Widespread divergent transcription from bacterial and archaeal promoters is a consequence of DNA sequence symmetry

Warman, Emily; Forrest, David; Guest, Thomas; Haycocks, James; Wade, Joseph T.; Grainger, David

DOI:
10.1038/s41564-021-00898-9

License:
None: All rights reserved

Document Version
Peer reviewed version

Citation for published version (Harvard):
Warman, E, Forrest, D, Guest, T, Haycocks, J, Wade, JT & Grainger, D 2021, 'Widespread divergent transcription from bacterial and archaeal promoters is a consequence of DNA sequence symmetry', *Nature Microbiology*, vol. 6, no. 6, pp. 746-756. https://doi.org/10.1038/s41564-021-00898-9

Link to publication on Research at Birmingham portal

March 4th 2021

Widespread divergent transcription from bacterial and archaeal promoters is a consequence of DNA sequence symmetry

Emily A. Warman[1], David Forrest[1], Thomas Guest[1], James J.R.J. Haycocks[1],
Joseph T. Wade[2,3], David C. Grainger[1]*

[1]Institute for Microbiology and Infection, School of Biosciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

[2]Wadsworth Centre, New York State Department of Health, Albany, NY, 12208, USA

[3]Department of Biomedical Sciences, University at Albany, Albany, NY, 12201, USA

*for correspondence, d.grainger@bham.ac.uk Tel: +44 (0)121 4145437

ABSTRACT

Transcription initiates at promoters, DNA regions recognised by a DNA-dependent RNA polymerase. We previously identified horizontally acquired *Escherichia coli* promoters where the direction of transcription was unclear. Here, we show that more than half of these promoters are bidirectional. Using genome-scale approaches, we demonstrate that 19% of all transcription start sites detected in *E. coli* are associated with a bidirectional promoter. Bidirectional promoters are similarly common in diverse bacteria and archaea and have inherent symmetry: specific bases required for transcription initiation are reciprocally co-located on opposite DNA strands. Bidirectional promoters enable co-regulation of divergent genes and are enriched in both intergenic and horizontally acquired regions. Divergent transcription is conserved among bacteria, archaea and eukaryotes, but the underlying mechanisms for bidirectionality are different.

INTRODUCTION

Promoters are sections of duplex DNA that interact with RNA polymerase (RNAP) to stimulate transcription initiation[1]. In most organisms, promoters consist of ordered core elements with distinct roles[2,3]. For instance, the bacterial -10 element (consensus 5'-TATAAT-3') is usually indispensable and interacts with the housekeeping RNAP σ^{70} subunit (σ^{A} in some bacteria). The second and sixth positions of -10 elements are most critical; non-template strand bases interact with σ^{70} to stabilise DNA unwinding[4,5]. Position one is also important, and defines the upstream boundary of DNA melting[5]. Less conserved ancillary sequences can aid RNAP recruitment. For instance, the -35 element (consensus 5'-TTGACA-3') also contacts σ^{70}. Following initiation, σ^{70} is evicted from the elongation complex. In many eukaryotes and archaea, the TATA box functions analogously to the bacterial -10 element; TATA binding protein (TBP) facilitates DNA unwinding and serves as a scaffold for recruiting the transcriptional apparatus[6].

It has long been assumed that promoter sequences are directional, driving transcription in a single orientation determined by promoter element arrangement[2,7]. This view has been challenged in eukaryotes[8–11]. In addition to driving the production of a canonical sense mRNA, many RNAP II promoters simultaneously stimulate antisense transcription[12]. Permissive chromatin plays a key role; nucleosome-depleted DNA allows fortuitous binding of transcriptional activators that permit divergent transcription[12–15]. Thus, permissive sections of eukaryotic chromatin, not core promoters *per se*, give rise to bidirectionality[9,12,13,16]. The phenomenon is particularly prevalent for recently evolved promoter regions suggesting that, if beneficial, selection can fix mutations that favour unidirectional transcription[16]. In prokaryotic organisms, particularly the bacteria, chromosomes are not folded into structures reminiscent of eukaryotic chromatin[17]. The consensus view remains that transcription from bacterial promoter sequences is unidirectional[18]. Here, we show that bidirectional prokaryotic promoter sequences, resulting in divergent transcription, are in fact commonplace. However, the underlying molecular mechanisms are fundamentally different to those in eukaryotes.

RESULTS

Identification of bidirectional promoter sequences in horizontally acquired Escherichia coli genes

Transcription start sites (TSSs) in *Escherichia coli* have been mapped by detecting triphosphorylated RNA 5' ends[19]. These can be assigned to σ^{70} binding events identified using ChIP-seq[19]. We noticed that not all σ^{70} binding was associated with detectable RNA synthesis. This was particularly evident for horizontally acquired genes silenced by histone-like nucleoid structuring (H-NS) protein (Extended Data Fig. 1). We reasoned that RNAP might initiate transcription, but produce unstable RNAs, at these sites; similar to cryptic unannotated transcripts in eukaryotes[12]. To test this, we transcriptionally fused 33 such σ^{70} targets, derived from H-NS silenced genes, to *lacZ*. Translation prevents Rho mediated transcription termination. Hence, any RNAs produced should be stabilised and detectable[20]. As transcription orientation cannot be directly inferred from σ^{70} ChIP-seq data, DNA sequences were cloned in both directions (Fig. 1a). Over half of the fragments were transcriptionally active (≥ 2-fold above the background control) regardless of orientation (Fig. 1b). We designated the direction of highest *lacZ* expression as "forward". On average, "reverse" transcription neared half the "forward" activity (Fig. 1c). We repeated the experiment with 25 well-characterised promoter DNA fragments[21]. Importantly, we selected only promoters that did not contain a detectable TSS on the opposite DNA strand[19,22,23]. Such DNA fragments could only drive *lacZ* expression in the "forward" orientation (Extended Data Fig. 2). For an arbitrary subset of TSS pairs, we mapped RNA 5' ends (Fig. 1d). This allowed annotation of promoter elements (Fig. 1e). For the *yib*A2 DNA fragment, TSSs were convergent. For all other DNA fragments, TSSs were divergent. Hence, promoter elements for oppositely oriented transcripts mapped, partially or completely, to the same section of DNA (Fig. 1e and Extended Data Fig. 3). Mutation of these shared promoter sequences (Fig. 1e and Extended Data Fig. 3) reduced expression in both orientations (Fig. 1f).

Bidirectional promoter sequences are widespread and obey precise organisational rules

To understand global patterns of divergent transcription we analysed TSSs mapped by RNA 5' polyphosphatase sequencing (PPP-seq), dRNA-seq or cappable-seq[19,22,23]. Oppositely orientated TSSs tended to co-locate (Fig. 2a). To increase sensitivity, we merged the datasets (Fig. 2a, combined). This identified 5,292 divergent TSSs, defined as being separated by between 25 and 7 bp; 19 % of all detected TSSs in *E. coli* (Table S1). We refer to the associated promoter sequences as "bidirectional" and the corresponding TSSs as "divergent TSS pairs". The most common distance between divergent TSS pairs was 18 bp (Fig. 2a, top expansion). This corresponds to transcription initiation, on opposite

DNA strands, either side of the same promoter -10 region (Fig. 2a, top expansion). Presumably, promoter element symmetry must play a major role in creating bidirectional promoter sequences. To test this, we made a position weight matrix (PWM) describing all *E. coli* promoter sequences. The PWM was aligned with its reverse complement across a range of spacings (i.e. altered stagger between forwards and reverse PWM). We then calculated a symmetry score for each spacing. If the PWM resembled the same section of DNA in both orientations, the symmetry score increased. There was strong correlation between experimentally detected divergent TSS pairs and those predicted on the basis of symmetry score ($R^2 = 0.85$; Fig. 2a bottom expansion). Consistent with this, a DNA sequence logo generated by aligning divergent TSSs, separated by 18 bp, was symmetrical (Fig. 2b). Contrastingly, TSSs with no divergent transcript generated an asymmetrical motif (Fig. 2c). Recall that the second and sixth positions of σ^{70} promoter -10 elements are crucial[5]. Non-template strand bases at these positions, relative to the orientation of RNAP binding, are sequestered by σ^{70} to stabilise initial duplex melting[5]. At divergent TSS pairs offset by 18 bp, these key bases reciprocally coincide on opposite DNA strands. Hence, these positions are strongly conserved (Fig. 2b). An example of a -10 element with such symmetry is shown in Fig. 2d; the critical bases at positions two and six are underlined. Divergent transcription was also enriched for TSS pairs offset by 23, 12, 10 or 7 bp (Extended Data Fig. 4a). These configurations also correspond to reciprocal base pairing between key -10 element nucleotides and TSSs (Extended Data Fig. 4b). We note that, of the divergent TSS pairs detected within horizontally acquired sequences, two match one of the common configurations (Fig. 1). For instance, the divergent TSSs within *yigG* are 7 bp apart. To test whether the symmetrical sequences were intrinsically able to drive divergent transcription we used *in vitro* transcription assays. Bidirectional promoter sequences were cloned, in either the forward or reverse orientation, upstream of the λ*oop* terminator in plasmid pSR. Transcripts terminated by λ*oop* have a defined length and can be detected using electrophoresis. In all cases, regardless of the cloning orientation, bidirectional promoter sequences produced detectable transcripts terminated by λ*oop* (Extended Data Fig. 4c,d). Since *in vitro* transcription assays use no protein factors other than RNAP, divergent transcription must be an intrinsic property of bidirectional promoter DNA sequences.

Molecular basis for promoter sequence bidirectionality: a dual role for transcription start sites

In *E. coli* transcription preferentially initiates at an adenine (Fig. 2c). For divergent TSSs 18 bp apart, the +1 nucleotide corresponds to position -18 on the opposite DNA strand. Hence, -18 is most often a thymine (Fig. 2b). A thymine at position -18 can increase transcription by altering DNA bending[24]. This change in DNA conformation enhances the interaction between the nearby σ^{70} residue R451 and the DNA backbone[24] (Fig. 3a). Importantly, this can negate the need for a -35 element[25]. We speculated that the +1/-18 overlap could explain why this configuration is most frequently detected. To test this, we cloned a bidirectional promoter sequence, with 18 bp between TSSs, in both orientations upstream of the λ*oop* transcriptional terminator (Fig. 3bi). We also made derivatives where the A•T at each +1/-18 position was replaced with C•G (Fig. 3bii-iii). Note that cloning in both orientations was necessary because transcription directed away from λ*oop* is not precisely terminated. Hence, a discrete transcript is not produced. As expected, altering the TSS reduced production of the associated RNA (Fig. 3c, compare lane 1 with 5 and 3 with 11); the same mutations also reduced transcription in the opposite direction (compare lane 1 with 9 and 3 with 7). Though σ^{70} RA451 was defective at the bidirectional promoter sequences (even lane numbers to 12) it was unimpaired at a control promoter not requiring this contact (lanes 13-14).

Bidirectional promoter sequences are overrepresented at sites of mRNA synthesis

Of all divergent TSS pairs, 48% located to intergenic DNA (Fig. 2e). Of the resulting transcripts, 75 % are expected to be mRNAs, based on the orientation of the flanking genes. By comparison, only 29 % of directional TSSs were in intergenic regions, near a gene 5' end, with 89 % of associated transcripts expected to be mRNAs (Extended Data Fig. 5d). This suggests many bidirectional promoter sequences control gene expression. Hence, we determined if divergent TSS pairs mapped to well-characterised promoters, known to control mRNA production, listed in RegulonDB[26]. First, we examined the RegulonDB set of 317 mRNA TSSs identified by 5' RACE[27]. Of these TSSs, 311 were in our combined TSS dataset; a 156-fold enrichment compared to random genome co-ordinates. Enrichment was significantly more pronounced for divergent TSS pairs (252-fold) than directional TSSs (136-fold) ($P = 0.002$, Fisher's exact test). RegulonDB lists a further 3,330 mRNA TSSs, identified using numerous approaches, according to RNAP σ factor specificity[26]. The majority are σ^{70} dependent[26]. Of the 1,994 σ^{70} dependent TSSs, 1,410 are in our combined TSS dataset (a 113-fold enrichment). Moreover, σ^{70} dependent TSSs are significantly overrepresented amongst divergent TSS pairs (133-fold enrichment) compared to directional TSSs (108-fold enrichment) ($P = 0.032$, Fisher's exact test). Conversely, RegulonDB described promoters for alternative σ factors do not preferentially map to divergent TSS pairs (Table S2). This is consistent with divergent TSS pairs mapping to sequences resembling the σ^{70} -10 element (Fig. 2b and Extended Data Fig. 4b).

Length and stability of transcripts arising from bidirectional promoter sequences

To investigate length and stability of transcripts from bidirectional promoter sequences we used RNA-seq. We focused on divergent TSS pairs in non-coding regions; overlapping mRNA synthesis confounds analysis of intragenic promoters. After grouping intergenic loci according to adjacent gene orientation, we generated aggregate RNA coverage plots (Fig. 3d and 3e). At bidirectional promoter sequences between co-oriented genes, RNAs generated in each direction had different properties (Fig. 3d). Whilst non-coding antisense transcripts were detectable, coding transcripts were more abundant and longer. For bidirectional promoter sequences between divergent genes, two coding RNAs are expected. Hence, transcript abundance and length was similar in both directions (Fig. 3e). Fig. 3f illustrates examples of RNAs derived from divergent TSS pairs. Note that cappable-seq detects only RNA 5' ends whilst RNA-seq detects all RNA sequences.

Bidirectional promoter sequences are widespread in bacteria

Widespread divergent transcription from bacterial promoters has not been reported previously. However, a prior study did identify a modest number of divergent TSS pairs offset by 18 bp in *Pseudomonas aeruginosa*[28]. To determine the prevalence of bidirectional promoter sequences across the bacterial kingdom we analysed TSS maps for proteobacteria[19,22,23,28–31], actinobacteria[32,33], and a firmicute[34]. We also mapped TSSs in an additional firmicute, *Bacillus subtilis*, using cappable-seq (Extended Data Fig. 5 and Table S3). Co-localised divergent TSSs were abundant in all bacteria analysed (Fig. 4a). Proteobacteria and actinobacteria were most similar; divergent TSS pairs were usually offset by 18 or 19 bp as in *E. coli* (Extended Data Fig. 6). Firmicutes used the same range of -10 element configurations illustrated in Extended Data Fig. 4 for *E. coli* but with little preference for a single arrangement (Extended Data Fig. 6).

Bidirectional promoter sequences in archaea and bacteria are analogous

Archaeal transcription is closely related to that of eukaryotes; promoters have a TATA box and B recognition element (BRE; 5'-CGAAA-3'), located a narrow range of distances from the TSS[35]. Previously, Grünberger and co-workers noted divergent transcription from sites either side of a shared TATA box in *Pyrococcus furiosus*[36]. This resembles the scenario presented here for bacteria. We speculated that bidirectional promoter sequences should be widespread in archaea with multiple spacing preferences evident. We analysed TSS maps for the archaea *Thermococcus kodakarensis* and *Haloferax volcanii*[37,38]. We observed strong signatures of promoter sequence bidirectionality (Fig. 4a). In *T. kodakarensis*, divergent TSS pairs were predominantly offset by 52 bp, and located either side of a shared TATA box element (5'-TTATAAA-3') (Fig. 4b,c and Extended Data Fig. 7a). Less frequently, divergent TSS pairs were offset by 36 bp (Fig. 4b and Extended Data Fig. 7a). In this situation, the BRE is positioned so the initial C•G bp can also act as the TSS on the opposite DNA strand (Fig. 4c). Similar observations were made for *H. volcanii* despite the unusual TATA box consensus (5'-TTWT-3') of haloarchaea (Extended Data Fig. 7b,c).

Promoter sequences acquired by horizontal gene transfer are more frequently bidirectional

In eukaryotes, bidirectional promoters occur more frequently in recently acquired DNA[16]. We have shown that horizontally acquired bacterial genes, by virtue of their high AT-content, are enriched for promoter -10 elements and TSSs[19,39]. We reasoned that many such sites could represent bidirectional promoter sequences. Indeed, our initial analysis of 33 σ^{70} binding events, within horizontally acquired

DNA, is consistent with this view (Fig. 1). As predicted, detection of divergent TSS pairs using PPP-seq, increased in cells lacking H-NS; a protein that suppresses transcription at horizontally acquired DNA (Extended Data Fig. 8a). Parallel DNA sequence analysis demonstrated elevated promoter symmetry in foreign genes (Extended Data Fig. 8b). To understand other bacteria we utilised the TSS datasets described above. For both directional and bidirectional promoter sequences we determined the percentage of associated TSSs mapping to horizontally acquired sections of the cognate genome. Bidirectional promoter sequences were enriched in horizontally acquired regions for 6 of the 8 genomes analysed (Extended Data Fig. 8c).

Bidirectional promoter sequences allow coordinated regulation of divergent operons

The widespread occurrence of bidirectional promoter sequences has implications for our understanding of gene regulation. In the bacterium *Vibrio cholerae,* the genes VC1303 and VC1304 encode a para-aminobenzoate synthetase and a fumarate hydratase respectively. The divergent coding sequences share the same gene regulatory region (Fig. 5a). Examination reveals a divergent transcription start site pair with 23 bp spacing; the second most common configuration in both *E. coli* and *V. cholerae* (Fig. 5a and Extended Data Fig. 4). Here, reciprocal base pairing is observed between -10 element positions one and two (underlined in Fig. 5a). The intergenic region is also a target for the cyclic-di-GMP responsive transcription factor VpsT (identified using ChIP-seq, data to be presented elsewhere). Binding of VpsT was confirmed using DNAseI footprinting. As expected, in the absence of cyclic-di-GMP, VpsT was unable to bind the regulatory DNA (Fig. 5b, lanes 1-5). Conversely, in the presence of cyclic-di-GMP, VpsT protected a ~50 bp section of DNA from digestion (Fig. 5b, lanes 6-10). The expansion in Fig. 5a illustrates that the VpsT footprint overlaps the bidirectional -10 element. To investigate the impact of VpsT on transcription in each orientation, we first used *in vitro* transcription assays. We cloned the regulatory DNA, in either orientation, upstream of the λ*oop* terminator in plasmid pSR. In the absence of VpsT, transcripts of the expected size were detected in each orientation (Fig. 5c, lanes 1 and 3). When VpsT was added, production of both transcripts was greatly reduced (Fig. 5c, lanes 2 and 4). To understand the effect of VpsT *in vivo* we cloned the same promoter DNA fragment, in either orientation, upstream of *lacZ* in plasmid pRW50T. In both orientations, promoter activity was significantly reduced in *V. cholerae* expressing VpsT (Fig. 5d).

RNA polymerase complexes compete at bidirectional promoter sequences

At bidirectional promoter sequences RNAP can bind the same section of duplex DNA in two possible orientations. This binding cannot be simultaneous; structural constraints preclude this[40]. Instead, RNAP molecules likely compete to access the DNA duplex. We hypothesised that increased RNAP binding in one orientation would reduce transcription in the opposite direction. Since the promoter -35 element stabilises RNAP binding we introduced this sequence either side of a bidirectional -10 region. Our initial attempts to clone such DNA fragments in plasmid pSR failed. Specifically, we could not isolate recombinants with DNA inserts expected to generate high levels of reverse transcription. We reasoned that such transcription might interfere with expression of the upstream *bla* gene. Hence, we utilised a derivative of pSR with a λ*oop* terminator positioned upstream, as well as downstream, of the cloned DNA. The DNA constructs generated are shown in Fig. 5e. The presence of two transcriptional terminators allowed simultaneous detection of both forward and reverse RNA products following *in vitro* transcription (Fig. 5f, lane 1) dependent on σ^{70} side chain R451 (lane 2). Addition of a -35 element upstream of the -10 sequence increased transcription in the forward direction (lane 3, lower band). Concurrently, transcription in the reverse direction was reduced (lane 3, upper band). The inverse result was obtained if the -35 element was introduced downstream of the -10 region (compare lanes 3 and 5). When both -35 elements were present levels of divergent transcription increased in both directions (lane 7). However, increases were smaller than those detected with individual -35 elements (compare lanes 3, 5 and 7). Note that promoter -35 elements removed the requirement for σ^{70} residue R451 (compare lane 2 with lanes 4, 6 and 8). Indeed, the σ^{70} R451A derivative was moderately more active in such instances. Most likely, the R451-DNA contact hinders escape from near consensus promoters.

DISCUSSION

We demonstrate that divergent transcription from promoter sequences is a process conserved in all domains of life. The phenomenon is similarly frequent in diverse prokaryotes (Extended Data Fig. 9) and superficially resembles the situation in eukaryotes. However, the mechanistic basis is fundamentally different (Fig. 6). In eukaryotes, chromosomal regions associated with divergent transcription are large; bidirectionality is generated by nucleosome-depleted DNA and fortuitous binding of transcriptional activators[12–15]. Hence, divergent transcripts initiate from easily distinguishable sites separated by hundreds or thousands of base pairs, with no distance optimal. Accordingly, each TSS is associated with a distinct RNAP binding event involving non-overlapping DNA regions[9,12,13,16]. By contrast, bidirectional promoter sequences in bacteria have inherent symmetry. Hence, RNAP can bind the same section of duplex DNA in either orientation. Our global TSS analysis shows that symmetrical -10 elements are the main driver of divergent transcription (Fig. 2). This is consistent with the unique role of this promoter motif. Thus, whilst other promoter sequences stabilise RNAP binding, the -10 element also facilitates DNA opening and transcription initiation. Accordingly, ancillary promoter sequences are ineffective without an appropriately positioned -10 motif. We show that -10 elements, with inherent symmetry, can function independently to drive divergent transcription (Fig. 3c and Extended Data Fig. 4c). In the most common situation, the +1 and -18 positions on opposite strands align. This enhances the ability σ^{70} side chain R451 to stabilise RNAP binding. Interestingly, one example of a bidirectional -35 element was identified within the horizontally acquired *ygaQ* gene (Fig. 1e). We speculate that such configurations are more likely to arise in foreign DNA; the high AT-content ensures many potential -10 sequences are available. As in bacteria, divergent TSS pairs in archaea are separated by preferred distances, corresponding to key bases for transcription initiation overlapping on opposite DNA strands (Fig. 4c and Extended Data Fig. 7). The separation of TSSs by 34 bp and 36 bp in *H. volcanii* and *T. kodakarensis* respectively corresponds to alignment of the BRE (important for RNAP binding) and the +1 site of initiation on the opposite DNA strand. This is similar to alignment of positions -18 and +1 in bacteria.

Remarkably, despite the differences between prokaryotes and eukaryotes, our data suggest divergent transcription is often a property of newly acquired DNA in both kingdoms. Thus, nascent promoters can be inherently bidirectional. In bacteria, this is likely a consequence of both the DNA motif for divergent transcription, and horizontally acquired loci, having a high AT-content[19]. The abundance of non-canonical promoter elements is also likely to play a role[41]. Most sites of transcription within horizontally acquired genes are associated with non-coding RNA production. However, bidirectional promoter sequences elsewhere drive mRNA synthesis. Indeed, compared to directional promoters, divergent TSS pairs are more frequently found in intergenic regions, particularly between divergent genes (Fig. 2e and Extended Data Fig. 5d). Hence, divergent transcription must also have important implications for gene expression. For instance, we show that transcriptional repressors can co-regulate divergent operons by binding sites that overlap a bidirectional promoter sequence (Fig. 6). We also show that frequency of transcription in a given orientation impacts divergent RNA synthesis (Fig. 5f). Hence, bidirectional promoter sequences have inbuilt regulatory properties. Speculatively, divergent transcription could also displace adjacently bound transcription factors or generate asRNAs impacting adjacent genes. In conclusion, the widespread occurrence of bidirectional promoter sequences has important implications for understanding gene regulation in all prokaryotes.

MATERIALS AND METHODS

Strains, plasmids and oligonucleotides

All strains plasmids and oligonucleotides used are listed in Table S4. Standard procedures for strain and DNA manipulation were used throughout. All bacterial cultures were grown in LB media.

β-galactosidase assays

Assays were done according to the method of Miller[42]. Cells were grown in LB media supplemented with appropriate antibiotics to mid-log phase. Values shown are the mean of three independent experiments. Error bars represent the standard deviation of three independent experiments. Promoters were characterised as active if they stimulated β-galactosidase activity >2-fold over background levels generated by promoterless *lacZ*.

Identification of transcription start sites by primer extension

Transcript start sites were mapped for individual promoters using primer extension as described by Haycocks and Grainger[43]. The RNA was purified from indicated *E. coli* strains carrying different DNA fragments cloned in pRW50. The 5' end-labelled primer D49724, which anneals downstream of the *Hin*dIII site in pRW50, was used in all experiments. Primer extension products were analysed on denaturing 6% polyacrylamide gels, calibrated with size standards, and visualized using a Fuji phosphor screen and Bio-Rad Molecular Imager FX.

Genome-wide identification of divergent transcription start site pairs

Divergent TSS pairs at bidirectional promoter sequences were identified by calculating the distance between each TSS on the top and bottom DNA strands. The TSS were classified as divergent pairs if the bottom strand TSS was between 7 and 25 bp upstream of the top strand TSS. If a TSS on a given DNA strand could couple with multiple TSSs on the opposite DNA strand these were each counted as separate TSS pairs. Similarly, if directional promoter sequences were associated with multiple TSSs these also were individually counted. To compare TSSs in wild type *E. coli*, and the Δ*hns* derivative, we used our previously generated data[19] and remapped TSSs. This was done using TSSpredator (version 1.06)[44] with the following settings: step height 0.1, step height reduction 0.09, step factor 1.5, step factor reduction 0.5, enrichment factor 3, normalisation percentile 0.9, enrichment normalisation percentile 0.5, UTR length 300 and antisense UTR length 100. Cluster method was set to HIGHEST and all other parameters were set to 0. For all other datasets, we used TSS locations provided by the original studies. We designated TSSs as likely to drive mRNA synthesis if they were intergenic and in the correct orientation upstream of a gene. Note that previous PPP-seq analysis[19] was done according to the protocol of Singh and Wade[45].

Promoter symmetry scoring

To determine symmetry scores, we derived a PWM corresponding to sequences from -100 to +50 bp relative to each TSSs for each species test. We refer to this as the “forward PWM”. (Note that for the heatmap in Fig. 2a, the forward PWM was derived from sequences from -100 to +100 to facilitate analysis over a longer range of spacings; importantly, this does not impact the calculated scores). We then made a “reverse PWM” that corresponds to the reverse complement of the forward PWM, but was limited to sequences from -37 to +5 relative to the TSSs, since this is the range that includes all key promoter elements for all species tested. We aligned the forward and reverse PWMs across all possible spacing combinations. For each spacing, we calculated a symmetry score by (i) multiplying the fraction of each of the four nucleotides at each position of the forward PWM with the fraction of each of the complementary nucleotides at the overlapping position of the reverse PWM, and (ii) multiplying this value by 4, taking the log (base 2), and summing for all positions within the overlapping PWM positions. Stated R^2 values are Pearson product-moment correlation coefficients generated by comparing symmetry scores with TSS spacing abundance across the spacing range shown. Symmetry scores were also calculated for individual *E. coli* promoter sequences, to compare promoter sequences in horizontally acquired versus non-horizontally acquired regions, and to compare promoter sequences in H-NS-bound versus unbound regions. In these cases, we analysed individual promoter regions from position -100 to +50 relative to the TSS. We aligned the reverse PWM for *E. coli* (derived as described above) with each promoter sequence across all possible spacings. For each spacing, we determined the frequency of the nucleotide found in the promoter with the corresponding nucleotide frequency in the reverse PWM. We then multiplied these values for every position within the PWM. The final symmetry score for each promoter sequence was calculated as the maximum score across all possible spacings multiplied by a constant (to avoid extremely small numbers).

Promoter sequence analysis

To determine the distance between TSSs and promoter -10 elements we searched for the sequence 5'-TANNNT-3' in the 17 bp region upstream of the TSS. If this sequence did not occur, or occurred multiple times, the TSS was excluded to avoid ambiguities. To generate DNA sequence motifs we used Weblogo[46]. For directional *E. coli* promoters we created two alignments, anchored by either the position of the TSS or -10 element, that were then spliced together in the intervening DNA. This was required because the spacing between the +1 and -10 entities is variable (Extended Data Fig. 5a) and results in improper alignment unless taken into account (compare Fig. 2c and Extended Data Fig. 5b). This adjustment was not required for bidirectional promoters with TSSs separated by 18 bp (Fig. 2b). In this situation, juxtaposition of the TSSs and -10 elements are “locked” in place in accordance with Fig. 3 and the associated description.

Proteins

The *V. cholerae* RNAP holoenzyme was purified as described previously[47]. To facilitate overexpression, *vpsT* was cloned in pET28a and the resulting construct used to transform T7 express cells. Resulting transformants were used to inoculate 20 ml of LB media that was incubated overnight with shaking at 37 °C. Subsequently, this culture was used to inoculate 2 L of fresh LB media. The resulting culture was incubating at 37 °C with shaking until the OD_{650} reached 0.8. Overexpression of the encoded His_6-VpsT fusing was induced with 1 mM IPTG for 16 hours at 18 °C. Cells were then recovered by centrifugation, resuspended in 25 mM Tris-HCl pH 7.5, 550 mM NaCl, 20 mM Imidazole, and lysed by sonication. The cleared lysate was applied to a HisTrap (Amersham) column and bound proteins eluted with an imidazole concentration gradient up to 500 mM. Fractions containing VpsT were pooled and transferred into 25 mM Tris-HCl pH 7.5, 100 mM NaCl, 5 % (*v/v*) glycerol by dialysis. Contaminating proteins were removed using a HiTrap heparin HP (Amersham) column and pure His_6-VpsT was eluted with concentration gradient up to 1 M NaCl. Fractions containing the pure protein were pooled and concentrated to 1 mg/ml using a Vivaspin (Sartorius) concentrator. Precipitated protein was removed by filtration and the His_6 tag removed by thrombin digestion. The cleaved tag was separated from VpsT in a final HisTrap chromatography step. The pure VpsT was concentrated to 1 mg/ml and glycerol added to a final concentration of 50 % (*v/v*) for storage.

in vitro transcription assays

In vitro transcription reactions used the method of Kolb *et al.*[48] as described by Savery *et al.*[49]. Plasmid template DNA was isolated from *E. coli* transformed with pSR containing the appropriate promoter DNA fragment. Reaction buffer contained 20 mM Tris pH 7.9, 200 mM GTP/ATP/CTP, 10 mM UTP, 5 µCi (α^{32}P) UTP, 500 mM DTT, 5 mM $MgCl_2$, 100 µg ml^{-1} BSA and 0.2 mM cAMP. Template DNA (at a final concentration of 16 µg ml^{-1}) was incubated with RNAP holoenzyme, derived from either *E. coli* or *V. cholerae* as appropriate, to start the reaction.

Comparison with mRNA transcription start sites listed by RegulonDB

Matches between the combined TSS list, and TSSs listed in RegulonDB, were identified using the COUNTIF function in Microsoft Excel. When matching TSSs in RegulonDB to the combined list of *E. coli* TSSs we allowed a +/- 2 bp leeway. This was done because positions of equivalent TSSs, identified using different methods, often vary slightly. Additionally, RNAP can initiate transcription from a single promoter at one of several adjacent nucleotides. To calculate fold enrichment we first determined how many known TSSs listed in RegulonDB matched TSSs in our combined dataset. We then determined how many matches were identified if the positions of TSSs in our combined dataset were randomised to any position in the genome. To calculate the stated fold enrichment the former result was divided by the latter. We used the same approach with subsets of the combined TSS dataset corresponding to directional or bidirectional promoter sequences. We note that, in all cases, similar results were obtained if the positions of TSSs in our combined dataset were instead randomised to equivalent genomic contexts (i.e. intragenic and intergenic promoters to coding and non-coding regions respectively). This was expected because 19 % of TSSs in the combined list were in intragenic regions. This is comparable to the 12 % of the *E. coli* genome annotated as non-coding. To test for significant enrichment of different TSSs groups in RegulonDB, amongst the combined list of TSSs presented here, we used a hypergeometric test. For this test, the number of successful draws was the number of RegulonDB TSSs identified amongst the set of divergent 5,292 TSSs or the set of 23,813 directional TSSs (i.e. the total number of draws). The population size was 4,639,675 (i.e. the number of bp in the *E. coli* U00096.2 genome) and the number of successes in the population was the number of TSSs in the RegulonDB group being tested. To determine if there was a significant difference between the number of RegulonDB TSSs found in the lists of divergent TSSs and directional TSSs we used Fisher's exact test. Our null hypothesis was no difference in enrichment. To determine the expected number of RegulonDB promoters amongst the bidirectional TSSs, we calculated the relative frequency of RegulonDB promoters in the set of 23,813 directional TSSs. We then multiplied the relative frequency value by 5,292 (the number of divergent TSSs). These values were then compared to the experimental data.

Mapping global transcript abundance by RNA-seq

Duplicate cultures of *E. coli* strain MG1655 were grown until mid-exponential phase in LB media with shaking at 37 ºC. Cells were harvested by centrifugation, flash frozen in liquid nitrogen, and lysed by RNAsnap[50]. Total RNA was then purified from lysates using the Qiagen Mini RNeasy kit. Library preparation and sequencing was done by Vertis Biotechnologie AG. Briefly, RNA molecules were fragmented by sonication before ligation to oligonucleotide adapters at their 3' end. First-strand cDNA synthesis was done using M-MLV reverse transcriptase and the 3' adapter as primer. The first-strand cDNA was purified and the 5' Illumina TruSeq adapter was attached at the 3' end. After PCR amplification the cDNA was purified using the Agencourt AMPure XP kit (Beckman Coulter Genomics) and analysed by capillary electrophoresis. Libraries were sequenced on an Illumina Nextseq 500 system with a read length of 75 bp. Fastq files were deposited in Array Express (accession number E-MTAB-9655). Individual sequence reads were mapped using Bowtie2[51]. The reference genome for *E. coli* was that assigned Genbank accession number U00096.3. Resulting Binary Alignment Map (BAM) files were used to generate wiggle plots using bam2wig.py[52,53]. These data, were used to generate the aggregate plots shown in Fig. 3. For each dataset, we calculated the 10 % trimmed mean of the read depth, in 10 bp bins, across all 3 kb regions centred on selected TSSs. We focused our analysis on TSSs in non-coding regions to avoid confounding signals from overlapping mRNA transcripts.

Identification of transcription start sites by cappable-seq

To map TSSs globally we used cappable-seq. Duplicate cultures of *B. subtilis* strain 168 ca were grown until mid-exponential phase in LB media with shaking at 37 ºC. Cells were harvested and flash frozen in liquid nitrogen. Total RNA was isolated as described previously with the exception that RNA concentration and quality was determined on an Agilent 2200 Tapestation following the manufacturer's instructions[54]. Library preparation and sequencing was done by Vertis Biotechnologie AG according to the protocol described by Ettwiller *et al*.[22]. Briefly, 5' triphosphorylated RNA was capped with 3'-desthiobiotin-TEG-guanosine 5' triphosphate (DTBGTP) using Vaccinia capping enzyme (New England Biolabs). Biotinylated RNA was captured and eluted from streptavidin beads to obtain 5' fragments of primary transcripts. These transcripts were poly(A) tailed with poly(A) polymerase before conversion of the 5' CAP moiety to a 5' monophosphate using CAP-clip Acid pyrophosphatase (Cellscript). An RNA adapter was ligated to the 5' monophosphate and cDNA synthesis was done with an oligo(dT)-adapter primer and M-MLV reverse transcriptase. cDNAs were amplified by PCR to a final concentration of 10-20 ng µl^{-1}. Full length cDNAs were fragmented and immobilised with streptavidin magnetic beads for blunting and ligation of the 3' Illumina sequencing adapter. The immobilised cDNA fragments were amplified via PCR. The sample libraries were mixed in equimolar amounts 200-500 bp fragments were purified from an agarose gel after electrophoresis. The libraries were sequenced on an Illumina Nextseq 500 system with a read length of 75 bp. Individual sequence reads were mapped using Bowtie2[51]. The *B. subtilis* reference genome was that assigned Genbank accession numbers NC_000964.3. Resulting Binary Alignment Map (BAM) files were used to generate wiggle plots using bam2wig.py[52,53]. For each strand of the chromosome, we assigned TSSs to base positions where the read depth increased more than 3-fold, compared to the previous base, in both experimental replicates.

DNAse I footprinting

DNA fragments were excised from pSR using *AatII* and *Hin*dIII. After end-labelling using γ^{32}-ATP and T4 PNK (NEB), footprints were done as previously described in buffer containing 40 mM Tris acetate pH 7.9, 50 mM KCl, 5 mM $MgCl_2$, 500 µM DTT and 12.5 µg/ml Herring Sperm DNA[47]. Resulting DNA fragments were analysed on a 6 % denaturing gel. Subsequently, dried gels were exposed to a Biorad phosphorscreen that was scanned using a Biorad Personal Molecular Imager.

Assignment of transcription start sites to horizontally acquired DNA

To identify horizontally acquired genomic regions in different bacteria we used DarkHorse with genus level phylogenetic granularity[55]. Sections of DNA with high or low H-NS binding were identified using the ChIP-seq analysis of Kahramanoglou *et al*[56].

DATA AVAILABILITY

The data that support these findings are available from the corresponding author on request. The *E. coli* RNA-seq, and *B. subtilis* cappable-seq, data are available in Array Express using accession numbers E-MTAB-9655 and E-MTAB-8582 respectively.

FIGURE LEGENDS

Figure 1: Transcription start site pairs within horizontally acquired genes. a) β-galactosidase activity derived from cryptic RNAP binding sites. Data are presented as mean values (n = 3 independent experiments) +/- SD and individual data points are overlaid as dot plots. b) Direction of transcription from cloned DNA fragments. c) Average forward or reverse β-galactosidase activity of all DNA fragments. d) Start sites mapped by primer extension for selected DNA fragments (orientations labelled a or b). Primer extension products in lanes 1 to 10, sizes in nucleotides (nt). Lanes 11-14 are sequencing reactions for calibration. e) Schematic representation of transcription start site pairs. Core promoter element sequences in the forward or reverse orientation are indicated by solid or open rectangles respectively. Speckled shading indicates converge of promoter elements on the same section of DNA. Transcription start sites shown as bent arrows. The positions of mutations (x) or deletions (Δ) are indicated. f) Effect of mutating shared core promoter elements. Data are presented as in panel a.

Figure 2: Widespread divergent transcription from bidirectional promoter sequences in *Escherichia coli*. a) Heatmaps made using global transcription start site (TSS) data [19,22,23] or position weight matrix analysis. TSSs on the top chromosome strand are aligned at the centre of the heatmap (bent arrow, labelled +1). Heatmap colour indicates abundance of bottom strand TSSs at that position.

The expansion shows the occurrence of bottom strand TSSs in a 50 bp window either side of all top strand promoters. b) Predominant DNA sequence motif associated with bidirectional or c) directional promoters. The x-axis break indicates the variable distance between -10 element and TSS at directional promoters. Each sequence motif was generated from 638 aligned promoters. d) a bidirectional promoter sequence between the *E. coli pfs* and *dgt* genes. TSSs are in uppercase. Promoter -10 elements are bold. Key sites of -10 element symmetry are underlined and correspond to the strongly conserved bases in panel b. The non-template strand bases at these positions, relative to the direction of transcription, are sequestered by σ^{70} to stabilise initial DNA unwinding[5]. e) Categorisation of bidirectional *E. coli* promoters according to nearby gene organisation. Percentages indicate the proportion of bidirectional promoters in each genomic context. For comparison, 89 % of the *E. coli* genome is coding whist 6 %, 3 % and 2 % is intergenic DNA between co-directional, divergent and convergent genes respectively.

Figure 3: Reciprocal stimulation between divergent transcription start sites. a) Structure of RNAP bound to DNA (PDB: 6CA0)[57]. Relevant features labelled. b) DNA templates used for *in vitro* transcription. Sequences of promoter -10 elements (labelled) and TSSs (bent arrows) are shown. Plasmid vector DNA is shown by black lines and opposing DNA strands of the cloned bidirectional promoter sequence are shown by teal or grey lines. Interaction of σ^{70} R451 and the DNA backbone is indicated by dashes. Note that only transcription towards the λ*oop* terminator produces an RNA of defined length, detectable as a discrete band, following electrophoresis. Hence, to detect transcription in the opposite direction, it was necessary to invert the orientation of the cloned DNA sequence. c) Products of *in vitro* transcription (using templates in panel b) using either σ^{70} or the R451A derivative. The RNAI transcript is derived from the replication origin of the plasmid DNA template. The transcript of interest/RNAI signal intensity is 0.16, 0.05, 0.56, 0.11, 0.12, 0.06, 0.17, 0.09, 0.06, 0.05, 0.15, 0.06, 1.24 and 1.30 for lanes 1 to 14 respectively. The control promoter has the sequence 5′-TTGGCATATGAAATTTTGAGGATTATACTACACTTA-3′. A representative example of two separate experiments is shown. d,e) Aggregate profiles of transcription detected by genome-wide RNA-seq experiments. Each plot illustrates averaged sequence read depth across all 3 kb regions centred on bidirectional promoter sequences in non-coding DNA. Shaded areas of plots indicate signals above the background level f) A 17.5 kb section of the *E. coli* genome aligned with cappable-seq and RNA-seq reads mapping to the top (teal) or bottom (grey) DNA strands. Genes are denoted by red block arrows. Transcription start sites (TSSs) are denoted by gridlines and bent back arrows. Double arrow heads indicate divergent TSS pairs at bidirectional promoter sequences.

Figure 4: Bidirectional promoter sequences are widespread in prokaryotes. a,b) Heatmaps indicate abundance and position of TSSs on the bottom DNA strand, relative to the nearest top strand promoter (bent arrow). Species and phylogenetic relationships are indicated to left of heatmaps. c) DNA sequence motifs derived from divergent TSSs in *T. kodakarensis*.

Figure 5: Coordinated regulation of divergent transcription units from bidirectional promoter sequences. a) Organisation of the region between VC1303 and VC1304 in *Vibrio cholerae*. Transcription start sites are shown by bent arrows (+1) and the region footprinted by VpsT is underlined. The bidirectional promoter -10 region is bold with key positions of symmetry underlined. Position numbers indicate distances from the downstream end of the cloned DNA fragment subsequently used. b) Pattern of DNAse I digestion with or without VpsT (2, 3, 4 or 5 μM) and cyclic-di-GMP (50 μM). The gel is calibrated with a Maxam-Gilbert GA ladder. The region protected by VpsT marked by a blue bar (triangles indicate VpsT induced DNAse I hypersensitivity). A representative example of three experiments is shown. c) Transcripts generated from the VC1303-VC1304 intergenic region by RNA polymerase *in vitro* with or without 2 μM VpsT and 50 μM cyclic-di-GMP. The transcript of interest/RNAI signal intensity is 0.11, 0.02, 0.18 and 0.05 for lanes 1 to 4 respectively. A representative example of two separate experiments is shown. d) β-galactosidase activity derived from the VC1303-VC1304 intergenic region cloned in either orientation upstream of *lacZ*. Cells were supplied with VpsT

from plasmid pAMNF. Empty plasmid was used as a control. Bars are mean values (n = 3 independent experiments) +/- SD with individual data points overlaid as dot plots. *P* was derived from a two-sided paired student's *t*-test e) DNA templates to assess competition between RNA polymerase molecules during transcription *in vitro*. Promoter -10 and -35 elements are shown by black rectangles. TSSs are indicated by bent arrows. Plasmid vector DNA shown as black lines and opposing DNA strands of cloned bidirectional promoter sequences as teal or grey lines. f) Products of *in vitro* transcription using templates in panel e. The RNAI transcript is derived from the replication origin of the plasmid DNA template. The 134 nt RNA/129 nt RNA signal intensity is 0.68, not detectable, 0.09, 0.09, 4.48, 5.96, 0.39 and 0.32 in lanes 1 to 8 respectively. A representative example of two separate experiments is shown.

Figure 6: Promoter bidirectionality has a different basis in prokaryotes and eukaryotes.

REFERENCES

1. Mejía-Almonte, C. *et al.* Redefining fundamental concepts of transcription initiation in bacteria. *Nat. Rev. Genet.* (2020) doi:10.1038/s41576-020-0254-8.

2. Browning, D. F. & Busby, S. J. W. The regulation of bacterial transcription initiation. *Nat. Rev. Microbiol.* **2**, 57–65 (2004).

3. Haberle, V. & Stark, A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat. Rev. Mol. Cell Biol.* **19**, 621–637 (2018).

4. Bae, B., Feklistov, A., Lass-Napiorkowska, A., Landick, R. & Darst, S. A. Structure of a bacterial RNA polymerase holoenzyme open promoter complex. *Elife* **4**, (2015).

5. Feklistov, A. & Darst, S. A. Structural basis for promoter -10 element recognition by the bacterial RNA polymerase σ subunit. *Cell* **147**, 1257–1269 (2011).

6. Kramm, K., Engel, C. & Grohmann, D. Transcription initiation factor TBP: old friend new questions. *Biochem. Soc. Trans.* **47**, 411–423 (2019).

7. Butler, J. E. F. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev.* **16**, 2583–2592 (2002).

8. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008).

9. Seila, A. C. *et al.* Divergent transcription from active promoters. *Science* **322**, 1849–1851 (2008).

10. Preker, P. *et al.* RNA exosome depletion reveals transcription upstream of active human promoters. *Science* **322**, 1851–1854 (2008).

11. He, Y., Vogelstein, B., Velculescu, V. E., Papadopoulos, N. & Kinzler, K. W. The antisense transcriptomes of human cells. *Science* **322**, 1855–1857 (2008).

12. Neil, H. *et al.* Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature* **457**, 1038–1042 (2009).

13. Scruggs, B. S. *et al.* Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. *Mol. Cell* **58**, 1101–1112 (2015).

14. Rege, M. *et al.* Chromatin Dynamics and the RNA Exosome Function in Concert to Regulate Transcriptional Homeostasis. *Cell Rep.* **13**, 1610–1622 (2015).

15. Wu, X. & Sharp, P. A. XDivergent transcription: A driving force for new gene origination? *Cell* vol. 155 990 (2013).

16. Jin, Y., Eser, U., Struhl, K. & Churchman, L. S. The Ground State and Evolution of Promoter Region Directionality. *Cell* **170**, 889-898.e10 (2017).

17. Remus T. Dame, F.-Z. M. R. and D. C. G. Chromosome organization in bacteria: mechanistic insights into genome structure and function. *Nat. Rev. Genet.* (2019).

18. Browning, D. F. & Busby, S. J. W. Local and global regulation of transcription initiation in bacteria. *Nat. Rev. Microbiol.* **14**, 638–650 (2016).

19. Singh, S. S. *et al.* Widespread suppression of intragenic transcription initiation by H-NS. *Genes Dev.* **28**, 214–219 (2014).

20. Mitra, P., Ghosh, G., Hafeezunnisa, M. & Sen, R. Rho Protein: Roles and Mechanisms. *Annu. Rev. Microbiol.* **71**, 687–709 (2017).

21. Keseler, I. M. *et al.* The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12. *Nucleic Acids Res.* **45**, D543–D550 (2017).

22. Ettwiller, L., Buswell, J., Yigit, E. & Schildkraut, I. A novel enrichment strategy reveals unprecedented number of novel transcription start sites at single base resolution in a model prokaryote and the gut microbiome. *BMC Genomics* **17**, 199 (2016).

23. Thomason, M. K. *et al.* Global transcriptional start site mapping using differential RNA sequencing reveals novel antisense RNAs in Escherichia coli. *J. Bacteriol.* **197**, 18–28 (2015).

24. Singh, S. S., Typas, A., Hengge, R. & Grainger, D. C. Escherichia coli σ 70 senses sequence and conformation of the promoter spacer region. *Nucleic Acids Res.* **39**, 5109–5118 (2011).

25. Warman, E., Forrest, D., Wade, J. T. & Grainger, D. C. Widespread divergent transcription from prokaryotic promoters. *bioRxiv* **44**, 2020.01.31.928960 (2020).

26. Santos-Zavaleta, A. *et al.* A unified resource for transcriptional regulation in Escherichia coli K-12 incorporating high-throughput-generated binding data into RegulonDB version 10.0. *BMC Biol.* **16**, (2018).

27. Mendoza-Vargas, A. *et al.* Genome-Wide Identification of Transcription Start Sites, Promoters and Transcription Factor Binding Sites in E. coli. *PLoS One* **4**, e7526 (2009).

28. Gill, E. E. *et al.* High-throughput detection of RNA processing in bacteria. *BMC Genomics* **19**, 223 (2018).

29. Papenfort, K., Förstner, K. U., Cong, J. P., Sharma, C. M. & Bassler, B. L. Differential RNA-seq of Vibrio cholerae identifies the VqmR small RNA as a regulator of biofilm formation. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E766–E775 (2015).

30. Kröger, C. *et al.* The primary transcriptome, small RNAs and regulation of antimicrobial resistance in Acinetobacter baumannii ATCC 17978. *Nucleic Acids Res.* **46**, 9684–9698 (2018).

31. Sharma, C. M. *et al.* The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature* **464**, 250–255 (2010).

32. Cortes, T. *et al.* Genome-wide Mapping of Transcriptional Start Sites Defines an Extensive Leaderless Transcriptome in Mycobacterium tuberculosis. *Cell Rep.* **5**, 1121–1131 (2013).

33. Jeong, Y. *et al.* The dynamic transcriptional and translational landscape of the model antibiotic producer Streptomyces coelicolor A3(2). *Nat. Commun.* **7**, 11605 (2016).

34. Fan, B. *et al.* dRNA-Seq Reveals Genomewide TSSs and Noncoding RNAs of Plant Beneficial Rhizobacterium Bacillus amyloliquefaciens FZB42. *PLoS One* **10**, e0142002 (2015).

35. Decker, K. B. & Hinton, D. M. Transcription regulation at the core: similarities among

bacterial, archaeal, and eukaryotic RNA polymerases. *Annu. Rev. Microbiol.* **67**, 113–39 (2013).

36. Grünberger, F. *et al.* Next Generation DNA-Seq and Differential RNA-Seq Allow Re-annotation of the Pyrococcus furiosus DSM 3638 Genome and Provide Insights Into Archaeal Antisense Transcription. *Front. Microbiol.* **10**, 1603 (2019).

37. Babski, J. *et al.* Genome-wide identification of transcriptional start sites in the haloarchaeon Haloferax volcanii based on differential RNA-Seq (dRNA-Seq). *BMC Genomics* **17**, 629 (2016).

38. Jäger, D., Förstner, K. U., Sharma, C. M., Santangelo, T. J. & Reeve, J. N. Primary transcriptome map of the hyperthermophilic archaeon Thermococcus kodakarensis. *BMC Genomics* **15**, 684 (2014).

39. Lamberte, L. E. *et al.* Horizontally acquired AT-rich genes in Escherichia coli cause toxicity by sequestering RNA polymerase. *Nat. Microbiol.* **2**, 16249 (2017).

40. Chen, J. *et al.* Stepwise Promoter Melting by Bacterial RNA Polymerase. *Mol. Cell* **78**, 275-288.e6 (2020).

41. Warman, E. A., Singh, S. S., Gubieda, A. G. & Grainger, D. C. A non-canonical promoter element drives spurious transcription of horizontally acquired bacterial genes. *Nucleic Acids Res.* **48**, 4891–4901 (2020).

42. Miller, J. Experiments in Molecular Genetics. (1972).

43. Haycocks, J. R. J. & Grainger, D. C. Unusually situated binding sites for bacterial transcription factors can have hidden functionality. *PLoS One* **11**, (2016).

44. Dugar, G. *et al.* High-Resolution Transcriptome Maps Reveal Strain-Specific Regulatory Features of Multiple Campylobacter jejuni Isolates. *PLoS Genet.* **9**, e1003495 (2013).

45. Singh, N. & Wade, J. T. Identification of regulatory RNA in bacterial genomes by genome-scale mapping of transcription start sites. *Methods Mol. Biol.* **1103**, 1–10 (2014).

46. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: A Sequence Logo Generator. doi:10.1101/gr.849004.

47. Haycocks, J. R. J. J. *et al.* The quorum sensing transcription factor AphA directly regulates natural competence in Vibrio cholerae. *PLoS Genet.* **15**, e1008362 (2019).

48. Kolb, A., Kotlarz, D., Kusano, S. & Ishihama, A. Selectivity of the Escherichia coli RNA polymerase Eσ38 for overlapping promoters and ability to support CRP activation. *Nucleic Acids Res.* **23**, 819–826 (1995).

49. Savery, N. J. *et al.* Transcription activation at class II CRP-dependent promoters: Identification of determinants in the C-terminal domain of the RNA polymerase α subunit. *EMBO J.* **17**, 3439–3447 (1998).

50. Stead, M. B. *et al.* RNAsnap^{TM}: A rapid, quantitative and inexpensive, method for isolating total RNA from bacteria. *Nucleic Acids Res.* **40**, e156 (2012).

51. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

52. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* **46**, W537–W544 (2018).

53. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).

54. Forrest, D., James, K., Yuzenkova, Y. & Zenkin, N. Single-peptide DNA-dependent RNA polymerase homologous to multi-subunit RNA polymerase. *Nat. Commun.* **8**, 15774 (2017).

55. Podell, S., Gaasterland, T. & Allen, E. E. A database of phylogenetically atypical genes in archaeal and bacterial genomes, identified using the DarkHorse algorithm. *BMC Bioinformatics* **9**, 419 (2008).

56. Kahramanoglou, C. *et al.* Direct and indirect effects of H-NS and Fis on global gene expression control in Escherichia coli. *Nucleic Acids Res.* **39**, 2073–2091 (2011).

57. Narayanan, A. *et al.* Cryo-EM structure of Escherichia coli σ 70 RNA polymerase and promoter DNA complex revealed a role of σ non-conserved region during the open complex formation. *J. Biol. Chem.* **293**, 7367–7375 (2018).

ACKNOWLEDGEMENTS

FUNDING: This work was funded by a Leverhulme Trust project grant (RPG-2018-198) and Wellcome Trust Investigator Award (212193/Z/18/Z) to DCG. **AUTHOR CONTRIBUTIONS:** The work was conceived and supervised by D.C.G. The *B. subtilis* TSS mapping and, *in vitro* analysis of different TSS spacings, was done by D.F. Analysis of the *V. cholerae* VC1303-VC1304 regulatory region was done by T.G. and J.R.J.H. All other experimental work was done by E.A.W. Computational analysis was done by D.C.G., J.T.W. and D.F. All authors contributed to data analysis and interpretation. The manuscript was written by D.C.G. with input from all authors.

COMPETING INTERESTS STSTEMENT

The authors declare that there are no competing interests.

Figure 1

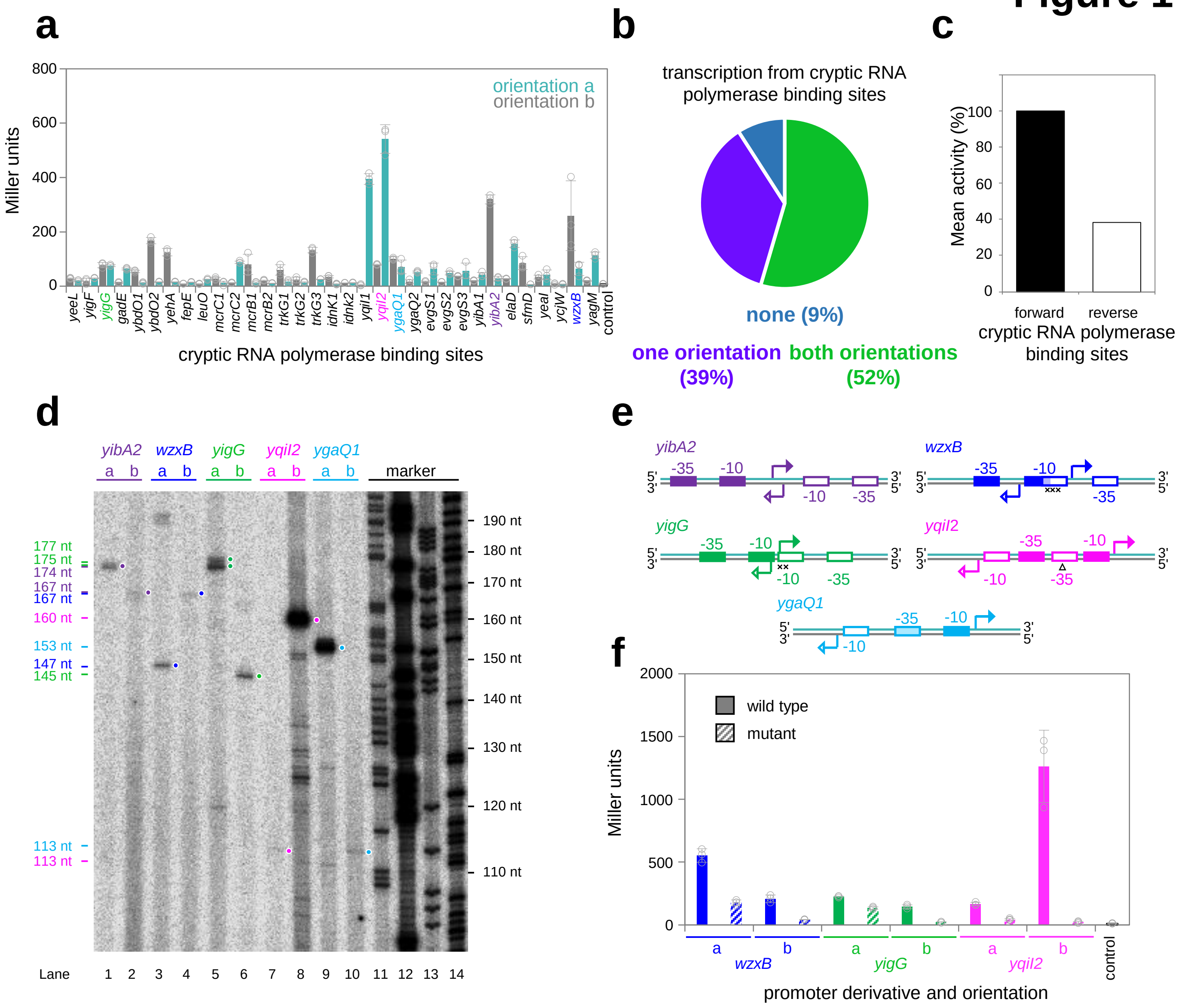

a
800
600
400
200
0
Miller units
orientation a
orientation b
yeeL
yigF
yigG
gadE
ybdO1
ybdO2
yehA
fepE
leuO
mcrC1
mcrC2
mcrB1
mcrB2
trkG1
trkG2
trkG3
idnK1
idnK2
yqiI1
yqiI2
ygaQ1
ygaQ2
evgS1
evgS2
evgS3
yibA1
yibA2
elaD
sfmD
yeaI
ycjW
wzxB
yagM
control
cryptic RNA polymerase binding sites
b
transcription from cryptic RNA polymerase binding sites
none (9%)
one orientation (39%)
both orientations (52%)
c
100
80
60
40
20
0
Mean activity (%)
forward
reverse
cryptic RNA polymerase binding sites
d
yibA2
wzxB
yigG
yqiI2
ygaQ1
a
b
marker
177 nt
175 nt
174 nt
167 nt
167 nt
160 nt
153 nt
147 nt
145 nt
113 nt
113 nt
190 nt
180 nt
170 nt
160 nt
150 nt
140 nt
130 nt
120 nt
110 nt
Lane
1
2
3
4
5
6
7
8
9
10
11
12
13
14
e
yibA2
wzxB
yigG
yqiI2
ygaQ1
5'
3'
-35
-10
f
2000
1500
1000
500
0
Miller units
wild type
mutant
a
b
wzxB
yigG
yqiI2
control
promoter derivative and orientation

Figure 2

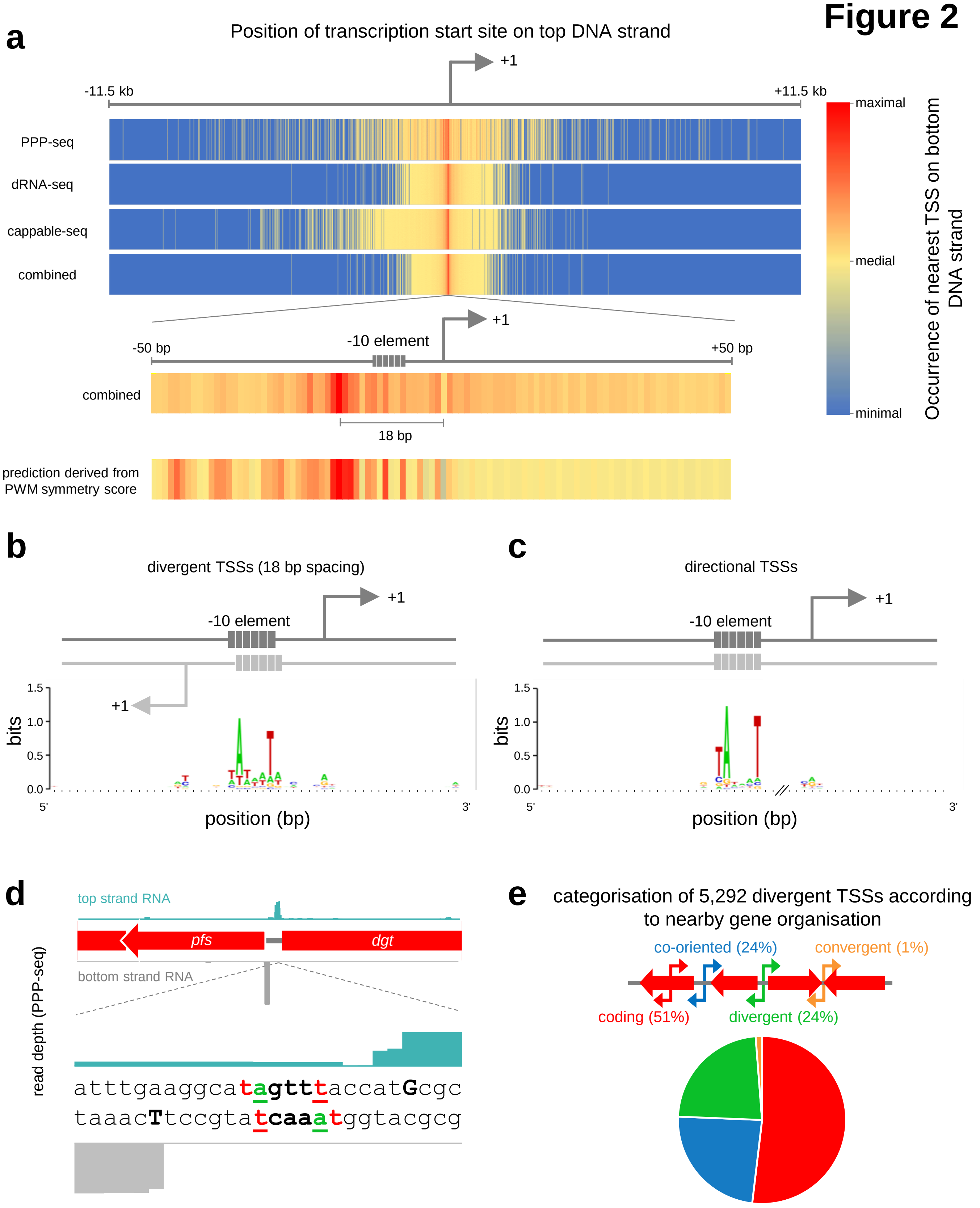

a
Position of transcription start site on top DNA strand
+1
-11.5 kb
+11.5 kb
PPP-seq
dRNA-seq
cappable-seq
combined
maximal
medial
minimal
Occurrence of nearest TSS on bottom DNA strand
-50 bp
-10 element
+1
+50 bp
combined
18 bp
prediction derived from PWM symmetry score
b
divergent TSSs (18 bp spacing)
+1
-10 element
+1
1.5
1.0
0.5
0.0
bits
5'
3'
position (bp)
c
directional TSSs
-10 element
+1
1.5
1.0
0.5
0.0
bits
5'
3'
position (bp)
d
top strand RNA
pfs
dgt
bottom strand RNA
read depth (PPP-seq)
atttgaaggcatagtttaccatGcgc
taaacTtccgtatcaaatggtacgcg
e
categorisation of 5,292 divergent TSSs according to nearby gene organisation
co-oriented (24%)
convergent (1%)
coding (51%)
divergent (24%)

Figure 3

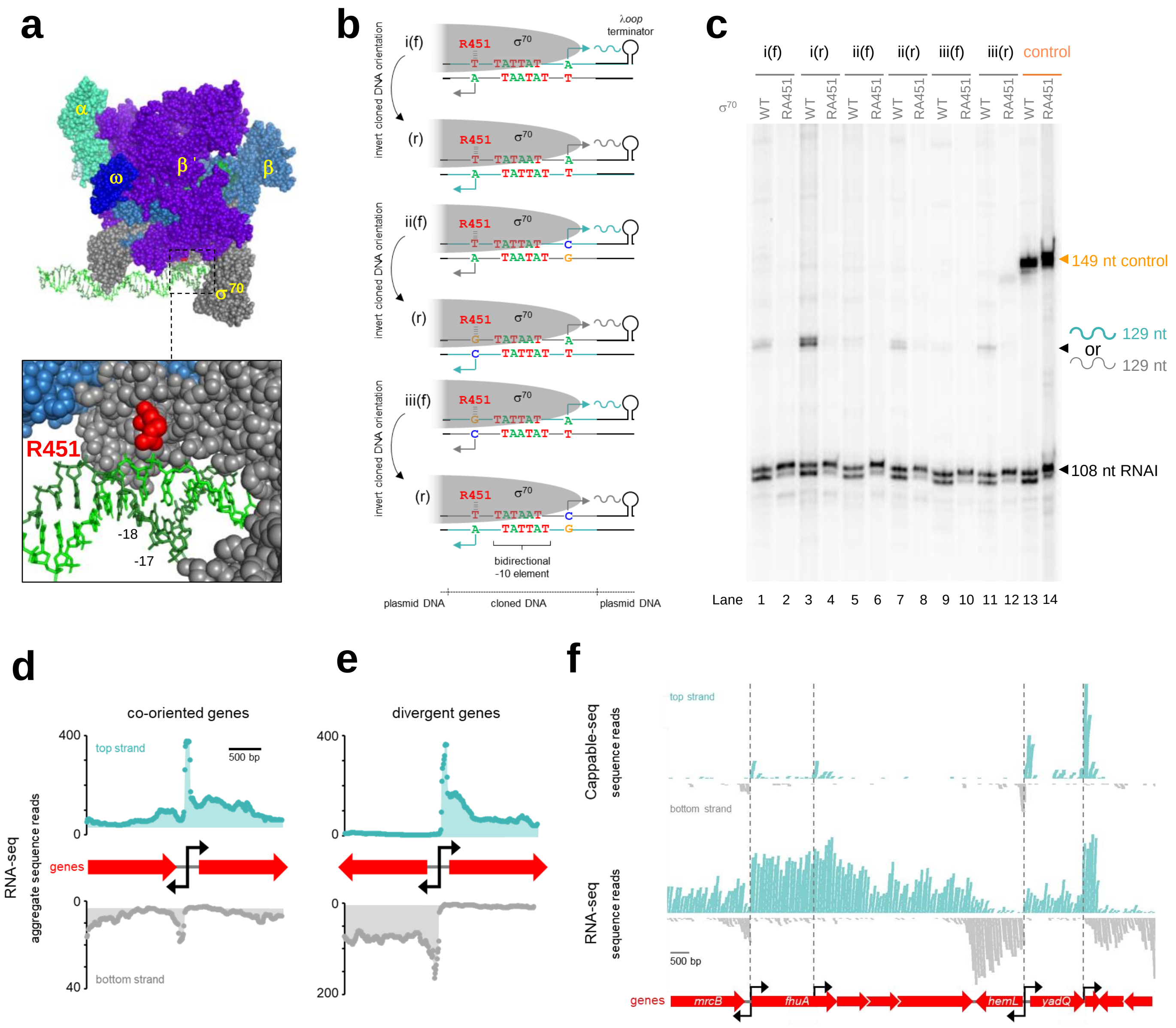

a
α
ω
β'
β
σ70
R451
-18
-17
b
i(f)
(r)
ii(f)
(r)
iii(f)
(r)
invert cloned DNA orientation
invert cloned DNA orientation
invert cloned DNA orientation
λoop terminator
R451
σ70
T—TATTAT—A
A—TAATAT—T
T—TATAAT—A
A—TATTAT—T
T—TATTAT—C
A—TAATAT—G
G—TATAAT—A
C—TATTAT—T
G—TATTAT—A
C—TAATAT—T
T—TATAAT—C
A—TATTAT—G
bidirectional -10 element
plasmid DNA
cloned DNA
plasmid DNA
c
i(f)
i(r)
ii(f)
ii(r)
iii(f)
iii(r)
control
σ70
WT
RA451
149 nt control
129 nt
or
129 nt
108 nt RNAI
Lane 1 2 3 4 5 6 7 8 9 10 11 12 13 14
d
co-oriented genes
RNA-seq
aggregate sequence reads
400
0
0
40
top strand
500 bp
genes
bottom strand
e
divergent genes
400
0
0
200
f
Cappable-seq
sequence reads
RNA-seq
sequence reads
top strand
bottom strand
500 bp
genes
mrcB
fhuA
hemL
yadQ

Figure 4

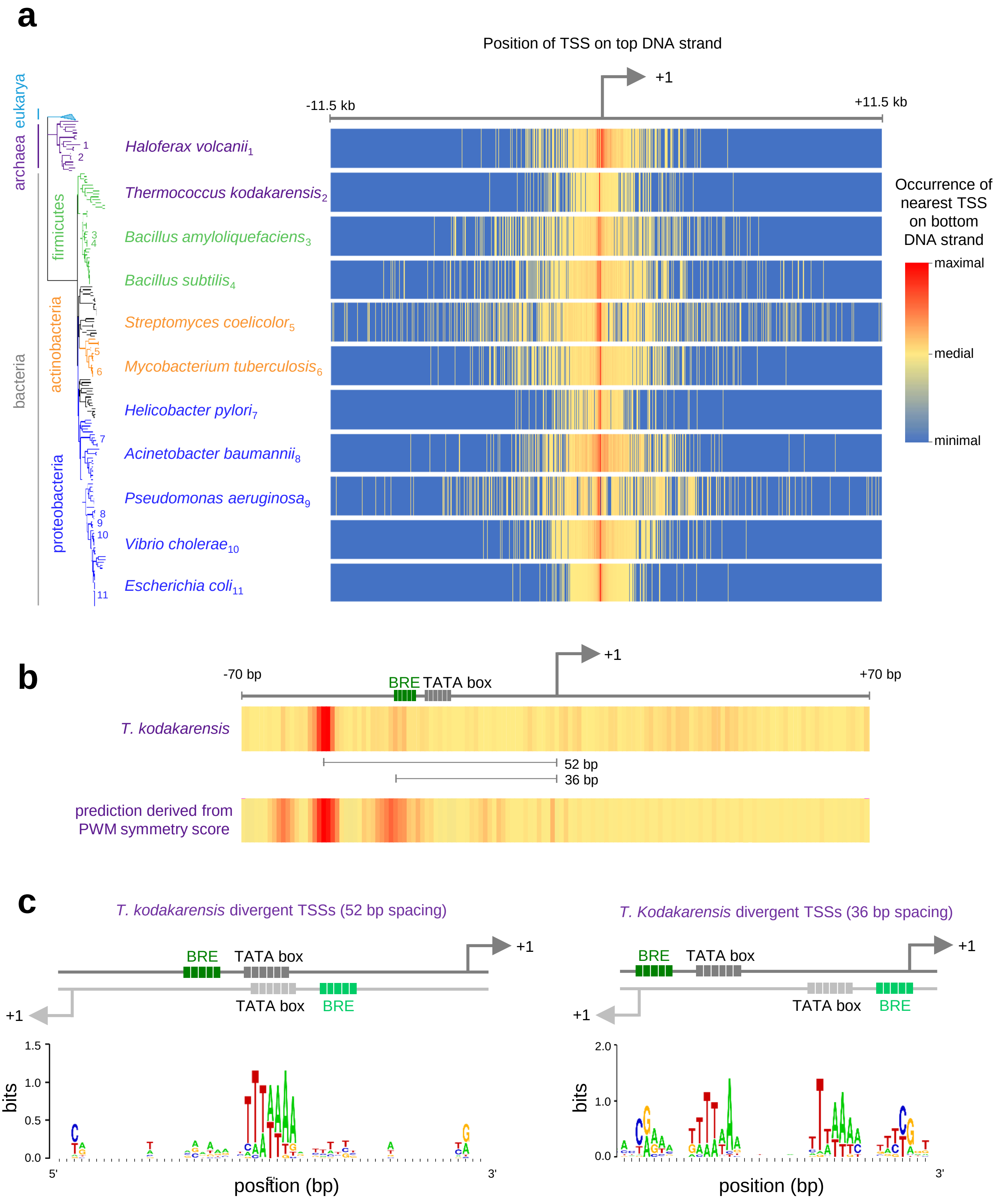

a
Position of TSS on top DNA strand
+1
-11.5 kb
+11.5 kb
eukarya
archaea
bacteria
firmicutes
actinobacteria
proteobacteria
Haloferax volcanii1
Thermococcus kodakarensis2
Bacillus amyloliquefaciens3
Bacillus subtilis4
Streptomyces coelicolor5
Mycobacterium tuberculosis6
Helicobacter pylori7
Acinetobacter baumannii8
Pseudomonas aeruginosa9
Vibrio cholerae10
Escherichia coli11
Occurrence of nearest TSS on bottom DNA strand
maximal
medial
minimal
b
-70 bp
+70 bp
BRE
TATA box
+1
T. kodakarensis
52 bp
36 bp
prediction derived from PWM symmetry score
c
T. kodakarensis divergent TSSs (52 bp spacing)
T. Kodakarensis divergent TSSs (36 bp spacing)
BRE
TATA box
+1
TATA box
BRE
+1
bits
1.5
1.0
0.5
0.0
5'
3'
position (bp)
2.0
1.0
0.0
3'
position (bp)

Figure 5

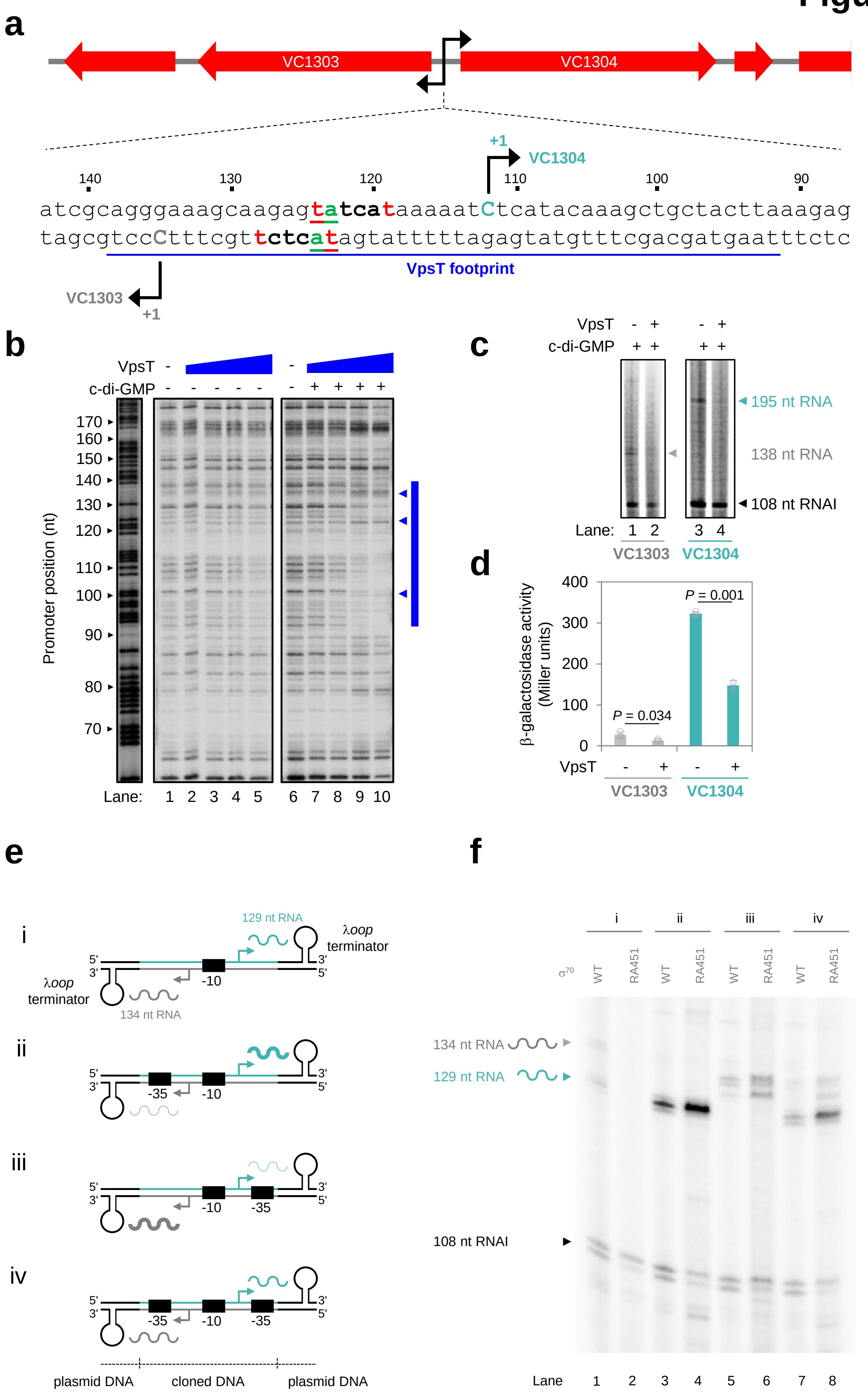
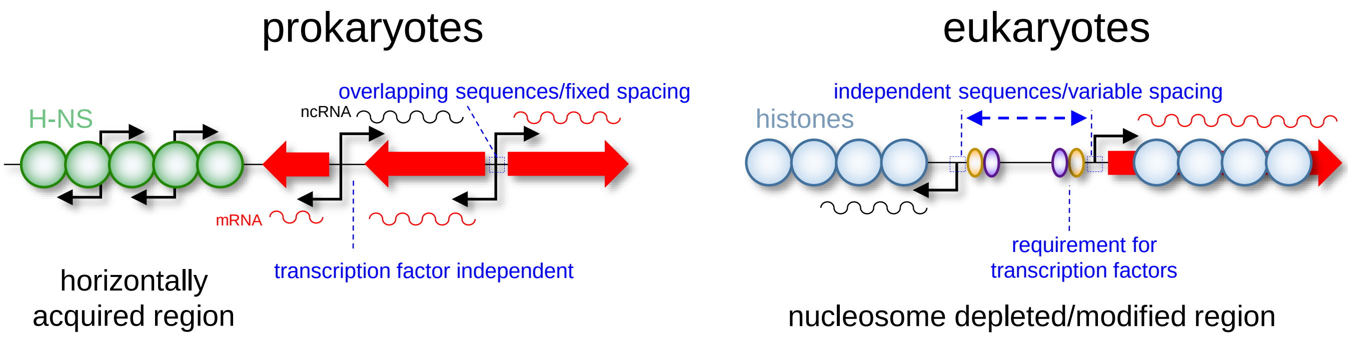

a
VC1303
VC1304
+1
VC1304
140
130
120
110
100
90
atcgcagggaaagcaagagtatcataaaaatCtcatacaaagctgctacttaaagag
tagcgtccCtttcgttctcatagtatttttagagtatgtttcgacgatgaatttctc
VpsT footprint
VC1303
+1
b
VpsT
c-di-GMP
- - - - -
- + + + +
Promoter position (nt)
170
160
150
140
130
120
110
100
90
80
70
Lane: 1 2 3 4 5 6 7 8 9 10
c
VpsT - + - +
c-di-GMP + + + +
195 nt RNA
138 nt RNA
108 nt RNAI
Lane: 1 2 3 4
VC1303
VC1304
d
β-galactosidase activity (Miller units)
400
300
200
100
0
P = 0.034
P = 0.001
VpsT - + - +
VC1303
VC1304
e
i
ii
iii
iv
129 nt RNA
λoop terminator
λoop terminator
134 nt RNA
5'
3'
-35
-10
plasmid DNA
cloned DNA
plasmid DNA
f
i
ii
iii
iv
σ70
WT
RA451
134 nt RNA
129 nt RNA
108 nt RNAI
Lane 1 2 3 4 5 6 7 8

Figure 6


prokaryotes
H-NS
ncRNA
overlapping sequences/fixed spacing
mRNA
horizontally
acquired region
transcription factor independent
eukaryotes
independent sequences/variable spacing
histones
requirement for
transcription factors
nucleosome depleted/modified region