A pause sequence enriched at translation start sites drives transcription dynamics in vivo
Matthew H. Larson et al.
Science 344, 1042 (2014);
DOI: 10.1126/science.1251871

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by clicking here.

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines here.

The following resources related to this article are available online at www.sciencemag.org (this information is current as of June 3, 2014):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:
http://www.sciencemag.org/content/344/6187/1042.full.html

Supporting Online Material can be found at:
http://www.sciencemag.org/content/suppl/2014/04/30/science.1251871.DC1.html

This article cites 57 articles, 22 of which can be accessed free:
http://www.sciencemag.org/content/344/6187/1042.full.html#ref-list-1

This article appears in the following subject collections:
Molecular Biology
http://www.sciencemag.org/cgi/collection/molec_biol
transcription by RNA polymerase (RNAP) is interrupted by pauses that play diverse regulatory roles. Although individual pauses have been studied in vitro, the determinants of pauses in vivo and their distribution throughout the bacterial genome remain unknown. Using nascent transcript sequencing, we identified a 16-nucleotide consensus pause sequence in *Escherichia coli* that accounts for known regulatory pause sites as well as ~20,000 new in vivo pause sites. In vitro single-molecule and ensemble analyses demonstrate that these pauses result from RNAP–nucleic acid interactions that inhibit next-nucleotide addition. The consensus sequence also leads to pausing by RNAPs from diverse lineages and is enriched at translation start sites in both *E. coli* and *Bacillus subtilis*. Our results thus reveal a conserved mechanism unifying known and newly identified pause events.

The number of mapped reads at each genomic position is proportional to the number of RNAP molecules at that position. We observed well-defined single-nucleotide peaks within transcribed regions at known regulatory pause sites, including sites that synchronize transcription with translation, mediate RNA folding, or recruit translation factors. The NET-seq profiles also revealed a large number of other highly reproducible peaks in RNAP density throughout the genome (example gene in Fig. 1C). In total, we identified ~20,000 previously undocumentsed pause sites across well-transcribed genes, representing an average frequency of 1 per 100 base pairs (bp) (Fig. 1D). Thus, known regulatory pause sites represent a tiny fraction of actual pause events. We found that in vivo pause propensity depended strongly on the sequence identity at the 3′ end of the transcript (87% of paused transcripts end with either cytosine or uracil), as well as on the identity of the incoming nucleoside triphosphate (NTP) substrate (70% of pause sites occur before addition of guanosine 5′-triphosphate (GTP)) (Fig. 2A). Sequence dependence extends outside the RNAP active site to 11 nucleotides (nt) upstream and 5 nt downstream of the pause position, consistent with the extent of core nucleic-acid contacts made within the elongation complex (8). To determine the contribution of each base to pause duration, we used the density of reads in the NET-seq profile to calculate the relative dwell time of RNAP at each well-transcribed position in the genome. Modeling the addition of the next nucleotide as a process with a single activation barrier, we calculated the effective energetic barrier to nucleotide addition as the logarithm of the RNAP occupancy signal (supplementary materials). We used these values to determine the sequence dependence of this
A disparate group of pause sequences derive from a "same threshold correctly classified the group of above the threshold (Fig. 2C). Furthermore, the (fig. S5) and found that most pause sequences lay threshold for distinguishing these two populations characteristic analysis, we determined the optimal sequences (Fig. 2C). Using a receiver-operating well separated in sequence space from nonpause tions shows that pause-associated sequences were histogram of the energetics for the two popula-

which pausing was undetectable. A cumulative which pausing was observed, and sequences for were grouped into two categories: sequences for which pausing was observed, and sequences for which pausing was undetectable. A cumulative histogram of the energetics for the two populations shows that pause-associated sequences were well separated in sequence space from nonpause sequences (Fig. 2C). Using a receiver-operating characteristic analysis, we determined the optimal threshold for distinguishing these two populations (fig. S5) and found that most pause sequences lay above the threshold (Fig. 2C). Furthermore, the same threshold correctly classified the group of "canonical" regulatory pauses previously identified in E. coli, suggesting that this seemingly disparate group of pause sequences derive from a single consensus sequence. Intriguingly, the HIV-1 TAR pause element, which affects mammalian RNAP II (9), resembles our consensus sequence (Fig. 2C).

To understand the minimal requirements for pausing, we modified a high-resolution optical-trapping technique to measure sequence-resolved nucleotide addition by individual RNAP molecules in vitro (10, 11). By limiting the concentration of GTP, which is the nucleotide most frequently associated with pausing in vivo, its addition became rate limiting for elongation, allowing us to determine the absolute alignment of single-molecule records with the transcribed sequence. In this fashion, we measured the nucleotide addition rate for E. coli RNAP at more than 300 unique positions in a segment of the E. coli rpoB gene (Fig. 2D). These position-specific rates, which ranged over two to three orders of magnitude, yielded activation-energy barriers well correlated to those computed from NET-seq (Fig. 2, E and F). Moreover, they are qualitatively consistent with an in vitro consensus proposed previously from a small set of pause-inducing elements (12). This agreement suggests that interactions of RNAP with the DNA template and nascent transcript are sufficient for pausing in vivo and that these interactions largely dictate genome-wide pause patterns.

To probe individual elements of the consensus pause sequence, we reconstituted transcription complexes on a series of short, artificial nucleic-acid scaffolds. These scaffolds encoded either the consensus pause or an anti-consensus pause, in which the nucleotide at each position from -11 to +5 (excepting the highly conserved -1/+1 active-site positions) was altered to be the nucleotide predicted to cause the shortest dwell time (Fig. 3A). Strong pausing was observed at the expected position on the short consensus scaffold (Fig. 3A), and also on a template with the same consensus sequence embedded in a long DNA template (fig. S6). The consensus pause was roughly five times as long as the his pause (τ = 2 s at saturating GTP, Fig. 3B), even though the his pause is stabilized by a nascent RNA hairpin. Pausing was undetectable at the equivalent position on the anti-consensus scaffold (Fig. 3A). Thus, sequence elements upstream and downstream of the RNAP active site, although less enriched in our analysis, are essential for generating a pause signal. Consistent with prior proposals that discrete pause elements act together to form a multipartite pause signal (13), substitutions that disrupt RNA:DNA base-pairing at the -11 or -10 positions, remove the +1 non-template strand base, or alter the downstream DNA at positions +2 to +4 were found to reduce pause strength significantly (Fig. 3C; see fig. S7 for additional analysis of sequence dependence).

RNAP has the ability to "backtrack," shifting the transcript 3′ end downstream from the -1/+1 positions of the active site into the NTP-entry pore. Backtracking is resolved by cleavage of two or more nucleotides from the RNA, generating a

Fig. 1. Bacterial NET-seq provides a genome-wide view of transcription dynamics. (A) Nascent RNA is isolated from bacteria and converted to a DNA library sequenced with deep coverage. Reads are aligned to the reference genome and mapped according to their 3′ end, which corresponds to the RNAP active site. (B) An example of RNAP density in the his leader region (hisL) shows a peak at a single site that matches the previously mapped regulatory pause position (underlined). (C) Biological replicates along the ribosomal L10 protein subunit (rplJ). (D) Histogram of pause frequency for highly transcribed genes (n = 1984, gene average >1 read/bp) within the protein-coding sequence.
new 3' end in the active site. To determine whether RNAP backtracked at the consensus pause, we tested for transcript cleavage at the active site. Pause complexes reconstituted on the consensus scaffold cleaved only a single nucleotide, consistent with no backtracking, clearly different from the 2-nt cleavage observed with complexes prepared with an obligately backtracked scaffold, and also from complexes prepared with an anti-consensus scaffold (Fig. 3D). GreA, a cleavage factor in *E. coli* known to relieve backtracking, stimulated a 2-nt cleavage of the RNA at the consensus pause, but failed to reduce the pause dwell time (Fig. 3C and fig. S8), suggesting that the consensus pause sequence leads to a predominantly pretranslocated register that may be poised to backtrack, but that such backtracking does not principally determine the barrier to pause escape. It is likely that variations of the consensus sequence may lead to pauses that backtrack more readily. The observed pause profiles in vivo were unaffected by the deletion of GreA and GreB (Fig. 3E), suggesting that most transcriptional pauses in *E. coli* lead to an elemental non-backtracked pause state.

Pausing at the consensus sequence is conserved across diverse lineages, as demonstrated...
in vitro with RNAPs derived from *Rhodobacter sphaeroides* (Rsp), *Mycobacteria bovis* (Mbo), and *Thermus thermophilus* (Tth), which paused on the consensus template, but not on the anti-consensus template (Fig. 3C and figs. S9 and S10). Mammalian RNAPII (*B. taurus*, *Bta*) also responded to the consensus sequence (Fig. 3C), but exhibited a somewhat different pattern, involving pausing at the consensus position and even stronger pausing 1 nt downstream (fig. S11). Addition of the cleavage factor TFIIS converted the downstream pause to a strong pause at the consensus position, suggesting that the consensus pause leads to backtracking by RNAPII. This result is consistent with other evidence indicating a greater proclivity for eukaryotic RNAPII to backtrack as compared to bacterial RNAP (15).

**Fig. 3.** Pause consensus sequence leads to a long-lived, non-backtracked pause in vitro. (A) Purified *E. coli* RNAP was reconstituted on a nucleic-acid scaffold containing either the consensus pause sequence or an anti-consensus sequence. RNA nucleotides in lowercase were added after initial reconstitution by extension with α-32P-labeled or unlabeled NTPs. Full sequences are shown in fig. S7. A strong pause is observed at the predicted position on the consensus pause scaffold, but does not occur on the anti-consensus scaffold. (B) Consensus pause escape rate (SD of ≥3 replicates) as a function of GTP concentration reveals a maximal escape rate about one-fifth of that for the his pause. (C) Relative pause strengths for variants of the consensus pause (yellow), in the presence of transcription regulators, or with diverse RNAPs (SD of ≥3 replicates). (D) RNA active site–catalyzed hydrolytic cleavage of nascent RNA in complexes reconstituted with a 3’ mismatch forcing a backtracked register (left), at the pause site on the consensus pause scaffold (middle), and at the equivalent position on the anti-consensus scaffold (right). (E) Mean cross-correlation between NET-seq profiles for wild-type (WT) *E. coli* and *AgreA* (green), *AgreB* (red), or *AgreA/AgreB* (blue) strains for well-transcribed genes (*n* = 1240, gene average >1 read/bp for each sample). The mean autocorrelation for the WT strain is shown for comparison (black).
The average RNAP density across all genes exhibited a sharp peak within the start codon (Fig. 4A), at the juxtaposition of the ribosome-binding sequence (RBS; AGGAGG) and the ATG start codon, which are separated by an average spacing of 10 nt in *E. coli* (16) and consequently define the ends of a consensus pause sequence (Fig. 4B). Indeed, RBS substitutions abolished the start-codon pause for the *lacZ* gene in vivo (fig. S12). Similar to *E. coli*, we observed frequent pausing throughout the genome of the Gram-positive bacterium *B. subtilis*, with a consensus pause sequence characterized by −11G−10G and a −1 pyrimidine, but with A rather than G as the preferred +1 nt (fig. S13, A and B). Start-codon pausing also occurred in *B. subtilis*, just before the A of the ATG codon, placing it 2 nt earlier than the *E. coli* start-codon pause (Fig. 4C). The *B. subtilis* RBS, which generates the −11G−10G of the start-codon consensus pause, is, on average, 2 nt further upstream from the ATG codon than in *E. coli* (Fig. 4D) (16). Thus, the change in the consensus pause sequence in *B. subtilis* may reflect an evolved alteration that compensates for the 2-nt upstream shift of the RBS relative to the start codon (Fig. 4D).

In addition to start-codon pausing, RNAP also exhibits a pronounced tendency to pause within the first 100 nt of expressed genes, even though consensus pause sequences are not statistically overrepresented within these regions (Fig. 4A; compare RNAP density to predicted density). This 5′-proximal RNAP pausing may be increased until a ribosome can initiate translation and inhibit pausing during coupled transcription-translation (4, 5) (Fig. 4A), which likely explains the promoter-proximal buildup of *E. coli* RNAP previously observed by chromatin immunoprecipitation (17).

We have defined a consensus pause sequence that temporarily halts transcription at more than 20,000 unique sites in *E. coli*. Pauses are over-represented at ATG translation start codons, and this could direct folding of the 5′-untranslated region into structures that preserve accessibility of the RBS once transcription resumes (fig.
S14), consistent with the known ability of paused RNAP to influence nascent RNA folding (1) and the correlation between RBS accessibility and the rate of translation initiation (18, 19). The enhanced pausing downstream of the start codon (in the first 100 nt of genes) may also help preserve the unstructured RBS by limiting synthesis of additional RNA until translation starts. More generally, the conservation of pause sequences across diverse lineages suggests that consensus-sequence pausing may have evolved early in primitive organisms and was subsequently co-opted to control transcription in a variety of regulatory contexts, accounting for the diverse functions of transcriptional pausing observed today.

REFERENCES AND NOTES

ACKNOWLEDGMENTS
We thank O. Brandman, V. Chu, D. Larson, G. Li, C. McLean, and E. Simmons for critical reading of the manuscript and J. Lund and E. Chow for assistance with sequencing. This research was supported by the Center for RNA Systems Biology (J.S.W.), the Howard Hughes Medical Institute (J.S.W.), a Ruth L. Kirschstein National Research Service Award (M.H.L., J.M.P.), and grants from the NIH to C.A.G., S.M.B., W.J.G., and R.L. All data are deposited in Gene Expression Omnibus (accession number GSE56720). Jonathan Weissman and Shirling Churchman have submitted a patent on the NET-seq technology.

SUPPLEMENTARY MATERIALS
www.sciencemag.org/content/344/6187/1042/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S14
Tables S1 to S4
References (20–57)
6 February 2014; accepted 18 April 2014
Published online 1 May 2014;
10.1126/science.1251871