Unbiased Map of Transcription Termination Sites in *Haloferax volcanii* reveals manifold termination patterns

Sarah J. Berkemer[1,2], Lisa-Katharina Maier[3], Fabian Amman[9,10], Stephan H. Bernhart[1,5], Julia Wörtz[3], Pascal Märkle[3], Friedhelm Pfeiffer[4], Peter F. Stadler*[1,2,6,7,8,9,11,12], and Anita Marchfelder*[3]

1 Bioinformatics Group, Department of Computer Science - and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany
2 Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany
3 Biology II, Ulm University, 89069 Ulm, Germany
4 Computational Biology Group, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany
5 Transcriptome Bioinformatics, Interdisciplinary Center for Bioinformatics, Leipzig University, Germany
6 German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig
7 Competence Center for Scalable Data Services and Solutions, and Leipzig, Research Center for Civilization Diseases, University Leipzig, Germany
8 Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia
9 Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Vienna, Austria
10 Division of Cell and Developmental Biology, Medical University Vienna, Schwarzspanierstraße 17, 1090 Vienna, Austria
11 Center for RNA in Technology and Health, Univ. Copenhagen, Grønnegårdsvej 3, Frederiksberg C, Denmark
12 Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501

* corresponding authors:
studla@bioinf.uni-leipzig.de, anita.marchfelder@uni-ulm.de

Abstract

Currently, little is known about transcription termination in archaea. Archaeal genomes are densely packed; thus, correct transcription termination is an important factor for orchestrated gene expression, since faulty termination, for instance, can result in aberrant transcription of downstream genes. A systematic analysis of transcription termination sites (TTS) using RNAseq data has hitherto only been performed for two archaea, *Methanosarcina mazei* and *Sulfolobus solfataricus*. In this study, only regions directly downstream of annotated genes were analysed, and thus, only a part of the genome was investigated. Furthermore, the analysis was performed on reads originating from a total RNA fraction, representing a mixture of 3´ ends resulting from transcription termination as well as from processing. Here, we developed a novel algorithm (Internal Enrichment-Peak Calling) that allows an unbiased, genome-wide identification of TTS independent of annotation. In addition, we treated the RNA fraction used for cDNA preparation with terminator exonuclease, to enrich original transcription termination ends. We employed this approach in the model archaeon *Haloferax volcanii* and identified 1,543 transcription termination sites. Approximately half of the termination sites were located in intergenic regions, and the remainder were found in coding regions. We determined distinct termination motifs for intergenic and coding regions. In intergenic regions, the termination signal consisted of stretches of Ts upstream of the termination, followed in most cases by a C at the termination site. Termination signals in coding regions differed, and they were prevalently associated with the motif AGATC, which was located directly downstream of the termination site. *In vivo* reporter gene tests of selected TTS confirmed termination at these sites. In several cases, more than one termination site was present downstream of a gene, resulting in transcripts with different lengths of the 3´ untranslated region (3´ UTR). The median 3´ UTR length was 97 nucleotides, which is longer than the median 3´ UTR length recently reported for the two archaea *M. mazei* and *S. solfataricus* and the bacterium *Bacillus subtilis*.

Introduction

Archaeal RNA synthesis is generally considered to be more closely related to transcription in eukaryotes than to bacterial transcription. The archaeal RNA polymerase is similar to the eukaryotic RNA polymerase II, and the basal promoter elements in Archaea are similar to their eukaryotic pendants (TATA box and BRE); general transcription factors TBP (TATA binding protein), TFB (transcription factor B) and TFE (transcription factor E) resemble the eukaryotic proteins TBP, TFIIB and TFEII, respectively (for a review see: Fouqueau et al.[1]). Transcriptional regulators, however, seem to be more similar to those in bacteria[2]. Thus, the archaeal transcription machinery consists of a mixture of bacterial-like and eukaryotic-like components. Whereas some data have been reported on transcription initiation and elongation in archaea, very little is known about transcription termination. Controlled transcription termination is important to avoid aberrant RNA molecules and to help with RNA polymerase recycling. Generally, the genes in archaeal chromosomes are densely packed, so that proper termination is also important to prevent transcription from continuing into downstream genes. The process of transcription termination is not trivial because the very stable transcription elongation complex must be destabilised and dissociated during termination. In bacteria, two major classes of termination signals have been described: intrinsic termination and factor-dependent termination[3,4]. Intrinsic termination occurs either at a stretch of Ts or at hairpin structures that fold in the newly synthesised RNA; both trigger dissociation of the elongation complex. Factor-dependent termination occurs upon interaction with a specific protein such as the bacterial termination factor Rho[5]. Protein factor-assisted termination is especially important in regions where strong selective pressure on the DNA sequence does not allow encoding of intrinsic termination signals. This may be the case when termination must occur in the coding region of the downstream gene. In eukaryotes, several RNA polymerases synthesise the different RNA classes, and each polymerase has different modes of termination. RNA polymerase I requires protein factors for termination[6], whereas RNA polymerase III terminates efficiently and precisely at a stretch of Ts[7]. Termination of RNA polymerase II is quite complex, involving modification of the RNAP as well as interactions with additional protein factors, and seems to be coupled to co-transcriptional RNA processing[8]. Whereas archaeal transcription initiation resembles eukaryotic RNA polymerase II initiation[9], transcription elongation and termination seem to be more similar to the eukaryotic RNA polymerase III pathway, which is independent of RNA secondary structures and protein co-factors[10].

Compared to the determination of transcription start sites, the identification of termination sites is more complex. Termination is often leaky and encompasses several consecutive sites, and degradation by exonucleases renders the 3´ ends heterogeneous and less clear. Data reported on archaeal termination so far show that intrinsic termination occurs with a run of Ts[10-14] and is potentially also influenced by secondary structure elements[10,14]. Factor-dependent termination was predicted based on the results of an *in vivo* reporter assay in the archaeon *Thermococcus kodakarensis*[15]. A recent study confirmed this hypothesis, reporting the discovery of the first archaeal termination factor[16].

Recently, the termination sites for two archaea (*Sulfolobus solfataricus* and *Methanosarcina mazei*) were investigated systematically using RNAseq data[17]. The applied algorithm analysed only regions directly downstream of annotated genes, and thus only a part of the genome was covered. Transcription termination sites were found for 707 and 641 transcriptional units in *S. solfataricus* and *M. mazei*, respectively. For more than a third of the genes analysed, multiple consecutive terminators were identified, resulting in 3´ untranslated regions (UTRs) with different lengths[17]. In some cases, the terminator of an early gene in an operon was shown to be located in the downstream gene, allowing gene-specific regulation within an operon. The study also revealed lineage-specific features of termination for both archaea, confirming the requirement that more data from different archaeal organisms are required to learn more about transcription termination in this domain.

Here, we describe the transcription termination sites of the halophilic model archaeon *Haloferax volcanii. H. volcanii* has been used for a plethora of biological studies[18,19], including the determination of nucleosome coverage[20] and a genome-wide identification of transcription start sites (TSS)[21]. *H. volcanii* requires high salt concentrations for optimal growth, and due to the high intracellular salt concentrations, RNA-protein interactions -including modes of transcription termination- may differ from those in mesophilic archaea.

We used a new algorithm to identify transcription termination sites (TTS) genome-wide in an unbiased manner, independent of annotation and on the basis of reads enriched for original transcription termination ends. Applying this algorithm to RNAseq data, we found 1,543 TTS for the *Haloferax* genome. Sequences around the termination sites were analysed to detect primary and/or secondary structure motifs as termination signals. The length and distribution of 3´ UTRs was determined, operon composition was obtained and selected termination motifs were confirmed using an *in vivo* reporter gene system.

Results

Identification of transcription terminators downstream of annotated regions

To determine TTS, RNA was isolated from *H. volcanii* cells (from three biological replicates) and cDNA libraries were generated to allow RNA 3´ end determination. Libraries were subjected to paired-end next-generation sequencing (NGS), resulting in 49 million reads for each library on average. 89-98% of the reads obtained mapped to the *Haloferax* genome (Supplementary Table 10).

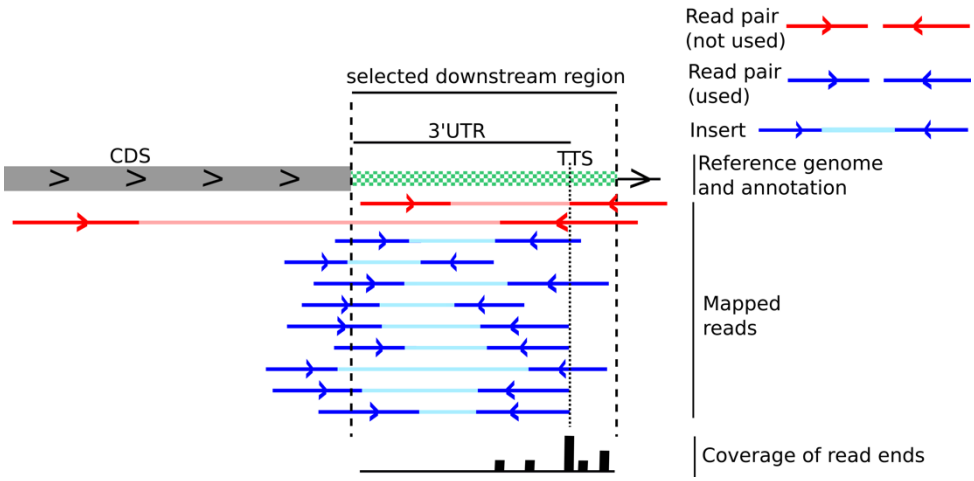

Figure 1. Principle of the Dar-Sorek-Method. As a first step in the DSM method, read pairs with an insert overlapping an annotated region are selected (red and blue lines)[17]. Inserts that do not overlap or have a length of more than 500 nucleotides are discarded (red read pairs in the Figure). From the selected read pairs (blue), the average length over all inserts is calculated as described in Dar et al.[17]. This value is used to determine the length of the selected downstream region (green white region in the Figure). The coverage of all read ends in this region is retrieved (bottom line), and the position with the highest coverage is identified as the TTS.

Mapped reads were initially analysed using a self-implemented version of the method described by Dar et al.[17], which will be referred to as the Dar-Sorek-Method (DSM). This method identifies TTS in a defined region downstream of annotated genes at the position with the highest coverage of mapped read ends. The length of the downstream region is determined by the average insert length of corresponding paired-end reads (Figure 1). In our dataset, the median length of the analysed region was 126 bp. The resulting 3´ UTRs were mostly shorter than 100 nucleotides, with a median length of 58 nucleotides (Supplementary Figure 1). The length restriction given in the DSM approach might be too strict for genes with long 3´ UTRs. In addition, the method cannot determine TTS independent of an annotation (see also the paragraph below "Comparison between DSM and IE-PC algorithms"). Using DSM, we identified 3,155 termination sites for the complete *Haloferax* genome of which 85% were in intergenic and the remainder in coding regions (Supplementary Table 7 and 11). A typical termination site is shown in Supplementary Figure 2 for the *pilA2* gene (HVO_2062).

New approach for the identification of transcription termination sites

To determine the TTS the DSM approach used reads from a total cellular RNA fraction, that contains RNA 3´ ends derived from transcription termination as well as 3´ ends derived from processing. Thus the 3´ ends identified by DSM are not all TTS but also processing sites (PS). A similar problem exists for the determination of original transcription 5´ ends, where a well established method for the reliable identification of transcription start sites (TSS) has been developed, termed differential RNAseq (dRNAseq). Here, data derived from an RNA sample treated with terminator exonuclease (+TEX) are compared with data obtained from an untreated sample (-TEX)[22,23]. TEX treatment of an RNA sample enriches primary transcripts containing the original 5´ triphosphate end. Comparison of data from the +TEX sample with the -TEX sample helps to identify TSS. We used this dRNAseq approach to determine original termination ends. RNA still containing their original 5´ end have a higher probability to also still contain their original 3´ end. Therefore, we isolated -in addition to the total RNA fraction already obtained- a second fraction that was treated with 5´ terminator exonuclease (TEX) to enrich primary transcripts and thereby original termination ends. After cDNA library generation from the TEX treated RNA, NGS was performed, resulting in an average of 40 million reads for each of the three libraries. Reads obtained were analysed with our newly established algorithm as described below.

Development of a novel algorithm, "Internal Enrichment-Peak Calling", as a tool to identify transcription termination sites genome-wide

DSM analysis only includes sequences downstream of annotated genes and, thus, only a fraction of the genome[17]. To overcome this restriction, we developed a novel approach to interpret the RNAseq data obtained that we termed Internal Enrichment-Peak Calling (IE-PC). To this end, we utilised the nature of the fragmentation process in the course of library preparation, which leads to an enrichment of unprocessed 3´ fragment ends over 5´ ends of fragments that are generated via random fragmentation during library preparation. Since we used the latter as the normalisation background for the former, we called this method internal enrichment (IE) (Figure 2). This approach requires mate-pair sequencing data, from which a set of read-pairs is selected that contains the same original first-strand cDNA 5´ ends (which reflect the RNA 3´ end) (Figure 2). The corresponding 3´ ends of the mate-pair reads are then evaluated. Sequencing the cDNAs generated from these first-strand cDNA fragments in paired-end mode preserved information about fragment ends. After mapping the read-pairs to the reference genome, the hallmark of an original RNA 3´ end was its high coverage of read ends, with its associated mate ends originating from a multitude of genomic sites. It is highly unlikely that independent clones end at an identical position (Figure 2). Multiple fragments with a heterogeneous 5´ end but a common 3´ end thus are indicative of likely TTS candidates.

However, this method cannot distinguish between original termination sites and RNA processing sites. Therefore, we used here the data obtained from the +TEX fraction. First, reads from the TEX RNA fraction (+TEX), containing enriched original

transcription terminated 3´ ends were used to reliably call TTS. In a second analysis reads from the untreated TEX fraction (-TEX) that contained RNA 3´ ends from transcription termination as well as from processing, was investigated.

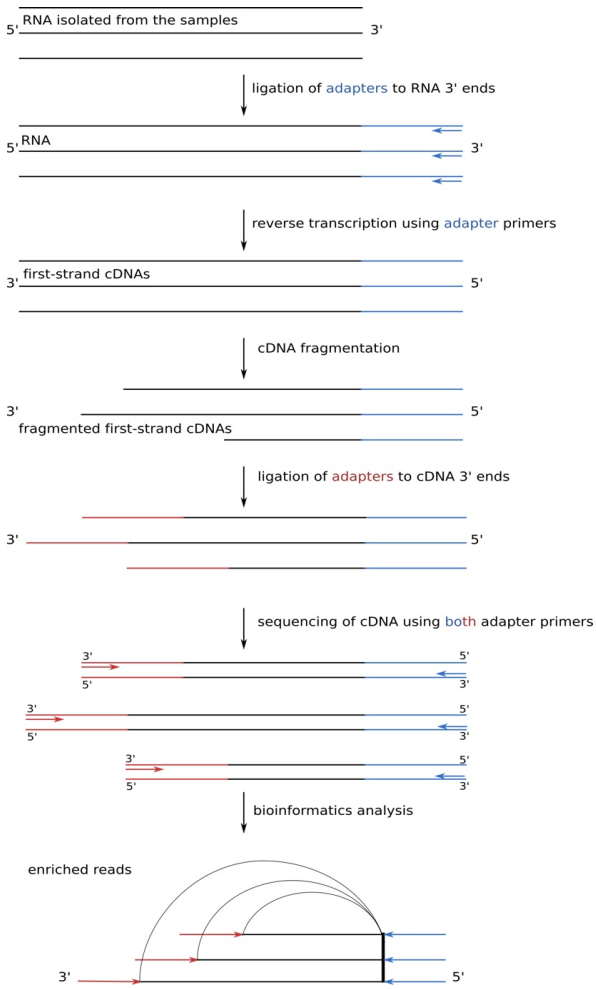

Figure 2. Principle of internal enrichment. After RNA isolation, adapter primers were ligated to the RNA 3´ end and the RNA was reverse transcribed. First-strand single-stranded cDNA was fragmented prior to the addition of the adapter primer at the cDNA 3´ end (for details, see Materials and Methods and Supplementary Methods). The break points were considered to be random, leading to an enrichment of original RNA 3´ ends over 5´ fragmentation ends. Sequencing the cDNAs generated from these first-strand cDNA fragments in paired-end mode preserved information about fragment ends, even if they were longer than the read length. After mapping the read-pairs to the reference genome, the hallmark of an original 3´ end was its high coverage of read ends, with its associated mate ends originating from a multitude of genomic sites.

A sudden decrease in read coverage was indicative of a potential TTS (found by a peak calling approach we termed PC) (Figure 3). The two methods (IE and PC) were sequentially run on the data, and only sites that were found by both approaches, IE as well as PC, were considered to be bona-fide TTS. We allowed a maximal distance of 10 nucleotides when computing overlapping sites for IE and PC. The advantage of this algorithm is that it is independent of genome annotation and thus analyses the complete genome sequence rather than a restricted region allowing the identification of all TTS of a genome.

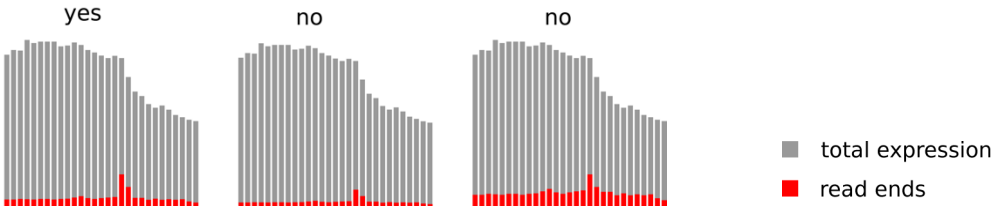

Figure 3. Principle of the applied peak calling procedure. In a sliding window approach, positions where the number of read stops (red) exceeded the mean number of read stops over the whole window by a z-score above the threshold 2 were treated as potential endings (left), while potential peaks of the same height but below a z-score of 2 were discarded (right). In addition, if the number of read stops was below a 10% threshold relative to the total coverage (grey), the peak was discarded (centre).

Identification of original transcription termination sites and processed sites

Application of the IE-PC algorithm to the data from the TEX treated samples consisting of enriched original transcription termination sites identifies 1,543 TTS (Table 1), whereas 6,284 RNA 3´ ends are identified when using the -TEX data set. 1,220 TTS are found in both data sets.

Data sets	TTS (+TEX data)	TTS and PS (-TEX data)	TTS found in both fractions
intergenic	807	2,856	670
coding	736	3,428	550
total	1,543	6,284	1,220

Table 1. TTS and PS identified with IE-PC. The IE-PC algorithm was run on two data sets: data from TEX treated samples (column TTS) as well as on data from untreated RNA fractions (column TTS and PS). Most of the original TTS identified in the +TEX data are also present in the data set from the untreated RNA fractions (column TTS found in both fractions).

Clustering of termination signals

Inspection of the TTS obtained revealed closely spaced TTS, that were 11 to 150 nucleotides apart. These closely spaced TTS were subclassified into primary TTS and non-primary TTS: a primary TTS is located directly downstream of a 3´ gene end on the same strand (Figure 4A & B); a non-primary TTS is located downstream of another TTS, with no other features (like TSS or 3´ gene end) in between, as shown in Figure 4C. For the +TEX samples we found in total 1,056 primary TTS and 487 non-primary TTS (data for the -TEX samples are in Supplementary Table 12). We also found very closely spaced TTS that were less than 10 nucleotides apart , these were assumed to reflect stuttering of the RNA polymerase. Thus, this was reported as only a single TTS (the one with the highest coverage). It should be noted that stuttering may happen several times, so that the overall length covered by such a termination region may significantly exceed 10 nucleotides.

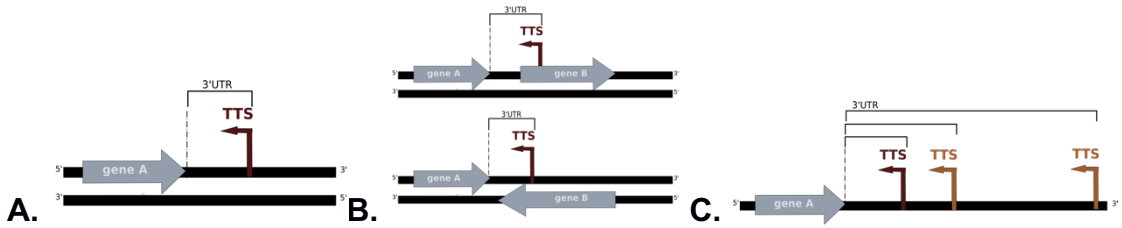

Figure 4. Location of TTS. A. TTS is located in an intergenic region. **B.** Location of the TTS in an annotated gene that is located on the sense or antisense strand. The length of the UTR was in all cases measured as the distance between the TTS and the 3´ end of the upstream annotated gene on the same strand. **C.** A primary (dark brown) and two non-primary TTS (light brown) are shown.

Comparison between DSM and IE-PC algorithms

Our newly established tool IE-PC is able to determine TTS covering the complete genome and thus can identify all TTS of an organism. Comparison of signals obtained with both methods can only be done with the regions that are included in the DSM analysis. The original DSM analysis was based on regions with a median length of 126 nucleotides downstream of an annotated 3´ end, with the TTS located at the position with the highest coverage. In addition, the DSM analysis is based on -TEX data, thus for comparison IE-PC data with the -TEX set were used. We further compared individual TTS found with DSM and/or IE-PC in downstream regions defined by the DSM analysis in more detail. Examples are shown in Figures 5 and 6, Table 2 lists the number of TTS that were identified by both methods (allowing for discrepancy of up to 10 nucleotides) as well as TTS that appeared in only one of the data sets. Altogether, we found 1,664 TTS being present in both data sets (Table 2).

chromosome	IE-PC only	overlapping	DSM only
main	3,184	1,296	1,128
pHV1	231	53	39
pHV3	394	97	125
pHV4	811	218	199
total	**4,620**	**1,664**	**1,491**

Table 2. Comparison of TTS found with DSM and IE-PC in downstream regions covered by DSM. The column "IE-PC only" lists all TTS identified by IE-PC but not with those identified by DSM method. TTS identified with IE-PC that were also found with DSM are listed in the column "overlapping". Sites that were detected in the defined region only by DSM and that did not overlap with IE-PC sites are listed in the column “DSM only”. The IE-PC data shown here are the ones calculated on the basis of the -TEX data set, since the DSM data are also based on the -TEX data.

Comparing the DSM data with IE-PC results as well as with given coverage data, we see that chosen downstream regions in DSM were frequently too short. This is also clear when comparing the median 3´ UTR length determined by both methods. While DSM found a median 3´ UTR length of 58 nucleotides, IE-PC determined it to be 96

nucleotides (for primary TTS in the -TEX data set). A comparison of TTS found with DSM and IE-PC for the HVO_1876s gene is shown as an example (Figure 5). The lower panel (DSM) in Figure 5 shows the results obtained with DSM, coverage is shown as the 3´ end coverage of corresponding reads, the TTS position for the transcript is also shown.

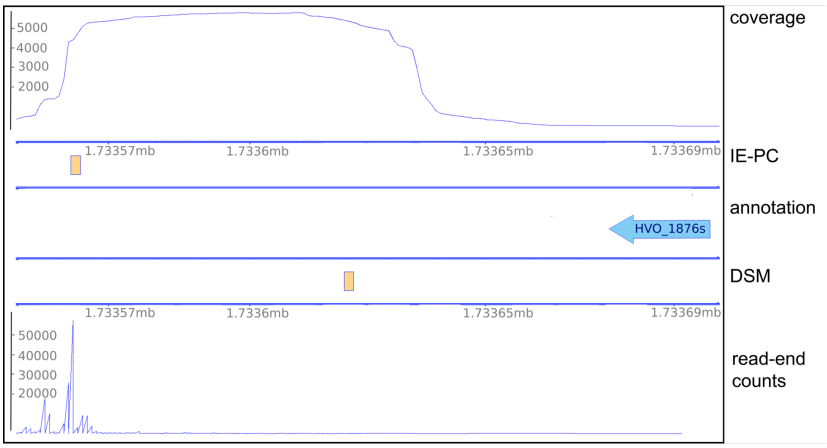

Figure 5. TTS comparison for DSM and IE-PC analyses for the transcript HVO_1876s. The data in the lower panel (DSM) show the DSM results with TTS and corresponding read end counts from the DSM data set. The upper panel (IE-PC) shows the TTS locations determined by IE-PC and the total coverages, corresponding to the -TEX data set. Since sequencing starts at the 3´ end, coverage starts at the 3´ end and runs continuously for 75 bp due to the read length. The annotation and genome coordinates are shown in the middle (coordinates given in Mb). TTS are shown as orange rectangles.

The upper panel (IE-PC) shows IE-PC results, with TTS location and read coverage above. The drop in coverage values is clearly visible, matching the TTS location as identified by IE-PC. Corresponding read end coverage of the DSM data set was present at the same position. However, in the DSM analysis, every gene had a downstream region with an individual length, since for each gene the average insert length of the corresponding reads was calculated. The length for the downstream region of HVO_1876s, was calculated with only 64 nucleotides, and therefore the high read end coverage was beyond the region selected for analysis. As the algorithm is bound to report a TTS (there is always a position with highest coverage in the selected region), DSM reported a false positive in this case.

For the genes *trmY* and $tRNA^{Pro}$, both methods identified one identical TTS position and two different TTS positions (Figure 6). The IE-PC analysis identified two TTS downstream of the $tRNA^{Pro}$ gene (upper panel). The corresponding coverage values for DSM (lower panel) showed similar signals, but due to its specific algorithm, DSM assigned one TTS downstream of the *trmY* gene (HVO_1989) and one downstream of the $tRNA^{Pro}$ gene (lower panel).

Taken together comparison between the DSM approach and the IE-PC algorithm clearly shows the limitation of the DSM approach.

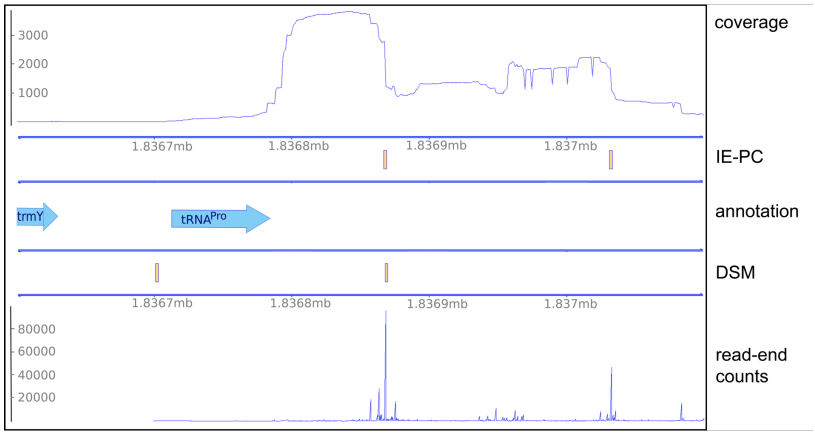

Figure 6. TTS comparison of DSM and IE-PC for the tRNA^{Pro} gene. In the lower panel, termination sites and corresponding read end coverages from the DSM data set are shown, assigning a TTS downstream of the *trmY* gene and another one downstream of the tRNA^{Pro} gene (shown as orange rectangles). The upper panel shows the TTS location determined by IE-PC and the total coverages, corresponding to the -TEX data set, identifying two TTS downstream of the tRNA^{Pro} gene. Since sequencing starts at the 3´ end, coverage starts at the 3´ end and runs continuously for 75 bp due to the read length. The secondary TTS of the tRNA^{Pro} gene was not reported by the DSM algorithm as this algorithm systematically reports only a single TTS for each annotated gene.

The *Haloferax* genome contains 1,543 transcription termination sites

As expected, the IE-PC algorithm employed on the +TEX data set identified less TTS as with the -TEX data set and less than DSM, that was also run with a -TEX data set. Altogether, 1,543 transcription termination sites were found (Table 1). Slightly more than half of the sites were found in intergenic regions (807 TTS, 52 %) and the remainder in coding regions (736 TTS, 48%). Detailed information for each TTS can be found in Supplementary Table 1. The following results are based on the small, conservative analysis of the +TEX data set, that contains enriched TTS, if not stated otherwise. Supplemental table 1 lists TTS detected for each data set.

Analysis of the regions up- and downstream of the TTS were performed separately for TTS located in coding and intergenic regions (Figure 7), and in both, an increase in hybridisation energy at the TTS similar to the increase identified in the TTS set obtained with DSM was found (Supplementary Figure 3). The pattern of nucleotide enrichment for sites located in intergenic regions showed that Ts were prevalent at the TTS (Figure 7A). The analysis was also done with the -TEX data as shown in Supplementary Figure 4.

We next analysed sequences 15 nucleotides up- and five nucleotides downstream of the 807 intergenic termination sites for common sequence motifs. For 748 sites, we found similarities in the sequences, such as a conserved dinucleotide TC as part of the motif as well as a stretch of T's of variable length upstream of the termination site (Figure 8A, Supplementary Table 2).

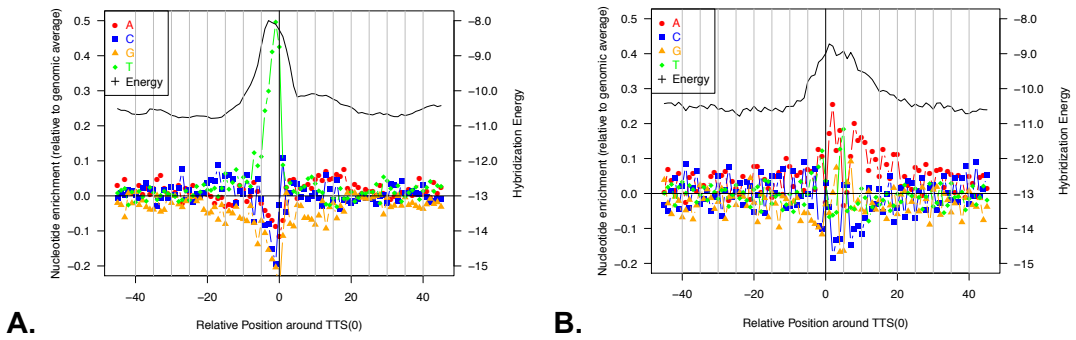

Figure 7. Analysis of up- and downstream regions for TTS identified with IE-PC. A. Intergenic and B. coding regions were investigated. Forty-five nucleotides up- and downstream of the termination site were analysed for (1) nucleotide enrichment at each position (left y-axis) and (2) the hybridisation energy (right y-axis). x-axis: nucleotide position (upstream -, downstream +). The colour scheme for the four nucleotides is shown at the upper left, and the energy data are shown with a black line. Hybridisation energies were calculated based on the binding energies between the DNA template and RNA in the area behind the RNA polymerase. The regular pattern in panel B is due to the statistical characteristics of coding regions in GC-rich organisms[24,25].

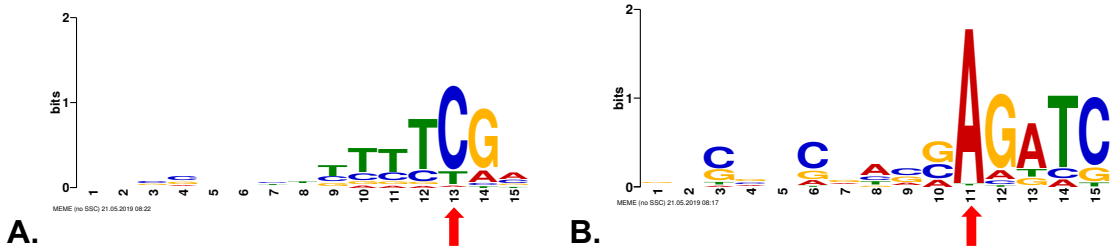

Figure 8. Enriched sequence motifs around the TTS identified with IE-PC. A. Sequence motif close to TTS located in intergenic regions. The TTS is located at position 13 (red arrow). **B.** Sequence motif close to TTS located in coding regions. The TTS is located at position 11 (red arrow) . Motifs were detected using MEME [26].

Sequences 15 nucleotides up- and five nucleotides downstream of the 736 TTS in coding regions were likewise investigated for common sequence motifs (Figure 8B). The prominent C residue at every third position is typical for coding regions in *Haloferax*. The third codon position combines an enrichment for GC (due to the GC-rich genome) and pyrimidines. For 456 TTS in coding regions, we found the downstream motif AGATC (Figure 8B). The analysis was also done with the -TEX data (Supplementary Table 3).

Taken together, we could identify distinct termination motifs that were specific for intergenic and for coding regions.

<u>Secondary structures can act as termination signals</u>

To identify potential secondary structure motifs, we conducted a search for accessible and inaccessible regions around the TTS found with the +TEX data using RNAplfold[27] (Vienna RNApackage[28,29]). We plotted accessibilities against a background distribution

of shuffled dinucleotides. However, no clear signals were found to indicate significantly increased or decreased accessibility. In a second attempt to identify secondary structures, we applied graphclust 2.0[30] within Galaxy[31] to sequences 100 nucleotides upstream of all TTS (Supplementary Figure 5). Graphclust is a tool that clusters input sequences based on their secondary structure(s). It will cut the sequences based on a window size parameter and align and fold the sequences into secondary structures using RNAalifold of the ViennaRNA package[28,29]. Using this approach we found hairpin structures upstream of 503 of the TTS (up to 10 nucleotides distance), an example is shown in Figure 9.

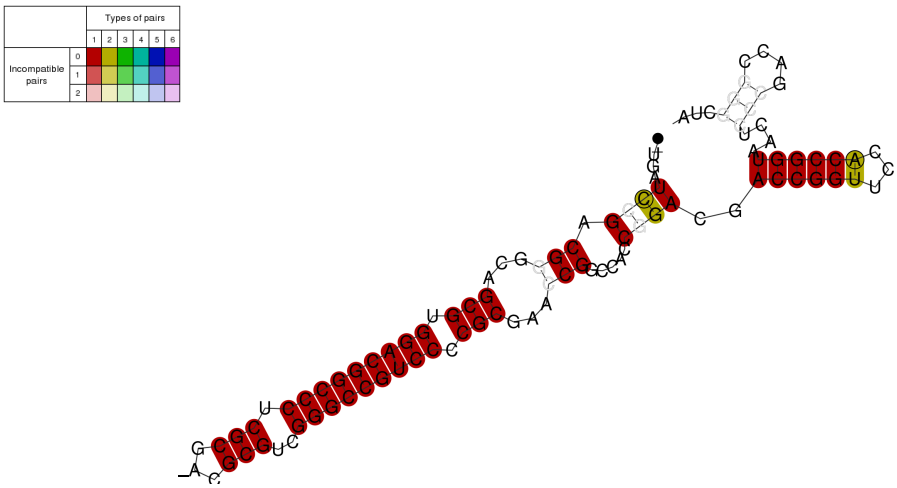

Figure 9. Hairpin structure found upstream of TTS. The secondary structure was plotted with RNAalifold, and it's 5' end is denoted by a small dot at the end of the line. The colour coding is at the top left; whereas darker colours show compatible pairs, and the number of types of pairs shows how many types are found at this position. Dark red colours indicate base pairs that are compatible and conserved.

Thus, in some cases secondary structures are present that might influence transcription termination.

Experimental confirmation of selected termination signals

We selected four TTS identified with our new IE-PC approach in intergenic and coding regions to test their termination activity with an *in vivo* reporter gene assay. Termination regions (the TTS and up- and downstream regions) were cloned into a fragment of the reporter gene β-galactosidase (sequences are listed in Supplementary Table 9). Termination activity was monitored with northern blot analyses (Figure 10, Supplementary Table 13). If termination occurred in the inserted fragment, the RNA was shorter than that in the control construct (Figure 10). The T stretch identified here and in earlier studies with archaea terminated efficiently at the expected site with 95% termination and some read-through, showing that this assay worked well (Figure 10B, Supplementary Table 13). Next, we tested two TTS (A1 and A2) found in coding regions with the downstream motif AGATC. Both terminated at the expected site,

confirming the activity of this newly identified downstream motif. However, termination is not as efficient as for the T stretch motif (11% for A1 and 27% for A2) but allows considerable read-through (Figure 10C, Supplementary Table 13). Furthermore, a hairpin motif found upstream of TTS in coding regions (S12, shown in Figure 9) was tested in the assay (Figure 10D, Supplementary Table 13). Again, termination was detected at the expected site, 59% of the transcripts are terminated at the structural motif. To exclude that RNAs were processed at the sequence or structure motifs, northern blots were hybridised with a probe against the downstream fragment. If transcription would not terminate at the insert sequence but promote to the final terminator present in the vector and the transcript would be subsequently processed, the downstream processing product would be picked up with this hybridisation. However, no additional RNA fragments were detected (Supplementary Figure 8).

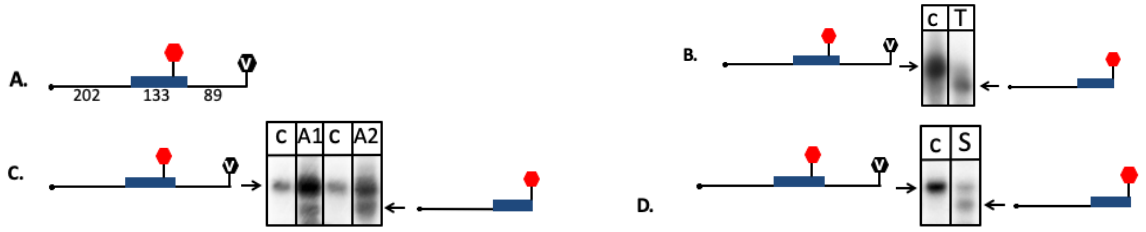

Figure 10. *In vivo* reporter gene test of termination sites. Termination sites were cloned into a reporter gene (the gene for β-galactosidase) construct to confirm their termination activity. RNA was isolated from strains transformed with reporter gene plasmids, separated by size and subsequently transferred to a nylon membrane. Membranes were hybridised with a probe against the β-galactosidase mRNA. Active termination sites generated shorter RNA molecules than the control sequences, indicated by an arrow. Experiments were carried out in triplicate. **A.** The construct used is shown schematically. A full length transcript terminating at the vector termination site is 424 nucleotides long. The insert containing the termination site is 133 nucleotides long (the insert is shown in blue, the location of the termination site in the insert is indicated with a red hexagon), the transcribed plasmid sequences are 202 nucleotides (upstream of the insert) and 89 nucleotides (downstream of the insert) long. The termination site in the plasmid is indicated with a with a V in the black hexagonal. **B.** The T_5C termination sequence was inserted into the reporter gene. Such a T_5C sequence can be found for example in the intergenic sequence downstream of the *ftsz2* gene. Transcripts terminating at the T_5C motif are 321 nucleotides long. The T stretch terminated efficiently (indicated with an arrow) with 95% termination at the T stretch and 5% termination at the plasmid encoded terminator (lane T), lane co: control insert without a termination site. **C.** Two sequences with the downstream motif AGATC (A1 and A2) were inserted into the reporter gene. Both are located in coding regions; A1 is located in the gene HVO_2724, A2 is located in the gene HVO_0455. Termination at these motifs results in 309 nucleotide long RNAs. Both sequences terminated, however considerable read-through is present (termination at A1: 11% and at A2: 27%). Lane A1: AGATC termination site A1, lane A2: AGATC termination site A2. **D.** A sequence with the potential to fold into a hairpin structure (S12) was inserted into the reporter gene. The structure S12 is located in a coding region (gene HVO_A0420). Termination at this motif results in a 309 nucleotide long RNA. S12 terminated, but considerable read-through was observed (termination at S12: 59%). Lane S: sequence S12.

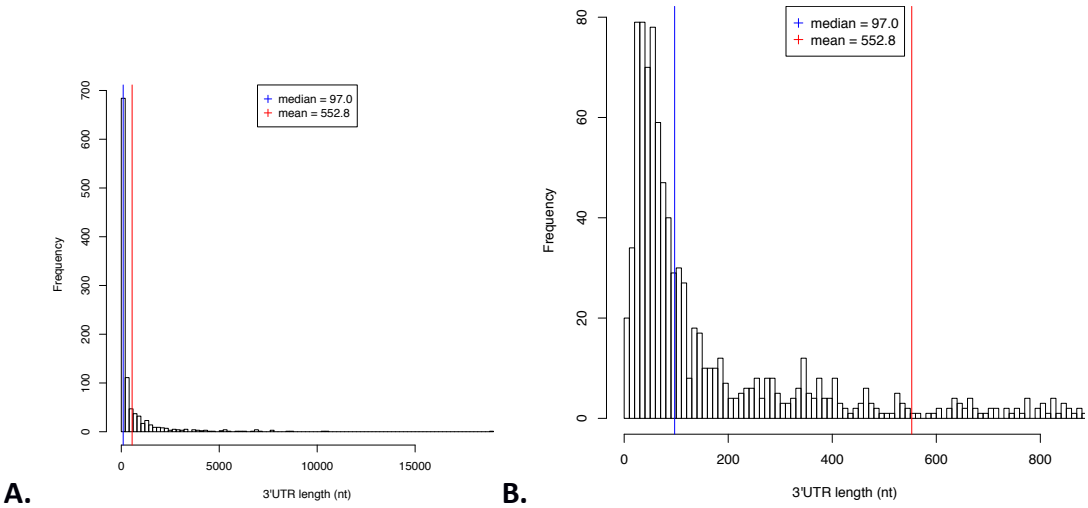

Figure 11. Histogram of 3´ UTR lengths. The 3´ UTR lengths for all primary TTS are shown as calculated from the IE-PC data on the basis of the +TEX reads. **A.** Without any restriction on the length of the 3' UTR, we found a high number of 3' UTRs with lengths below 1,000 nucleotides and very few 3' UTRs that were longer than 1,000 nucleotides. The histogram in panel **B.** shows the frequency of 3' UTR lengths for the region up to 800 nucleotides in length.

3´ UTR length for primary TTS and identification of unannotated genes

To determine 3´ UTR lengths (for the +TEX data set), the distance between the primary TTS and the 3´ end of the preceding annotated gene was determined. The primary TTS had a median 3´ UTR length of 97 nucleotides (Figure 11) (the median 3´ UTR length for all TTS is 189). To confirm that the 3´ UTR regions were part of the same transcripts as the upstream coding regions, we generated transcriptome data. To that end we, performed next-generation sequencing of a cDNA library generated from total RNA for RNAseq. An average of 42 million reads were obtained for three independent cDNA libraries, and with these data, we confirmed that the 3´ UTR regions were part of the same transcripts as the upstream coding regions, since we found continuous RNAseq reads over the complete putative 3´ UTRs (Supplementary Table 4).
Figure 11 shows that some longer 3´ UTRs with lengths of 1 kb and more were present. A possible explanation for this observation is that there may be some as-yet unannotated genes between the next upstream gene end and the TTS. We performed a detailed analysis of these UTR sequences with the +TEX data set, revealing that in 18 cases a transcription start site is located in the 3´ UTR, between the 3´ end of the annotated gene and the TTS (Supplementary Figure 8). Thus, 18 potential new genes are located in these 3´ UTRs defined by the presence of a TSS and a TTS (Supplementary Table 4). The same analysis for the -TEX TTS resulted in 356 TSS-TTS pairs and thus 356 potential new genes.

Interaction of identified sRNAs and long 3´ UTRs

The observed long 3´ UTRs would allow for interactions with regulatory molecules such as proteins and RNA. Small RNAs that can potentially act as regulatory RNAs have been detected in *Haloferax*[32,33], and we analysed whether they have the potential to interact with 3´ UTRs. Twenty-one previously detected sRNAs[32] were investigated for potential interactions with long 3´ UTR regions using RNAplex[34], version 2.4.9 (Vienna

RNApackage). RNAplex allows users to check for interactions of short RNA sequences with longer target sequences, and we first analysed the 3´ UTRs belonging to the TTS identified from the +TEX data set. In addition, we used the 2,631 3´ UTR sequences discovered with the -TEX data set as target sequences. If several UTRs were detected for one gene, we used the longest UTR for the analysis. RNAplex was applied to each sRNA sequence against the long 3´ UTR sequences. To have comparable values, we additionally applied RNAplex to each sRNA against a set of random sequences with a similar length distribution as the set of 3´ UTRs. Only sRNA interactions with 3´ UTRs that showed more favourable binding energies than those with random sequences were taken into account. While we did not find any interactions of sRNA with the +TEX derived 3´ UTRs we found clear signals for potential interactions of sRNAs with 3´ UTR sequences of the -TEX data set: six different sRNAs and nine different interactions (Supplementary Table 5 and Supplementary Figure 6)).

chromosome	**no. of operons**	**max no. genes**	**med no. genes**	**min no. genes**	**median length**
main	442	25	2	2	1,742
pHV1	16	4	2	2	854
pHV3	58	11	2	2	2,276
pHV4	116	6	2	2	1,597

Table 4. Summary of operon information. The number of operons and the number of genes per operon as the maximum, median and minimum for all operons is shown (columns "no of operons", "max no genes", "med no genes", "min no genes"). The last column shows the median length of operons (in bp).

Identification of operons

To identify genes that are transcribed together as multicistronic RNAs, we used the transcriptome data obtained by RNAseq. RNAseq data were analysed together with TSS data already obtained earlier[21] and the TTS data obtained in this work. Genes that are transcribed as a transcriptional unit should, as the simplest example, contain one promoter (a single TSS) and one terminator (a single TTS), and in between, transcriptome data should be continuously present. To identify operons in the *H. volcanii* genome, we used Rockhopper[35], a tool that maps prokaryotic RNA sequencing data onto a reference genome. Given an annotation, the output of the program includes a list of operons retrieved from the RNAseq data. We used the latest version of the *Haloferax* genome annotation[36] as a customised input for Rockhopper. Altogether, we found 632 operons and 2,752 genes that were monocistronic. The majority of operons contained two genes, and the highest number of genes in an operon was 25 (Table 4). Operons are listed in Supplementary Table 6, and operon annotation is also included in Supplementary Table 1, which lists all TTS detected with IE-PC together with corresponding genes, operons and TSS. Additionally, we counted the number of TTS located inside and downstream of an operon for the +TEX data. In total, we detected

326 TTS that were located inside an operon and 207 TTS that were located downstream of an operon.

Discussion

The new IE-PC approach allows comprehensive, genome-wide identification of transcription termination sites

Using the new IE-PC algorithm, we were able to determine the TTS genome-wide for the archaeon *H. volcanii*. Altogether, 1,543 TTS were found with our new algorithm, representing the first unbiased, truly genome-wide approach. The dRNAseq method was originally established for the identification of TSS but the enrichment of primary transcripts also allows enrichment of original 3´ ends as shown here. The number of TTS identified with the dRNASeq approach is clearly lower than the one identified with -TEX but is enriched for transcription termination sites and depleted for processing sites. Additional TTS might be present in the -TEX data set but more experiments are required to differentiate between the TTS and the PS in the -TEX data set.

Motifs for termination

Amongst all TTS, two general sequence motifs for termination were identified, each being used in about half of the termination events. One was specific for intergenic regions, and the other one for coding regions. Both types of termination motifs (T stretch and AGATC) were confirmed using *in vivo* assays. Motifs in intergenic regions consisted of a stretch of Ts, while termination occurred at a C residue. This motif is similar to those found by *in vitro* studies in other archaea[10,12-14,37] and for *S. solfataricus* and *M. mazei* using the DSM algorithm[17]. The presence of TTS in coding regions has previously been observed in archaea[17,38] and the prevailing motif we have found in coding regions was AGATC, located downstream of the termination site. The DNA sequence downstream of the TTS has been shown to influence termination efficiency in bacteria and eukaryotes[13]. Furthermore, it has been reported that the archaeal RNA polymerase interacts with downstream duplex DNA[39]. Thus, it is entirely possible that a downstream termination motif exists. It might also be possible that a protein binds to the downstream motif, acting as a minor roadblock for the RNA polymerase and thereby inducing pausing and in some cases subsequently terminating transcription. The *in vivo* termination test confirmed that AGATC is used as termination motif but revealed that it is not very efficient. That might be characteristic for motifs located in coding regions, to allow transcription of the AGATC containing gene. The problem with terminating transcription in a coding region is that termination should not be too efficient, otherwise the RNA of the underlying coding region will not be transcribed in its full length. A termination event as strong as T_5C would be detrimental for the expression of the terminator containing gene. In addition, it would be very difficult to place such a T stretch in a coding region since the coding region is governed by the requirement to contain the specific triplets for the encoded protein, and especially a G/C rich genome like the one from *Haloferax* can rarely contain a T stretch in a coding region. Thus, to be able to terminate in coding regions other signals have to be used.

Indeed, a T stretch was never found in the +TEX data set as TTS in coding regions. Thus, it seems that the transcription machinery has learned to stop at other motifs.
Comparison of both methods for TTS identification, IE-PC and DSM, showed that IE-PC could identify TTS genome-wide, whereas DSM was restricted to regions downstream of annotated genes. Furthermore, since DSM was designed to include the average insert length of the reads obtained in some cases false TTS were identified.
The search for secondary structure motifs located close to the TTS was successful for a subset of the TTS and revealed hairpin structures upstream of the TTS, suggesting that these structures might influence transcription termination. One of the identified structures was tested in the *in vivo* termination experiments and indeed terminated transcription, however not as efficiently as the T stretch. A certain amount of folding potential upstream of TTS was also found in *M. mazei* but was considered not to be essential, whereas no structures were identified near TTS in *S. acidocaldarius*[17]. Possibly, termination signals in different archaeal phyla are more variable.

Different 3´ UTR lengths are generated by termination at several TTS

With our genome-wide TTS determination approach, we can present the first comprehensive determination of 3´ UTR lengths in *H. volcanii*. The median 3´ UTR length for primary TTS was 97 nucleotides, which is longer than the median 3´ UTR length recently reported for two archaea (with 55 nucleotides for *S. solfataricus* and 85 nucleotides for *M. mazei*) and the bacterium *B. subtilis* (40 nucleotides)[17]. The presence of transcripts with long UTRs was confirmed by RNAseq data. One explanation for these long UTRs could be that as yet un-detected genes are located between the 3´ ends of the upstream genes and the TTS. In particular, non-coding genes are very difficult to identify, and might not have been detected yet. While approaches to identify small non-coding RNAs have been performed with archaea, attempts to find long non-coding RNAs shown to exist in eukaryotes[40] (f.i. 17-kb-long XIST RNA) are, to our knowledge, still missing for archaea. Our detailed analysis of 3´ UTR sequences indeed found 356 potential new genes between the 3´ end of an annotated gene and a TTS. These new genes could encode non-coding RNAs. Further biological studies are required to confirm this hypothesis.
In many cases, we found several TTS downstream of one gene, showing that mRNAs with different 3´ UTR lengths are generated. The lengths of the UTRs as well as the fact that some genes have several 3´ UTRs provide a good platform for interactions with regulatory molecules, such as sRNAs. This is similar to the situation in eukaryotes, in which termination at different sites generates different 3´ UTRs, that can interact with different regulatory proteins or RNAs such as miRNAs. Different UTR lengths were also found in *M. mazei* and *S. acidocaldarius*[17].
Several small RNAs previously identified in *Haloferax* were shown to be able to interact with 3´ UTRs. Thus, gene regulation via the interaction of sRNAs with 3´ UTR sequences would be possible in *Haloferax*. A similar situation was found for *M. mazei*[17], suggesting that this phenomenon might be a general characteristic of archaea that is possibly used for gene regulation.

Transcription produces antisense RNA

If a gene is located on the opposite strand of a TTS (Figure 4B, lower panel), an antisense RNA against this oppositely encoded gene is generated. Of the 1,543 TTS detected, 14.4% (222 TTS) of them were located in in a gene on the opposite strand, thus resulting in antisense RNAs. Up to 75% of all genes were found to be associated with antisense RNAs in bacteria[41-43]. Studies with archaea painted a similar picture[21,44]. In *S. acidocaldarius*, 301 gene pairs with convergent orientations were investigated, and in 52% of them, the terminator of one gene was located in the coding region of a gene on the opposite strand[17] (as in Figure 4B lower panel); however, in *M. mazei*, only 8% of the convergent genes showed such an overlap. The potential functions and the physiological relevance of these antisense transcripts must be uncovered in the future.

Low numbers of operons in *Haloferax*

The number and composition of operons for *Haloferax* were determined based on RNAseq data. In total, 632 operons were identified with an average of two genes per operon, resulting in 1,687 genes in operons for *Haloferax* and 2,749 monocistronic genes. The operon with the highest number of genes contained 25 genes, almost all of which encode ribosomal proteins (HVO_2565-2541). The published TSS data revealed that the transcription start site for a gene can be located in an upstream gene in an operon[21]. Similarly, the TTS data reported here show that transcription can be terminated in a downstream gene in an operon. Together, this allows fine-tuning of the expression of individual genes in an operon and thus challenges the simplistic view that all genes of an operon are co-transcribed under all conditions. This phenomenon was confirmed experimentally, for instance for the dicistronic operon on the minus strand encoding the proteins Lsm and L37e (HVO_2723 and HVO_2722). Here, a promoter for expression of the *l37e* gene is located in the upstream *lsm* gene (Supplementary Figure 7)[45]. In addition, the upstream *lsm* gene can be terminated within the downstream *l37e* gene. A dicistronic transcript and a *l37e* mRNA were identified by northern blot analysis[45].

Conclusion

Taken together, we showed that our new IE-PC approach used with dRNAseq is well-suited to identifying the complete set of TTS for a genome without any a priori limitations on the search space due to genome annotation. We confirmed the T stretch termination motif detected 30 years ago, but we also identified an additional new motif specific for coding regions as well as a hairpin structure for termination. The presence of multiple 3´ UTRs for a gene provides a platform for regulatory mechanisms similar to those described in eukaryotic systems.

Materials & Methods

Haloferax volcanii culture conditions and RNA extraction

H. volcanii strains H119 (Δ*leuB,* Δ*pyrE2,* Δ*trpA*)[46] and HV55 (Δ*leuB,* Δ*pyrE2,* Δ*trpA,* Δ*bga*H) (this work, see below) were grown aerobically at 45°C in Hv-YPC or Hv-Ca medium[46] to an OD_{600} of 0.8-0.9 (for a list of strains used see Supplementary Table 9). Total RNA was isolated from three biological replicates using TRIzol (ThermoFisher scientific). RNA fractions were sent to vertis (vertis Biotechnologie AG, Martinsried, Germany) for further treatment, cDNA library preparation and high throughput sequencing. RNA preparations were made for each of the three different RNAseq approaches: (1) To obtain an RNA fraction enriched in primary transcripts RNA was treated with terminator exonuclease (this fraction was termed +TEX), (2) RNA from one preparation was left untreated representing the complete RNA pool (this fraction was termed -TEX). RNA preparations (1) and (2) were performed such that the original RNA 3´ end was maintained (for details see Supplementary methods). For RNA preparation (3) RNA was isolated for obtaining transcriptome data (this fraction was termed RNAseq data). In all sequencing approaches primers were designed such that the identification of the RNA strand was possible. Details for library construction and sequencing are reported in Supplementary methods.

E. coli culture conditions

E. coli strains DH5α (Invitrogen) and GM121[47] were grown aerobically at 37°C in 2YT medium.

Read mapping

Raw reads were clipped and trimmed using cutadapt version 1.10[48] based on fastqc version 0.11.4[49] quality control reports. Reads were then mapped with segemehl (version 0.2.0)[50,51]. We used -A 94 to require higher accuracy in order to account for prokaryote mapping instead of mapping eukaryotic genomes. Mapped reads were afterwards processed using samtools version 1.3[52]. In order to calculate genome coverage and intersection of data sets, we used bedtools (bedtools v2.26.0)[53].

DSM analysis

Based on the Dar-Sorek-Method (DSM)[17,54], putative TTS were retrieved from the mapping data. We only used read pairs and at least a coverage of 4 for every valid position. For each annotated region in the genome, all inserts overlapping with the annotated region were collected. Then, for each position downstream of a gene, all collected reads that end at this position were summarised, excluding inserts longer than 500 bp. The average insert length of all the corresponding read pairs was defined as the length of the target downstream region. For all read pairs where the insert overlapped an annotated region, the position with highest read-end coverage inside the target downstream region of an annotated 3´ end was reported as a transcription termination site (TTS) (Table 1 and Figure 1). Collected TTS with DSM are listed in Supplementary Table 11. The used implementation of this method is available at Bioinformatics Leipzig (http://www.bioinf.uni-leipzig.de/publications/ supplements/18-059).

Internal Enrichment algorithm

To detect sites with a significant enrichment of sequenced and mapped fragment ends a sound background without enrichment is desired. In the current setting, we used the intrinsic properties of a paired-end sequencing run to directly deduce the following information (Figure 2). We also took advantage of the fact, that the cDNA library preparation applies a fragmentation step after ligation of the 3´ end primer. Since each fragment which results from an individual fragmentation event (in contrast to PCR duplicated fragments) is very unlikely to have the exact same length, truly enriched sites can be expected to be associated with sequenced fragments, all ending at the respective site but starting at different positions. Therefore, the more different mates (mapping to different positions) are associated with the different reads ending at a particular site the higher the enrichment of read end signal at that particular position can be considered. To capture this, we calculate for each position a score S as

$$S_i = \frac{C_i}{\left(\prod_{j=1}^{n} C_j\right)^{1/n}} \forall j \exists R(i \circ j)$$

Thereby, C_i denotes the number of fragment ends at position i, C_j the number of fragment starts at position j, and $R(i \circ j)$all position i,j which are associated via at least one read-mate-pair R. To get an expected background distribution of these scores, we again use the nature of paired-end reads. Since we expect only fragment ends, in contrast to fragment starts, to be enriched, we can use the distribution of the reciprocally defined scores for the fragment start S_j as a background distribution. Based on the distributions, native fragment end scores can be assigned an empirical p-value, evaluating its likelihood to occur by chance alone. The above detailed software, named Internal-Enrichment (IE) was implemented in perl and is available at Bioinformatics Leipzig (http://www.bioinf.uni-leipzig.de/publications/supplements/18-059). We ran IE using all position showing signals for starts and ends (-mr 1), we omitted read fragments with length more than 100 nt (-mf 100), and the geometric mean as a method to calculate a score for the given signals (-mode GeomMean). We only included signals that were present in all three replica such that each showed a maximal empirical p-value of 0.05. The geometric mean of all empirical p-values for one position is used as the final score for the given signal.

Peak Calling

A complementary approach to find transcription termination sites is identifying peaks in read stops (Figure 3). In order to find these peaks, we first computed the strand specific read coverage at every position in the genome, which we used as a background.

We then used a sliding window approach with a window size of 150 nt and an overlap of 50 nt to find positions in the respective windows with a significantly higher number of read stops than the rest of the window. This was done by computing the mean number of stops as well as the standard deviation by window and then calculating the z-score for the number of read ends at every position of the window. Subsequently, we required a minimum number of 5 reads stopping, at least 10% of all reads covering

position i-1 must end at position i as well as a minimum z-score of +2 at a site to report it as a putative TTS. As a consequence of the overlapping of the windows, a site is evaluated multiple times and must fulfil the criteria in at least one contexts evaluated.

Terminator identification

To be able to differentiate between 3´ end processing sites and termination sites we isolated RNA and enriched this RNA for primary transcripts by digestion with the terminator exonuclease (TEX). Primary transcripts contain 5´ triphosphates and are not digested by TEX. In order to see which loci are related to transcription termination focusing on primary transcripts, we calculated the overlap between the resulting RNAseq data from the +TEX library applied to IE and the results of the peak calling procedure applied on +TEX as well. Our final set of TTS is the intersection of resulting positions of both methods, PC and IE. More information is available at Bioinformatics Leipzig (http://www.bioinf.uni-leipzig.de/publications/supplements/18-059).

We defined our set of putative TTS as the intersection of the resulting positions of the IE and PC approaches. In most cases, both methods reported the same sites. For a few cases, more than one position was reported within a distance of 10 nt. In this case, the positions with the highest coverage was chosen as TTS.

The same approach to identify TTS was carried out with the -TEX data set.

Terminator sequence and structure analysis

We calculated percentages of nucleotide enrichment and hybridization energies for regions of 45 nt upstream and downstream of the TTS. Hybridization energies are calculated as the energy that stabilizes the RNA-DNA hybrid during transcription[27], using RNAplfold, a program part of the ViennaRNA package[29] (version 2.4.9) with special energy parameters for RNA-DNA hybrids. The lower (the more negative) the hybridization energy, the more stable is the RNA-DNA hybrid.

Motif and structure analysis for 3´ UTR sequences

Motifs were detected using MEME[26] (version 5.0.1) by scanning all sequences from 15 nt upstream to 5 nt downstream of the TTS (see Figure 8 for an example). Structure search was conducted using graphclust 2.0[30] within Galaxy[31]. Input sequences were sequences of 100 nt length upstream to the TTS. Graphclust was used with default parameter and additionally a window size of 110 (such that our sequences fit in one window to avoid duplicated sequences), a bitscore of 15 for the results of cmscan, an upper threshold of 50 clusters and 20 top sequences in each alignment for the visualization. The application of graphclust resulted in 37 clusters providing covariance models (CMs) for each. Covariance models are probabilistic models that are created based on given sequence and structure motifs and can be used to search for sequence and structure homologies. To scan sequences for these given motifs, we applied cmscan (a part of the infernal program suite[55], version 1.1.1) on our input sequences after calibrating the Cms (using cmcalibrate, part of the infernal program suite) and additionally on all the transcripts in order to get a background model. Given cmscan results, we filtered the secondary structures detected in the TTS-related sequences such that the structures were at most 10 nt upstream from the TTS.

Analysis of 3´ UTR regions: coverage RNAseq, interactions with sRNAs

We confirmed 3´ UTR regions by checking if coding region and 3´ UTR are completely covered by RNAseq reads. If so, this would be indicative of uninterrupted transcription

until the assigned TTS. Concerning the -TEX data: out of 2,631 3´ UTR sequences, 2,518 show a continuous coverage within RNAseq reads (95.7%) (Supplementary Table 4). In order to get an impression of continuously covered sequences throughout intergenic regions in the genome, we create a randomized set of sequences with similar length distribution as our set of 3´ UTR sequences. Randomized 3´ UTR sequences were created using bedtools shuffle[53]. This results in 75.9% of the randomized sequences being continuously covered by RNAseq reads. This shows that read coverage is generally very high, but the coverage for the 3´ UTR sequences is with 95.7% clearly higher than the result for randomized sequences. The interaction of 21 previously detected sRNAs[33] with long 3´ UTR regions was investigated using RNAplex[34], a program of the ViennaRNA package, that allows to check for interactions of short RNA sequences with longer target sequences. In total, 2,631 3´ UTR sequences were used, for genes with several corresponding TTS, the longest 3´ UTR sequence was chosen. These are based on TTS detected by IE-PC with the -TEX data set. The same approach was performed with the TTS detected by IE-PC with the +TEX data set which resulted in 1,056 3´ UTRs. RNAplex was applied to each sRNA sequence against all 3´ UTR sequences. In order to have comparable values, we additionally applied RNAplex to each sRNA against a set of random sequences with similar length distribution as the set of 3´ UTRs. We then checked for each of the sRNA if there is a clear separation between binding energies of interactions with 3´ UTR sequences or random sequences.

Code availability

The code is available at: (http://www.bioinf.uni-leipzig.de/publications/supplements/18-059)(see also data availability statement below).

Reporter gene investigations of identified TTS

Construction of bgaH deletion mutant HV55. To allow the use of a plasmid carrying the ß-galactosidase reporter gene (a fusion of ß-galactosidase genes from *Haloferax alicante* and *Haloferax volcanii*)(*bga*Ha)[56,57], the *H. volcanii* ß-galactosidase gene (*bga*H) was deleted in strain H119. H119 was transformed with the *bga*H deletion plasmid pTA617 (for a list of plasmids used see Supplementary Table 9)[45,58] and grown in Hv-Ca medium with tryptophan (final concentration 0.25 mM) to generate the deletion strain[59]. Homozygous knock-out clones were verified via southern-blot using *Sal*I digested genomic DNA and probe bgaHaDO (primers: bgaHKODO-for and bgaHKODO-rev (for a list of primers used see Supplementary Table 9); template genomic DNA of *H. volcanii*). Probe labelling and detection of the blot were carried out using the DIG-DNA labelling mix and detection reagents (Anti-Digoxigenin-AP) (Roche) according to the manufacturer´s protocol. The resulting deletion strain was termed HV55.

Construction of pTA231-termtest constructs. To construct the terminator-test plasmid a p.*syn* expression cassette (Anice Sabag-Daigle and Charles J. Daniels, in preparation) synthesised by lifetechnologies™ (Thermofisher Scientific) was introduced via *Not*I/*Eco*RI into pTA231[46], resulting in pTA231psyn (for sequences see Supplementary Table 9). This plasmid was subsequently cured of the *Bam*HI site by digestion with *Bam*HI, blunting by Klenow fragment and religation. Then a C-terminal fragment of the reporter gene *bga*Ha was inserted at the *Nde*I site. After *Nde*I digestion,

the vector was treated with Pfu polymerase to fill-in the *Nde*I site. The inserted *bga*Ha fragment was generated by PCR using primers bgaHatermifw and bgaHatermirev and pTA599 as template[58]. The correct orientation of the inserted fragment was confirmed by sequencing, the resulting plasmid was termed pTA231-termtest. The candidate terminator sequences were inserted into the *Bam*HI site of this plasmid. Terminator fragments were generated with PCR using primers as listed in Supplementary Table 9 and templates pTA599 (for the control construct) or genomic DNA of *H. volcanii* (for TTS-A1, TTS-A2, TTS-S12 and TTS-ftsZ). HV55 cells were transformed with pTA231termtest constructs, and grown in Hv-Ca medium with uracil (final concentration 0.45 mM). Cells were grown to an OD_{650} of 0.6-1.0 and RNA was isolated using NucleoZOL (Machery-Nagel) according to manufacturer´s instructions.
Northern-blot analysis of Hv55xpTA231-termtest RNA. To analyse the RNA levels, total RNA was isolated from *H. volcanii* as described above. Ten µg RNA was separated on a 1,5% agarose gel and subsequently transferred to a nylon membrane (Biodyne® A, PALL). After UV-crosslinking the membrane was hybridised with a radioactively labelled probe against the 5'-part of the *bga*Ha mRNA to detect termination events. The probe was generated by PCR (primers bgaHatermi fw and TermiVectorrev; template pTA599) and the purified PCR fragment was labelled using α-^{32}P-dCTP and the random primed DNA labelling kit DECAprime™II (invitrogen). Experiments were done in triplicate. For detection of a potential processing fragment downstream of the termination site, a radioactively labelled probe against the 3'-part of the *bga*Ha mRNA was used. The probe was generated and labelled as described above using primers BetaHinten1 and BetaHinten2 and pTA231-termtestA1 as template. For quantification of northern blot signals, membranes were exposed to phosphorimaging plates(FujiFilm) and the resulting signals detected using a Typhoon imager (GE). Analysis of three replicates was carried out using the ImageQuant TL software (GE).

Acknowledgements
AM, LKM, JW and PM would like to thank Britta Stoll, Elli Bruckbauer, Irma Merdian and Susanne Schmidt for expert assistance, as well as the DFG (MA1538/21-1) for funding. FA was supported by the Austrian Science Fund (FWF) (SFB F43).

Author contributions. AM conceived the study and designed the experiments. LKM, JW and PM performed the experiments. SJB, FA, SHB analysed the RNAseq data and wrote the code. SJB, LKM, FA, SHB, FP, PFS and AM analysed the data. SJB and AM wrote the manuscript with support from FP, FA, SHB and PFS, all authors commented and revised the manuscript.

Competing interests. The authors declare no competing interest.

Materials & Correspondence. Indicate the author(s) to whom correspondence and material requests should be addressed.
studla@bioinf.uni-leipzig.de, anita.marchfelder@uni-ulm.de

Data availability

The data supporting our findings are available in the Supplementary Information, in addition the code is available under (http://www.bioinf.uni-leipzig.de/publications/supplements/18-059), RNAseq data are available at the European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena) with the following assigned experiment accession numbers.

Experiment replicon	Experiment accession number at ENA
+TEX, replicon 1	ERX3014341
+TEX, replicon 2	ERX3014376
+TEX, replicon 3	ERX3014377
-TEX, replicon 1	ERX3014378
-TEX, replicon 2	ERX3014379
-TEX, replicon 3	ERX3014380
transcriptome, replicon 1	ERX3014381
transcriptome, replicon 2	ERX3014382
transcriptome, replicon 3	ERX3014383

References

1 Fouqueau, T. *et al.* The cutting edge of archaeal transcription. *Emerging Topics in Life Sciences* (2018).

2 Bell, S. D. Archaeal transcriptional regulation – variation on a bacterial theme? *Trends in microbiology* **13**, 262-265, doi:10.1016/j.tim.2005.03.015 (2005).

3 Ray-Soni, A., Bellecourt, M. J. & Landick, R. Mechanisms of Bacterial Transcription Termination: All Good Things Must End. *Annu Rev Biochem.* **85:319-47.**, 10.1146/annurev-biochem-060815-014844. Epub 062016 Mar 060817. (2016).

4 Porrua, O., Boudvillain, M. & Libri, D. Transcription Termination: Variations on Common Themes. *Trends Genet.* **32**, 508-522. doi: 510.1016/j.tig.2016.1005.1007. Epub 2016 Jun 1028. (2016).

5 Peters, J. M., Vangeloff, A. D. & Landick, R. Bacterial Transcription Terminators: The RNA 3′ -End Chronicles. *Journal of molecular biology* **412**, 793-813, doi:10.1016/j.jmb.2011.03.036 (2011).

6 Nemeth, A. *et al.* RNA polymerase I termination: Where is the end? *Biochimica et biophysica acta* **1829**, 306-317, doi:10.1016/j.bbagrm.2012.10.007 (2013).

7 Arimbasseri, A. G., Rijal, K. & Maraia, R. J. Transcription termination by the eukaryotic RNA polymerase III. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1829**, 318-330, doi:10.1016/j.bbagrm.2012.10.006 (2013).

8 Kuehner, J. N., Pearson, E. L. & Moore, C. Unravelling the means to an end: RNA polymerase II transcription termination. *Nature Reviews. Molecular Cell Biology* **12**, 283-294, doi:10.1038/nrm3098 (2011).

9 Fouqueau, T., Blombach, F. & Werner, F. Evolutionary Origins of Two-Barrel RNA Polymerases and Site-Specific Transcription Initiation. *Annu Rev Microbiol.* **71:331-348.**, 10.1146/annurev-micro-091014-104145. Epub 092017 Jun 091028. (2017).

10 Santangelo, T. J. & Reeve, J. N. Archaeal RNA polymerase is sensitive to intrinsic termination directed by transcribed and remote sequences. *J Mol Biol.* **355**, 196-210. doi: 110.1016/j.jmb.2005.1010.1062. Epub 2005 Nov 1019. (2006).
11 Hirtreiter, A. *et al.* Spt4/5 stimulates transcription elongation through the RNA polymerase clamp coiled-coil motif. *Nucleic Acids Res.* **38**, 4040-4051. doi: 4010.1093/nar/gkq4135. Epub 2010 Mar 4042. (2010).
12 Santangelo, T. J., Cubonova, L., Skinner, K. M. & Reeve, J. N. Archaeal intrinsic transcription termination in vivo. *J Bacteriol.* **191**, 7102-7108. doi: 7110.1128/JB.00982-00909. Epub 02009 Sep 00911. (2009).
13 Spitalny, P. & Thomm, M. A polymerase III-like reinitiation mechanism is operating in regulation of histone expression in archaea. *Mol Microbiol.* **67**, 958-970. doi: 910.1111/j.1365-2958.2007.06084.x. Epub 02007 Dec 06019. (2008).
14 Thomm, M., Hausner, W. & Hethke, C. Transcription Factors and Termination of Transcription in Methanococcus. *Systematic and Applied Microbiology* **16**, 648-655, doi:10.1016/S0723-2020(11)80336-X (1993).
15 Santangelo, T. J. *et al.* Polarity in archaeal operon transcription in Thermococcus kodakaraensis. *J Bacteriol.* **190**, 2244-2248. doi: 2210.1128/JB.01811-01807. Epub 02008 Jan 01811. (2008).
16 Walker, J. E., Luyties, O. & Santangelo, T. J. Factor-dependent archaeal transcription termination. *Proc Natl Acad Sci U S A.* **114**, E6767-E6773. doi: 6710.1073/pnas.1704028114. Epub 1704022017 Jul 1704028131. (2017).
17 Dar, D., Prasse, D., Schmitz, R. A. & Sorek, R. Widespread formation of alternative 3' UTR isoforms via transcription termination in archaea. *Nat Microbiol.* **1**, 16143. doi: 16110.11038/nmicrobiol.12016.16143. (2016).
18 Leigh, J. A., Albers, S. V., Atomi, H. & Allers, T. Model organisms for genetics in the domain Archaea: methanogens, halophiles, Thermococcales and Sulfolobales. *FEMS Microbiol Rev* **35**, 577-608 (2011).
19 Pohlschroder, M. & Schulze, S. Haloferax volcanii. *Trends in microbiology*, doi:10.1016/j.tim.2018.10.004 (2018).
20 Ammar, R. *et al.* Chromatin is an ancient innovation conserved between Archaea and Eukarya. *eLife* **1**, e00078, doi:10.7554/eLife.00078 (2012).
21 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: 610.1186/s12864-12016-12920-y. (2016).
22 Sharma, C. M. *et al.* The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature.* **464**, 250-255. Epub 2010 Feb 2017. (2010).
23 Vogel, J. & Sharma, C. M. How to find small non-coding RNAs in bacteria. *Biol Chem.* **386**, 1219-1238. (2005).
24 Pfeiffer, F. *et al.* Genome information management and integrated data analysis with HaloLex. *Arch Microbiol.* **190**, 281-299. Epub 2008 Jul 2001. (2008).
25 Veloso, F., Riadi, G., Aliaga, D., Lieph, R. & Holmes, D. S. Large-scale, multi-genome analysis of alternate open reading frames in bacteria and archaea. *Omics* **9**, 91-105, doi:10.1089/omi.2005.9.91 (2005).
26 Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic acids research* **37**, W202-W208 (2009).
27 Lorenz, R., Hofacker, I. L. & Bernhart, S. H. Folding RNA/DNA hybrid duplexes. *Bioinformatics* **28**, 2530-2531 (2012).
28 Bernhart, S. H., Hofacker, I. L., Will, S., Gruber, A. R. & Stadler, P. F. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC bioinformatics* **9**, 474 (2008).

29 Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms for Molecular Biology* **6**, 26 (2011).
30 Miladi, M. *et al.* RNAscClust: clustering RNA sequences using structure conservation and graph based motifs. *Bioinformatics* **33**, 2089-2096 (2017).
31 Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic acids research* **46**, W537-W544 (2018).
32 Heyer, R. *et al.* High throughput sequencing reveals a plethora of small RNAs including tRNA derived fragments in Haloferax volcanii. *RNA Biol* **9**, 1011-1018 (2012).
33 Straub, J. *et al.* Small RNAs in haloarchaea: Identification, differential expression and biological function. *RNA Biol* **6**, 281-292 (2009).
34 Tafer, H. & Hofacker, I. L. RNAplex: a fast tool for RNA–RNA interaction search. *Bioinformatics* **24**, 2657-2663 (2008).
35 Tjaden, B. De novo assembly of bacterial transcriptomes from RNA-seq data. *Genome Biol.* **16:1.**, 10.1186/s13059-13014-10572-13052. (2015).
36 Pfeiffer, F. & Oesterhelt, D. A manual curation strategy to improve genome annotation: application to a set of haloarchael genomes. *Life (Basel, Switzerland)* **5**, 1427-1444, doi:10.3390/life5021427 (2015).
37 Brown, J. W., Daniels, C. J., Reeve, J. N. & Konisky, J. Gene Structure, Organization, And Expression In Archaebacteria. *CRC Critical Reviews in Microbiology* **16**, 287-337, doi:10.3109/10408418909105479 (1989).
38 Koide, T. *et al.* Prevalence of transcription promoters within archaeal operons and coding sequences. *Mol Syst Biol.* **5:285.**, 10.1038/msb.2009.1042. Epub 2009 Jun 1016. (2009).
39 Werner, F. & Grohmann, D. Evolution of multisubunit RNA polymerases in the three domains of life. *Nature Reviews Microbiology* **9**, 85-98, doi:10.1038/nrmicro2507 (2011).
40 Brosnan, C. A. & Voinnet, O. The long and the short of noncoding RNAs. *Current opinion in cell biology* **21**, 416-425, doi:10.1016/j.ceb.2009.04.001 (2009).
41 Amman, F. *et al.* Primary transcriptome analysis reveals importance of IS elements for the shaping of the transcriptional landscape of Bordetella pertussis. *RNA biology*, 1-9 (2018).
42 Georg, J. & Hess, W. R. cis-antisense RNA, another level of gene regulation in bacteria. *Microbiology and molecular biology reviews : MMBR* **75**, 286-300, doi:10.1128/mmbr.00032-10 (2011).
43 Thomason, M. K. & Storz, G. Bacterial antisense RNAs: how many are there, and what are they doing? *Annual review of genetics* **44**, 167-188, doi:10.1146/annurev-genet-102209-163523 (2010).
44 Cohen, O. *et al.* Comparative transcriptomics across the prokaryotic tree of life. *Nucleic Acids Res.* **44**, W46-53. doi: 10.1093/nar/gkw1394. Epub 2016 May 1096. (2016).
45 Maier, L. K. *et al.* Deletion of the Sm1 encoding motif in the lsm gene results in distinct changes in the transcriptome and enhanced swarming activity of Haloferax cells. *Biochimie* **6**, 00058-00059 (2015).
46 Allers, T., Ngo, H. P., Mevarech, M. & Lloyd, R. G. Development of additional selectable markers for the halophilic archaeon Haloferax volcanii based on the leuB and trpA genes. *Appl Environ Microbiol* **70**, 943-953 (2004).
47 Allers, T., Barak, S., Liddell, S., Wardell, K. & Mevarech, M. Improved strains and plasmid vectors for conditional overexpression of His-tagged proteins in Haloferax volcanii. *Appl Environ Microbiol* **76**, 1759-1769 (2010).

48 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, pp. 10-12 (2011).
49 Andrews, S. FastQC: a quality control tool for high throughput sequence data. (2010).
50 Hoffmann, S. *et al.* Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS computational biology* **5**, e1000502 (2009).
51 Otto, C., Stadler, P. F. & Hoffmann, S. Lacking alignments? The next-generation sequencing mapper segemehl revisited. *Bioinformatics* **30**, 1837-1843 (2014).
52 Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
53 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
54 Dar, D. *et al.* Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. *Science.* **352**, aad9822. doi: 9810.1126/science.aad9822. (2016).
55 Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933-2935 (2013).
56 Holmes, M. L. & Dyall-Smith, M. L. Sequence and expression of a halobacterial beta-galactosidase gene. *Mol Microbiol.* **36**, 114-122. (2000).
57 Delmas, S., Shunburne, L., Ngo, H. P. & Allers, T. Mre11-Rad50 promotes rapid repair of DNA damage in the polyploid archaeon Haloferax volcanii by restraining homologous recombination. *PLoS Genet.* **5**, e1000552. Epub 1002009 Jul 1000510. (2009).
58 Large, A. *et al.* Characterization of a tightly controlled promoter of the halophilic archaeon Haloferax volcanii and its use in the analysis of the essential cct1 gene. *Mol Microbiol.* **66**, 1092-1106. Epub 2007 Oct 1031. (2007).
59 Bitan-Banin, G., Ortenberg, R. & Mevarech, M. Development of a gene knockout system for the halophilic archaeon Haloferax volcanii by use of the pyrE gene. *J Bacteriol* **185**, 772-778 (2003).