nature microbiology

SUPPLEMENTARY INFORMATION

ARTICLE NUMBER: 15007 | DOI: 10.1038/NMICROBIOL.2015.7

NusA-dependent transcription termination prevents misregulation of global gene expression

Library preparation. Six independent replicates of the NusA depletion strain (PLBS802) were grown in minimal-ACH medium supplemented with 12.5 µg/ml tetracycline ± 30 µM IPTG and RNA was isolated using the RNeasy kit (Qiagen) as described previously (1). Ten µg of RNA from each sample was dephosphorylated with calf intestinal alkaline phosphatase (New England Biolabs) and RNA was recovered by phenol-chloroform extraction/ethanol precipitation. Five µg of dephosphorylated RNA from each sample was subjected to rRNA depletion using the Ribo-Zero rRNA Removal Kit for Gram-Positive Bacteria according to the manufacturer's instructions (Epicenter). To ligate the 3' RNA oligo (5'AAUGAGACACUGAGAUCCAGUCGAUGAGCUAddC3') to the RNA, 0.5 µg of RNA from each sample was incubated with 150 pmol oligo in 20 µl reactions containing RNA ligase buffer, 20% PEG-8000, and 20 U T4 RNA ligase 1 (New England Biolabs) at 16°C for 16 hr. The ligated RNA was subsequently purified using RNeasy columns (Qiagen) and examined with a Bioanalyzer (Agilent). Differentially barcoded libraries were generated using the TruSeq stranded mRNA library kit (Illumina) according to the manufacturer's instructions, except that the polyA selection step was omitted. The barcoded libraries were examined with a bioanalyzer and qPCR for size and concentration. Equal amounts of the libraries were pooled and subsequently sequenced on one lane of the Illumina HiSeq 2500 in Rapid Run mode using 100 x 100 paired end sequencing.

Data analysis. Illumina sequencing generated ~120 million reads for 12 samples (6 biological replicates for –NusA and +NusA each). The reads were parsed into separate samples based on barcode (~10 million paired-end reads per sample) and the illumina adaptors were trimmed using the trimmomatic tool (2). A subset of the dataset consisted of the reads with the unique oligo sequence ligated to their 3' ends, thus preserving the native 3' ends. Such reads were extracted and the oligo sequence was removed using the discard-untrimmed option of the cutadapt program (3). Thus, two clean data sets were obtained; one with all reads for standard RNA-seq (whole), and the other with just the reads that had the oligo sequence (oligo-only) for 3' end-mapping. Both whole and oligo-only data sets were mapped separately to the reference genome (*Bacillus subtilis str.168*, NC_000964.3, downloaded from NCBI) using the Burrows-Wheeler Aligner (BWA) short read aligner (4). Next, the 6 samples from each category (+IPTG and –IPTG) were separately merged together using SAMTools (5) and genome coverage graphs for the merged samples were generated using BEDTools (6). To make coverage values comparable across samples, the read counts at each position were normalized to library size by dividing them by the total number of mapped reads (~104×10^6 reads for +IPTG, ~109×10^6 reads for –IPTG) from the merged whole mapping dataset. Visualization of the mapping was performed using Integrative Genomics Viewer (IGV, Broad Institute). Reads that were mapped to the forward strand and reverse strands were separated using SAMTools. The strand-specific mapping files were used for all subsequent analyses.

DOI: 10.1038/NMICROBIOL.2015.7

Identification of 3' ends. The oligo-only mapping datasets were used to identify the native 3' end positions. For each position in the genome, the coverage variation (C_V) at that position was calculated as the difference of the average coverage of the 15 nt upstream (C_U) and the 15 nt downstream (C_D) relative to that position ($C_V = C_U - C_D$). Hence, the read coverage of the oligo-only dataset was transformed into a coverage variation dataset where every genomic location is characterized by change of read abundance around that location (C_V). In this transformed dataset, a large negative C_V corresponds to an increase in read abundance whereas a large positive C_V corresponds to a decrease of read abundance. We only processed locations with positive C_V values because our objective was to identify the 3' ends of the transcripts. Since locations with relatively high positive coverage variation (C_V) indicate sites of oligo-ligation, we performed peak detection to identify these sites. The heights of the peaks, described as peakheight hereafter, directly correspond to the coverage change that is present at the coordinate of the peak center. A step-size of >3 was used to calculate the local maxima that precisely identified the ligation sites (native 3' ends). The positions of the peaks in the +IPTG and –IPTG conditions were highly similar; in >99% of the cases the peak position was found to be within a window of ±3 nt. Peak positions that were not present within 5 nt in both +IPTG and –IPTG samples were considered nonspecific and removed. This method identified 14,455 peaks (7756 in the forward and 6699 in the reverse strand). To further reduce noise, we only considered peaks with peakheight higher than 10, resulting in 5332 considered peaks (2931 in the forward and 2401 in the reverse strand).

Terminator screening and characterization. The peakheight reflects the relative abundance of the 3' ends that are dependent on expression of the transcript and the efficiency of the mechanism by which the 3' end is generated. A peak with a high peakheight value indicates a highly abundant 3' end. In bacteria, abundant 3' ends can be generated by transcription termination, RNA processing, or via stabilization of RNA degradation intermediates, while the peakheight is directly related to the efficiency of these mechanisms. 3' ends can also be generated by abortive transcription, transcriptional arrest or by mechanical shearing during the process of RNA isolation. We initially screened the peaks using FindTerm (SoftBerry) (7) using a modified configuration file to find intrinsic terminators. Several of the peaks were crosschecked visually using Mfold (8) and RNAfold (9) to confirm the presence of a terminator.

Calculation of termination efficiency. To characterize the global effect of NusA on intrinsic terminators, we excluded peaks with a peakheight below 30 (transcripts with low abundance and/or terminators with very low efficiency) to avoid analyzing termination efficiencies with low statistical significance. We thus considered a total of 1492 terminators with a U-tract containing at least 3 U residues within the first 9 nt following a hairpin, of which 726 were from the forward strand and 766 from the reverse strand. In many cases, the 3' end occurred within a 2-3 nt window rather than at a fixed position. In such cases, the nucleotide at which maximum ligation occurred was considered as the point of termination (POT), although exonucleolytic trimming of the U-tracts following termination likely contributed to the 2-3 nt window in some cases. Termination efficiencies were determined for the standard RNA-seq merged datasets and for individual datasets (to determine the statistical significance), using an algorithm to compare the number of

terminated versus readthrough reads at each POT. The average number of reads in 10 nt windows were calculated both upstream and downstream of the POT. Three nt on both sides of the POT were excluded from this calculation to eliminate the impact of the 2-3 nt window described above. Percent termination (%T) was calculated using the equation:

$$\%T = \frac{U - D}{U} X\ 100$$

where, T is termination efficiency, U is the average number of reads in a 10 nt window upstream of the POT and D is the average number of reads in a 10 nt window downstream of the POT. Negative %T values were obtained for 19 terminators in which an overlapping promoter was present such that the transcription start site was within 14 nt of the POT, causing interference with the calculation. Such terminators were ignored for assessment of the effect of NusA on termination.

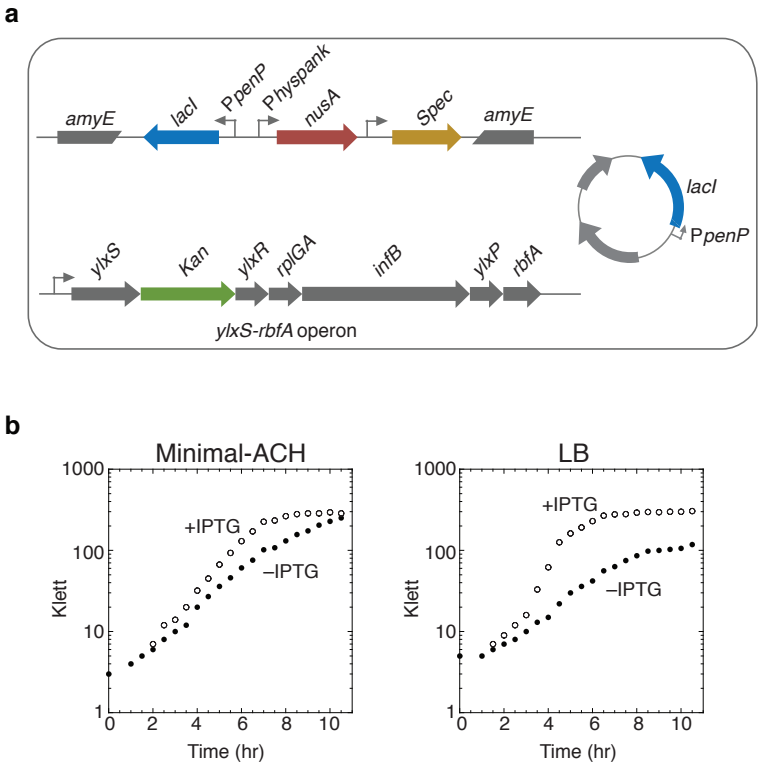
Differential expression and GO-term analysis. The expression levels of genes were measured and differentially expressed genes were identified using a combination of the EDGE-pro and DESeq programs. EDGE-pro (10) is designed to quantify the gene expression explicitly in prokaryotes. All trimmed reads were mapped to the reference genome using EDGE-pro in the paired-end mode. The output was converted to a format usable in DESeq (11) by using the edgeToDeseq.pl script included in the EDGE-pro package. DESeq is an R package designed to calculate the statistical significance of expression differences measured in RNA-seq data based on a negative binomial distribution. In this method, raw p-values are adjusted (p-adj) for multiple testing using the Benjamini-Hochberg method (12), which controls the false discovery rate. The read counts that were mapped to each gene in each sample were used as DESeq input. The top differentially expressed (twofold or more, p-adj < 0.05) genes were inspected manually in IGV to identify the mechanism by which it is differentially expressed. Genes that are regulated directly by NusA-dependent termination were analyzed using the DAVID (Database for Annotation, Visualization and Integrated Discovery, v6.7) program for GO-term enrichment using default settings (medium stringency). Enrichment scores above 1.3 are considered highly significant (13, 14).

DNA templates, plasmids and proteins. The *B. subtilis trp* 5'UTR was used for testing the effect of NusA on WT and mutant *trp* leader terminators (T_{trpL}). All other terminators were cloned in a plasmid from which DNA templates were amplified by PCR. Templates of λtR2 and naturally occurring *B. subtilis* terminators contained the common sequence 5'-GTCATTGACAAAAATACTGAATTGTGTTATAATAAGAACAGGTTAGAAATACACAAGAGTGTGT ATAAAGCAATCTGCAGCGCCGGGATCCGAAAGGATCTGCGCTGAAACTATtgccgcgcatgtcgcggcat tgttttcatggaagacgaaTATATAGTATTTTATCCTCTCATGTCATCTTCTCATTCTCC-3' (the -35 and -10 regions of the promoter are underlined) except that the terminator sequences (lowercase font) were different. Templates for λtR2 mutants were generated by PCR amplification using primers that introduced the mutation. The *ylxS*-leader T_1T_2 construct contained the two terminators in tandem. *B. subtilis* His-tagged RNAP (15), His-tagged NusA (16) and TRAP (17) proteins were purified as described previously. *E. coli* RNAP holoenzyme was obtained from Epicenter.

DOI: 10.1038/NMICROBIOL.2015.7

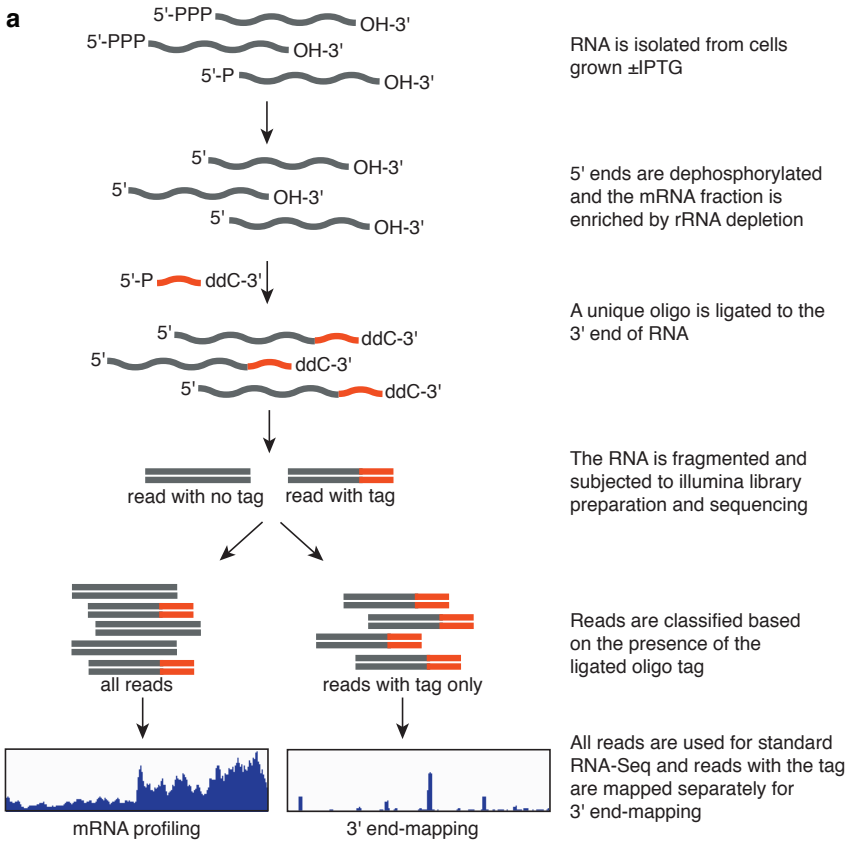
Supplementary References

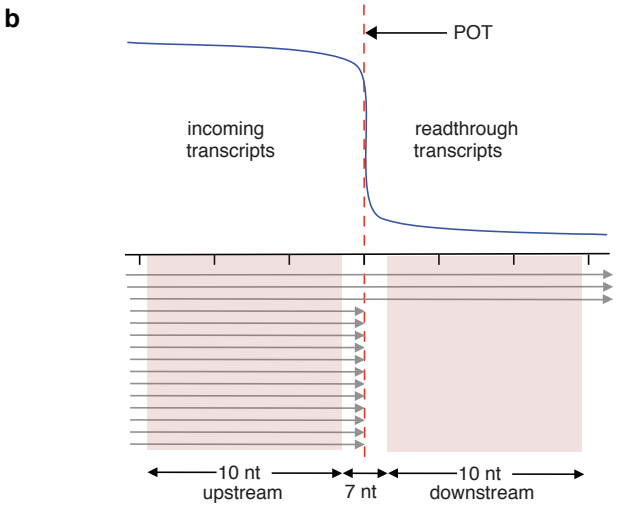
1. Yakhnin, H. *et al.* Complex regulation of the global regulatory gene *csrA*: CsrA-mediated translational repression, transcription from five promoters by $E\sigma^{70}$ and $E\sigma^{S}$, and indirect transcriptional activation by CsrA. *Mol. Microbiol.* **81,** 689–704 (2011).
2. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics* **30,** 2114–2120 (2014).
3. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal* **17,** 10–12 (2011).
4. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* **25,** 1754–1760 (2009).
5. Li, H. *et al.* The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
6. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
7. Solovyev, V. & Salamov, A. Automatic Annotation of Microbial Genomes and Metagenomic Sequences. *Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies*, ed Li, R. W., pp. 61–78 (Nova Science Publishers , 2011).
8. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31,** 3406–3415 (2003).
9. Bellaousov, S., Reuter, J. S., Seetin, M. G. & Mathews, D. H. RNAstructure: Web servers for RNA Secondary Structure Prediction and Analysis. *Nucleic Acids Res.* **41,** W471–W474 (2013).
10. Magoc, T., Wood, D. & Salzberg, S. L. EDGE-pro: Estimated Degree of Gene Expression in Prokaryotic Genomes. *Evol. Bioinform. Online* **9,** 127–136 (2013).
11. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
12. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57**, 289–300 (1995)
13. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc.* **4,** 44–57 (2009).
14. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37,** 1–13 (2009).
15. Yakhnin, A. V., Yakhnin, H. & Babitzke, P. RNA polymerase pausing regulates translation initiation by providing additional time for TRAP–RNA interaction. *Mol. Cell* **24,** 547–557 (2006).
16. Yakhnin, A. V. & Babitzke, P. (2002) NusA-stimulated RNA polymerase pausing and termination participates in the *Bacillus subtilis trp* operon attenuation mechanism in vitro. *Proc. Natl. Acad. Sci. USA* **99,** 11067–11072.
17. Yakhnin, A. V., Trimble, J. J., Chiaro, C. R. & Babitzke, P. Effects of mutations in the L-tryptophan binding pocket of the *trp* RNA-binding attenuation protein of *Bacillus subtilis*. *J. Biol. Chem.* **275,** 4519-4524 (2000).



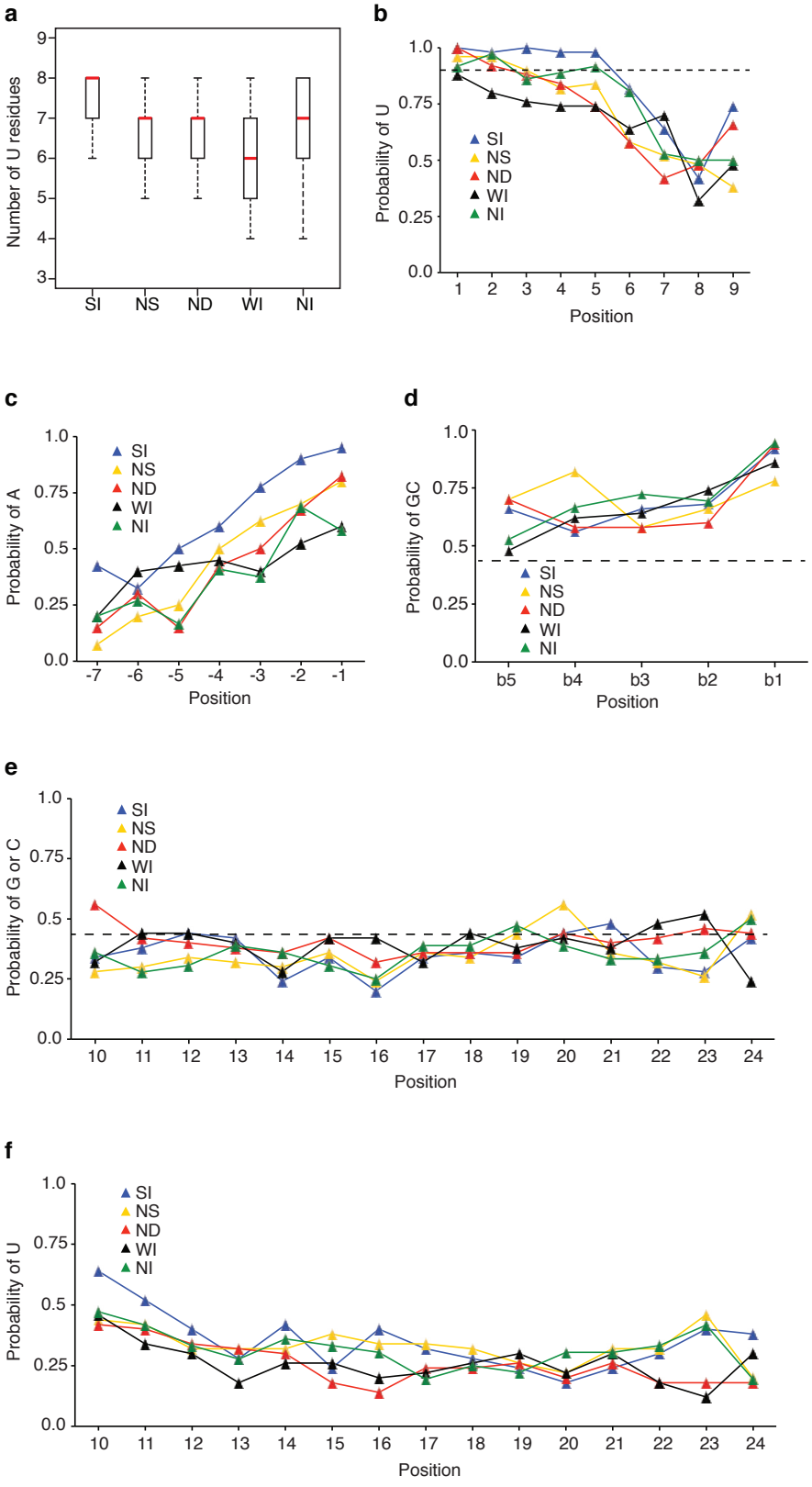
Supplementary Figure 1. The *B. subtilis* NusA depletion strain PLBS802. **a**, His-tagged *nusA* (red) under the control of an IPTG-inducible promoter (Physpank), along with *lacI* expressed from a constitutive promoter and a spectinomycin resistance gene, was integrated into the *amyE* locus. The chromosomal copy of *nusA* was then replaced with a kanamycin resistance gene. The strain also contains a plasmid expressing *lacI* for tighter repression. **b**, Growth curves of the NusA depletion strain. Cells were grown in minimal-ACH medium or LB in the absence (filled circles) or presence (open circles) of 30 μM IPTG.

DOI: 10.1038/NMICROBIOL.2015.7



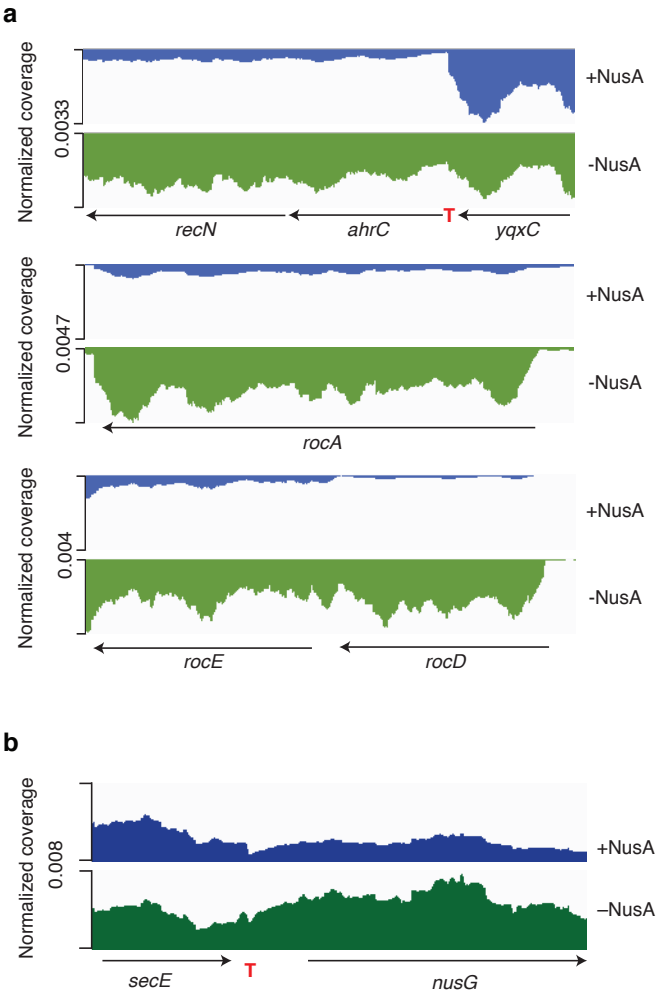


Supplementary Figure 2. Library construction and determination of termination efficiency (%T). **a**, Schematic representation of the library preparation for simultaneous mRNA profiling and 3' end-mapping. **b**, Calculation of termination efficiency (%T). The average number of reads in 10 nt windows were calculated both upstream and downstream of the point of termination (POT), leaving a 3 nt gap on both sides. %T was calculated as the percent of total reads that extend past the POT.

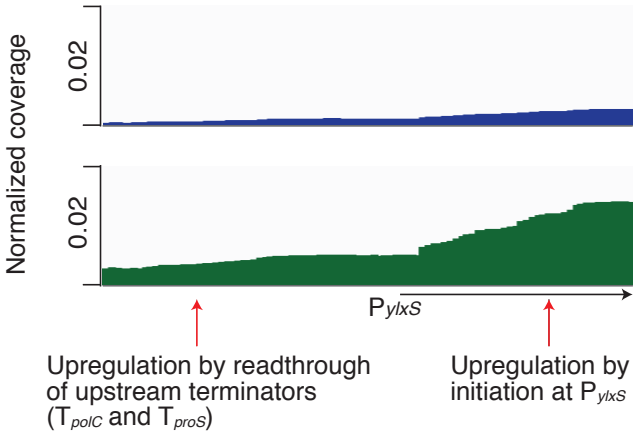


Supplementary Figure 3. Sequence features of terminators that contribute to stimulation by NusA. **a**, Uracil content of the U-tract (positions 1-9). **b**, Probability of U at positions 1-9 of the U-tract. **c**, Probability of A at positions –7 to –1 of the A-tract. **d**, Probability of G or C at positions b5-b1 of the stem. **e**, Probability of G or C at positions 10-24 of the downstream element. **f**, Probability of U at positions 10-24 of the downstream element. **b**, **c**, **d**, **e**, **f**, The top 50 terminators from the SI, WI, NS and ND classes were used, while 35 were used for NI.

DOI: 10.1038/NMICROBIOL.2015.7

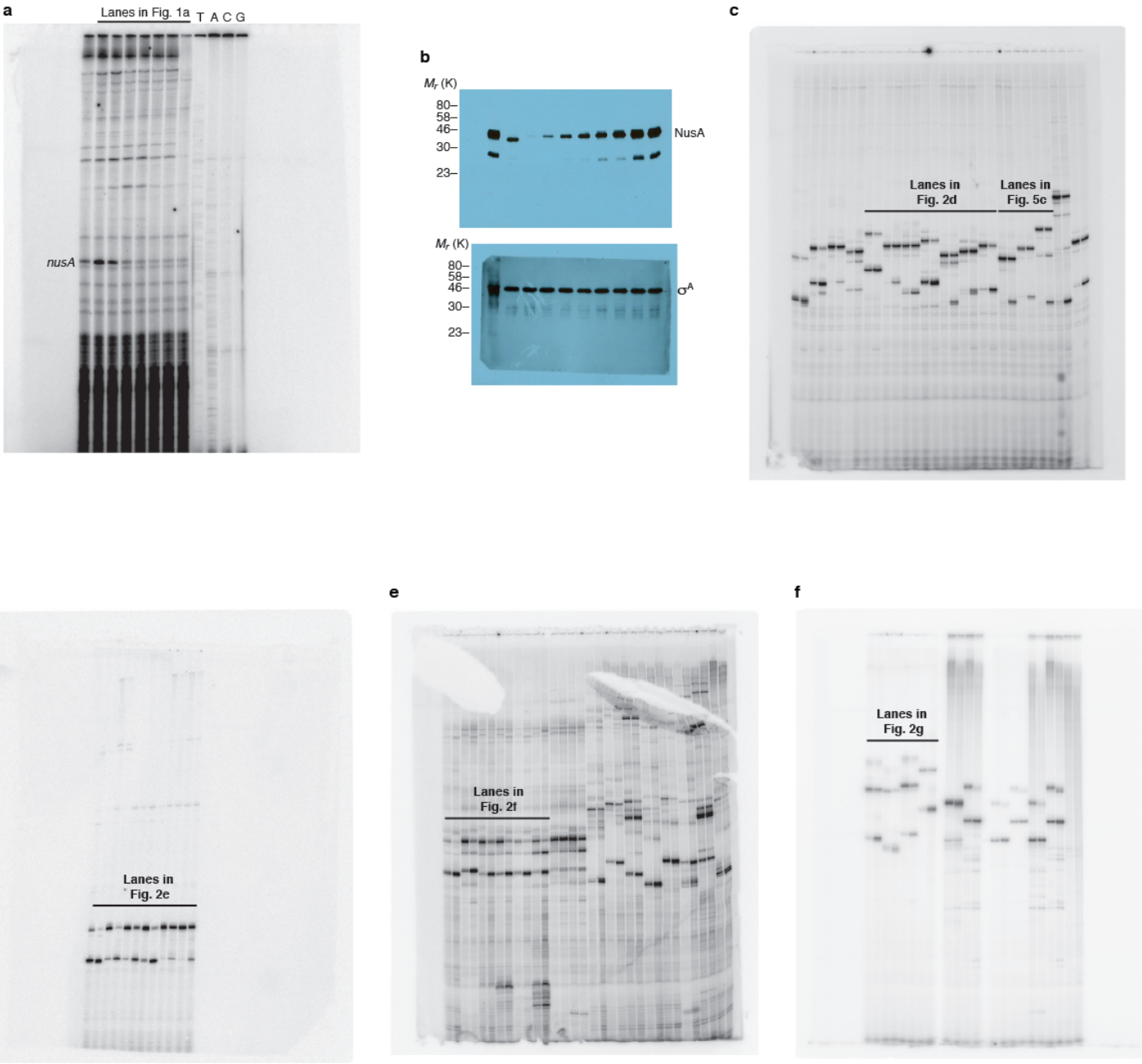


Supplementary Figure 4. NusA depletion alters gene expression directly and indirectly. a, Possible indirect effect of NusA on transcription initiation. An activator of arginine metabolism, *ahrC*, is overexpressed by enhanced readthrough of a NusA-dependent terminator. Overexpressed AhrC then activates transcription of the *rocABC* and *rocDE-argI* operons. **b**, The general elongation factor NusG is controlled by an upstream NusA-dependent terminator.



Supplementary Figure 5. NusA depletion causes increased transcription of the *nusA* operon via readthrough of upstream terminators and by increased transcription initiation at P*ylxS*. Transcriptional readthrough of the upstream NusA-dependent (T*proS*) and NusA-stimulated terminators (T*polC*) result in increased transcription into the *nusA* operon. There is also an indirect increase in transcription initiation at P*ylxS*. Normalized coverage ± NusA is at the same scale.

DOI: 10.1038/NMICROBIOL.2015.7



Supplementary Figure 6. Full size gels and blots from Figures 1, 2 and 5. a, Primer extension analysis from Fig. 1a. Sequencing ladder is shown on the right. **b**, Western blot analysis from Fig 1b. Molecular weight standards are shown on the left. A NusA degradation product is observed below the full length protein (top). **c**, *In vitro* transcription assays from Fig. 2d and Fig. 5c. **d**, *In vitro* transcription assays from Fig. 2e. **e**, *In vitro* transcription assays from Fig. 2f. **f**, *In vitro* transcription assays from Fig. 2g.