Ubiquitous mRNA decay fragments in *E. coli* redefine the functional transcriptome

Lydia Herzel, Julian A. Stanley, James C. Taggart and Gene-Wei Li[#]

Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142, USA

[#]corresponding author
gwli@mit.edu

running title: RNA demographics in *E. coli*

keywords: RNA sequencing, translation, mRNA decay

Abstract

Bacterial mRNAs have short life cycles, in which transcription is rapidly followed by translation and degradation within seconds to minutes. The resulting diversity of mRNAs impacts their functionality but has remained unresolved. Here we quantitatively map the 3′ status of cellular RNAs in *Escherichia coli* during steady-state growth and report a large fraction of molecules (median>60%) that are fragments of canonical full-length mRNAs. The majority of RNA fragments are decay intermediates following endonuclease cleavage by RNase E and yet-unknown nucleases, whereas nascent RNAs contribute to a smaller fraction. Despite the prevalence of decay intermediates in total RNA, they are underrepresented in the pool of ribosome-associated transcripts and can thus distort quantifications for the abundance of full-length, functional mRNAs. The large heterogeneity within mRNA molecules *in vivo* highlights the importance in discerning functional transcripts and provides a lens for studying the dynamic life cycle of mRNAs.

Introduction

Cellular mRNAs are present in different forms – nascent, full-length, partially degraded, and others. Whereas full-length molecules are often depicted as the canonical mRNAs in the central dogma of molecular biology, the other forms are necessary intermediates of their birth-death cycle and have potential to (mis)interact with the same pool of factors that regulate mRNA functionality, such as translation. A complete molecular census of the mRNA population is thus critical for quantifying the functional transcriptome, while also providing an opportunity to reveal the life cycle of mRNAs *in vivo*.

In bacteria, functional half-lives of mRNAs are often short, which may give rise to a large heterogeneity among the transcripts that co-exist at the same time (Bernstein et al., 2002; Vargas-Blanco and Shell, 2020). After transcription initiates, mRNAs remain nascent for about a minute (considering a typical operon of 2 kilobases (kb) and a chain elongation rate of 45 nucleotides (nt)/s (Proshkin et al., 2010; Yu et al., 2006)). Meanwhile, RNA endonucleases, such as RNase E in *E. coli*, typically cleave mRNAs within a few minutes (Hui et al., 2014). Following endonuclease cleavage, the upstream RNA fragments are digested by exonucleases, such as PNPase and RNase

II, whereas the downstream fragments are further cleaved endonucleolytically. The timescales of these subsequent events are less characterized but can nevertheless determine the extent of decay intermediates present in the transcriptome (Fazal et al., 2015).

Amidst the heterogeneous RNA pool, translation takes place concurrently during transcription and mRNA decay in bacteria (Irastortza-Olaziregi and Amster-Choder, 2021). Whereas translation of nascent mRNAs is beneficial in *E. coli* by preventing Rho-dependent transcription termination, translation of mRNA decay intermediates may produce aberrantly truncated proteins and require rescue of ribosomes that are stalled at the end of RNA fragments (Keiler, 2015; Müller et al., 2021). In the meantime, some mRNA decay intermediates are translation-incompetent but often included as total RNA by RNA quantification methods. These considerations highlight the need to quantitatively resolve the forms of mRNA molecules across the transcriptome.

Existing methodologies for quantification primarily rely on regional information and are thus blind to distinct forms of RNAs. RT-qPCR and conventional RNA-seq detect ~100 bases at a time. Although Northern blotting could distinguish RNAs of different lengths, it only detects transcripts containing the short stretch of complementary sequence that is probed. Further, partial RNA fragments in Northern blots are often spread out over a large size range that dilutes the signal compared to full-length RNAs. Recently, several high-throughput single-molecule strategies have been employed to explore isoform structure, and could also be used to probe for transcriptome heterogeneity (SEnd-seq, Nanopore, SMRT-Cappable-seq) (Grünberger et al., 2021; Ju et al., 2019a; Yan et al., 2018). However, these methods involve at least one size-selection step that restricts the range of RNAs (or cDNAs) analyzed. Therefore, the full spectrum of RNA molecules present in the cell remains unmapped.

Here, we present a molecular census of RNA 3′ ends across the *E. coli* transcriptome. The 3′ terminal position of RNAs varies during their life cycle, and the relative representations can be quantified using methods that are not restricted to a limited size range in principle (Dar and Sorek, 2018). We find that, during steady-state growth, the majority of RNA molecules do not possess the mature 3′ ends of the corresponding transcription units (TUs). Instead, the 3′ termini are widely distributed across positions internal to TUs. Most RNA fragments at steady state are

decay intermediates and not nascently transcribed. The impact of the ubiquitous mRNA decay fragments on translation and transcriptome quantification were further investigated.

Results

A molecular census of *in vivo* mRNA states by 3′ end sequencing

To obtain a comprehensive molecular census of mRNA *in vivo*, two experimental conditions must be met. First, both the method RNA extraction and downstream detection should be unbiased towards different length and terminal nucleotide moieties. Second, the method of RNA extraction must minimize RNA fragmentation both during and after cell lysis. Here we describe our approach to meet these criteria.

We aimed to quantify all RNA states of production and decay to gain a molar representation of the total mRNA transcriptome (Fig. 1A). In principle, high throughput sequencing allows for single molecule and nucleotide resolution that is necessary to achieve this goal. Specifically, RNA 3′ end profiles obtained by RNA sequencing allow grouping of transcripts based on the location of their 3′ ends within a transcription unit (TU). 3′ ends mapping internally to individual transcription units (TUs) reflect either nascent or decaying transcripts, whereas 3′ ends mapping to the mature 3′ end reflect full-length transcripts and to some extent also decaying transcripts that vary in their 5′end (Fig. 1A). Thus, quantifying 3′ ends yields an upper bound for the molar fraction of full-length RNA Φ_{end} that can be refined in more detail by further analysis and experiments. An analogous representation could be achieved focusing on RNA 5′ ends. However, these can have different chemical properties, e.g. mono-, triphosphate or otherwise modified 5′ ends based on their point in the mRNA lifecycle (Bird et al., 2016; Chao et al., 2017; Ju et al., 2019a), which complicates uniform capture and thus is mainly amenable to relative comparisons, but is not quantitative for all populations. Strategies that capture both transcript ends have been used to annotate full-length transcripts (Grünberger et al., 2021; Ju et al., 2019a; Yan et al., 2018), but limitations in the captured size range by reverse transcription and downstream sequencing prevent a quantitative analysis of the transcriptome heterogeneity. Thus, we focused our study on RNA 3′ end distributions across individual TUs that provide an upper bound of the molar fraction of full-length RNA.

To capture the total mRNA transcriptome, we chose an RNA extraction method that retains molecules of various lengths and minimizes *in vitro* RNA fragmentation. We used several measures to evaluate the quality of our RNA extraction and library preparation, such as the total RNA profile, 23S and 5S to 16S rRNA ratios, spike-in quality, and correlation of Northern blot analysis with sequencing data and among replicates (Fig. 1B-C, E-G). We and others found that RNAsnap and Phenol-based RNA extraction methods accurately capture a diverse length range of RNAs in *E. coli* (Stead et al., 2012) (Fig. 1B). Except for Trizol extraction, we detected rRNA ratios distributing around 1.02 ± 0.22 (mean ± standard deviation), matching the expected 1:1 ratio *in vivo* (Fig. 1B) (Bremer, 1996; Dong et al., 1996). Additionally, we found that the quality of our five RNA spike-ins ranging from 28 to 1800nt stays high during RNAsnap and library preparation (Fig. 1C). Hence, we continued using RNAsnap for total RNA preparations and hot phenol-based RNA extraction for ribosome-associated RNA analysis. We conclude that our RNA extraction and sequencing strategy is largely unbiased and can accurately report on *in vivo* RNA end distributions.

Full-length transcripts constitute a small molar fraction of total mRNA

To analyze RNA 3′ end distributions without convoluting effects of alternative transcript isoforms, we focused our initial analysis on simple TUs, namely TUs that have a single transcript isoform as defined by RNA-end enriched RNA sequencing (Rend-seq) in our condition (Lalanne et al., 2018) (Methods). The simple TU harboring the protein-coding gene *ycaO* is shown in Fig. 1D. As expected, the mature 3′ end peak is the most prominent signal and illustrates the likely presence of full-length transcripts. However, we also observed substantial 3′ end coverage internal to the TU that suggests a large transcriptome heterogeneity with $\Phi_{end} = 0.57$. Confirming this, Northern blot analysis of the same TU detected full-length and partial transcripts of diverse lengths (Fig. 1E). We extended this analysis to other simple TUs that were selected based on their high expression and length in the range of 1 to 2kb, a range typical for protein-coding transcripts and well-suited for Northern blot analysis. We found good correspondence between full-length RNA quantification by Northern blot and 3′ end-seq for 8 simple TUs and between biological replicates (Fig. 1F, G). The distribution of Φ_{end} revealed that a typical transcription unit has at most

40% of full-length transcripts (median 0.4, Fig. 1H), reflecting substantial transcriptome heterogeneity based on intermediates in the lifecycle of an mRNA.

Confirmatory evidence comes from previously published studies in *E. coli*. We obtained similar results when analyzing *E. coli* Term-seq or SEnd-seq data that have been prepared using different RNA extraction and library design (Dar and Sorek, 2018; Ju et al., 2019a) (Fig. S1B-C, F). SEnd-seq enabled simultaneous detection of 5′ and 3′ ends, constrained by experimentally captured size of the underlying RNA transcript (Fig. S1G). Within the boundaries of accurate length representation (500-1500nt) it was possible to analyze the fraction of full-length transcripts and partially decaying transcripts that both have a mature 3′ end (Fig. S1F). For all simple TUs $\Phi_{end,\ SEnd\text{-}seq}$ was a combination of partial and full-length transcripts as initially expected, with some examples where full-length mRNA was completely absent. This reemphasizes that Φ_{end} reflects an upper bound for the fraction of full-length mRNAs and the actual fraction of full-length transcripts is often smaller than Φ_{end}.

Different bacterial species might arrive at different amounts of full-length transcripts at steady-state that we sought to test by comparing Φ_{end} in simple TUs of *E. coli* and *B. subtilis*. Interestingly, our analogous analysis of 3′ ends in *B. subtilis* yielded that more than half of all transcripts are possibly full-length (Fig. S1D-E), suggesting that different RNA synthesis and decay dynamics and mechanisms can yield species-specific steady-state levels of full-length mRNAs (Hui et al., 2014; Johnson et al., 2020). We conclude that the molar fraction of full-length RNA that we obtained by RNA sequencing and Northern blot analysis is small and to some extent species dependent.

Majority of partial transcripts are not nascent

Next, we quantified the relative contribution of nascent and decaying transcripts to the fraction of partial transcripts. If most partial transcripts were nascent, we expected internal 3′ ends from total RNA and nascent RNA to show strong similarities. Nascent transcripts are associated to RNA polymerase as part of the biochemically stable transcription elongation complex and can be purified by immunoprecipitation, followed by sequencing (Larson et al., 2014; Vvedenskaya et al., 2014). We enriched for nascent RNAs by immunoprecipitation of Flag-

tagged RNA polymerase prior 3′ end-seq (Fig. 2A, S2). As illustrated for the gene *rplQ*, internal 3′ end profiles differed strongly between total and nascent RNA (Fig. 2B). Furthermore, we recapitulated the elemental pause sequence identified in nascent RNA, but not in total RNA (Imashimizu et al., 2015; Larson et al., 2014; Vvedenskaya et al., 2014) (Fig. 2C). These two observations indicated that most partial transcripts were not nascent.

To estimate the pool of nascent RNAs from our data, we quantified the 3′ end signal within elemental pauses per TU in both total and nascent RNA. Identified RNA polymerase pauses are bonafide positions in the genome that should primarily yield RNA 3′ ends derived from nascent RNAs and the 3′ end signal at these sites in total RNA then reflect nascent RNAs in the pool of all transcripts. Based on this assumption, the ratio of relative signal at pause sites in total RNA to the nascent RNA sample provided an upper bound for nascent transcripts in the total RNA pool. The distribution median was 0.15, indicating that nascent RNAs contribute 15% of the diversity of a typical transcription unit (Fig. 2D). We conclude that most transcripts are not nascent and are thus decaying.

Majority of partial transcripts are decay intermediates

The high prevalence of RNA decay intermediates demands a correlation to mRNA stability and detection of *in vivo* signatures of RNA decay. If partial transcripts are largely RNA decay intermediates, then Φ_{end} should be maximal for stable RNAs independent of their length. Indeed, very long-lived RNAs spanning 2kb of size showed higher Φ_{end} than simple TUs (Fig. 3A). Additionally, Φ_{end} of simple TUs correlated positively with mRNA half-life (Fig. S3A, R (Pearson) = 0.46). The correlation to RNA half-life suggests a possible connection between 3′end profiles and RNA decay dynamics.

In *E. coli,* RNA decay occurs through cleavage and degradation by endonucleases and 3′-5′ exonucleases. RNase E is the main endonuclease and is essential in cells. PNPase is the main exonuclease whose function can be partly compensated by other 3′-5′ exonucleases, such as RNase II (Hui et al., 2014). To test, if RNase E cleavages could explain the internal 3′ end profiles, we compared our 3′ end-seq data to RNase E sites, defined by reanalysing data from (Clarke et al., 2014). 1449 of these 13244 sites and their vicinity had sufficient read coverage in 3′ end-seq

data for 3′ end profile analysis and a sequence motif, similar to what has been reported for *S. enterica* and *E. coli* previously (Chao et al., 2017). We observed a modest enrichment of 3′ ends at RNase E sites that was more pronounced in our PNPase deletion strain (Fig. 3B, S3B). Furthermore, partial transcripts were stabilized upon PNPase deletion, leading to a smaller fraction of mature 3′ ends (Fig. 3C-D). This suggests that most 3′ ends are generated by exonuclease digestion following endonucleolytic cleavage.

If endo- and exonucleolytic RNA decay intermediates contributed strongly to the pool of partial transcripts then adenylation of 3′ ends, another signature that influences these degradation steps, should be detectable. RNA 3′ end adenylation has been identified as a decay enhancing posttranscriptional RNA modification in *E. coli*, especially at transcription terminators (Kushner, 2015). Our approach captured these signatures and revealed that most steady-state A-tails are as short as a single nucleotide (Fig. 3E-G, S3). Due to the sequencing read length, our method is blind towards tails longer than 32nt but the tail length distribution and parsing the unmapped read fraction suggested that very few such tails exist at steady-state in fast growing *E. coli* cells (Fig. S3D). *In vivo* adenylation was absent in the deletion of the responsible polyA polymerase *pcnB* and increased as well as extended to internal sites of the transcription units when the exonuclease PNPase is missing (Fig. 3E-G, S3). The evidence presented in this section, namely the correlation of Φ_{end} with mRNA stability, the detection of *in vivo* adenylation and RNase cleavage signatures, supports the conclusion that most partial transcripts reflect RNA decay intermediates.

Some internal 3′ends localized in peaks, suggesting slower decay of these transcript fragments and the possibility for *cis*- and *trans*-regulatory features stabilizing fragments (Fig. 1E, 3E). RNA secondary structure can stabilize such intermediates (Dar and Sorek, 2018). Enhanced GC-content at stem loop structures is indicative of this and is detectable for 3′ end peaks outside of CDSs (Fig. 4A). For internal 3′peaks we also saw a GC-content enhancement, albeit at a shorter distance, and a pronounced spike 1nt upstream of the 3′peak. This could be consistent with a short stem loop upstream in some or all internal peaks and possibly a specific sequence associated with some peaks.

To resolve this, we plotted individual nucleotide abundances at positions near the peaks (Fig. 4B). This revealed a G-rich region and a preference for AUG one nucleotide upstream of the peak position. Classifying peaks by their trinucleotide preference one nucleotide upstream of the peak further substantiated that AUG is a preferred trinucleotide together with the triplet AAG and CUG and GUG. AAG, CUG and GUG differ from AUG in one nucleotide, and CUG and GUG are present as initiation codons for protein-coding genes in bacteria (Fig. 4C, S4A)(Hecht et al., 2017). AUG+1 peaks accounted for 10% of all 4210 internal peaks (28% considering all AUG/AAG/GUG/CUG+1), were distinct from RNA polymerase pause sites (3.8% overlap, $p = 1$, hypergeometric test), lacked upstream SD-like sequences, and showed equal presence in short and long partial transcripts (Fig. S4C-D). To test whether these ends are products of endo- or exonucleases, we combined 5′end- and 3′ end-seq and obtained matching peaks, strongly suggesting endonucleolytic activity (Fig. 4D). Interestingly, we found that this AUG+1 signature is conserved between *E. coli* and *B. subtilis* (Fig. S4A). Hence, it is likely that a similar mechanism generated these 3′ ends in both species and independently of species-specific endonucleases, such as RNase E/G and toxin RNases. In summary, we found prevalent and conserved AUG+1 peaks across transcription units that could be created in cases of improper translation initiation.

Translation primarily occurs on the small fraction of transcripts that are full-length (5)

To probe for the molar fraction of full-length and partial transcripts that are ribosome-associated we performed Northern blot analysis and 3′ end-seq across sucrose density gradients. This allowed for separation of the cytoplasm by density: fractions of low density are ribosome-free or contain ribosomal subunits, whereas increasing density separates monosomes from disomes and higher polysomes (Fig. 5A). Northern blot analysis for four TUs showed that all full-length transcripts associated with these TUs are ribosome-associated (Fig. 5A, S5A). Partial transcripts increased in size with increasing density in Northern blot analysis and by 3′ end-seq (Fig. 5A, B). To analyze the relative Φ_{end} difference between fractions, we normalized to Φ_{end} in total RNA. This shows that Φ_{end} increases beyond the for total RNA measured value in polysomes (Fig. 5C). These observations are consistent with the predominant translation of nascent and full-length mRNAs.

If indeed nascent and full-length transcript are the primary templates for translation, RNA polymerase pause intensity should be enriched in ribosome-bound fractions, as we observed (Fig. 5D). AUG+1 peaks were also more readily detected in higher polysome fractions substantiating that these partial transcripts are translated and possibly reflect improperly initiating ribosomes (Fig. S5B). However, these signatures corresponded to a small molar fraction in the polysome extracts and are thus rare events. Lastly, RNA decay signatures such as RNase E cleavage sites should be less prevalent in fractions enriched for ribosome-associated RNAs. Indeed, we found a substantially reduced RNase cleavage pattern for ribosome-associated fractions compared to the ribosome-free RNA (Fig. 5D). In conclusion, we present supporting evidence that, despite a high molar fraction of decaying transcripts in total RNA, most translation occurs on nascent and full-length transcripts.

Quantitation of the functional transcriptome requires discerning decay intermediates

Initiation of RNA decay co-translationally downstream of ribosomes can yield no-go or non-stop transcripts that stall ribosomes (Keiler, 2015; Müller et al., 2021). This can lead to incomplete protein synthesis and needs to be resolved to allow for future rounds of complete translation. Stalled ribosomes are rescued by *trans*-translation mediated by the RNA *ssrA* and the protein SmpB or cellular backup systems, that together are essential to the cells (Chadani et al., 2011, 2010). The cellular backup system mediated by ArfA is downregulated posttranscriptionally through the *trans*-translation pathway. However, when the *trans*-translation pathway is working at capacity, downregulation is relieved and *arfA* mRNA stabilized (Abo and Chadani, 2014).

Upon deletion of the exonuclease PNPase partial transcripts were stabilized (Fig. 3D), that could increase the need for ribosome rescue by elevated translation and ribosome binding of decay intermediates. Consistent with possibly diminished functional translation required for growth, the strain Δ*pnp* grew two times slower than wildtype (Fig. S6A). Furthermore, *ssrA* was three times more prevalent in ribosome-associated gradient fractions and alternative ribosome rescue factor *arfA* was strongly induced (Fig. S6B-F), suggesting a pronounced need for ribosome rescue due to decaying transcripts.

Despite our evidence towards translation stress in the strain Δ*pnp* previous work did not see pronounced differences in RNA half-lives and expression in that strain background (Bernstein et al., 2004; Lalanne et al., 2018; Pobre and Arraiano, 2015), leading us to question whether standard population-based RNA-seq is the best method to assess functional transcriptome changes based on perturbation of RNA decay. When we compared RNA-seq expression and Φ_{end} changes between wildtype and Δ*pnp*, we found that RNA-seq largely underestimated the pronounced effect on RNA decay that was visible in Φ_{end} changes for individual genes, as well as globally (Fig. 6A-C, S6G-I). Inhibition of mRNA translation, which generates ribosome-free stretches of RNA that could enable initiation of endonucleolytic RNA decay, also led to drastic Φ_{end} changes. However, co-occuring expression changes were less pronounced (Fig. 6D). This highlights that changes in the transcriptome diversity can occur in absence or precede measurable expression differences, e.g. in scenarios of altered mRNA decay or translation-targeting antibiotic stress. In conclusion, our study shows that end-based mapping approaches are important to study RNA decay and that prevalent decay intermediates that are thought to be cleared away quickly can intervene with functional gene expression and its quantification.

Discussion

We devised an approach to obtain a largely unbiased representation of *in vivo* transcriptome snapshots that reflect on synthesis and decay dynamics and their capabilities for translation. We find that fast growing *E. coli* cells have only few mature full-length mRNAs at a given time, which - when further reduced - has a substantial impact on the translation capacity of the cell. Most partial transcripts reflect *in vivo* RNA decay intermediates, albeit they are less prevalent in the ribosome-bound transcriptome. This means that the presence of decay intermediates in total RNA can distort quantifications for the abundance of full-length, functional mRNAs in RNA-seq measurements.

By profiling individual transcript ends of total RNA we obtain a single-nucleotide perspective of transient RNA decay dynamics that are possibly linked to other aspects of gene regulation. 3′ ends generated by endonucleoytic RNase cleavage are very unstable in *E. coli* due to the presence of 3′ to 5′ exonucleases (Hui et al., 2014). Hence, we were initially surprised to

pick up such cleavage signatures, but it also highlights that our approach generates snapshots of *in vivo* dynamics and transient intermediates of RNA decay. Consistent with previous studies, we detected a sizeable fraction of adenylated transcripts (Maes et al., 2017; Mohanty and Kushner, 2006) and found that very short tails of non-templated As are present at RNA 3′ ends at steady-state (Waters et al., 2017), probably reflecting the balance of concomitant adenylation and deadenylation by poly(A) polymerase PcnB and exonucleases, such as PNPase. Because of the high sensitivity of our assay and conservation of *in vivo* 3′ end signatures, we were able to detect novel endonucleolytic cleavage signatures that suggest an interconnection between spurious translation initiation and RNA decay (Fig. 4, S5). More experiments are needed to elucidate whether RNA cleavage precedes ribosome association or vice versa. However, the prevalence of AUG among other triplets that have been previously shown to serve as initiation codons suggests a link to translation initiation (Hecht et al., 2017; Sussman et al., 1996). Mutations in the initiation factor 3 *infC* were linked to increased initiation from non-AUG codons (Sussman et al., 1996), which makes it tempting to speculate that improper assembly of initiating ribosomes could precede the endonucleolytic cleavage.

Changes to transcriptome heterogeneity based on the abundance of mRNA decay intermediates are conceivable in conditions where mRNA decay is perturbed. This could influence how much protein can be synthesized. Upon deletion of non-essential enzymes involved in mRNA decay in *E. coli* and inhibition of translation initiation (Schneider et al., 1978), we detected a decreasing Φ_{end} with the strongest effect for the deletion of PNPase, which also showed a strong growth defect. PNPase is an enzyme with functions beyond RNA decay, e.g. in sRNA metabolism, bacterial immune defense by CRISPR, and is conserved across evolution, hence understanding its transcriptome dynamics associated with its function is important for bacterial physiology and beyond, e.g. also in eukaryotic organelles (Cameron et al., 2019; Dendooven et al., 2021; dos Santos et al., 2018; Pobre and Arraiano, 2015; Sesto et al., 2014). Despite its prevalent role in RNA decay and strong growth defect, average transcript half-lives and expression do not change substantially in exponentially growing *E. coli* (Bernstein et al., 2004; Lalanne et al., 2018; Pobre and Arraiano, 2015) (Fig. 6). However, whether transcriptome heterogeneity changes upon PNPase deletion reduce the capacity for translation by leading to

more stalled non-stop translation complexes was not explored. We found that the capacity for translation was reduced, which was reflected in the increased demand for ribosome rescue by *trans*-translation and through the alternative ribosome rescue factor *arfA* (Fig. S6). Similar responses have been detected when depleting the ribosome recycling factor RRF or overexpression of the ribosome-associated mRNA cleaving toxin RelE (Hwang and Buskirk, 2017; Saito et al., 2020a). In general, environmental responses and stress conditions that result in a pronounced ribosome rescue response might be derived from changes in transcriptome heterogeneity, e.g. due to altered mRNA decay.

The unchanged total RNA expression levels measured by RNA-seq, but strong changes in Φ_{end} upon PNPase deletion reveal limitations in RNA-seq to capture transcriptome heterogeneity based on RNA metabolism (Lalanne et al., 2018; Pobre and Arraiano, 2015)(Fig. 6). Conventional RNA-seq does not distinguish between partial and full-length transcripts and thus reports average expression values that can arise from multiple possible combinations of partial and full-length transcripts. Many partial transcripts are longer than typically applied small RNA size cutoffs, such as 200nt in many column-based RNA extraction methods (Fig. 1E, S1, 3C). Thus, this population can substantially contribute to the expression estimate despite being non-functional for protein production in many cases.

The high transcriptome diversity due to RNA decay intermediates requires consideration in all biological systems. In bacteria specifically, the life of an mRNA is shaped by co-occurrence of all steps of gene expression with little spatial separation in the cytoplasm, and thus with ample opportunities for crosstalk and interference. We quantified the resulting transcriptome distribution in this study in the two bacterial model systems *E. coli* and *B. subtilis*. Also in eukaryotes high transcriptome diversity due to RNA decay intermediates has been previously identified and utilized to reveal translation dynamics (Pelechano et al., 2015). Considering the endosymbiotic origin of organelles, gene expression there is subject to similar constraints as in bacteria. Hence, the observed heterogeneity of cellular mRNAs plays important roles in gene expression and cell physiology across species.

Methods

Strains, cell growth and harvest

Strains used in this study are listed in Table S1. To generate fresh gene deletions in the strain background MG1655, P1 transduction was done as described previously (Thomason et al., 2007). To harvest cells in fast exponential growth, cells were backdiluted from overnight cultures in the respective growth media to enable ~10 doublings prior harvest at an OD smaller or equal to 0.3 in 30ml (for total RNA analysis) or 500ml (for nascent RNA and polysome analysis) of growth media at 37°C. *E. coli* cells were grown in LB or MOPS Complete media supplemented with 0.2% Glucose. *B. subtilis* cells were grown in MCCG media. For total RNA analysis, 5ml of cells were transferred into 0.55ml of ice-cold stop solution (95:5 Ethanol:Phenol) in a 15ml conical tube, quickly inverted and spun at 3220g at 4°C for 5-10min. The supernatant was discarded and cell pellets flash-frozen in liquid nitrogen and stored at -80C until lysis and RNA extraction. For nascent RNA and polysome analysis, 250ml of cells were filter harvested at 37°C and scraped into liquid nitrogen (Johnson and Li, 2018). The cell paste was kept at -80°C until lysis and RNA extraction.

Genome version, annotation and simple transcription unit definition

For all experiments described here, the NCBI Reference Sequence and CDS annotation for the *E. coli* genome NC_000913.2 and the *B. subtilis* genome NC_000964.3 were used.

To quantify the abundance of intermediates that are generated during transcription and RNA decay, we decided to focus on simple transcription units (simple TUs) that have only one isoform, defined by a single 5′ end and a single 3′ end that we refer to as mature 5′ and 3′ end. The 5′ end is presumably a transcription start site, whereas the 3′ end can be generated by intrinsic termination, rho-dependent termination or a mix between those two or even RNA processing after termination at a site downstream of the mature 3′end. In order to classify regions in the *E. coli* and *B. subtilis* genome as simple TUs we utilized high-coverage Rend-seq data from *E. coli* MG1655 grown in the same conditions as for our 3′end-seq experiments (MOPS-Complete and LB, harvest at OD 0.3) and *B. subtilis* 168 grown in LB until OD 0.3 (Lalanne et al., 2018)(GSM2500131, GSM2500127, GSMXXX). This type of RNA-seq data allows to map transcript

isoform ends (peaks) with nucleotide precision (Lalanne et al., 2018). We called 5′ and 3′ peaks as described in (Lalanne et al., 2018) and a Z-score of 12 and a window size of 50nt for Z-score calculation. 5′ or 3′ peaks that were within +/-2nt in different datasets were considered as 1 peak as this corresponds to the expected peak width from these data and can have maxima that vary by +/-2nt around the center (Lalanne et al., 2018).

To classify transcription units into simple and complex TUs, we define a TU as a region in the genome that shows continuous Rend-seq coverage, has a 5′ and 3′ peak at its boundary and no Rend-seq coverage (defined by the median) until the next 5′ or 3′ peak, respectively. Therefore, we matched all 5′ ends that are upstream of any 3′ end and also identified the closest 5′ end downstream to a 3′ end. From our pooled Rend-seq data across the two growth conditions we calculated the median coverage across the window between 3′ end and closest 5′ end downstream. If this median was 0 the regions were kept separate, otherwise they were joined as one transcription unit to pair any 5′ and 3′ peaks within that unit to reflect transcript isoforms. Next, any possible combination of 5′ and 3′ peaks within such a TU was saved as possible isoform. This approach tends to overestimate transcription unit complexity. This was intended to then filter for TUs that have only one 5′ and 3′ end and are thus true simple TUs. Lastly, we size restricted isoforms to the size range within the 2-98% quantile of all isoforms (75-20872 nt for *E. coli*). This resulted in 7635 isoforms for *E. coli* from 3617 unique 5′ peaks and 1681 unique 3′ peaks. We excluded rRNA and tRNA regions from further analysis by keeping only non-overlapping isoforms using bedtools intersect (rRNA & tRNA annotation Table S3, bedtools/2.25.0 (Quinlan and Hall, 2010)). That reduced the number of isoforms to 5676.

To obtain a candidate set for simple TUs we overlapped the isoforms with each other using bedtools intersect and only kept the 215 isoforms that overlapped exclusively to each other. Manual inspection and curation in the genome browser IGV and MochiView followed (Homann and Johnson, 2010; Robinson et al., 2011), which reduced the number of simple TUs to 142 that were used for the 3′ end analysis reported in the main text (Table S4). The same approach (omitting the curation in the genome browser) yielded 418 simple TUs in *B. subtilis* that were used for the analysis shown in Fig. S1 (Table S5). It is likely that there are more simple TUs that we did not capture using our approach due to the restriction on looking at transcripts that are expressed

in exponential phase in rich media conditions. Furthermore, many 3′ ends of transcription units are less well defined compared to *E. coli* 5′ ends and thus some 5′ ends cannot be confidently linked to a 3 ′ end.

For analysis of 3′ and 5′ ends profiles with regard to RNase E cleavages, RNA polymerase pauses and non-templated adenylation, we focused on protein-coding gene regions. To this end, we grouped CDSs into TUs based on their distance to each other (<121nt) and strand orientation. To include expressed UTR region we extended these coarse-grained TUs by 60nt at each end (Table S6).

***In vitro* transcription and RNA spike-in generation**

Three of five RNA spike-ins were generated by *in vitro* transcription as outlined below. RNA spike-ins of short length were ordered as PAGE-purified RNA oligos from IDT (Table S2). To gain the ~0.1kb RNA spike-in, genomic DNA from *S. cerevisiae* was PCR amplified targeting the gene PDR5 with oligos oLH245 and oLH238 (Table S2), agarose gel extracted and column-purified with the Oligo Clean & Concentrator kit (Zymo Research). The DNA template was *in vitro* transcribed using the HiScribe T7 high yield RNA synthesis kit (NEB) in a 20µl reaction for 10h at 37°C. TurboDNase digestion with a final concentration of 0.1U/µl in 200µl, clean-up using the RNA Clean & Concentrator-25 kit (Zymo Research), gel extracted from a 10% polyacrylamide gel as described in (Johnson and Li, 2018) and finally quantified with the fluorometric Qubit RNA BR assay.

To generate a RNA spike-in control of ~1kb length, the pTXB1vector (NEB) was digested with BamHI, gel purified and *in vitro* transcribed using the HiScribe T7 high yield RNA synthesis kit (NEB) in a 20µl reaction for 10h at 37°C. TurboDNase digestion with a final concentration of 0.1U/µl in 200µl, clean-up using the RNA Clean & Concentrator-25 kit (Zymo Research), agarose gel analysis and quantification with the Qubit RNA BR assay followed.

To obtain a ~2kb RNA spike-in, the FLuc Control Template from the HiScribe T7 high yield RNA synthesis kit (NEB) was *in vitro* transcribed and purified as described the linearized pTXB1 vector.

Aliquots of each spike-in RNA in 10mM Tris pH 7.0 at the final concentration for pooling were kept at -80°C (1 spike-in sample for 1 RNA extraction contained approximately 6ng 28nt spike-in, 57ng 60nt spike-in, 53ng 120nt spike-in, 70ng pTXB1 spike-in, 170ng Fluc spike-in). Each pooled batch was sequenced by 3' end-sequencing without spiking it into a cell suspension to serve as an input control for quality normalization (4 'spike-in only' samples in Fig. 1C, normalized to mean Φ_{end}).

RNA purification

RNAsnap

RNA extraction using RNAsnap was done as previously described with small adjustments (Stead et al., 2012). Cell pellets were thawed for 4min in a 4°C pre-cooled tabletop centrifuge while spinning at 3220g. Tubes were transferred on ice and residual media was removed by pipetting. The cell pellets were resuspended in 500 µl RNA extraction solution (RES, 18mM EDTA, 0.025% SDS, 1% BME, 95% Formamide, RNA spike-ins). The RES was prepared fresh each time from stock components and RNA spike-in aliquots were added just before use. The RES-cell suspension was vortexed for 50s and incubated for 7min at 95°C. For *B. subtilis* cells, the RES-cell suspension was added to 0.2ml cold 0.1mm Zirconia beads, vortexed for 5min at maximum speed and incubated for 7min at 95°C. A 5min spin at room temperature at 20,000g followed. 400µl of supernatant were transferred to a fresh tube containing 1.6ml of DEPC-water. Subsequently, this 1:4 dilution was split into 4x 500µl aliquots for sodium acetate / isopropanol precipitation (details in 'RNA cleanup' section). Right after precipitation the sample was further purified using the RNA Clean & Concentrator-5 kit (Zymo Research) and DNase treated subsequently.

Acid-Phenol:Chloroform extraction (Hot phenol)

500µl Acid-Phenol:Chloroform, pH 4.5 (with IAA, 125:24:1, Thermo Fisher Scientific) containing 29 µl of 20% SDS was prewarmed to 65°C prior adding it to 500µl sample. The solution was incubated for 5min at 65°C in an Eppendorf Thermomixer at 1400rpm. Subsequently the solution was transferred to ice and chilled for 5min. After a 2min spin at room temperature and

20,000g 90% of the aqueous layer were transferred to a fresh tube and precipitated with Sodium Acetate / Isopropanol. As for RNAsnap, the sample was further purified using the RNA Clean & Concentrator-5 kit (Zymo Research) and DNase treated subsequently.

RiboPure RNA Purification Kit

RNA extraction with the RiboPure RNA Purification Kit (Thermo Fisher Scientific) was done as recommended by the manufacturer. Cell pellets were thawed and cleared off residual media as described for RNAsnap.

RNeasy RNA Purification Kit

Cell pellets were thawed and cleared off residual media as described for RNAsnap. Cells were resuspended in 100µl of 4 mg/ml lysozyme in 10mM Tris, pH 8.0 with spike-in RNA and incubated for 5min at 37°C. RNA was extracted as described in the manual without the use of gDNA columns. Instead, the sample was DNase treated and purified as described below. rRNA was removed from 20µg of RNA with 2 reactions of the MICROB*Express* Bacterial mRNA Enrichment Kit (ThermoFisher).

Trizol extraction

Cell pellets were thawed and cleared off residual media as described for RNAsnap. Cells were resuspended in 100µl of 4 mg/ml lysozyme in 10mM Tris, pH 8.0 and incubated for 5min at 37°C. 1.5ml Trizol was added and mixed by inversion. After 5min incubation at room temperature, 300ul of chloroform were added. After vigorous shaking for 15-30s the lysate was spun for 15min at 12,000g in a pre-cooled 4°C centrifuge. About 800µl of the aqueous phase were transferred to a fresh tube and precipitated with sodium acetate / isopropanol. As for RNAsnap, the sample was further purified using the RNA Clean & Concentrator-5 kit (Zymo Research) and DNase treated subsequently.

RNA cleanup and DNase treatment

To concentrate and cleanup RNA samples, Sodium Acetate - Ethanol or Isopropanol precipitation was carried out. 3M Sodium Acetate (pH 5.4, Life Technologies) was added to make up 1/10th of the sample volume. 1-2µl of GlycoBlue (Invitrogen) was used as co-precipitant. Depending on the sample volume either 3 volumes of ice-cold 100% Ethanol (<300ul) or 1 volume of ice-cold 100% Isopropanol was added. After thorough mixing, samples were precipitated at -80°C for at least 30min and spun for at least 60min in the pre-cooled 4°C centrifuge at 18213g. RNA pellets were washed with at least 250µl of ice-cold 80% Ethanol and spun for at least 4min. After aspiration by pipetting RNA pellets dried at room temperature for 3-5 minutes and were resuspended in 10mM Tris, pH 7.0. To cleanup samples further and remove traces of short DNA oligos post DNase treatment RNA samples were purified by RNA Clean & Concentrator-5 kit (Zymo Research). If this kit was used prior DNase treatment, elution was done with 85µl DEPC-water. Otherwise, samples were eluted into 10mM Tris, pH 7.0. For DNase treatment, 5µl 10x Turbo DNase buffer and 5µl 2U/µl TurboDNase (Thermo Fisher Scientific) were added to 85µl of sample and incubated for 20-30min at 37°C. RNA was cleaned up either with the RNA Clean & Concentrator-5 kit (Zymo Research) or by Ethanol precipitation.

RNA size selection for 3′ end analysis of transcripts in particular size ranges

To test whether some internal 3′ ends, e.g. AUG+1 peaks, were associated with particular transcript sizes, we probed RNAs in the length ranges from 100-400nt and greater than 400nt by 3′ end sequencing (Fig. S4C). To this end, we gel extracted 6.6 µg total RNA obtained using RNAsnap from a 6% TBE-Urea polyacrylamide gel (Invitrogen, ThermoFisher, 70 min 180V in 1xTBE) in the range of 100-400nt and cut everything above 400nt. Compared to a standard gel extraction during library preparation for smaller oligonucleotides, we extended the incubation period at 70°C in an Eppendorf reaction tube incubator at 1200rpm from 10 to 50 and 70min, respectively, to enhance extraction efficiency. After separation of gel debris and liquid we precipitated all samples using isopropanol precipitation. As quantified using the Qubit HS and BR RNA assay, our extraction efficiency was 78% in weight, with RNA > 400nt making up 67% of the 6.6 µg input. We proceeded with 0.4 µg of 100-400nt RNA and 1.8 µg of > 400nt RNA for 3′ end-seq library preparation. The dataset for RNAs > 400nt was used to compare Φ_{end} measured

through our approach with $\Phi_{end, SEnd}$ that we obtained from reanalyzing published SEnd-seq data (Fig. S1C).

Protein analysis

4-12% SDS polyacrylamide gel electrophoresis in combination with SyproRuby staining or membrane transfer, blocking and antibody incubation were done as described in the NuPAGE Technical Guide from Invitrogen / Life Technologies. Western blots of different fractionations during the RNA polymerase immunoprecipitation and polysome gradient fractionation were performed to assay efficiency of flag-tagged RNA polymerase enrichment using the monoclonal Anti-Flag M2 antibody from mouse (F1804, Sigma).

RNA polymerase immunoprecipitation

The RNA polymerase immunoprecipitation was carried out as described in (Larson et al., 2014) with the following modification. Nascent RNAs were extracted from the magnetic-bead bound fraction by adding 1.5ml Trizol to the sample and proceeding as described above. For comparison to total RNA, 100µl of lysate were diluted to 250µl in lysis buffer and RNA was extracted using Trizol as described above. Following DNase treatment and sample cleanup, rRNA was removed using the MICROB*Express* Bacterial mRNA Enrichment Kit (ThermoFisher) and precipitated with isopropanol. RNA levels were quantified fluorometrically with the Qubit RNA BR assay and 1.8µg of RNA was used for 3′ end-seq library preparation and 3-5µg for Rend-seq library preparation.

Polysome gradient centrifugation

Cell paste from 250ml of cells filter-harvested at or below OD 0.3 was cryo-lysed as done for ribosome profiling (Johnson and Li, 2018). 1M sodium chloride (Mohammad et al., 2019), 200U/ml SUPERaseIn RNase Inhibitor and 100U/ml TurboDNase were included in the lysis buffer in addition to the standard composition of 20mM Tris pH 8.0, 100mM ammonium chloride, 0.4% Triton X-100, 0.1% NP-40, 10mM magnesium chloride and 5mM calcium chloride. 250µl of lysate were loaded onto a 10-55% sucrose gradient and spun for 156min at 151263g with

a SW41 Ti rotor at 4°C. 12 gradient fractions were collected into disposable glass tubes for fraction collection with the same instrumentation as for ribosome profiling (Johnson and Li, 2018). The remaining liquid and pellet in the tube were either taken as a separate fraction (Fig. S6B) or pooled as a final fraction (Fig. S6A). Each fraction was transferred to a fresh tube and precipitated with ice-cold isopropanol as described above. The pellet was resuspended in 500µl 10mM Tris, pH 7.0 containing 0.5% SDS and RNA was extracted subsequently using hot phenol extraction as described above. Following DNase treatment and sample cleanup, RNA levels were quantified fluorometrically with the Qubit RNA BR assay and 3-5µg were used for Northern blot analysis and 1.8µg for 3′ end-seq library preparation.

PCR & RT-qPCR

Reverse transcription of RNA was done using SuperScript III reverse transcriptase (ThermoFisher). The protocol recommended for the enzyme was used in either 10 or 20 µl reactions. 100 ng of RNA were used as input. Random hexamers (ThermoFisher, 1ul per 20ul) were used for priming. A 'no RT primer' and a 'no enzyme' control were also included to assess random priming from residual short gDNA oligonucleotides and presence of contaminating gDNA, respectively. RNA, dNTPs and RT primer were denatured at 80°C for 3 min and immediately cooled on ice for at least 1 min. Afterwards RT buffer, 0.1M DTT and SuperaseIn (ThermoFisher) were added in the respective amounts. Superscript III enzyme was added last (or water for 'no enzyme' controls). Samples were incubated at room temperature for 5 min, then 55°C for 45 min. RNA was hydrolyzed at 95°C with 0.1M NaOH (final concentration) for 5 min and neutralized with HCl. cDNA samples were diluted 1:10 for further use in qPCR. For subsequent qPCRs, Kappa qPCR reaction mix (Roche) was used. In each well of a 96-well plate 5 µl 2x Kappa qPCR reaction mix, 3 µl of the 1 µM primer solution and 2 µl of the 1:10 diluted cDNA solution were added. Each sample, 'no RT primer', 'no enzyme' and a 'no cDNA' control were assayed in 2-3 technical replicates. Each run was done as recommended by the manufacturer. A difference in 10 or more between control and sample Ct value was considered valid. To quantify gene expression differences the double-Ct method was employed (Rao et al., 2013). Namely, Ct values were normalized by the expression of *cysG* that did not change in the

assayed conditions (Ct values highly similar in samples with same amount of RNA) and was used previously in similar experiments (Pobre and Arraiano, 2015). Following the internal normalization expression in the target strain was normalized to wildtype.

For conventional PCRs, Q5 DNA polymerase (NEB) was used with recommended reaction settings and reagent concentrations. 35-40 cycle PCR reactions were carried out. PCR products were analyzed by conventional 1.5% Agarose gel electrophoresis. Templates for *in vitro* transcription reactions were gel purified using the QIAquick Gel Extraction kit (Qiagen).

Northern blot

Northern blot analysis was done as described in (Lalanne et al., 2018) and DNA oligos used as probes are given in Table S2. For total RNA analysis 2% agarose gels were used. 1pmol of DNA probe was used for 5'terminal γ-^{32}P labeling followed by cleanup with ProbeQuant G-50 columns (Sigma) when assessing mRNA levels. To probe highly abundant RNAs, such as *ssrA* and *rnpB* 0.01pmol of DNA probe sufficed and enabled full removal of signal by stripping of transcript specific probes with three washes with boiling 0.1% SDS followed by an overnight.

Labeled membranes were exposed to a phosphor storage screen (GE Life Science) for hours to days depending on the signal strength and imaged with a 635nm laser scanner (Typhoon FLA9500, GE Life Sciences or Amersham Typhoon) with 750V and 100µm pixel size. The Amersham Typhoon Scanner Control Software 2.0.0.6 converts the original 32-bit data into a 16-bit data format for saving as tif-file that was used for further analysis in Fiji (Schindelin et al., 2012). Images were cropped to only include the membrane of interest and intensity profiles plotted per lane with a linewidth of 50 pixel. These profiles are shown in Fig. 1E, S1A, 3C, S6D. The fraction of full-length RNAs was quantified semi-automatically for Northern blots that showed uniform background signal across the full membrane and showed uniform sample and run characteristics for all lanes on the gel. The median background signal above the full-length signal was subtracted. The Northern blot shown in Fig. 1E showed a background distribution that increased linearly from top to bottom. Hence, we determined the median background signal above the full-length product and at the bottom of the blot, where no RNA was present and subtracted the linear profile between these two background intensities.

Upon log-transformation of the background-subtracted intensities for all lanes on one membrane, 2 inflection points, indicating the boundaries of full-length RNA to 1) no signal above and 2) partial transcript signal below, were defined. The signal between these two points reflects full-length RNA signal (FL) (shading in Fig. 1F). The signal below the 2nd inflection point reflects any signal of partial transcripts (P). The ratio of FL / (FL+P) was defined as the fraction of full-length RNAs and compared to Φ_{end} derived from 3′ end-seq (Fig. 1G).

RNA-seq, 5′ end-sequencing, Rend-seq

Oligos used for library preparations are listed in Table S2. RNA-seq for expression quantification in Fig. S6G-I was executed as described in (Parker et al., 2020). 5′ end-sequencing of 5′ phosphorylated RNAs was done modifying the RNA-seq protocol slightly. Namely, prior RNA fragmentation a 5′ end RNA oligo was ligated for 18h at 22°C to the RNA 5′ end of 1.8 μg total RNA using T4 RNA ligase 1 (NEB) and precipitated using isopropanol precipitation. RNA was fragmented for 2min and precipitated again. Subsequently, fragmented RNA was separated by size on a 10% TBE-Urea polyacrylamide gel and size extracted in the size range of 41-71nt and precipitated. 3′ end dephosphorylation and adenylation to enable uniform reverse transcription, reverse transcription, cDNA size selection and PCR were done as described in (Parker et al., 2020).

Rend-seq libraries were prepared as described in (Lalanne et al., 2018) from 20 μg of RNA prior rRNA removal. RNA fragmentation was done for 25s.

Final libraries were analyzed for size, quality and concentration using qPCR and the Fragment analyzer at the Biomicrocenter at MIT and sequenced on Illumina HiSeq 2000 or NextSeq500 machines. RNA-seq, 5′ end-seq and Rend-seq data processing and mapping was done as described in (Lalanne et al., 2018; Parker et al., 2020).

3′ end-sequencing

For 3′ end sequencing the Rend-seq protocol was modified in the following ways: 3′ end ligation with Linker-I was done with 1.8 μg of non-fragmented total or nascent RNA, followed by RNA precipitation using isopropanol. This ligation targets hydroxylated RNA 3′ ends that are typical products of transcription and canonical RNA decay (Yang, 2011). Phosphorylated RNA 3′

ends that can arise by non-enzymatic fragmentation or RNase A contamination are not ligated using this approach (Herzel, 2015; Yang, 2011). RNA was fragmented for 2min following ligation, which was followed again by RNA precipitation using isopropanol. 32-63 nt fragmented 3′ end-ligated RNA was size selected from a 15% TBE-Urea polyacrylamide gel and precipitated for reverse transcription, circularization and final PCR amplification with 6-8 cycles as described in (Lalanne et al., 2018).

In analogy to Rend-seq, 3′ end linker sequences were trimmed. To deal with non-template addition during reverse transcription, all reads were further trimmed with the fastx_trimmer of the fastxtoolkit/0.0.13 by 1nt at their 5′end and only reads of 15nt length or longer were kept for mapping with bowtie/1.2 using the options -v 0 -k 1 (Langmead et al., 2009). For further data analysis mapped data were transformed into bam-format and bedgraph-format using samtools/1.5 and bedtools/2.25.0 (Li et al., 2009; Quinlan and Hall, 2010). Coverage data include only 3′ terminal single nucleotide counts per read. To identify 3′adenylated transcript ends, the unmapped reads were trimmed using cutadapt/1.16 (Martin, 2014). First, 3′terminal CCAs were removed with the following settings: -a "CCA$" --minimum-length 15 -O 3 --trimmed-only. Second, A-tails of any length were trimmed (-a "A{38}" --minimum-length 15 -O 1 --trimmed-only). These trimmed reads were mapped and further converted with the same settings and code as untrimmed reads above. For total RNA 3′ end analysis mapped trimmed (i.e. A-tailed) and untrimmed 3′ end coverage profiles were summed.

Spike-in mapping

Unmapped reads after mapping 3′ end-sequencing to the *E. coli* genome (described above) were mapped with the same settings and code as above, but to the 5 spike-in template sequences (Table S2). In most cases T7 RNA polymerase adds a few non-templated nucleotides at the 3′ end of the template. To include these 3′ ends for full-length RNA analysis, unmapped reads after mapping to the spike-in template were five times iteratively trimmed by 1nt, filtered for length to only include reads of 15nt length or longer (fastx_trimmer -Q 33 -t 1 -m 15) and remapped to the template. After trimming 5nt only background mapping was observed and thus trimming

was stopped. Trimmed mapped reads and reads that mapped to the spike-in template prior trimming were merged and converted to bedgraph files using sam- and bedtools.

Expression, Φ_{end} calculation, rRNA ratio calculation

To quantify mRNA levels (rpkm) per coding sequence (CDS) from Rend-seq data, coverage at each base was divided by the total number of mapped reads in million yield reads per million (rpm). Next, the total coverage per CDS was calculated excluding 50 nt at both ends of the CDS to avoid conflating effects by end enrichment in Rend-seq and divided by the length of the CDS in kb minus 0.1 kb, yielding rpkm. Depending on the depth of the RNA-seq dataset either a minimum of 20 reads or 100 reads was required to consider an rpkm for further analysis.

The fraction of mature 3′ ends Φ_{end} was calculated as the ratio of 3′ ends mapping to the mature 3′ end of simple TUs (+/- 2nt based on the expected possible peak width (Lalanne et al., 2018)) and all 3′ end counts mapping to that TU (Table S7). A minimum of 20 reads per simple TU was required for further analysis. In Fig. S1C we report Φ_{end} for a dataset of > 400 nt size-selected 3′ end-seq data to roughly match the underlying size distribution of transcripts in SEnd-seq data (Fig. S1C, Fig. S4C).

The three very differently sized ribosomal RNAs (rRNAs) 5S (0.12kb), 16S (1.5kb) and 23S (2.9kb) are expressed at a molar stoichiometry of 1 and very stable with half-lives beyond the doubling time of *E. coli* (Bremer, 1996; Dong et al., 1996). Hence, we reasoned that these can be used as internal size standards and their mature 3′ end signal should reflect the 1:1 ratio in samples that were not depleted of rRNAs prior library preparation. That is the case for our RNAsnap and hot phenol extracted RNA samples (Fig. 1B). In addition to this analysis, Φ_{end} of rRNAs are also high above 0.8 even for the 23S rRNA (Fig. 3A).

Peak calling

3′ end or 5′ end coverage peaks are called based on Z-score calculation across a window of 100 nt. Specifically, read coverage was extracted in a region +/- 50 nt of any position covered by a read in the dataset and normalized by the total amount of reads in this window. To calculate mean and standard deviation of the background the center position was excluded. The Z-score

reflects the normalized coverage at the position of interest minus the background mean divided by the standard deviation. Z-scores above 7 and with 15 or more reads were considered peaks. These could be within or outside of CDSs.

To call RNase E sites from remapped data from (Clarke et al., 2014) we required further that Z-scores in the wildtype strain were above 7 and at least four times lower in the 5′ end-seq data in the rne-3071 ts strain (temperature-sensitive mutant of RNase E) and within our coarse-grained protein-coding transcription units. The RNase-dependent sites are listed in Table S9.

Quantification of nascent RNA fraction in total RNA

For the nascent RNA 3′ end data, 21355 positions had Z-scores above the cutoff of 7 and sufficient read coverage that were 2.1% of all positions passing read cutoff of 15 in the window around the position of interest. Peaks overlayed with CDS annotation using bedtools intersect and classified according to their relative position as in- or outside of CDSs (71% inside CDSs). To visualize the nucleotide distribution around peaks internal to CDSs, the underlying sequence 10nt upstream and 5nt downstream around peak positions was extracted from the genome fasta sequence and visualized as sequence logo with Weblogo 3 (Crooks et al., 2004).

To refine the classification of RNA polymerase pause positions for the quantification of nascent RNAs in total RNA, we calculate a pause motif score across the 16nt of interest and exclude the 25% lowest scoring sites from any further analysis. To this end, we calculate the position weight matrix from all sites in CDSs that have a peak with a Z-score greater 7. From that we generate a log-odds matrix by comparing to the nucleotide frequencies within protein-coding genes. For each individual 16nt peak sequence the pause motif score was calculated by summing up log-odds matrix entries at the respective sequence position (column) and base identity (row). Peaks with pause scores below 2 (25% lowest scoring peaks) do not show the sequence preference as given in Fig. 2C when visualized as sequence logo and were not used for the quantification of nascent RNAs in total RNA described below.

The peak distribution and sequence comparison between total and nascent RNA sample suggest little nascent RNA in the pool of total RNA (Fig. 2B-C). To estimate the nascent RNA fraction in total RNA we reasoned that the amount of total RNA 3′ ends at positions of RNA

polymerase pause sites provides an estimate of the contribution of nascent RNAs to the total transcriptome. However, not all nascent RNA 3'ends fall into these pause sites. Hence, we calculate the fraction of nascent RNA signal in pause sites compared to the whole transcription unit (total TU counts); the same is done for total RNA (without the adjustment for biochemical enrichment, see below). The ratio of the two fractions (total/nascent) is considered to reflect the "Nascent RNA fraction in total RNA", which is an upper bound for nascent RNA in the pool of total RNAs as also 3' ends at pause positions could come from decaying transcripts that happen to be caught at that position. Transcription units were defined in a coarse-grained way by grouping same-strand CDSs of less than 121nt distance and extending the ends of such grouped genes by 60nt to include the UTRs. Some 3'end signal across the whole transcription unit originates from total RNA due to the incomplete biochemical enrichment of nascent RNAs. To account for this, we multiply the total TU counts with 0.91, the median of the internal end fraction for simple TUs with 30 or more 3' end reads in the nascent RNA sample that indicates that at least 9% of 3'ends come from total RNA. The distribution of the "Nascent RNA fraction in total RNA" is given in Fig. 2D and Table S8 for simple TUs and all TUs. The medians of these distributions are 0.14 and 0.16, respectively.

Analysis around RNase E sites and 3' end peaks

Sequence logo of RNase E sites (Fig. 3B)

To visualize the nucleotide distribution around 5' end peaks that differ between the wildtype strain and the condition, where RNase E activity is perturbed based on a temperature-sensitive *rne* allele, we extracted the underlying sequence 5nt upstream and 3nt downstream at peak positions from the genome fasta sequence using bedtools getfasta and visualized them as sequence logo with Weblogo 3 (Crooks et al., 2004; Quinlan and Hall, 2010).

GC content, nucleotide distribution, SD-site identification (Fig. 4, S4)

To test whether RNA secondary structures could explain some or all 3' end peaks internal to CDSs we calculated the GC-nucleotide content around 3' end peaks in +/- 40 nt windows. The underlying sequence around peak positions was extracted from the genome fasta sequence using

bedtools getfasta (Quinlan and Hall, 2010). The number of G or Cs at each position was counted for all peaks and divided by the total number of peaks in the respective group and plotted.

Analogously, we calculated GC-content given in Fig. S4B for the 20 nt window upstream of positions defined in the different categories shown in Fig. S4B. Random sites refer to random positions within the genome, "CDS_AUG" to canonical start codon AUGs, "peak_start" to AUG+1 peaks that align with a canonical start codon, "peak_internal" to AUG+1 peaks that align to AUGs within CDSs, "any_internal" to AUGs that are internal to CDSs, but do not have an AUG+1 peak, "any_external" to AUGs that are outside of CDSs and also do not have an AUG+1 peak, "peak_BWG" to 3′ end peaks that are internal to CDSs and have either at CTG, GTG or AAG as triplett and peak_other to 3′ end peaks that have no specific sequence signature associated with them.

To visualize the nucleotide distribution around 3′ end peaks we extracted the underlying sequence from the genome sequence as described above and plotted the counts of any of the four nucleotides relative to the total number of nucleotides at this position around the peaks. In order to visualize the distribution of triplets one nucleotide upstream of 3′ end peaks, we count the abundance of any of the 64 tripletts among all 3′ end peaks internal to CDSs in different datasets of *E. coli* and *B. subtilis* and plot it as rank-ordered distribution. As control, the same number as peaks, but randomly localized within the genome were chosen and +1 nt triplets extracted and ordered in the same way as the actual peaks.

If AUG+1 peaks were consequences of initiating ribosomes, SD-like sequence motifs (SD = Shine Dalgarno) might be present in the proximity. Alternatively, no presence of SD-like motifs could justify that these sites are non-canonical points of translation initiation that might lead to truncation of the message, generating AUG+1 peaks. To distinguish between these possible scenarios, we sought to assess the likelihood of SD-like motifs in the 15 nt region upstream of the AUGs linked to AUG+1 peaks. First, we derived a SD sequence logo and underlying position weight matrix for canonical start codons using MEME (Bailey et al., 2015).

Top MEME motif frequency table to use a basis for log-odds matrix derivation

position	1	2	3	4	5	6

A	1203	1685	0	0	2008	648
C	854	250	0	0	0	203
G	581	253	2638	2638	337	1388
T	0	450	0	0	293	399

From that we generate a log-odds matrix by comparing to the nucleotide frequencies within protein-coding genes. For each position in the 15 nt window upstream the AUG from an AUG+1 peak, we calculated the SD-like motif score by summing up log-odds matrix entries at the respective sequence position (column) and base identity (row). Based on the SD-like motif score distribution for canonical AUGs we chose a SD-like motif score cutoff of 2 and count all peaks that show at least one SD-like motif in the region upstream of AUG+1 peaks. The error bars were estimated with 1000 iterations of bootstrapping (standard deviation from mean).

Normalized mean distributions around RNase E sites / nascent and total RNA 3' end peaks and relative to TU starts and start codons

For alignment of 3' and 5' end coverage relative to RNase E sites, RNA polymerase pauses and AUG+1 peaks, we focused on sites within protein-coding genes including adjacent UTRs (see coarse-grained TU definition above). We defined symmetrical 60 nt windows around these sites and extract coverage information for all positions around these sites. For further analysis, only sites were considered that met our read cutoff criteria (0.5 reads per position in wildtype & Δ*pnp* for RNase E sites in Fig. 3, S3, 0.15 reads per position in polysome gradient samples for RNA polymerase pause sites, 0.15 reads per position in polysome gradient samples for RNA polymerase pause sites, RNase E sites and AUG+1 peaks in Fig. 5D, S5B, 0.15 reads per position in 3' and 5' end data for AUG+1 peaks in Fig. 4D). To be able to consider different sites together despite differences in (local) expression, coverage around each site was normalized by the mean coverage in the 60 nt window. We refer to this as relative coverage. A value of 1 resembled the region mean, everything above 1 was increased relative to the mean and everything between 0 and 1 decreased. For visualization, the composite mean of all relative coverages is plotted for each position centered around the sites of interest. In Fig. 3B, S3 the full 60 nt window is shown, in Fig. 4D we focus on the +/- 10 nt window for visualization and in Fig. 5D, S5B we show the mean

signal at the respective site with error bars as standard error of the mean. For visualization as heatmap, the relative coverage for each site was plotted ranked by the RNase E site Z-score (see peak calling for the data in (Clarke et al., 2014)).

For alignment of 3′ end coverage relative to the start of CDSs, we focused on 816 first CDSs in our coarse-grained set of TUs. We added 100 nt at the start of each CDS to include 5′ UTRs in the coverage analysis. We filter CDSs based on length and read coverage (CDS > 150 nt, mean read coverage per 100 nt 5′UTR + CDS at least 0.01 in the lowest coverage dataset (fraction 1)). As described above for sites in a 60 nt window, we normalized coverage in each CDS by the mean coverage in the whole region. Next, we applied a running average smoothing per region within a 100 nt window. The smoothed relative coverage data were winsorized to the 95% quantile for each region. Finally, we computed the mean of the smoothed coverage at every position for -100 nt to 1900 nt across all CDSs that were at least as long as the respective position. Thus, more CDSs are included in the smaller distances to the start codon than towards later positions and the profile became noisier towards later positions (Fig. 4B).

Published data integration and considerations for SEnd-seq data analysis

Data from the following studies were remapped and integrated into the analysis described above:

- Term-seq from *E. coli* in LB at OD 0.5 (ERR2433552, ERR2433553, ERR2433554) and *B. subtilis* in LB OD 0.1-0.2 (ERR1232477, ERR1248362, ERR1248363) (Dar et al., 2016; Dar and Sorek, 2018).
- 5′ end-seq data rne-3071 ts (GSM1405878), rne wild-type (GSM1405877) (Clarke et al., 2014)
- Total wildtype 4-6 (SRR8449221, SRR8449222, SRR8449223) SEnd-seq (Ju et al., 2019a)
- Rend-seq data from *E. coli* MG1655 (GSM2500131) and *B. subtilis* 168 (GSM2500127)(Lalanne et al., 2018)
- Scripts used for read processing and mapping can be found on https://github.com/gwlilabmit/Herzel2022_RNAdemographics.

We used the reported mRNA half-lives from (Moffitt et al., 2016) to compare to our Φ_{end} distribution (Fig. S3A). Only CDSs were considered with half-lives differing less than twofold between τ-seq replicates.

SEnd-seq enables to sequence transcripts and retain 5′ and 3′ end information. Hence, a comparison to our 3′ end focused data has the potential to strengthen our conclusions. For full-length transcript analysis with SEnd data we remapped the published data, overlapped SEnd-seq transcripts with our simple TUs and utilized 3′ ends of SEnd-reads to compute $\Phi_{end,\ SEnd}$. During this process we noticed that SEnd reads are generally longer than 400nt in the wildtype samples. That limits the scope in which the fraction of mature 3′ ends Φ_{end} can be accurately reported from these data. Nevertheless, a comparison to a 3′ end-seq dataset, where we size-selected transcripts greater than 400nt by gel extraction shows a decent correlation (Fig. S1C).

To be able to consider the information gained from an analysis of the joined 5′ and 3′ end information present in SEnd data, we wanted to understand whether SEnd data show also limitations in the representation of longer transcripts. To this end, we assessed the size distribution of partial SEnd transcripts that we expected to scale linearly with the length of the underlying TU based on our data (Northern blot analysis and overall uniform 3′ end distributions in total RNA (Fig. S1A, S4C, 5B)). Thus, the comparison of SEnd partial transcript lengths should yield a linear correlation, if the SEnd length distribution is dominated by its *in vivo* length distribution. We correlated the median partial transcript length for individual simple TUs with the simple TU length (Fig. S1G). This correlation is linear in the range of 500-1500nt, but plateaus at shorter and longer sizes, which suggests that quantitative recovery of different *in vivo* transcript sizes is limited to the range between 500-1500nt in SEnd-seq. During SEnd-seq library preparation RNA and cDNA cleanup steps with cutoffs for short RNA/DNA molecules result in the underrepresentation of short transcripts. For longer than 1500nt transcripts incomplete reverse transcription, size limitations during intramolecular circularization and/or dropout and underamplification of long DNA molecules during PCRs could explain the length bias.

To determine the fraction of partial transcripts under the above given constraints, but nevertheless with both 5′ and 3′ end information, we classified transcripts into 4 classes: 1) full-length, when 5′ and 3′ end were within the 5nt of the mature ends as defined by our simple TUs,

2) transcripts with a mature 5′ end and varying 3′ end, 3) transcripts with a mature 3′ end and varying 5′ end or 4) internal transcripts that varied both in their 5′ and 3′ end. We show the fraction of all 4 classes in Fig. S1F. We required at least 20 SEnd reads per simple TU and that simple TUs are longer than 600nt and indicated the 1500nt boundary with a dashed line in Fig. S1F. This shows that the fraction of full-length transcripts decreases with TU size (also true for our Φ_{end} quantified by 3′ end-seq) and all TUs have also a fraction of transcripts that are bounded by a mature 3′ end, but vary in their 5′ end (light green). Some do not have any full-length transcripts at all, consistent with a predication made by modeling transcription and RNA decay in (Chen et al., 2015). That emphasizes that Φ_{end} is indeed an upper bound to estimate the fraction of full-length transcripts and a typical TU has less full-length transcripts than the by us reported median of 40%, which includes both partial transcripts with a mature 3′ end and full-length transcripts.

Data analysis presented in this manuscript was done using bash scripts and software as cited above, python 2.7, matlab R2021b, Rstudio (Version 1.2.5042) and on the high-performance cluster LURIA maintained by the Biomicrocenter at MIT.

Data access

The sequencing data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSEXXX. Scripts used for read processing and mapping can be found on github.com/gwlilabmit/Herzel2022_RNAdemographics.

Acknowledgements

The authors thank Matthew H. Larson and Jason Peters for sharing the Flag-tagged RpoC strains, Tania A. Baker, Christopher B. Burge, Eliezer Calo, Joseph H. Davis, Alan D. Grossman, Rebecca L. Lamason and Michael T. Laub and their labs for sharing equipment and strains, Jean-Benoit Lalanne and Cassandra Schaening for providing python and matlab scripts for peak calling,

Chun-Chen Yao for help with initial testing of the RNAsnap protocol and 3′ end library preparation in different *E. coli* wildtype strain backgrounds, Mirae Parker for assistance with the initial execution of the in-house RNA-seq protocol, Matthew Z. Tien for the initial version of the in-house Northern blot protocol and explanation, and Michael R. Weber for comments on the manuscript. The work was supported by funding by the National Science Foundation (NSF) Career award to GWL. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NSF. During part of this work, LH was a postdoctoral fellow of the Helen Hay Whitney foundation.

Author contributions

LH and GWL conceived the study. JAS performed a P1 tranduction to obtain replicate strains for *Δpnp*, measured the bacterial growth shown in Fig. S6A, performed polysome gradient centrifugation and Northern blot analysis of one wildtype and *Δpnp* replicate (Fig. S6B-D), and performed the +/- Kasugamycin treatment and Rend- and 3′ end-seq library preparation, data processing, mapping and expression quantification (Fig. 6D). JCT developed the in-house 5′ end-seq protocol that LH used to generate the data in Fig. 4D. LH performed all other experiments, analysis and prepared all the figures. LH wrote the first draft of the manuscript. GWLi and JAS edited the manuscript.

Disclosure declaration

The authors declare no conflict of interest.

Supplemental Material

Supplemental Table 1: Strains used in this study.
Supplemental Table 2: Oligonucleotides used in this study.
Supplemental Table 3: rRNA annotation used in this study.
Supplemental Table 4: Simple transcription unit annotation *E. coli*.
Supplemental Table 5: Simple transcription unit annotation *B. subtilis*.
Supplemental Table 6: Coarse-grained transcription unit annotation *E. coli*.
Supplemental Table 7: Φ_{end} from 3′ end-seq.
Supplemental Table 8: Nascent RNA fraction.
Supplemental Table 9: RNase E-dependent 5′ends.

References

Abo, T., Chadani, Y., 2014. The fail-safe system to rescue the stalled ribosomes in Escherichia coli. Front. Microbiol. 5, 156. https://doi.org/10.3389/fmicb.2014.00156

Bailey, T.L., Johnson, J., Grant, C.E., Noble, W.S., 2015. The MEME Suite. Nucleic Acids Res. 43, W39–W49. https://doi.org/10.1093/nar/gkv416

Bernstein, J.A., Khodursky, A.B., Lin, P.-H., Lin-Chao, S., Cohen, S.N., 2002. Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. Proc. Natl. Acad. Sci. U. S. A. 99, 9697–9702. https://doi.org/10.1073/pnas.112318199

Bernstein, J.A., Lin, P.-H., Cohen, S.N., Lin-Chao, S., 2004. Global analysis of Escherichia coli RNA degradosome function using DNA microarrays. Proc. Natl. Acad. Sci. U. S. A. 101, 2758–2763. https://doi.org/10.1073/pnas.0308747101

Bird, J.G., Zhang, Y., Tian, Y., Panova, N., Barvík, I., Greene, L., Liu, M., Buckley, B., Krásný, L., Lee, J.K., Kaplan, C.D., Ebright, R.H., Nickels, B.E., 2016. The mechanism of RNA 5′ capping with NAD(+), NADH and desphospho-CoA. Nature. https://doi.org/10.1038/nature18622

Bremer, D., 1996. Modulation of Chemical Composition and Other Parameters of the Cell by Growth Rate.

Cameron, T.A., Matz, L.M., Sinha, D., De Lay, N.R., 2019. Polynucleotide phosphorylase promotes the stability and function of Hfq-binding sRNAs by degrading target mRNA-derived fragments. Nucleic Acids Res. 47, 8821–8837. https://doi.org/10.1093/nar/gkz616

Chadani, Y., Ono, K., Kutsukake, K., Abo, T., 2011. Escherichia coli YaeJ protein mediates a novel ribosome-rescue pathway distinct from SsrA- and ArfA-mediated pathways. Mol. Microbiol. 80, 772–785. https://doi.org/10.1111/j.1365-2958.2011.07607.x

Chadani, Y., Ono, K., Ozawa, S., Takahashi, Y., Takai, K., Nanamiya, H., Tozawa, Y., Kutsukake, K., Abo, T., 2010. Ribosome rescue by Escherichia coli ArfA (YhdL) in the absence of trans-translation system. Mol. Microbiol. 78, 796–808. https://doi.org/10.1111/j.1365-2958.2010.07375.x

Chao, Y., Li, L., Girodat, D., Förstner, K.U., Said, N., Corcoran, C., Śmiga, M., Papenfort, K., Reinhardt, R., Wieden, H.-J., Luisi, B.F., Vogel, J., 2017. In Vivo Cleavage Map Illuminates the Central Role of RNase E in Coding and Non-coding RNA Pathways. Mol. Cell 65, 39–51. https://doi.org/10.1016/j.molcel.2016.11.002

Chen, H., Shiroguchi, K., Ge, H., Xie, X.S., 2015. Genome-wide study of mRNA degradation and transcript elongation in Escherichia coli. Mol. Syst. Biol. 11, 808. https://doi.org/10.15252/msb.20159000

Clarke, J.E., Kime, L., Romero A., D., McDowall, K.J., 2014. Direct entry by RNase E is a major pathway for the degradation and processing of RNA in Escherichia coli. Nucleic Acids Res. 42, 11733–11751. https://doi.org/10.1093/nar/gku808

Crooks, G.E., Hon, G., Chandonia, J.-M., Brenner, S.E., 2004. WebLogo: A Sequence Logo Generator. Genome Res. 14, 1188–1190. https://doi.org/10.1101/gr.849004

Dar, D., Shamir, M., Mellin, J.R., Koutero, M., Stern-Ginossar, N., Cossart, P., Sorek, R., 2016. Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. Science 352. https://doi.org/10.1126/science.aad9822

Dar, D., Sorek, R., 2018. Extensive reshaping of bacterial operons by programmed mRNA decay. PLoS Genet. 14, e1007354. https://doi.org/10.1371/journal.pgen.1007354

Dendooven, T., Sinha, D., Roeselová, A., Cameron, T.A., Lay, N.R.D., Luisi, B.F., Bandyra, K.J., 2021. A cooperative PNPase-Hfq-RNA carrier complex facilitates bacterial riboregulation. Mol. Cell 81, 2901-2913.e5. https://doi.org/10.1016/j.molcel.2021.05.032

Dong, H., Nilsson, L., Kurland, C.G., 1996. Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. J. Mol. Biol. 260, 649–663. https://doi.org/10.1006/jmbi.1996.0428

dos Santos, R.F., Quendera, A.P., Boavida, S., Seixas, A.F., Arraiano, C.M., Andrade, J.M., 2018. Chapter Five - Major 3′–5′ Exoribonucleases in the Metabolism of Coding and Non-coding RNA, in: Teplow, D.B. (Ed.), Progress in Molecular Biology and Translational Science. Academic Press, pp. 101–155. https://doi.org/10.1016/bs.pmbts.2018.07.005

Fazal, F.M., Koslover, D.J., Luisi, B.F., Block, S.M., 2015. Direct observation of processive exoribonuclease motion using optical tweezers. Proc. Natl. Acad. Sci. 112, 15101–15106. https://doi.org/10.1073/pnas.1514028112

Grünberger, F., Ferreira-Cerca, S., Grohmann, D., 2021. Nanopore sequencing of RNA and cDNA molecules expands the transcriptomic toolbox in prokaryotes. https://doi.org/10.1101/2021.06.14.448286

Hecht, A., Glasgow, J., Jaschke, P.R., Bawazer, L.A., Munson, M.S., Cochran, J.R., Endy, D., Salit, M., 2017. Measurements of translation initiation from all 64 codons in E. coli. Nucleic Acids Res. 45, 3615–3626. https://doi.org/10.1093/nar/gkx070

Herzel, L., 2015. Co-transcriptional splicing in two yeasts (Dissertation). Technical University Dresden, Dresden, Germany.

Homann, O.R., Johnson, A.D., 2010. MochiView: versatile software for genome browsing and DNA motif analysis. BMC Biol. 8, 49. https://doi.org/10.1186/1741-7007-8-49

Hui, M.P., Foley, P.L., Belasco, J.G., 2014. Messenger RNA Degradation in Bacterial Cells. Annu. Rev. Genet. 48, 537–559. https://doi.org/10.1146/annurev-genet-120213-092340

Hwang, J.-Y., Buskirk, A.R., 2017. A ribosome profiling study of mRNA cleavage by the endonuclease RelE. Nucleic Acids Res. 45, 327–336. https://doi.org/10.1093/nar/gkw944

Imashimizu, M., Takahashi, H., Oshima, T., McIntosh, C., Bubunenko, M., Court, D.L., Kashlev, M., 2015. Visualizing translocation dynamics and nascent transcript errors in paused RNA polymerases in vivo. Genome Biol. 16, 98. https://doi.org/10.1186/s13059-015-0666-5

Irastortza-Olaziregi, M., Amster-Choder, O., 2021. Coupled Transcription-Translation in Prokaryotes: An Old Couple With New Surprises. Front. Microbiol. 11, 3532. https://doi.org/10.3389/fmicb.2020.624830

Johnson, G.E., Lalanne, J.-B., Peters, M.L., Li, G.-W., 2020. Functionally uncoupled transcription–translation in Bacillus subtilis. Nature 1–5. https://doi.org/10.1038/s41586-020-2638-5

Johnson, G.E., Li, G.-W., 2018. Chapter Ten - Genome-Wide Quantitation of Protein Synthesis Rates in Bacteria, in: Carpousis, A.J. (Ed.), Methods in Enzymology, High-Density Sequencing Applications in Microbial Molecular Genetics. Academic Press, pp. 225–249. https://doi.org/10.1016/bs.mie.2018.08.031

Ju, X., Li, D., Liu, S., 2019a. Full-length RNA profiling reveals pervasive bidirectional transcription terminators in bacteria. Nat. Microbiol. 4, 1907–1918. https://doi.org/10.1038/s41564-019-0500-z

Ju, X., Li, D., Liu, S., 2019b. Full-length RNA profiling reveals pervasive bidirectional transcription terminators in bacteria. Nat. Microbiol. 4, 1907–1918. https://doi.org/10.1038/s41564-019-0500-z

Keiler, K.C., 2015. Mechanisms of ribosome rescue in bacteria. Nat. Rev. Microbiol. 13, 285–297. https://doi.org/10.1038/nrmicro3438

Kushner, S.R., 2015. Polyadenylation in E. coli: a 20 year odyssey. RNA 21, 673–674. https://doi.org/10.1261/rna.049700.115

Lalanne, J.-B., Taggart, J.C., Guo, M.S., Herzel, L., Schieler, A., Li, G.-W., 2018. Evolutionary Convergence of Pathway-Specific Enzyme Expression Stoichiometry. Cell 173, 749-761.e38. https://doi.org/10.1016/j.cell.2018.03.007

Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25. https://doi.org/10.1186/gb-2009-10-3-r25

Larson, M.H., Mooney, R.A., Peters, J.M., Windgassen, T., Nayak, D., Gross, C.A., Block, S.M., Greenleaf, W.J., Landick, R., Weissman, J.S., 2014. A pause sequence enriched at translation start sites drives transcription dynamics in vivo. Science 344, 1042–1047. https://doi.org/10.1126/science.1251871

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Maes, A., Gracia, C., Innocenti, N., Zhang, K., Aurell, E., Hajnsdorf, E., 2017. Landscape of RNA polyadenylation in E. coli. Nucleic Acids Res. 45, 2746–2756. https://doi.org/10.1093/nar/gkw894

Martin, M., 2014. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal 17 h ttp. journal. embnet. org/index. php/embnetjournal/article/ ….

Moffitt, J.R., Pandey, S., Boettiger, A.N., Wang, S., Zhuang, X., 2016. Spatial organization shapes the turnover of a bacterial transcriptome. eLife 5, 636. https://doi.org/10.7554/eLife.13065

Mohammad, F., Green, R., Buskirk, A.R., 2019. A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution. eLife 8. https://doi.org/10.7554/eLife.42591

Mohanty, B.K., Kushner, S.R., 2006. The majority of Escherichia coli mRNAs undergo post-transcriptional modification in exponentially growing cells. Nucleic Acids Res. 34, 5695–5704. https://doi.org/10.1093/nar/gkl684

Molecular Biology of the Cell [WWW Document], n.d. URL https://wwnorton.com/books/9780815344322 (accessed 11.18.21).
Müller, C., Crowe-McAuliffe, C., Wilson, D.N., 2021. Ribosome Rescue Pathways in Bacteria. Front. Microbiol. 12, 558. https://doi.org/10.3389/fmicb.2021.652980
Parker, D.J., Lalanne, J.-B., Kimura, S., Johnson, G.E., Waldor, M.K., Li, G.-W., 2020. Growth-Optimized Aminoacyl-tRNA Synthetase Levels Prevent Maximal tRNA Charging. Cell Syst. 11, 121-130.e6. https://doi.org/10.1016/j.cels.2020.07.005
Pelechano, V., Wei, W., Steinmetz, L.M., 2015. Widespread Co-translational RNA Decay Reveals Ribosome Dynamics. Cell 161, 1400–1412. https://doi.org/10.1016/j.cell.2015.05.008
Pobre, V., Arraiano, C.M., 2015. Next generation sequencing analysis reveals that the ribonucleases RNase II, RNase R and PNPase affect bacterial motility and biofilm formation in E. coli. BMC Genomics 16, 72. https://doi.org/10.1186/s12864-015-1237-6
Proshkin, S., Rahmouni, A.R., Mironov, A., Nudler, E., 2010. Cooperation between translating ribosomes and RNA polymerase in transcription elongation. Science 328, 504–508. https://doi.org/10.1126/science.1184939
Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. https://doi.org/10.1093/bioinformatics/btq033
Rao, X., Huang, X., Zhou, Z., Lin, X., 2013. An improvement of the 2ˆ(–delta delta CT) method for quantitative real-time polymerase chain reaction data analysis. Biostat. Bioinforma. Biomath. 3, 71.
Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., Mesirov, J.P., 2011. Integrative genomics viewer. Nat. Biotechnol. 29, 24–26. https://doi.org/10.1038/nbt.1754
Saito, K., Green, R., Buskirk, A.R., 2020a. Ribosome recycling is not critical for translational coupling in Escherichia coli. eLife 9, e59974. https://doi.org/10.7554/eLife.59974
Saito, K., Green, R., Buskirk, A.R., 2020b. Translational initiation in E. coli occurs at the correct sites genome-wide in the absence of mRNA-rRNA base-pairing. eLife 9. https://doi.org/10.7554/eLife.55002
Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.-Y., White, D.J., Hartenstein, V., Eliceiri, K., Tomancak, P., Cardona, A., 2012. Fiji: an open-source platform for biological-image analysis. Nat. Methods 9, 676–682. https://doi.org/10.1038/nmeth.2019
Schneider, E., Blundell, M., Kennell, D., 1978. Translation and mRNA decay. Mol. Gen. Genet. MGG 160, 121–129. https://doi.org/10.1007/BF00267473
Sesto, N., Touchon, M., Andrade, J.M., Kondo, J., Rocha, E.P.C., Arraiano, C.M., Archambaud, C., Westhof, É., Romby, P., Cossart, P., 2014. A PNPase Dependent CRISPR System in Listeria. PLOS Genet. 10, e1004065. https://doi.org/10.1371/journal.pgen.1004065
Stead, M.B., Agrawal, A., Bowden, K.E., Nasir, R., Mohanty, B.K., Meagher, R.B., Kushner, S.R., 2012. RNA snap ™: a rapid, quantitative and inexpensive, method for isolating total RNA from bacteria. Nucleic Acids Res. 40, e156–e156. https://doi.org/10.1093/nar/gks680
Sussman, J.K., Simons, E.L., Simons, R.W., 1996. Escherichia coli translation initiation factor 3 discriminates the initiation codon in vivo. Mol. Microbiol. 21, 347–360. https://doi.org/10.1046/j.1365-2958.1996.6371354.x

Thomason, L.C., Costantino, N., Court, D.L., 2007. E. coli genome manipulation by P1 transduction. Curr. Protoc. Mol. Biol. Ed. Frederick M Ausubel Al Chapter 1, Unit 1.17-1.17.8. https://doi.org/10.1002/0471142727.mb0117s79

Vargas-Blanco, D.A., Shell, S.S., 2020. Regulation of mRNA Stability During Bacterial Stress Responses. Front. Microbiol. 11. https://doi.org/10.3389/fmicb.2020.02111

Vvedenskaya, I.O., Vahedian-Movahed, H., Bird, J.G., Knoblauch, J.G., Goldman, S.R., Zhang, Y., Ebright, R.H., Nickels, B.E., 2014. Interactions between RNA polymerase and the "core recognition element" counteract pausing. Science 344, 1285–1289. https://doi.org/10.1126/science.1253458

Waters, S.A., McAteer, S.P., Kudla, G., Pang, I., Deshpande, N.P., Amos, T.G., Leong, K.W., Wilkins, M.R., Strugnell, R., Gally, D.L., Tollervey, D., Tree, J.J., 2017. Small RNA interactome of pathogenic E. coli revealed through crosslinking of RNase E. EMBO J. 36, 374–387. https://doi.org/10.15252/embj.201694639

Yan, B., Boitano, M., Clark, T.A., Ettwiller, L., 2018. SMRT-Cappable-seq reveals complex operon variants in bacteria. Nat. Commun. 9, 3676. https://doi.org/10.1038/s41467-018-05997-6

Yang, W., 2011. Nucleases: diversity of structure, function and mechanism. Q. Rev. Biophys. 44, 1–93. https://doi.org/10.1017/S0033583510000181

Yu, J., Xiao, J., Ren, X., Lao, K., Xie, X.S., 2006. Probing Gene Expression in Live Cells, One Protein Molecule at a Time. Science.

Figures & legends

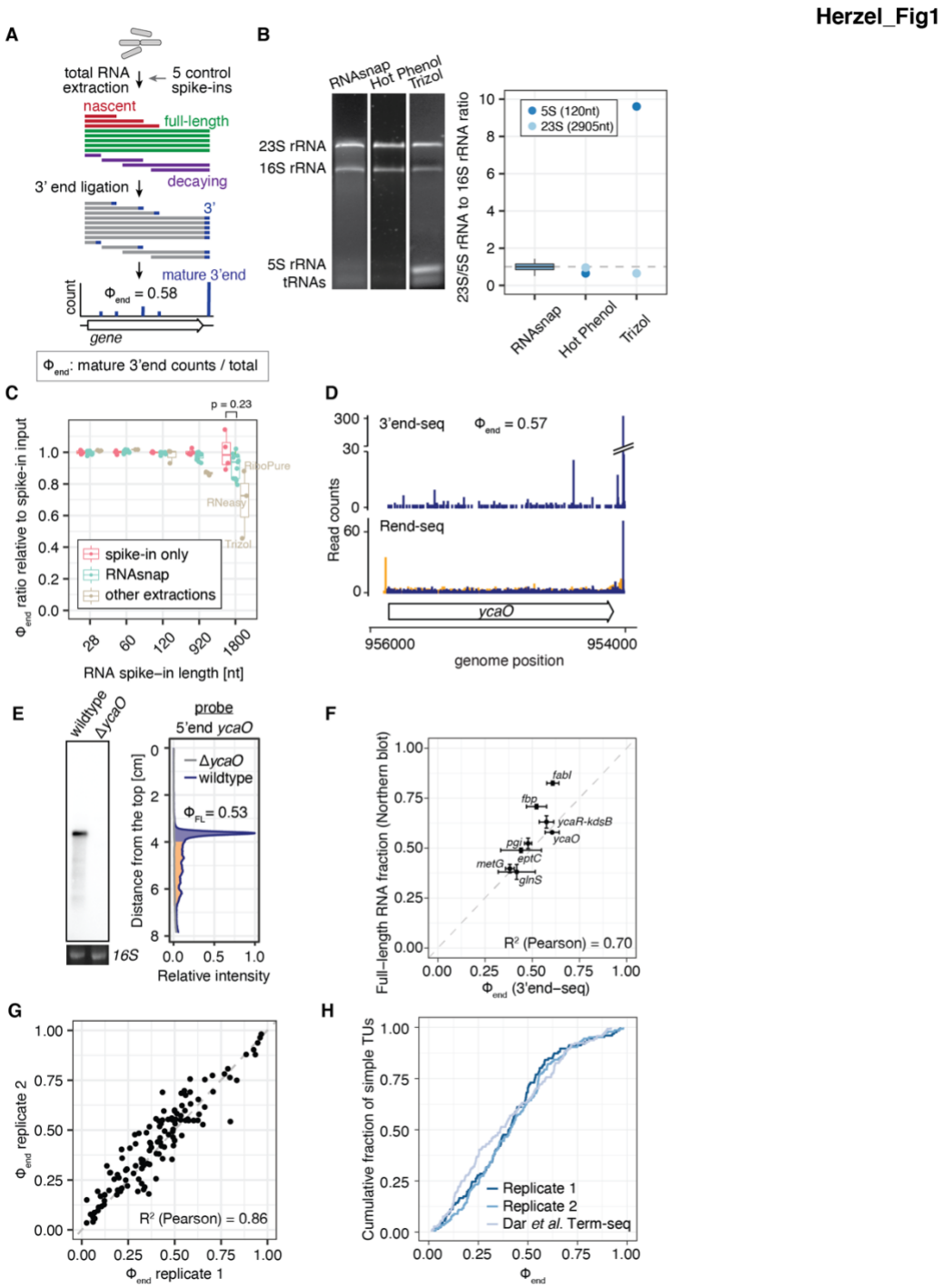

Figure 1: *In vivo* RNA 3′ end profiles reveal that the majority of transcripts are not full-length. (A) Schematic of 3′ end sequencing and hypothetical data derived from the suite of nascent, decaying and full-length transcripts. The measure Φ_{end} is an upper bound for the fraction of full-length mRNA. (B) Representative total RNA profiles from 3 different RNA extraction methods and a comparison of RNA extraction procedures in their ability to capture *in*

vivo 1:1 rRNA stoichiometry. The mean ratio for RNAsnap samples is 1.02 ± 0.22. The dashed line indicates the ratio of 1. (C) RNA spike-in quality is largely retained upon RNAsnap total RNA extraction. Boxplot showing the ratio of spike-in Φ_{end} after addition to the sample and without sample contact for each of the 5 differently sized spike-ins. (D) 3′ end- and Rend-seq data for the simple transcription unit (TU) encoding *ycaO*. (E) Northern blot and relative intensity profile for the same TU as in D. (F) Proportions of partial transcripts for 8 simple TUs, ranging from 0.9-2.1kb, are consistent when measured by 3′-seq and Northern blot. Error bars reflect standard deviation of 3 and 2 replicates from Northern blotting and 3′ end-seq, respectively. Full northern blots are given in Fig. S1A. (G) High reproducibility of Φ_{end} by 3′ end sequencing of RNAsnap extracted total RNA. Scatter plot correlating Φ_{end} from simple TUs between two biological replicates. (H) Cumulative distributions of Φ_{end} for two biological replicates and published data from Dar *et al*. 2018 with medians around 0.4.

Herzel_FigS1

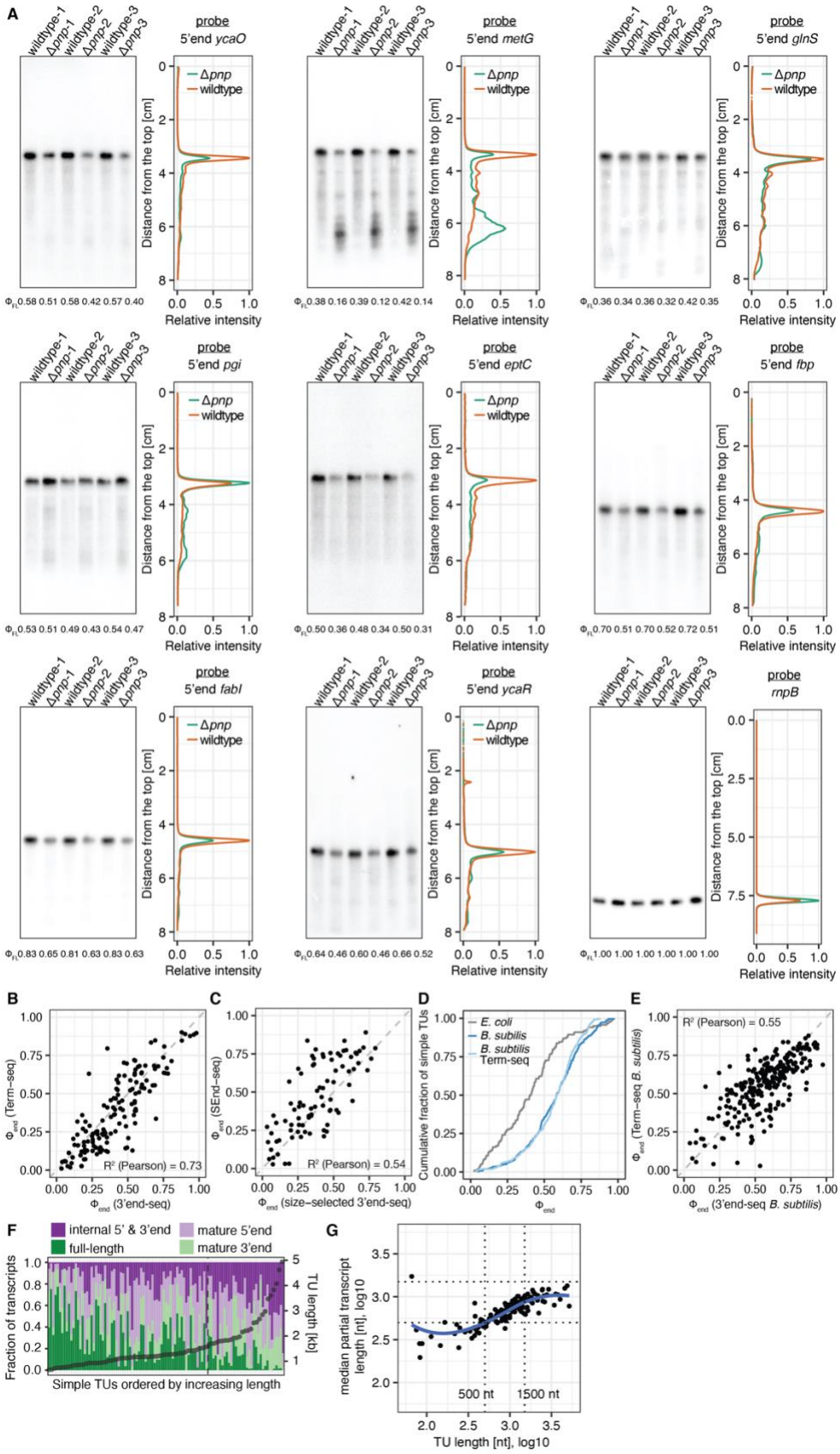

Figure S1: Measuring transcriptome integrity and heterogeneity by RNA-seq and Northern blot. (A) Northern blots and intensity profiles used for quantification of the full-length fraction in wildtype in Fig. 1F. 5 µg of total RNA from 3 biological replicates per strain are analyzed. Blotting against the stable RNA *rnpB* serves as loading control. (B-C) High reproducibility of Φ_{end} between different studies. Scatter plot correlating Φ_{end} from simple TUs from this study, Term-seq (Dar and Sorek, 2018) and Send-seq (Ju et al., 2019b). (D) Cumulative distributions of Φ_{end} for simple TUs in *E. coli* and *B. subtilis*. (E) Scatter plot correlating Φ_{end} from *B. subtilis* simple TUs from this study and Term-seq (Dar et al., 2016). (F) Partial and full-length transcript diversity for individual simple TUs obtained by SEnd-seq. The TU length distribution ranging 0.6-5kb is overlayed. The dashed line separates TUs at 1.5kb, above 1.5kb the transcript diversity might be strongly affected by length constraints introduced during sample preparation as shown in G. (G) Partial transcript length increases non-linearly with TU size. Below 500nt most RNAs are depleted during sample preparation. A similar effect could explain depletion above 1500nt.

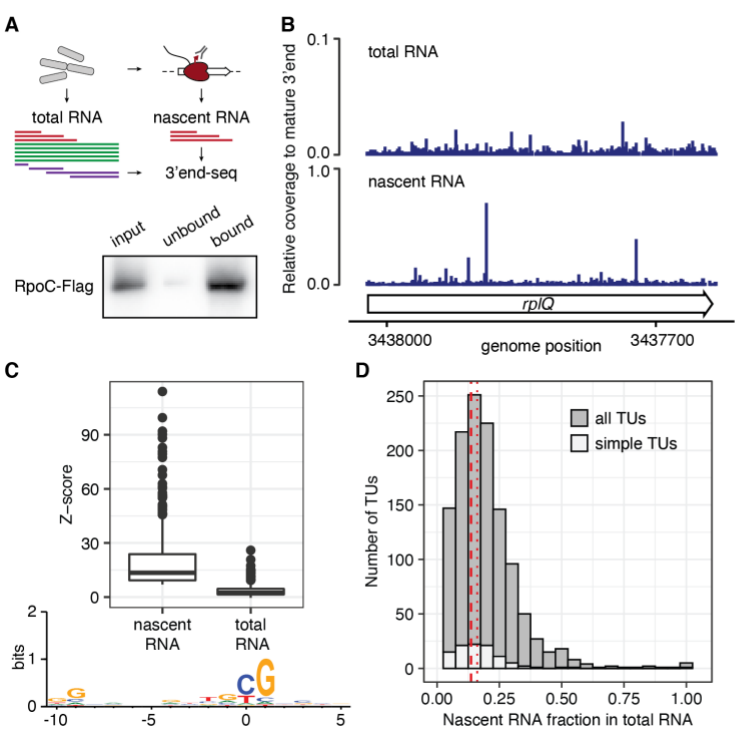

Figure 2: Few cellular transcripts are RNA polymerase associated. (A) Schematic of 3′ end sequencing of nascent and total RNA. Nascent RNA enrichment prior library preparation by immunoprecipitation of Flag-tagged RNA polymerase transcription elongation complexes. Lower panel shows efficient isolation of RNA polymerases by western blot analysis with the Flag-antibody detecting RpoC-FLAG prominently in the total lysate (input) and eluate (bound), but less in the unbound fraction after immunoprecipitation. Full western in Fig. S2. (B) Nascent and total RNA differ in their 3′ end profiles. 3′ end read counts relative to 3′ peak intensity at the end of the transcription unit containing *rplQ*. (C) Elemental pauses are strongly detected in nascent RNA, but not total RNA. Nascent peak Z-scores compared to total RNA Z-scores at the

same sites. Sequence logo for regions around nascent RNA peaks in CDSs. (D) The transcriptome of a typical TU contains ~15% nascent RNA. Histogram representing the signal in RNA polymerase pauses in total RNA relative to nascent RNA for simple TUs and all TUs. Red lines mark distributions medians (dashed – simple TU, dotted – all TUs).

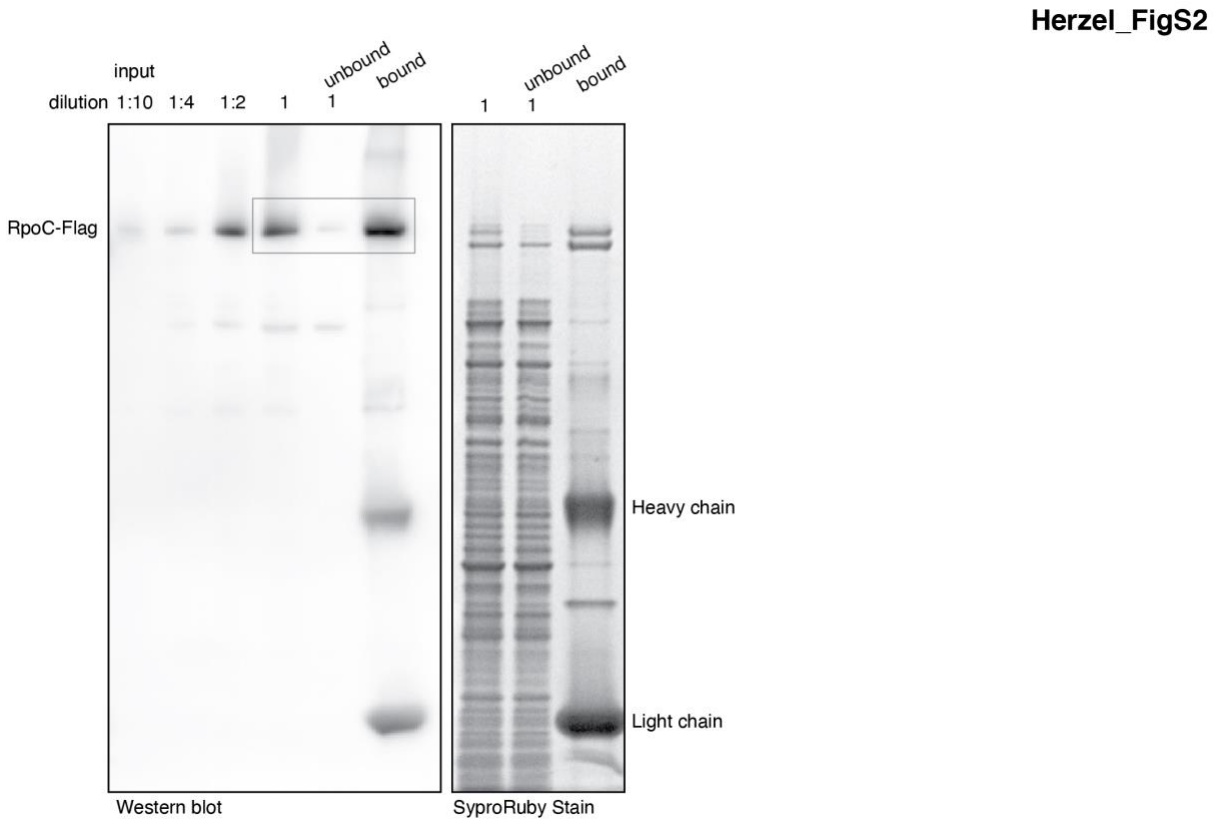

Figure S2: Probing for RNA polymerase enrichment. Full western blot probing for Flag-tagged RpoC and corresponding to Fig. 2A. Boxed area corresponds to cutout shown in Fig. 2A. A dilution series was loaded for the input sample to estimate the remaining RpoC after immunoprecipitation (10-25%).

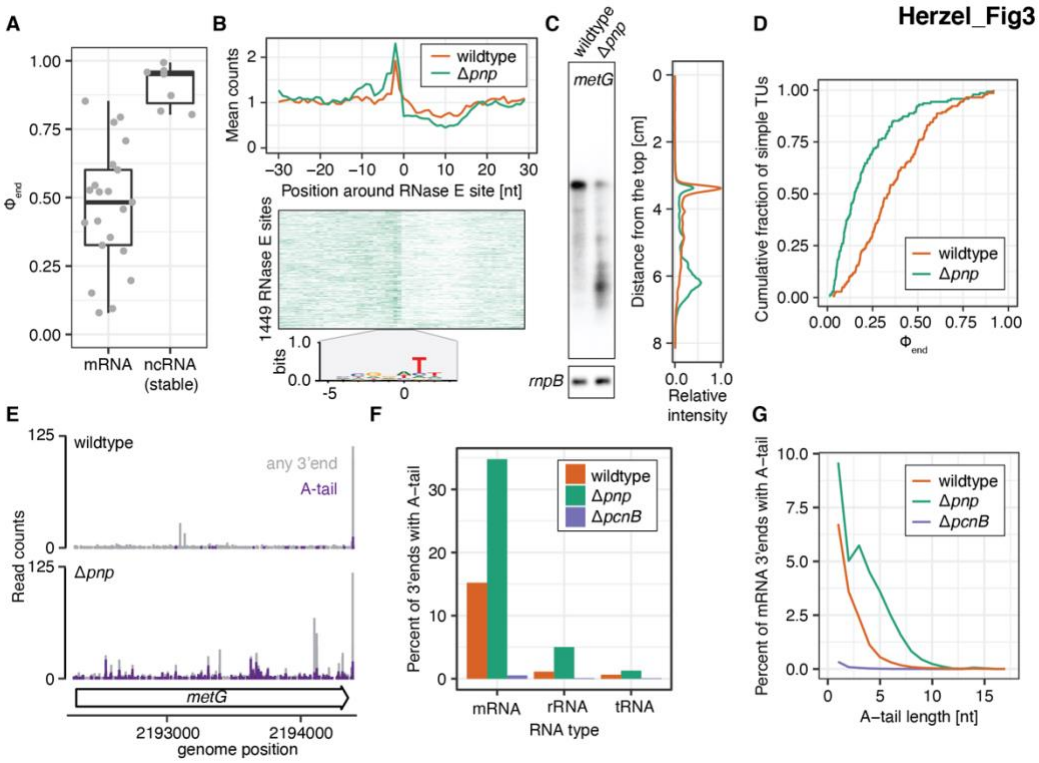

Figure 3: *In vivo* RNA decay signatures corroborate high amounts of decaying transcripts. (A) Stable RNAs have more full-length RNAs. Boxplot of Φ_{end} for stable RNAs (*ffs*, *gcvB*, *rnpB*, 5S, 16S, 23S rRNAs, *ssrA*) and mRNAs of the same length range as stable RNAs. (B) Some 3′ ends arise from RNase E cleavage. Distribution of 3′ ends relative to 1449 RNase E sites with 0.5 reads / nt in a 60nt window. Average signal shown in line plot for wildtype and Δ*pnp* 3′ end-seq. Heatmap for Δ*pnp* 3′ end-seq, sorted by RNase E site Z-score (Clarke et al., 2014), wildtype heatmap in Fig. S3. The sequence logo at these RNase E sites matches the *S. enterica* motif (Chao et al., 2017). (C) Stabilization of partial transcripts and reduction of full-length mRNA by northern blot analysis of simple TU *metG* in wildtype and Δ*pnp*. Probe against stable ncRNA *rnpB* as loading control. Further simple TU analysis by northern blot in Fig. S1A. (D) Global decrease of full-length RNAs in the exonuclease deletion (Δ*pnp*) compared to wildtype. Cumulative distributions of Φ_{end} for Δ*pnp* and wildtype. (E-G) Prevalent detection of short PolyA polymerase (PcnB)-dependent polyA tails. (E) 3′ end profiles across the simple TU *metG* in wildtype and Δ*pnp*. (F) Percent of 3′ ends with A-tails for 3 types of RNAs. Adenylation is abolished upon deletion of the PolyA polymerase (*ΔpcnB*) and increases upon deletion of the exonuclease PNPase (*Δpnp*). (G) Percent of mRNA 3′ ends with A-tails of different lengths. The average A-tail length is very short and increases upon deletion of the exonuclease PNPase (*Δpnp*).

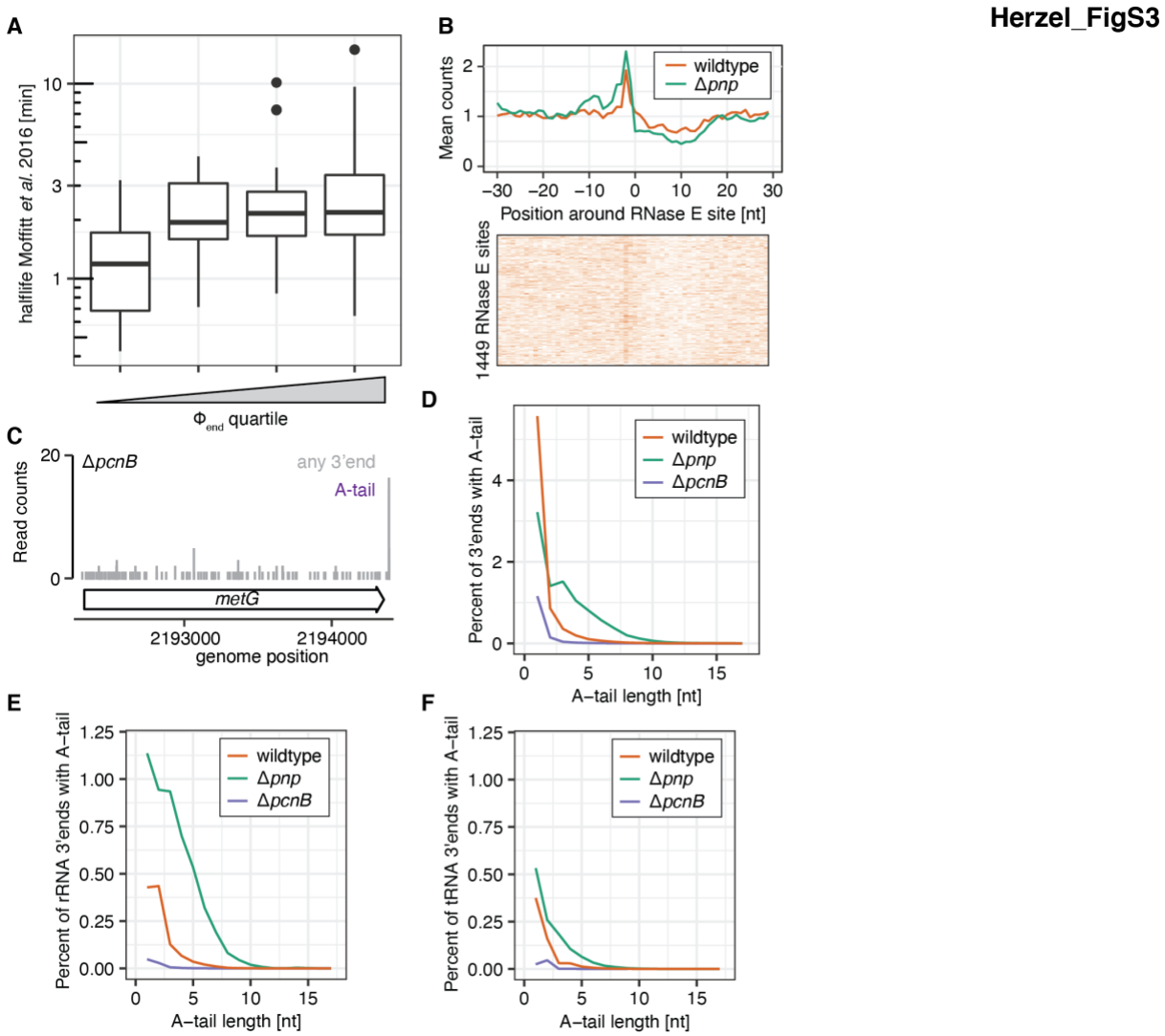

Figure S3: RNA decay signatures in mRNA, tRNA and rRNA. (A) mRNAs with low Φ_{end} are particularly short-lived. Boxplot showing quartiles of Φ_{end} from simple TUs and their mRNA halflife distribution (Moffitt et al., 2016). (B) Some 3′ ends arise from RNase E cleavage. Distribution of 3′ ends relative to 1449 RNase E sites with 0.5 reads / nt in 60nt window (same as in Fig. 3B). Average signal shown in line plot for wildtype and Δ*pnp* 3′ end-seq. Heatmap for wildtype 3′ end-seq, sorted by RNase E site Z-score (Clarke et al., 2014). (C) Absence of A-tails in the strain Δ*pcnB* missing PolyA polymerase. 3′ end profiles across the simple TU *metG*, compare to Fig. 3E. (D-F) Percent of 3′ ends with A-tails of different lengths for (D) all mapped and unmapped reads, (E) in rRNA and (F) in tRNA for wildtype, *Δpnp* and *ΔpcnB*. As shown for mRNA in Fig. 3G, the average A-tail length is very short and increases upon deletion of the exonuclease PNPase (*Δpnp*).

Herzel_Fig4

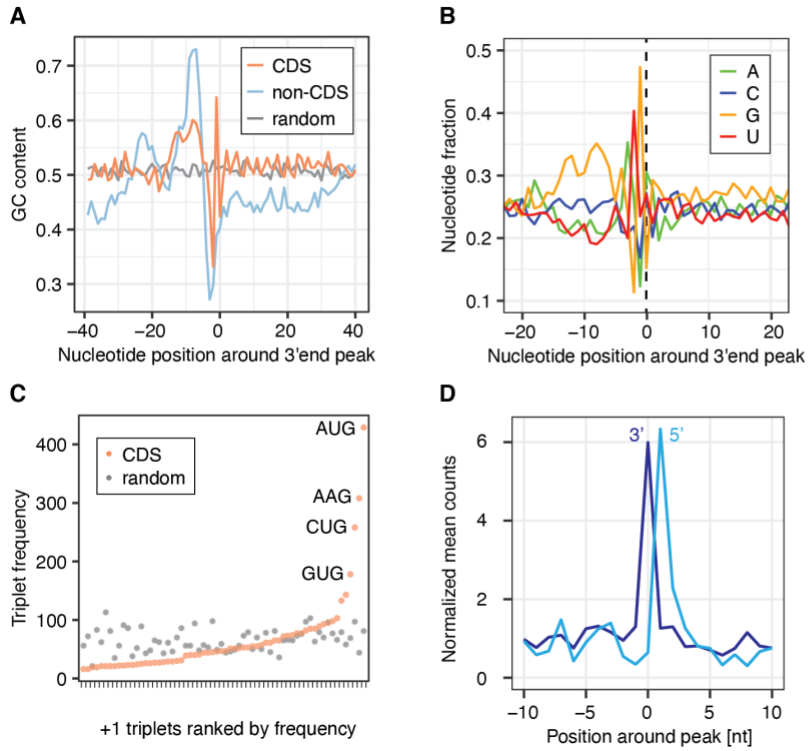

Figure 4: Prevalent AUG+1 peaks suggest interconnection of improper translation initiation and RNA decay. (A) GC-content near 3′ end peaks within and outside of CDSs. High GC-content upstream and low GC-content downstream of non-CDS peaks is consistent with secondary structure and polyU tracts of intrinsic terminators. CDS peaks differ in their profile with a broad peak a few nucleotides upstream and a GC-spike one nucleotide upstream the 3′ end peak. GC-content around random sites in the genome shows no GC-profile variation. (B) Nucleotide distribution around CDS peaks shows prevalence for Gs a few nucleotides upstream and AUG one nucleotide upstream of the 3′ end peak. (C) Frequency of triplets one nucleotide upstream of the CDS 3′ end peak shows enrichment of translation initiation codons relative to random sites in the genome. (D) Matched 5′ and 3′ end analysis yields evidence for endonucleolytic cleavage generating AUG+1 peaks.

Herzel_FigS4

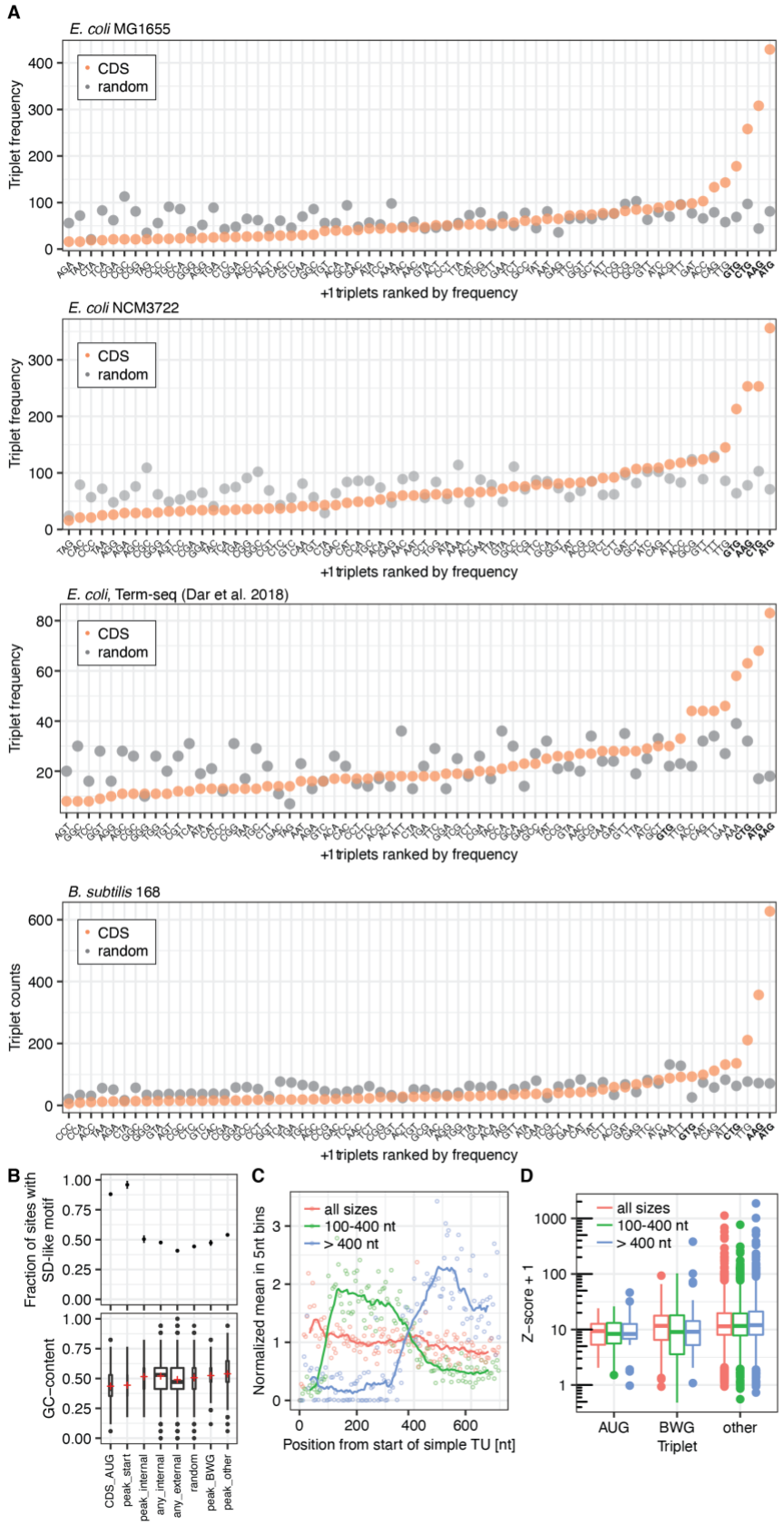

Figure S4: AUG+1 peak prevalence and characteristics. (A) Frequency of triplets one nucleotide upstream of the CDS 3′ end peaks shows enrichment of translation initiation codons relative to random sites in the genome for different *E. coli* wildtype strains, Term-seq data from (Dar and Sorek, 2018) and for *B. subtilis*. (B) SD-like motif and GC-content analysis. SD-like motifs are prevalent upstream of canonical start codons and start codons that have an AUG+1 peak (peak_start), but not particularly enriched at internal AUG+1 peaks. Standard deviation is estimated by bootstrapping. The overall higher GC content across CDSs could explain the occasional presence of SD-like sequences (Saito et al., 2020b). Red plus indicates the distribution mean. (C-D) 3′ end-seq of total RNA selected for different sizes shows that AUG+1 peaks originate from RNAs of diverse lengths. (C) Average 3′ end distribution relative to the start of simple TUs for samples with different RNA sizes. As expected for the 100-400nt RNA 3′ ends map primarily to the respective region at the beginning of simple TUs and 3′ ends of RNAs larger than 400nt start accumulating ~400nt downstream of TU starts. (D) Z-score distribution of sites that pass read cutoffs in all 3 samples and have a Z-score greater 7 in at least one sample. No difference in Z-scores is detectable among different size ranges, indicating that 3′ end peak detection is independent of RNA size. 3′ end peaks were classified as AUG+1 peaks, BWG peaks (GUG, CTG or AAG codon 1nt upstream of 3′ end peak) or any other peak.

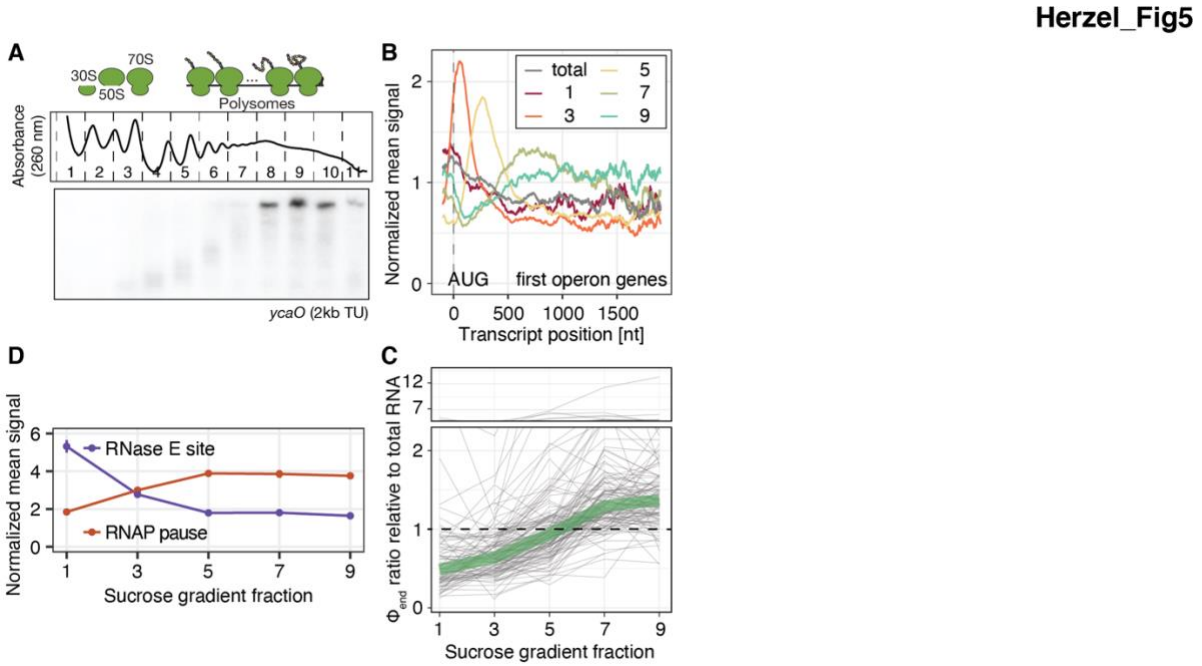

Figure 5: Enrichment of protein synthesis-competent RNAs in polysomes. (A) Polysome gradient centrifugation and Northern blot analysis of the simple TU *ycaO*. (B) 3′ end distance to CDS starts increases for higher polysome fractions. Average 3′ end signal for the same subset of expressed CDSs plotted with respect to distance to the start codon for individual fractions. (C) Global increase of Φ_{end} in denser gradient fractions. The median ratio of Φ_{end} in individual fractions for simple TUs relative to the input Φ_{end} covers a 3-fold range (green). (D) Average 3′ end signal in fractions across the polysome gradient around 396 RNA polymerase pause sites and 460 RNase E sites. Error bars reflect standard error of the mean.

Herzel_FigS5

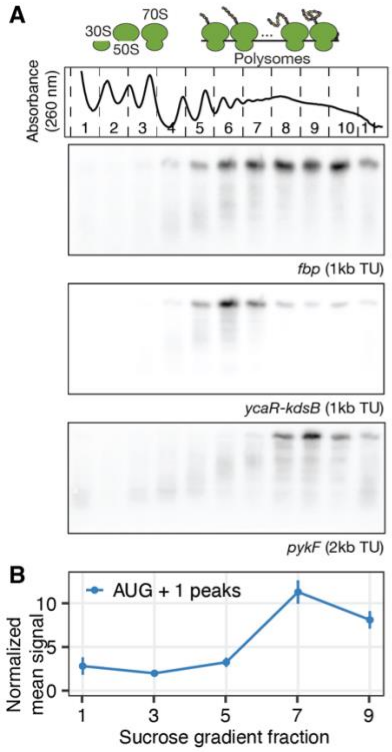

Figure S5: Polysomal northern blot analysis and AUG+1 peak distribution. (A) Full-length and partial transcript profile of the simple TU *fbp*, *ycaR-kdsB*, *pykF*, probing the 5′ region of the TU and for *pykF* the 3′ region. Polysome gradient same as in Fig. 5A and Fig. S6A.(B) Average 3′ end signal in fractions across the polysome gradient around 50 common AUG+1 peaks. Error bars reflect standard error of the mean.

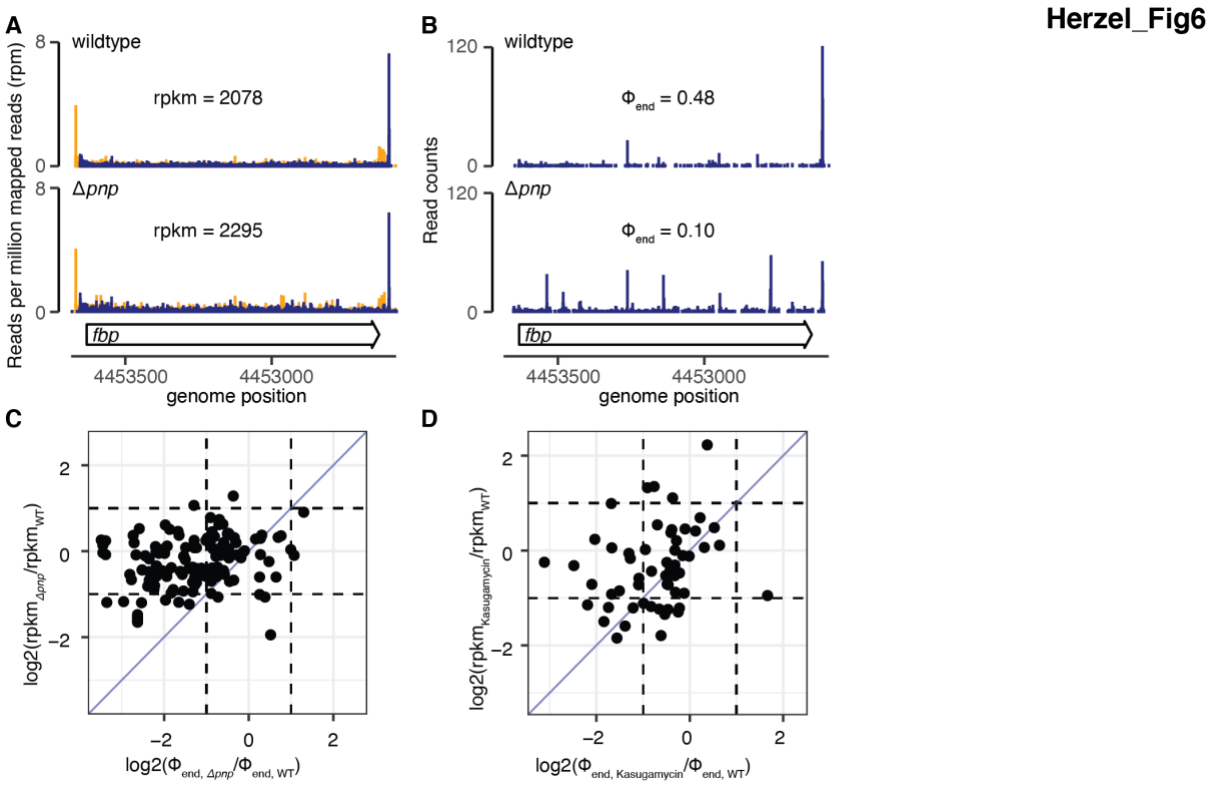

Figure 6: Transcriptome heterogeneity changes in the absence of average expression differences. (A) Rend-seq and (B) 3′ end-seq coverage and rpkm for the simple TU *fbp*. (C-D) Scatterplot of log2-fold changes in Φ_{end} and mRNA expression from RNA-seq for simple TUs between (C) Δ*pnp* and wildtype and (D) upon inhibition of translation initiation by Kasugamycin for 15min.

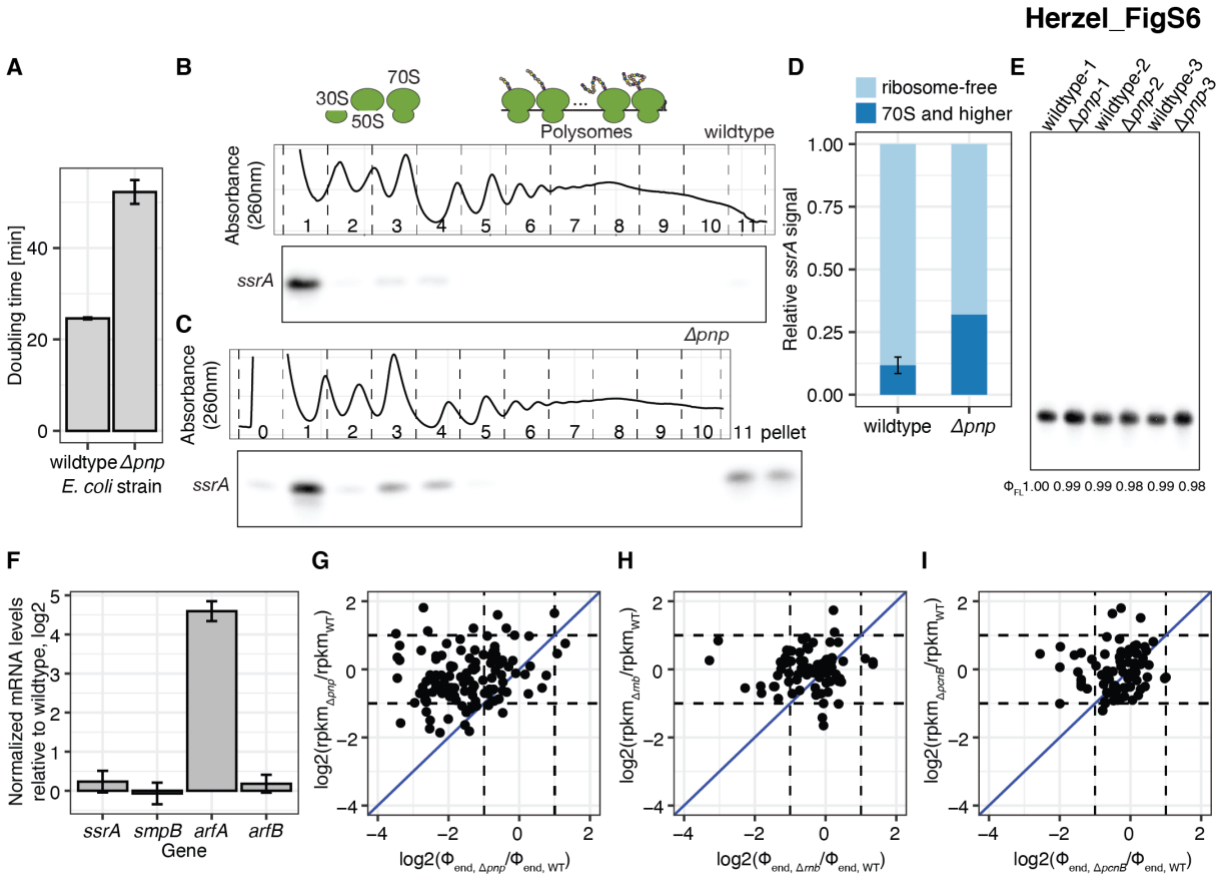

Figure S6: Higher engagement of *trans*-translation in *Δpnp* and transcriptome heterogeneity changes in deletions of RNA decay associated enzymes. (A) Doubling times in exponential phase in MOPS-Complete of three independent bacterial cultures. Error bars correspond to standard deviation. (B) Northern blot profile of the RNA-component of the *trans*-translation machinery *ssrA* across polysomes for wildtype. Linear intensity profile capped at maximum pixel intensity in fraction 1. (C) Similar to A, but for strain *Δpnp*. Linear intensity profile capped at maximum pixel intensity in fraction 2. (D) Quantification of *ssrA* signal in ribosome-free (fraction 1,2 and 0-2, respectively) and ribosome-associated (fraction >2) fractions from panel A and 2 more replicates and B shows 3-fold increased signal in ribosome-associated fractions in *Δpnp*. Error bar indicates standard deviation of 3 biological replicates in wildtype. (E) Total RNA northern blot probing for *ssrA* shows small, not significant increase in *ssrA* levels in *Δpnp* (F) mRNA expression of alternative ribosome rescue factor A *arfA* is upregulated in Δ*pnp* as measured by RT-qPCR for 3 biological replicates. Error bars indicate standard deviation. (G-I) Scatterplot of log2-fold changes in Φ_{end} and mRNA expression from RNA-seq for simple TUs between wildtype and (G) Δ*pnp* (RNA-seq replicate generated with an alternative library preparation compared to Fig. 6C), (H) Δ*rnb* or (I) Δ*pcnB*.