A global analysis of transcription reveals two modes of Spt4/5 recruitment to archaeal RNA polymerase

Katherine Smollett[1], Fabian Blombach[1], Robert Reichelt[2], Michael Thomm[2] and Finn Werner[1*]

[1]University College London, Institute for Structural and Molecular Biology, Gower Street, London, WC1E 6BT, UK

[2]Institute of Microbiology and Archaea Center, Universität Regensburg, 93053 Regensburg, Germany

*Correspondence: f.werner@ucl.ac.uk

Abstract

The archaeal transcription apparatus is closely related to the eukaryotic RNA polymerase (RNAP) II system, while archaeal genomes are more similar to bacteria with densely packed genes organised in operons. This makes understanding transcription in archaea vital, both in terms of molecular mechanisms and evolution. Very little is known about how archaeal cells orchestrate transcription on a systems level. We have characterised the genome-wide occupancy of the *Methanocaldococcus jannaschii* transcription machinery and its transcriptome. Our data reveal how the TATA and BRE promoter elements facilitate the recruitment of the essential initiation factors TBP and TFB, respectively, which in turn are responsible for the loading of RNAP into the transcription units. The occupancy of RNAP and Spt4/5 strongly correlate with each other and with the RNA levels. Our results show that Spt4/5 is a general elongation factor in archaea since its presence on all genes matches RNAP. Spt4/5 is recruited proximal to the TSS on the majority of transcription units, while on a subset of genes including rRNA and CRISPR loci, Spt4/5 is recruited to the transcription elongation complex during early elongation within 500 bp of the TSS, and akin to its bacterial homolog NusG.

Keywords: transcription, archaea, promoter, TBP, TFB, RNAP, Spt4/5

Introduction

Transcription is a fundamental process in biology and RNA polymerases (RNAP) are closely related in all domains of life[1]. The archaeal and eukaryotic systems are near-identical in terms of RNAP subunit composition and architecture, regarding transcription initiation, elongation factors and the molecular mechanisms that govern their activity[2]. The universally conserved core of RNAP resembles a crab claw-like structure made of the large catalytic subunits Rpo1/2 and the assembly platform including Rpo3/11. The archaeal RNAP shares five to six additional subunits with eukaryotic RNAPII that are absent in bacterial RNAP[3]. This includes the Rpo4/7 stalk module that protrudes from the core enzyme, binds to the nascent RNA and modulates transcription processivity and termination[4]. Archaeal transcription has been studied extensively in vitro, but relatively little is known about the genome-wide distribution of RNAP and basal transcription factors, and how this correlates with promoter elements and transcription output. A limited number of archaeal promoters have been functionally characterised, and seem to rely TATA boxes, B-recognition- (BRE) and Initiator elements (Inr)[5,6]. The former two are binding sites for the two basal transcription factors TBP and TFB, respectively[2]. Both are strictly required for promoter-directed transcription in vitro[7], and homologous to eukaryotic TBP and TFIIB with identical functions but faster dynamics in terms of promoter binding[8]. The third basal transcription factor TFE is homologous to TFIIE, it enhances the stability of the transcription preinitiation complex (PIC) by catalysing the isomerisation of closed to open complex, during which the DNA strands are separated and the template strand is loaded into the active site of RNAP[9,10]. The elongation factor Spt4/5, NusG in bacteria, is the only RNAP-associated factor that is conserved throughout the three domains of life. Spt4/5 enhances transcription processivity and possibly functions during promoter escape[11]. Interestingly, in vitro experiments revealed that Spt4/5 and NusG are denied access to the preinitiation complex (PIC) by TFE and σ^{70}, respectively[10,12]. Chromatin immunoprecipitation (ChIP) experiments show that yeast Spt4/5 is recruited to RNAP proximal to the promoter, suggesting a role in transition from initiation to elongation[13], whereas *E. coli* NusG is recruited to RNAP during elongation in a stochastic fashion[14].

We applied Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) in order to characterise the whole genome distribution of *Methanocaldococcus jannaschii* (Mja) RNAP and initiation factors TBP and TFB, and to examine the recruitment patterns of Spt4/5 in archaea. To orientate the transcription machinery within the genome, we mapped and analysed global TSSs and steady-state RNA levels. We identified positive correlations between: BRE/TATA motif strength; binding of TBP and TFB to the promoter; occupancy of RNAP and Spt4/5 within the gene; and RNA levels. The elongation factor Spt4/5 showed two different modes of recruitment: early, promoter-proximal recruitment to RNAP similar to yeast Spt4/5; and a later recruitment during early elongation on rRNA and CRISPR loci more akin to bacterial NusG.

Results

Organisation of the Mja transcriptome. The workflow of the RNA-seq analysis is illustrated in Supplementary Figure 1a. To characterise the Mja transcriptome we first mapped the genome-wide transcription start sites (TSSs) using a terminator exonuclease (TEX) RNA-seq approach. We mapped 1508 TSSs (see supplementary materials) and used our TSS map to annotate 976 transcription units (TUs) that we defined as the sequence spanning from the primary TSS to the stop codon (on mRNA genes) or the annotated 3' end (on noncoding RNA genes) of the last cistron. A further 138 TUs were predicted based on gene orientation but were not associated with a TSS. We identified several novel genes encoding ORFs and ncRNAs that are listed in Supplementary tables 3 and 4. Mja TUs are organised into a combination of single- and multicistronic operons (Supplementary Fig. 2e). The majority of protein-encoding genes encode long untranslated leader regions (5'-UTR) with only 16 mRNAs (1.9%) being defined as leaderless (<5 nt, Fig. 1a). Within the 5' UTRs we identified ribosome binding sites (RBS) in 54% of mRNA genes (Fig. 1a). To determine the global steady-state RNA levels, we next calculated RPKM (Reads Per Kilobase of transcript per Million mapped reads) values for each TU. Using a cut off value of RPKM > 1, we defined 63% of the TUs as transcriptionally active (adjusted P value < 0.05, Supplementary text and Supplementary Table 3). The two ribosomal rRNA operons had the highest RPKM values and account for 80% of all mapped reads. Several small ncRNA genes including tRNAs were detected at low levels but may be misrepresented due to loss during size selection of library preparation. We could detect antisense transcription in Mja (Fig. 1b), however, the majority of antisense transcripts were not associated with a TSS, possibly due to their rapid degradation. We identified twelve antisense TUs with assigned TSS, including the Mja histone A3 gene (Fig. 1c, Supplementary Table 4). Both sense and antisense A3 transcripts were highly abundant, hinting at a possible regulation of A3 expression by antisense transcription. Northern blotting confirmed the presence of both sense and antisense A3 transcripts covering the histone A3 ORF (Supplementary Fig. 2f).

Promoter sequence elements and start site selection. Alignment of DNA sequences surrounding the TSSs identified two regions with a sequence bias, corresponding to the BRE/TATA elements, and the initially melted region (IMR) that includes the initiator (Inr) surrounding the TSS (Fig. 1d). Sequence motif analysis of these DNA sequences revealed a global BRE/TATA consensus (Fig. 1e). These elements could be identified upstream of 76% of TSSs using a stringent motif confidence score (motif P value < 10^{-3}, Supplementary Fig. 3a), including all primary TSSs of TUs defined as transcriptionally active. BRE/TATA motifs are centred on register +24 relative the TSS; this distance is conserved from archaea to metazoans[15] (Fig. 1f). During open complex formation the two DNA strands of the initially melted region (IMR) of the promoter from -12 to +2 are separated[9,16-18]. Alignments show that this region is enriched in A and T residues (80 ± 12% AT, genome average 69 % AT, Fig. 1g). The AT content of the IMR does not correlate with RNA levels (Supplementary Fig. 3b). The Inr element formed by the bases surrounding the TSS showed a strong bias for the sequence T(A/G) at position -1/+1 (Fig. 1d) but, similar to the IMR, did not correlate with RNA levels (Supplementary Fig. 3c). Examining the di-nucleotide frequency within this region revealed that TA and TG are not only highly enriched at position -1/+1 (combined > 60%, compared to the genome average of 15%), but also strongly disfavored at the neighboring positions (-2/-1 and +1/+2, Fig. 1h). The conservation of the T(A/G) motif is independent of the distance between the TATA box and the TSS (Supplementary Fig. 3d). Since these results suggest that the Inr dictates TSS selection, we analysed the TSS specificity on promoters with and without Inr motif. Promoters with an Inr sequence T(A/G) showed up to four-fold lower levels of transcription initiation at neighboring positions compared to promoters without the T(A/G) motif (Fig. 1i). In summary, while the BRE/TATA motifs facilitate the transcription preinitiation complex assembly, the Inr fine-tunes TSS selection. A comparison with other archaeal promoters[19-24] (Supplementary Fig. 4) reveals that the TATA consensus is largely conserved across the archaea, while the significance of IMR and Inr are subject to variation[25].

TBP and TFB binding to the Mja BRE/TATA motifs. We determined the global occupancy of the essential initiation factors TBP and TFB by chromatin immunoprecipitation using polyclonal antibodies raised against recombinant proteins followed by high-throughput sequencing (ChIP-seq). The workflow and detailed methods are described in supplementary materials. Figure 2a-e show the ChIP-seq profiles of four representative promoters, ranging from promoters that show a distinct and defined increased TBP and TFB occupancy centred on the BRE/TATA motifs (*mcrB* and *ftr,* panel b and c), those that display broader profiles, but are distinct from the mock control (*sla*, d), to promoters that do not show any increased occupancy at all (*rrnA*, e). Averaging the TBP/TFB occupancy profiles centred on the TSS of the top 25% expressed mRNA TUs (by RPKM) shows distinct TBP and TFB peaks (Fig. 2f). The apex of both peaks concurs with the location of the BRE/TATA motifs, which confirms the validity of our TBP/TFB profiling analysis (Fig. 1f). The profile of the mock IP control demonstrates that, while the mock shows a slight increase in signal, both TBP and TFB signals are above the background (Fig. 2f). In order to validate our results we compared our data to a subset of experimentally characterised promoters. 19 tRNA and 12 mRNA Mja promoters have been analysed quantitatively in vitro with respect to the formation of DNA-TBP-TFB complexes using EMSAs[6]. There is a strong correlation between the published in vitro binding data and the in vivo occupancy across their promoter regions (TSS ± 250 bp, TBP R = 0.7, P value = 1.1×10^{-5}, TFB R = 0.61, P value = 2.6×10^{-4} Supplementary Fig. 5c), which also implies that in vitro EMSAs are a good indicator for the binding of TBP and TFB to promoters in vivo. In order to relate strength of the TBP/TFB binding to the sequence of the BRE/TATA motifs, we compared the confidence score (P value) of the BRE/TATA motif of each promoter to the TBP and TFB ChIP signal (Fig. 2g). The BRE/TATA score showed a weak but significant correlation to the TBP/TFB occupancy (TBP R = -0.23, P value = 6×10^{-8}, TFB R = -0.30, P value < 10^{-10}, mock R = -0.08, P value = 0.03), but only a very weak correlation to TU RNA levels (Fig. 2h, TBP R = 0.15, P value

= 1 x 10^{-4}, TFB R = 0.15, P value 8.1 x 10^{-5}, no correlation with mock, P value > 0.05).

RNAP occupancy correlates with RNA levels. We characterised the global occupancy of RNAP with two polyclonal antibodies directed against two distinct RNAP subcomplexes. The pair-wise genome-wide correlation between occupancy of Rpo4/7 stalk and Rpo3/11 assembly platform subunits was calculated using 250 bp windows with a 50 bp overlap. The Rpo4/7 and Rpo3/11 signals correlate very strongly with each other (R = 0.95, P value < 10^{-10}, Fig. 3a). In order to visualise the RNAP occupancy within TUs we plotted the ChIP-seq profile as occupancy per nucleotide across the genome. The RNAP ChIP seq profiles of individual loci emphasise very diverse profiles on different genes (Fig. 3b-e, and figure 5), e.g. while occupancy is high on the *sla* and *mcr* TUs (Figure 3b,d), it is low on the *tuf* and *rpo* operons (Figure 3c,e). A metadata analysis averaging the RNAP occupancy centred on the TSS reveals that the Rpo4/7 signal appears approximately 100 bp upstream of the Rpo3/11 signal (Fig. 3f). Promoter-bound TBP and TFB are strictly required for the recruitment and subsequent loading of RNAP into the TU in vitro. In good agreement, the occupancy of TBP and TFB at the promoter correlated with RNAP occupancy within the TU (Fig. 3g, Rpo4/7 compared to TBP R = 0.37, P value < 10^{-10}, Rpo4/7 to TFB R = 0.3, P value < 10^{-10}, mock R = 0.1, P value = 0.02). Finally, the RNAP occupancy within TUs correlated moderately well with RNA levels (Fig. 3h Rpo4/7 R = 0.45, P value < 10^{-10}, Rpo3/11 R = 0.48, P value < 10^{-10}, mock R = -0.15, P value 3.4 x 10^{-4}).

In vitro preinitiation complex assembly. Surprisingly the two Mja rRNA promoters (*rrnA* and *rrnB*) have no identifiable BRE/TATA motifs and do not show strong TBP/TFB ChIP signal (Fig. 2a,e). This suggests that they are weak promoters which is in stark contrast to the high RNAP occupancy and RNA levels. In order to probe the strength of Mja *rrn* promoters in vitro, we monitored PIC formation on the *rrnA* promoter using EMSA, and promoter activity using transcription assays. For comparison, we included a representative Mja mRNA promoter (*rpl12*), which is associated with high RNAP occupancy and RNA level, an Mja CRISPR promoter, which has high RNAP occupancy and the well-characterised viral SSV T6 promoter (Fig. 4a)[7,16,26,27].

The SSV T6 and CRISPR promoters recruit RNAP in a TBP/TFB-dependent fashion, and the addition of TFE stimulated the PIC in EMSA experiments (Fig. 4b). The *rpl12* promoter, that has a similar BRE/TATA consensus but lower IMR AT% than the CRISPR promoter formed a weak PIC in the absence of TFE. In contrast, the *rrnA* promoter was not able to form a stable PIC. Heteroduplex promoter variants include a 4 bp noncomplementary region (-3 to +1), mimic the open complex and enhance PIC stability[16,26]. These variants enabled PIC formation at all four promoters, including *rrnA* (Fig. 4c). Introducing mutations into the TATA sequence abolished or dramatically reduced PIC formation on all promoters (Supplementary Fig. 6a-b). We used promoter-directed in vitro transcription experiments to complement the promoter-binding experiments. The results from both assays mirrored each other; while the SSV T6, *rpl12* and CRISPR promoters resulted in large amounts of transcripts with the correct size, the *rrnA* promoter was inactive

(Fig. 3d). In conclusion, in contrast to the in vivo analysis, the in vitro transcription experiments show a direct link between promoter motifs, the recruitment of stable PIC and promoter strength.

Spt4/5 is a general elongation factor with two distinct recruitment modes. We carried out a ChIP-seq analysis to characterise the global occupancy of the transcription elongation factor Spt4/5. The pair-wise correlation between genome-wide occupancies of Spt4/5 and RNAP is very strong (Fig. 5a, Rpo3/11 R = 0.96, P value < 10^{-10}; Rpo4/7 R = 0.95, P value < 10^{-10}, mock R = 0.035, P value < 10^{-10}). Furthermore, a comparison of RNAP and Spt4/5 ChIP-seq profiles on individual TUs (by plotting their per nucleotide occupancy) demonstrates that Spt4/5 closely mirrors the undulating pattern of RNAP occupancy that likely reflects pausing and varying transcription processivity (Fig. 5b,c). This behavior suggests that Spt4/5 stably associates with the transcription elongation complex (TEC) in vivo. In order to detect any potential heterogeneity in the genome occupancy of RNAP and Spt4/5 we identified genome locations characterised by a lower Spt4/5:RNAP occupancy ratio (red dots in Fig. 5a). The overlapping 250 bp windows were merged to identify 23 separate genome regions with significantly lower Spt4/5 than RNAP occupancy (adjusted P value < 0.05, Supplementary Table 5). These regions included 18 of the 20 CRISPR loci, both ribosomal rRNA operons (*rrnA* and *rrnB*), two annotated small non-coding RNA genes, and *mj0496* (uncharacterised ORF). Closer scrutiny of these regions revealed that the lower Spt4/5:RNAP occupancy ratio is restricted to the promoter-proximal region of the gene, with the Spt4/5 profile matching that of RNAP from ~500 bp downstream of the promoter onwards (Fig. 5d,e, Supplemental Table 5). The bacterial Spt5 homologue NusG aids the coupling of transcription and translation by interacting with the RNAP and the ribosome[28,29]. Similarly transcription and translation are coupled in archaea[30]. We tested whether the recruitment of Spt4/5 to TECs on protein-encoding genes was influenced by the recruitment of the ribosome to the RBS by analysing Spt4/5 occupancy on mRNA genes with long 5' UTRs. The 5' UTR of the *korB* gene is 162 bp long, but Spt4/5 is recruited symmetrically with RNAP close to the TSS and not further downstream at the RBS (Fig. 5f). To explore this globally we subtracted the RNAP- from the Spt4/5 occupancy at each mRNA promoter and plotted the value against the length of the 5'-UTR. If Spt4/5 recruitment was aided by the ribosome we would expect the difference in occupancy to increase with 5'-UTR length, however no difference was observed (Fig. 5g). In conclusion, Spt4/5 follows two modes of recruitment (Fig. 6), in proximity of the promoter on the majority of TUs, and several hundred bp downstream of the TSS on a subset of genes.

Discussion

We present the first comprehensive genome-wide analysis of transcription in archaea by characterising the (i) occupancy of RNAP and basal transcription factors, (ii) the transcriptome including a TSS map, and (iii) a promoter motif analysis all in the same organism.

We identified 1508 TSSs in *Mja*, and could account for 88% of TSS of the 1114 predicted TU. The TSS analysis reveals that *Mja* mRNAs have long 5' UTRs indicative of extensive riboregulation by sRNA and riboswitches. This pattern is similar to other methanogens including *M. mazei*, *M. psychrophilus, T. kodakarensis* and *P. furiosus*, and different from Sulfolobales and halophilic archaea that are characterised by leaderless mRNAs[19-23,31-34]. The assembly of the PIC in vitro is strictly dependent on the binding of TBP and TFB to TATA and BRE motifs of archaeal promoters, respectively. Our in vivo analysis reveals the prevalence of BRE/TATA motifs, suggesting that they are the dominant promoter elements in archaea. This is in contrast to eukaryotes where conventional TATA motifs are absent at the majority of promoters[35]. We also reveal the importance of downstream sequences including the IMR and the 3-bp Inr element that increases the accuracy of TSS selection, while not correlating with the RNA levels. Thus far the role of the archaeal Inr has only been studied in vitro, mainly with mutated variants of the viral SSV1 T6 model promoter[36,37]. Our systems data reveal that the Mja Inr has a bias for T(A/G) at registers -1/+1. This preference for pyrimidine and purine nucleotides is a universally conserved promoter feature, which reflects the high degree of conservation between the RNAP active site architectures in the three domains of life[15,38,39]. The elevated AT content of the IMR favors local

DNA melting, and experimental evidence shows that the IMR sequence affects promoter strength at individual promoters in vitro[9,25]. However, on a global level the AT content of the Mja promoter IMR does not correlate with RNA levels, and it is thus unlikely that the IMR's AT content alone limits promoter strength in vivo.

Having explored the sequence characteristics of archaeal promoters we characterised the association of RNAP, TBP, TFB and the elongation factor Spt4/5 with the genome. The averaged occupancy profiles of highly expressed genes illustrate the early stages of the archaeal transcription cycle with the step-wise assembly of the PIC, RNAP and Spt4/5 recruitment, and promoter escape (Fig. 6). The individual RNAP profiles in different TUs are very diverse, including regions of high and low occupancy proximal to the promoter motifs and within TUs, which likely reflects variations in promoter recruitment, efficiency of escape, processivity and pausing[40]. It has been proposed that the yeast RNAPII RPB4/7 stalk reversibly associated with the RNAP core. Our ChIP-seq results demonstrate that both Rpo4/7 and Rpo3/11 are colocalised across the genome suggesting that the stalk remains associated with the RNAP core as it progresses through the transcription cycle. The fact that Rpo4/7 is slightly off-set upstream from Rpo3/11 signals at TSSs is likely due to epitope occlusion of the latter in the PIC[11,16]. The molecular mechanisms of archaeal Spt4/5 have been characterised in some detail in vitro[10,17,41]. Our ChIP-seq results demonstrate that Spt4/5 associates with elongating RNAPs throughout the genome behaving like an 'honorary' RNAP subunit on all genes, protein-encoding as well as non-coding RNA genes, meaning that Spt4/5 fulfills the criteria of a general elongation factor.

By comparing the ChIP-seq profiles of RNAP and Spt4/5 two distinct modes of Spt4/5 recruitment become apparent, either (1) proximal to promoter and just off-set from the TSS or (2) further downstream within the first 500 bp of the TU (Fig. 6). All multisubunit RNAP face a similar mechanical engineering challenge: a network of interactions between promoter-bound initiation factors (TBP/TFB/TFE) and RNAP is crucial to enable efficient recruitment of RNAP during early initiation, however, these interactions need to be disrupted to allow RNAP to escape from the promoter[11]. As Spt4/5 and the initiation factor TFE bind to the RNAP clamp in a mutually exclusive manner in vitro[10,11], Spt4/5 recruitment proximal to the TSS could assist promoter escape of RNAP by displacing TFE. Our attempts to ChIP TFE were unsuccessful despite the use of several independent antibody preparations, therefore we could not directly characterise the swapping of Spt4/5 and TFE in vivo. However, Spt4/5 mode (1) does support the recruitment during promoter escape - and not during elongation. ChIP analyses from eukaryotic systems are in agreement with promoter-proximal recruitment of Spt4/5[13] and the swapping with TFIIE proximal to the promoter[42,43]. Our results show notable exceptions to mode (1); in mode (2) the Spt4/5 occupancy does not match RNAP occupancy until several hundred bp downstream of the TSS; these include the two ribosomal RNA operons that account for 80% of the total RNA in the cell, and the abundant CRISPR loci. In contrast to Mja Spt4/5, *E. coli* NusG is recruited during elongation at most TUs, but proximal to rRNA promoters due to the assembly of antitermination complexes including NusA, B and E, other ribosomal proteins, some of which are conserved in archaea[14,44]. rRNA operons and CRISPR regions differ from coding genes as templates for transcription in several regards such as absence of coupled translation, strong secondary-structure content, co-transcriptional processing and ribosome biogenesis. Unidentified rRNA and CRISPR promoter-specific transcription activators could enhance RNAP recruitment, stabilise the PIC, or interact with the RNAP clamp and possibly enhance promoter escape. This notion is supported by our finding that Mja rRNA promoters have a surprisingly poor BRE/TATA motifs and have very low activity in vitro, in apparent conflict with the high steady-state levels of rRNA and RNAP occupancy on rRNA operons in vivo. The *Sulfolobus solfataricus* and *Pyrococcus furiosus* rRNA promoters have defined BRE/TATA motifs, and are very strong in vitro[9,27,45], while bacterial rRNA promoters tend to form unstable PICs, making them more amenable to regulation[46].

A quantitative analysis of the transcriptome reveals that 700 of the 1114 TU (63 %) contain detectable transcript, under optimal growth conditions used. We found only a weak correlation between BRE/TATA motif scores or TBP/TFB occupancy, and no correlation with RNA levels. Steady-state RNA levels do not take into account factors such as RNA stability, however as a good correlation was found between RNAP occupancy and RNA levels it seems a reasonable proxy for transcription output for most Mja genes. The lack of a strong correlation between promoter motifs and RNA levels illustrates the importance of additional factors such as the chromatin context as well as gene-specific regulators[47]. For example, TBP recruitment to the Mja *rb2* promoter TATA element is enhanced by the adjacent binding of the Ptr2 activator in vitro[48]. Based on the BRE/TATA score of the *rb2* promoter the relative TBP promoter occupancy can be predicted by linear regression as

0.14 Log_2(IP/input), while the observed value is much higher at 1.01, in line with a Ptr2-enhancement of TBP binding in vivo. A nascent elongating transcript (NET)-seq[49,50] approach would allow a direct determination of transcription output in vivo, and could provide insights into the manifold factors that regulate transcription within archaea in the future.

Methods

Culture conditions. Mja strain DSM 2661[51] were grown in large scale 100 l fermenters in a minimal media containing 0.3 mM K_2HPO_4, 0.4 mM KH_2PO_4, 3.6 mM KCl, 0.4 M NaCl, 10 mM $NaHCO_3$, 2.5 mM $CaCl_2$, 38 mM $MgCl_2$, 22 mM NH_4Cl, 31 μM $Fe(NH_4)_2(SO_4)_2$, 1 mM $C_6H_9NO_6$, 1.2 μM $MgSO_4$, 0.4 mM $CuSO_4$, 0.3 μM $MnSO_4$, 36 nM $FeSO_4$, 36 nM $CoSO_4$, 3.5 nM $ZnSO_4$, 4 nM $KAl(SO_4)_2$, 16 nM H_3BO_3, 42 μM Na_2SeO_4, 0.3 nM Na_2WO_4, 11 μM $NaMoO_4$, 44 μM $(NH_4)_2Ni(SO_4)_2$ and 2 mM Na_2S. Fermenters were mixed at 250 rpm and with H_2:CO_2 gas at 4:1 ratio at 85°C.

RNA preparation. RNA for sequencing was prepared from Mja cell pellets by Vertis Biotechnologies AG using the mirVana RNA isolation kit (Ambion). For TSS mapping total RNA was treated with Terminator exonuclease (TEX, Epicentre) to remove 5' mono-phosphate RNA. RNA for Northern blot analysis was prepared from Mja cell pellets using peqGOLD TriFast reagent (PeQlab) as per manufacturers instructions.

Chromatin immunoprecipitation. All antibodies used in ChIP experiments were rabbit antisera produced by Davids Biotechnologie GmbH using recombinant proteins prepared as in[52]. Specificity of antibodies was determined by Western blot. Mock control IPs used pre-immune sera. ChIP was performed on cultures of Mja that were grown to late log phase as measured by a cell count of ~ 1 x 10^8 cells/ml, and cross-linked by addition of 0.1% formaldehyde for 1 min before quenching with 12.5 mM glycine. Similar cross-linking conditions have been used successfully for the thermophile Pyrococcus[53,54]. Fixed cell pellets were washed three times in PBS and then resuspended in lysis buffer (0.1% sodium deoxycholate, 1 mM EDTA, 50 mM

HEPES pH 7.5, 140 mM NaCl, 1% Triton-X-100) plus 10% glycerol and protease inhibitor (cOmplete mini, EDTA-free protease inhibitor cocktail, Roche). DNA was sheared by sonication to approximately 300 bp fragments using a cup horn sonicator (Qsonica Q700) before mixing overnight at 4°C with the appropriate antibody prebound to Dynabeads M-280 sheep anti-rabbit IgG (Life Technologies). Beads were washed twice with lysis buffer, once with lysis buffer 500 (0.1% sodium deoxycholate, 1 mM EDTA, 50 mM HEPES pH 7.5, 500 mM NaCl, 1% Triton-X-100), once with LiCl buffer (0.5% sodium deoxycholate, 1 mM EDTA, 250 mM LiCl, 0.5% nonidet P-40, 10 mM Tris pH 8) and a final wash with TE buffer (10 mM Tris pH 7, 0.1 mM EDTA). DNA-protein complexes were eluted with ChIP elution buffer (10 mM EGTA, 1% SDS, 50 mM Tris pH 8) at 65°C for 10 min and remaining complexes eluted in TE (10 mM Tris pH 7, 0.1 mM EGTA) containing 0.67% SDS. Input samples were prepared by mixing sheared DNA-protein mix with TE (10 mM Tris pH 7, 0.1 mM EGTA) containing 1% SDS. Crosslinks were reversed and protein removed by treatment of samples with 0.05 mg ml^{-1} RNase A and 0.5 mg ml^{-1} proteinase K at 37°C for 2-4 hrs followed by overnight incubation at 65°C. DNA fragments were purified using MinElute columns (Qiagen) and quantified using the Qubit ds DNA HS kit (Life Technologies).

Illumina sequencing. For summary of steps see Supplementary Fig. 1. Library preparation and Illumina sequencing of total- and TEX treated RNA was performed by Vertis Biotechnologies. For the TEX treated samples RNA adapters were ligated to the 5' ends and 3' ends were poly(A) tailed before first-strand cDNA synthesis and PCR amplification. Resulting cDNA was fractionated by ultrasound and 5' ends selected and further amplified after ligation of TruSeq 3' end adapter primer (Illumina). For RNA-seq of total RNA samples were fragmented with ultrasound and first-strand cDNA synthesis was performed using randomised N6 primer before ligation of strand-specific TruSeq adapters (Illumina) to the 5' and 3' end of the cDNA and PCR amplification. cDNA samples were pooled, subjected to size selection of 150-500 bp using Agencourt AMPure XP beads (Beckman Coulter) and sequenced on an Illumina HiSeq 2000 with single-end 50 bp read length followed by adapter trimming and filtering by quality score. ChIP-seq library preparation was performed using NEBNext ChIP-seq library preparation set for Illumina and NEBNext multiplex adaptor oligos (New England Biolabs) including size selection to ~250 bp using Agencourt AMPure kit and sequenced on an Illumina HiSeq (library 1) or MiSeq (libraries 2 and 3) with single-end 50 nt read length followed by adapter trimming and quality filter. The quality of the sequences was further assessed by FastQC[55].

TSS mapping. For TSS analysis TEX treated RNA sequences were aligned to the Mja genome using Bowtie[56] allowing for no mismatches in the first 28 nt of the read and filtering out any read that aligned to more than one location, (mapping statistics in Supplementary Table 1). BedTools[57] was used to create strand specific nucleotide resolution histograms of the 5' nucleotide of each read across the entire genome for each replicate. The R statistical program[58] with findPeaks function from package quantmod was used to determine the genome positions containing TSS as peaks, i.e. the highest position in any continuous sequence of counts. These TSS were further filtered as detailed in Supplementary Text and identified TSS are listed in Supplementary Table 2 along with the read count for each replicate at the TSS coordinate.

TU mapping. The TSS list and list of annotated and novel genes (Supplementary Tables 2-4) was used to determine the transcription units (TU) for single gene cistrons, multi gene operons and non-coding RNA genes. TU co-ordinates were defined as the TSS to the stop codon of the last cistron for coding TU, or the annotated end for non-coding RNA. Where multiple TSS occur for a single TU the primary TSS, i.e. that with the highest read count, was used (details in Supplementary Text).

Fidelity of TSS selection. To assay fidelity of TSS the TSS were first filtered so that where multiple assigned TSS occurred within 5 nt the one with the highest read count was retained. Then the number of reads from the TEX treated samples whose 5'-end mapped to each position -5 to +5 relative to the assigned TSS was determined and averaged over the two replicates. For each individual region the read count was normalised to the read count at the +1 position of the assigned TSS. Significance between the same relative positions for assigned TSS with an Inr of T(A/G) compared to those without was determined by Wilcoxon rank sum test.

Transcriptome analysis. For transcriptome analysis random primed RNA sequences were aligned to the Mja genome using Bowtie[56] allowing for no mismatches in the first 28 nt of the read. Reads that align to more than one location were found to only effect 1.8% of the genome so these were included and each mapped to one location so that regions containing repeats (such as the ribosomal rRNA operons) were not misrepresented in the data set. Mapping statistics in Supplementary Table 1. For expression analysis the number of strand specific reads across the length of each TU was determined using BedTools[57] and used to calculate the strand specific RPKM (reads per kilobase per million mapped reads). RPKM values were averaged over the two replicates (Supplementary Table 3). To assess if a TU contains detectable transcript sense RPKM values for each replicate were first log transformed to approximate a normal distribution, then applied a one-sample t-test for Log_{10}(RPKM) greater than 0 (i.e. RPKM greater than 1) followed by Benjamini Hochberg false discovery rate adjustment. An adjusted P value < 0.05 was used to define detectable transcript.

ChIP occupancy analysis. An outline of the sequencing analysis is shown in Supplementary Fig. 1b. ChIP sequenced reads were aligned to genome using Bowtie[56] allowing for no mismatches within the first 28 nt. BAM files were read into the R statistical program[58] with packages ShortRead and GenomicRanges. The package chipseq was used to extend the 50 bp reads in the sense orientation to reflect the average fragment size of 250 nt. Mapping statistics are shown in Supplementary Table 1 (for additional details see Supplementary text).

Genome wide occupancy: overlapping windows across entire genome. For pair-wise genome-wide comparison of occupancies the genome was split into overlapping windows of 250 bp to reflect the average DNA fragment length of the ChIP fragments. The reads per window for each IP and input sample was determined using BedTools[65] and normalised to individual read depth by dividing by total mapped reads per sample, and multiplying by 1,000,000. Each IP sample was divided by the input resulting in the normalised (IP/input) read count. The normalised read count was averaged across replicates and log transformed to provide the Log_2(IP/input) for each region.

Genome wide occupancy: TU occupancy. To determine the TU occupancy each TU with detectable transcript levels (sense RPKM >1 with adjusted P value < 0.05) was first separated into a promoter region corresponding to TSS ± 250 nt (average fragment length), and a intra-TU region starting at the TSS + 250 nt to the end of the TU, and excluding those TU smaller than 250 nt. The reads per segment for each IP and input sample was determined using BedTools[65] and normalised to individual read depth by dividing by total mapped reads per sample, and multiplying by 1,000,000. Each IP sample was divided by the input resulting in the normalised (IP/input) read count. The normalised read count was averaged across replicates and log transformed to provide the Log_2(IP/input) for each region.

Occupancy at specific loci. For comparison of specific genomic intervals BedTools[65] was used to create per nucleotide read count for the extended reads of IP and input samples across the entire genome. The reads were normalised to individual read depth at each position by dividing by total mapped reads per sample, and multiplying by 1,000,000. Each IP sample was divided by the input resulting in the normalised (IP/input) read count. The normalised read count was averaged across replicates and log transformed to provide the Log_2(IP/input) for each position. For individual genomic intervals the histograms at specific genome coordinates were extracted, replicates were averaged, and plots smoothed using sliding 40 bp windows.

Meta-data analysis plots. To prepare average occupancy profiles, the read counts surrounding the regions of interest (e.g. TSS for top 25% of mRNA genes by RPKM) were extracted from the per nucleotide occupancy histograms normalised to read depth and input. The occupancy at each position relative to the site of interest was averaged across each TU. Replicates were averaged and plots smoothed by averaging over sliding 60 bp windows.

Occupancy RNAP vs Spt4/5. In order to detect variations in Spt4/5 recruitment pattern on different TUs, we calculated the difference between Spt4/5 and RNAP occupancy for each 250 bp window across the genome as described above. We extracted the coordinates for windows with a difference < -1, i.e. where Spt4/5 Log_2(IP/input) occupancy was at least 1 lower than RNAP occupancy. Overlapping windows were merged to determine coordinates of theses regions of difference and the read counts for each complete region of difference was calculated and normalised to read depth and input as described above. The significance between RNAP and Spt4/5 occupancies at these regions was determined by applying the Welch's t-test followed by Benjamini Hochberg false discovery rate adjustment. In order to determine whether differences between RNAP and Spt4/5 related to 5'-UTR length of coding TU genome-wide, the difference between Spt4/5 and RNAP occupancy were calculated for each mRNA TU promoter region (see above for calculation of promoter occupancy) and correlated to the length of the 5'-UTR.

Sequence motif analysis. To identify promoter elements, the DNA sequences ranging from -50 to +10 nt relative to the identified TSS were extracted using BedTools[57] and direct alignments were visualised using WebLogo 3[59]. Putative promoter motifs were determined using MEME-ChIP (Motif Analysis of Large Nucleotide Datasets)[60] restricting the search to motifs 6-15 nt wide on the sense strand. The position weight matrix of the resulting

15 nt BRE/TATA motif was used with FIMO (Find Individual Motif Occurrences)[60] to identify matches in the sequences upstream of the TSS and provide confidence scores as P values. Due to high AT content of Mja genome, FIMO was also used to identify matches to the BRE/TATA motif in a control set of 7 randomly generated sets of 1508 sequences of the same length from the Mja genome using BedTools[57] (Supplementary Fig. 3a). For identification of the Mja RBS motif the DNA sequences corresponding to -20 to +20 surrounding the start codons were analysed using MEME-ChIP and restricting the search to motifs of 4-5 nt on the sense strand. For analysis of the dinucleotide frequencies, the proportion of TA or TG at each position relative to the TSS was calculated. This was compared to the genome average occurrence of TA/TG dinucleotides using Fisher's exact test of significance. For analysis of the IMR the percentage of AT at positions -12 to +2 relative to the TSS was calculated using BedTools[57] and significance calculated by Wilcoxon signed rank test.

EMSA and in vitro transcription assays. Recombinant mjRNAP was prepared as in[52] and EMSA assays performed as in[61]. Oligonucleotides are listed in Supplemental Table 6. In vitro transcription reactions with plasmids bearing Mja promoters fused to C-less cassettes were carried out analogous to[9] with the promoter region including 15 bp upstream of the identified BRE/TATA motifs and 8-13 bp downstream of the TSS. For construction of the C-less fusions the following oligos (Supplemental Table 6) were used: *rrnA* fw, CRISPR TSS1 fw, CRISPR TSS2 fw, and *rpl12* fw all with the C-less rev. Buffer conditions and Mja transcription factor concentrations for Mja in vitro transcription assays were as described in[61] with 300 ng of SacIlinearised plasmid, heparin concentration reduced to 5 µg/ml and a single incubation step at 65 °C for 15 min. A recovery marker was included in order to monitor possible losses during the nucleic acid purification prior to gel loading.

Northern blotting. Northern blotting was carried out as in[62] using low range RiboRuler RNA ladder (Fermentas) and probes constructed from oligonucleotide templates A3 sense and A3 antisense (Supplemental Table 6).

Statistical analysis. All graphs were produced using GraphPad Prism version 5 and The R Statistical program[58] and package ggplot2[63]. Correlations and statistical tests were performed using R base install, specific tests are detailed as appropriate throughout the manuscript.

Data availability

The sequencing datasets generated during this study have been deposited in in the NCBI sequence read archive (SRA) with accession codes SRP089683 (ChIP) and SRP089689 (RNA). The supplementary information includes TSS and promoter mapping data (Supplementary table 2) and Mja operon organisation, gene expression and occupancy data (Supplementary table 3) in excel spreadsheet format. The data that support the findings of this study are available from Finn Werner (f.werner@ucl.ac.uk) upon request.

Acknowledgements

We are very grateful to Jürg Bähler and Daniel Bitton for helpful advice throughout this project. We would like to thank Tine Arnvig, Dina Grohman and other members of the RNAP lab for encouragement and critical reading of the manuscript. Research in the RNAP laboratory at University College London is funded by Wellcome Trust Investigator Award WT096553MA (to FW).

Author contributions

KS designed and performed experiments, analysed data and wrote the manuscript. FB performed experiments and wrote the manuscript. RR and MT helped with fermenter growth, cross-linking and provided biomass. FW conceptualised the study, designed experiments and wrote the manuscript.

Competing financial interests

The authors declare no competing financial interests.

References

1 Werner, F. Structural evolution of multisubunit RNA polymerases. *Trends Microbiol* **16**, 247-250, doi:10.1016/j.tim.2008.03.008 (2008).

2 Werner, F. & Grohmann, D. Evolution of multisubunit RNA polymerases in the three domains of life. *Nat Rev Microbiol* **9**, 85-98, doi:10.1038/nrmicro2507 (2011).

3 Korkhin, Y. *et al.* Evolution of complex RNA polymerases: the complete archaeal RNA polymerase structure. *PLoS Biol* **7**, e1000102, doi:10.1371/journal.pbio.1000102 (2009).

4 Hirtreiter, A., Grohmann, D. & Werner, F. Molecular mechanisms of RNA polymerase - the F/E (RPB4/7) complex is required for high processivity in vitro. *Nucleic Acids Res* **38**, 585-596, doi:10.1093/nar/gkp928 (2010).

5 Li, E., Reich, C. I. & Olsen, G. J. A whole-genome approach to identifying protein binding sites: promoters in *Methanocaldococcus* (*Methanococcus*) *jannaschii*. *Nucleic Acids Res* **36**, 6948-6958, doi:10.1093/nar/gkm499 (2008).

6 Zhang, J., Li, E. & Olsen, G. J. Protein-coding gene promoters in *Methanocaldococcus* (*Methanococcus*) *jannaschii*. *Nucleic Acids Res* **37**, 3588-3601, doi:10.1093/nar/gkp213 (2009).

7 Werner, F. & Weinzierl, R. O. A recombinant RNA polymerase II-like enzyme capable of promoter-specific transcription. *Mol Cell* **10**, 635-646 (2002).

8 Gietl, A. *et al.* Eukaryotic and archaeal TBP and TFB/TF(II)B follow different promoter DNA bending pathways. *Nucleic acids Res* **42**, 6219-6231, doi:10.1093/nar/gku273 (2014).

9 Blombach, F. *et al.* Archaeal TFEalpha/beta is a hybrid of TFIIE and the RNA polymerase III subcomplex hRPC62/39. *Elife* **4**, e08378, doi:10.7554/eLife.08378 (2015).

10 Grohmann, D. *et al.* The initiation factor TFE and the elongation factor Spt4/5 compete for the RNAP clamp during transcription initiation and elongation. *Mol Cell* **43**, 263-274, doi:10.1016/j.molcel.2011.05.030 (2011).

11 Werner, F. A nexus for gene expression-molecular mechanisms of Spt5 and NusG in the three domains of life. *J Mol Biol* **417**, 13-27, doi:10.1016/j.jmb.2012.01.031 (2012).

12 Sevostyanova, A., Svetlov, V., Vassylyev, D. G. & Artsimovitch, I. The elongation factor RfaH and the initiation factor sigma bind to the same site on the transcription elongation complex. *Proc Natl Acad Sci U S A* **105**, 865-870, doi:10.1073/pnas.0708432105 (2008).

13 Mayer, A. *et al.* Uniform transitions of the general RNA polymerase II transcription complex. *Nat Struct Mol Biol* **17**, 1272-1278, doi:10.1038/nsmb.1903 (2010).

14 Mooney, R. A. *et al.* Regulator trafficking on bacterial transcription units in vivo. *Mol Cell* **33**, 97-108, doi:10.1016/j.molcel.2008.12.021 (2009).

15 Kadonaga, J. T. Perspectives on the RNA polymerase II core promoter. *Wiley Interdiscip Rev Dev Biol* **1**, 40-51, doi:10.1002/wdev.21 (2012).

16 Nagy, J. *et al.* Complete architecture of the archaeal RNA polymerase open complex from single-molecule FRET and NPS. *Nature Commun* **6**, 6161, doi:10.1038/ncomms7161 (2015).

17 Schulz, S. *et al.* TFE and Spt4/5 open and close the RNA polymerase clamp during the transcription cycle. *Proc Natl Acad Sci U S A* **113**, E1816-1825, doi:10.1073/pnas.1515817113 (2016).

18 Bell, S. D., Jaxel, C., Nadal, M., Kosa, P. F. & Jackson, S. P. Temperature, template topology, and factor requirements of archaeal transcription. *Proc Natl Acad Sci U S A* **95**, 15218-15222 (1998).

19 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, doi:10.1186/1471-2164-15-684 (2014).

20 Jäger, D. *et al.* Deep sequencing analysis of the *Methanosarcina mazei* Go1 transcriptome in response to nitrogen availability. *Proc Natl Acad Sci U S A* **106**, 21878-21882, doi:10.1073/pnas.0909051106 (2009).

21 Li, J. *et al.* Global mapping transcriptional start sites revealed both transcriptional and post-transcriptional regulation of cold adaptation in the methanogenic archaeon *Methanolobus psychrophilus*. *Sci Rep* **5**, 9209, doi:10.1038/srep09209 (2015).

22 Wurtzel, O. *et al.* A single-base resolution map of an archaeal transcriptome. *Genome Res* **20**, 133-141, doi:10.1101/gr.100396.109 (2010).

23 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, doi:10.1186/s12864-016-2920-y (2016).

24 Seitzer, P., Wilbanks, E. G., Larsen, D. J. & Facciotti, M. T. A Monte Carlo-based framework enhances the discovery and interpretation of regulatory sequence motifs. *BMC Bioinformatics* **13**, 317, doi:10.1186/1471-2105-13-317 (2012).

25 Blombach, F., Smollett, K. L., Grohmann, D. & Werner, F. Molecular Mechanisms of Transcription Initiation-Structure, Function, and Evolution of TFE/TFIIE-Like Factors and Open Complex Formation. *J Mol Biol* **428**, 2592-2606, doi:10.1016/j.jmb.2016.04.016 (2016).

26 Werner, F. & Weinzierl, R. O. Direct modulation of RNA polymerase core functions by basal transcription factors. *Mol Cell Biol* **25**, 8344-8355, doi:10.1128/MCB.25.18.8344-8355.2005 (2005).

27 Qureshi, S. A., Bell, S. D. & Jackson, S. P. Factor requirements for transcription in the Archaeon *Sulfolobus shibatae*. *Embo J* **16**, 2927-2936, doi:10.1093/emboj/16.10.2927 (1997).

28 Burmann, B. M. *et al.* A NusE:NusG complex links transcription and translation. *Science* **328**, 501-504, doi:10.1126/science.1184953 (2010).

29 Proshkin, S., Rahmouni, A. R., Mironov, A. & Nudler, E. Cooperation between translating ribosomes and RNA polymerase in transcription elongation. *Science* **328**, 504-508, doi:10.1126/science.1184939 (2010).

30 French, S. L., Santangelo, T. J., Beyer, A. L. & Reeve, J. N. Transcription and translation are coupled in Archaea. *Mol Biol Evol* **24**, 893-895, doi:msm007 [pii] 10.1093/molbev/msm007 (2007).

31 Brenneis, M., Hering, O., Lange, C. & Soppa, J. Experimental characterization of Cis-acting elements important for translation and transcription in halophilic archaea. *PLoS Genet* **3**, e229, doi:10.1371/journal.pgen.0030229 (2007).

32 Torarinsson, E., Klenk, H. P. & Garrett, R. A. Divergent transcriptional and translational signals in Archaea. *Environ Microbiol* **7**, 47-54, doi:10.1111/j.1462-2920.2004.00674.x (2005).

33 Koide, T. *et al.* Prevalence of transcription promoters within archaeal operons and coding sequences. *Mol Syst Biol* **5**, 285, doi:10.1038/msb.2009.42 (2009).

34 Toffano-Nioche, C. *et al.* RNA at 92 degrees C: the non-coding transcriptome of the hyperthermophilic archaeon *Pyrococcus abyssi*. *RNA Biol* **10**, 1211-1220, doi:10.4161/rna.25567 (2013).

35 Yang, C., Bolotin, E., Jiang, T., Sladek, F. M. & Martinez, E. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene* **389**, 52-65, doi:10.1016/j.gene.2006.09.029 (2007).

36 Qureshi, S. A. Role of the *Sulfolobus shibatae* viral T6 initiator in conferring promoter strength and in influencing transcription start site selection. *Can J Microbiol* **52**, 1136-1140, doi:10.1139/w06-073 (2006).

37 Bell, S. D. & Jackson, S. P. The role of transcription factor B in transcription initiation and promoter clearance in the archaeon *Sulfolobus acidocaldarius*. *J Biol Chem* **275**, 12934-12940, doi:DOI 10.1074/jbc.275.17.12934 (2000).

38 Shultzaberger, R. K., Chen, Z., Lewis, K. A. & Schneider, T. D. Anatomy of *Escherichia coli* sigma70 promoters. *Nucleic Acids Res* **35**, 771-788, doi:10.1093/nar/gkl956 (2007).

39 Basu, R. S. *et al.* Structural basis of transcription initiation by bacterial RNA polymerase holoenzyme. *J Biol Chem* **289**, 24549-24559, doi:10.1074/jbc.M114.584037 (2014).

40 Ehrensberger, A. H., Kelly, G. P. & Svejstrup, J. Q. Mechanistic interpretation of promoter-proximal peaks and RNAPII density maps. *Cell* **154**, 713-715, doi:10.1016/j.cell.2013.07.032 (2013).

41 Hirtreiter, A. *et al.* Spt4/5 stimulates transcription elongation through the RNA polymerase clamp coiled-coil motif. *Nucleic Acids Res* **38**, 4040-4051, doi:10.1093/nar/gkq135 (2010).

42 Diamant, G., Bahat, A. & Dikstein, R. The elongation factor Spt5 facilitates transcription initiation for rapid induction of inflammatory-response genes. *Nature Commun* **7**, 11547, doi:10.1038/ncomms11547 (2016).

43 Larochelle, S. *et al.* Cyclin-dependent kinase control of the initiation-to-elongation switch of RNA polymerase II. *Nat Struct Mol Biol* **19**, 1108-1115, doi:10.1038/nsmb.2399 (2012).

44 Arnvig, K. B. *et al.* Evolutionary comparison of ribosomal operon antitermination function. *J Bacteriol* **190**, 7251-7257, doi:10.1128/JB.00760-08 (2008).

45 Micorescu, M. *et al.* Archaeal transcription: function of an alternative transcription factor B from *Pyrococcus furiosus*. *J Bacteriol* **190**, 157-167, doi:10.1128/JB.01498-07 (2008).

46 Jensen, K. F. & Pedersen, S. Metabolic growth rate control in *Escherichia coli* may be a consequence of subsaturation of the macromolecular biosynthetic apparatus with substrates and catalytic components. *Microbiol Rev* **54**, 89-100 (1990).

47 Peeters, E., Driessen, R. P., Werner, F. & Dame, R. T. The interplay between nucleoid organization and transcription in archaeal genomes. *Nat Rev Microbiol* **13**, 333-341, doi:10.1038/nrmicro3467 (2015).

48 Ouhammouch, M., Dewhurst, R. E., Hausner, W., Thomm, M. & Geiduschek, E. P. Activation of archaeal transcription by recruitment of the TATA-binding protein. *Proc Natl Acad Sci U S A* **100**, 5097-5102, doi:10.1073/pnas.0837150100 (2003).

49 Churchman, L. S. & Weissman, J. S. Native elongating transcript sequencing (NET-seq). *Curr Protoc Mol Biol* **Chapter 4**, Unit 4 14 11-17, doi:10.1002/0471142727.mb0414s98 (2012).

50 Nojima, T. *et al.* Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell* **161**, 526-540, doi:10.1016/j.cell.2015.03.027 (2015).

51 Jones, W. J., Leigh, J. A., Mayer, F., Woese, C. R. & Wolfe, R. S. *Methanococcus jannaschii* Sp. Nov., an extremely thermophilic

methanogen from a submarine hydrothermal vent. *Arch Microbiology* **136**, 254-261 (1983).

52 Smollett, K., Blombach, F. & Werner, F. Transcription in Archaea: preparation of *Methanocaldococcus jannaschii* transcription machinery. *Methods Mol Biol* **1276**, 291-303, doi:10.1007/978-1-4939-2392-2_17 (2015).

53 Reichelt, R., Gindner, A., Thomm, M. & Hausner, W. Genome-wide binding analysis of the transcriptional regulator TrmBL1 in *Pyrococcus furiosus*. *BMC Genomics* **17**, doi:10.1186/s12864-015-2360-0 (2016).

54 Liu, W., Vierke, G., Wenke, A. K., Thomm, M. & Ladenstein, R. Crystal structure of the archaeal heat shock regulator from *Pyrococcus furiosus*: A molecular chimera representing eukaryal and bacterial features. *J Mol Biol* **369**, 474-488, doi:10.1016/j.jmb.2007.03.044 (2007).

55 Andrews, S. FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc. (2010).

56 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10** (2009).

57 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

58 R Core Team. R: A language and environment for statistical computing. http://www.R-project.org/. (2014).

59 Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res* **14**, 1188-1190, doi:10.1101/gr.849004 (2004).

60 Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**, W202-208, doi:10.1093/nar/gkp335 (2009).

61 Smollett, K., Blombach, F. & Werner, F. Transcription in Archaea: in vitro transcription assays for mjRNAP. *Methods Mol Biol* **1276**, 305-314, doi:10.1007/978-1-4939-2392-2_18 (2015).

62 Arnvig, K. B. & Young, D. B. Identification of small RNAs in *Mycobacterium tuberculosis*. *Mol Microbiol* **73**, 397-408, doi:10.1111/j.1365-2958.2009.06777.x (2009).

63 Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2009).

Figures and legends

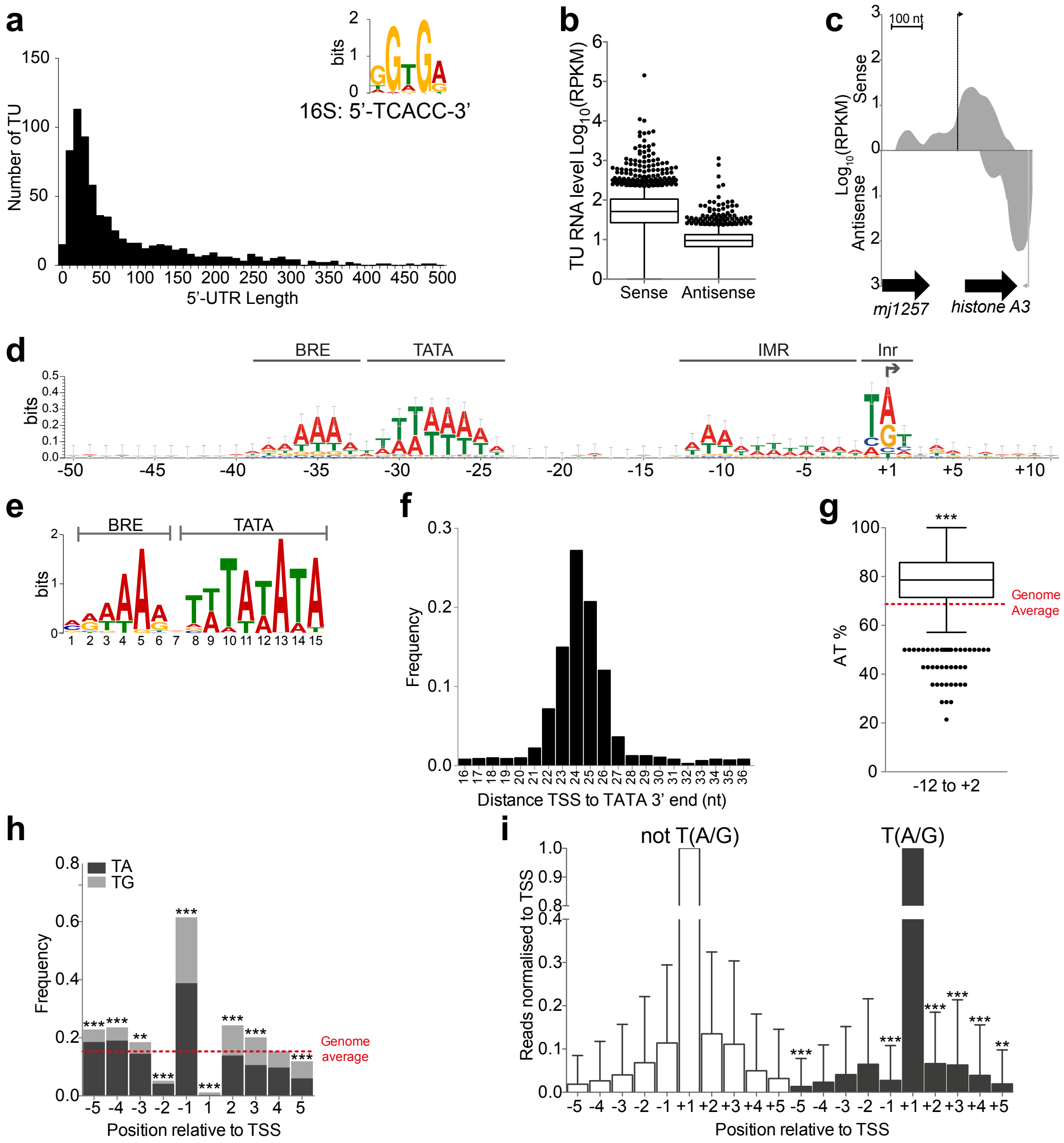


Figure 1: Transcription start site map and promoter motif analysis. a, The 5'-UTR distance distribution from the primary TSS to the start codon of Mja mRNAs (n = 689). The insert shows the ribosome binding site (RBS) sequence motif identified by the MEME algorithm; for comparison the complementary sequence of the Mja 16S RNA is shown. **b,** Comparison of sense and antisense RNA levels at all TUs (n= 1138) the whiskers indicate 1.5X the interquartile range, individual RPKM values represent average of two biological replicates. **c,** Strand-specific RNA profiles reveals sense- and

antisense transcripts on the histone A3 locus. The grey arrows indicate TSS, average of two biological replicates. **d,** Promoter DNA sequence alignments centred on the TSS (n = 1508) reveal regions with a sequence bias corresponding to the BRE/TATA elements, the initially melted region (IMR), and the initiator (Inr) of the promoter. **e**, The BRE/TATA consensus motif identified by MEME-ChIP. **f,** the distance between the 3' end of the TATA motif and the TSS is centred on 24 nt (TATA at a P value of < 10^{-3}, n = 1129). **g**, The AT content distribution of the IMR that exceeds the genome average of 68.7% (red dotted line), the whiskers indicate 1.5X the interquartile range. Significance according to a Wilcoxon signed rank test (P < 10^{-10}, n = 1508). **h**, The di-nucleotide frequency of TA and TG motifs surrounding the TSS. The red dotted line indicates the genome wide frequency of 0.15, and the significance was assessed by Fisher's exact test (n = 1507). **i**, The T(A/G) motif increases the precision of TSS selection. The read count of all 5'- ends from TEX-treated RNA surrounding assigned TSSs were identified (averaged across the two biological replicates), and the reads normalised to the TSS at each position. Data shows mean + standard deviation, n = 447 not T(A/G) or 762 T(A/G). Initiation immediately upstream and downstream is four- and two-fold lower, respectively, for TSS with T(A/G) compared to those without by Wilcoxon rank sum test. P value: * <0.05; ** <0.01; *** <0.001).

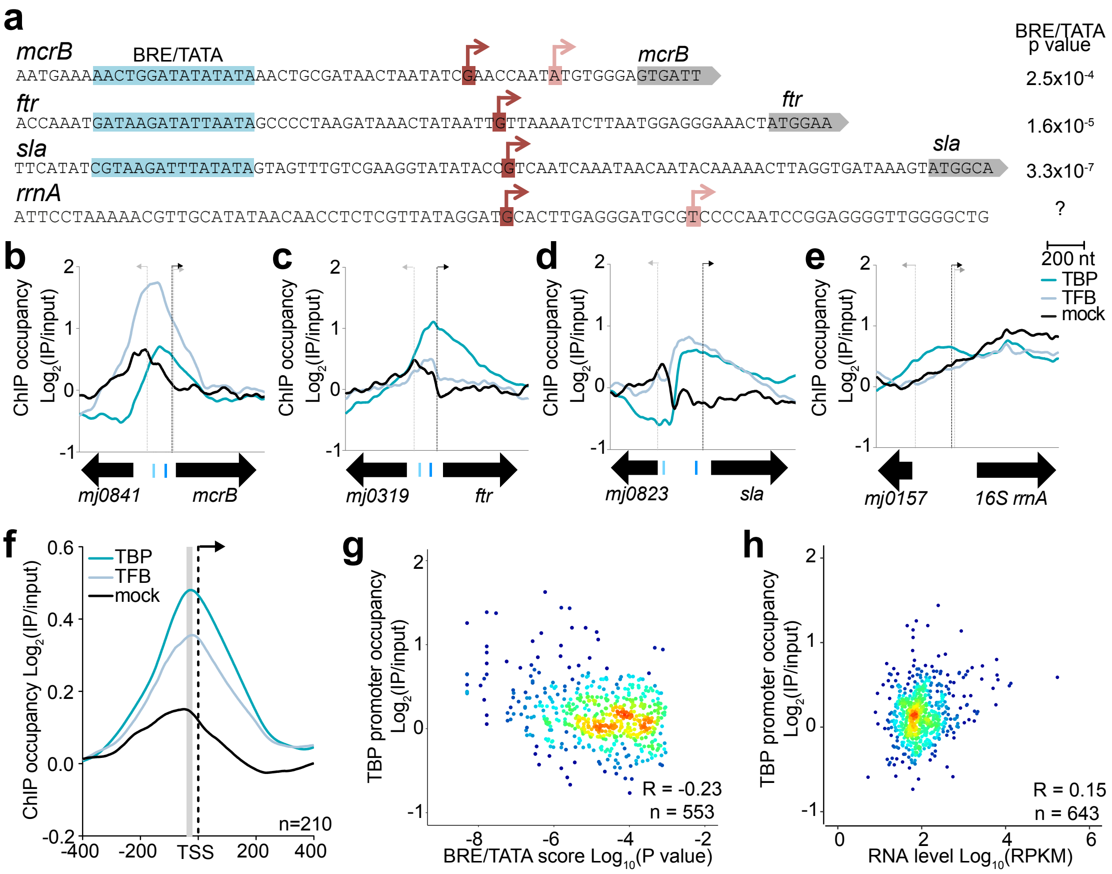


Figure 2: Correlation between TBP/TFB binding to the promoter and RNA levels. a, The BRE/TATA motifs (highlighted in blue), primary and secondary TSS (red and pink, respectively), and the coding region (grey) of three selected mRNA (*mcrB*, *ftr* and *sla*) and the rRNA promoter. The confidence score (P value) for the BRE/TATA motif is indicated to the right of the sequence. **b-e**, TBP, TFB and mock control occupancy profiles at the *mcrB* (**b**), *ftr* (**c**), *sla* (**d**) and *rrnA* (**e**) promoter. TSS are indicated as arrows, with the primary TSS in black. **f**, A metadata analysis shows that the averaged occupancy profiles of TBP and TFB of the top 25% of mRNA TU (by sense RPKM, n = 210) collocate with the predicted BRE/TATA motif (grey). **g**, Correlation between the BRE/TATA score (P value) and TBP occupancy. Spearman correlations are indicated TBP R = -0.23, P value = 6 x 10^{-8}, n =

553, points are coloured using a density gradient (ranging from blue-low to red-high). **h,** Correlation between the TBP occupancy and RNA levels (sense RPKM for all TU with detectable transcript, average of two biological replicates). Spearman correlations indicated on TBP R = 0.15, P value = 1 x 10^{-4}, n = 643. Occupancy data in panels **a-h** represent the average of four (TBP) or two (TFB and mock) technical replicates.

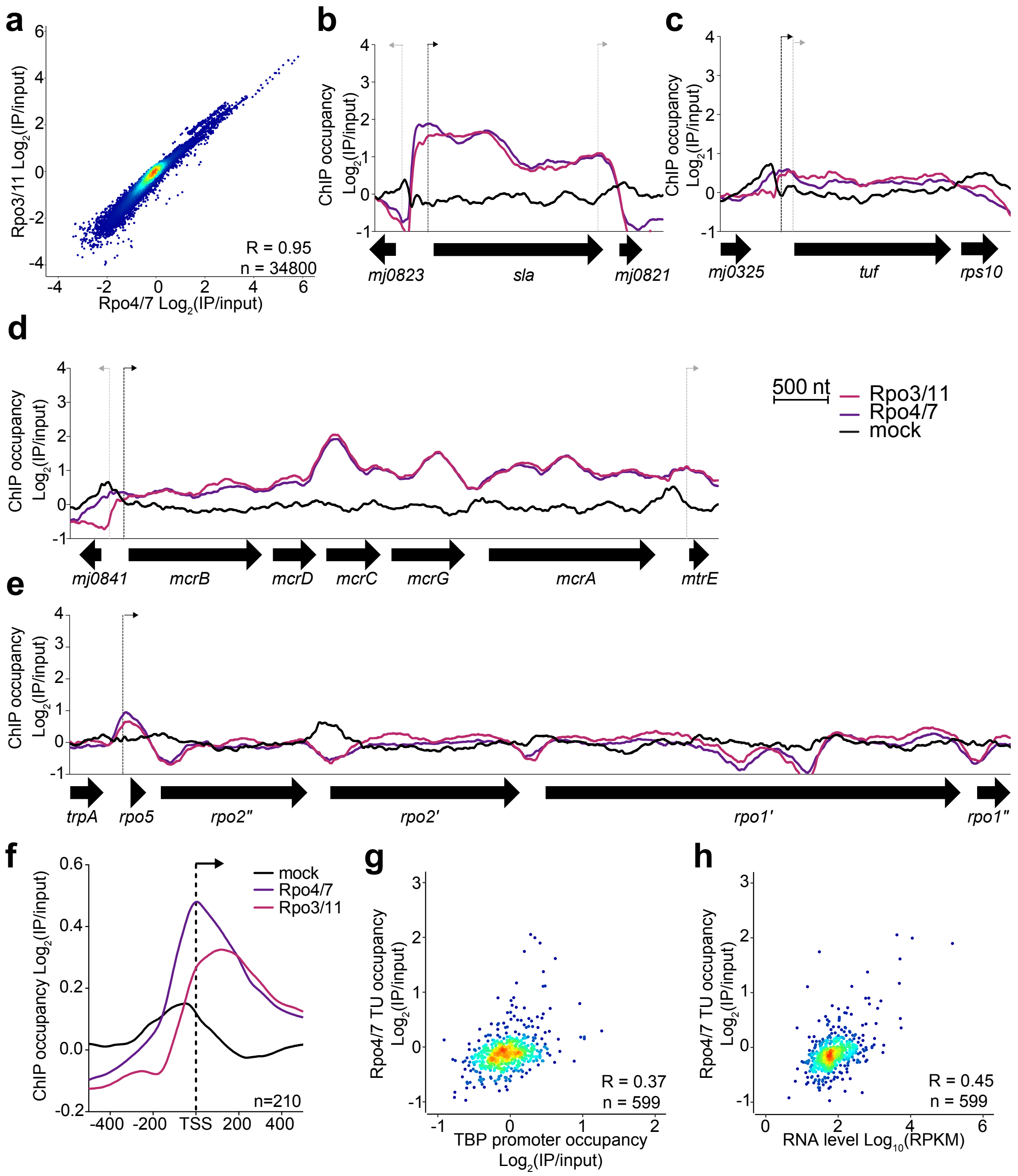


Figure 3: The Rpo4/7 stalk and RNAP core remain associated through the transcription cycle. a, The correlation between the occupancy of RNAP subunit complexes Rpo4/7 and Rpo3/11 is very strong across the genome, Spearman correlations indicated (P value < 10^{-10}, n = 34800). **b-e**, RNAP occupancy profiles on representative TUs: the *sla* (**b**), *tuf* (**c**), *mcr* (**d**), and RNAP subunit operon (**e**). Arrows indicate TSS (primary in black). **f**, Averaged occupancy profiles of Rpo4/7, Rpo3/11 and mock control at the top 25% of

mRNA TU (by sense RPKM, n = 210). **g**, Correlation between the TBP promoter occupancy (TSS +/- 250 bp) and RNAP TU occupancy (TSS + 250 to TU end) for all Tus (RPKM > 1). Spearman correlations TBP R = 0.37, P value < 10^{-10}, n = 599. **h,** Correlation between steady-state RNA levels (sense RPKM for all TU RPKM > 1, average of two biological replicates), and RNAP (Rpo4/7) occupancy within the body of each TU, Spearman correlations Rpo4/7 R = 0.45, P value < 10^{-10}, n = 599. Occupancy data in panels a-h represent the average of four (Rpo4/7), three (Rpo3/11) or two (mock) technical replicates.

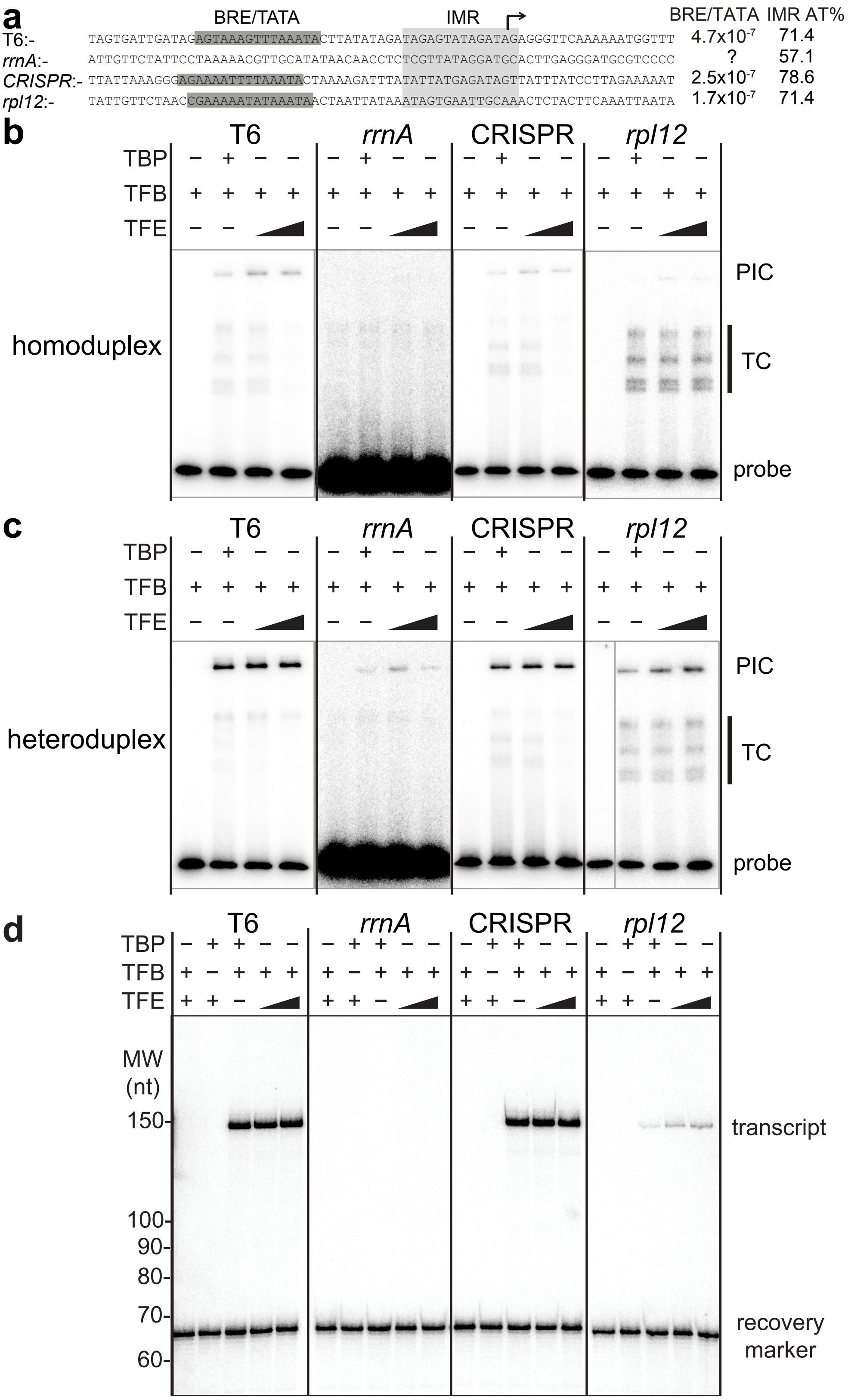

a
BRE/TATA
IMR
BRE/TATA IMR AT%
T6:- TAGTGATTGATAGAGTAAAGTTTAAATACTTATATAGATAGAGTATAGATAGAGGGTTCAAAAAATGGTTT 4.7x10-7 71.4
rrnA:- ATTGTTCTATTCCTAAAAACGTTGCATATAACAACCTCTCGTTATAGGATGCACTTGAGGGATGCGTCCCC ? 57.1
CRISPR:- TTATTAAAGGGAGAAAATTTTAAATACTAAAAGATTTATATTATGAGATAGTTATTTATCCTTAGAAAAAT 2.5x10-7 78.6
rpl12:- TATTGTTCTAACCGAAAAATATAAATAACTAATTATAAATAGTGAATTGCAAACTCTACTTCAAATTAATA 1.7x10-7 71.4
b
T6
rrnA
CRISPR
rpl12
TBP
TFB
TFE
homoduplex
PIC
TC
probe
c
T6
rrnA
CRISPR
rpl12
TBP
TFB
TFE
heteroduplex
PIC
TC
probe
d
T6
rrnA
CRISPR
rpl12
TBP
TFB
TFE
MW (nt)
150
100
90
80
70
60
transcript
recovery marker

Figure 4: PIC formation and promoter strength in vitro. **a**, Alignment of SSV T6 model promoter and representative Mja promoters including ribosomal RNA (*rrnA*), CRISPR and mRNA (ribosomal protein *rpl12*) promoters. The BRE/TATA motifs are shown in dark gray with P values indicated, the IMR is highlighted in light grey with AT% indicated. **b**, EMSA showing preinitiation complex (PIC) formation on promoter templates shown in (**a**). **c**, EMSAs using heteroduplex promoter variants. PIC indicates the transcription preinitiation complex, and TC the ternary DNA-TBP-TFB complexes. Exposure is adjusted to account for diverse signal intensities. **d**, Promoter-directed in vitro transcription assays. Promoter templates shown in (**a**) were fused to C-less cassette resulting in transcripts of 150 nt (T6), 157 nt (*rrnA*) and 152 nt (CRISPR and *rpl12*) length. A representative example of two technical replicates are shown.

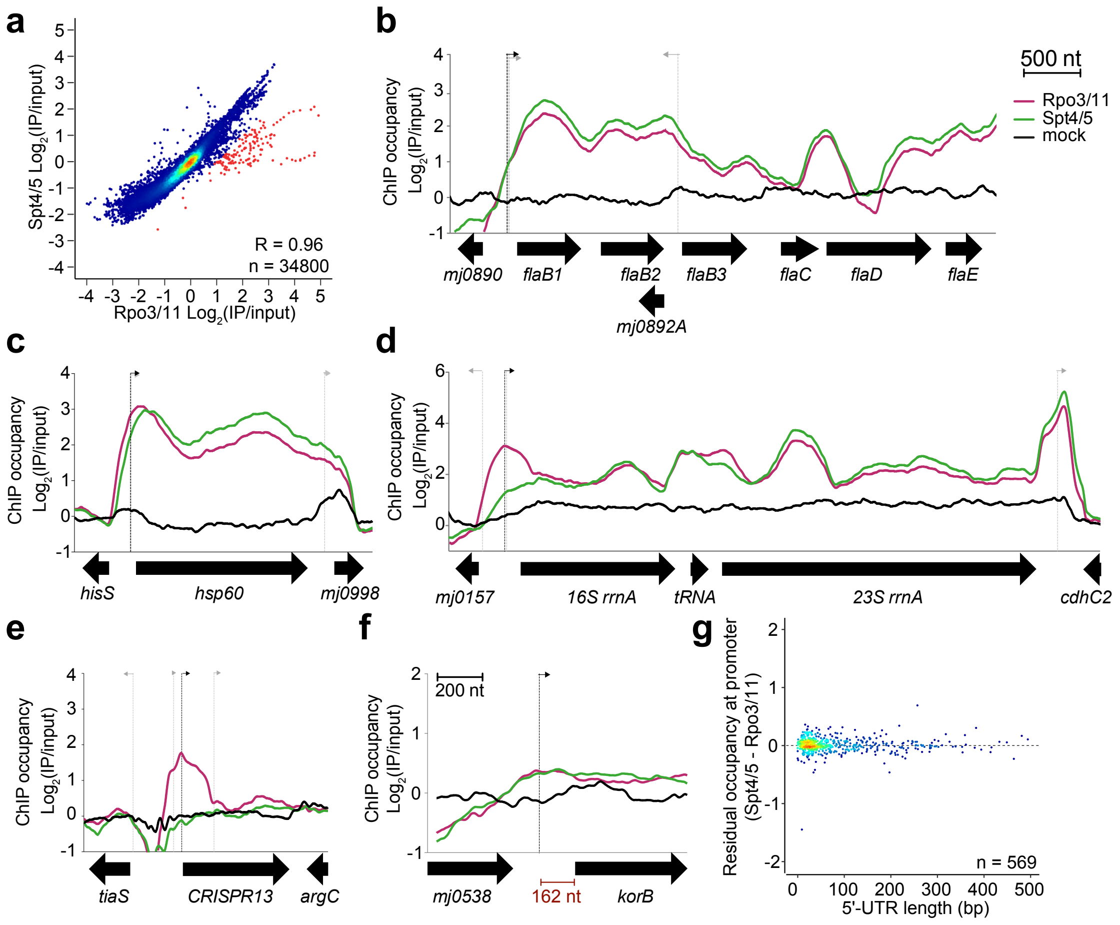


Figure 5: Archaeal Spt4/5 is a general elongation factor that is recruited to RNAP via two distinct modes. a, Spt4/5 and RNAP occupancy correlates very strongly across the whole genome. Data points of substoichiometric Spt4/5:RNAP occupancy, with Spt4/5 occupancy more than 1 Log$_2$(IP/input) lower than RNAP occupancy, are indicated in red, Spearman correlations R = 0.96, P value < 10^{-10}, n = 34800. **b-f**, The Spt4/5 occupancy profiles reflect two recruitment modes of Spt4/5 exemplified by the archaealleum (**b**) and rRNA operons (**d**). Representative RNAP and Spt4/5 occupancy profiles on the *fla* (**b**), *hsp60* (**c**), *rrnA* (**d**), *CRISPR13* operon (**e**) and larger scale plot of the long 5' UTR gene *korB* gene (**f**). Arrows indicate TSS. **g**, The 5'-UTR length does not affect the difference between Spt4/5 and Rpo3/11 occupancy proximal to the promoter of TUs (RPKM > 1), n = 569. Occupancy data in

panels **a-g** represent the average of three (Rpo3/11 and Spt4/5) or two (mock) technical replicates.

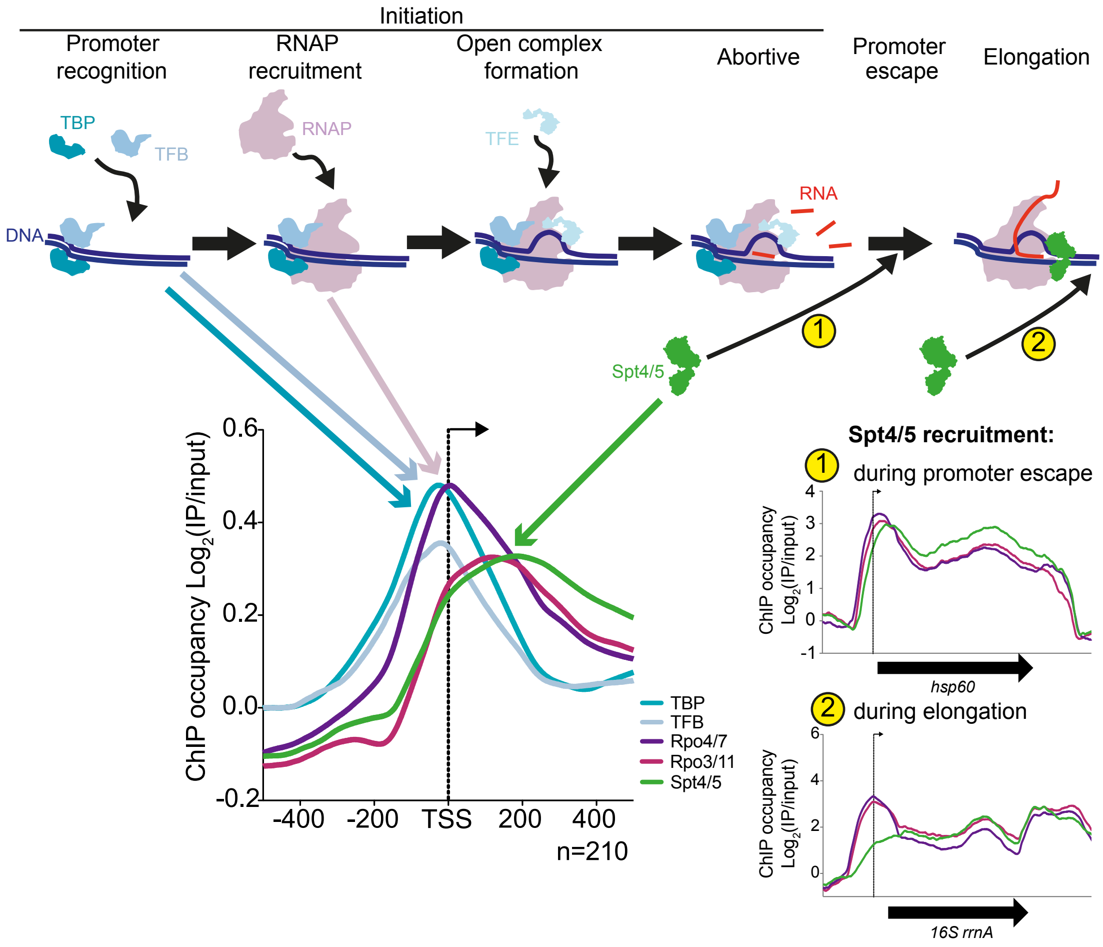


Figure 6: The initial stages of the transcription cycle in archaea. The average occupancy profiles of TBP, TFB, RNAP and Spt4/5 on the top 25% of mRNA TUs reflect the initial stages of the transcription cycle. TBP and TFB are bound to the TATA and BRE promoter elements 24 nt upstream of the TSS, which in turn recruit RNAP to form the preinitiation complex (PIC). Subsequently, two modes of Spt4/5 recruitment could be distinguished on different genes: 1. On the majority of genes Spt4/5 is recruited 'early', likely during promoter escape; 2. On the ribosomal rRNA operons and CRISPR Spt4/5 is recruited 'later' offset from TSS in the downstream direction, likely occurring during transcription elongation.

Supplementary Information:

Supplementary Notes

In order to carry out a comprehensive systems analysis of transcription in *M. jannaschii* (Mja) we combined genome-wide analysis of transcription start site (TSS) mapping, examination of total, steady-state, RNA levels (RNA-seq), and chromatin immunoprecipitation (ChIP-seq) of TBP, TFB, RNAP and Spt4/5. The steps of our analysis are outlined in Supplementary Fig. 1, and the sequencing mapping statistics are shown in Supplementary Table 1.

Transcription start site mapping of *Methanocaldococcus jannaschii.* To identify genome-wide transcription start sites (TSSs) we used the terminator exonuclease (TEX)-treated RNA combined with high-throughput sequencing. TEX treatment removes RNA containing 5'-monophosphates that result from RNA processing while retaining nascent RNA due to their triphosphate moiety at the 5' terminus. Total RNA was isolated from Mja grown under optimal growth conditions. The RNA 5' position of each 50 nt single-end read of the TEX treated samples were mapped to the Mja genome and used to generate strand specific histograms. The genomic positions of the maximum in any continuous sequence of counts were extracted as the 'peak position' (for examples see Supplemental Fig. 2a-c). We defined TSSs as peak positions that had > 50 reads for at least one of the two biological replicates.

Each predicted transcription unit (TU, see below) was limited to a maximum of 4 TSS, ranked according to their read depth. Where a TSS was not identified for a predicted TU first cistron, below threshold peaks (i.e. < 50 reads) were used where possible to enable more complete TSS mapping resulting in an additional 250 TSS. Using this methodology we identified 1508 TSS in the Mja genome (Supplemental Table 2). As our method assigned the local maximum as a TSS, rather than the most upstream signal, we validated our TSS assignment using a data set of 134 TSS that were previously identified by primer extension and 5'-RACE[1]. We were able to identify 103 of these, of which 77% matched the precise genomic position in our map and 93% were accurate to within 2 bp (Supplemental Fig. 2d). In conclusion, our TEX mapping provides a faithful picture of genome wide TSS landscape.

Organisation of the Mja transcription units. Using the TSS map and the annotated Mja open reading frames (ORF) we characterised the genome-wide TU organisation of Mja. All TUs (both coding and non-coding) were defined as regions spanning from the TSS to the stop codon of the last cistron for coding TU, or annotated 3' end of non-coding RNA. In case of TUs with multiple TSS, the primary TSS (i.e. with the highest read count) was used (Supplemental Table 3). Mja TUs are organised into a combination of single cistron genes (70% of TU) and multicistronic operons (with 47 operons containing 5 cistrons or more) as in other archaea, and similar to the bacterial gene organisation (Supplemental Fig. 2e). 55 TSS were mapped immediately downstream of an annotated start codon, but comparison with homologous genes in other species revealed that this was most likely the result of a start codon mis-annotation; the start codons of these Mja genes were altered accordingly (Supplemental Table 3). We were also able to account for TSS of 26 of the 37 small ncRNA that have been predicted computationally[2-4].

There were several TSSs that did not associate with annotated genes. We defined novel genes where the total RNA-seq revealed an increase in reads downstream of the TSS. Due to the high stringency of the TSS cut-off used in our analyses it is possible that additional TUs are present in the Mja genome, but for the purpose of this investigation only high confidence novel TUs were included. This led to the identification of 19 potential new ORFs, 17 new intergenic or antisense ncRNAs (Supplemental Table 4), and 1 new CRISPR repeat region. The Mja genome encodes 20 CRISPR loci that facilitate prokaryotic adaptive immunity[5]. Interestingly, the majority of CRISPR RNA precursors had two TSS located within the leader at positions -20 and -90 relative to the first repeat, in contrast to the single TSS found for *E. coli* and *Sulfolobus* CRISPR systems[6,7]. For some of the Mja CRISPR loci we identified a third TSS upstream of the leader, which suggests a more complex promoter organisation. In addition we identified 41 TSSs positioned within TUs in sense orientation, which could potentially result in synthesis of N-terminally truncated proteins or regulatory noncoding (nc) RNAs. We were able to detect BRE/TATA motifs for 89% of the predicted new ORFs, 75% of the newly annotated antisense and intergenic ncRNAs, and 78% of the intragenic TSSs indicating that these are likely to be real promoters.

Ribosome binding sites (RBS). Sequence analysis surrounding the annotated start codon of coding genes identified the RBS consensus GGWGR (W = A or T; R = A or G) 4-6 nt upstream of the ATG, which is complementary to the Mja 16S rRNA sequence 5'-TCACC-3' (Fig. 1a) as has been described elsewhere[1,8]. We identified potential RBSs matching the consensus for 54% of the protein encoding genes. In some cases the identified RBS was found to overlap or be slightly downstream of the annotated start codon. Similar to the TSS mapping, sequence comparison revealed that in most cases this was likely due to a misannotation of the start codons, which were updated accordingly (Supplementary Table 3). Using the updated ORF maps, 36% of the genes with reannoated start codons now included an RBS immediately upstream of the start codon. In total 300 genes had their start sites changed due to TSS or RBS mapping. Overall we defined 1114 different TUs, 976 (88%) of which we were able to assign a TSS while the remaining 138 TUs were identified based on divergent orientation of the upstream gene.

The Mja transcriptome. To gain a picture of the transcript profile of Mja we calculated the sense strand RPKM (reads per kilobase per million, the values for the two biological replicates were averaged) for all 1114 TUs (Supplementary table 3). In order to assess the Mja transcript profile, the sense RPKM values for each replicate were first log transformed to approximate a normal distribution, then subjected to a one sample t-test for Log_{10}(RPKM) greater than 0 (i.e. RPKM greater than 1) followed by Benjamini Hochberg FDR adjustment. An adjusted P value < 0.05 was used to define a TU as being expressed. Of the total of 1114 TUs, 700 (63%) could be identified as expressed. We used these 700 TU for the downstream analysis. We were able to detect antisense transcription, albeit at much lower levels compared to sense transcription (Fig. 1b). When the highly stringent statistical analysis we used for sense transcript was applied to the antisense signals, none of the TUs contained antisense transcript with RPKM > 1 and adjusted P value < 0.05. This is likely due to the majority of antisense RNAs only partially covering the length of the TU, while RPKM calculations were based on the TU size (in lieu of well-defined borders for the antisense RNAs). This could also be due to rapid degradation of antisense RNAs, a hypothesis that is supported by the fact that most antisense RNAs were not found associated with a TSS, and that smaller transcripts are likely to have been depleted in the sizing step of the library preparation. Antisense transcription at higher abundance has been noted in other archaeal species, which suggests that this is a common phenomenon in archaea. However, several of these studies were specifically aimed at characterising sRNAs and included enrichment steps for smaller RNA species, rather using conditions that would deplete them[9-16]. Only a modest number of small ncRNA were identified (Supplementary Table 3), in agreement with computational predictions[4], although, as with antisense transcripts, larger numbers may be discovered by enriching for small transcripts.

Occupancy profiling of the Mja general transcription machinery using ChIP-seq. Mja was cultured under optimal growth conditions and chemically cross-linked at the physiologically relevant temperature of 85˚C with formaldehyde for 1 minute before quenching with glycine and cooling of the sample[17,18]. For the immunoprecipitations we used polyclonal antibodies raised against recombinant proteins including the RNAP subcomplexes Rpo4/7 (4 technical replicates) and Rpo3/11 (3 technical replicates), the transcription initiation factors TBP (4 technical replicates) and TFB (2 technical replicates) and the elongation factor Spt4/5 (3 technical replicates), as well as a mock control antibody (pre-immune sera, 2 technical replicates). Resulting ChIP DNA samples and input control were subjected to high-throughput, single-end sequencing on a Illumina MiSeq and HiSeq platforms. Each read covered 50 nt of the 5'-end of the sequenced DNA fragment. To provide a more accurate representation of the genomic DNA fragments the reads were extended to 250 nt, reflecting the average fragment length of the initial sequenced library, and therefore the resolution of the ChIP analysis. The Mja genome was split into overlapping windows of 250 bp (total windows = 34,800) and the reads that map to each window were calculated for each sample. The reads per window for each IP and input sample was determined and normalised to individual read depth by dividing by total mapped reads per sample, and multiplying by 1,000,000 (chosen arbitrarily to obtain a convenient order of magnitude for the numbers). Each IP sample was divided by the input resulting in the normalised (IP/input) read count. The normalised read count was averaged across replicates and log transformed to provide the Log_2(IP/input) for each region. The occupancy distribution across all windows shows little variability (interquartile range 0.17) for the mock control, which indicates that the overall level of noise is low; the ChIP samples are much more variable (Supplemental Fig. 5a). Additionally, correlations between the mock and different IP samples is extremely weak, indicating that the ChIP signals differ dramatically from the noise (Supplemental Fig. 5b). Plotting the per nucleotide occupancy of the mock control illustrates the background noise on an individual gene level (Fig. 3b-e, Fig. 5b-f). Here instead of splitting the genome into windows, the extended reads are used to calculate the reads per base across the genome, before normalising as above to give the Log_2(IP/input). The graphs were smoothed by averaging each position with that of the 20 bp on either side. Little variation is seen with the mock, with the different ChIP samples fluctuating more widely. On more highly occupied TU such as the *mcr* operon (Fig. 3d) or *sla* (Fig. 3b) RNAP occupancy is enriched well above the background throughout the TU. TUs such as the *rpo* operon or *tuf* gene reveal an RNAP occupancy comparable with the mock within intragenic regions, but above background proximal to the promoter (Fig. 3c,e). The initiation factors TBP and TFB have an occupancy profile more similar to that of the mock overall, with higher correlation shown when comparing overlapping windows across the genome, particularly for TFB (Supplemental Fig. 5b). When scrutinising individual loci this similarity to the mock is seen within the TU body, where TBP and TFB are not predicted to bind, while higher and specific occupancy is observed proximal to some but not all TSS (Fig. 2b-e).

Supplementary Table 1: Mapping statistics for RNA and ChIP samples.

Sample	Replicate	Total reads (millions)	Mapped reads (millions)*
RNA			
Total	1	29.7	27.0 (90.8%)
	2	33.1	30.5 (92.2%)
TEX-treated	1	16.5	5.9 (36.0%)† [15.5 (93.7%)]
	2	13.9	3.2 (22.7%)† [13.1 (94.3%)]
ChIP			
Input	1	13.0	12.8 (98.8%)
	2	2.3	2.3 (98.5%)
	3	1.0	1.0 (99.3%)
Mock	1	1.5	0.6 (41.0%)
	2	1.0	0.8 (81.3%)
TBP	1	16.8	16.1 (95.9%)
	2	1.0	0.9 (89.8%)
	3	2.8	2.6 (93.6%)
	4	0.8	0.6 (76.9%)
TFB	1	1.7	0.5 (30.6%)
	2	1.2	1.0 (86.4%)
Rpo4/7	1	12.0	11.6 (96.5%)
	2	2.6	2.1 (80.0%)
	3	1.9	1.8 (92.0%)
	4	0.5	0.4 (81.0 %)
Rpo3/11	1	1.6	1.2 (73.4%)
	2	1.7	4.5 (88.4%)
	3	0.9	0.7 (79.8%)
Spt4/5	1	2.7	2.2 (82.1%)
	2	1.6	1.5 (90.1%)
	3	1.2	0.8 (68.2%)

*Reads were mapped to Mja genome using Bowtie[19] allowing for no mismatches in the first 28 nt of the read. Reads that aligned to multiple locations were mapped to one position unless otherwise stated. †For these samples apparent low read mapping is due to filtering, mapped reads without filtering is shown in square brackets.

Supplementary Table 2 [separate file]: Identified TSS and their promoter elements.

Supplementary Table 3 [separate file]: Gene organisation of Mja.

Supplemental Table 4: Candidate new ORF and intergenic and antisense ncRNA.

Name	Coordinates	Strand	Amino acids	Notes
ORF				
Mj0002A	4456-4566	+	36	Possible transporter protein
Mj0156A	154062-154616	-	184	In antisense orientation to *cdhC2*/*mj0156*
Mj0272A	257664-257782	+	39	GCN5-related N-acetyltransferase. Contains two frame shifts
Mj0273A	258583-258717	-	44	Candidate ORF
Mj0356A	325298-325526	+	75	Conserved in other Methanocaldococcales species
Mj0356B	325522-325653	-	43	Candidate ORF.
Mj0360A	328313-328534	-	73	Candidate ORF
Mj0360B	328547-328672	-	41	Candidate ORF.
Mj0431A	387568-387708	-	46	Conserved in other Methanocaldococcales species
Mj0510A	451479-451724	-	81	Similarity to LAGLIDADG_3 superfamily protein
Mj0590A	524218-524505	-	95	HesB related selenoprotein.
Mj0892A	822725-823045	-	106	In antisense orientation to *flaB*/*mj0892*
Mj0992A	921336-921506	-	56	Candidate ORF
Mj1144A	1084636-1084873	+	79	Predicted membrane protein
Mj1223A	1166055-1166168	+	37	Candidate ORF
Mj1388A	1336438-1336890	+	83	Candidate ORF
MJECL08A*	7531-7740	+	69	Contains spoVT_AbrB domain
MJECL33A*	40060-40311	+	83	In antisense orientation to *mjecl33*.
MJECS05A†	6823-6948	-	41	Contains similarity to adenylate cyclase.
intergenic RNA				
mjpred20	489769-489966	+		Between *mj0533* (acylphosphatase-like protein) and *mj0554*
mjpred36	1150331-1150382	+		Between *tRNA-gly2* and *mj1207* (uncharacterized N-acetyltransferase)
mjpred42	1422318-1422474	-		Between *mj1451* and *mj1452*
mjeclpred03*	35351-35439	+		Between *mjecl28* and *mjecl29* (potential archaeal histone)
antisense RNA				
mjpred05	118265-119157	+		Antisense to *mj0122* (ribose1,5-bisphosphate isomerase)
mjpred07	124486-124998	-		Antisense to *mj0129*
mjpred13	324366-324697	+		Antisense to *mj0355*
mjpred14	350177-350265	-		Antisense to *cas8a2*/*mj0385* (CRISPR associated protein
mjpred22	591436-591853	+		Antisense to *mj0666* (putative molybdopterin biosynthesis protein)
mjpred29	873866-875199	+		Antisense to *mj0943*
mjpred32	986082-986116	+		Antisense to *cnr13* (candidate ncRNA identified by Schattner 2002)
mjpred35	1112085-1113698	+		Antisense to *pyrG*/*mj1174* (CTP synthase)
mjpred39	1200672-1200875	-		Antisense to *histone A3*/*mj1258* (potential archaeal histone)
mjpred15	1483247-1483518	-		Antisense to *mj1511*
***mjecspred01*†**	8708-11929	+		Antisense to *mjecs*08
***mjecspred02*†**	11930-12635	+		Antisense to *mjecs*09

*Encoded on large plasmid; †encoded on small plasmid.

Supplemental Table 5: TUs characterised by the alternative Spt4/5 recruitment mode.

Coordinates	Length	Associated gene	Length of region within TU[†]	Spt4/5[‡]	Rpo4/7[‡]	Rpo3/11[‡]
0-550	550	*CRISPR1*	258	0.88	1.85**	1.72*
29450-29750	300	*sR02* snoRNA	162	-0.01	0.90***	0.54*
49100-49400	300	*CRISPR2*	33	0.40	1.40**	1.08*
92100-92900	800	*CRISPR3*	449	-0.05	2.16**	1.62*
117750-118550	800	*cnr1*	431	0.42	1.82**	1.46***
132500-133350	850	*CRISPR4*	513	0.39	2.14***	1.76**
159250-160000	750	*rrnA* operon	337	1.04	2.24**	2.16**
236300-236600	300	*CRISPR5*	120	-1.03	0.43*	-0.24
352000-352950	950	*CRISPR6*	485	-0.05	3.72***	2.96*
438250-438500	250	*mj0496*	285	-0.20	0.92**	0.49*
471250-472050	800	*CRISPR9*	441	1.23	3.73***	2.97*
501050-501750	700	*CRISPR10*	277	0.15	1.32**	0.99*
506900-507600	700	*CRISPR11*	354	0.25	1.46***	1.01*
623850-624650	800	*CRISPR21*	487	0.48	2.25***	1.83***
637900-638550	650	*rrnB* operon	353	1.14	1.89*	1.92*
857950-858750	800	*CRISPR12*	412	-0.27	1.69***	1.13*
1034450-1035200	750	*CRISPR13*	398	-0.12	1.04**	0.72*
1049000-1049350	350	*CRISPR14*	-11	-0.12	1.09*	0.60
1219900-1220550	650	*CRISPR16*	350	0.85	1.81***	1.41*
1266650-1267350	700	*CRISPR17*	322	0.49	1.64*	1.26
1457200-1457600	400	*CRISPR18*	220	0.24	1.36*	0.81
1569950-1570700	750	*CRISPR19*	344	0.10	1.33**	0.91*
1575150-1575750	600	*CRISPR20*	316	0.06	1.01**	0.74*

[†] Distance from TSS of the TU to the end of the region of difference. Where multiple TSS for a single TU are identified the primary TSS is used.

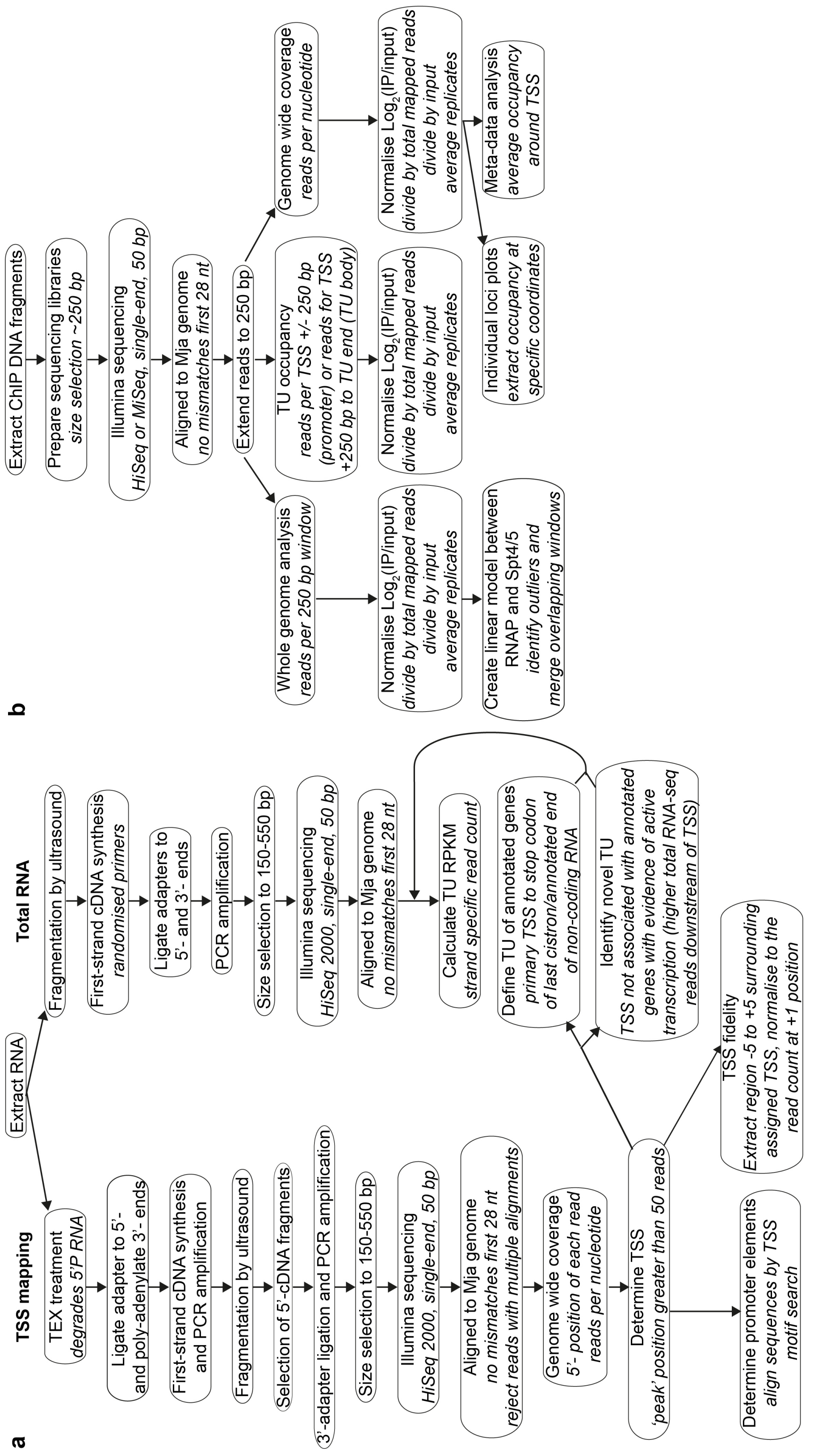
[‡] Normalised occupancy over regions as Log_2(IP/input) difference between Spt4/5 and RNAP, average of four (Rpo4/7) or three (Rpo3/11, Spt4/5) technical replicates. Significant differences to Spt4/5 occupancy for the RNAP subcomplexes by Welch's t-test followed by Benjamini Hochberg correction. Adjusted P value: * = <0.05; ** = <0.01; *** = <0.001 n = 3 (Rpo3/11 and Spt4/5) or 4 (Rpo4/7).

Supplemental Table 6: Oligonucleotides used in this study.

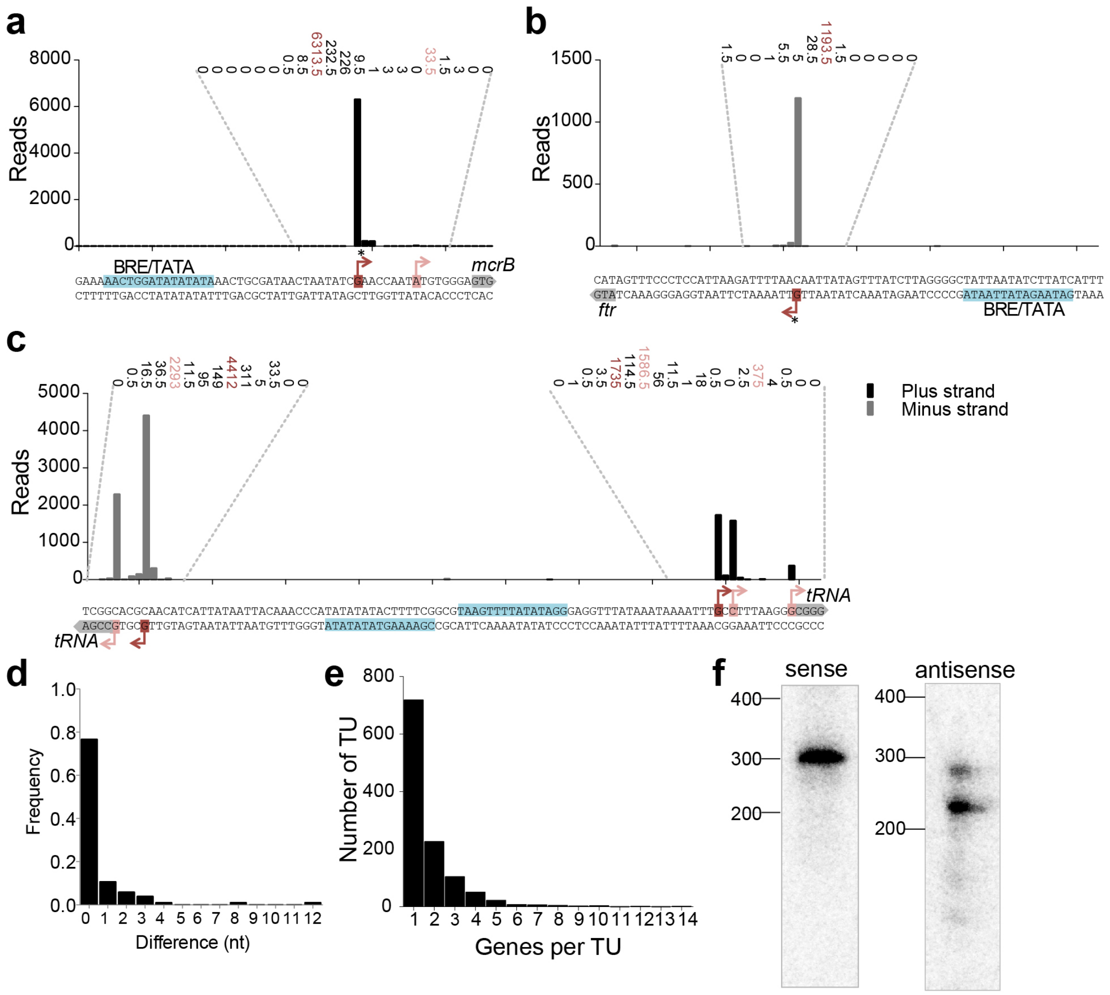
Name	Sequence (5’-3’)
T6 NTS	GATTGATAGAGTAAAGTTTAAATACTTATATAGATAGAGTATAGATAGAGGGTTCAAAAAATGGTT
T6 TS	AACCATTTTTTGAACCCTCTATCTATACTCTATCTATATAAGTATTTAAACTTTACTCTATCAATC
T6 bubble	AACCATTTTTTGAACCCTCCGCTTATACTCTATCTATATAAGTATTTAAACTTTACTCTATCAATC
T6 NTS TATA mut	GATTGATAGAGTAAAGTTTGCATACTTATATAGATAGAGTATAGATAGAGGGTTCAAAAAATGGTT
T6 TS TATA mut	AACCATTTTTTGAACCCTCTATCTATACTCTATCTATATAAGTATGCAAACTTTACTCTATCAATC
rrnA NTS	ATTGTTCTATTCCTAAAAACGTTGCATATAACAACCTCTCGTTATAGGATGCACTTGAGGGATGCGTCCCC
rrnA TS	GGGGACGCATCCCTCAAGTGCATCCTATAACGAGAGGTTGTTATATGCAACGTTTTTAGGAATAGAACAAT
rrnA bubble	GGGGACGCATCCCTCAAGTGTGCTCTATAACGAGAGGTTGTTATATGCAACGTTTTTAGGAATAGAACAAT
rrnA NTS TATA mut	ATTGTTCTATTCCTAAAAACGTTGCGGCCGGCAACCTCTCGTTATAGGATGCACTTGAGGGATGCGTCCCC
rrnA TS TATA mut	GGGGACGCATCCCTCAAGTGTGCTCTATAACGAGAGGTTGCCGGCCGCAACGTTTTTAGGAATAGAACAAT
CRISPR NTS	TTATTAAAGGGAGAAAATTTTAAATACTAAAAGATTTATATTATGAGATAGTTATTTATCCTTAGAAAAAT
CRISPR TS	ATTTTTCTAAGGATAAATAACTATCTCATAATATAAATCTTTTAGTATTTAAAATTTTCTCCCTTTAATAA
CRISPR bubble	ATTTTTCTAAGGATAAATAATCGCCTCATAATATAAATCTTTTAGTATTTAAAATTTTCTCCCTTTAATAA
CRISPR NTS TATA mut	TTATTAAAGGGAGAAAATGGCGGCCGCTAAAAGATTTATATTATGAGATAGTTATTTATCCTTAGAAAAAT
CRISPR TS TATA mut	ATTTTTCTAAGGATAAATAATCGCCTCATAATATAAATCTTTTAGCGGCCGCCATTTTCTCCCTTTAATAA
rpl12 NTS	TATTGTTCTAACCGAAAAATATAAATAACTAATTATAAATAGTGAATTGCAAACTCTACTTCAAATTAATA
rpl12 TS	TATTAATTTGAAGTAGAGTTTGCAATTCACTATTTATAATTAGTTATTTATATTTTTCGGTTAGAACAATA
rpl12 bubble	TATTAATTTGAAGTAGAGTTCATGATTCACTATTTATAATTAGTTATTTATATTTTTCGGTTAGAACAATA
rpl12 NTS TATA mut	TATTGTTCTAACCGAAAAATAGCGCTAACTAATTATAAATAGTGAATTGCAAACTCTACTTCAAATTAATA
rpl12 TS TATA mut	TATTAATTTGAAGTAGAGTTCATGATTCACTATTTATAATTAGTTAGCCGTATTTTTCGGTTAGAACAATA
rrnA fw	TTATTCAAATTATTGTTCTATTCCTAAAAACGTTGCATATAACAACCTCTCGTTATAGGATGGAGTTGAGGGATGatggaaggagaatataattgag
CRISPR TSS1 fw	TGTTTTATTAAAGGGAGAAAATTTTAAATACTAAAAGATTTATATTATGAGATAGTTATTTATatggaaggagaatataattgag
CRISPR TSS2 fw	AGAAAAATGTGGTGTAGAAAAGCTTAAATATTAGGAGAGTAGTATAAATTATATTGTGGATAAGatggaaggagaatataattgag
rpl12 fw*	CCCTATTGTTCTAACCGAAAAATATAAATAACTAATTATAAATAGTGAATTGCAAAGTGTAGatggaaggagaatataattgag
C-less rev	TCATTCACTCTCATCCCCTCTT
A3 sense	GAAGCTTTGGAAGAAATTGCCTTAGAGATAGCAAGTTTCCTGTCTC
A3 antisense	TGCTATCTCTAAGGCAATTTCTTCCAAAGCTTCAGTTTCCTGTCTC

For EMSA templates NTS oligonucleotides were hybridized with either the corresponding TS (for homoduplex templates) or “bubble” oligonucleotides (for heteroduplex templates). To generate in vitro transcription templates with Mja promoters fused to a synthetic C-less cassette, the respective fw* primers were combined with primer C-less rev to amplify the C-less cassette by PCR as described previously**.** C to G mutations introduced to generate a C-less cassette are underlined and sequences derived from the synthetic C-less cassette are shown in minor case.

Supplementary figures and legends

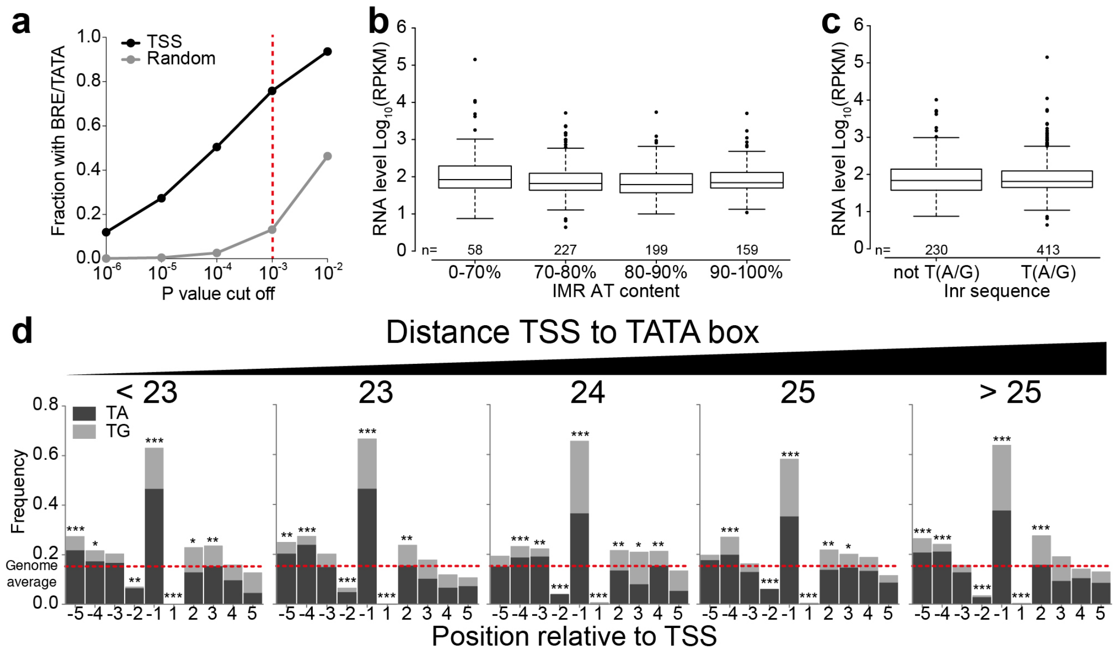

a
TSS mapping
Extract RNA
TEX treatment
degrades 5'P RNA
Ligate adapter to 5'- and poly-adenylate 3'- ends
First-strand cDNA synthesis and PCR amplification
Fragmentation by ultrasound
Selection of 5'-cDNA fragments
3'-adapter ligation and PCR amplification
Size selection to 150-550 bp
Illumina sequencing
HiSeq 2000, single-end, 50 bp
Aligned to Mja genome
no mismatches first 28 nt
reject reads with multiple alignments
Genome wide coverage
5'- position of each read
reads per nucleotide
Determine TSS
'peak' position greater than 50 reads
Determine promoter elements
align sequences by TSS
motif search
TSS fidelity
Extract region -5 to +5 surrounding assigned TSS, normalise to the read count at +1 position
Total RNA
Fragmentation by ultrasound
First-strand cDNA synthesis
randomised primers
Ligate adapters to 5'- and 3'- ends
PCR amplification
Size selection to 150-550 bp
Illumina sequencing
HiSeq 2000, single-end, 50 bp
Aligned to Mja genome
no mismatches first 28 nt
Calculate TU RPKM
strand specific read count
Define TU of annotated genes
primary TSS to stop codon of last cistron/annotated end of non-coding RNA
Identify novel TU
TSS not associated with annotated genes with evidence of active transcription (higher total RNA-seq reads downstream of TSS)
b
Extract ChIP DNA fragments
Prepare sequencing libraries
size selection ~250 bp
Illumina sequencing
HiSeq or MiSeq, single-end, 50 bp
Aligned to Mja genome
no mismatches first 28 nt
Extend reads to 250 bp
Whole genome analysis
reads per 250 bp window
Normalise Log$_2$(IP/input)
divide by total mapped reads
divide by input
average replicates
Create linear model between RNAP and Spt4/5
identify outliers and merge overlapping windows
TU occupancy
reads per TSS +/- 250 bp (promoter) or reads for TSS +250 bp to TU end (TU body)
Normalise Log$_2$(IP/input)
divide by total mapped reads
divide by input
average replicates
Individual loci plots
extract occupancy at specific coordinates
Genome wide coverage
reads per nucleotide
Normalise Log$_2$(IP/input)
divide by total mapped reads
divide by input
average replicates
Meta-data analysis
average occupancy around TSS

Supplementary Figure 1: Flow diagram outlining steps in RNA-seq (a) and ChIP-seq (b) analysis.



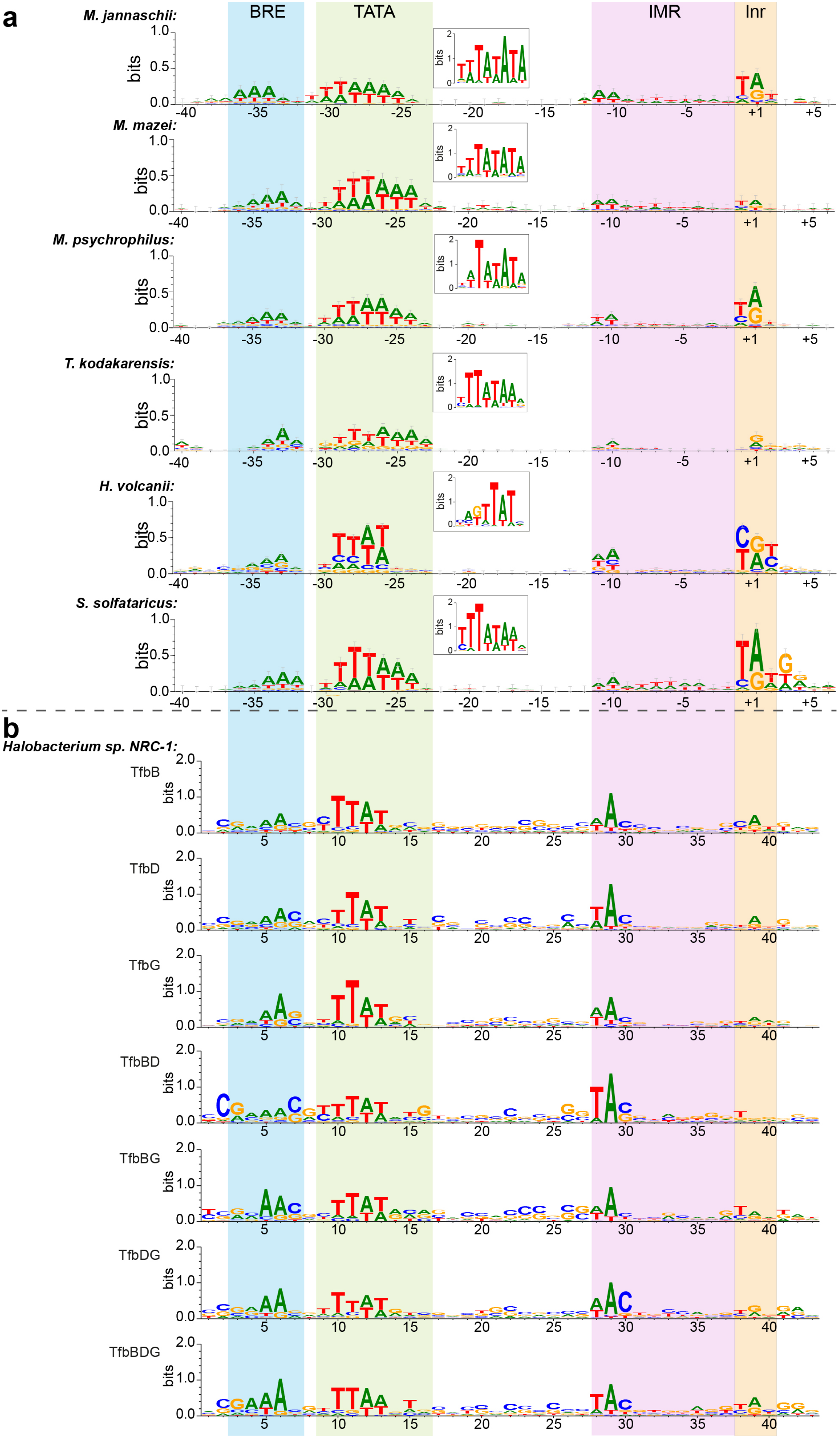
Supplementary Figure 2: TSS mapping and TU organisation in Mja. a-c, Representative TSS mapping of the *mrcB* (**a**), *ftr* (**b**) and two divergent tRNA promoters (**c**). For each 50 bp read of the terminator exonuclease-treated RNA sample the 5' end only was plotted give a histogram of reads per base across the genome. The, primary TSS is highlighted in dark red, secondary TSS in pink, and TSS confirmed in vitro[1] are marked with an asterisk. Coding and annotated ncRNA regions are highlighted in grey and identified BRE/TATA motifs are highlighted in blue. The read counts at each base position are indicated above the columns (average of two biological replicates). Note that some of the positions identified as secondary TSSs for tRNA genes are likely to correspond rather to processed RNA 5' ends. **d,**

Frequency plot comparing the position of mapped TSS to previously determined TSS[1], n = 103. **e,** Operon organisation in Mja. Frequency plot showing the number of genes per TU, n = 1114. **f,** Northern blotting confirms antisense transcription and indicates the sizes of both sense and antisense transcripts at histone A3 loci (Fig, 1c). A representative example of two biological replicates is shown (for additional replicate see Supplemental Figure 7).

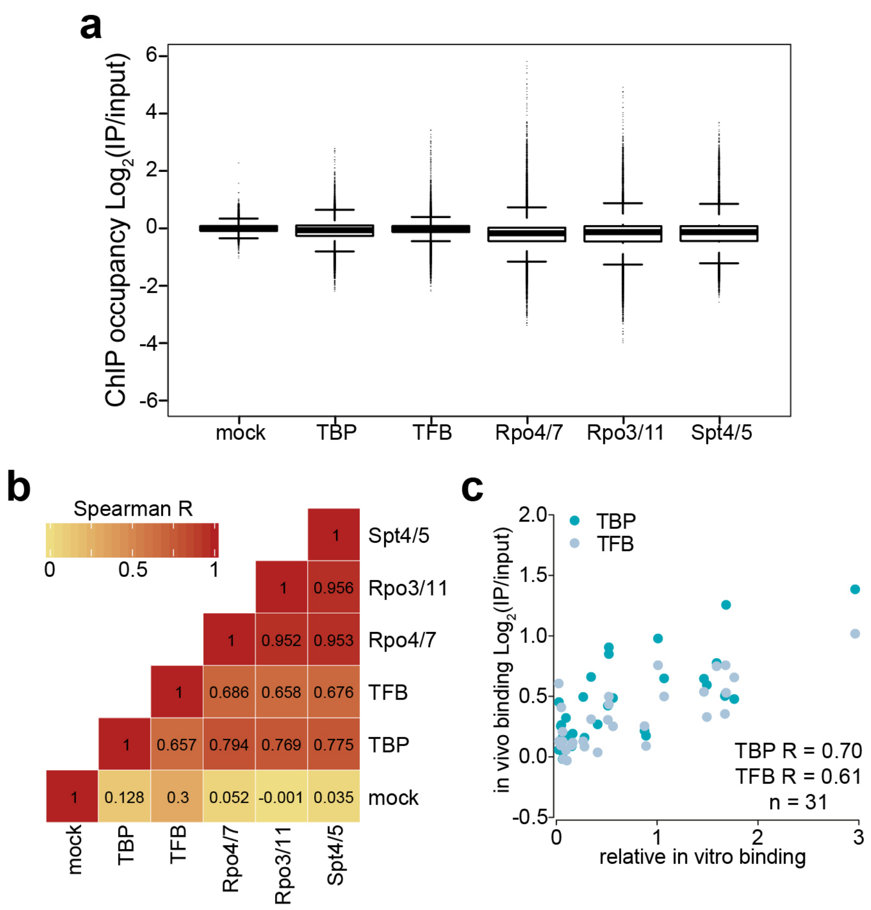


Supplementary Figure 3: BRE/TATA consensus, IMR base composition, and dinucleotide frequency surrounding the TSS. a, Specificity of the BRE/TATA motif prediction (Fig. 1e) in the AT-rich Mja genome. BRE/TATA motif confidence scores were calculated using the FIMO algorithm from the MEME suite[20,21]. The fraction of DNA sequences in which the BRE/TATA motif can be identified in the -50 to -15 relative to each TSS, compared to seven sets of randomly selected Mja sequences. Red dotted line indicates P value cut off chosen to include BRE/TATA. **b-c,** Sequence content of IMR (**b**) and Inr (**c**) has no effect on RNA levels. Distribution of RNA levels, as sense RPKM per TU for all TU with detectable transcript (average of two biological replicates), for the different sequence elements, individual n values indicated on graph, the whiskers indicate 1.5X the interquartile range. **d**, T(A/G) di-nucleotide frequency surrounding the TSS grouped according to the distances between the TSS and the 3'- end of the TATA box. Red line indicates genome average of 0.15, significance by Fisher's exact test. P value: * <0.05; ** <0.01;

*** <0.001, n = 157, 169, 307, 234 and 262 for TATA distances of <23, 23, 24, 25 and >25 bp respectively.


a
b
BRE
TATA
IMR
Inr
M. jannaschii:
M. mazei:
M. psychrophilus:
T. kodakarensis:
H. volcanii:
S. solfataricus:
bits
1.0
0.5
0.0
-40
-35
-30
-25
-20
-15
-10
-5
+1
+5
Halobacterium sp. NRC-1:
TfbB
TfbD
TfbG
TfbBD
TfbBG
TfbDG
TfbBDG
2.0
5
10
15
20
25
30
35
40

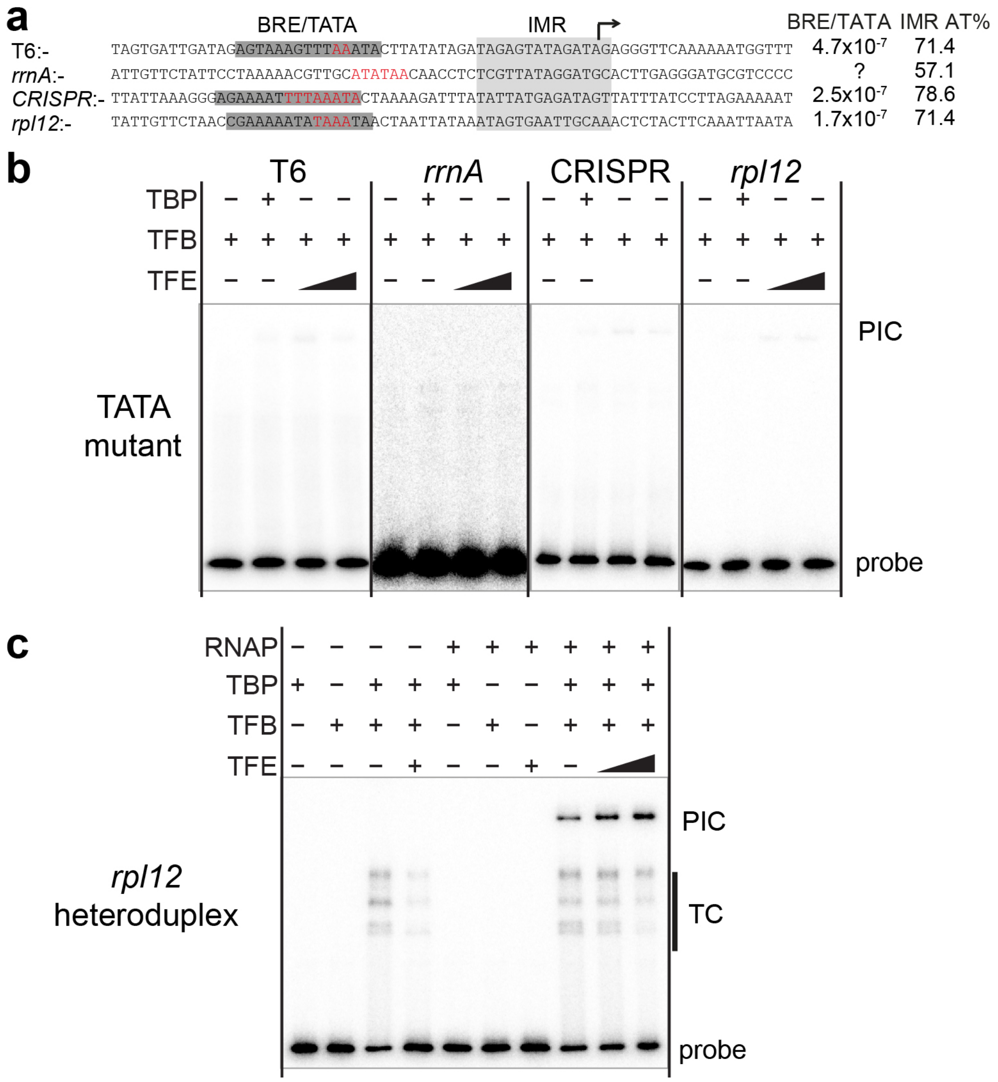
Supplemental Figure 4: Comparison of promoter elements for archaea. **a,** Alignment of the DNA sequences upstream of TSS identified on a genome-wide scale identifies individual promoter elements including the TFB recognition element (BRE), the TATA box, the initially melted region (IMR) and the initiator (Inr) surrounding the TSS. Alignment of primary TSSs identified by whole genome sequencing or *M. jannaschii* (our data), *Methanosarcina mazei*[9], *Methanolobus psychrophilus*[15], *Thermococcus kodakarensis*[13], *Haloferax volcanii*[16] and *Solfolobus solfataricus*[11]. Alignment visualised using WebLogo 3 adjusting to the background GC content for each organism (31.3% *M. jannaschii*, 41.5% *M. mazei*, 44.6% *M. psychrophilus*, 52% *T. kodakarensis*, 35.8% *S. solfataricus*,). Insert shows TATA box motif determined using MEME. Adapted from[22]. **b,** Alignment of motifs generated by ChIP-seq of different TFB variants B, D and G of *Halobacterium Sp. NRC-1* shows similar promoter features. Based on and with permission from the authors[23].



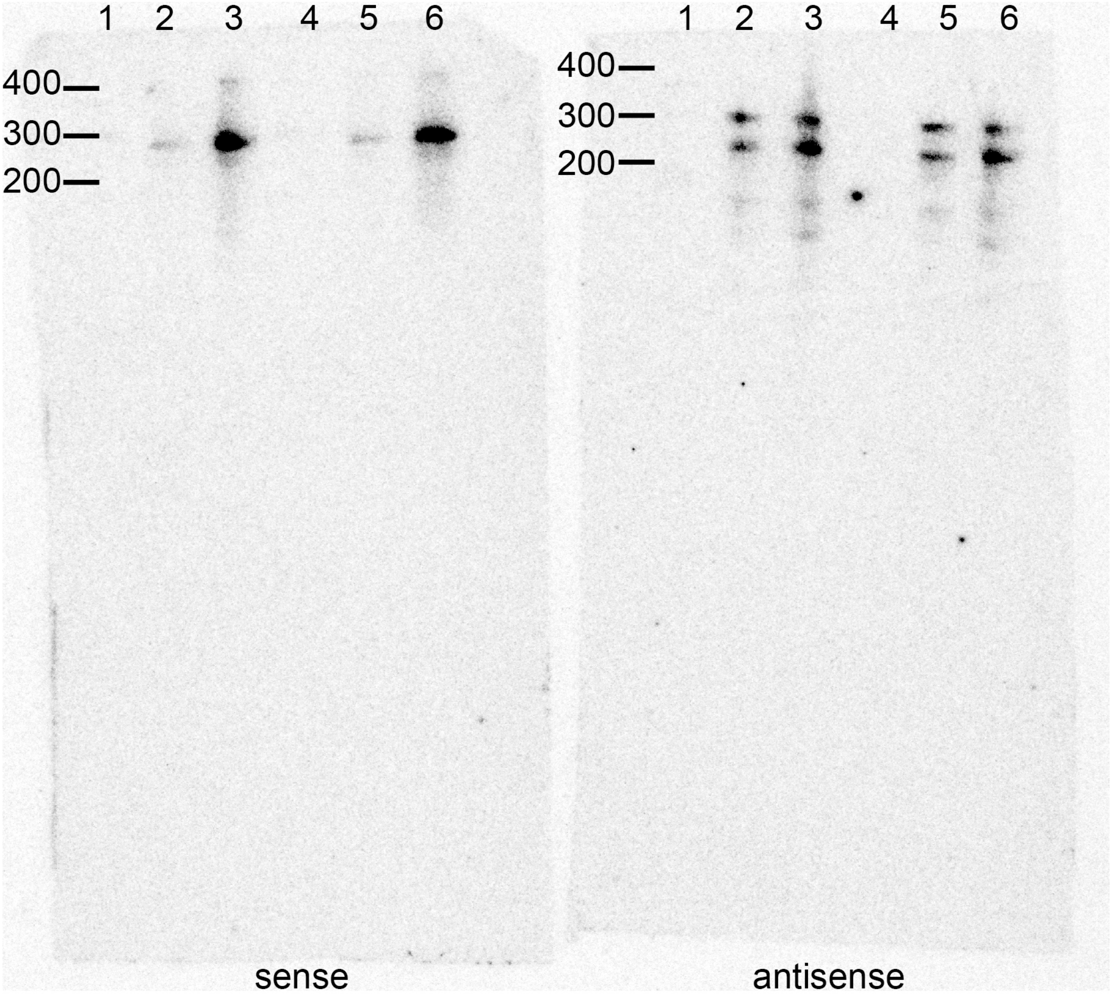
	mock	TBP	TFB	Rpo4/7	Rpo3/11	Spt4/5
Spt4/5						1
Rpo3/11					1	0.956
Rpo4/7				1	0.952	0.953
TFB			1	0.686	0.658	0.676
TBP		1	0.657	0.794	0.769	0.775
mock	1	0.128	0.3	0.052	-0.001	0.035

Supplemental Figure 5: Comparison of ChIP occupancy for different antibodies with mock control. a, The spread of occupancy for samples genome-wide. The genome was split into 50 bp overlapping windows of 250 bp and the occupancy per window calculated for each sample. Boxplot shows the distribution of occupancy for all windows (n = 34800, whiskers indicate 1.5X the interquartile range). **b**, Correlation between different ChIP samples genome-wide. Pairwise Spearman correlations performed between all samples on genome-wide windowed occupancy values in (**a**). P values to mock are all < 10^{-10} except Rpo3/11 where P value > 0.05. **c**, Correlation between ChIP occupancy at the promoter (in vivo binding) and relative in vitro binding determined by competition EMSA in[1] for TBP and TFB, Spearman Correlation indicated on graph TBP R = 0.7, P value = 1.1 x 10^{-5}, TFB R = 0.61, P value = 2.6 x 10^{-4}, n = 31. Panels a-c represent the average of four

(TBP, Rpo4/7), three (Rpo3/11, Spt4/5) or two (TFB, mock) technical replicates.


a
BRE/TATA
IMR
BRE/TATA
IMR AT%
T6:- TAGTGATTGATAGAGTAAAGTTTAAATACTTATATAGATAGAGTATAGATAGAGGGTTCAAAAAATGGTTT 4.7x10-7 71.4
rrnA:- ATTGTTCTATTCCTAAAAACGTTGCATATAACAACCTCTCGTTATAGGATGCACTTGAGGGATGCGTCCCC ? 57.1
CRISPR:- TTATTAAAGGGAGAAAATTTTAAATACTAAAAGATTTATATTATGAGATAGTTATTTATCCTTAGAAAAAT 2.5x10-7 78.6
rpl12:- TATTGTTCTAACCGAAAAATATAAATAACTAATTATAAATAGTGAATTGCAAACTCTACTTCAAATTAATA 1.7x10-7 71.4
b
T6
rrnA
CRISPR
rpl12
TBP
TFB
TFE
TATA mutant
PIC
probe
c
RNAP
TBP
TFB
TFE
rpl12 heteroduplex
PIC
TC
probe

Supplemental Figure 6: Effect of TATA box mutations on PIC formation in vitro. a, alignment of SSV T6, *rrnA*, CRISPR and *rpl12* promoters. The BRE/TATA motifs are shown in dark gray with P values indicated; the IMR is highlighted in light grey with AT% indicated. TATA box mutations used to abrogate TBP binding are indicated in red. **b**, EMSA showing PIC formation on promoter templates shown in (**a**). The TATA mutant templates were tested in the context of homoduplex (T6 promoter), and heteroduplex templates (the three Mja promoters). Contrast adjusted to aid visualisation. **c**, EMSA showing PIC formation on *rpl12* heteroduplex template. Ternary complex (TC) formation is dependent on both TBP and TFB. A representative example of two technical replicates is shown for both panels b and c.



Supplemental Figure 7: Full length image of Supplemental Figure 2f.

Northern blot autoradiogram probed against histone A3 in either the sense (left blot) or antisense (right blot) orientation. Indicated lanes are as follows: 1 - RiboRuler Low Range RNA Ladder (Thermo Fisher Scientific), now radiolabeled; 2 - biological replicate 1 (10 µg RNA); 3 - biological replicate 2 (25 µg RNA); 4-6 - technical repeats of lanes 1-3. Lanes 6 was used in Supplemental Figure 2f. The contrast was adjusted so that membrane boundaries become visible. Note that the markers bands are not visible on the autoradiogram but were visualised by methylene blue staining and the positions of marker bands are indicated on the autoradiogram.

Supplementary references

1 Zhang, J., Li, E. & Olsen, G. J. Protein-coding gene promoters in *Methanocaldococcus* (*Methanococcus*) *jannaschii*. *Nucleic Acids Res* **37**, 3588-3601, doi:10.1093/nar/gkp213 (2009).

2 Schattner, P. Searching for RNA genes using base-composition statistics. *Nucleic Acids Res* **30**, 2076-2082 (2002).

3 Schattner, P., Brooks, A. N. & Lowe, T. M. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res* **33**, W686-689, doi:10.1093/nar/gki366 (2005).

4 Klein, R. J., Misulovin, Z. & Eddy, S. R. Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proc Natl Acad Sci U S A* **99**, 7542-7547, doi:10.1073/pnas.112063799 (2002).

5 Grissa, I., Vergnaud, G. & Pourcel, C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**, 172, doi:10.1186/1471-2105-8-172 (2007).

6 Lillestol, R. K. *et al.* CRISPR families of the crenarchaeal genus Sulfolobus: bidirectional transcription and dynamic properties. *Mol Microbiol* **72**, 259-272, doi:10.1111/j.1365-2958.2009.06641.x (2009).

7 Pul, U. *et al.* Identification and characterization of E. coli CRISPR-cas promoters and their silencing by H-NS. *Mol Microbiol* **75**, 1495-1512, doi:10.1111/j.1365-2958.2010.07073.x (2010).

8 Hayes, W. S. & Borodovsky, M. Deriving ribosomal binding site (RBS) statistical models from unannotated DNA sequences and the use of the

RBS model for N-terminal prediction. *Pac Symp Biocomput*, 279-290 (1998).

9 Jäger, D. *et al.* Deep sequencing analysis of the *Methanosarcina mazei* Go1 transcriptome in response to nitrogen availability. *Proc Natl Acad Sci U S A* **106**, 21878-21882, doi:10.1073/pnas.0909051106 (2009).

10 Tang, T. H. *et al.* Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus*. *Mol Microbiol* **55**, 469-481, doi:10.1111/j.1365-2958.2004.04428.x (2005).

11 Wurtzel, O. *et al.* A single-base resolution map of an archaeal transcriptome. *Genome Res* **20**, 133-141, doi:10.1101/gr.100396.109 (2010).

12 Straub, J. *et al.* Small RNAs in haloarchaea: identification, differential expression and biological function. *RNA Biol* **6**, 281-292 (2009).

13 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, doi:10.1186/1471-2164-15-684 (2014).

14 Toffano-Nioche, C. *et al.* RNA at 92 degrees C: the non-coding transcriptome of the hyperthermophilic archaeon *Pyrococcus abyssi*. *RNA Biol* **10**, 1211-1220, doi:10.4161/rna.25567 (2013).

15 Li, J. *et al.* Global mapping transcriptional start sites revealed both transcriptional and post-transcriptional regulation of cold adaptation in the methanogenic archaeon *Methanolobus psychrophilus*. *Sci Rep* **5**, 9209, doi:10.1038/srep09209 (2015).

16 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, doi:10.1186/s12864-016-2920-y (2016).

17 Liu, W., Vierke, G., Wenke, A. K., Thomm, M. & Ladenstein, R. Crystal structure of the archaeal heat shock regulator from *Pyrococcus furiosus*: A molecular chimera representing eukaryal and bacterial features. *J Mol Biol* **369**, 474-488, doi:10.1016/j.jmb.2007.03.044 (2007).

18 Reichelt, R., Gindner, A., Thomm, M. & Hausner, W. Genome-wide binding analysis of the transcriptional regulator TrmBL1 in *Pyrococcus furiosus*. *BMC Genomics* **17**, doi:10.1186/s12864-015-2360-0 (2016).

19 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10** (2009).

20 Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**, W202-208, doi:10.1093/nar/gkp335 (2009).

21 Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-1018, doi:10.1093/bioinformatics/btr064 (2011).

22 Blombach, F., Smollett, K. L., Grohmann, D. & Werner, F. Molecular Mechanisms of Transcription Initiation-Structure, Function, and Evolution of TFE/TFIIE-Like Factors and Open Complex Formation. *J Mol Biol* **428**, 2592-2606, doi:10.1016/j.jmb.2016.04.016 (2016).

23 Seitzer, P., Wilbanks, E. G., Larsen, D. J. & Facciotti, M. T. A Monte Carlo-based framework enhances the discovery and interpretation of regulatory sequence motifs. *BMC Bioinformatics* **13**, 317, doi:10.1186/1471-2105-13-317 (2012).