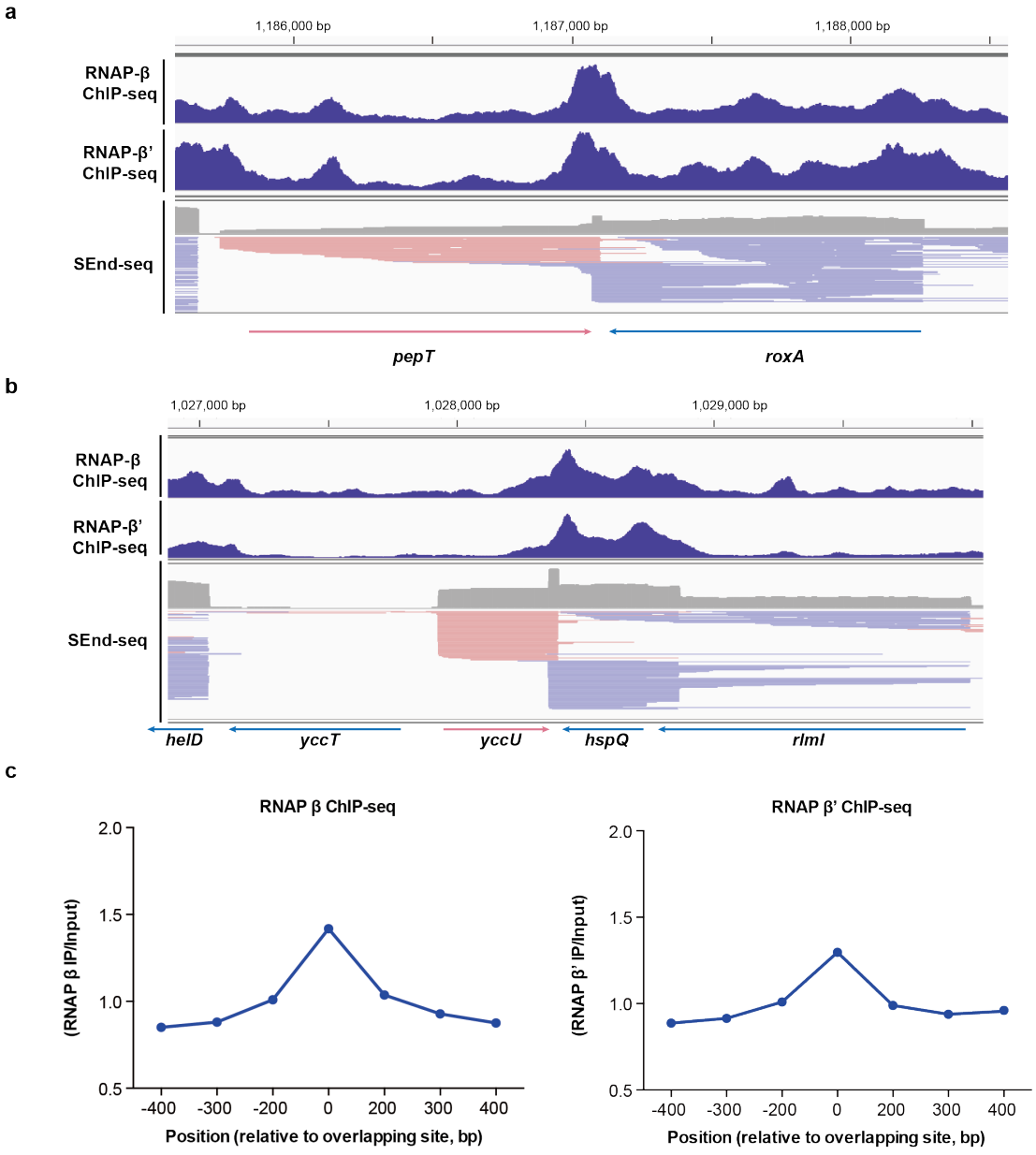


Supplementary Figure 20 | Examples of bidirectional termination caused by convergent transcription in vitro. a, Three DNA templates that all contain the *macB-cspD* overlapping TTS were constructed. The first two templates (Forward and Reverse) contain one T7A2 promoter at either end. In the Dual construct, each end contains a promoter sequence. **b**, Predicted secondary structure of the overlapping bidirectional TTS between *macB* and *cspD*. **c**, Gel showing the size of RNA products from in vitro transcription with the different templates described in (**a**) in the absence or presence of NusA. Data are representative of three independent experiments. **d**, Fraction of readthrough transcripts for the different conditions described in (**c**). Data are mean ± s.d. *P* values were determined by two-sided unpaired

Student's *t*-tests. **e**-**h**, The same experiments as in (**a**-**d**) were repeated with the *yfhL-ansP* overlapping TTS. Note that in this case, there is neither a strong 3' U-tract nor a strong 5' A-tract flanking the hairpin as shown in (**f**). Consistently, substantial readthrough was observed for both directions. Convergent transcription enhances bidirectional termination efficiency, which is further improved by NusA. Data are mean ± s.d. from three independent experiments. *P* values were determined by two-sided unpaired Student's *t*-tests.



Supplementary Figure 21 | Example demonstrating that convergent transcription contributes to bidirectional termination in vivo. a, SEnd-seq data track for the *ynaJ-uspE* convergent gene pair and schematic of in vivo genome editing. **b**, Predicted secondary structure of the overlapping bidirectional TTS between *ynaJ* and *uspE*. **c**, qPCR results showing the relative abundance of *ynaJ* readthrough transcripts across the overlapping region when *uspE* transcription is abolished (*ΔuspE*). Alternatively, transcription of *hfq* or *yeaQ* was disrupted as negative controls. Data are mean ± s.d. from three independent experiments. *P* values were determined by two-sided unpaired Student's *t*-tests. **d**,**e**, Reciprocal results showing that disrupting *ynaJ* transcription (*ΔynaJ*) increases the level of *uspE* readthrough across the overlapping TTS. Data are mean ± s.d. from three independent experiments. *P* values were determined by two-sided unpaired Student's *t*-tests.



Supplementary Figure 22 | RNAP occupancy is enriched around overlapping bidirectional TTS. **a**,**b**, Example RNAP ChIP-seq and primary SEnd-seq data tracks showing the enrichment of RNAP occupancy around the overlapping bidirectional TTS between a convergent gene pair [*pepT-roxA* in (**a**), *yccU-hspQ* in (**b**)]. Data are representative of two independent experiments. **c**, Cumulative RNAP ChIP-seq signal around the overlapping bidirectional TTS summed over those sites identified by SEnd-seq using stationary-phase samples. ChIP-seq was conducted using antibodies against the RNAP β or β' subunit.

Figure 5c:

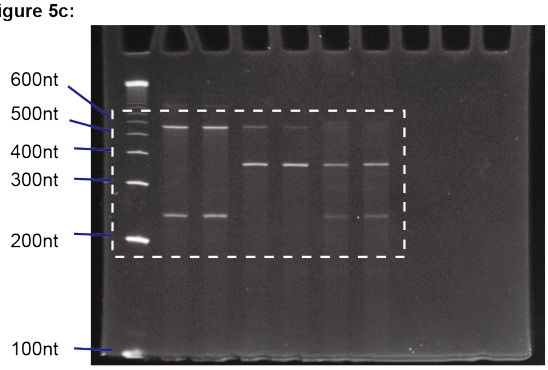
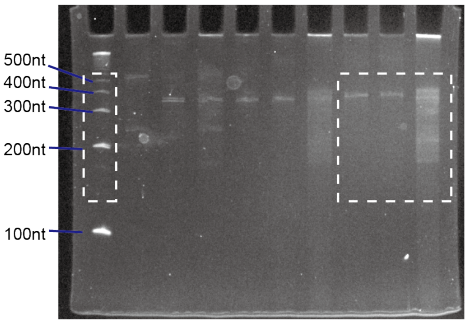
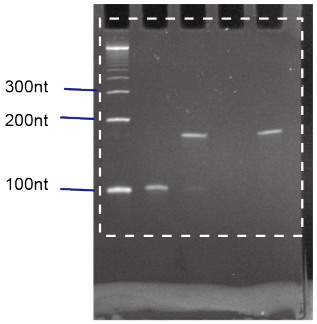


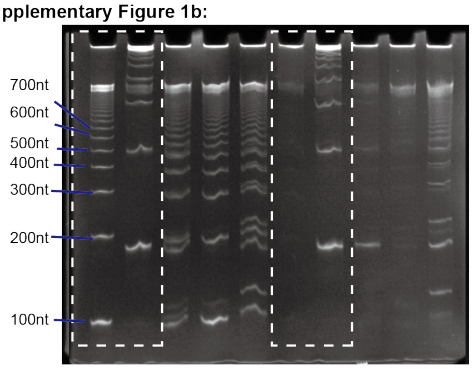
Figure 5g:



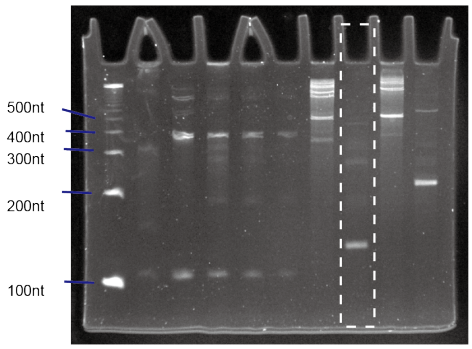
Supplementary Figure 1a:



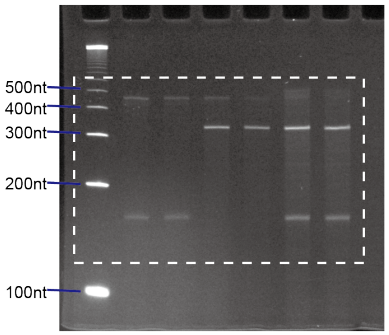
Supplementary Figure 1b:



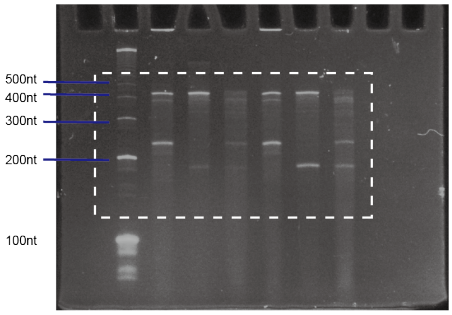
Supplementary Figure 19c:



Supplementary Figure 20c:



Supplementary Figure 20g:



Supplementary Figure 23 | Full-length gels. Boxes indicate the cropped areas used to generate the corresponding figures.

Supplementary Table 5 | Oligonucleotides used in this study.

1) Oligonucleotides for SEnd-seq library preparation

Name	Sequence
5' adaptor	/5SpC3/rArArA rCrUrC rUrCrC rArCrG rUrNrN rNrNrN rN
3' adaptor-1 (barcode1)	/5Phos/rNrNrNrNrArArCrCrUrGrCrUrArUrCrArArC rUrG/3ddC/
3' adaptor-2 (barcode2)	/5Phos/rNrNrNrNrGrCrUrUrCrCrUrGrCrUrArUrCrA rArCrUrG/3ddC/
3' adaptor-3 (barcode3)	/5Phos/rNrNrNrNrNrNrArUrA rGrCrA rArCrC rUrGrC rUrArU rCrArA rCrUrG /3ddC/
Biotinylated RT primer	GGG CAG T/iBiodT/G ATA GCA GG

/5SpC3/: 5' C3 Spacer modification
/5Phos/: 5' Phosphorylation modification
/3ddC/: 3' Dideoxycytidine (ddC) modification
/iBiodT/: Internal biotin dT
'r': ribonucleotide
Sequences are represented in the 5'-3' orientation. All of these oligonucleotides were HPLC purified after synthesis.

2) Oligonucleotides for in vitro RNA synthesis by *E. coli* RNA polymerase

To incorporate the T7A2 promoter to both ends of a target DNA, the genome DNA template was first amplified by the half promoter set of primers. The cleaned template was further amplified by the two full-length T7A2 promoter primers (T7A2 full-1 forward, T7A2 full-2 reverse), which contain the same promoter sequence.

Name	Sequence
T7A2 half-1 1878781 forward	GCT AGG TAA CAG ACG CAA TCA ATG GGT GCG GTA TTA TTT TTC CGG TTG TAA TC
T7A2 half-2 1879297 reverse	GCT AGG TAA CAT GAT GAT TGT GCG TGC TAC AGC GAC TAA CCC GCT AAT GCA AAA CAG
T7A2 full-1 forward	TGG TCG GCG TTT TTA AAA CAG GTA TTG ACA ACA TGA AGT AAC ATG CAG TAA GAT ACA AAT CGC TAG GTA ACA GAC GCA ATC AAT GGG TGC
T7A2 full-2 reverse	CGG CGC CAT TTT TAA AAC AGG TAT TGA CAA CAT GAA GTA ACA TGC AGT AAG ATA CAA ATC GCT AGG TAA CAT GAT GAT TGT GCG TGC TAC
T7A2 half-1 922200 forward	GCT AGG TAA CAG ACG CAA TCA ATG GGT GCC TCT GCT CGA CGG TCA CCG GG
T7A2 half-2 922600 reverse	GCT AGG TAA CAT GAT GAT TGT GCG TGC TAC ATG TCG AAG CAT GGA AAA GGG
T7A2 half-1 1879000 forward	GCT AGG TAA CAG ACG CAA TCA ATG GGT GCT AAG CGT TAA CTT TTA ATC TTC
T7A2 half-2 1879300 reverse	GCT AGG TAA CAT GAT GAT TGT GCG TGC TAC ATC AGC GAC TAA CCC GCT AAT G
T7A2 half-1 1028290 forward	GC TAG GTAACAGACGCAAT CAATGGGTGCCCG GGT TAA ATG TGG TGA TGG
T7A2 half-2	GC TAG GTAACATGATGAT TGTGC GTGCT ACG GAC GAT AAC GGC CTA CCG G

1028540 reverse	
T7A2 half-1 1863700 reverse	GC TAG GTAACAGACGCAAT CAATGGGTGCGTG GTA CGA CAA CGA AAC CGG
T7A2 full-1 2699900 forward	GCT AGG TAA CAG ACG CAA TCA ATG GGT GCA CCA CGC GGA TAA AAT TTA ATA TTC
T7A2 full-2 2700210 reverse	GCT AGG TAA CAT GAT GAT TGT GCG TGC TAC CGA ATG ACT GAT TGC CGA TAC C
T7A2 full-1 1524276 forward	GCT AGG TAA CAG ACG CAA TCA ATG GGT GCA TGG CAA AAC GCT GTA TGT CAG
T7A2 full-2 1524600 reverse	GCT AGG TAA CAT GAT GAT TGT GCG TGC TAC CTG GTT ATA GGC TGG TTT GGT G
1863950 reverse	GGA TGA TCG ACC ACA ATG AGG TCC

The numbers indicate genome positions.

3) Oligonucleotides for in vitro RNA synthesis by phage T7 RNA polymerase

Name	Sequence
Spike-in RNA (C gene 500nt) forward	TAC GTA ATA CGA CTC ACT ATA GGG GTG ACA GCA GAG CTG CGT AAT CTC
Spike-in RNA (C gene 500nt) reverse	GCC GGG AGG CGG CAC TGG CAA G
Spike-in RNA (B gene 1000nt) forward	TAC GTA ATA CGA CTC ACT ATA GGG ATG AAA ACG CCC ACC ATT CCC AC
Spike-in RNA (B gene 1000nt) reverse	CCG TAT CCT GAG CCG TCT GCA G
Spike-in RNA (A gene 1500nt) forward	TAC GTA ATA CGA CTC ACT ATA GGG GTG AAT ATA TCG AAC AGT CAG G
Spike-in RNA (A gene 1500nt) reverse	GGC CAC CGG CTT TCC GTA GAC GG
Spike-in RNA (H gene 2000nt) forward	TAC GTA ATA CGA CTC ACT ATA GGG ATG GCT GAA CCG GTA GGC GAT CTG
Spike-in RNA (H gene 2000nt) reverse	TCA TCA TGG ACA GCA CGG AAC GGG

4) Oligonucleotides for genome editing and qPCR

The template of kanamycin-resistance cassette flanked by FLP recognition target sites was originally amplified by PCR from the *E. coli* gene deletion strain generated in a previous paper[8] with TR kana stop site set primers, which would yield a strong intrinsic terminator at either end after modification. The different genome target primer sets were then used to amplify the template for promoter deletion.

For *pnp-, rnr-* and *rnb*-knockout strains, the corresponding primer set was used to directly amplify DNA template for target gene deletion.

Name	Sequence
TR kana stop site forward	GCA TAC ATA AAA AAC CTG CCA GCG ATG GCA GGT TTT TTT TAT TCC GGG GAT CCG TCG ACC
TR kana stop site reverse	AAC CAC GTT TAA AAA AGG TGC TCA ATG AGC ACC TTT TTT CTG TAG GCT GGA GCT GCT TCG
yoak PD_stop forward	TTT CGT CGC TGC GAG AAC GCA TTC AGC AGC AAA TAC GTG GGC ATA CAT AAA AAA CCT GCC
yoak PD_stop reverse	CTT TAT GCT CCC GGG GCA GCA TAG CCG CCA ATA AAA AAC CAA CCA CGT TTA AAA AAG GTG
ynaj PD_stop forward	ATC GAA TAC ATG AAA GCC AGC ATC CGG GCC AGG GTG GAG CGC ATA CAT AAA AAA CCT GCC
ynaj PD_stop reverse	AAA TTT CTT CCC TTT CGC TGA CTT CAG TTT CGC CAT AAT CAA CCA CGT TTA AAA AAG GTG
uspe PD_stop forward	TCT TCT GGC AGA CCT TTT TCT ACG TGC GTC ATG TTT TCA TGC ATA CAT AAA AAA CCT GCC
uspe PD_stop reverse	AAA TTT TTC TGA TTT ATT GAT CTG GCA GAA GGT TCA TCA CAA CCA CGT TTA AAA AAG GTG
hspq PD_stop forward	CCA GCT CAT CAG GCG ACG GTT CAG AGA GCG AAT AAA CCG GGC ATA CAT AAA AAA CCT GCC
hspq PD_stop reverse	TAT TGC ACG GCA GAA CGT TGA GCT GAA CAA ACT GGA TCT GAA CCA CGT TTA AAA AAG GTG
yoak PD_S pcr forward	ACC CAT CAT AAT AAT CGT TGT T
yoak PD_S pcr reverse	TTC ATT ACA GCG GTC GTA
yoak PD_S RT primer	TAA TCT TCA TTA CAG CG
ynaj PD_S pcr forward	GTA AAG ACC AGG CTA TAA ACA A
ynaj PD_S pcr reverse	TAT TCC GCT GTC TGA GTG
ynaj RT primer	AAT TTC TCT TTG GTT TG
uspe PD_S qpcr forward	TCT TCT TCG TCA TCC AGT T
uspe PD_S qpcr reverse	CGG TAT TTC AGC AGC ATT
uspe PD_S RT primer	GGT GAT TCC TGA TTT GG
hspq PD_S qpcr forward	ATG GCT GTT CAG GAT GTT
hspq PD_S qpcr reverse	GTA GTG ATG GAG GAC GAT AA
hspq PD_S RT primer	CGG TGA ATG ACG AGC TTC
pnp del forward	CCC CCC GCC GCA GCG GAG GGC AAA TGG CAA CCT TAC TCG CAT TCC GGG GAT CCG TCG ACC
pnp del reverse	TTC CCA CTG ACC GTT AAC TAT CAG GAG CGT ACC TAC GCT GTG TAG GCT GGA GCT GCT TCG
rnr del forward	ATC CTT TCC AGG AAC GCG AAG CTG AAA AAT ACG CGA ATC CAT TCC GGG GAT CCG TCG ACC
rnr del reverse	CCT CTT CTT TTA AAG AGG GTA TTG ATC ACT CTG CCA CTT TTG TAG GCT GGA GCT GCT TCG
rnb del forward	CGC GAC CGG GCG CGC AAT AAT GCT GCG GGT TTC CAT GCG GAT TCC GGG GAT CCG TCG ACC

rnb del reverse	GTT TCA GGA CAA CCC GCT GCT AGC GCA GCT TAA ACA GCA ATG TAG GCT GGA GCT GCT TCG

Supplementary References

1. Ettwiller, L., Buswell, J., Yigit, E. & Schildkraut, I. A novel enrichment strategy reveals unprecedented number of novel transcription start sites at single base resolution in a model prokaryote and the gut microbiome. *BMC Genomics* **17**, 199 (2016).
2. Kim, D. et al. Comparative analysis of regulatory elements between Escherichia coli and Klebsiella pneumoniae by genome-wide transcription start site profiling. *PLoS Genet* **8**, e1002867 (2012).
3. Deutscher, M.P. Maturation and degradation of ribosomal RNA in bacteria. *Prog Mol Biol Transl Sci* **85**, 369-91 (2009).
4. Conway, T. et al. Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing. *MBio* **5**(2014).
5. Yan, B., Boitano, M., Clark, T.A. & Ettwiller, L. SMRT-Cappable-seq reveals complex operon variants in bacteria. *Nat Commun* **9**, 3676 (2018).
6. Dar, D. & Sorek, R. High-resolution RNA 3'-ends mapping of bacterial Rho-dependent transcripts. *Nucleic Acids Res* **46**, 6797-6805 (2018).
7. Sedlyarova, N. et al. sRNA-Mediated Control of Transcription Termination in E. coli. *Cell* **167**, 111-121 e13 (2016).
8. Baba, T. et al. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* **2**, 2006 0008 (2006).