Molecular Systems Biology 8; Article number 585; doi:10.1038/msb.2012.16
Citation: *Molecular Systems Biology* 8:585

www.molecularsystemsbiology.com

molecular systems biology

REPORT

Transcription start site associated RNAs in bacteria

This article has been corrected since Online Publication. The name of the seventh author has been corrected.

Eva Yus[1,4], Marc Güell[1,4], Ana P Vivancos[2], Wei-Hua Chen[3], María Lluch-Senar[1], Javier Delgado[1], Anne-Claude Gavin[3], Peer Bork[3] and Luis Serrano[1,4,*]

[1] Center for Genomic Regulation (CRG), UPF, Barcelona, Spain, [2] Translational Research Program, Vall d'Hebron Institute of Oncology, Barcelona, Spain, [3] European Molecular Biology Laboratory, Heidelberg, Germany and [4] Institució Catalana de Recerca i estudis Avançats (ICREA), Barcelona, Spain
[4]These authors contributed equally to this work
* Corresponding author. EMBL-CRG Systems Biology Unit, Center for Genomic Regulation (CRG-UPF), c. Dr Aiguader 88, Barcelona 08003, Spain. Tel.: +34 933160247; Fax: +34 933160099; E-mail: luis.serrano@crg.es

Received 22.3.11; accepted 24.4.12

Here, we report the genome-wide identification of small RNAs associated with transcription start sites (TSSs), termed tssRNAs, in *Mycoplasma pneumoniae*. tssRNAs were also found to be present in a different bacterial phyla, *Escherichia coli*. Similar to the recently identified promoter-associated tiny RNAs (tiRNAs) in eukaryotes, tssRNAs are associated with active promoters. Evidence suggests that these tssRNAs are distinct from previously described abortive transcription RNAs. ssRNAs have an average size of 45 bases and map exactly to the beginning of cognate full-length transcripts and to cryptic TSSs. Expression of bacterial tssRNAs requires factors other than the standard RNA polymerase holoenzyme. We have found that the RNA polymerase is halted at tssRNA positions *in vivo*, which may indicate that a pausing mechanism exists to prevent transcription in the absence of genes. These results suggest that small RNAs associated with TSSs could be a universal feature of bacterial transcription.
Molecular Systems Biology **8**: 585; published online 22 May 2012; doi:10.1038/msb.2012.16
Subject Categories: RNA; microbiology & pathogens
Keywords: non-coding RNAs; small RNAs; transcription; transcriptomics

Introduction

Systematic transcriptomic analyses have unveiled a variety of non-coding RNAs ranging in size from a few to several thousand bases (Miura *et al*, 2006; Chekanova *et al*, 2007; Shi *et al*, 2009). Recently, a new class of small (13–26 bases), transcription initiation-associated RNAs, termed tiny RNAs (tiRNAs), has been found in fruit fly, human, and chicken (Taft *et al*, 2009; Cserzo *et al*, 2010). tiRNAs are produced at sites associated with stalled RNA polymerase as well as at transcription start sites (TSSs) and splice sites (Taft *et al*, 2009, 2011). tiRNAs can be differentially regulated during development and in a tissue-specific manner (Taft *et al*, 2011) and may play a role in epigenetic regulation (Taft *et al*, 2009). tiRNAs differ from the so-called abortive transcripts (of 6–17 bases) that are associated with *in vitro* and *in vivo* initiation of transcription, which result from abortive cycles of the RNA polymerase before the promoter is cleared (Hsu, 2002; Kapanidis *et al*, 2006; Goldman *et al*, 2009).

We recently reported that a number of transcriptional features are similar between bacteria and eukaryotes, including the existence of a significant number of non-coding RNAs (Guell *et al*, 2009). We therefore decided to look for possible existence of small RNAs associated with TSSs. By combining various experimental approaches, we have now identified a new class of bacterial small RNAs that map to TSSs, somewhat similar to the tiRNAs detected in eukaryotes. For efficiency, we examined one of the smallest Firmicute bacteria, *Mycoplasma pneumoniae*, which is a model organism for genome-reduced bacteria (Guell *et al*, 2009; Kuhner *et al*, 2009). We then validated the generality of our findings with the most widely used bacterial model organism, the gram-negative, γ-proteobacterium *Escherichia coli*.

Results and discussion

To systematically investigate small RNAs possibly associated with TSSs in *M. pneumoniae* and *E. coli*, we specifically isolated non-fragmented RNAs ranging in size from 15 to 65 bases and subjected them to direct strand-specific sequencing (DSSS) (Figure 1A) (Vivancos *et al*, 2010).

The sequencing reads of the small (15–65 bases) RNAs from both *M. pneumoniae* (Figures 1A and 2A; Supplementary Figure S2) and *E. coli* (Supplementary Figure S1) showed a non-uniform distribution along the chromosome. The reads often appeared as well-defined, single peaks (Figure 1B) with

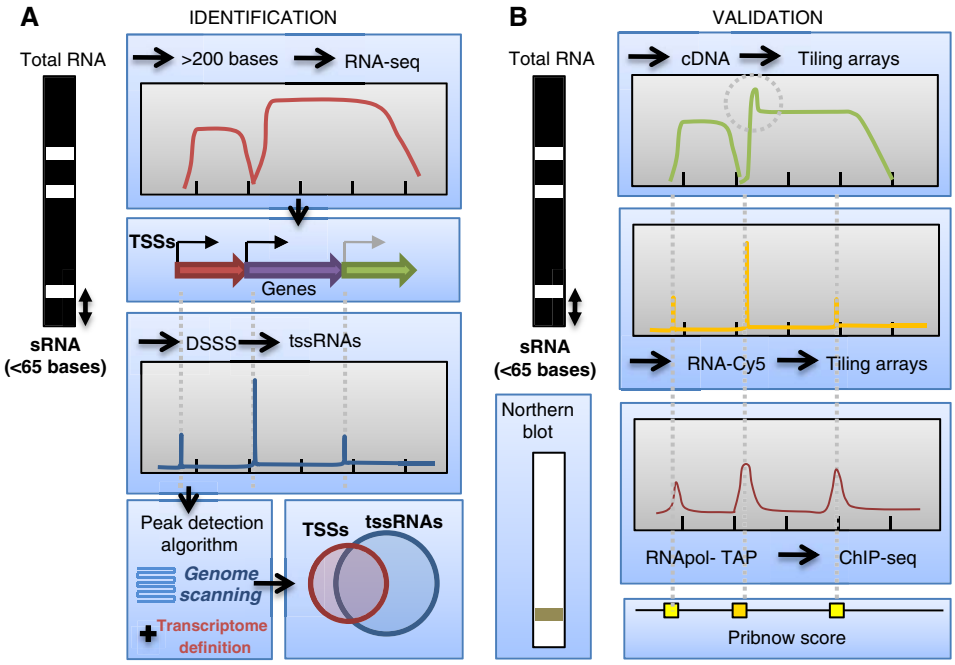

Figure 1 Bacterial tssRNA identification and validation workflow. (**A**) Methodology used to detect tssRNAs. A fraction of small (<65 bases) RNA was submitted to DSSS. The genome was scanned to detect novel small RNAs. Total RNA was used to define the transcriptome and putative TSSs. tssRNAs were shown to map preferentially to the promoter regions. (**B**) Various methodologies, both experimental (e.g., detection of the 5′ peaks on tiling arrays of total cDNA, Cy5-labeling of RNA of <65 bases, and binding of RNA polymerase subunits, among others) and computational (by correlating the levels of the tssRNAs with the Pribnow score used to identify the presence of promoters), were used to identify tssRNAs. Determining tssRNAs allowed TSSs to be identified that would otherwise escape identification, such as those located in overlapping genes or in suboperons (exemplified by the third TSS on the A scheme).

an abrupt raise and a flat plateau (Supplementary Figure S2A). We also observed more complex patterns, such as two or more overlapping distributions, indicative of multiple promoters (Supplementary Figure S2B).

In order to quantify these small RNAs automatically and consistently, we developed a computational method that identifies narrow peaks with a flat plateau significantly above the background (Figure 1A; Table I; Supplementary information). Automatic analysis of *M. pneumoniae* small transcripts obtained from stationary phase cells (Yus *et al*, 2009) allowed 1339 ± 116 small RNAs to be identified (using data from three independent biological replicates; Supplementary Tables S3 and S4), of which 457 ± 28 (~34%) were located <10 bases away from a manually annotated TSS (annotated in this study based on the published data in Guell *et al*, 2009) (Table I; Supplementary Table S1). In total, ~73% of the *M. pneumoniae* TSSs have an automatically assigned small RNA (Table I). Visual inspection of the missing TSSs reveals that all of them have an associated RNA that was not identified by the algorithm due to a complicated shape or low height (Supplementary Figure S4B). In this way, we identified 1371 small RNAs on the plus strand from the *E. coli* data set. Using the Regulon database (Gama-Castro *et al*, 2008) to extract a high-confidence group of sigma 70-dependent TSSs on the plus strand (Table I; Supplementary Table S2), we reproducibly found that a somewhat smaller (but still large) proportion of active TSSs have associated small RNAs in the stationary phase (44 ± 5%; Table I). In both species, the number of small RNAs decreased in the exponential phase (from 1239 to 818 for *M. pneumoniae*, and from 1371 to 684 for *E. coli*; Table I). The distance from the small RNA start position to the experimentally determined TSS of the cognate full-length transcript overlapped significantly ($P = 0.00015$; Supplementary Figure S4B), with differences between the small RNA starting positions and the annotated TSSs of −0.5 ± 10 bases in *M. pneumoniae*, and of −3.5 ± 12 bases in *E. coli*. At each TSS, we observed a dominant species of small RNAs, with some minor ones that started at the same point but had slightly different lengths (Supplementary Figure S5). These results suggest that these newly identified small RNAs are transcribed from the promoters of the corresponding cognate full-length transcripts. Hence, we named these 'transcription start site-associated' RNAs (tssRNAs). We propose that tssRNAs could be used as markers for promoters in uncharacterized genomes. They could also help to identify ambiguous TSSs, for example in regions where transcripts overlap or when no clear boundaries (e.g., start and end) can be observed at the RNA level (see scheme in Figure 1A, and example in Figure 2A).

We next applied a number of independent approaches to confirm the existence of tssRNAs and to rule out possible technical artifacts. We found that (i) tssRNAs were also observed when using RNA-seq protocols that do not require fractionation of RNAs by size (using TrueSeq, from Illumina) (Figures 1B and 2B); (ii) tssRNAs were unequivocally detected on tiling arrays hybridized with total cDNA (see Supplementary information; Figures 1B and 2D) and, similarly, deep sequencing of non-size-fractionated RNAs showed a clear peak at the TSSs from low expression genes (allowing co-detection of the cognate tssRNAs) (Supplementary Figure

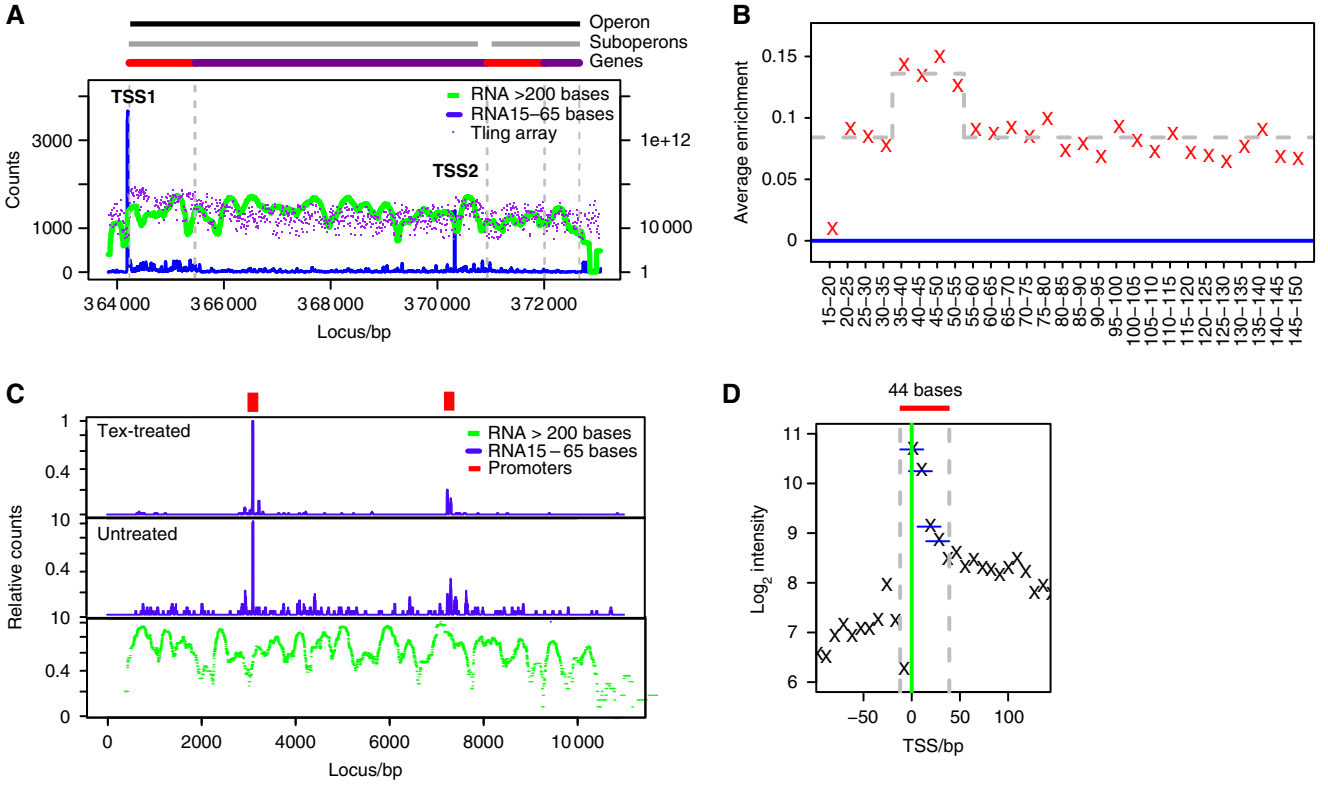

Figure 2 Properties of tssRNAs. (**A**) Example of tssRNA identification by DSSS of RNAs with fewer than 65 bases in a *M. pneumoniae* cytoadherence operon (*mpn309-mpn312*). ORFs are shown in alterative colors. A second, internal TSS (TSS2) was easily detected due to the presence of a tssRNA, and revealed the presence of a suboperon or alternative transcript. (**B**) Size distribution of tssRNAs by a RNA-seq method that does not involve RNA fragmentation (TrueSeq) showed an average size of 45 bases. Pile-ups were built by selecting reads containing the indicated sizes, and normalized by the mean number of counts. Untreated samples were substracted from TAP-treated samples to remove the background value (see Supplementary Methods). Average normalized counts (shown in red) were estimated for positions where tssRNAs start. To determine the null distribution for random selected points in the genome (shown in gray), 500 groups with equal sizes as the tssRNAs group were selected. (**C**) Upper panel: samples of total RNA were treated with terminator 5′-phosphate-dependent exonuclease (Tex) and analyzed as in (A). Second panel: untreated control. Treated samples contained a higher ratio of tssRNAs with respect to degradation products. This demonstrates that tssRNAs are capped with 5′-triphosphate and are thus nascent transcripts. Lower panel: mRNA levels of the genomic region as measured by DSSS. (**D**) Plot showing the distribution of log2 values for tiling array data around the TSS start site of 49 genes that showed a clear tssRNA over the mRNA level. There was a peak at the TSS proximity that spans about three probes in the tiling array (with an average length of 28 bases separated by about 8 or 9 bases) is clearly evident; this is equivalent to around a 44-base length, similar to that found by RNA-seq (see Materials and methods).

Table I Presence of tssRNAs at TSSs

Organism	TSS	TSS tssRNAs		Total tssRNAs	
		Stationary	Exponential	Stationary	Exponential
M. pneumoniae (+ strand)	309	210–236*	168	586–652*	418
M. pneumoniae (− strand)	317	219–249*	183	535–705*	400
E. coli (+ strand; active)[b]	220	97	70	1371	684

The number of TSSs identified in *M. pneumoniae*, and a subset of high confidence sites in the plus strand of *E. coli* as defined in the Regulon database[a] (Supplementary Table S2), are listed. [b]The number of *E. coli* promoters found to be active in the ultrasequencing analysis under the conditions tested is indicated. TSS tssRNAs indicates the number of TSSs for which an associated tssRNA was automatically detected. Total tssRNAs indicates the total number of tssRNAs that were automatically detected. (*) Values represent the range of two independent experiments.

S6B); (iii) direct hybridization of fluorescently labeled small RNAs (<65 bases) onto tiling arrays further substantiated the presence of tssRNAs (Figure 1B; Supplementary Figure S6A); and (iv) the existence of tssRNAs was directly confirmed by Northern blot analyses (Figure 3B; Supplementary Figure S12, see below).

Small RNAs associated with TSSs in eukaryotes (tiRNAs) are likely to be the result of endonucleolytic cleavage of the nascent RNA (Taft *et al*, 2009). To see if this is also the case in bacteria, we exposed the purified small RNAs to terminator 5′-phosphate-dependent exonuclease prior to cDNA synthesis (Figure 2C; Supplementary Figure S3). Since the 5′ ends of bacteria transcripts bear a triphosphate, while RNA degradation products have a single phosphate, this treatment should remove all RNA degradation products (Sharma *et al*, 2010). tssRNAs remained largely unaffected by this treatment, while

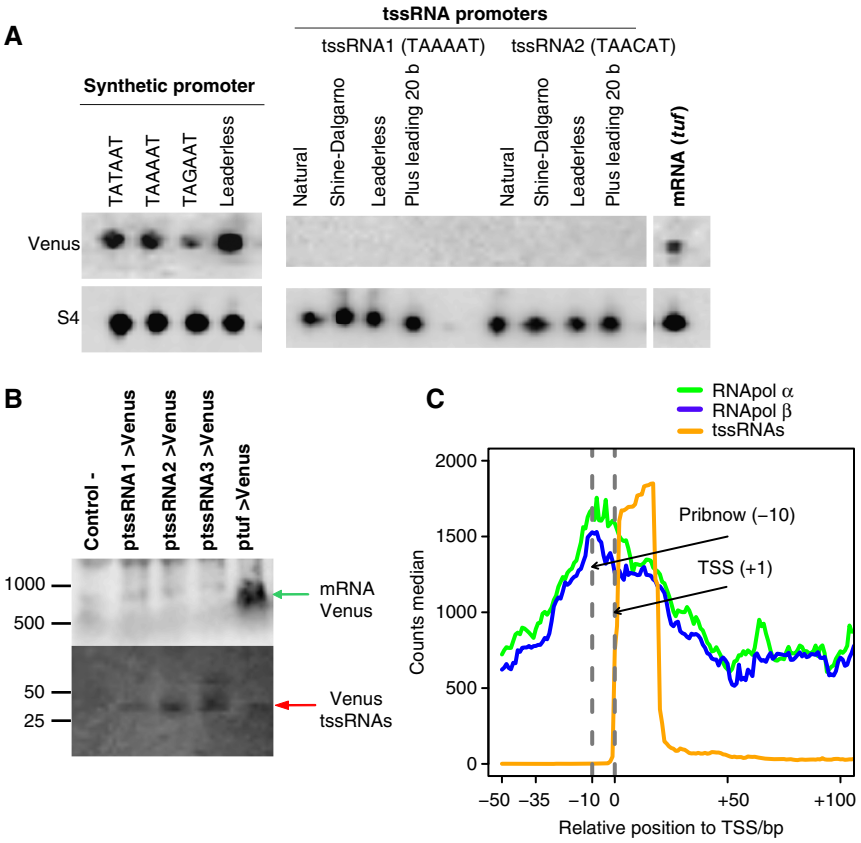

Figure 3 tssRNA promoter analysis and function. (**A**) Western blot against YFP-Venus under the control of synthetically designed promoters with good Pribnow boxes (see Supplementary information), or the promoter of two tssRNAs with no associated full-length transcript. The tssRNA promoters were either not modified (natural) or contain a modification to improve expression of a 'Shine-Dalgarno' (ribosome-binding site: GGAGGA), 'Leading 20 bases', the first 20 bases after the TSS, or 'Leaderless', without a 5′ UTR (see Supplementary Table S5). (**B**) Northern blot analysis of cell lines expressing three tssRNA promoters driving a YFP reporter (see also Supplementary Figure S12). On the upper panel, full-length YFP is detected in total RNA extracts, only when driven by an endogenous gene promoter. In the lower panel, RNA of <50 kDa were analyzed. (**C**) Average binding of the RNA polymerase subunits around the TSS, as detected by ChIP-seq. The closest Pribnow box (-10) and the TSS ($+1$) are marked with vertical lines. In addition to the well-known binding at the promoter region (which is responsible for generating abortive transcripts), an overlapping peak is observed inside the gene.

the level of other (background) small RNAs was reduced (Figure 2C; Supplementary Figure S3), consistent with the view that tssRNAs are primary transcripts and not the result of endonucleolytic cleavage. However, this experiment cannot distinguish from other possible mechanisms that produce such a 5′ end, like specific endonucleolytic cleavage near the 5′ end, or 3′-to-5′ RNase activity with some degree of protection of the first 40–50 bases. The fact that tssRNAs are observed to be isolated entities, without a corresponding long transcript (Supplementary Figure S11), would exclude that they are generated by degradation, since the RNAs we observed could not arise from a longer RNA. Even if secondary structure of the 5′ untranslated region (UTR) could explain their resistance to degradation in some cases, it is very unlikely that this could apply to all the detected tssRNAs. In fact, we did not observe any particular enrichment in three-dimensional structures at the 5′ of *M. pneumoniae* genes (Supplementary Figure S14). In sum, this supports the idea that they are primary transcripts resulting from transcription and not the result of degradation.

To further determine the length of tssRNAs, we analyzed the following various experimental results from *M. pneumoniae*: (i) standard electrophoresis of total RNA gave an apparent size of 35–40 bases (Supplementary Figure S3); (ii) Northern blot of tssRNA promoters driving YFP (see below) were within similar range (Supplementary Figure S12); (iii) tiling array showed a size of around 44 bases (Figure 2D); (iv) deep sequencing of RNA fractionated in various size ranges (<65, 65–100, 100–150, 150–200, and >200 bases) showed a consistent enrichment of tssRNAs in the size range 15–65 bases (Supplementary Figure S6E); and (v) using the TrueSeq kit from Illumina, an improved DSSS method that does not involve any size selection step (see Supplementary information), we verified that the tssRNAs were in the size range of 35–55 bases (Figure 2B). All methods offer a congruent view of the tssRNAs ranging in length from 35 to 55 bases. In *E. coli*, we observed a similar pattern, but with slightly smaller sizes for the tssRNAs (33–40 bases; Supplementary Figure S8). Thus, the bacterial tssRNAs are clearly larger in size than the described abortive transcripts found *in vitro* and *in vivo* (which range from 6 to 17 bases; see Goldman *et al*, 2009). Moreover, *in-vitro* transcription (IVT) performed with *M. pneumoniae* genomic DNA resulted in a pattern similar to that observed *in vivo* (Supplementary Figure S7B) but did not reveal the presence of any tssRNA

Table II Pribnow box analysis of tssRNA upstream regions

Category	Pribnow_score	Distance TSS bases
TSS	3.61 (1.0)	9.5 ± 2.9
Internal_ATG	4.70 (1.7)	10.3 ± 3.6
Intergenic	4.54 (1.1)	11.0 ± 3.7
Intragenic	5.16 (1.5)	11.2 ± 4.2
AverageGenome	9.70 (2.1)	

The scores for the best Pribnow box sequences within an interval of 7–25 bases 5′ from the TSSs of the tssRNAs are indicated. The numbers in brackets correspond to the standard deviation of the value. Pribnow_score is the sum of the Pribnow_sum plus 33% of the Pribnow_modifiers score. 'Distance TSS bases' is the distance from the TSS of a tssRNA to the third base of the best Pribnow box sequence within the interval described above. TSSs are transcription start positions for mRNA, rRNA, and tRNAs. Internal_ATG indicates that the tssRNAs were found within 50 bases from the ATG of a gene inside an operon. Intergenic indicates tssRNAs located between open reading frames, and intragenic peaks are inside open reading frames but >50 bases away from the initial ATG. AverageGenome reflects the average values over the whole chromosome.

(Supplementary Figure S7A). In sum, these results indicate that tssRNA are distinct from abortive transcripts, and that tssRNA synthesis requires the endogenous RNA polymerase machinery.

The majority of TSSs in *M. pneumoniae* have an associated tssRNA (Table I). However, we also found a large number of tssRNAs at other genomic positions that are not associated with the start of a full-length transcript. One explanation could be that these are synthesized from 'cryptic' promoters, that is, promoter-like sequences that appear randomly in the genome. We thus scored the quality of putative RNA polymerase recognition sites −10 regions (Pribnow boxes), which are the most conserved regions in *M. pneumoniae* promoters (Guell *et al*, 2009) along the chromosome ('Pribnow_score'; Supplementary Methods; Table II). The score was based on an analysis of the sequences upstream of the manually annotated TSSs (as determined by transcriptome analysis; Guell *et al*, 2009) (Supplementary Table S1), or by 5′ sequencing (Weiner *et al*, 2000) (see Supplementary Methods). Our analysis showed that all tssRNA upstream regions have a better Pribnow score than the average value of a random sequence (taking the whole *M. pneumoniae* genome into account), meaning that they have promoter-like features. Analyzing the 25 bases upstream of the tssRNA start sites for the best-scored Pribnow boxes showed that they are located at the right distance of around 10 bases upstream (according to the previously determined distance of 9 ± 3 bases; see (Shultzaberger *et al*, 2007) (Table II). Moreover, these regions have classic Pribnow sequences (of TANAAT, where N can be any base; Supplementary Figure S9), indicating they are true cryptic promoters. Consistent with this, we found RNA polymerase to be bound to them (see below).

Of the tssRNAs that did not map to the TSS of a long transcript, 34% are close to a translation start codon, about 21% are intragenic, and 48% are intergenic (in stationary phase; Table II). We analyzed the upstream promoter-like sequences to determine whether intragenic tssRNAs can be considered to be background and can thus be used to distinguish the true positives. However, all three sets had a good Pribnow score (although worse than that of the TSS-associated tssRNAs) at the right distance to the TSS (Table II; Supplementary Figure S9), and thus all could represent true TSSs. We additionally observed a positive correlation between the Pribnow score and the expression level of the tssRNA (Supplementary Figure S10). These results suggest that tssRNAs found at intergenic and intragenic regions reflect true TSSs from promoters that are likely to originate from random sequences, or from promoters that are activated under specific conditions. Considering the base composition of *M. pneumoniae*, we estimated a probability of having 1562 Pribnow boxes (TANAAT) in the genome, a figure that is around 30% larger than the actual one (of 1131 TANAAT sequences in the genome).

To demonstrate that tssRNAs not associated with long transcripts are the result of spurious transcription, we made three *M. pneumoniae* tssRNA promoter constructs that had a good Pribnow score and supported a high level of expression (Supplementary Figure S12A–C) but that did not produce a corresponding full-length transcript (Supplementary Table S5). These promoters were fused to the yellow fluorescent protein (YFP-Venus) gene. We did not detect any YFP expression from any of the constructs (as shown for two cases; Figure 3A), even when they were trimmed to a minimum length (i.e., they were 'leaderless', which improves the signal for a synthetic promoter) or when a ribosomal recognition sequence (Shine-Dalgarno) was added (Figure 3A). Adding the first 20 bases of the tssRNA did not influence the expression levels from either of the two tssRNAs promoters. We confirmed by Northern blot that these promoters did not yield full-length transcripts but rather only tssRNAs (Figure 3B; Supplementary Figure S12D). On the other hand, promoters that produce mRNAs and associated tssRNAs (Figure 3A), or even rRNAs, produced detectable Venus expression and long transcripts, as well as tssRNAs (Figure 3B; Supplementary Figure S12). These results suggest that although a good promoter will support RNA polymerase recruitment and tssRNA production, DNA features other than the Pribnow box are needed to produce full-length RNAs from productive transcription.

So far, we have confirmed the existence of native tssRNAs that are associated with full-length transcripts in both *E. coli* and *M. pneumoniae*. However, it is still unclear whether these are co-regulated by the same promoter sequences and thus expressed to the same extent, or whether the tssRNAs could be independently expressed. In *M. pneumoniae,* tssRNA expression levels correlate weakly with that of the corresponding mRNA ($R = 0.54$; see Supplementary Figure S13). However, when comparing the expression levels of tssRNAs with those of the cognate full-length transcripts in *M. pneumoniae* in both exponential and stationary phases, we observed an important relative increase of tssRNAs expression only in the stationary phase ($P = 3.52 \times 10^{-32}$, two-sample *t*-test), when transcription is known to be repressed (Table I; Supplementary Figure S15 and Table S6). This could indicate that tssRNA production is driven independently from its associated full-length RNA, and/or that it depends on other protein factors that determine transcription. To test for these possibilities, we first performed chromatin immunoprecipitation analyses of the two RNA polymerase subunits, α and β, in *M. pneumoniae* (MPN191|RpoA and MPN515|RpoB, respectively), followed by DNA ultrasequencing (ChIP-seq) and DNase protection

assays. These results revealed that the RNA polymerase indeed binds to both the productive (i.e., associated with long associated transcripts) and unproductive (isolated) tssRNAs (Supplementary Figure S16). The RNA polymerase was found to be located both at the promoter region (-10), a position at which it is known to stall and produce abortive transcripts prior to initiation of transcription elongation (Goldman *et al*, 2009), and at some nucleotides downstream of the TSS, where it could produce tssRNAs (around the $+30$ position; Figure 3C).

Positioning at downstream regions is more prominent in unproductive, isolated tssRNAs (Supplementary Figure S17), despite the fact that the overall affinity of the RNA polymerase is generally lower at these promoters, which on average have slightly worse Pribnow scores (Table II). This would indicate that RNA polymerase pausing is more likely to occur in non-productive promoters. Altogether, these results suggest that, once elongation has started, RNA polymerase pausing could be a mechanism to avoid spurious transcription at any place where a Pribnow box sequence is present. tssRNAs could thus represent a transcriptional checkpoint to ensure that the RNA polymerase machinery is correctly assembled (e.g., that the sigma factor is lost and the correct elongation factors are recruited) (Roberts *et al*, 2008; Yang and Lewis, 2010; Burmann and Rosch, 2011). This would further guarantee that there is no unnecessary transcription, avoiding the energy expense and preventing unwanted products, such as truncated proteins or transcripts that are antisense to essential genes (Supplementary Figure S18).

Conclusions

We identified and validated a putative new and distinct class of bacterial RNAs that are associated with TSSs, which we have termed tssRNAs. These have an average size of ~45 bases and exhibit dynamic behavior not necessarily concomitant with that of the cognate gene. The absence of tssRNA synthesis *in vitro* indicates that their expression requires additional native factors that would ensure the accurate elongation/termination of transcription (Nudler and Gottesman, 2002). While the results of our experiments indicate that tssRNAs could be primary transcripts, based on their 5′ triphosphate, it is still possible that other mechanisms could produce the 5′ ends, such as specific endonucleolytic cleavage near the 5′ end or 3′-to-5′ RNase activity with some degree of protection of the first 40–50 bases. However, we consider this to be unlikely, since: (i) almost all mRNAs are associated with tssRNAs; (ii) isolated tssRNAs are not associated with longer RNAs; (iii) promoters of isolated tssRNAs fused to the Venus protein did not produce full-length RNAs, while mRNA promoters did; (iv) tssRNAs overlap with pausing sites, as shown by RNA polymerase ChIP-seq. We hypothesize that the incorrect assembly of processive RNA polymerase complexes could lead to premature termination of RNA transcripts, which in turn could result in deleteriously truncated proteins. tssRNAs could be part of a regulatory mechanism that prevents transcription from starting before the correct RNA polymerase complex is assembled. It is still unclear which additional factors are involved in this process, or which sequences determine that a promoter will ensure productive transcription. In addition, it cannot be ruled out that tssRNAs have a role on their own. We expect that the data presented here will inspire future studies to address these questions. One practical and immediate application of tssRNAs is in high-throughput studies, where tssRNAs (identified by sequencing the small RNA fraction) could be analyzed in combination with transcriptomic data to identify promoters in bacterial species.

Materials and methods

M. pneumoniae RNA (in the size groups of ~15–65, 65–100, 100–150, 150–200, and over 200 bases) and *E. coli* small RNA (~15–65 bases) from cells in exponential and stationary phases were subjected to direct strand-specific sequencing as previously described (Vivancos *et al*, 2010). tssRNAs were identified with an algorithm that takes into consideration their particular shape and context. High-resolution tiling arrays and Pribnow score calculations were performed as previously described (Guell *et al*, 2009). IVT, chromatin immunoprecipitation, and northern and western blotting are described in the Supplementary Methods. Error intervals represent the standard deviation.

Sequencing and tiling array data have deposited in the NCBI Short Read Archive (http://www.ncbi.nlm.nih.gov/sra) and Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) as data sets SRA051821 and GSE14019, respectively.

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

This work was supported by an European Research council (ERC) advanced grant, the Fundación Marcelino Botín, the Spanish Ministry of Research and ICREA (Institució Catalana de Recerca i Estudis Avançats). We thank the CRG ultrasequencing facility (Rebecca Curley and Heinz Himmelbauer) and the EMBL Genomics Core (Tomi Bähr-Ivacevic and Vladimir Benes) for technical advice. We kindly thank Veronica Raker for editorial help.

Author contributions: MG and EY designed the study, carried out the experiments, analyzed the data, prepared the figures, and wrote the manuscript. WH helped in the bioinformatics analysis of the tssRNAs and contributed to the software developed in this work. APV helped in the ultrasequencing data of the tssRNAs. MLS helped with the sample preparation. JD helped with the data processing. LS designed the study, performed the simulations, analyzed the data, discussed the results, and commented on the manuscript. ACG and PB contributed to the study design, discussed the results, and commented on the manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Burmann BM, Rosch P (2011) The role of E. coli Nus-factors in transcription regulation and transcription:translation coupling: from structure to mechanism. *Transcription* **2:** 130–134

Chekanova JA, Gregory BD, Reverdatto SV, Chen H, Kumar R, Hooker T, Yazaki J, Li P, Skiba N, Peng Q, Alonso J, Brukhin V, Grossniklaus U, Ecker JR, Belostotsky DA (2007) Genome-wide high-resolution mapping of exosome substrates reveals hidden features in the Arabidopsis transcriptome. *Cell* **131:** 1340–1353

Cserzo M, Turu G, Varnai P, Hunyady L (2010) Relating underrepresented genomic DNA patterns and tiRNAs: the rule behind the observation and beyond. *Biol Direct* **5:** 56

Gama-Castro S, Jimenez-Jacinto V, Peralta-Gil M, Santos-Zavaleta A, Penaloza-Spinola MI, Contreras-Moreira B, Segura-Salazar J, Muniz-Rascado L, Martinez-Flores I, Salgado H, Bonavides-Martinez C, Abreu-Goodger C, Rodriguez-Penagos C, Miranda-Rios J, Morett E, Merino E, Huerta AM, Trevino-Quintanilla L, Collado-Vides J (2008) RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res* **36**(Databaseissue): D120–D124

Goldman SR, Ebright RH, Nickels BE (2009) Direct detection of abortive RNA transcripts *in vivo*. *Science* **324:** 927–928

Guell M, van Noort V, Yus E, Chen WH, Leigh-Bell J, Michalodimitrakis K, Yamada T, Arumugam M, Doerks T, Kuhner S, Rode M Suyama M, Schmidt S, Gavin AC, Bork P, Serrano L (2009) Transcriptome complexity in a genome-reduced bacterium. *Science* **326:** 1268–1271

Hsu LM (2002) Promoter clearance and escape in prokaryotes. *Biochim Biophys Acta* **1577:** 191–207

Kapanidis AN, Margeat E, Ho SO, Kortkhonjia E, Weiss S, Ebright RH (2006) Initial transcription by RNA polymerase proceeds through a DNA-scrunching mechanism. *Science* **314:** 1144–1147

Kuhner S, van Noort V, Betts MJ, Leo-Macias A, Batisse C, Rode M, Yamada T, Maier T, Bader S, Beltran-Alvarez P, Castano-Diez D, Chen WH, Devos D, Guell M, Norambuena T, Racke I, Rybin V, Schmidt A, Yus E, Aebersold R *et al* (2009) Proteome organization in a genome-reduced bacterium. *Science* **326:** 1235–1240

Miura F, Kawaguchi N, Sese J, Toyoda A, Hattori M, Morishita S, Ito T (2006) A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *Proc Natl Acad Sci USA* **103:** 17846–17851

Nudler E, Gottesman ME (2002) Transcription termination and anti-termination in E. coli. *Genes Cells* **7:** 755–768

Roberts JW, Shankar S, Filter JJ (2008) RNA polymerase elongation factors. *Annu Rev Microbiol* **62:** 211–233

Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, Chabas S, Reiche K, Hackermuller J, Reinhardt R, Stadler PF, Vogel J (2010) The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature* **464:** 250–255

Shi Y, Tyson GW, DeLong EF (2009) Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature* **459:** 266–269

Shultzaberger RK, Chen Z, Lewis KA, Schneider TD (2007) Anatomy of Escherichia coli sigma70 promoters. *Nucleic Acids Res* **35:** 771–788

Taft RJ, Glazov EA, Cloonan N, Simons C, Stephen S, Faulkner GJ, Lassmann T, Forrest AR, Grimmond SM, Schroder K, Irvine K, Arakawa T, Nakamura M, Kubosaki A, Hayashida K, Kawazu C, Murata M, Nishiyori H, Fukuda S, Kawai J *et al* (2009) Tiny RNAs associated with transcription start sites in animals. *Nat Genet* **41:** 572–578

Taft RJ, Hawkins PG, Mattick JS, Morris KV (2011) The relationship between transcription initiation RNAs and CCCTC-binding factor (CTCF) localization. *Epigenetics Chromatin* **4:** 13

Vivancos AP, Guell M, Dohm JC, Serrano L, Himmelbauer H (2010) Strand-specific deep sequencing of the transcriptome. *Genome Res* **20:** 989–999

Weiner 3rd J, Herrmann R, Browning GF (2000) Transcription in Mycoplasma pneumoniae. *Nucleic Acids Res* **28:** 4488–4496

Yang X, Lewis PJ (2010) The interaction between bacterial transcription factors and RNA polymerase during the transition from initiation to elongation. *Transcription* **1:** 66–69

Yus E, Maier T, Michalodimitrakis K, van Noort V, Yamada T, Chen WH, Wodke JA, Guell M, Martinez S, Bourgeois R, Kuhner S, Raineri E, Letunic I, Kalinina OV, Rode M, Herrmann R, Gutierrez-Gallego R, Russell RB, Gavin AC, Bork P *et al* (2009) Impact of genome reduction on bacterial metabolism and its regulation. *Science* **326:** 1263–1268