nature microbiology

ARTICLES

https://doi.org/10.1038/s41564-019-0500-z

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

Xiangwu Ju[1], Dayi Li[1,2] and Shixin Liu[1*]

The ability to determine full-length nucleotide composition of individual RNA molecules is essential for understanding the architecture and function of a transcriptome. However, experimental approaches capable of capturing the sequences of both 5′ and 3′ termini of the same transcript remain scarce. In the present study, simultaneous 5′ and 3′ end sequencing (SEnd-seq)—a high-throughput and unbiased method that simultaneously maps transcription start and termination sites with single-nucleotide resolution—is presented. Using this method, a comprehensive view of the *Escherichia coli* transcriptome was obtained, which displays an unexpected level of complexity. SEnd-seq notably expands the catalogue of transcription start sites and termination sites, defines unique transcription units and detects prevalent antisense RNA. Strikingly, the results of the present study unveil widespread overlapping bidirectional terminators located between opposing gene pairs. Furthermore, it has been shown that convergent transcription is a major contributor to highly efficient bidirectional termination both in vitro and in vivo. This finding highlights an underappreciated role of RNA polymerase conflicts in shaping transcript boundaries and suggests an evolutionary strategy for modulating transcriptional output by arranging gene orientation.

It has become widely appreciated that RNA is not merely the messenger that relays genetic information from DNA to protein, but also itself carries out diverse regulatory roles in cell physiology[1]. The function of an RNA transcript is fundamentally determined by its constituent sequence elements, including those residing at the 5′ and 3′ ends. Next-generation RNA sequencing (RNA-seq) is a revolutionary tool for profiling a transcriptome—the set of all RNA molecules in a cell[2]. However, Illumina-based, short-read RNA-seq—the most commonly used platform for transcriptomic analysis—requires strand fragmentation, which decouples the 5′ end sequence of an RNA molecule from its 3′ end sequence. As such, the resultant transcriptome map reports ensemble RNA levels, but information on the end-to-end nucleotide composition of individual transcripts is inevitably lost. Although various methods have been developed to delineate the 5′ or 3′ extremities of transcripts[3–9], they cannot concomitantly sequence both ends of RNA. On the other hand, single-molecule, long-read sequencing platforms possess the ability to read an RNA molecule from one end to the other. None the less, their read depth, error rate and cost still compare unfavourably with the Illumina platform[10].

Prokaryotic transcriptomes were once considered to be simple due to their small size and limited splicing. This view is rapidly changing due to the growing list of RNA-based gene regulatory mechanisms found in prokaryotes[11,12]. However, counterintuitively, prokaryotic transcriptomic analyses have lagged behind eukaryotic counterparts. In particular, transcription termination sites (TTSs), which mark the 3′ ends of primary transcripts, have remained incompletely annotated even in model organisms such as *Escherichia coli*[13]. Intramolecular ligation of 5′- and 3′-RNA termini allows for simultaneous capture of the sequences of both ends, thereby representing a promising strategy for inferring full-length sequences of prokaryotic RNA. However, existing methods that employ this strategy suffer from strong length bias[14,15]. In addition, they are not readily applicable to prokaryotic transcripts because of their reliance on 3′ end polyadenylate tails for the generation of complementary cDNA. Thus, a method capable of comprehensively profiling full-length transcripts in prokaryotes is still urgently needed.

In the present study, a method—termed 'simultaneous 5′ and 3′ end sequencing' (SEnd-seq)—was developed to read both ends of cellular transcripts concurrently. This method enabled the determination of the correlated occurrence of transcription start sites (TSSs) and TTSs with single-nucleotide resolution across the whole transcriptome. Using SEnd-seq, a large number of previously unannotated TSSs and TTSs were identified in *E. coli*. Strikingly, SEnd-seq unveiled prevalent occurrence of overlapping bidirectional TTSs between head-to-head gene pairs or between a gene and an opposing non-coding RNA (ncRNA). Further in vitro and in vivo experiments were conducted to support the model in which convergent transcription is an important contributor to highly efficient bidirectional termination.

Results

Simultaneous 5′ and 3′ end capture by SEnd-seq. The general workflow of SEnd-seq is depicted in Fig. 1a. The key step is the circularization of cDNA by a single-stranded ligase that strongly favours intramolecular ligation[16] (see Supplementary Fig. 1a). Importantly, this step circularizes DNA of varying lengths with uniformly high efficiencies (see Supplementary Fig. 1b). After fragmentation, the biotin-labelled pieces containing the 5′–3′ junction are isolated and prepared for paired-end sequencing. The 5′ and 3′ end sequences are extracted and mapped to the reference genome (Fig. 1b). The full-length composition of individual transcripts is then inferred by connecting the two termini (see Supplementary Fig. 1c). Besides total RNA SEnd-seq, workflows were also developed to selectively enrich primary (5′-triphosphorylated) or processed (5′-monophosphorylated) transcripts (see Supplementary Fig. 2).

Evaluation of the performance of SEnd-seq. SEnd-seq was applied to *E. coli* cells collected under different growth conditions (see Supplementary Fig. 3a,b). The read coverage on each gene is highly

[1]Laboratory of Nanoscale Biophysics and Biochemistry, The Rockefeller University, New York, NY, USA. [2]Present address: The Kimmel Center for Biology and Medicine of the Skirball Institute, New York University School of Medicine, New York, NY, USA. *e-mail: shixinliu@rockefeller.edu

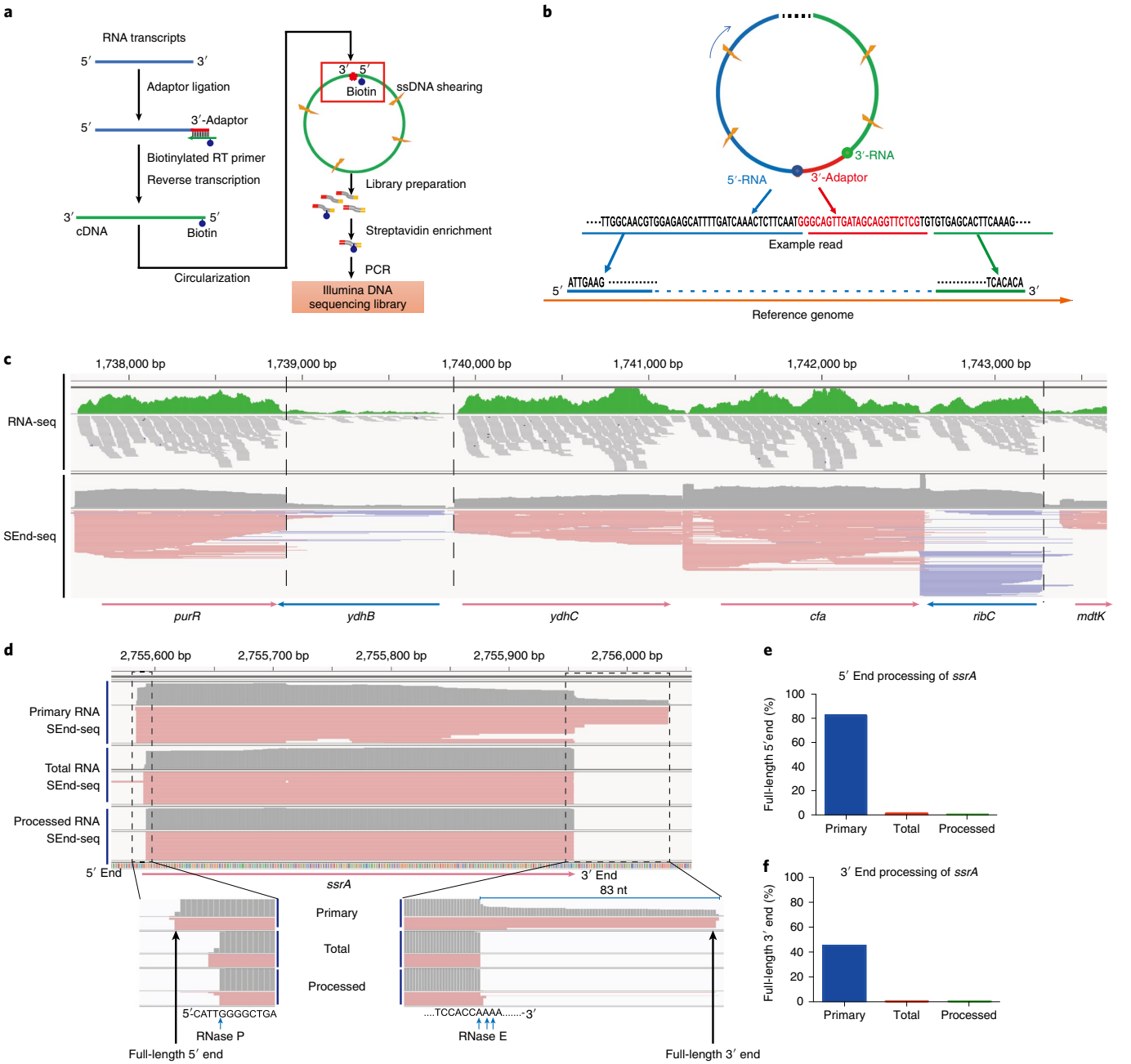

Fig. 1 | Simultaneous capture of 5′ and 3′ end sequences of bacterial transcripts by SEnd-seq. a, Workflow of SEnd-seq. See Methods for details. **b**, An example read illustrating how to infer the full-length sequence of individual transcripts by extracting correlated 5′ and 3′ end sequences and mapping them to the reference genome. **c**, A sample data track of the log-phase *E. coli* transcriptome showing the comparison between standard RNA-seq and SEnd-seq. Dashed lines highlight the sharp boundaries of transcripts delineated by SEnd-seq, which are obscured in standard RNA-seq. bp, basepair. **d**, SEnd-seq reads mapped to the *ssrA* gene in primary, total and processed RNA datasets. **e**, Ratio of *ssrA* transcripts with an intact, unprocessed 5′ end in different datasets. **f**, Ratio of *ssrA* transcripts with an intact 3′ end in different datasets.

correlated between SEnd-seq replicates (see Supplementary Fig. 3c), and between SEnd-seq and standard RNA-seq (see Supplementary Fig. 3d). The transcriptome dataset yielded by SEnd-seq exhibits no severe nucleotide bias at either the 5′ or the 3′ end of RNA (see Supplementary Fig. 3e,f).

The advantage of SEnd-seq over standard RNA-seq in mapping the boundaries of individual RNA molecules is apparent in a direct comparison of their respective data tracks (Fig. 1c). To assess the ability of SEnd-seq to reproduce the precise ends of input transcripts, a mixture of in vitro synthesized RNA was added to the cellular RNA and these were together subjected to SEnd-seq analysis. Correct lengths were recovered for all tested spike-in RNA species (see Supplementary Fig. 4a–c), indicating minimal sample deterioration during the procedure. The read coverage on each spike-in RNA species matches the ratio at which it was added to the mixture (see Supplementary Fig. 4d), arguing against any notable length bias of SEnd-seq.

It was also demonstrated that SEnd-seq can recover the boundaries of endogenous transcripts with single-nucleotide resolution. For example, intact 5′ and 3′ ends of the 452-nucleotide (nt) *ssrA*

RNA precursor were enriched in the primary RNA sample, whereas the processed and total RNA datasets predominantly yielded the mature 365-nt *ssrA* species, with its termini corresponding exactly to the known RNase cleavage sites[17] (Fig. 1d–f). As another example, the 1,861-nt 16S ribosomal RNA (rRNA) precursor, and the major intermediates in its maturation pathway, were successfully detected by SEnd-seq (see Supplementary Fig. 5).

Identification of transcription start sites. The single-nucleotide resolution afforded by SEnd-seq allows precise annotation of TSSs and TTSs in the same assay. Using primary RNA datasets, 4,358 and 4,038 TSSs were identified for log-phase and stationary-phase *E. coli* cells, respectively, of which 2,884 are common sites (Fig. 2a and see Supplementary Table 1). These sites are located both within intergenic regions and inside gene bodies (Fig. 2b,c). Most of them display a characteristic bacterial promoter sequence in the 5′ flank[18] (Fig. 2d).

SEnd-seq not only reproduced the vast majority of TSSs previously annotated by other 5′ end mapping methods[19,20], but also identified thousands of TSSs that were unknown until now (see Supplementary Fig. 6). A subset of these start sites was selected and validated by primer-extension assays (see Supplementary Fig. 7). The study found 2,133 genes that feature alternative TSSs upstream of their coding regions (Fig. 2e,f), indicating that those genes are each controlled by multiple promoters. In many cases, the usage of alternative TSSs is dependent on the growth condition (Fig. 2g,h and see Supplementary Fig. 8). For the genes that employ multiple TSSs, the fraction of transcripts initiated from the upstream TSS versus the downstream TSS (for example, *yajQ* (Fig. 2i,j)) were analysed. It was found that the most downstream TSS (that is, the one closest to the start codon) tends to make the largest contribution to the overall RNA expression level (Fig. 2k,l). The upstream and downstream TSS regions share a similar bacterial promoter −10 element, while exhibiting minor differences in the −35 element (see Supplementary Fig. 9).

Identification of TTSs. Two major transcription termination mechanisms have been well documented in bacteria: intrinsic termination, which is mediated by a hairpin structure formed in the nascent RNA followed by a U-rich tract, and factor-dependent termination, which relies on the rho ATPase[21]. The identification of TTSs is more challenging than that of TSSs because of the lack of chemical distinction between bona fide termination sites and processed 3′ ends, resulting in far fewer annotated TTSs in the existing database. To exclude post-processing cleavage sites, single-deletion *E. coli* strains were created in which each of the three genes (*pnp*, *rnb*, *rnr*) that encodes a major 3′ to 5′ exoribonuclease is knocked out[22]. Only those RNA 3′ ends that were not affected by any of these knockouts were annotated as TTSs, notwithstanding the caveat that these RNases probably have redundant roles. The study identified 1,285 TTSs that are common between log-phase and stationary-phase *E. coli* cells, as well as 255 growth-stage-specific ones (Fig. 3a and see Supplementary Table 2). SEnd-seq recaptures most of the TTSs annotated by other 3′ end mapping methods[20,23], but also finds a large number of previously unknown sites (see Supplementary Fig. 10). It was found that TTSs predominantly reside within intergenic regions (89%), although there are cases where termination occurs prematurely within the 5′-untranslated region (UTR) of a gene (see Supplementary Fig. 11).

TTS sites identified in the present study tend to form stable secondary structures (Fig. 3b). The termination efficiency, derived from the level of readthrough transcripts across the termination site, varies widely (see Supplementary Fig. 12a). The study assigned 709 TTSs as rho-dependent terminators based on their sensitivity to the rho-specific inhibitor bicyclomycin (BCM)[24] (Fig. 3c–f). Among the other TTSs, which are less sensitive to BCM treatment, many display sequence characteristics of an intrinsic terminator (a GC-rich hairpin followed by a 7- to 8-nt U-rich tract)[21] (Fig. 3g–i). As the number of uridines decreases, the termination efficiency drops—consistent with previous results[25]—and can be further reduced by rho inhibition (see Supplementary Fig. 12b). This result suggests that the intrinsic and rho-dependent termination mechanisms are not mutually exclusive and can act on the same site. Alternatively, such apparent overlap could result from RNase trimming after rho action downstream of the hairpin[23,26], despite the aforementioned exonuclease knockout not substantially changing the 3′ end pattern of these sites.

Taking advantage of the ability of SEnd-seq to simultaneously determine the 5′ and 3′ ends of the same transcript, the question was raised of whether the TSS selection—especially for those genes that employ multiple start sites—influences the termination efficiency at the corresponding TTS. It was found that 71 TTSs had their termination efficiency altered by at least 40%, depending on the choice of TSS (Fig. 3j,k), implying crosstalk between the two termini, as previously proposed[27,28].

Annotation of transcription units and antisense transcripts. The concomitant mapping of TSSs and TTSs enabled us to define 3,578 unique transcription units (TUs) in the *E. coli* transcriptome (see Supplementary Fig. 13a,b and Supplementary Table 3). Most TUs have their boundaries located within intergenic regions. We did detect 323 TUs with TSSs in a gene-coding region, yielding a shorter RNA product (see Supplementary Fig. 13c). We also found 452 TUs with an intragenic TSS that drives transcription of a downstream gene (see Supplementary Fig. 13d).

The ability of SEnd-seq to comprehensively profile full-length RNA of different sizes also allowed us to analyse the genome-wide distribution of antisense transcripts, the prevalence and importance of which in bacteria are increasingly being appreciated[29,30]. It was found that a substantial fraction of transcripts (~15%) are derived from the complementary strand of protein-coding genes. These antisense transcripts are mostly located towards the 3′ end of a coding region or within a 3′-UTR, and have a wide range of lengths (see Supplementary Fig. 14).

Prevalent overlapping bidirectional TTSs revealed by SEnd-seq. As demonstrated above, SEnd-seq provides an unprecedented inventory of the *E. coli* transcriptome. In the following the focus was on one of the most striking findings that emerged from the SEnd-seq dataset. There are 658 pairs of neighbouring genes in *E. coli* that are oriented in a head-to-head manner (see Supplementary Fig. 15). Unexpectedly, it was discovered that two opposing TTSs frequently overlap with each other between a pair of convergent genes (284 of 658 pairs) (Fig. 4a–d and see Supplementary Table 4). In addition, 115 cases were found in which a TTS of an unopposed gene overlaps with that of an antisense RNA (Fig. 4b,d). These overlapping regions are largely hidden from the standard RNA-seq dataset due to its lack of coverage around RNA ends (Fig. 4a,b).

Overlapping bidirectional TTSs are, on average, ~80% efficient in both directions. The termination efficiency tends to be even higher for the sites that are sandwiched between two highly expressed genes (Fig. 4e). The length of the overlapping region ranges from 18 nt to 60 nt (Fig. 4f). The vast majority of these overlapping sequences are predicted to form RNA stem-loop structures (Fig. 4g–i). However, only a minor fraction (~16%) exhibit features of a canonical bidirectional intrinsic terminator[25,31], that is, a short GC-rich hairpin flanked by an A-tract and a U-tract on either side (Fig. 4j,k). Most overlapping regions feature a non-specific flanking sequence on at least one side of the hairpin. Moreover, the stems tend to be longer than those of typical intrinsic terminators, and often contain mismatches and bulges (see Supplementary Fig. 16). These bidirectional terminators do not appear to be primarily rho

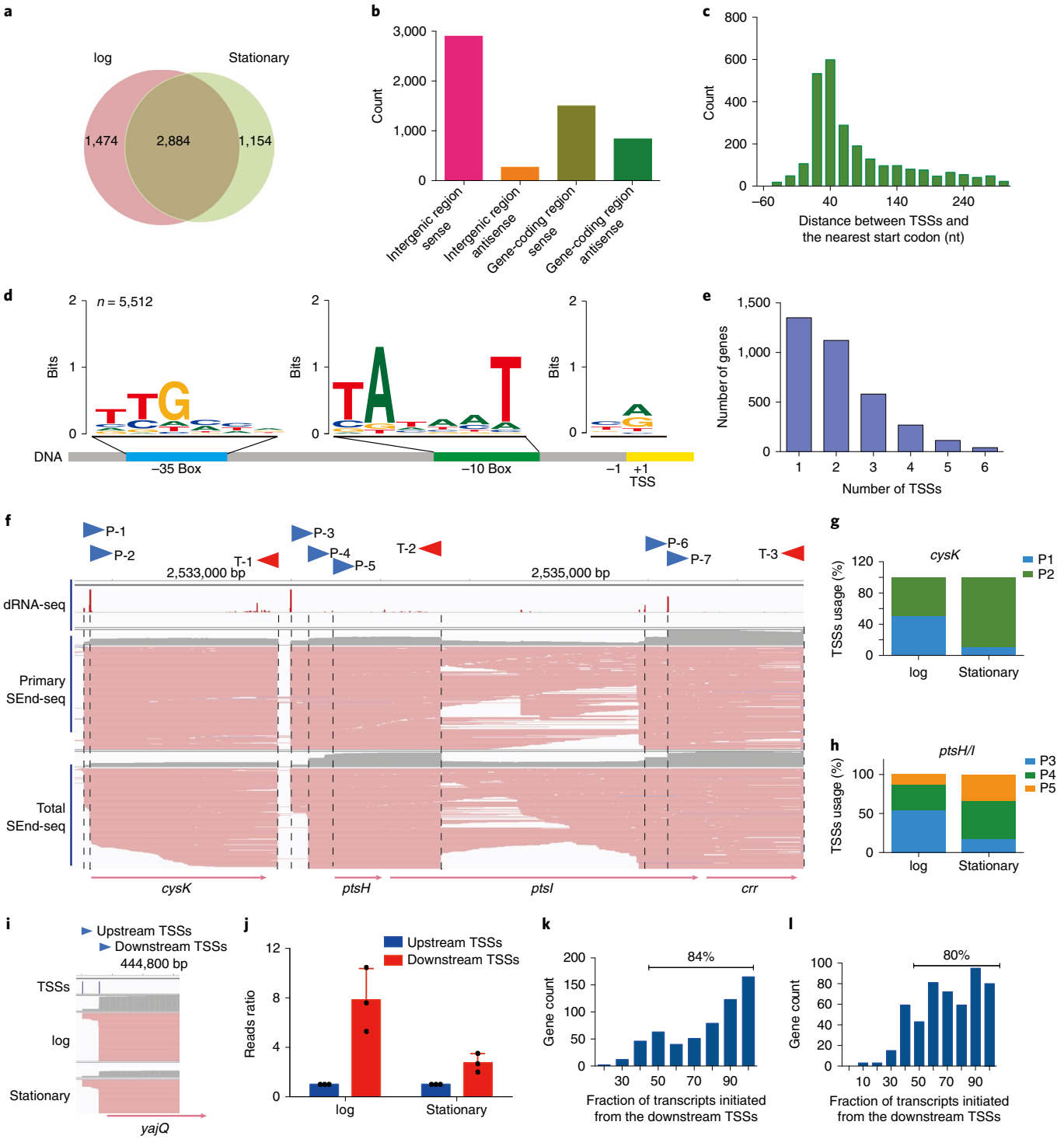

Fig. 2 | Identification of TSSs. a, Venn diagram showing the number of TSSs identified by SEnd-seq for *E. coli* cells growing in log phase versus stationary phase. **b**, Number of TSSs located within intergenic regions or inside annotated genes (in either the sense or the antisense orientation). **c**, Distribution of the distance between an identified TSS and the start codon of its nearest annotated coding region (cutoff is 300 nt). **d**, Motif analysis of the +1 site, −10 element and −35 element from all TSSs detected by SEnd-seq in log-phase *E. coli* cells. **e**, Distribution of the number of alternative TSSs for a given annotated gene. **f**, Log-phase SEnd-seq data track for the *cysK–ptsH–ptsI–crr* operon that shows multiple TSSs (P-1 to P-7) and TTSs (T-1 to T-3). TSSs identified by dRNA-seq are shown at the top for comparison[51]. **g**,**h**, Bar graphs displaying the differential usage of alternative TSSs for the *cysK* (**g**) and *ptsH/I* (**h**) genes during different growth stages. **i**, SEnd-seq data track showing two TSSs controlling the expression of the *yajQ* gene. **j**, Bar graphs displaying the number of *yajQ* transcripts initiated from the upstream versus the downstream TSSs. Values are normalized to the upstream TSS transcript level for each experimental replicate. Data are mean ± s.d. from three independent replicates. **k**,**l**, Histogram of the percentage of detected transcripts initiated from the most downstream TSS for any gene employing multiple TSSs that use cells harvested from the log phase (**k**) or stationary phase (**l**) of growth.

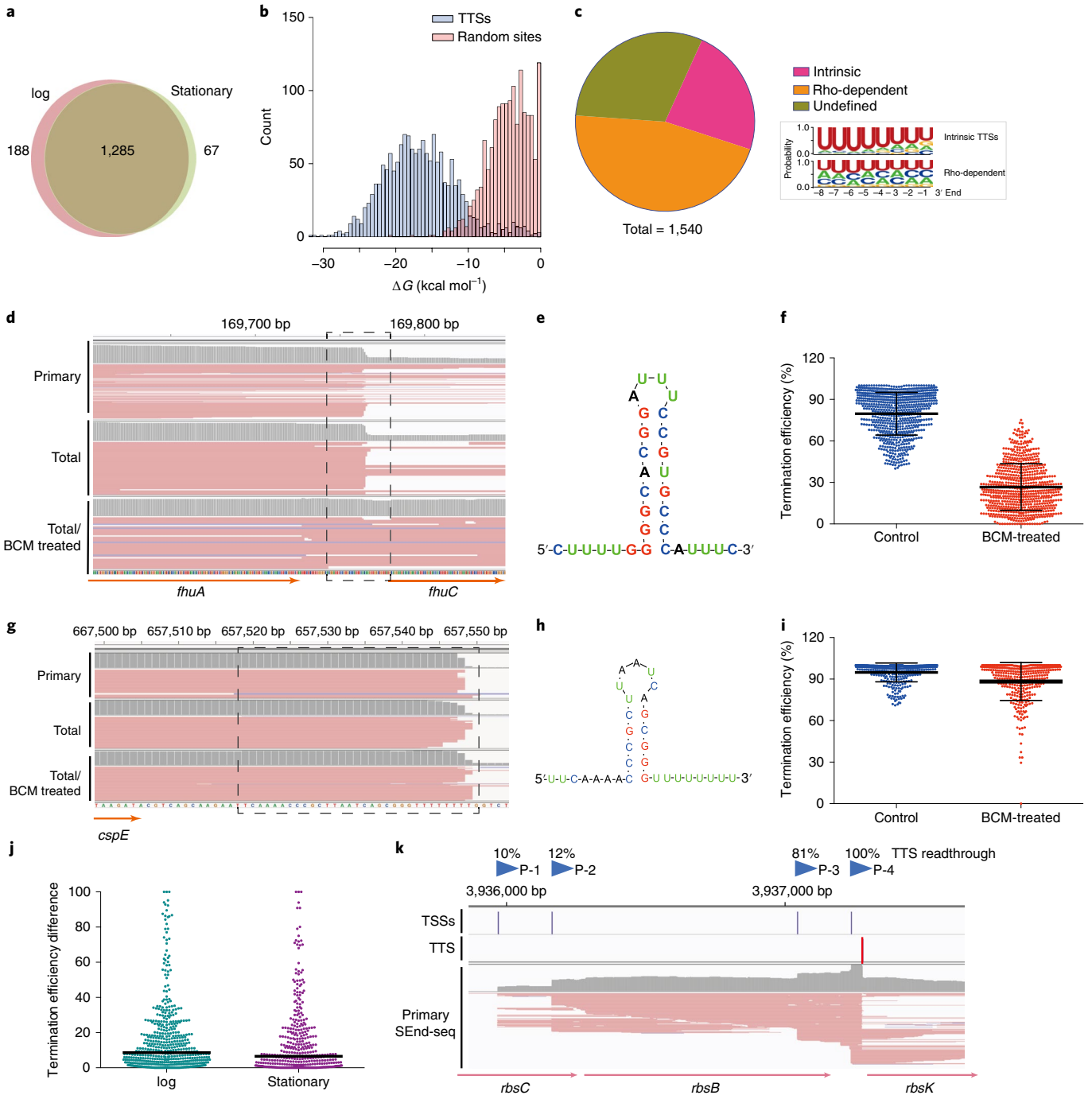

Fig. 3 | Identification of TTSs. a, Venn diagram showing the number of identified TTSs for log- versus stationary-phase *E. coli* cells. **b**, Distribution of the RNA-folding energy for identified TTS sequences (blue bars), compared with that for sequences of identical length randomly selected from the *E. coli* genome (red bars). **c**, Pie chart showing the fraction of intrinsic and rho-dependent terminators identified by SEnd-seq (left). Nucleotide profiles for the 3′ end sequences of intrinsic and rho-dependent TTSs (right). Data represent two independent experiments. **d**, SEnd-seq data track for an example rho-dependent terminator located downstream of the *fhuA* gene. When treated with the rho inhibitor BCM, the fraction of readthrough transcripts notably increased. **e**, Predicted secondary structure of the *fhuA* terminator. **f**, Average termination efficiency of all identified rho-dependent terminators without or with BCM treatment; $n = 709$ (number of terminators analysed). Error bars denote s.d. Data represent two independent experiments. **g**, SEnd-seq data track for an example intrinsic terminator located downstream of the *cspE* gene. **h**, Predicted secondary structure of the *cspE* terminator. **i**, Average termination efficiency of all identified intrinsic terminators without or with BCM treatment; $n = 357$. Error bars denote s.d. Data represent two independent experiments. **j**, Scatter plot showing the span of termination efficiency for each TTS that is linked to multiple TSSs. For example, a data point at 50% means that, for this TTS, the maximal termination efficiency and the minimal efficiency—depending on the choice of TSS—differ by 50%; $n = 520$ for the log-phase dataset and $n = 395$ for the stationary-phase dataset. The black bars indicate median values. **k**, An example SEnd-seq data track illustrating that the alternative usage of TSSs can induce differential termination efficiencies at the same TTSs. The fractions of readthrough transcripts initiated from any given TSS (P-1 to P-4) are indicated.

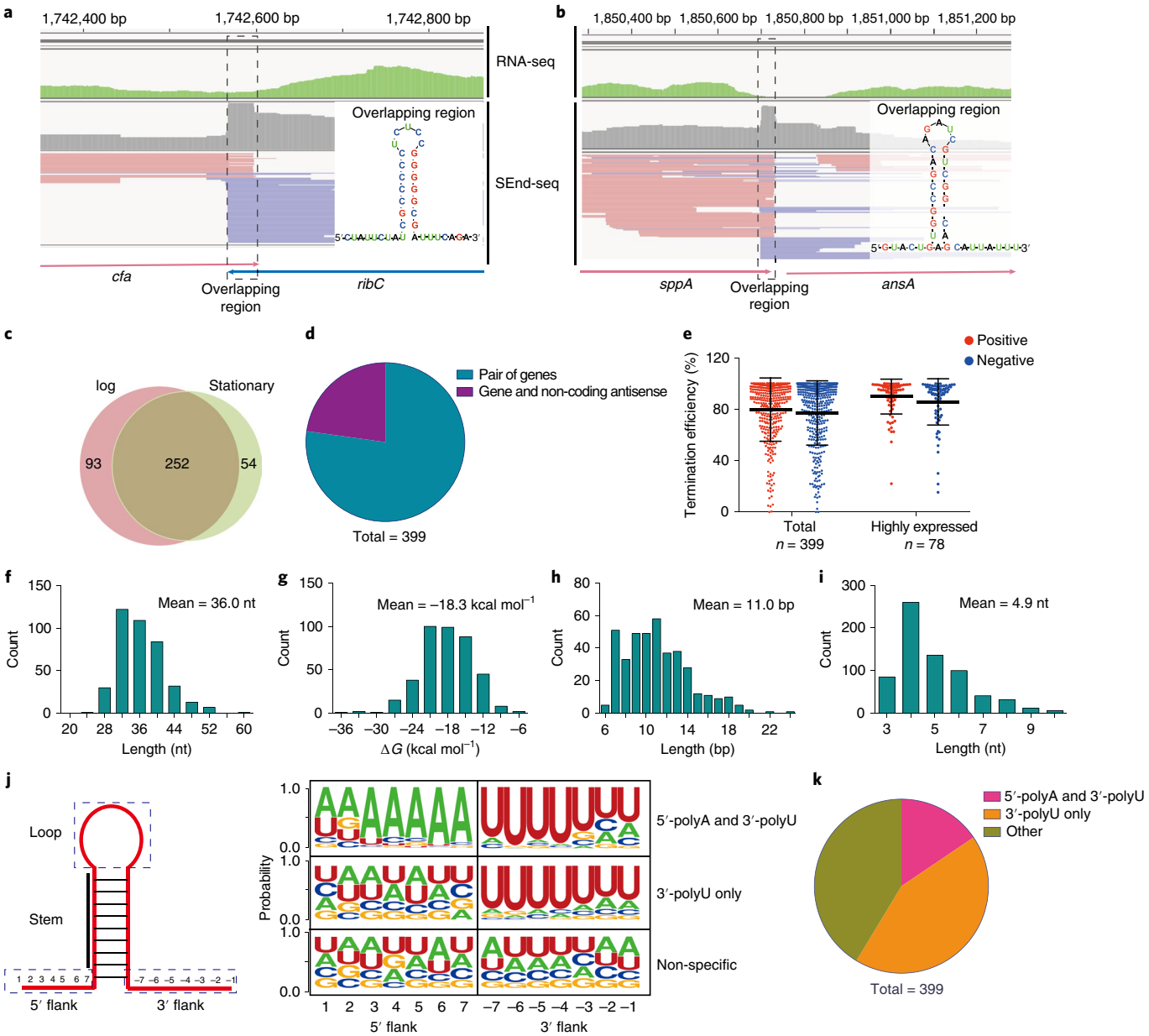

Fig. 4 | Pervasive bidirectional overlapping TTSs revealed by SEnd-seq. a, SEnd-seq data track for an example convergent gene pair (*cfa–ribC*) exhibiting overlapping TTSs. A standard RNA-seq data track is shown in green for comparison and the inset shows the predicted secondary structure for the overlapping region. Data represent three independent experiments. **b**, SEnd-seq data track and predicted secondary structure of an example overlapping TTS between a coding gene (*sppA*; red reads) and a non-coding antisense RNA (blue reads). Data represent three independent experiments. **c**, Venn diagram showing the number of overlapping bidirectional terminators identified for log- versus stationary-phase *E. coli* cells. **d**, Pie chart showing the fraction of overlapping TTSs located between either a gene pair or a gene and an antisense ncRNA. **e**, Average termination efficiency for all identified overlapping bidirectional terminators in either orientation (positive direction in red; negative direction in blue) (left); $n = 399$. Average termination efficiency for those bidirectional TTSs that are located between a pair of highly expressed genes (right); $n = 78$. Error bars denote s.d. Data represent two independent experiments. **f**–**i**, Distributions of the length (**f**), folding energy (**g**), predicted stem size (**h**) and loop size (**i**) for the overlapping TTS. **j**, Schematic of the stem-loop structure formed in the overlapping region (left). Nucleotide profiles for the 5′- and 3′-flanking sequences of the stem-loop within an overlapping region (right). Such profiling allows for classification of the overlapping TTSs into three categories. **k**, Pie chart showing the fraction of each category described in **j**.

dependent either, because the BCM inhibitor confers only a minor effect on their termination efficiency (see Supplementary Fig. 17).

The present study found that the patterns of these overlapping regions in the RNase-knockout strains (Δ*pnp*, Δ*rnb*, Δ*rnr*) are largely similar to those in the wild-type strain (see Supplementary Fig. 18), suggesting that the boundaries of these regions are genuine termination sites rather than products of RNase trimming. In further support of this notion, the overlapping sequences identified in the present study almost always contain single-stranded regions flanking the stem loop, unlike decay products that are usually processed until the edge of the protective hairpin stem[32].

Convergent transcription drives bidirectional termination in vitro. As neither intrinsic termination nor rho-mediated

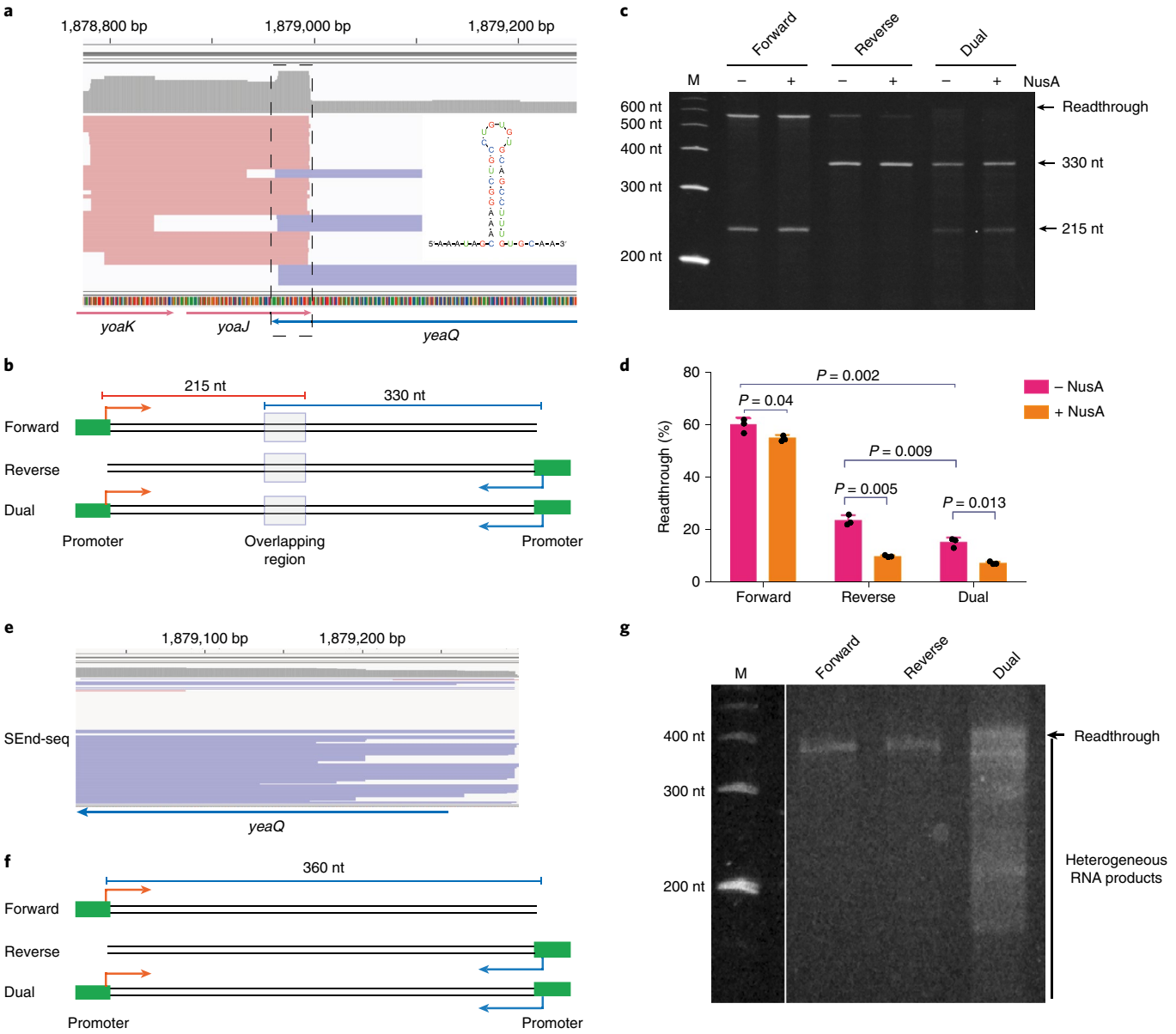

Fig. 5 | Convergent transcription is required for bidirectional termination in vitro. a, SEnd-seq data track for the *yoaJ–yeaQ* gene pair showing an overlapping TTS. Data represent three independent experiments. **b**, Schematic of DNA templates harbouring the *yoaJ–yeaQ* overlapping TTS regions that are used in the in vitro transcription assay. **c**, Gel showing the RNA products transcribed from the different templates shown in **b** in the absence or presence of NusA. Data represent three independent experiments. M, molecular weight marker. **d**, Quantification of the fraction of readthrough transcripts for the different templates. Data are mean ± s.d. from three independent experiments. *P* values were determined using two-sided, unpaired Student's *t*-tests. **e**,**f**, SEnd-seq data track for part of the *yeaQ* gene (**e**) and DNA templates derived from this region which lack a terminator sequence (**f**). The templates contain either one or two promoters to allow unidirectional or convergent transcription, respectively. **g**, Gel showing predominant readthrough for unidirectional transcription (forward and reverse templates) and heterogeneous RNA products for convergent transcription (dual template). Data represent three independent experiments.

termination can fully explain the widespread occurrence of overlapping TTSs between convergent TU pairs, we postulated that head-on collisions between opposing transcription machineries may cause termination in both directions. To test this hypothesis, we performed in vitro transcription assays with *E. coli* RNA polymerase (RNAP) on synthetic DNA templates harbouring a convergent gene pair. The genomic sequence around the *yoaJ–yeaQ* locus was copied into the template (Fig. 5a,b). This region contains a 34-nt overlapping TTS sequence and displays strong bidirectional termination in vivo. When a T7A2 promoter that controls transcription initiation by *E. coli* RNAP was placed at one end of the template, unidirectional transcription was permitted, which resulted in substantial readthrough (Fig. 5c,d). This result confirms the notion that the overlapping TTS sequence alone cannot cause efficient termination. In comparison, in vitro transcription using a strong intrinsic terminator yielded much lower readthrough (see Supplementary Fig. 19).

Importantly, when a promoter was incorporated into both ends of the template, to support convergent transcription, the readthrough level was notably reduced (Fig. 5c,d). The sizes of the RNA products are consistent with termination occurring at positions demarcating the overlapping region. Similar results were

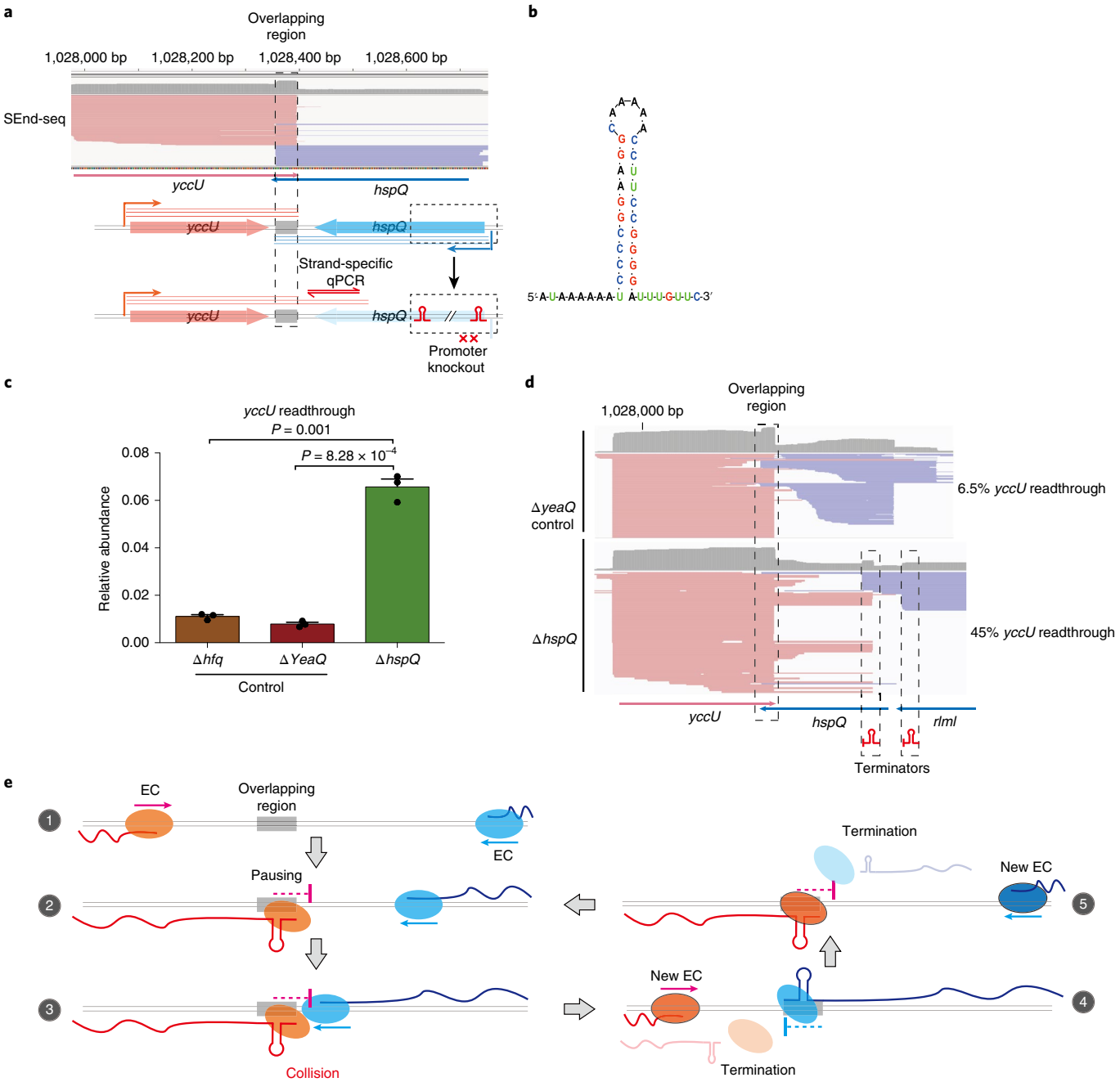

Fig. 6 | Convergent transcription contributes to bidirectional termination in vivo. a, SEnd-seq data track (top) and schematic of in vivo genomic modification (bottom) for the *yccU–hspQ*-convergent gene pair. To disrupt *hspQ* transcription, the promoter and part of the gene body of *hspQ* were replaced with two strong intrinsic terminators. Data represent three independent experiments. **b**, Predicted secondary structure for the overlapping TTS between *yccU* and *hspQ*. **c**, Quantitative PCR (qPCR) results showing the relative abundance of *yccU* readthrough transcripts across the overlapping region when *hspQ* transcription has been abolished (Δ*hspQ*). Genes outside the convergent pair were also edited using the same procedure (Δ*hfq* and Δ*yeaQ*) as controls. Data are mean ± s.d. from three independent experiments. *P* values were determined using two-sided, unpaired Student's *t*-tests. **d**, SEnd-seq data track around the *yccU–hspQ* region for the Δ*yeaQ* (top) or Δ*hspQ* strain (bottom). The fraction of *yccU* readthrough transcripts for each strain is indicated. Data represent two independent experiments. **e**, Model illustrating that head-on collisions between converging RNA polymerases drive bidirectional termination. The overlapping region produces an RNA hairpin that traps the transcription machinery, which is dislodged by another elongation complex (EC) travelling from the opposite direction—either through direct physical interaction or via torsional stress accumulated in the DNA. This process occurs repeatedly, resulting in highly efficient termination in both directions.

obtained with sequences taken from other convergent gene pairs (see Supplementary Fig. 20). These in vitro results strongly suggest that RNAP conflicts alone—without other cellular factors—can induce bidirectional termination.

The N utilization substance protein A (NusA) is known to stimulate bacterial transcription termination[33]. The influence of NusA on convergent transcription was examined and it was found that NusA further enhanced the bidirectional termination efficiency

(Fig. 5c,d and see Supplementary Fig. 20). Therefore, the effect of NusA and the effect of RNAP conflicts on the termination efficiency can be additive.

How do transcription complexes originating from stochastic initiation events always meet at the overlapping region? Given that the formation of RNA hairpins often contributes to RNAP pausing[34], it was posited that the stem-loop structures formed in the overlapping regions—although they do not lead to termination as such—cause RNAP to pause for an extended period of time such that another polymerase travelling from the opposite direction causes interference at the pausing site. To test this idea, in vitro transcription assays were conducted using DNA templates that lack an overlapping TTS sequence (Fig. 5e,f). As expected, unidirectional transcription yielded predominantly readthrough transcripts (Fig. 5g). It is interesting that, when convergent transcription was allowed, readthrough decreased, but the RNA products were heterogeneous in length (Fig. 5g), indicating promiscuous collision sites. This is in contrast to the uniform RNA products released from templates harbouring an overlapping TTS sequence (Fig. 5c). Therefore, the overlapping TTS sequence—and hence the pausing signal—is required to synchronize the converging transcription complexes, causing them to interfere with each other at well-defined positions.

Convergent transcription contributes to bidirectional termination in vivo. To seek further evidence that converging transcription elongation complexes contribute to their own termination inside the cell, in vivo genome editing was performed to disrupt transcription from one direction in an opposing gene pair. The *yccU–hspQ* convergent pair, which displays a 40-nt overlapping TTS (Fig. 6a,b), was targeted. To disrupt *hspQ* transcription, the Δ*hspQ* strain was created by deleting the promoter sequence for *hspQ* and inserting two strong intrinsic terminators around the original TSS of *hspQ*. Then the extent of *yccU* readthrough was assessed across the overlapping region using strand-specific quantitative PCR. As predicted, the Δ*hspQ* strain showed a notable increase in the abundance of *yccU* readthrough transcripts (Fig. 6c). Disrupting the transcription of other genes at distal genomic locations did not confer the same effect on *yccU* readthrough (Δ*hfq* and Δ*yeaQ* in Fig. 6c). Furthermore, SEnd-seq was performed with the Δ*hspQ* strain and the transcript profile around the *yccU–hspQ* region was examined (Fig. 6d). First, *hspQ* transcription was indeed abolished. Second, the *yccU* readthrough level markedly increased in the Δ*hspQ* dataset compared with the control dataset (45% versus 6.5%). Similar results were obtained from genome editing experiments on other convergent gene pairs (see Supplementary Fig. 21).

Together, these in vitro and in vivo results support a model in which the stem-loop structure formed near the 3′ ends of two converging TUs causes pausing of the elongation complex and, subsequently, transcription termination when an opposite elongation complex collides with it (Fig. 6e). This model predicts that RNAP occupancy is enriched at the overlapping bidirectional TTSs due to pausing. RNAP chromatin immunoprecipitation (ChIP)-seq experiments were thus performed using antibodies against the β- or β′-subunit. Indeed, stronger ChIP signals were observed around the overlapping TTS sites compared with nearby regions (see Supplementary Fig. 22).

Discussion

Despite the reinvigorated interest of the scientific community in RNA biology and the myriad RNA-seq technologies, methods capable of defining the boundaries of all transcripts in a transcriptome still remain scarce. Transcript isoform sequencing, which was developed to analyse eukaryotic transcript isoforms[15], ligates the termini of double-stranded DNA—as opposed to single-stranded DNA in SEnd-seq—and displays a strong bias towards short transcripts. Recently, a method based on PacBio long-read sequencing was reported[20]. But this method involves size-selection steps that remove any RNA shorter than 1,000 nt, and therefore is blind to all small RNAs and a substantial fraction of messenger RNA (mRNA). In contrast, SEnd-seq comprehensively profiles RNAs of different sizes in a single assay with reduced length bias. It is worth noting that the conversion from RNA to full-length cDNA in SEnd-seq is critically dependent on the performance of reverse transcription. A reverse transcriptase with enhanced processivity was used in the present study[35]. Continued enzyme engineering could further enhance the transcriptome coverage of SEnd-seq.

SEnd-seq enabled the determination of the correlated occurrence of TSSs and TTSs, so that it was possible to discern the crosstalk between promoters and terminators that control the same transcript. Future experiments are needed to elucidate the origin of such crosstalk. The method in the present study uses the sequences of 5′ and 3′ termini to infer the full-length composition of each distinct transcript. Thus, it is most ideally suited to studying organisms with limited splicing. SEnd-seq could also be employed for meta-transcriptomics analysis with RNA pooled from multi-species communities.

The sharp transcript boundaries defined by SEnd-seq led us to identify a widespread, but previously underappreciated, mechanism of transcription termination driven by head-on interference between transcription complexes. The unique ability of SEnd-seq to determine the 5′ end origin of terminated RNA and the full sequence of the overlapping region helped to uncover this mechanism. Transcriptional interference resulting from convergent promoters has been well documented in bacteria[36–38]. However, studies of transcriptional interference have thus far mainly focused on its negative impact on gene activity due to promoter occlusion or random RNAP collisions during elongation[39]. The present study shows that such interference can be exploited to precisely terminate transcription, thereby limiting undesired readthrough and fine-tuning the transcriptional output. Moreover, although overlapping bidirectional terminators have been reported for a few individual genes[40,41], the extent to which they occur genome wide was unexplored. In the present study this phenomenon was shown to be pervasive, which raises the intriguing scenario that head-to-head gene pairs are functionally related, akin to co-directional genes within the same polycistronic operon. In the cases where an opposing gene is absent, antisense transcription can also suppress the readthrough of sense transcription, which adds to the functional repertoire of non-coding RNA.

In the present study the strong T7A2 promoter was used for the in vitro transcription experiments, where efficient bidirectional termination was observed. In vivo, the likelihood of an RNAP head-on encounter is influenced by additional factors, notably the promoter strength[42]. For highly expressed convergent gene pairs, the frequent physical interference between RNAP is probably a major contributor to the bidirectional termination, although alternative, but not mutually exclusive, mechanisms cannot be excluded that may play a role in shaping the transcript 3′ boundaries, such as antisense RNA-mediated attenuation[43]. Moreover, given the known effect of ribosome movement on RNAP pause release[44,45], the uncoupling between transcription and translation downstream of the stop codon may enhance RNAP pausing and termination at intergenic bidirectional TTSs. With regard to RNAP collisions, further studies are needed to elucidate whether termination is induced by direct contacts between the converging motors or by the accumulation of torsional stress in DNA when they approach[46,47]. Finally, considering that convergent genes and polymerase conflicts are also found in eukaryotes[48–50], it will be interesting to investigate whether the transcription termination mechanism documented in the present study is conserved across kingdoms of life.

Methods

Bacterial strains and growth conditions. *E. coli* K-12 MG1655 and K-12 SIJ_488 (Addgene no. 68246; a gift from A. Nielsen) were cultured in lysogeny broth medium (10 g l^{-1} tryptone, 5 g l^{-1} yeast extract, 10 g l^{-1} NaCl, pH 7.4) under aerobic conditions at 37 °C. To inhibit rho activity, cells were cultured in lysogeny broth medium with 50 μg ml^{-1} BCM (Santa Cruz, sc-391755) at 37 °C for 15 min under the indicated growth conditions. *Δpnp*, *Δrnb* and *Δrnr* strains were generated using a previously reported protocol based on the arabinose-inducible lambda red recombineering system and the rhamnose-inducible flippase recombinase[52]. PCR primers listed in Supplementary Table 5 were used to amplify the kanamycin-resistant gene in pKD13 and the DNA product was transformed into the K-12 SIJ_488 strain. After selection for positive colonies, the inserted kanamycin-resistant gene was excised by culturing with L-rhamnose. To knock out a gene in a convergent gene pair, two strong intrinsic terminators were put into the insert DNA to replace the promoter region of the target gene.

SEnd-seq pipeline. *Cellular RNA isolation.* The overnight culture medium was diluted 1:50 into fresh medium and grown to an optical density OD_{600} of 0.4–0.6 for the log-phase sample or an OD_{600} >2.0 for the stationary-phase sample. *E. coli* cells were quenched by adding 0.5× vol. of cold stop buffer (5% phenol in ethanol) to the culture medium immediately before harvest, and placed on ice for 15 min. Cell pellets were collected by centrifugation (6,000*g* for 5 min at 4 °C), thoroughly resuspended in 100 μl of lysozyme solution (2 mg ml^{-1} in TE buffer (10 mM Tris-HCl and 1 mM ethylenediaminetetraacetic acid (EDTA))), and incubated for 2 min. The cells were then immediately lysed by adding 1 ml of TRIzol Reagent (Invitrogen, 15596) and subsequently pipetted vigorously until the solution was clear. After incubation for 5 min at room temperature, 200 μl of chloroform was added and the sample was gently inverted several times until it reached homogeneity. The sample was then incubated for 15 min at room temperature before centrifugation at 12,000*g* for 10 min. The upper phase (~600 μl) was gently collected and mixed at a 1:1 ratio with 100% isopropanol. The sample was incubated for 1 h at −20 °C and then centrifuged at 14,000*g* for 10 min at 4 °C. The pellet was washed twice with 1 ml of 75% ethanol, air dried for 5 min and dissolved in nuclease-free water. RNA integrity was assessed with 1% agarose gel and Agilent 2100 Bioanalyzer System.

3′-Adaptor ligation. RNA with or without 5′-adaptor ligation (see below) was subjected to 3′-adaptor ligation by mixing 12 μl of RNA (<5 μg) with 1 μl of 100 μM 3′-adaptor (see Supplementary Table 5), 0.5 μl of 50 mM ATP, 2 μl of dimethylsulfoxide (DMSO), 5 μl of 50% polyethylene glycol (PEG) 8000, 1 μl of RNase Inhibitor (New England BioLabs, M0314), 1 μl of High Concentration T4 RNA Ligase 1 (New England BioLabs, M0437) and 2.5 μl of T4 RNA Ligase Reaction Buffer. After incubation at 23 °C for 5 h, the reaction was diluted to 40 μl with water and purified twice with 1.5× vol. of Agencourt RNAClean XP beads (Beckman Coulter, A63987) to remove excess RNA adaptors. The sample was subsequently eluted in 12 μl of water.

Ribosomal RNA removal and reverse transcription. The eluted RNA was subjected to an optional step of rRNA removal with Ribo-Zero rRNA Removal Kit (Illumina, MRZB12424). The RNA was then recovered by ethanol precipitation. Eluted RNA, 11.5 μl, was incubated with 0.5 μl of 100 μM biotinylated reverse transcription primer (see Supplementary Table 5) and 1 μl of 10 mM deoxynucleotide solution mix (New England BioLabs, N0447) at 65 °C for 5 min, and then placed on ice for 2 min. Then, 1 μl of the maturase reverse transcriptase from *Eubacterium rectale* (recombinantly purified from *E. coli*; a gift from A.M. Pyle, Yale University)[35], 4 μl of 5× maturase buffer, 2 μl of 100 mM dithiothreitol (DTT) and 0.5 μl of RNase inhibitor were added to the reaction and incubated at 42 °C for 90 min. The reaction was then terminated by incubation at 85 °C for 10 min. After reverse transcription, 10 μl of 1 M NaOH solution was added and incubated at 70 °C for 15 min to remove the RNA templates. After neutralization by adding 10 μl of 1 M HCl solution, the reaction was diluted to 100 μl with TE buffer and cleaned twice with 100 μl of TE-saturated phenol:chloroform:isoamyl alcohol (25:24:1, v/v/v) (Thermo Fisher Scientific, 15593031). The cDNA was purified by ethanol precipitation, dissolved in TE buffer and cleaned once with 1.5× vol. of Agencourt AMPure XP beads (Beckman Coulter, A63881). The cDNA was then eluted with 30 μl of water and subjected to 5′-phosphorylation by adding 2 μl of T4 polynucleotide kinase (New England BioLabs, M0201), 4 μl of polynucleotide kinase reaction buffer and 4 μl of 10 mM ATP. After incubation at 37 °C for 60 min and 65 °C for 20 min, the cDNA was cleaned with 1.5× vol. of AMPure beads again and eluted with 20 μl of 0.1× TE buffer. The cDNA concentration was determined using the Qubit ssDNA Assay Kit (Invitrogen, Q10212).

Enrichment of primary transcripts. Primary transcripts were enriched following a protocol adapted from a previously published method[8]. Total RNA, 5 μg, was mixed with 5 μl of 10× VCE Buffer (New England BioLabs, M2080) in a total volume of 40 μl, incubated for 2 min at 70 °C, and then placed on ice. 3′-Desthiobiotin-GTP (5 μl; New England BioLabs, N0761) and 5 μl of vaccinia virus capping enzyme (New England BioLabs, M2080) were added to the reaction and incubated at 37 °C for 30 min. After purification with 1.5× RNAClean beads, the capped RNA was eluted and subjected to 3′-adaptor ligation as described above. The RNA was cleaned twice with 1.5× RNAClean beads and then enriched with Hydrophilic Streptavidin Magnetic Beads (New England BioLabs, S1421). After washing thoroughly four times with binding buffer (10 mM Tris-HCl, pH 7.5, 2 M NaCl, 1 mM EDTA) and three times with washing buffer (10 mM Tris-HCl, pH 7.5, 0.25 M NaCl, 1 mM EDTA), the RNA was eluted with 26 μl of biotin buffer (10 mM Tris-HCl pH 7.5, 0.5 M NaCl, 1 mM EDTA, 1 mM biotin) and incubated at 37 °C for 25 min on a rotator. Then 14 μl of binding buffer was added and incubated for another 4 min. The RNA was cleaned with 1.5× RNAClean beads and eluted in 12 μl of H_2O. The 5′-capped and 3′-ligated RNA was reverse transcribed by the maturase as described above.

Enrichment of processed transcripts. Processed RNA in a total RNA sample was selectively ligated to a 5′-adaptor. Briefly, 5 μg total RNA (12 μl) was incubated for 2 min at 70 °C and then placed on ice; 1 μl of 100 μM 5′-adaptor (see Supplementary Table 5), 0.5 μl of 50 mM ATP, 2 μl of DMSO, 5 μl of 50% PEG 8000, 1 μl of RNase Inhibitor, 1 μl of high concentration T4 RNA ligase 1 and 2.5 μl of T4 RNA Ligase Reaction Buffer were added to the sample. After incubation at 23 °C for 5 h, the sample was diluted with water and cleaned twice with 1.5× vol. of Agencourt RNAClean XP beads. After the SEnd-seq pipeline, a customized shell script was used to search for the adaptor-labelled reads, thereby specifically extracting processed RNA ends.

Circularization. Then, 50 ng cDNA (30 μl) was mixed with 2 μl of CutSmart Buffer (New England BioLabs, B7204), 2 μl of 50 mM $MnCl_2$, 2 μl of 0.1 M DTT, 2 μl of 5 M betaine (Affymetrix, 77507) and 2 μl of TS2126 RNA Ligase I (a gift from K. Ryan, City College of New York)[16]. The reaction was incubated at 37 °C for 5–16 h. Subsequently, the reaction was supplemented with 1 μl of 10 mM deoxynucleotide solution mix and diluted to 100 μl with TE buffer and 0.1% sodium dodecylsulfate. Then 100 μl of TE-saturated phenol:chloroform:isoamyl alcohol (25:24:1, v/v/v) was added and incubated for 1 h with occasional vortexing. After centrifugation, the water phase was cleaned again with phenol:chloroform:isoamyl alcohol. Finally, the circularized cDNA was ethanol precipitated and dissolved in 130 μl of TE buffer.

Library preparation. Circularized cDNA was fragmented by acoustic shearing in a microTUBE (Covaris, 520045) with a Covaris S220 focused-ultrasonicator under the condition of Peak145 for 90 s. After ethanol precipitation, the single-stranded DNA was converted into double-stranded DNA using the Second Strand cDNA Synthesis Kit (New England BioLabs, E6114) at 16 °C for 2 h. The product was cleaned with 1.8× vol. of AMPure beads and eluted in 50 μl of 0.1× TE buffer. The DNA ends were prepared and ligated to the Illumina sequencing adaptor with the NEBNext Ultra II DNA Library Prep Kit (New England BioLabs, E7645). The ligated product was cleaned twice with 1× vol. of AMPure beads and eluted in 50 μl of 0.1× TE buffer. Biotin-labelled DNA strands were bound to the Dynabeads M-280 Streptavidin (Invitrogen, 11205D) and cleaned four times with washing buffer (5 mM Tris-HCl, pH 7.5, 1 M NaCl, 0.5 mM EDTA) and twice with TE buffer. The beads were resuspended thoroughly with the Q5 High-Fidelity 2× Master Mix (New England BioLabs, M0492). The DNA library was then amplified for 13 (total RNA SEnd-seq) to 17 cycles (primary RNA SEnd-seq) following the manufacturer's protocol. The final library was cleaned twice with 1× vol. (50 μl) of AMPure beads, and its concentration and size distribution were determined using Agilent 2200 TapeStation (Agilent, 5067-5576).

Spike-in RNA preparation. A T7 promoter sequence was incorporated upstream of four DNA sequences with different lengths taken from the bacteriophage λ genome. After PCR amplification and gel excision/clean-up, the DNA templates were subjected to in vitro transcription by T7 RNA Polymerase (New England BioLabs, M0251). DNA was removed by adding 1 μl of TURBO DNase (Life Technologies, AM2238) and incubated at 37 °C for 15 min. Full-length RNA products were purified by polyacrylamide gel electrophoresis. After clean-up and concentration measurement, all spike-in RNA species were pooled together. Typically the spike-in RNA mix was added to the total bacterial RNA at a mass ratio of 1:300.

RNA-seq. For standard RNA-seq, ~5 μg RNA was treated with TURBO DNase and recovered by ethanol precipitation. Ribosomal RNA was depleted with the Ribo-Zero rRNA Removal Kit. The sequencing library was prepared with the TruSeq Stranded mRNA Library Prep Kit (Illumina, RS-122-2101) following the manufacturer's instructions.

RNAP ChIP-seq. The ChIP-seq workflow is adapted from a previously published ChIP-microarray study[53]. Briefly, cells were grown to the stationary stage and crosslinked by the addition of formaldehyde (1% final concentration) with continued shaking at 37 °C for 10 min before quenching with glycine (100 mM final concentration). Cells were then lysed and DNA was sheared by sonication, followed by treatment with micrococcal nuclease (New England BioLabs, M0247S) and RNase A (Thermo Fisher Scientific, EN0531). Antibodies against the RNAP β- or β′-subunit (BioLegend, 663903 or 662904) were used for immunoprecipitation. RNAP-DNA crosslinks were enriched by protein A/G beads (Thermo Fisher Scientific, 26159). Enriched immunoprecipitated DNA and input DNA-sequencing libraries were prepared using NEBNext Ultra II DNA Library Prep Kit.

Primer-extension assay. RNA (~5 µg) was treated with TURBO DNase, cleaned three times with phenol:chloroform:isoamyl alcohol (25:24:1, v/v/v) and recovered by ethanol precipitation. Subsequently the RNA was denatured at 70 °C for 2 min and then treated with terminator 5′-phosphate-dependent exonuclease (Illumina, TER51020) at 30 °C for 1 h. After ethanol precipitation, the recovered RNA was treated with RppH (New England BioLabs, M0356S) at 37 °C for 1 h. The RNA was cleaned by 1.5 × vol. of Agencourt RNAClean XP beads (Beckman Coulter, A63987) and ligated to a 5′-adaptor as described above. After reaction, the RNA was cleaned with 1.5 × vol. of Agencourt RNAClean XP beads. The eluted RNA was then reverse transcribed to cDNA with pooled reverse-transcribed primers by the maturase. Subsequently, 10 µl of 1 M NaOH solution was added and incubated at 70 °C for 15 min to remove the RNA templates. The second strand DNA was synthesized with an oligo complementary to the 5′-adaptor. The resultant double-stranded DNA was used for sequencing library preparation and sequencing was performed on MiSeq.

Data analysis. *Sequencing data collection and processing.* SEnd-seq data were collected using the Illumina MiSeq or NextSeq 500 platform in a paired-end mode (150 nt × 2). After quality filter and adaptor trimming, the paired-end reads were merged to single-end reads by using the FLASh software. The correlated 5′ end and 3′ end sequences were extracted using the custom script fasta_to_paired.sh. The full-length sequences were inferred by mapping to the reference *E. coli* genome NC_000913.3 using Bowtie 2. Reads with an insert length greater than 10,000 nt were discarded. For each sample we obtained over 2 million usable reads (that is, those harbouring at least 15 nt on each end of the same transcript). RNA-seq and ChIP-seq data were collected using the Illumina MiSeq or NextSeq 500 platform in a paired-end mode (75 nt × 2). After quality filter, the sequencing data were analysed using the Rockhopper software[54]. The wig files and SAM files were further analysed by customized Perl scripts. The results were visualized with the Integrative Genome Viewer.

Gene coverage quantification. For SEnd-seq data, each read was first mapped to the genome. Each position within the intervening region of the read (from the start site to the end site) was considered to be effective coverage. For RNA-seq data, the coverage of each nucleotide position was directly extracted from the wig files generated by the Rockhopper software[54]. Gene coverage was quantified by summing the coverage of all nucleotide positions spanned by each gene. Only genes longer than 200 nt were used for the correlation analysis between SEnd-seq and RNA-seq.

TSS identification. TSSs were identified from the primary transcript SEnd-seq data with a customized Perl script. Only positions with more than 10 reads starting at that position, and with an increase of at least 30% in read coverage from its upstream to its downstream, were retained. Candidate TSS positions within five bases in the same orientation were clustered together, and the position with the largest amount of read increase was used as the representative TSS position. Motif analysis around the TSS regions (−40 nt to +1 nt) was performed by MEME[55].

TTS identification. Based on previous work[56] and the observation that transcripts with intact, unprocessed 3′ termini are enriched in the primary RNA SEnd-seq dataset, it was reasoned that TTSs should be reproducible between the total RNA and primary RNA datasets. In practice, it was first identified from the total RNA SEnd-seq data positions with more than 10 reads, ending at that position (outside rRNA genes) and with a reduction of more than 40% in read coverage from its upstream to its downstream. Then the site in the primary SEnd-seq dataset was cross-checked, also with RNase-knockout strains (Δ*pnp*, Δ*rnb*, Δ*rnr*). Candidate TTS positions within five bases in the same orientation were clustered together, and the position with the largest amount of read reduction was used as the representative TTS position. Only the TTS sites identified from at least two samples were used for further analysis. The terminators are classified into rho-dependent terminators (those showing a readthrough percentage increase >30% on BCM treatment), intrinsic terminators (those showing a readthrough percentage <30% in the control sample, a readthrough percentage increase <15% on BCM treatment, and harbouring at least five uracils out of the eight bases in the 3′ flank region of the terminator hairpin), or undefined.

Overlapping bidirectional TTS identification. Overlapping bidirectional termination sites were identified by screening for two opposing TTSs using a customized Perl script. Only those with an overlapping region shorter than 60 nt and yielding a stem-loop structure were retained for further analysis. Highly expressed convergent gene pairs were defined as those with >20 read counts for each gene in the pair.

RNA secondary structure analysis. The sequence from 45 nt upstream to 9 nt downstream of an identified TTS was used for RNA secondary structure prediction with RNAfold[57] combined with customized Perl scripts.

Motif analysis. The −45 nt to +9 nt TTS regions and overlapping bidirectional TTS regions were used for motif analysis. Nucleotide logos around TTSs were generated by WebLogo[58].

Transcription unit annotation. Transcription units were identified using a customized Perl script based on the defined TSSs, TTSs and read coverage. Only those with a continuous coverage of more than five reads were retained for further analysis. Also excluded were units with a length shorter than 80 nt.

ChIP-seq data analysis. The RNAP ChIP-seq signal at each nucleotide position was calculated and normalized to the input sample data using a customized script. The normalized ChIP/input ratio was used for downstream analysis.

Previously deposited datasets. Differential RNA (dRNA)-seq datasets (SRR1411276 and SRR1411277 for log- and stationary-phase *E. coli* RNA, respectively)[19] and an SMRT-Cappable-seq dataset (accession number GSE117273)[20] were used for comparison with the SEnd-seq results from the present study.

In vitro transcription. DNA templates for T7 RNAP were amplified from the FLuc Control Template (New England BioLabs, E2040S). DNA templates for *E. coli* RNAP were prepared by PCR from the *E. coli* genomic DNA with indicated primer sets (see Supplementary Table 5). The T7A2 promoter sequence was incorporated at one or both ends of the template. Purified *E. coli* RNAP and sigma factor σ^{70} holoenzyme (a gift from the Darst Lab at the Rockefeller University) was used for in vitro transcription reactions. The reaction mixture included 4 µl of 5 × reaction buffer (200 mM Tris-HCl, 600 mM KCl, 40 mM $MgCl_2$, 4 mM DTT, 0.04% Triton X-100, pH 7.5 at 25 °C), 0.5 µl of RNase Inhibitor, 0.5 pmol DNA template, 2 pmol *E. coli* RNAP holoenzyme and nuclease-free water to a final volume of 18 µl. When applicable, 20 pmol NusA (a gift from the Landick Lab at the University of Wisconsin–Madison) was added to the reaction mixture. The mixture was incubated at 37 °C for 30 min before 2 µl ribonucleoside triphosphates (final 50 µM of each) were added to initiate transcription. After 5 min of reaction (unless noted otherwise), reinitiation of transcription was prevented by adding heparin (Sigma-Aldrich, H4784) to a final concentration of 100 µg ml^{-1}. After incubation with 0.3 µl of TURBO DNase for 10 min, the RNA was separated by 5% urea polyacrylamide gel electrophoresis, stained by SYBR Gold Nucleic Acid Gel Stain (Thermo Fisher Scientific, S11494), scanned by Axygen Gel Documentation System (Corning, GD1000) and quantified by ImageJ (National Institutes of Health).

Quantitative PCR. First-strand cDNA was reverse transcribed from the total RNA of indicated samples with the High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems, 4368813) and strand-specific reverse-transcribed primers (see Supplementary Table 5). Control cDNA was reverse transcribed with random primers and the same amount of input RNA. Quantitative reverse-transcribed PCR was performed using the SYBR Green PCR Master Mix (Applied Biosystems, 4309155) and QuantStudio 6 Flex Real-Time PCR System (Thermo Fisher Scientific). The relative abundance of RNA is represented as the signal ratio between the target transcript and the reference *rnpB* gene from the same sample using the formula: $2^{-(\Delta CT)}$ ($\Delta CT = CT_{target} - CT_{rnpB}$, where CT is cycle threshold).

Statistics. Data are shown as mean ± s.d. unless noted otherwise. *P* values were determined using the two-sided, unpaired, Student's *t*-tests with GraphPad Prism v.6. The difference between two groups was considered statistically significant when $P < 0.05$ (*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$; ****$P < 0.0001$; ns, not significant).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

SEnd-seq and standard RNA-seq datasets from this study have been deposited in the Gene Expression Omnibus (GEO) with the accession number GSE117737.

Code availability

The custom scripts used in this study are available on Github (https://github.com/LiuLab-codes/SEnd_seq_analysis). Other data that support the findings of this study are available from the corresponding author upon request.

Received: 3 January 2019; Accepted: 29 May 2019;
Published online: 15 July 2019

References

1. Morris, K. V. & Mattick, J. S. The rise of regulatory RNA. *Nat. Rev. Genet.* **15**, 423–437 (2014).
2. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
3. Sharma, C. M. et al. The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* **464**, 250–255 (2010).
4. Wurtzel, O. et al. A single-base resolution map of an archaeal transcriptome. *Genome Res.* **20**, 133–141 (2010).
5. Dar, D. et al. Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. *Science* **352**, aad9822 (2016).
6. Babski, J. et al. Genome-wide identification of transcriptional start sites in the haloarchaeon *Haloferax volcanii* based on differential RNA-Seq (dRNA-Seq). *BMC Genom.* **17**, 629 (2016).

7. Lalanne, J. B. et al. Evolutionary convergence of pathway-specific enzyme expression stoichiometry. *Cell* **173**, 749–761 (2018).
8. 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 Genom.* **17**, 199 (2016).
9. Matteau, D. & Rodrigue, S. Precise identification of genome-wide transcription start sites in bacteria by 5′-rapid amplification of cDNA ends (5′-RACE). *Methods Mol. Biol.* **1334**, 143–159 (2015).
10. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
11. Hor, J., Gorski, S. A. & Vogel, J. Bacterial RNA biology on a genome scale. *Mol. Cell* **70**, 785–799 (2018).
12. Guell, M., Yus, E., Lluch-Senar, M. & Serrano, L. Bacterial transcriptomics: what is beyond the RNA horiz-ome? *Nat. Rev. Microbiol.* **9**, 658–669 (2011).
13. Gama-Castro, S. et al. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res.* **44**, D133–D143 (2016).
14. Ruan, X. & Ruan, Y. Genome wide full-length transcript analysis using 5′ and 3′ paired-end-tag next generation sequencing (RNA-PET). *Methods Mol. Biol.* **809**, 535–562 (2012).
15. Pelechano, V., Wei, W. & Steinmetz, L. M. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* **497**, 127–131 (2013).
16. Lama, L. & Ryan, K. Adenylylation of small RNA sequencing adapters using the TS2126 RNA ligase I. *RNA* **22**, 155–161 (2016).
17. Lin-Chao, S., Wei, C. L. & Lin, Y. T. RNase E is required for the maturation of ssrA RNA and normal ssrA RNA peptide-tagging activity. *Proc. Natl Acad. Sci. USA* **96**, 12406–12411 (1999).
18. Ruff, E. F., Record, M. T. Jr. & Artsimovitch, I. Initial events in bacterial transcription initiation. *Biomolecules* **5**, 1035–1062 (2015).
19. Conway, T. et al. Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing. *mBio* **5**, e01442–14 (2014).
20. Yan, B., Boitano, M., Clark, T. A. & Ettwiller, L. SMRT-Cappable-seq reveals complex operon variants in bacteria. *Nat. Commun.* **9**, 3676 (2018).
21. Ray-Soni, A., Bellecourt, M. J. & Landick, R. Mechanisms of bacterial transcription termination: all good things must end. *Annu. Rev. Biochem* **85**, 319–347 (2016).
22. Hui, M. P., Foley, P. L. & Belasco, J. G. Messenger RNA degradation in bacterial cells. *Annu Rev. Genet.* **48**, 537–559 (2014).
23. Dar, D. & Sorek, R. High-resolution RNA 3′-ends mapping of bacterial Rho-dependent transcripts. *Nucleic Acids Res.* **46**, 6797–6805 (2018).
24. Zwiefka, A., Kohn, H. & Widger, W. R. Transcription termination factor rho: the site of bicyclomycin inhibition in *Escherichia coli*. *Biochemistry* **32**, 3564–3570 (1993).
25. Chen, Y. J. et al. Characterization of 582 natural and synthetic terminators and quantification of their design constraints. *Nat. Methods* **10**, 659–664 (2013).
26. Wang, X. et al. Processing generates 3′ ends of RNA masking transcription termination events in prokaryotes. *Proc. Natl Acad. Sci. USA* **116**, 4440–4445 (2019).
27. Goliger, J. A., Yang, X. J., Guo, H. C. & Roberts, J. W. Early transcribed sequences affect termination efficiency of *Escherichia coli* RNA polymerase. *J. Mol. Biol.* **205**, 331–341 (1989).
28. Telesnitsky, A. P. & Chamberlin, M. J. Sequences linked to prokaryotic promoters can affect the efficiency of downstream termination sites. *J. Mol. Biol.* **205**, 315–330 (1989).
29. Thomason, M. K. et al. Global transcriptional start site mapping using differential RNA sequencing reveals novel antisense RNAs in *Escherichia coli*. *J. Bacteriol.* **197**, 18–28 (2015).
30. Dornenburg, J. E., Devita, A. M., Palumbo, M. J. & Wade, J. T. Concerns about recently identified widespread antisense transcription in *Escherichia coli*. *mBio* **1**, e00106–10 (2010).
31. Peters, J. M., Vangeloff, A. D. & Landick, R. Bacterial transcription terminators: the RNA 3′-end chronicles. *J. Mol. Biol.* **412**, 793–813 (2011).
32. Dar, D. & Sorek, R. Extensive reshaping of bacterial operons by programmed mRNA decay. *PLoS Genet.* **14**, e1007354 (2018).
33. Mondal, S., Yakhnin, A. V., Sebastian, A., Albert, I. & Babitzke, P. NusA-dependent transcription termination prevents misregulation of global gene expression. *Nat. Microbiol.* **1**, 15007 (2016).
34. Zhang, J. & Landick, R. A two-way street: regulatory Interplay between RNA polymerase and nascent RNA structure. *Trends Biochem. Sci.* **41**, 293–310 (2016).
35. Zhao, C., Liu, F. & Pyle, A. M. An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. *RNA* **24**, 183–195 (2018).
36. Callen, B. P., Shearwin, K. E. & Egan, J. B. Transcriptional interference between convergent promoters caused by elongation over the promoter. *Mol. Cell* **14**, 647–656 (2004).
37. Horowitz, H. & Platt, T. Regulation of transcription from tandem and convergent promoters. *Nucleic Acids Res.* **10**, 5447–5465 (1982).
38. Elledge, S. J. & Davis, R. W. Position and density effects on repression by stationary and mobile DNA-binding proteins. *Genes Dev.* **3**, 185–197 (1989).
39. Shearwin, K. E., Callen, B. P. & Egan, J. B. Transcriptional interference—a crash course. *Trends Genet.* **21**, 339–345 (2005).
40. Sameshima, J. H., Wek, R. C. & Hatfield, G. W. Overlapping transcription and termination of the convergent *ilvA* and *ilvY* genes of *Escherichia coli*. *J. Biol. Chem.* **264**, 1224–1231 (1989).
41. Postle, K. & Good, R. F. A bidirectional rho-independent transcription terminator between the *E. coli tonB* gene and an opposing gene. *Cell* **41**, 577–585 (1985).
42. Sneppen, K. et al. A mathematical model for transcriptional interference by RNA polymerase traffic in *Escherichia coli*. *J. Mol. Biol.* **346**, 399–409 (2005).
43. Brantl, S. & Wagner, E. G. An antisense RNA-mediated transcriptional attenuation mechanism functions in *Escherichia coli*. *J. Bacteriol.* **184**, 2740–2747 (2002).
44. Landick, R., Carey, J. & Yanofsky, C. Translation activates the paused transcription complex and restores transcription of the trp operon leader region. *Proc. Natl Acad. Sci. USA* **82**, 4663–4667 (1985).
45. Proshkin, S., Rahmouni, A. R., Mironov, A. & Nudler, E. Cooperation between translating ribosomes and RNA polymerase in transcription elongation. *Science* **328**, 504–508 (2010).
46. Ma, J., Bai, L. & Wang, M. D. Transcription under torsion. *Science* **340**, 1580–1583 (2013).
47. Crampton, N., Bonass, W. A., Kirkham, J., Rivetti, C. & Thomson, N. H. Collision events between RNA polymerases in convergent transcription studied by atomic force microscopy. *Nucleic Acids Res.* **34**, 5416–5425 (2006).
48. Hobson, D. J., Wei, W., Steinmetz, L. M. & Svejstrup, J. Q. RNA polymerase II collision interrupts convergent transcription. *Mol. Cell* **48**, 365–374 (2012).
49. Prescott, E. M. & Proudfoot, N. J. Transcriptional collision between convergent genes in budding yeast. *Proc. Natl Acad. Sci. USA* **99**, 8796–8801 (2002).
50. Eszterhas, S. K., Bouhassira, E. E., Martin, D. I. & Fiering, S. Transcriptional interference by independently regulated genes occurs in any relative arrangement of the genes and is influenced by chromosomal integration position. *Mol. Cell Biol.* **22**, 469–479 (2002).
51. Creecy, J. P. & Conway, T. Quantitative bacterial transcriptomics with RNA-seq. *Curr. Opin. Microbiol.* **23**, 133–140 (2015).
52. Jensen, S. I., Lennen, R. M., Herrgard, M. J. & Nielsen, A. T. Seven gene deletions in seven days: fast generation of *Escherichia coli* strains tolerant to acetate and osmotic stress. *Sci. Rep.* **5**, 17874 (2015).
53. Peters, J. M. et al. Rho directs widespread termination of intragenic and stable RNA transcription. *Proc. Natl Acad. Sci. USA* **106**, 15406–15411 (2009).
54. McClure, R. et al. Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Res.* **41**, e140 (2013).
55. Bailey, T. L. et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
56. Celesnik, H., Deana, A. & Belasco, J. G. Initiation of RNA decay in *Escherichia coli* by 5′ pyrophosphate removal. *Mol. Cell* **27**, 79–90 (2007).
57. Lorenz, R. et al. ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
58. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).

Acknowledgements

We thank S. Darst and E. Campbell for help with the in vitro transcription experiments and critical reading of the manuscript, K. Ryan, A. Pyle and R. Landick for sharing reagents and E. Cheng for help with data analysis. This work was supported by a C.H. Li Memorial Scholar Fund Award (X.J.), the Robertson Foundation, the Quadrivium Foundation, a Monique Weill-Caulier Career Scientist Award, a March of Dimes Basil O'Connor Starter Scholar Award, a Kimmel Scholar Award, and National Institute of Health grants R00GM107365 and DP2HG010510 (S.L.).

Author contributions

S.L. conceived of and oversaw the project. X.J. performed the experiments and data analysis. D.L. contributed to the development of SEnd-seq workflow. S.L. and X.J. wrote the manuscript.

Competing interests

The Rockefeller University has filed a provisional patent application encompassing aspects of the SEnd-seq technology on which S.L. and X.J. are listed as inventors.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41564-019-0500-z.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to S.L.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Shixin Liu

Last updated by author(s): May 19, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	SEnd-seq and RNA-seq raw data were collected on Illumina Next-seq 500 or Mi-seq platforms at the Rockefeller University Genomics Resource Center.
Data analysis	After quality filter and Illumina sequencing adaptor trimming, the paired-end reads raw data were merged to single-end reads by using FLASh software (v1.2.11, T. Magoc et al., 2011). The correlated 5'-end and 3'-end sequences were extracted by the custom script fasta_to_paired.sh. The inferred full-length reads were mapped to the reference E. coli genome NC_000913.3 by using Bowtie2 (2.3.3.1, Langmead B et al., 2012), and reads with an insert length greater than 10,000 nt were discarded. The mapping results were visualized by the genome viewer IGV (v2.4.10, James T. Robinson et al., 2011). All downstream analyses were performed with custom perl scripts. SEnd-seq coverage was calculated per nucleotide position in the genome. Custom codes are available here: https://github.com/LiuLab-codes/SEnd_seq_analysis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

SEnd-seq data and standard RNA-seq data have been deposited at the GEO database under accession number GSE117737.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical test was used to predetermine sample size. Experiments for each condition were repeated two to four times.
Data exclusions	Low quality reads in the raw data and paired-end reads extracted from SEnd-seq analysis with alignment insert length more than 10,000 nt were discarded and not included in the results of this study.
Replication	All results described in this manuscript were reliably reproduced.
Randomization	RNA samples were collected under different growth conditions (O.D. 0.4-0.6 for log phase and O.D. >2.0 for stationary phase) and repetitive samples of the same condition were collected on different days to mimic randomization.
Blinding	Researchers were not blinded during experiments or data analysis since all of the findings are supported by quantitative data analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☐	☒ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Antibodies against RNAP β or β' subunit (BioLegend 663903 or 662904) were used for immunoprecipitation.
Validation	These commercial antibodies have been validated by the vendor and confirmed by several published studies [e.g., Peters JM, Mooney RA, Kuan PF, Rowland JL, Keles S, Landick R (2009) Rho directs widespread termination of intragenic and stable RNA transcription. Proc Natl Acad Sci USA 106: 15406–15411].

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	ChIP-seq data were deposited at the GEO database under accession number GSE117737.
Files in database submission	ChIP_seq_Stationary-Ab1_R1_rep1.fastq ChIP_seq_Stationary-Ab1_R1_rep2.fastq ChIP_seq_Stationary-Ab1_R2_rep1.fastq ChIP_seq_Stationary-Ab1_R2_rep2.fastq ChIP_seq_Stationary-Ab2_R1_rep1.fastq ChIP_seq_Stationary-Ab2_R1_rep2.fastq

ChIP_seq_Stationary-Ab2_R2_rep1.fastq
ChIP_seq_Stationary-Ab2_R2_rep2.fastq
ChIP_seq_Stationary_input_R1_rep1.fastq
ChIP_seq_Stationary_input_R1_rep2.fastq
ChIP_seq_Stationary_input_R2_rep1.fastq
ChIP_seq_Stationary_input_R2_rep2.fastq
ChIP_seq_Stationary-Ab1_rep1.wig
ChIP_seq_Stationary-Ab1_rep1_chip_vs_input.wig
ChIP_seq_Stationary-Ab1_rep2.wig
ChIP_seq_Stationary-Ab1_rep2_chip_vs_input.wig
ChIP_seq_Stationary-Ab2_rep1.wig
ChIP_seq_Stationary-Ab2_rep1_chip_vs_input.wig
ChIP_seq_Stationary-Ab2_rep2.wig
ChIP_seq_Stationary-Ab2_rep2_chip_vs_input.wig
ChIP_seq_Stationary_input_rep1.wig
ChIP_seq_Stationary_input_rep2.wig

Genome browser session
(e.g. UCSC)

Genome browser files (wig) (reference genome: NC_000913.3) were deposited at the GEO database under accession number GSE117737.

Methodology

Replicates

The ChIP-seq experiment was repeated twice with cells collected from the same growth stage.

Sequencing depth

All ChIP-seq samples were sequenced with the Mi-seq platform (paired-end 75 bp). On average about 2 million paired-end reads were collected for each sample.

Antibodies

Antibodies against the RNAP β (Ab1) or β' (Ab2) subunit (BioLegend 663903 or 662904) were used.

Peak calling parameters

We compared the signal intensities around the overlapping TTS sites to those at adjacent regions.

Data quality

After sequencing, only high-quality paired-end reads were used for mapping. Only appropriately mapped paired-end reads were used for downstream analysis.

Software

After read alignment with Bowtie2, the ChIP-seq signal coverage was calculated across the genome with custom Perl scripts, which are available here: https://github.com/LiuLab-codes/SEnd_seq_analysis.