Functionally uncoupled transcription–translation in *Bacillus subtilis*

Grace E. Johnson1,3, Jean-Benoît Lalanne1,2,3, Michelle L. Peters1 & Gene-Wei Li1

Tight coupling of transcription and translation is considered a defining feature of bacterial gene expression1–3. The pioneering ribosome can both physically associate and kinetically coordinate with RNA polymerase (RNAP)4–11, forming a signal-integration hub for co-transcriptional regulation that includes translation-based attenuation12,13 and RNA quality control2. However, it remains unclear whether transcription–translation coupling—together with its broad functional consequences—is indeed a fundamental characteristic of bacteria other than *Escherichia coli*. Here we show that RNAPs outpace pioneering ribosomes in the Gram-positive model bacterium *Bacillus subtilis*, and that this ‘runaway transcription’ creates alternative rules for both global RNA surveillance and translational control of nascent RNA. In particular, uncoupled RNAPs in *B. subtilis* explain the diminished role of Rho-dependent transcription termination, as well as the prevalence of mRNA leaders that use riboswitches and RNA-binding proteins. More broadly, we identified widespread genomic signatures of runaway transcription in distinct phyla across the bacterial domain. Our results show that coupled RNAP–ribosome movement is not a general hallmark of bacteria. Instead, translation-coupled transcription and runaway transcription constitute two principal modes of gene expression that determine genome-specific regulatory mechanisms in prokaryotes.

Transcription of mRNA in *E. coli* is accompanied by a closely trailing ribosome, whose ability to modulate the fate of the transcribing RNAP establishes a key paradigm for bacterial gene regulation. At operon leaders, ribosome pausing provides a signal for transcriptional attenuators, such as those of the *trp* and *his* biosynthetic operons13,14. Within coding regions, the proximity between RNAPs and ribosomes further enables the termination factor Rho to selectively abrogate transcription if a premature stop codon is present, thereby initiating a surveillance mechanism analogous to nonsense-mediated decay in eukaryotes2,15. If a premature stop codon is present, thereby initiating a surveillance mechanism analogous to nonsense-mediated decay in eukaryotes2,15. Rho also suppresses antisense transcription by pervasively targeting RNAPs that lack an accompanying ribosome16. As such, tight transcription–translation coupling is a fundamental feature of the central dogma in *E. coli* and related species.

However, the interplay between transcription and translation has been largely unexplored in other species, notably in the extensively studied Gram-positive bacterium *B. subtilis*. Previous work has shown that *B. subtilis* and *E. coli* often have markedly divergent regulatory mechanisms. For example, although many *B. subtilis* operons are regulated via transcriptional attenuators similar to their *E. coli* counterparts, the underlying mechanisms primarily rely on riboswitches or RNA-binding proteins, not a nearby ribosome14,17–19. In addition, the factors that monitor translation-coupled transcription in *E. coli*—Rho and its adaptor NusG—are dispensable in *B. subtilis*, with their absence causing only mild phenotypes19. Furthermore, nonsense-mediated polarity (Rho-dependent transcription termination of mRNAs containing premature stop codons) is thought to be rare in *B. subtilis*20–22.

In this work, we show that these differences stem from genome-wide ‘runaway transcription’, in which RNAPs transcribe far ahead of trailing ribosomes.

**RNAP outpaces the pioneering ribosome**

To determine whether RNAPs and pioneering ribosomes are kinetically coupled, we first used classic assays based on isopropyl β-D-1-thiogalactopyranoside (IPTG)-inducible lacZ2 adapted from *E. coli* (1,020 amino acids (aa)) (Supplementary Data 1) to measure the times of first appearance for its mRNA and protein4,8,23–25 (Fig. 1a). At each time point after induction, an aliquot of cell culture was added to a stop solution (phenol-ethanol for mRNA or chloramphenicol and erythromycin (for *B. subtilis* for protein). Because protein folding takes place after cell collection, it does not prohibit the measurement of the much faster translation kinetics4,8,23,25. If the pioneering ribosome were closely following the RNAP, as previously demonstrated4,8 and independently confirmed here for *E. coli* (Fig. 1b), the first full-length mRNA product would appear at the same time (τTX) as the first protein product (τTL). Unexpectedly, in *B. subtilis* there was a substantial delay between the protein and mRNA signals during rapid growth (τTL − τTX = 40 ± 4 s at a growth rate of 2 h⁻¹) (Fig. 1b). Whereas the translation time is similar to that measured in *E. coli* (τTL = 77 ± 2 s versus τTX = 78 ± 1 s in *E. coli*), the transcription time is much shorter (τTX = 37 ± 3 s versus τTX = 79 ± 1 s in *E. coli*). There was also a time delay during slow growth (τTL − τTX = 32 ± 1 s at a growth rate of 0.65 h⁻¹) (Extended Data Fig. 1). Thus, the RNAP ‘runs...
Fig. 1 | Fast RNAP movement results in runaway transcription. a, Schematic of inducible lacZ expression system in B. subtilis. Region probed by quantitative PCR with reverse transcription (qRT–PCR) is labelled in magenta. LacI, lacZ repressor; Spankp, modified lacZ promoter. b, Induction time courses of full-length lacZ mRNA (top, transcription) and protein (bottom, translation), measured by qRT–PCR and β-galactosidase assays, respectively. After the first appearance times (τL and τT), mRNAs accumulate quadratically with time whereas proteins accumulate quadratically with time. Lines indicate linear fits of transformed data after signals rise. Shaded regions indicate time difference between τL and τT. Uncertainties are s.e.m. among biological replicates (three for B. subtilis β-galactosidase assay, two for all others). c, Schematic of lacZa complementation reporter for endogenous genes. Endogenous 5′ UTR and gene are indicated in green. xylAp, modified xylose promoter. d, As in b, but for endogenous genes. Translation efficiencies for pycA and tkt are at the 50th and 93rd percentiles, respectively, among B. subtilis genes. Three biological replicates for pycA–lacZa β-galactosidase assay, two for all others. e, Table of first appearance times for truncated pycA constructs (schematic on left). Uncertainties are s.e.m. among two biological replicates. f, Left, schematic showing ribosome–RNAP distance: bottom right, estimated terminal ribosome–RNAP distance as a function of gene length. Elongation rates are based on e, and translation initiation times are assumed to be negligible. Histogram (top right) shows distribution of gene lengths in B. subtilis. See also Extended Data Figs. 1–3.

runaway' from the pioneering ribosome and the two do not reach the end of lacZ at the same time.

Runaway transcription occurs for endogenous B. subtilis genes as well as for the exogenous lacZ. To measure the first appearance times of endogenous mRNAs and proteins, we adapted an α-complementation-based strategy by attaching a short region encoding LacZα (100 aa) to the C terminus of endogenous genes23 (Fig. 1c). The fusion gene was placed under an IPTG-inducible promoter, while keeping the native B. subtilis ribosome binding site to maintain translation efficiency. After protein synthesis was inhibited, protein levels were estimated by allowing fusion proteins to complement the β-galactosidase activity of pre-expressed LacZα (Fig. 1c). The translation time of the full-length reporters (τT) was reproducibly longer than their transcription time (τL): τL − τT = 34 ± 2 s for the pycA–lacZα fusion (1,255 aa) and 16 ± 0.5 s for the shorter tkt–lacZα fusion (774 aa) (Fig. 1d). By the time the protein signal started to accumulate, mRNA levels had already increased by 5- to 20-fold (Fig. 1d). Different methods for cell collection and inhibition of translation reproducibly yielded the same time delay (Extended Data Fig. 2b). These results show that pioneering ribosomes lag far behind each RNAP, potentially owing to mismatches in their respective elongation rates.

To estimate the elongation rates of transcription and translation, we measured τL and τT for a truncated pycA–lacZα construct (255 aa) under the same promoter and native 5′ untranslated region (UTR). Using the length of truncation (ΔL) and the reduction in first appearance times (ΔτL and ΔτT), we obtained average transcription and translation elongation rates ($v_T^L$ and $v_T^T$, respectively) of 73 ± 2 nucleotides per second (nt s$^{-1}$) and 47 ± 3 nt s$^{-1}$, respectively (assuming the same transcription and translation initiation rates between long and truncated constructs) (Fig. 1e, Extended Data Fig. 2c). The transcription elongation rate for this region is nearly twice as fast as mRNA transcription elongation in E. coli4. By contrast, the translation elongation rate is similar to previous estimates in E. coli4,8,24,25. Together, these results show that RNAPs consistently outpace pioneering ribosomes in B. subtilis. The mismatch in elongation rates (Δμ = 26 ± 4 nt s$^{-1}$) creates gaps between RNAPs and the trailing ribosomes along nascent mRNAs, reaching about 360 nt after RNAP has transcribed a gene of around 1 kb (or a larger gap if translation initiation is slow) (Fig. If, Supplementary Discussion SN3).

The difference in transcription elongation rates between E. coli and B. subtilis could arise from a variety of sources. Because the lacZ transcription time differs between these species, the differential speed is unlikely to be due to nucleic acid sequences or secondary structures (Fig. 1b). We also ruled out contributions of non-essential RNAP subunits and transcription elongation factors, as B. subtilis cells lacking the α-subunit, the ε-subunit, GreA or NusG do not exhibit substantially increased transcription times (Extended Data Fig. 3, Supplementary Discussion SN4). It remains possible that the accelerated speed in B. subtilis is driven by differences in the RNAP pause signals26 or other components of the RNAPα11 (Extended Data Fig. 4, Supplementary Discussion SN5). Irrespective of the underlying driver, runaway transcription and the long ribosome-free nascent mRNA indicate that the
The stop-to-stem distances \( \text{patA} \) (80th percentile for translation efficiency). The stop-to-stem distances \( d \) for the native and extended constructs are 26 nt and 14 nt, respectively. b, c. Northern blots against readthrough isoforms (top) and control for \( \text{pugP} \) expression (bottom) for constructs indicated in a, ND, not detected. For gel source data, see Supplementary Fig. 1. Northern blotting was performed twice for \( \text{T2} \) (top) and control for \( \text{T2} \), terminator (99.97% termination efficiency); \( \text{T1} \), translation efficiency). tkt \( x \), (80th percentile for \( \text{pugP} \) translated. When we forced translation through a strong intrinsic terminator, we observed a significant reduction in translation efficiency compared to the control. d. Example of a Rho-terminated asRNA (cssS\( x \)). Rend-seq data for wild-type and \( \Delta \rho \) cells show region of potential termination sites (orange: 5′-end mapped reads, blue: 3′-end mapped reads), d. Quantification of mRNA levels with variants of cssS\( x \) insertions (with seven mutations to replace in-frame stop codons with sense codons, see Supplementary Data I) for sequence). Relative mRNA expression measured as in a, e. Quantification of C:G ratios (number of C residues divided by number of G residues) in 100 nt moving windows of cssS\( x \). See also Extended Data Figs. 6–9, Supplementary Data 3.

**B. subtilis** genome must have different rules for co-transcriptional regulation than those of \( E. coli \).

**RNAP is insensitive to translation**

A major prediction of runaway transcription is that \( B. subtilis \) RNAPs should be insensitive to translation, in contrast to the pervasive use of translation-controlled termination of transcription in \( E. coli \). Consistent with this prediction, intrinsic terminator transcription in \( B. subtilis \) were effective even when the entire terminator hairpin was translated. When we forced translation through a strong intrinsic terminator downstream of a highly translated gene (\( \text{pugP} \)) by replacing its stop codon (Fig. 2a, Extended Data Fig. 5a), transcription readthrough remained undetectable (Fig. 2b). By contrast, the same open reading frame (ORF)-extended construct in \( E. coli \) showed completely abrogated terminator activity, consistent with the current paradigm that the closely trailing ribosomes blocks terminator hairpin formation (Fig. 2c). Thus, unlike \( E. coli \), the pioneering ribosomes in \( B. subtilis \) do not strongly modulate intrinsic transcription termination signals (except under some circumstances\(^{29–39} \)). This fundamental difference may contribute to the divergence of regulatory mechanisms between these species in two major ways. First, it enables intrinsic terminators to generate functional nonstop mRNAs for highly translated genes in

**Rho-termination has alternative roles**

This pervasive runaway transcription indicates that the physiological roles of Rho in \( B. subtilis \) do not involve surveillance of nascent mRNA translation, as this would render most transcription events unproductive. Consequently, \( B. subtilis \) may lack an important

---

**Fig. 2** Lack of translational control on transcription. a. Schematics of ORF-extension constructs and controls for pugP (80th percentile for translation efficiency). T1, pugP terminator (99.97% termination efficiency); T2, sodA terminator (99.9% termination efficiency). Stars indicate mutations. The stop-to-stem distances \( d \) for the native and extended constructs are 26 nt and 14 nt, respectively. b, c. Northern blots against readthrough isoforms (top) and control for pugP expression (bottom) for constructs indicated in a, ND, not detected. For gel source data, see Supplementary Fig. 1. Northern blotting was performed twice for B. subtilis (biological replicates) and once for E. coli. Results for both species were independently confirmed (biological replicates) by qRT–PCR (Methods). d. Examples of terminator stem-loops overlapping with stop codons (\( \text{patA} \) d = –12 nt, ispg d = –5 nt), Peaks in Rend-seq data show sites of termination. Terminator stems are highlighted. Stop codons are indicated in red. Translation efficiencies for \( \text{patA} \) and ispg are at the 63rd and 90th percentiles, respectively, in B. subtilis. e. Genome-wide distribution of stop-to-stem distances \( d \) (inset) for high-confidence intrinsic terminators in B. subtilis (top, \( n = 1,228 \)) and E. coli (bottom, \( n = 409 \)). ORF-overlapping terminators (\( d \leq 0 \) nt) are in dark magenta, and ribosome-overlapping terminators (\( d \leq 12 \) nt) are in medium and dark magenta, with the respective fraction of terminators indicated. See also Extended Data Fig. 5, Supplementary Data 2.

---

**Fig. 3** Signals of Rho-dependent termination. a. Quantification of mRNA levels with and without premature stop codons (indicated by \( x \)). mRNA levels quantified by qRT–PCR for the lacZ region (blue) relative to gyrA. Comparison of mRNA levels between cells without Rho (green) and wild-type cells (WT) (magenta) is shown. b. Distributions of mRNA level changes between two wild-type replicates (magenta) and between wild-type and \( \Delta \rho \) (green) as measured by Rend-seq. Expression changes for asRNAs nested within operons (Extended Data Fig. 6) and pseudogenes (Extended Data Figs. 7, 8) are indicated below, \( n \), number of cases. Box plots are defined by median, 25th and 75th percentiles. c. Example of a Rho-terminated asRNA (cssS\( x \)). Rend-seq data for wild-type and \( \Delta \rho \) cells show region of potential termination sites (orange: 5′-end mapped reads, blue: 3′-end mapped reads), d. Quantification of mRNA levels with variants of cssS\( x \) insertions (with seven mutations to replace in-frame stop codons with sense codons, see Supplementary Data I for sequence). Relative mRNA expression measured as in a, e. Quantification of C:G ratios (number of C residues divided by number of G residues) in 100 nt moving windows of cssS\( x \). See also Extended Data Figs. 6–9, Supplementary Data 3.

**Bacillus**\(^{12} \), such as the mRNA for an alternative ribosome rescue factor whose analogue in \( E. coli \) is generated post-transcriptionally by RNase III cleavage\(^{13,14} \). Second, the lack of coupling is likely to lead to avoidance of ribosome-controlled transcriptional attenuators, instead favouring the riboswitch- or protein-based mRNA leaders that are widely observed in \( B. subtilis \)\(^{14,27,28} \).

The large RNAP–ribosome spacing in \( B. subtilis \) also markedly influences the positions of intrinsic transcription terminators at the ends of operons. Using end-enriched RNA-seq (Rend-seq) (Supplementary Discussion SNI) to map active intrinsic terminators\(^{30} \), we found that many terminator stem-loops directly overlapped with stop codons of the last genes of operons (107 out of 1,228 with ‘stop-to-stem distance’ of 0 nt or less) (Fig. 2d), a configuration that would cause antitermination in \( E. coli \) by the pioneering ribosomes\(^{12} \). Furthermore, the majority (72%) of intrinsic terminators in \( B. subtilis \) are positioned within half a ribosome footprint downstream of the stop codon (stop-to-stem distance 12 nt or less) (Fig. 2e, Supplementary Data 2). In \( E. coli \), only 24% of intrinsic terminator stems are within 12 nt of the stop codon, and these have significantly reduced termination activity compared to other terminators (\( P < 10^{-4} \)) (Extended Data Fig. 5b–e). By contrast, the gene-proximal terminators in \( B. subtilis \) do not show significantly weaker termination (\( P > 0.3 \)) (Extended Data Fig. 5f–i), indicating a lack of translational interference in terminator hairpin formation. These results further demonstrate that most operons in \( B. subtilis \) are transcribed without a closely trailing ribosome.
We considered how, then, Rho targets specific RNAs for termination in *B. subtilis*. Although Rho does not affect most mRNA transcription, it selectively terminates several operons and removes many asRNAs. Without being influenced by translation, Rho-termination may be solely dependent on *cis*-encoded elements. Indeed, when we forced translation of a 678-nt Rho-terminated asRNA (antisense to cssS, designated cssS<sup>AS</sup>) (Fig. 3c) by inserting it into the reading frame of the *tkt–lacZα* mRNA reporter (with the seven *cis*-encoded elements, including regions of high C-to-G ratios, in the *tkt* reading frame replaced by sense codons (Supplementary Data 1)), Rho activity was not abrogated (Fig. 3d). The more than 100-fold decrease in downstream mRNA level was restored by deletion of *rho* (Fig. 3d). Rho-termination is specified by a 339-nt window within a high C-to-G ratio, a feature of Rho utilization (Fig. 3d). Thus, the sequence of this asRNA alone was sufficient to promote Rho termination, independent of genomic location or transcription. More broadly, we found that C-to-G ratio distinguishes Rho-terminated mRNAs and asRNAs from those not terminated by Rho (Extended Data Fig. 9). Together, these results support our prediction that Rho-termination is mechanistically independent of translation in *B. subtilis*. Instead, we propose that Rho is guided by strategically placed *cis*-elements, including regions of high C-to-G ratios, in the *B. subtilis* genome to prevent both pervasive transcription in the antisense direction and premature termination in the sense direction.

---

**Fig. 4** | Phylogenomic distribution of uncoupling. **a** Phylogenetic tree (centre) is overlaid with greyscale heat map representation of the distributions of stop-to-stem distances (*d*) for each species (middle ring, range in *d* shown from −20 to 120 nt). Full and dashed lines mark *d* = 0 nt and *d* = 12 nt, respectively. The species-specific fraction *F* of high-confidence terminators with *d* ≤ 12 nt is shown in the outer ring. Number of species per phylum with at least 30% of terminators with *d* ≤ 12 nt is indicated under the phylum name. Species without Rho homologues are marked with lines next to the tree (grey, no homologue; red, partial homologue or pseudogene). The 1,434 representative or reference genomes (with RefSeq are included. Tandem terminators are excluded. See Extended Data Fig. 10, Supplementary Data 4, 5, Methods. Insets (*L. monocytogenes*, *n* = 705 identified terminators, *F* = 71.2% of terminators with *d* ≤ 12 nt; *P. aeruginosa*, *n* = 216, *F* = 6.9%) show representative examples of bioinformatically determined stop-to-stem distributions (compare with Fig. 2e) with their heat map representation (above) shown in the middle ring. Dark and light portions of the histograms in insets highlight terminators with *d* ≤ 12 nt and *d* > 12 nt, respectively. **b, c**, Schematics of transcription-coupled (b) and runaway transcription (c) and some of their functional consequences.
Runaway transcription across eubacteria

Last, we evaluated the prevalence of runaway transcription in other bacterial species. Using the short distance between intrinsic terminator hairpins and the preceding stop codons as a conservative signature for lack of kinetic coupling, we systematically annotated the positions of intrinsic terminators for sequenced bacterial genomes by developing a computational classifier with a low false discovery rate (1% or less) (Extended Data Fig. 10a-c, Supplementary Data 4, 5, Supplementary Discussion SN2). The phylogenetic tree bifurcates into phyla with either many or few short-distance terminators (stop-to-stem 12 nt or less) (Fig. 4). More than half of Firmicutes (182 out of 358 analysed), which many or few short-distance terminators (stop-to-stem 12 nt or less) (Extended Data Fig. 10d, e). Coincidentally, most Firmicutes are resistant to the Rho inhibitor bicyclomycin38 and many lack Rho altogether39, signifying the diminished role of Rho in this phylum (Fig. 4a). We note that the stop-to-stem distance analysis has limited power for some species within Firmicutes that have few identifiable intrinsic terminators, such as Mycoplasma pneumoniae, in which coupling has been reported40 and no short-distance terminators were found (Extended Data Fig. 4b, c). Other distant clades of Gram-negative bacteria (for example, Campylobacterota and Thermotogota) also contain a substantial fraction of short-distance terminators (Fig. 4a), although transcription termination in these species has not been well characterized. By contrast, Actinobacteria, Bacteroidetes, and different phyla of Proteobacteria have most terminators far from the stop codon (Fig. 4a). The prevalence of stop–terminator overlap across diverse bacterial species suggests that runaway transcription is a common feature in distinct phyla and is not unique to B. subtilis.

Conclusion

Our understanding of prokaryotic gene regulation has been guided by pioneering examples such as the E. coli trp and his operons, which shed light on an intimate relationship between transcription and translation40 (Fig. 4b). However, this study shows that in B. subtilis, transcription and translation are fundamentally disjointed. A much faster RNAP speed not only helps to explain the different co-transcriptional regulatory strategies in Firmicutes (Fig. 4c), but also raises important questions regarding mRNA quality control in these bacteria. For example, understanding how B. subtilis and related species tolerate the accumulation of aberrant transcripts with premature stop codons could provide insights into proteome homeostasis and the evolution of new gene functions. More broadly, our results illustrate how a simple kinetic property of the central dogma can markedly transform the regulatory genome.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2638-5.
Methods

Strains

*B. subtilis* strains were generated from 168 (bearing the trpC2 mutation) and *E. coli* strains from MG1655. Linearized plasmids, gDNA, and linear PCR products were transformed into 168 using standard protocols relying on natural competence. Sequences of plasmids and integrations were confirmed by Sanger sequencing. All strains are listed in Supplementary Data 6. All plasmids were generated in *E. coli* DH5a cells (except for ORF extension experiments in *E. coli*) using standard protocols and are listed with additional details in Supplementary Data 6.

List of oligonucleotides

Supplementary Data 7 contains the list of oligonucleotides used for strain construction, qRT–PCR primers, and probes for northern blots.

High-throughput expression datasets used

The following published gene expression (ribosome profiling and Rend-seq) datasets were used in the present work: ref. [41](accession GSE33767) for ribosome profiling of *E. coli* in rich defined medium; ref. [35](accession GSE95211) for ribosome profiling of *B. subtilis* grown in lysogeny broth (LB); Rend-seq of *B. subtilis* grown in LB (wild-type, ΔppA, Δrhol), Rend-seq of *E. coli* grown in rich defined medium (wild-type, ΔppA, Δrhol), Rend-seq of *Vibrio natriegens* grown in MOPS complete medium + 3% NaCl, and Rend-seq of *Caulobacter crescentus* grown in peptone-yeast extract medium (PYE); ref. [34](accession GSE108295) for Rend-seq of *Staphylococcus aureus* grown in tryptic soy broth (TSB).

Cell growth

To measure transcription and translation kinetics, pre-cultures were started from single colonies picked into 5 ml LB and grown at 37 °C. After 3 h, cultures were diluted to OD600 = 0.005 in 200 ml pre-warmed LB in a 1-l beaker. For a-complementation assays, xylose was also added to 200 ml LB to 2% w/v. Cultures were grown with vigorous shaking (>200 rpm) at 37 °C. The experiment was performed at OD600 = 0.2. For measurements of transcription and translation kinetics in slow growth, cells were grown as above in MOPS minimal medium + 0.4% maltose (100 ml 10× MOPS mixture, 10 ml 0.132 M K2HPO4, 20 ml 10% glutamate, 10 ml 10 mg/ml tryptophan, 80 ml 5% maltose, 780 ml water).

For northern blot analysis of *B. subtilis* samples, cells were grown in LB at 37 °C with vigorous shaking. *E. coli* samples were grown with additional 50 μg/ml kanamycin and 200 mM anhydrotetracycline to maintain and induce expression from the pSC101-derived plasmid. Pre-cultures were started in pre-warmed LB from fresh colonies. Once OD600 reached 0.1, cultures were diluted in the fresh, pre-warmed medium to OD600 = 2 × 10^4, and collected when OD600 = 0.3.

For measurements of Rho termination (Fig. 3a, d), single colonies were picked into 5 ml LB and grown for 3 h at 37 °C with shaking. Cultures were back diluted to OD600 = 0.005 into prewarmed 20 ml LB and 1 mM IPTG and grown at 37 °C with vigorous shaking. The experiments were performed at OD600 = 0.2.

Measurements of protein induction kinetics

Protein induction measurements were as described^8,21,41 with modifications and are detailed below.

Cells were grown as described above until they reached OD = 0.20–0.30. Three 1-ml pre-induction cultures were collected into ice-cold stop solution and mixed vigorously before being placed on ice. Expression was then induced by mixing the remaining culture with 10 ml IPTG in pre-warmed, shaking LB, such that the final IPTG concentration was 5 mM. Following a brief delay (depending on the length of the construct), 1 ml of culture was collected every 5–10 s for a total of 12 time points. Cells were collected into ice-cold stop solution and mixed vigorously before being placed on ice. The stop solution was 50 μl 20 mg/ml chloramphenicol for *E. coli* and 100 μl buffer containing 10 mg/ml chloramphenicol and 10 mg/ml erythromycin for *B. subtilis*.

For measurements using full-length lacZ, collected cells were pelleted at 4 °C at 18,213g for 4 min and flash frozen in liquid nitrogen. Cell pellets were stored at −80 °C until use. For *B. subtilis* measurements, cells were resuspended in 1 ml Z-buffer containing chloramphenicol, permeabilized with 15 μl toluene and vortexed. Cells (50 μl) cells were then added to 450 μl Z-buffer. For *E. coli* measurements, cells were resuspended in 1.3 ml Z-buffer containing chloramphenicol and 200 μl cells were then added to 300 μl Z-buffer. For measurements using lacZ fusions, collected cultures were incubated at 37 °C for 1 h to allow complementation between LacZo and LacZa. Cells were pelleted at 4 °C at 18,213g for 4 min and resuspended in 500 μl Z-buffer and chloramphenicol, permeabilized with 7.5 μl toluene and vortexed.

Beta-galactosidase activity was measured using a sensitive fluorescent substrate, 4-methylumbelliferyl-β-D-galactopyranoside (MUG). For all assays, following resuspension in Z-buffer, 50 μl 2 mg/ml MUG was added, and samples were incubated at 37 °C for 30 min before the reaction was stopped with 250 μl 1 M Na2CO3. A volume of 50 μl of the total reaction was moved to a 96-well plate and fluorescence measured on a BioTek Synergy H1 microplate reader with a blue filter set (EX 360/40, EM 460/40, DM 400) with 365 nm excitation and 450 nm emission. To confirm that the signal measured was within the linear range of the instrument, induced cultures of GLBS03 (full-length lacZ) at steady-state were collected and fluorescence of dilutions measured (Extended Data Fig. 2a). OD normalized signal measured from the lowest dilution was similar to signal from later time points in induction curves.

Average pre-induction background was subtracted from each point and the square root of signal plotted against time (Schleif plot) to obtain a linear fit with the intercept representing τTL. Points below background or before signal began to increase consistently were excluded from the fit. For lacZa-complementation assays, points after which signal stopped increasing quadratically were excluded and determined as follows. A slope was calculated from the first two points included in the fit and the angle between this line and the next point calculated. This calculation was done for all subsequent pairs of points until an angle >20° between the resulting fit line and the following point was measured. This point and all subsequent points were excluded from the final fit. Error bars shown are the s.e.m. between biological replicates.

In addition to the use of chloramphenicol and erythromycin to stop translation in *B. subtilis*, three additional stop solutions were tested by measuring τTL for pycA–lacZa fusions. Additional stop solutions were as follows. (1) Immediately after collection into chloramphenicol and erythromycin, cells were flash frozen in liquid nitrogen and subsequently thawed in a water bath before complementation as described above. (2) Toluene (15 μl) was added to chloramphenicol and erythromycin and cells were vortexed immediately after collection. Cells were pelleted and resuspended in 500 μl Z-buffer and complemented for 1 h at 37 °C. Cells were then permeabilized and the assay performed as described above. Reactions were run for 2.5 h rather than 30 min to account for lower β-galactosidase activity in these samples. (3) Lincomycin (50 μl, 12.5 mg/ml) was added to chloramphenicol and erythromycin solution. Following collection, assays were performed as described above. The measured τTL for each of these stop solutions was less than τTL measured for the stop solution containing only chloramphenicol and erythromycin (Extended Data Fig. 2b), suggesting that this stop solution was sufficient to quickly stop translation.

Measurements of mRNA induction kinetics

Cells were grown as described above until they reached an OD of 0.20–0.30. A 1-ml pre-induction culture was collected into 1 ml of ice-cold stop solution (60% ethanol, 2% phenol pH 8, 10 mM EDTA) and mixed vigorously before being placed on ice. Expression was induced as described for protein-induction measurements. Following a brief delay
upstream from the start of the hairpin stem, suggesting that they nucleotide mutations. These mutations were at least 20 nucleotides all the way inside the loop of its terminator by introducing two single mutations by northern blot. Importantly, the ORF of pupG could be extended with in our condition and presented a simple operon structure (bicistronic lation on intrinsic transcription termination. reading frame, providing a stringent test of lack of interference of trans- transcription and translation could be coupled at the end of the open con- consecutive U residues). The resulting fusion strains, primers at the end of lacZa were used (Fp: lacZa-fp; Rp: lacZa-rp; Supplementary Data 7); for lacZa fusion strains, primers at the end of lacZa were used (Fp: lacZa-fp; Rp: lacZa-rp; Supplementary Data 7). The resulting C values were normalized to time zero and all points with >1.5-fold increases in expression were fit to a line. τv was calculated from the intersect of this line and y = 1. Error bars reported are the s.e.m. between biological replicates.

Steady-state analysis of mRNA by qRT–PCR

Cells were grown as described above until they reached an OD of 0.2–0.3. One millilitre of culture was collected into 1 ml ice-cold methanol. Cells were pelleted at 4 °C at 18,213g for 4 min and flash frozen in liquid nitrogen. Cell pellets were stored at −80 °C until use. RNA was then extracted and gDNA depleted using RNeasy Plus mini kit (Qiagen) according to the manufacturer’s instructions. RNA (1 μg) was reverse-transcribed with M-MuLV RT and random hexamer priming. RNA was hydrolysed by adding 10 μl 1 M NaOH to the 20-μl RT reaction. The reaction was subsequently neutralized with 10 μl 1 M HCl and brought to a final volume of 200 μl with water. qPCR was preformed using 10 μl KAPA SYBR qPCR Master Mix, 6 μl 1 mM primer mix, and 4 μl cDNA and run on a Light Cycler 480 II Real-Time PCR Machine. For full-length lacZa: primers annealing to the 3'-end of the transcript were used (forward primer (Fp): P3105-F; reverse primer (Rp): P3105-R4; Supplementary Data 7); for lacZa fusion strains, primers at the end of lacZa were used (Fp: lacZa-fp; Rp: lacZa-rp; Supplementary Data 7). The resulting C values were normalized to time zero and all points with >1.5-fold increases in expression were fit to a line. τv was calculated from the intersect of this line and y = 1. Error bars reported are the s.e.m. between biological replicates.

ORF extension candidate gene selection

The following criteria were considered to select a maximally con- straining target for extending the ORF into a strong terminator in B. subtilis: 1. Stop-to-stem distance larger than 25 nt for the upstream open reading frame. 2. High translation efficiency (75th percentile or higher) of upstream gene as determined by combination of ribosome profiling and Rend-seq (see section ‘Translation efficiency percentile determination’). 3. Efficient intrinsic terminators according to the following criteria: (a) less than 1/1,000 readthrough fraction as measured by Rend-seq in wild-type, ΔpnpA, and Δrho59. (b) Strong RNA hairpin secondary structure (ΔG < −17 kcal/mol), (c) Perfect hairpin (all bases paired in the stem, no bulges). (d) Long U-tract (seven or more consecutive U residues).

The above criteria were selected to find candidate genes for which transcription and translation could be coupled at the end of the open reading frame, providing a stringent test of lack of interference of trans- lation on intrinsic transcription termination.

The pupG terminator met all the above criterion, was well expressed in our condition and presented a simple operon structure (bicistronic with drm) without alternative mRNA isoforms, permitting measurement by northern blot. Importantly, the ORF of pupG could be extended all the way inside the loop of its terminator by introducing two single nucleotide mutations. These mutations were at least 20 nucleotides upstream from the start of the hairpin stem, suggesting that they would not affect terminator function. The terminated transcript in the ORF extended configuration still has a stop codon, avoiding con- founding issues with non-stop-mediated mRNA degradation.

ORF extension constructs in B. subtilis

To assess the influence of translation on transcription termination12,28,44–47 in B. subtilis, we constructed three strains to measure transcriptional readthrough at the pupG intrinsic terminator in different contexts: T1+ (pupG terminator intact, endogenous context), T1− (pupG terminator disrupted: two mutations in the stem resulting in shift from ΔGt1+ = −17.5 kcal/mole (free energy of folding of endogenous terminator hairpin) to ΔGt1− = −4.6 kcal/mole (free energy of folding of disrupted terminator hairpin), and one mutation in the U-tract: UUUUUU → UUUCUU, and ORF extension (pupG terminator intact, two upstream mutations to extend the ORF into the terminator hairpin loop). These are illustrated in Extended Data Fig. 5a.

In these three different contexts at the pupG terminator, a strong intrinsic terminator (the B. subtilis endogenous sodA intrinsic terminator), labelled T2, was appended downstream (start of the sodA terminator region positioned 25 nt downstream of the 3’ end of the pupG terminator, with 25 nt upstream of the start of the sodA hairpin included). This second terminator was placed to capture putative readthrough products from the pupG terminator, and to allow measure- ment of the transcription termination readthrough, analogous to a published method40.

These modifications were inserted in the endogenous genetic context of pupG in B. subtilis by appending a downstream co-directional resistance cassette (spectinomycin<sup>48</sup>) that does not interfere with the upstream and downstream operon, and integrated using long-flanking homology. The constructs were generated by PCR with primers encoding mutations, joined by isothermal assembly, PCR amplified by outer primers and transformed directly into B. subtilis following standard protocols relying on natural competence40. All strains were confirmed by Sanger sequencing.

ORF extension constructs in E. coli

To perform similar measurements in E. coli, we transferred the three constructs (T1+, T1−, and ORF extended) on low copy number plasmid pSC101 under the control of the TetR repressor with a strong ribosome binding site (derived from pSC101-Bd<sup>49</sup>, a gift from J. Chin’s laboratory). The full region, from the beginning of the pupG gene to downstream of the T2 sodA terminator, was amplified by PCR and joined by isothermal assembly to the pSC101 backbone (see Supplementary Data 6 for details). The resulting plasmids were transformed into E. coli K12 MG1655 as described previously<sup>40</sup>, and assembly junctions were confirmed by Sanger sequencing.

Measurement of transcription readthrough

We added 7.5 ml of culture at OD<sub>600</sub> = 0.3 to 7.5 ml of ice-cold methanol, mixed it by inversion, and spun it down at 3,000 relative centrifugal force (rcf) for 10 min at 4 °C. The supernatant was decanted and the cell pellet frozen at −80 °C. RNA was extracted using the RNeasy kit (QIA- GEN) according to the manufacturer’s instruction. Total RNA (10 μg) was loaded onto a 1.2% TBE (tris-borate-EDTA) agarose gel containing 20 μL guanidine thiocyanate, run for 2 h 30 min at 5 V/cm at 4 °C, and transferred to a positively charged nylon membrane by downward capillary transfer. Membranes were cross-linked using UV light, hybridized with short single-stranded DNA probes labelled (T4 PNK, New England Biolabs) with ATP γ-<sup>32</sup>P (Perkin Elmer), and washed following the manufacturer’s instructions. Labelled membranes were exposed to a phosphor storage screen (GE Life Science) for 17 h and imaged with a laser scanner (Typhoon FLA9500, GE Life Science). Band intensities (background subtracted) were quantified using ImageJ. The probe binding to the region between T1 and T2 had some homology to the endogenous sodA region, which led to a strong band running at the
expected size for the sodA transcript, which provided an additional loading control.

To measure readthrough at the pupG terminator in the different translational contexts and strains, two northern blot probes were used: one upstream in the pupG gene for loading and expression control, and one between the T1 and T2 terminators to measure captured transcriptional readthrough products. For each RNA sample (corresponding to one strain), two lanes were loaded with the same RNA and blotted in parallel for the two different probes. The raw northern blot data are shown in Supplementary Fig. 1. The northern blot was repeated (biological replicate) for the B. subtilis constructs with similar results.

As independent confirmation of the northern blot measurements, we quantified readthrough by qRT–PCR on independent samples (biological replicates) for the various constructs from both B. subtilis and E. coli. cDNA samples were prepared as described above (“Steady-state analysis of mRNA by qRT–PCR”), and primer pairs quantifying the short and long (ojBL212+ojBL213 and ojBL209+ojBL210, Supplementary Data 7) and long (ojBL209+ojBL211 and ojBL224+ojBL225, Supplementary Data 7) mRNA isoforms from the pupG terminator were used for readthrough quantification. Readthrough (downstream/upstream signal) was normalized to the T1– construct, leading in B. subtilis to T1+/T1– = 0.0006 ± 0.0001, ORF ext./T1– = 0.0023 ± 0.0009, and in E. coli T1+/T1– = 0.06 ± 0.04, ORF ext./T1– = 0.7 ± 0.4, in which errors are s.e.m. from quantification from the four possible primer pair combinations.

High-confidence list of intrinsic terminators in E. coli and B. subtilis

To assess the proximity of intrinsic terminators to open reading frames (stop-to-stem distance), we leveraged our set of high-confidence intrinsic terminators based on end-enriched RNA-seq (Rend-seq) for wild-type and cells deficient in 3′-to-5′ exonuclease or Rho52, with additional quality control criteria described here.

In brief, the high-confidence terminators were obtained by first mapping in vivo RNA 3′ ends as positions with a combination of large Z-scores in both 3′ peak height and step size. Putative intrinsic terminators within this set of 3′ ends were identified on the basis of sequence features (presence of U-tract, and strong nearby upstream RNA hairpin). 3′ ends with upstream terminator-like sequences were further filtered on the basis of the presence of a corresponding 3′-end mapped peak in 3′-to-5′ exonuclease deletion strains (ΔpanpA in B. subtilis, Δnrnb and ΔpanpA in E. coli). In addition, putative Rho-dependent terminators (based on lack of 3′-mapped peak in a Δrho B. subtilis strain, or overlap with bicyclomycin in significant transcripts (BSTSs)52) were not retained as intrinsic terminators. To be conservative with Rho terminator removal in E. coli, we extended the BTSs with an upstream buffer of 300 nt.

Additional criteria were applied to ensure that identified intergenic intrinsic terminators were part of the simple 3′ UTR of their closest upstream gene. First, any terminator sharing the same upstream gene with one or more other terminator was not included (for example, tandem terminators). Second, any terminator with an intervening 5′ end peak (5′-mapped Z-score >12) between its 3′ end and the upstream gene was removed to avoid intra-operon regulatory elements such as riboswitches. Third, any terminator for which the average read density (100-nt travelling window) between the terminator and the closest upstream stop codon fell below 10% of the read density of the upstream gene was discarded. Finally, terminators for which the stop-to-stem distance was larger than 150 nt were not retained. Our final sets in E. coli and B. subtilis included 409 and 1,228 terminators, respectively (Supplementary Data 2).

The stop-to-stem distance for this high-confidence set (Fig. 2) was determined on the basis of the RNA structure of the hairpin obtained with a constraint of 6 nt unpaired bases from the mapped 3′ end (based on current understanding of the molecular mechanism of intrinsic termination52). This is slightly different from the folding constraint imposed on the purely computationally identified terminators (see ‘Classifier for putative intrinsic terminators’), for which all the consecutive U residues (with the end of the U-tract serving as a proxy for the RNA’s 3′ end) were required to be non-paired. We found no strong difference between the distributions of stem-to-stop distances from these two different methods (96% and 91% within 3 nt in E. coli and B. subtilis, respectively).

Correlation between transcription readthrough and stop-to-stem distance

To assess possible interference between translation and intrinsic termination, we compared the terminator readthrough (defined as the fraction of read densities after and before the terminator) measured by Rend-seq for our list of terminators39 (Supplementary Data 2) to stop-to-stem distances in E. coli and B. subtilis (Extended Data Fig. 5b–i). Readthrough could not reliably be estimated for 17 and 53 terminators in E. coli and B. subtilis from our final set, respectively (either 0 reads downstream or region for estimating readthrough less than 20 nt in size).

In E. coli, terminators close to ORFs (stop-to-stem d ≤12 nt) had significantly more readthrough than others (more readthrough corresponds to a poorer terminator). The 30th percentile (q30) of the readthrough distribution for the two sets of terminators (with d ≤12 nt and d >12 nt) has a fold change of F30 = q30 d≤12/q30 d>12 = 3.3, and P = 10−5 (the P value corresponds to the fraction of random bootstrap sub-samplings of the distribution with q30 d≤12 > q30 d>12, that is, the fraction of sub-samplings in which terminators with d >12 nt perform worse than terminators with d ≤12 nt); see Extended Data Fig. 5c, e, g, i for an illustration of quantities (q30 d≤12, q30 d>12, F30). The difference remained highly significant when we controlled for possible differences in terminator quality by performing the analysis on terminators with good U-tracts (free energy of RNA/DNA U-tract duplex ΔGn > −5 kcal/mol): F30 = 4.0, and P = 5 × 10−4. By contrast, the ORF-proximal intrinsic terminators in B. subtilis (stop-to-stem d ≤12 nt) do not have significantly more transcription readthrough than other terminators (F30 = 0.8, P = 0.97; controlling for U-tract (ΔGn > −5 kcal/mol), F30 = 1.0, P = 0.35).

We observed similar trends for terminator readthrough measured in the 3′-to-5′ exonucleases and Rho deletion strains (B. subtilis Δrho: F30 = 1.1, P = 0.46; B. subtilis ΔpnpA: F30 = 1.2, P = 0.16; E. coli Δnrnb: F30 = 3.6, P < 5 × 10−4; E. coli Δpnp: F30 = 2.6, P = 0.007; all values listed are controlled for U-tract quality, ΔGn > −5 kcal/mol).

Fold-change in rho-null versus wild-type bacteria

Genome-wide fold-changes in mRNA levels in Δrho and wild-type B. subtilis32 (green, Fig. 3b) were determined from Rend-seq data (GSE95211)33 as follows. For each gene, 40 bp on either end was excluded and the mean read density within the remainder of the gene body calculated in units of read per kilobase per million reads mapped to non-rRNA and tRNAs (rpkm). Reported fold-changes were calculated as the ratio of the rpkm value in Δrho to that in the wild type. Only genes with more than 50 mapped reads in both wild-type and Δrho strains were included to avoid counting noise larger than ~15% (3,345 genes in final comparison set). To account for possible slight differences in read depth normalization, the fold-changes were normalized such that the median fold-change was 1 (factor of 1.06). A similar procedure for biological replicates of wild-type cells was followed (magenta in Fig. 3b) to provide a measure of the reproducibility of the expression quantification using Rend-seq.

Translation efficiency percentile determination

Translation efficiency is the per mRNA initiation rate of ribosomes. The ribosome profiling (deep sequencing of ribosome protected fragments) read density divided by RNA-seq (or Rend-seq) read density for a given gene provides an estimate of translation efficiency for the mRNA corresponding to that gene. For genome-wide distributions and translation efficiency of intact genes, we used previously determined
translation efficiency data sets\textsuperscript{35,41}. For pseudogene regions (Extended Data Figs. 7, 8, Supplementary Data 3), translation efficiency was estimated by calculating the ribosome profiling and Rend-seq read density over the region normalized by the total number of million reads not mapping to rRNAs and tRNAs (rpkm) and calculating the ratio. The percentile was then determined by comparison to the genome-wide distribution.

**Nested antisense RNAs**

To look for long untranslated regions in the transcriptome, we searched for non-contiguous operons\textsuperscript{35,46} or excludons\textsuperscript{40,45}, which we refer to as nested asRNAs here, in \textit{B. subtilis} and \textit{E. coli}. These correspond to transcripts connecting two co-directional genes that are interrupted by one or more genes in the opposite direction.

To search for nested asRNAs, we listed all pairs of codirectional genes that were interrupted by 1–3 genes in the opposite direction. For each such codirectional gene pair, the moving average (50-nt window) of the read density from Rend-seq data (from the wild-type strain) between the midpoint of the two codirectional genes was computed. The codirectional pair was retained if at no position within this range did the average read density fall below 0.25 reads/nt, corresponding to continuous transcription from one gene to another. Fifty codirectional gene pairs (45 and 5 with one and two opposite intervening genes, respectively) met this threshold in \textit{B. subtilis}. Notably, few codirectional gene pairs (n = 3 and with one such pair interrupted by a very short 34-aa gene at the first percentile of gene sizes) met this threshold (for a similar total read depth) in \textit{E. coli}, suggesting that codirectional genes interrupted by genes on the opposite strand can only rarely be productively co-transcribed in that species.

To focus on untranslated RNAs with functional requirements, we restricted attention to nested asRNAs for which the mean read density between the two co-directional genes (from 50 nt after the end of the first gene to 50 nt before the start of the second gene, 5% winsorized) was equal to or larger than half of the read density (5% winsorized) of either the upstream or downstream gene, leading to 35 and 3 final candidates in \textit{B. subtilis} and \textit{E. coli}, respectively. These constitute our operational definition of nested asRNAs. The final list can be found in Supplementary Data 3, with representative examples from \textit{B. subtilis} shown in Extended Data Fig. 6. In \textit{B. subtilis}, 29 of 35 nested asRNAs had less than a twofold change in RNA levels (mean Rend-seq read density quantification) upon \textit{rho} deletion. In addition to antisense transcripts surrounded by sense genes on both 5′ and 3′ ends (nested), the final list also includes antisense regions inside the 3′ and 5′ UTRs of genes (29/35 nested, 4/35 in 5′ UTR, 2/35 in 3′ UTR). In \textit{E. coli}, one of the three identified nested asRNAs was sensitive to bicyclicycin (twofold change in mRNA level\textsuperscript{46}).

**Expressed pseudogenes in \textit{B. subtilis}**

To identify expressed pseudogenes in \textit{B. subtilis}, we used as a starting point all entities annotated as ‘pseudogenes’ (annotation file: GCF\_000009045.1_ASM904v1_genomic.gff), leading to 88 hits.

Given that ORF fragments interrupted by frameshifts are annotated as different entities, we clustered annotated pseudogenes spatially with a distance cutoff of 300 bp (end to start), resulting in two types of region: pseudogene clusters (containing more than one annotated pseudogene) and isolated pseudogenes. With these definitions, \textit{B. subtilis} had 28 pseudogene clusters (comprising 65 pseudogenes) and 23 isolated pseudogenes. Each type was considered separately.

For pseudogene clusters, the following criteria were used to restrict attention to expressed pseudogenes with interrupted translation. First, pseudogene clusters spanning less than 300 nt (from start of first pseudogene to end of last pseudogene in cluster) were not considered (too short, 4 of 28). Second, pseudogene clusters in which the first annotated pseudogene had a read density of less than 0.25 reads/nt (Rend-seq) in both wild-type and \textit{Delta}rho strains, or less than 0.25 reads/nt ribosome footprint density (ribosome profiling), were excluded (not measurably expressed, 16 of 24). Third, pseudogene clusters in which the ribosome footprint density changed by less than twofold from the first to the last pseudogene region were excluded (lack of translation interruption across cluster, 2 of 8). In the end, 6 of 28 pseudogene clusters satisfied these criteria.

For isolated pseudogenes, 15 of 23 were too short (less than 300 bp in size), and 8 of 23 were not measurably expressed (less than 0.25 reads/nt in both wild-type and \textit{Delta}rho Rend-seq data). Seven out of 23 isolated pseudogenes satisfied both criteria. Ribosome profiling expression cutoff was not applied a priori for isolated pseudogenes to prevent the exclusion of pseudogenes for which translation was interrupted upstream (see examples below). These were further investigated manually. The following isolated pseudogenes were not retained for polarity assessment: pseudogene \textit{trpC} was not retained as it constituted a full ORF (the allele is rendered non-functional by an in-frame 3-bp deletion which disrupts enzyme activity without interrupting translation\textsuperscript{48}); \textit{yuzB} constitutes a complete ORF paralogous to the flagellar protein gene \textit{hag} with no evidence of translational disruption; and \textit{yueE} constitutes a paralogue of \textit{gtbA} with no evidence of translation disruption (further, the pairs contain many identical regions, complicating the analysis of quantification on the basis of sequencing). Two isolated pseudogenes consisted of the C-terminal fragment ORFs interrupted by large-scale insertions: \textit{spmC} interrupted by SPβ prophage\textsuperscript{34}, and \textit{sigkC} interrupted by the Skin element\textsuperscript{58}. Ribosome profiling data indicated that \textit{sigkC} had no clear decrease in translation compared to the upstream gene \textit{yqab} and was overall expressed at very low levels. \textit{sigkC} was thus excluded.

Finally, isolated pseudogene \textit{yoyA} had evidence of a plausible short upstream unannotated ORF fragment with overlapping ribosome footprint density. It was thus retained for analysis and the two fragments renamed \textit{yoyAn} and \textit{yoyAc} here. The pseudogene cluster comprising \textit{ydzW} genes had two consecutive nonsense mutations throughout uninterrupted ORFs, concomitant with a sharp decrease in ribosome footprint density in ribosome profiling. That region was therefore split into three (\textit{ydzWn}, \textit{ydzWm} and \textit{ydzWc}).

Across all pseudogene transcripts, the translation efficiency percentile before and after the nonsense mutation was calculated to assess the decrease in translation (Extended Data Fig. 7, Supplementary Table 3).

In the end, we identified eight expressed pseudogenes with interrupted translation, providing additional independent examples to assess the prevalence of nonsense-mediated polarity in \textit{B. subtilis}. Each of these corresponds to a transcribed mRNA across which a sudden drop in translation occurs (experimentally confirmed by a decrease in ribosome footprint density).

We used two metrics to assess Rho-mediated polarity for these pseudogenes. First, we estimated the ratio of the read density in the first and last 15% of the regions to assess progressive decrease in mRNA level (schematically depicted in Extended Data Fig. 7a). The ratio spanned the range 0.85 to 1.42 (median 1.29) for the eight pseudogenes, suggesting a lack of Rho-dependent polarity over these regions. Second, the fold-change in read density (subtracting upstream read density corresponding to possible readthrough products) over the full region in \textit{Delta}rho versus wild-type bacteria was computed. The fold-changes for considered pseudogenes spanned 0.60 to 1.66 (median 0.70), corresponding to the 6th to 80th percentiles in the genome-wide mRNA level fold-change distribution in \textit{Delta}rho versus wild-type bacteria.

The lack of large 5′-to-3′ decreasing ramp RNA levels or increase in mRNA levels upon \textit{rho} deletion suggests the absence of nonsense-mediated polarity for the considered pseudogenes in \textit{B. subtilis}. See Fig. 3b, Extended Data Fig. 7 for a visual summary and Supplementary Data 3 for final candidates.

**Expressed pseudogene in \textit{E. coli}**

The analysis for pseudogenes in \textit{E. coli} was similar to that in \textit{B. subtilis}. There were 199 entities annotated as pseudogenes in \textit{E. coli} (GenBank
of these, 66 were in 30 clusters (300 nt distance cutoff for clustering), and the remaining 133 were isolated pseudogenes. Two pseudogene clusters passed thresholds of expression, ribosome footprint density decrease and size (29 of 30 > 300 bp, 3 of 29 measurably expressed with cutoff of 0.25 reads/nt in Rend-seq and ribosome profiling, 2 of 3 with at least a twofold drop in ribosome footprint read density between the first and last region). Twenty-three of 133 isolated pseudogenes were both longer than 300 bp (84 of 133) and measurably expressed (34 of 133 with >0.25 reads/nt Rend-seq read density).

Of expressed and long isolated pseudogenes, 12 of 23 resulting from insertion element rearrangements were excluded owing to the difficulty in read-mapping and possible rapid genomic changes near these regions. The remaining 11 of 23 regions were compared to other E. coli strains (O157, GCF_000008865.2; IA39, GCF_000026345.1; O83, GCF_000183345.1; O104, GCF_000299455.1) to provide context for the possible genetic changes in E. coli K-12 MG1655. Six of 11 had clear mutations leading to measurably disrupted translation as assessed by ribosome profiling (skIA, approximately 2-kb N-terminal portion of gene missing just downstream of promoter compared to O157; efeU, 1-bp deletion leading to frameshift in N-terminal portion of the ORF; gapC, nonsense mutation about 100 bp in the gene and 1-bp deletion leading to a stop codon at about 750 bp; bcsQ, nonsense mutation after 6 aa; ilgC, 2-bp deletion leading to stop codon about 1 kb inside the gene; cybc, 26-nt deletion leading to the ablation of the beginning of the ORF). Two of 11 had frameshifts leading to an early stop codon but no measurable decrease in ribosome footprint density across the stop codon (yabP and yifN). These were excluded from downstream analysis. Finally, some isolated pseudogenes were excluded because of hard-to-interpret features. ybcY, which has a 2-bp deletion in an ATG either leading to an early stop codon or removal of the start codon (depending on the start codon position), was excluded because the ribosome profiling density was not consistent with either scenario, suggesting a more complicated situation. ylbG was excluded despite clear evidence of decreasing 5′-to-3′ ramp at the transcriptional level given that it constituted an uninterrupted ORF. Finally, yibf (plausibly the C-terminal portion of a longer gene, based on comparison with E. coli O157, interrupted by the upstream gene yibA in K-12 MG1655) was excluded despite clear evidence of a decrease in transcription because of a hard-to-interpret short RNA nested in the pseudogene body. The pseudogene cluster comprising gapC segments had three consecutive nonsense mutations throughout uninterrupted ORFs, concomitant with a sharp decrease in ribosome footprint density in ribosome profiling. That region was therefore split in four (gapCn, gapCmt1, gapCmt2 and gapCcc).

Across all pseudogene transcripts, the translation efficiency percentile before and after the nonsense mutation was calculated to assess the decrease in translation (Extended Data Fig. 8, Supplementary Table 3). Note that with substantial polarity (decrease in mRNA level across nonsense mutation), the decrease in translation efficiency across the nonsense mutation will naturally be lower (as a reflection of the denominator of the mRNA level being lower).

In total, our final list of expressed pseudogenes with measurable interrupted translation in E. coli comprised eight examples. As in B. subtilis, we used decrease in mRNA read density across the pseudogene region as evidence of Rho-mediated polarity. The fold-change in RNA read density (Rend-seq) in the first and last 15% of each region was estimated (Extended Data Fig. 8a). Six of 8 regions had a fold-change start/end larger than 2 (range: 1.0 to 7.4, median 4.7). In addition, 5 of 8 showed a fold-change in mRNA levels (either pseudogenes or downstream genes in the same operon) of 2 or more upon treatment with bicyclomycin (an inhibitor of the transcription termination factor Rho) compared to untreated control in the study of Peters et al. (from the RNA-seq data in Supplementary Table 2 in ref. 49).

The decreasing mRNA level in a 5′ to 3′ fashion across the majority of considered pseudogenes with interrupted translation, together with responsiveness to bicyclomycin, confirms that nonsense-mediated polarity is common in E. coli, consistent with extensive prior literature50. See Extended Data Fig. 8 for a visual summary, and Supplementary Data 3 for final candidates.

Analysis of C-to-G Ratio
To categorize CDSs as either Rho-terminated or non-Rho-terminated, we considered CDSs with at least 150 reads in the Δrho dataset33, at least 15 reads within the first 10% of the gene in the wild-type data55, and that were at least 100 bp in length. For each gene that met these criteria, we calculated four values: the number of reads in the first and last 10% of the gene in wild type and Δrho (reads start_WT/reads end_WT, reads start_Rho/reads end_Rho), Rho-terminated CDSs were defined as those for which reads start_WT/reads end_WT > 4 (that is, exhibited an expression decrease along the gene body) and reads start_Rho/reads end_Rho < 2 (to filter out expression decreases resulting primarily from processing or intrinsic termination rather than Rho-dependent termination events). Eleven genes met both of these criteria (kinB, yfH4, aibE, comEC, trpE, msnG, msnE, sqhC, csbX, rapA and ywrK). msnG (amyC) contains several prominent 5′ and 3′ ends within the first 10% of the gene that are enriched in the wild type and that confound classification of this gene as Rho-terminated. This gene was therefore excluded from our analysis. Non-Rho-terminated genes were defined as genes where (reads start_WT/reads end_WT)/(reads start_Rho/reads end_Rho) < 1.5, of which there were 2,625 genes. For CDSs in both groups, we calculated the C-to-G ratio for all 100-nt windows and compared the distribution of maximum C-to-G ratios. The distribution of maximum C-to-G ratio for Rho-terminated CDSs was significantly higher than for non-Rho-terminated CDSs (P < 10−5, less than one in 104 random sub-samplings (n = 10) of non-Rho-terminated distribution had higher median maximum C-to-G ratio) (Extended Data Fig. 9a).

asRNAs terminated by Rho were identified using the same criteria used to identify Rho-terminated CDSs, looking instead in genomic regions antisense to CDSs. Of the 168 asRNAs that passed our expression threshold, 92 were classified as Rho-terminated asRNAs and 112 as non-Rho-terminated asRNAs. One of the Rho-terminated asRNAs (that antisense to cypX) contained a short transcript in the 5′ end of the asRNA driven by a promoter not present in Δrho. This asRNA was therefore excluded from analysis. For asRNAs in both groups, we calculated the C-to-G ratio for all 100-nt windows and compared the distribution of maximum C-to-G ratios. The distribution of maximum C-to-G ratios for Rho-terminated asRNAs was significantly higher than for non-Rho-terminated asRNAs (P < 10−5, less than one in 104 random sub-samplings (n = 10) of non-Rho-terminated distribution had higher median maximum C-to-G ratio compared to sub-sampling (n = 10) of Rho-terminated distribution) (Extended Data Fig. 9b).

For nested asRNAs, the C-to-G ratio in 100-nt moving windows was determined both for antisense regions and for the full region between the co-directional genes. We compared the distribution of maximum C-to-G ratios for the antisense regions within nested asRNAs to the distribution for all regions antisense to CDSs genome-wide. Random subsampling of the genome-wide antisense distribution suggested a highly significant decrease in the maximum C-to-G ratios for nested asRNAs (P < 10−5, less than one in 104 random sub-samplings of the genome-wide distribution had lower median maximum C-to-G ratio). Similar analysis was also performed with the maximum C-to-G ratio for the full region between codirectional genes for the nested asRNAs. The control set was all regions between codirectional with one or two intervening genes in the opposite directions, restricted to the same size range as our set of nested asRNAs (164 to 1,606 nt) leading to 484 regions. Again, the maximum C-to-G ratio for nested asRNAs was significantly lower than for the control set (P < 10−4, less than one in 104 random sub-samplings of control distribution had lower median maximum C-to-G ratio).
Species considered for intrinsic terminator identification

Prokaryotic reference and representative genomes from the RefSeq database were downloaded using Assembly from NCBI on 16 March 2019 using the query terms: "bacteria" [organism] AND "representative genome" [refseq category] OR "reference genome" [refseq category] AND (bacteria[filter] AND latest[filter]) AND "complete genome" [filter] AND all[filter] NOT anomalous[filter]. This returned 1,645 genomes, which were all searched for intrinsic terminators as described below.

Classifier for putative intrinsic terminators

Each genome (all sequence elements in the reference file: chromosome, plasmids, and so on) was analysed in isolation. Given the computationally intensive process of RNA secondary structure calculation, putative intrinsic terminators were identified by a two-step process: (1) restrict the attention to U-rich regions downstream of genes; and (2) fold upstream RNA and store hairpin characteristics. Stringent selection criteria were then applied to the resulting RNA structures and U-tract based on species-specific distribution in properties of RNA secondary structure from randomly selected genome positions.

Specifically, regions of the genome downstream of stop codons (on both strands, with strand-specific information retained), from \( x_{\text{f}} - 10 \) to \( x_{\text{f}} + 200 \) nt (in which \( x_{\text{f}} \) is the annotated position of the stop codon for ORF \( n \), and + refers to downstream of the gene in the 5′- to 3′ direction), were retained for all stop codons. If a downstream co-directional gene (\( n + 1 \)) was closer than 200 nt to the stop codon, region \( x_{\text{f}} - 10 \) to \( x_{n+1} + 30 \) nt was retained. From within this set of sequences, stretches of more than five consecutive Ts were identified as possible U-tract of intrinsic terminators for further analysis of upstream secondary structure motif (strong hairpin). For species with GC content exceeding 60%, stretches of four consecutive Ts were also retained.

For each putative U-tract identified, the minimum free energy RNA secondary structure for the upstream sequences of various lengths (30, 35, 40, 45 and 50 nt) ending at the end of the U-tract were obtained using RNAfold (option -C -p\(^{12}\), with the constraint that the U residues in the putative U-tract (that is, only the last stretch of consecutive U residues) remained unfolded. For each structure (and each folded length for a given position) at each putative U-tract, structure properties were extracted: \( N \) number of hairpins, \( S \) size of stem, \( L \) size of loop, \( \Delta G \) minimum free energy of folding, \( I \) distance between the 5′ end of the putative U-tract and the 3′ most base in the stem of the hairpin, and \( f \) fraction of bases paired in stem.

To mitigate differences in GC content across species with regards to selection threshold, and to account for other species-specific differences, we folded 10\(^{4}\) randomly chosen regions of 40 nt in the genome of each considered species. The hairpin properties for these random regions were stored, as for the U-tract selected regions above.

The selection criteria to identify putative intrinsic terminators were as follows. First, thresholds were applied to geometric features of the folded RNA secondary structure to select for appropriate hairpins: \( N = 1 \) (single hairpin for folded region), \( 5 \text{ bp} \leq N_{\text{loop}} \leq 15 \text{ bp} \) (sufficient stem), \( 3 \text{ nt} \leq \text{loop} \leq 8 \text{ nt} \) (non-anomalous loop), \( I = 0 \) nt (stem required to be immediately adjacent to U-tract given the importance of these U residues in termination\(^{11,12\text{a,b}}\)). If for a given U-tract position, more than one folded length hairpin passed these cuts, a single hairpin was selected as follows. Each hairpin was ranked on the basis of three properties: \( \Delta G_{\text{U-tract}}, \Delta G_{\text{U-tract}}/N_{\text{loop}} \) and \( f \). The selected hairpin for a U-tract was chosen as the hairpin with the highest number of best scoring ranks for these properties. In the case of a tie, the hairpin with the highest number of first and second ranks was chosen. In the case of a further tie, the hairpin arising from the shortest folded region was retained.

The final, species-specific thresholds on the hairpin strength \( \Delta G \) and fraction of bases paired in the stem were based on the properties of hairpins from randomly selected regions. For all random regions with \( N = 1 \) hairpin passing the same geometrical criteria as for putative intrinsic terminators (\( 5 \text{ nt} \leq N_{\text{loop}} \leq 15 \text{ nt}, 3 \text{ nt} \leq \text{loop} \leq 8 \text{ nt} \)), the free energy of folding \( \Delta G \) and \( \Delta G \) for which less than 1% and 1.5% of randomly folded regions' hairpins obeyed respectively \( (\Delta G \leq \Delta G_{\text{f}} \text{ and } f \geq 0.95) \), and \( (\Delta G \leq \Delta G_{\text{f}} \text{ and } f \geq 0.9) \) were identified. This determined \( \Delta G_{\text{f}} \) and \( \Delta G_{\text{f}} \) for each species. If \( \Delta G_{\text{f}} \) or \( \Delta G_{\text{f}} \) was higher than −6.5 kcal/mol, the free energy threshold was set to −6.5 kcal/mol. Hairpins upstream of the U-tract with satisfactory geometrical features (see previous paragraph) that further satisfied either \( (\Delta G \leq \Delta G_{\text{f}} \text{ and } f \geq 0.95) \) or \( (\Delta G \leq \Delta G_{\text{f}} \text{ and } f \geq 0.9) \) were considered putative intrinsic terminators. In the case of species with GC content higher than 60%, the additional threshold of requiring \( \Delta G_{\text{U-tract}} \leq -20 \text{ kcal/mol} \) for hairpins upstream of a 4-U residue U-tract was implemented to decrease the number of false positives (as assessed by our false discovery rate analysis on \( C. \) crescentus). The procedure is illustrated in Extended Data Fig. 10a–c for \( V. \) cholerae.

Some terminators were further excluded for reasons other than their intrinsic properties to ensure a high quality set for stop-to-stem distance assessment. First, terminators arising from genomic elements with size less than 5% of the maximal chromosome size (for example, small plasmids) were discarded. In addition, instances where multiple terminators were identified downstream of a gene were excluded. Repeated sequence/terminators (terminators with identical sequence and stop-to-stem distances) were excluded. Finally, terminators with stem-to-stop distance larger than 150 nt were excluded to avoid possible annotation errors.

In the end, we identified 301,817 terminators for which the stem-to-stop distance could be determined, with a median of 125 per species, and 1,434 species with at least 20 identified terminators. A summary for each starting species is in Supplementary Data 4, and all identified terminators that satisfied the above criteria are listed in Supplementary Data 5 (we recommend parsing this data file computationally). The Matlab scripts developed for the analysis have been deposited to GitHub (https://github.com/jbblalanne/intrinsic_trx_terminator_identifier).

False discovery rate estimation for putative intrinsic terminators

We took advantage of our Rend-seq datasets, which allow the identification of 3′ ends of transcripts, to directly assess the performance of our terminator identification algorithm in \( E. \) coli, \( B. \) subtilis, \( S. \) aureus, \( C. \) crescentus, and \( V. \) natriegens (the last species was not included in the RefSeq set, but the algorithm was run on this species as well).

For each species for which we had Rend-seq data, we computed the 3′ peak \( Z \)-score\(^{14\text{a,b}}\) in a ±5-nt neighbourhood from the end of the U-tract of putative intrinsic terminators identified by our algorithm. The position was considered to include a 3′ end if the maximum peak \( Z \)-score in that neighbourhood was above 7. Positions corresponding to insufficiently expressed genes (average read density <0.25 reads/nt in 100-nt window surrounding the position) were excluded from our analysis. For putative intrinsic terminators not containing a 3′ end at the end of their U-tract as determined automatically by our peak finding algorithm, we manually verified the presence or absence of clear ends of transcription units. This was important to avoid confounding false positives from the Rend-seq peak finding script with bona fide false positives in the terminator identification algorithm, and was particularly important for the \( S. \) aureus library prepared with lower end enrichment and \( C. \) crescentus (more noise in read coverage) samples. A position was deemed false positive if the identified position of the terminator was incorrect, or if there was no apparent termination based on our data.

For \( C. \) crescentus (GC content = 63%), we found that terminators identified using the usual thresholds and 4-U residue U-tract were more frequently false positives. Increasing the stringency of the free energy threshold to \( \Delta G \leq -20 \text{ kcal/mol} \) for these 4-U residue U-tract-derived hairpins reduced the false positives to the same rate as for other putative hairpins. We applied this criterion to all other species with GC content >60%.
We found false positive rates of: 1.1% in *C. crescentus* (2/187), 0.5% in *E. coli* (1/207), 0.1% in *B. subtilis* (1/733), 1.1% in *S. aureus* (2/175), and 0.7% in *V. natriegens* (4/606). Overall, these results suggest an experimentally validated false discovery rate for our terminator identification pipeline of close to or lower than 1% for these five diverse species.

**Generation of phylogenetic tree**

To generate a phylogenetic tree (Fig. 4) for display of the stop-to-stem distributions for the considered species, we first identified nearly universal and single-copy orthologous genes by using BUSCO (version 3)**a** (lineage file: bacteria_oddb) on each of our RefSeq species in protein assessment mode (n prot). We kept for downstream analysis only hits labelled as complete, and removed orthologues identified as complete in less than 95% of species, leading to a set of 84 unique orthologues in our final alignment. For each orthologue, the resulting sequences from the RefSeq species were aligned using MAFFT (version 7) (options:--retree 2 –maxiterate 0)**a**. The results for all 84 orthologues were concatenated horizontally. The resulting alignment was trimmed (removing columns with more than 50% gaps or whose consensus frequency was less than 25%). The final multiple sequence alignment contained 20,814 positions. To avoid double-counting identical species, species from the RefSeq whose above alignment was more than 99% identical were grouped together. Specifically, we grouped all members of connected clusters, where species were connected if their above described orthologues’ alignment was more than 99% identical. We found eight connected clusters of more than one species, including a total of 14 species (clusters are identified in Supplementary Data 4 with ‘grouped connected clusters, where species were connected if their above described orthologues’ alignment was more than 99% identical’). The resulting alignment was trimmed (removing columns with more than 50% gaps or whose consensus frequency was less than 25%). The multiple sequence alignment contained 20,814 positions. To avoid double-counting identical species, species from the RefSeq whose above alignment was more than 99% identical were grouped together. Specifically, we grouped all members of connected clusters, where species were connected if their above described orthologues’ alignment was more than 99% identical. We found eight connected clusters of more than one species, including a total of 14 species (clusters are identified in Supplementary Data 4 with ‘grouped connected clusters, where species were connected if their above described orthologues’ alignment was more than 99% identical’). The resulting alignment was trimmed (removing columns with more than 50% gaps or whose consensus frequency was less than 25%).

The final phylogenetic tree was generated for all species (with dereplication described above) with at least 20 identified terminators (n = 1,434) using FastTree (version 2.1.1)**b** with default options on the trimmed concatenated BUSCO MAFFT alignments. The tree was re-rooted at species GCF_000739455.1 and ladderized using phytools**b** (version 0.6-99). The resulting phyla were concordant with those of GTDB89 (only exception: GCF_000217795.1 and GCF_000734015.1 falling between Acquinocita and Deferrribacterota instead of with other Deltaproteobacteria). Plotting of the tree with stop-to-stem distributions was done in Matlab using custom scripts.

The following GTDB89 phyla**b** were grouped for display in Fig. 4 and phylogeny-stratified analysis (Extended Data Fig. 10d, thresholds for Fig. 4 correspond to thresholds given phylum). To highlight the phylum stratification of stop-to-stem distributions for the considered species, we first identified nearly universal and single-copy orthologous genes by using BUSCO (version 3)**a** (lineage file: bacteria_oddb) on each of our RefSeq species in protein assessment mode (n prot). We kept for downstream analysis only hits labelled as complete, and removed orthologues identified as complete in less than 95% of species, leading to a set of 84 unique orthologues in our final alignment. For each orthologue, the resulting sequences from the RefSeq species were aligned using MAFFT (version 7) (options:--retree 2 –maxiterate 0)**a**. The results for all 84 orthologues were concatenated horizontally. The resulting alignment was trimmed (removing columns with more than 50% gaps or whose consensus frequency was less than 25%). The final multiple sequence alignment contained 20,814 positions. To avoid double-counting identical species, species from the RefSeq whose above alignment was more than 99% identical were grouped together. Specifically, we grouped all members of connected clusters, where species were connected if their above described orthologues’ alignment was more than 99% identical. We found eight connected clusters of more than one species, including a total of 14 species (clusters are identified in Supplementary Data 4 with ‘grouped connected clusters, where species were connected if their above described orthologues’ alignment was more than 99% identical’). The resulting alignment was trimmed (removing columns with more than 50% gaps or whose consensus frequency was less than 25%).

**Sequence alignments**

Sequence alignments for NusA, NusG, and RpoB**b** (Extended Data Fig. 4, Supplementary Discussion SNS) were done using MAFFT, trimmed by removing columns at the inflection point of the gap distribution (≥95% gaps, such that protein domains present in less than 5% of species considered would not appear in our alignment), and ordered based on the phylogenetic tree of Fig. 4a. For NusA and NusG, domains annotated in Uniprot**b** for the *E. coli* protein were mapped to the positions in the alignment. For RpoB, the conserved sequences identified in ref. **b**(βb1 to βb16) were mapped to positions in the alignment.

**Downstream analysis on stop-to-stem distributions**

The stop-to-stem distributions in each species and the grouping by phylum represent two layers of statistical distribution (stop-to-stem distance within a given species, and distribution of the summary statistic of each species’ stop-to-stem distribution across species in a given phylum). To highlight the phylum stratification of stop-to-stem distributions of putative terminators, the fraction of species within a phylum for which more than a chosen fraction F of identified terminators had a stem-to-stop distance less than or equal to D was computed for all F and D thresholds (Extended Data Fig. 10d, e). For concreteness, note that Fig. 4 corresponds to thresholds F = 30%, and D = 12 nt. The fraction of species meeting the threshold on (D, F), as a function of F and D (Extended Data Fig. 10d, thresholds for Fig. 4) fell between 50% and 12 nt indicated by yellow star) is then an indication of tolerance of members of the phylum to proximity of intrinsic terminators to coding sequences. We see clear separation in the (D, F) space between different phyla highlighted in Fig. 4, confirming that our conclusion of phylogenetic prevalence of RNAP and ribosome uncoupling are not based on the specific thresholds selected for display in Fig. 4.

We also confirmed that the differences in stop-to-stem distance distributions were not strongly correlated to other properties, such as overall genome compaction fraction (defined as the fraction of the genome not encoding for genes, which can be both protein coding sequences and non-coding RNAs such as rRNA and tRNAs) and GC content (R² = 0.036 between median stop-to-stem distance and genome compaction fraction, and R² = 0.122 between median stop-to-stem distance and GC content).

**Reporting summary**

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

**Data availability**

All data generated and analysed during this study are included in this published article (and its Supplementary Information). The high-throughput sequencing datasets analysed during the current study are available from the Gene Expression Omnibus repository with accession numbers: GSE53767, GSE95211 and GSE108295 (see ‘High-throughput expression datasets used for details’). Uncropped gel source data for northern blots can be found in Supplementary Fig. 1. Source data are provided with this paper.

**Code availability**

Scripts for terminator identification have been deposited to GitHub (https://github.com/jblalanne/intrinsic_trx_terminator_identifier). Core Rend-seq analysis scripts used can be found on Github (https://github.com/jblalanne/Rend_seq_core_scripts). Other custom scripts used for data analysis are available upon request.

---

**References**

40. Harwood, C. R. & Cutting, S. M. *Molecular Biological Methods for Bacillus* (John Wiley, 1990).

41. Li, G.-W., Burkhardt, D., Gross, C. & Weissman, J. S. *Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources.* Cell 127, 624–635 (2014).

42. Deloughery, A., Lalone, J.-B., Losick, R. & Li, G.-W. *Maturation of polycistronic mRNAs by the coupling of translation to transcription.* EMBO J. 115, 2857–2861 (1984).

43. Zhu, M., Dai, X. & Wang, Y.-P. *Real-time determination of bacterial in vivo ribosome translation elongation speed based on LacZα complementation system.* Nucleic Acids Res. 44, gkw686 (2016).

44. Bonekamp, F., Clemen, K., Karlström, O. & Jensen, K. F. *Mechanism of UTP-modulated attenuation at the pyrE gene of Escherichia coli: an example of operon polarity control through the coupling of translation to transcription.* EMBO J. 3, 2857–2861 (1984).
45. Abe, H., Abo, T. & Alba, H. Regulation of intrinsic terminator by translation in Escherichia coli: transcription termination at a distance downstream. Genes Cells 4, 87–97 (1999).

46. Unniraman, S., Prakash, R. & Nagaraja, V. Alternate paradigm for intrinsic transcription termination in eubacteria. J. Biol. Chem. 276, 41850–41855 (2001).

47. Kobayashi, K., Kawaara, R. & Takamatsu, H. In vitro mRNA is missing a stop codon in the undomesticated Bacillus subtilis strain ATCC 6051. Microbiology 154, 54–63 (2008).

48. Guerout-Fleury, A. M., Shazand, K., Frandsen, N. & Stragier, P. Antibiotic-resistance cassettes for Bacillus subtilis. Gene 167, 335–336 (1995).

49. Rackham, O. & Chin, J. W. A network of orthogonal ribosome x mRNA pairs. Nat. Chem. Biol. 1, 159–166 (2005).

50. Chung, C. T., Vangeloff, A. D. & Landick, R. Bacterial transcription terminators: the RNA 3'-end chronicles. J. Mol. Biol. 412, 793–813 (2011).

51. Peters, J. M., Vangeloff, A. D. & Landick, R. Bacterial transcription terminators: the RNA 3'-end chronicles. J. Mol. Biol. 412, 793–813 (2011).

52. Bidnenko, V. et al. Termination factor Rho: from the control of pervasive transcription to complex gene regulation. J. Mol. Biol. 373, 779–808 (2007).

53. Sørensen, N. & Stragier, P. The sequence of the trp operon of Bacillus subtilis. Proc. Natl Acad. Sci. USA 86, 2172–2175 (1989).

54. Sesto, N., Wurtzel, O., Archambaud, C., Sorek, R. & Cossart, P. The excludon: a new concept in bacterial antisense RNA-mediated gene regulation. Nat. Rev. Microbiol. 11, 75–82 (2013).

55. Yan, B., Boitano, M., Clark, T. A. & Ettwiller, L. SMRT-Cappable-seq reveals complex operon variants in bacteria. Nat. Commun. 9, 3676 (2018).

56. Albertini, A. M. & Galizia, A. The sequence of the trp operon of Bacillus subtilis 168 (trpC2) revisited. Microbiology 145, 3319–3320 (1999).

57. Abe, H. et al. Developmentally-regulated excision of the SPβ prophage reconstitutes a composite gene for a developmental transcription factor. Science 260, 41850–41855 (2011).

58. Peters, J. M., Vangeloff, A. D. & Landick, R. Bacterial transcription terminators: the RNA 3'-end chronicles. J. Mol. Biol. 412, 793–813 (2011).

59. Peters, J. M., Vangeloff, A. D. & Landick, R. Bacterial transcription terminators: the RNA 3'-end chronicles. J. Mol. Biol. 412, 793–813 (2011).

60. Peters, J. M., Vangeloff, A. D. & Landick, R. Bacterial transcription terminators: the RNA 3'-end chronicles. J. Mol. Biol. 412, 793–813 (2011).

61. Lorenz, R. et al. Viennarna Package 2.0. Algorithms Mol. Biol. 6, 26 (2011).

62. Guaspari, I. & Nudler, E. The mechanism of intrinsic transcription termination. Mol. Cell 3, 495–504 (1999).

63. Sipos, K., Szigeti, R., Deng, X. & Turnbough, C. L., Jr. Systematic mutagenesis of the thyminde tract of the pyrB attenuator and its effects on intrinsic transcription termination in Escherichia coli. Microbiol. 66, 127–138 (2007).

64. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212 (2015).

65. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780 (2013).

66. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximate maximum-likelihood trees for large alignments. PLoS ONE 5, e9480 (2010).

67. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. 3, 217–223 (2012).

68. Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol. 36, 996–1004 (2018).

69. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 45 (D1), D158–D169 (2017).

70. Lane, W. J. & Darst, S. A. Molecular evolution of multisubunit RNA polymerases: sequence analysis. J. Mol. Biol. 395, 671–685 (2010).

71. Yang, X. et al. The structure of bacterial RNA polymerase in complex with the essential transcription elongation factor NusA. EMBO Rep. 10, 997–1002 (2009).
Extended Data Fig. 1 | Transcription and translation kinetics in slow growth. Induction time course of lacZ mRNA (top) and protein (bottom) as in Fig. 1b, d for WT B. subtilis grown in MOPS minimal media + 0.4% maltose (growth rate 0.65 h⁻¹). Lines indicate linear fits after signals rise. Uncertainties are standard error of the mean (s.e.m.) among biological replicates (2).
Extended Data Fig. 2 | Validation of β-gal assay. a, Measurement of linear range of microplate reader. Fluorescence relative to input of dilutions of an induced culture of GLB503 (full-length lacZ) at steady-state (Methods).

b, Effect of different stop solutions on stopping translation. Induction time courses of pycA-lacZα protein collected into a stop solution containing chloramphenicol and erythromycin (grey, all plots, from Fig. 1d) or with either flash freezing in liquid nitrogen (top), 15 μl toluene added to the stop solution (middle), or 50 μl 12.5 mg/ml lincomycin added to the stop solution (bottom), shown in red in each plot (as described in Methods). Lines indicate linear fits after signals rise and τT is indicated.

c, Induction time course of truncated pycA-lacZα mRNA (top) and protein (bottom) as in Fig. 1b, d. Lines indicate linear fits after signals rise. Uncertainties are standard error of the mean (s.e.m.) among biological replicates (2).
Extended Data Fig. 3 | Contribution of non-essential RNAP subunits and transcription factors to fast transcription. Induction time course of pycA-lacZα mRNA in various mutant backgrounds as in Fig. 1b, d. Time course of the same construct in WT from Fig. 1d also shown for reference. Lines indicate linear fits after signals rise. Uncertainties are standard error of the mean (s.e.m) among biological replicates (1 for ∆ykzG and 2 for all others). Time of appearance of full-length mRNA in mutants is not substantially different than that measured in WT (Supplementary Discussion SN4).
Extended Data Fig. 4 | See next page for caption.
Extended Data Fig. 4 | Phylogenetic distribution of domain architecture for NusG, NusA and RpoB. A. Multiple sequence alignments (Methods) for NusA (602 columns), NusG (325 columns), and the β subunit of the RNAP RpoB (1732 columns) for species shown in Fig. 4. The alignments are visualized in a binary fashion to highlight presence/absence of certain domains: white indicates presence of an amino acid in the alignment, and black indicates presence of a gap. The alignments were trimmed by removing columns with >95% gaps. Species with no homologues, partial or pseudogene homologues, or multiple homologues are shown as grey lines. Phylogenetic tree and fraction of terminators with stop-to-stem distances within 12 nt from Fig. 4 are reproduced in linearized form. The position of domains from the E. coli protein are identified by bars above the alignments. For RpoB, conserved bacterial regions identified by β16 (βb1 to βb16) are shown. The NusA C-terminal domain (orange box) is missing in a large fraction of Firmicutes (partly present in Mollicutes, which include Mycoplasma and Spiroplasma; red brace), Campylobacterota, Thermotogota, Fusobacteria, and Actinobacteria. NusG has a largely conserved domain architecture, with Actinobacteria showing N-terminal extension. As previously noted in detail, the β subunit of the RNAP has multiple insertion domains in diverse bacteria. Insertion domain βSI2, recently implicated in transcription-translation coupling is lineage-specific and absent in many clades of Gram-positive bacteria, as noted in. Dashed box in tree highlights clade containing Mycoplasma. B. Close-up view of our analysis of the clade containing Mycoplasma (indicated by black dots). Sub-tree includes species with n ≤ 20 identified terminators (marked in light red). Grayscale representation of stop-to-stem distributions and fraction of terminators with d ≤ 12 nt are the same as Fig. 4. M. pneumoniae is highlighted in cyan, and has no identified terminator (0/14) with d ≤ 12 nt. C. Cumulative distribution of stop-to-stem distance for bioinformatically identified terminators in M. pneumoniae.
Extended Data Fig. 5 | Details of ORF extension constructs and transcription terminator readthrough vs. stop-to-stem distances.

**a**. Sequence for terminators T1 and T2 for three variants (T1+: pupG original terminator, T1-: disrupted pupG terminator, ORF extension: original pupG with upstream ORF extended inside the loop of the terminator). For T1 and T2, blue and grey shading, respectively, marks the position of the terminator hairpin stems, with free energy of folding $\Delta G$ indicated. Black stars indicate introduced mutations. Downward carets ($\downarrow$) indicate the position of the 3′ ends associated with intrinsic terminators as determined by Rend-seq. Red dashed line indicates the complementary region of the northern blot probe to the readthrough product. **b**. Terminator readthrough fraction (defined as the Rend-seq read density after terminator divided by read density upstream of terminator, see ref.13 for details) as a function of stop-to-stem distance for *E. coli* intrinsic terminators from Fig. 2 for which readthrough could be reliably estimated ($n = 392$). Terminators with stop-to-stem distance $d \leq 12$ nt are highlighted in red. **c**. Cumulative distribution function of terminator readthrough for terminators far (black, $d \geq 12$ nt) from and close (red, $d \leq 12$ nt) to stop codons. Terminator close to genes have significantly more readthrough ($P < 10^{-3}$; $q_{30}^{d \leq 12} > q_{30}^{d > 12}$), $P$-value determined as the fraction of bootstrap random sub-samplings of the readthrough distributions with $q_{30}^{d > 12} > q_{30}^{d \leq 12}$, Methods.) **d**. Terminator readthrough as a function of $\Delta G_U$, the U-tract DNA/RNA hybrid free energy (measure of U-tract quality, with larger $\Delta G_U$ corresponding to U-rich U-tract). Grey shading indicates cutoff ($\Delta G_U > -5$ kcal/mol) to select good U-tract terminators. **e**. Same as c, but restricting to good U-tract terminators, still showing significantly less termination for terminators near ORF ($P < 10^{-3}$; same as above, Methods). **f–i**. Same as b–e, but with terminators from *B. subtilis*. Terminators close to ORF do not show less readthrough than their gene-distal counterparts ($P > 0.3$, $p$-value determined with same strategy as above, Methods).
Rho-termination in nested antisense RNAs: Dashed lines depict antisense regions for which mRNA level change in Δρho is calculated.

Extended Data Fig. 6 | See next page for caption.
Extended Data Fig. 6 | Examples of identified nested antisense RNAs. *B. subtilis* shows a number \( n = 35 \), see Methods for selection criteria of mRNAs with long untranslated regions fully encompassing genes in the antisense directions, which we call nested antisense RNAs (also termed non-contiguous operons\(^5\) or excludons\(^4\)). The majority \( n = 29/35 \) of these have a fold-change in mRNA level less than twofold upon \( \rho \) deletion (Fig. 3b). a, Schematic of a nested antisense RNA with corresponding Rend-seq signal, with orange peaks and blue peaks marking 5′ and 3′ boundaries of the transcript. b, Representative examples of nested antisense RNAs with mRNA level fold change upon \( \rho \) deletion less than 2. Rend-seq data (peak shadows removed, see\(^35\) for details on data processing) is shown. Orange and blue signal correspond to summed 5′-mapped reads and 3′-mapped reads, respectively (rpm: reads per million). Top trace corresponds to wild type, and bottom trace to \( \Delta \rho \). Horizontal size marker provides positional scale (200 bp) on each subpanel. Sense and antisense genes are shown in dark and light grey, respectively. Double line breaks (//) indicate truncated Rend-seq signal at peaks. Dashed lines mark regions for which fold-change in read density for \( \Delta \rho \)/WT was estimated. The fold-change for each instance is indicated on the graph. c, Same as b, with representative examples of nested antisense RNAs with increased expression upon \( \rho \) deletion (Fig. 3b). Three nested antisense RNAs were found in *E. coli* with identical criteria. See Supplementary Data 3 for a list of nested antisense RNAs identified.
Extended Data Fig. 7 | See next page for caption.
Extended Data Fig. 7 | Expressed pseudogenes with interrupted translation in *B. subtilis* show no polarity. Expressed pseudogenes endogenously present in the extant genome were used as additional independent experiments to assess the prevalence of Rho-mediated nonsense polarity in *B. subtilis* in situations of obligately uncoupled transcription and translation. Concomitant Rend-seq (mapping operon architecture) and ribosome profiling (measurement of translation) provides stringent data to determine translational status and transcript integrity of mRNAs. 

**a,** Schematic of analysis: for expressed pseudogenes (see Methods for selection criteria) with translation disruption, polarity was assessed by (1) comparing the mRNA read density at start and end of transcription unit, with large changes (<span class="MathJax数学公式">start/end ≫ 1</span>) indicative of polarity, and (2) fold change of pseudogene transcript upon *rho* deletion. Position of translation disrupting mutation is shown by ▲ and X. Dark and pale grey indicates region prior and after translation disruption mutation.

**b,** Rend-seq and ribosome profiling data for the 8 identified expressed pseudogenes. Each subpanel corresponds to a pseudogene region. Top traces show Rend-seq data (orange and blue signal correspond to summed 5′-mapped reads and 3′-mapped reads, peak shadows removed, see ref.36 for details on data processing). Orange peaks and blue peaks mark 5′ and 3′ boundaries of transcripts. Double line breaks (//) indicate truncated Rend-seq signal at peaks. Bottom traces show ribosome profiling data. Translation efficiency (ribosome profiling rpkm/Rend-seq rpkm) percentiles for each pseudogene sub-region (before and after translation disruption) are shown. Horizontal size marker provides positional scale (200 bp) on each subpanel. Nearby intact genes are shown in light blue. rpm: reads per million. Regions used to assess start to end decrease in RNA levels are marked by dashed lines. mRNA levels fold-changes (start/end, and ∆*rho*/WT) are shown. The *ydzW* region showed a second translation disruption the secondary frame, shown as a pale ▲ and X. See Methods, Fig. 3b and Supplementary Data 3 for details.
Polarity in pseudogenes:
Dashed lines depict regions at start and end of transcripts for which mRNA level is quantified and compared (start/end).

Translation efficiency percentiles:
- start/end=6.33
- start/end=3.81
- start/end=5.76
- start/end=1.05
- start/end=1.69

Extended Data Fig. 8 | See next page for caption.
Extended Data Fig. 8 | Most expressed pseudogenes with interrupted translation in *E. coli* show polarity. Similar to Extended Data Fig. 7. Expressed pseudogenes endogenously present in the extant genome were used as additional independent experiments to assess the prevalence of Rho-mediated nonsense polarity in *E. coli* in situations of obligately uncoupled transcription and translation. Concomitant Rend-seq (mapping operon architecture) and ribosome profiling (monitoring translation) provides stringent data to determine translational status and transcript integrity on mRNAs. 

**a.** Schematic of analysis: for expressed pseudogenes (see Methods for selection criteria) with translation disruption, polarity was assessed by comparing the mRNA read density at start and end of transcription unit, with large changes (start/end >1) indicative of polarity.

**b.** Rend-seq and ribosome profiling data for the identified expressed pseudogene with evidence of polarity. Each subpanel corresponds to a pseudogene region. Top traces correspond to Rend-seq data (orange and blue signal correspond to summed 5′-mapped reads and 3′-mapped reads, peak shadows removed, see ref.35 for details on data processing). Orange peaks and blue peaks mark 5′ and 3′ boundaries of transcripts. Double line breaks (//=) indicate truncated Rend-seq signal at peaks. Bottom traces show ribosome profiling data. Translation efficiency (ribosome profiling rpkm/Rend-seq rpkm) percentiles for each pseudogene sub-region (before and after translation disruption) are shown. Horizontal size marker provides positional scale (200 bp) on each subpanel. Light blue arrows correspond to nearby intact genes. rpm: reads per million. Regions used to assess start to end decrease in RNA levels are marked by dashed lines. mRNA levels fold-changes (start/end) are shown. The *gapC* region showed sequential translation disruptions secondary frames, shown as a pale ▲ and X.

**c.** Same as **b**, but for the two cases with no evidence of polarity. The translation disruptions mutation in *ykiA* and *cybC* are deletion of the beginning of ORFs. See Methods, Fig. 3b and Supplementary Data 3 for details.
Extended Data Fig. 9 | Analysis of C-to-G ratio for putative Rho-terminated RNAs. 

a. Cumulative distributions of maximum C-to-G ratio (“Max C:G”) of 100 nt sliding windows within non Rho-terminated coding sequences (CDSs, blue, $n = 2625$) and Rho-terminated CDSs (magenta, $n = 10$). Median of Max C:G is higher for Rho-terminated CDSs (magenta) than non Rho-terminated CDSs (blue) ($P < 10^{-5}$, less than one in $10^5$ random sub-samplings ($n = 10$) of non Rho-terminated distribution had higher median maximum C-to-G ratio). 

b. Cumulative distributions as in a for asRNAs that are not terminated by Rho (blue, $n = 112$) and asRNAs that are terminated by Rho (magenta, $n = 91$). Median of Max C:G is higher for Rho-terminated asRNAs than non Rho-terminated asRNAs ($P < 10^{-3}$, less than one in $10^3$ random sub-samplings ($n = 10$) of non Rho-terminated distribution had higher median maximum C to G ratio compared to sub-sampling ($n = 10$) of Rho-terminated distribution) (Methods).
Extended Data Fig. 10 | See next page for caption.
Extended Data Fig. 10 | Illustration of terminator identification pipeline and analysis of stem-to-stop distribution stratified by phyla. The terminator identification pipeline selects for strong hairpins immediately upstream of long U-tract found downstream of genes. Thresholds on hairpin folding free energy are determined on a species-by-species basis based on properties of randomly selected regions in respective genomes. The case of V. cholerae is illustrated in a–c. a, Results of folding $10^4$ regions of 40 nt chosen at random positions in the genome. Left panel shows the 2D distribution as a heatmap (dark positions corresponding to more density) of hairpin geometrical parameters (number of base pairs in stem $N_{bp}$, length of loop). Geometric thresholds are highlighted with blue dashes ($5 \text{ bp} \leq N_{bp} \leq 15 \text{ bp}$, $3 \text{ nt} \leq \text{Loop} \leq 8 \text{ nt}$) and retained region by blue shading. Right panel shows the 2D distribution as a heatmap (dark positions correspond to more density) of hairpin free energy of folding $\Delta G_{\text{hairpin}}$ and fraction of bases paired in stem $f$. Thresholds $\Delta G_1$ and $\Delta G_2$ on $\Delta G_{\text{hairpin}}$ are chosen such the total fraction of hairpin from random regions meeting geometrical (blue shading in left panel) and thermodynamic thresholds are 1% (orange, $\Delta G_{\text{hairpin}} \leq \Delta G_1$ and $f \geq 0.95$) and 1.5% (red, $\Delta G_{\text{hairpin}} \leq \Delta G_2$ and $f \geq 0.9$). b, Similar as for a, but for regions seeded by U-tracts (stretch of 5 or more consecutive T’s in the genome downstream of genes). Note the excess density of hairpins with strong energy of folding and large fraction of bases paired, corresponding to putative intrinsic terminators. c, Distribution of stop-to-stem distances for terminators passing thresholds shown in b. See Supplementary Data 2, Supplementary Data 3, and Methods for details of computational pipeline. d and e, Phylum stratified analysis on the stop-to-stem distribution. d, Each subpanel shows as a 2D greyscale the fraction of species within each phylum (shown in Fig. 4) for which more than fraction $F$ (y-axis) of terminators have stop-to-stem distances less than or equal to $D$ (x-axis). Black regions correspond to no species in the phylum, white all species. The contour line in the $(D,F)$ space marks points where 50% of species in the phylum have fraction $\geq F$ of their terminators with stop-to-stem distance $\leq D$. The yellow stars mark the thresholds used in Fig. 4 ($D = 12 \text{ nt}$, $F = 30\%$). For example, about 50% of species analysed in the Firmicutes have more than 30% of their terminators within 12 nt of upstream ORF (red contour line intersecting yellow star). e, The 50% species contour lines from d reported to the same panel, showing clear separation between phyla.
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.

- 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. mean) 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 wherever possible.
- 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

Software and code

Policy information about availability of computer code.

Data collection: Prokaryotic reference and representative genomes for RefSeq database were downloaded using Assembly from NCBI. Fluorescence measurements were collected using a BioTek Synergy H1 microplate reader. High-throughput sequencing data were from previous publications as referenced.

Data analysis: Northern blot intensities were quantified with ImageJ. Putative terminators were folded using RNAfold.

Sequencing data (already public, see “High-throughput expression datasets used” in Methods) was mapped to genomes using bowtie (version 1.0.1).

Generation of phylogenetic tree: Orthologous genes identified using BUSCO (version 3); orthologous orthologs from RefSeq species aligned using MAFFT (version 7); tree generated using FastTree (version 2.1.11); tree was re-rooted and ladderized using phytool (version 0.8-99).

Intrinsic terminators were identified using Matlab scripts that have been deposited on Github [blaine/Intrinsic_trx_terminator_identifier].

Read-seq data was analyzed using Matlab scripts deposited on Github [blaine/Read_seq_core_scripts].

Other data analysis and plotting was done using custom scripts written in either Python2.7 [calc_c_e_ratio.py] or Matlab2019b [diff_exp_mRNA_20190117.m, pseudo_gene_trace_display_20200316.m, non_contigious_operons_v2_20200429.m, display_FastTree_full_dist_rho_20200315.m, false_negatives_bioinfo_terms_20200314.m, correlation_stop_to_stem_readthrough_20191201.m, nested_asmRNA_trace_display_20200316.m]. These are available upon reasonable request.

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

Source data for each figure and extended data figure is included in this manuscript. High-throughput sequencing data used in this work are already available [see section "High-throughput expression datasets used" in Methods].

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 methods were used to pre-determine sample size. Each measurement is averaged over at least millions of cells. |
| Data exclusions | For mRNA and protein induction kinetic measurements, in order to obtain an accurate calculation of the induction time, we only fit points for which signal was increasing linearly (for mRNA) or quadratically (for protein). As such, some points were excluded from the fit as follows. For mRNA induction kinetic measurements, points with less than 1.5-fold increase over background were excluded from fit. For beta-galactosidase assays, points with signal below background and before signal began increasing consistently were excluded from fit. For lacZ complementation assays, points after which signal stopped increasing quadratically were determined as follows and excluded. A slope was calculated from the first two points included in the fit and the angle between this line and the next point calculated. This calculation was done for all subsequent pairs of points until an angle >20° between the resulting fit line and the following point was measured. This point and all subsequent points were excluded from the final fit. |

Given the diversity and heterogeneity of genomic elements annotated as pseudogenes, we had to manually curate these regions to ensure a high quality [expressed with clear translational disruption] set for polarity assessment. In B. subtilis, the following regions were annotated as pseudogenes but not applicable to our analysis: pseudogene tprC constituted a full ORF [the allele is rendered non-functional] by an in-frame 3 bp deletion which disrupts enzyme activity without interrupting translation, ywbZ constitutes a complete ORF paralogous to flagellar protein hag with no evidence of translational disruption, ywLB constitutes a paralog of gtaB with no evidence of translation disruption. In addition, two pseudogenes consisted of the C-terminal fragments ORFs interrupted by large-scale insertions: spsMc interrupted by SPβ prophage, sgkC interrupted by the 5′in element. Based on ribosome profiling data, sgkC had no clear decrease in translation compared to the upstream gene ywLb and was overall very lowly expressed. sgkC was thus not included in our final list of expressed pseudogenes. In E. coli, pseudogenes resulting from insertion element rearrangements were excluded due to difficulty in read-mapping and possible rapid genomic changes near these regions, pseudogene yabP and yfN had no measurable decrease in ribosome footprint density across the stop codon and were thus not included. In addition, three pseudogenes were excluded because of hard to interpret features: (1) ybcD has a 2 bp deletion in ATG either leading to an early stop codon or removal of the start codon [depending on the start codon position], was excluded because the ribosome profiling density was not consistent with either scenarios, suggesting a more complicated situation. (2) ybg was excluded despite clear evidence of decreasing 5′ to 3′ ramp [polarity] at the transcriptional level given that it constituted an uninterrupted ORF. (3) ybi [plausibly the C-terminal portion of a longer gene, based on comparison with E. coli O157, interrupted by the upstream gene ybaK in K-12 MG1655] was excluded despite clear evidence of decrease in transcription [polarity] because of a hard to interpret short RNA nested in the pseudogene body.

After manually inspecting all CDSs and aRNA identified as Rho-terminated, one identified Rho-terminated CDS (msmG) and one identified Rho-terminated aRNA (antisense to cydK) were excluded from analysis of CIR ratio as they contained WT-specific features within their 5′ ends that caused them to be identified by our analysis without showing clear signatures of Rho termination.

Some bioinformatically identified putative terminators were not included in our final set [independently of their stop-to-stem distance] for reasons other than their intrinsic properties to ensure highest quality set for stop-to-stem distance assessment. First, terminators arising from genomic elements with size less than 5% of the maximal chromosome size [e.g., small plasmid] were discarded. In addition, instances where multiple terminators were identified downstream of a gene were excluded. Repeated sequence/terminators [terminators with identical sequence and stop-to-stem distances] were excluded. Finally, terminators with stop-to-stem distance larger than 150 nt were excluded to avoid possible annotation errors.

| Replication | Transcription and translation kinetics were measured with n≥2 biological replicates [except for transcription kinetics in ΔyxKG background, which was repeated once], as all biological replicates gave consistent results. The specific sample size for each experiment is indicated in the figure legends. N=3 technical replicates were performed for RT-qPCR for transcription kinetic measurements and the mean is reported. Northern blotting (Fig. 2) was performed twice for B. subtilis [biological replicates] and once for E. coli. Results for both species were independently confirmed [biological replicates] by RT-qPCR. The sample size is noted in the figure legend. |
Measurements of steady state mRNA levels [measured by RT-qPCR, Fig. 3a, c] were performed for one biological replicate and three technical RT-qPCR replicates. Technical replicates were consistent and the mean reported.

Randomization
This study did not include experiments that required randomization, as there was no allocation of samples into groups.

Blinding
This study did not include experiments that required blinding, as all experiments were quantitative and there was no allocation of samples into groups.

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 | Methods |
|---------------------------------|---------|
| n/a | Involved in the study |
| 🗙 | Antibodies |
| ✗ | Eukaryotic cell lines |
| ✗ | Palaeontology |
| ✗ | Animals and other organisms |
| ✗ | Human research participants |
| ✗ | Clinical data |
| n/a | Involved in the study |
| ✗ | ChIP-seq |
| ✗ | Flow cytometry |
| ✗ | MRI-based neuroimaging |