Nanopore-based native RNA sequencing provides insights into prokaryotic transcription, operon structures, rRNA maturation and modifications

Felix Grünberger[1], Robert Knüppel[2], Michael Jüttner[2], Martin Fenk[1], Andreas Borst[3], Robert Reichelt[1], Winfried Hausner[1], Jörg Soppa[3], Sebastien Ferreira-Cerca[2*], and Dina Grohmann[1*]

[1] Institute of Biochemistry, Genetics and Microbiology, Institute of Microbiology and Archaea Centre, Single-Molecule Biochemistry Lab & Biochemistry Centre Regensburg, University of Regensburg, Universitätsstraße 31, 93053 Regensburg, Germany

[2] Institute for Biochemistry, Genetics and Microbiology, Biochemistry III, University of Regensburg, Universitätsstraße, 31, 93053 Regensburg, Germany.

[3] Goethe-University, Biocentre, Institute for Molecular Biosciences, Max-von-Laue-Str. 9, D-60439 Frankfurt, Germany

*For correspondence:

Sébastien Ferreira-Cerca
Biochemistry III – Institute for Biochemistry, Genetics and Microbiology, University of Regensburg, Universitätsstraße 31, 93053 Regensburg, Germany.
e-mail: sebastien.ferreira-cerca@ur.de
Tel.: 0049 941 943 2539
Fax: 0049 941 943 2474

Dina Grohmann
Department of Biochemistry, Genetics and Microbiology, Institute of Microbiology, University of Regensburg, Universitätsstraße 31, 93053 Regensburg, Germany
e-mail: dina.grohmann@ur.de
Tel.: 0049 941 943 3147
Fax: 0049 941 943 2403

Keywords: Nanopore, RNA-seq, next generation sequencing, transcription, ribosomal RNA, RNA modifications, transcriptome, archaea, bacteria

Abstract (186 words)

The prokaryotic transcriptome is shaped by transcriptional and posttranscriptional events that define the characteristics of an RNA, including transcript boundaries, the base modification status, and processing pathways to yield mature RNAs. Currently, a combination of several specialised short-read sequencing approaches and additional biochemical experiments are required to describe all transcriptomic features. In this study, we present native RNA sequencing of bacterial (*E. coli*) and archaeal (*H. volcanii*, *P. furiosus*) transcriptomes employing the Oxford Nanopore sequencing technology. Based on this approach, we could address multiple transcriptomic characteristics simultaneously with single-molecule resolution. Taking advantage of long RNA reads provided by the Nanopore platform, we could accurately (re-)annotate large transcriptional units and boundaries. Our analysis of transcription termination sites revealed that diverse termination mechanisms are in place in archaea. Moreover, we shed light on the poorly understood rRNA processing pathway in archaea and detected new processing intermediates. One of the key features of native RNA sequencing is that RNA modifications are retained. We could confirm this ability by analysing the well-known KsgA-dependent rRNA methylation sites. Notably, our analysis suggests that rRNA modifications are more abundant in a hyperthermophilic organism.

Introduction

In the last decade, next-generation sequencing (NGS) technologies[1] revolutionized the field of microbiology[2], which is not only reflected in the exponential increase in the number of fully sequenced microbial genomes, but also in the detection of microbial diversity in many hitherto inaccessible habitats based on metagenomics. Using transcriptomics, important advances were also possible in the field of RNA biology[3,4] that shaped our understanding of the transcriptional landscape[5,6] and RNA-mediated regulatory processes in prokaryotes[7]. RNA sequencing (RNA-seq) technologies can be categorized according to their platform-dependent read lengths and necessity of a reverse transcription and amplification steps, to generate cDNA[8]. Illumina sequencing yields highly accurate yet short sequencing reads (commonly 100-300 bp). Hence, sequence information is only available in a fragmented form, making full-length transcript- or isoform-detection a challenging task[9,10]. Sequencing platforms developed by Pacific Bioscience (PacBio) and Oxford Nanopore Technologies (ONT) solved this issue. Both sequencing methods are *bona fide* single-molecule sequencing techniques that allow sequencing of long DNAs or RNAs[11,12]. However, the base detection differs significantly between the two methods. PacBio-sequencers rely on fluorescence-based single-molecule detection that identifies bases based on the unique fluorescent signal of each nucleotide during DNA synthesis by a dedicated polymerase[11]. In contrast, in an ONT sequencer, the DNA or RNA molecule is pushed through a membrane-bound biological pore with the aid of a motor protein that is attached to the pore protein called a nanopore (Fig. 1a). A change in current is caused by the translocation of the DNA or RNA strand through this nanopore, which serves as a readout signal for the sequencing process. Due to the length of the nanopore (version R9.4), a stretch of approximately five bases contributes to the current signal. Notably, only ONT offers the possibility to directly sequence native RNAs without the need for prior cDNA synthesis and PCR amplification[13]. As RNAs are directly "read", RNA modifications are maintained and first attempts have been made to use ONT sequencing to identify RNA base modifications (e.g. methylations[14,15]). As ONT-based native RNA-seq holds the capacity to sequence full-length transcripts, to identify RNA base modifications and to detect molecular heterogeneity in a transcriptome, the technology found widespread attention[16]. Recently, the technology was exploited to sequence viral RNA genomes[17–20] to gain insights into viral and eukaryotic

transcriptomes[19,21–23] and to detect RNA isoforms in eukaryotes[24,25]. However, prokaryotic transcriptomes have not been characterized on the genome-wide level by native RNA-seq approaches so far as prokaryotic RNAs lack a poly(A) tail which is required to capture the RNA and feed it into the nanopore.

Here, we present a native RNA sequencing study of bacterial and archaeal transcriptomes using Nanopore technology. We employed an experimental workflow that includes the enzymatic polyadenylation of prokaryotic RNA to make them amenable for ONT´s direct RNA sequencing kit. We evaluated the accuracy and reliability of native RNA-seq in comparison to published Illumina-based sequencing studying prokaryotic transcriptomes of bacterial (*E. coli*) and archaeal (*Haloferax volcanii*, *Pyrococcus furiosus*) model organisms[26–31]. The transcriptomic analysis included determination of transcript boundaries, providing, among others, insights into termination mechanisms in archaea. We moreover demonstrate that the long RNA reads gathered on the ONT platform allow reliable transcriptional unit assignment. Strikingly, we gained insights into the so far poorly understood ribosomal RNA (rRNA) maturation pathway in archaea and detected putative new rRNA processing intermediates. As RNA modifications are retained when sequencing native RNAs, we explore rRNA modifications in prokaryotes. Notably, by expanding the approach, our analysis suggests that rRNA modifications are more abundant in an hyperthermophilic organism.

Material and Methods

Strains and growth conditions

Escherichia coli K-12 MG1655 cells were grown in LB medium (10 g tryptone, 5 g yeast extract, 10 g NaCl per liter) to an OD_{600nm} of 0.5 and harvested by centrifugation at 3,939 x g for 10 min at 4°C.

Pyrococcus furiosus strain DSM 3638 cells were grown anaerobically in 40 ml SME medium[32] supplemented with 40 mM pyruvate, 0.1 % peptone and 0.1 % yeast extract at 95°C to mid-exponential phase and further harvested by centrifugation at 3,939 x g for 45 min at 4°C.

Markerless deletion of *Haloferax volcanii* KsgA (Hvo_2746) was obtained using the pop-in/pop-out procedure[33]. Deletion candidates were verified by Southern blot and PCR analyses. Full characterization of this strain will be described elsewhere (Knüppel and Ferreira-Cerca, *in preparation*). Wildtype (H26) and Δ*ksgA* strains were grown in Hv-YPC medium at 42°C under agitation as described previously[34].

RNA isolation

E. coli total RNA was purified using the Monarch® Total RNA Miniprep Kit (New England Biolabs) according to manufacturer´s instructions including the recommended on-column DNAse treatment.

P. furiosus cell pellets were lysed by the addition of 1 ml peqGOLD TriFast™ (VWR) followed by shaking for 10 min at room temperature. After adding 0.2 ml 2 M sodium acetate pH 4.0, total RNA was isolated according to the manufacturer´s instructions. Contaminating DNA was removed using the TURBO DNA-free™ Kit (Thermo Fisher Scientific).

H. volcanii

H. volcanii total RNA was purified using the RNA easy kit (Qiagen) according to the manufacturer´s instructions. Alternatively, total RNA was isolated according to the method described by

Chomczynski and Sacchi[35], including a DNA-removal step with RNase-free DNase I (Thermo Scientific).

The Integrity of total RNA from *E. coli* and *P. furiosus* was assessed via a Bioanalyzer (Agilent) run using the RNA 6000 Pico Kit (Agilent). To evaluate the extent of remaining buffer and DNA contaminations, the RNA preparation samples were tested using standard spectroscopic measurements (Nanodrop One) and using the Qubit 1X dsDNA HS assay kit (Thermo Fisher Scientific). RNA was quantified using the Qubit RNA HS assay kit.

Primer extension analysis

5´ends determination of mature 16S and 23S rRNAs from *H. volcanii* by primer extension was performed as described previously (Knüppel et al, Method in Molecular Biology *in press*). In brief, reverse transcription was performed with the indicated fluorescently labeled primers (oHv396-DY682: 5'-CCCAATAGCAATGACCTCCG; oHv622-DY782: 5'-GCTCTCGAGCCGAGCTATCCACC) and SuperScript III reverse transcriptase using 1 µg of total RNA as template. The resulting cDNAs and reference dideoxy-chain termination sequencing ladder reactions were separated on a denaturing 14 % TBE-Urea (6 M)-PAGE. Fluorescence signals (700nm and 800nm) were acquired using a Li-COR Odyssey system.

RNA treatment and poly(A)-tailing

To prevent secondary structure formation, the RNA was heat incubated at 70°C for 3 min and immediately put on ice before TEX-treatment or poly(A)–tailing of the RNA samples. Partial digestion of RNAs that are not 5´-triphosphorylated (e.g. tRNAs, rRNAs) was achieved by incubation of the RNA with the Terminator 5´-Phosphate-Dependent Exonuclease (TEX, Lucigen). For this purpose, 10 µg of RNA was incubated with 1 unit TEX, 2 µl TEX reaction buffer (Lucigen) and 0.5 µl RiboGuard RNase Inhibitor (Lucigen) in a total volume of 20 µl for 60 minutes at 30°C. The reaction was stopped and the RNA was purified using the RNeasy MinElute Cleanup Kit (Qiagen). For *P. furiosus* and *E. coli* RNA samples, control reactions lacking the exonuclease

(NOTEX) were treated as described for TEX-containing samples. In the next step, a poly(A)-tail was added using the *E. coli* poly(A) polymerase (New England Biolabs) following a recently published protocol[36]. Briefly, 5 µg RNA, 20 units poly(A) polymerase, 2 µl reaction buffer and 1 mM ATP were incubated for 15 min at 37°C in a total reaction volume of 50 µl. To stop the reaction and to remove the enzyme, the poly(A)-tailed RNA was purified with the RNeasy MinElute Cleanup Kit (Qiagen).

Direct RNA library preparation and sequencing

Libraries for Nanopore sequencing were prepared from poly(A)-tailed RNAs according to the SQK-RNA001 Kit protocol (Oxford Nanopore, Version: DRS_9026_v1_revP_15Dec2016) with minor modifications for barcoded libraries (see Supplementary Fig. 1a). In this case, Agencourt AMPure XP magnetic beads (Beckman Coulter) in combination with 1 µl of RiboGuard RNase Inhibitor (Lucigen) were used instead of the recommended Agencourt RNAclean XP beads to purify samples after enzymatic reactions. The total amount of input RNA, the barcoding strategy and the number of flowcells used can be found in Supplementary Table 1. The efficiency of poly(A)-tailing was low. However, this could be compensated with a higher amount of input RNA. We added the ONT control RNA (RCS, yeast) to detect problems that arise from library preparation or sequencing. For the barcoded libraries, the RTA adapter was replaced by custom adapters described in https://github.com/hyeshik/poreplex and reverse transcription (RT) was performed in individual tubes for each library. After RT reactions, cDNA was quantified using the Qubit DNA HS assay kit (Thermo Fisher Scientific) and equimolar amounts of DNA for the multiplexed samples were used in the next step for ligation of the RNA Adapter (RMX) in a single tube. Subsequent reactions were performed according to the protocols recommended by ONT. The libraries were sequenced on a MinION using R9.4 flow cells and subsequently, FAST5 files were generated using the recommended script in MinKNOW.

Data analysis

Demultiplexing of raw reads, basecalling and quality control of raw reads

As some bioinformatic tools depend on single-read files we first converted multi-read FAST5 files from the MinKNOW output to single-read FAST5 files using the ont_fast5_api from Oxford Nanopore (https://github.com/nanoporetech/ont_fast5_api). To prevent actual good-quality reads from being discarded (this issue was reported previously[13,37]), we included both failed and passed read folders in the following steps of the analysis. Demultiplexing was done by poreplex (version 0.4, https://github.com/hyeshik/poreplex) with the arguments --trim-adapter, --symlink-fast5, --basecall and --barcoding, to trim off adapter sequences in output FASTQ files, basecall using albacore, create symbolic links to FAST5 files and sort the reads according to their barcodes. However, to ensure consistency between non-multiplexed and multiplexed samples and because of some major improvements in the current basecalling software (guppy), albacore files were not used. Instead demultiplexed FAST5 reads and raw FAST5 reads from non-multiplexed runs were locally basecalled using Guppy (Version 3.0.3) with --reverse_sequence, --hp_correct, --enable_trimming and --calib_detect turned on. After that, relevant information from the sequencing_summary.txt file in the Guppy output was extracted to analyse properties of raw reads (see Supplementary Fig. 2, see Supplementary Table 1).

Mapping of reads and quantification

Files were mapped to reference genomes from *Escherichia coli* K12 MG1655 (GenBank: U00096.2)[38], *Haloferax volcanii* (NCBI Reference Sequence NC_013967)[39] and *Pyrococcus furiosus* DSM3638 [31] using minimap2 (Release 2.17-r941, https://github.com/lh3/minimap2)[40]. Output alignments in the SAM format were generated with the recommended options for noisy Nanopore Direct RNA-seq (-ax splice, -uf, -k14) and also with (1) -p set to 0.99, to return primary and secondary mappings and (2) with --MD turned on, to include the MD tag for calculating mapping identities. Alignment files were further converted to bam files, sorted and indexed using SAMtools[41]. Strand-specific wig and bigwig files were finally created using bam2wig (Version 1.5, https://github.com/MikeAxtell/bam2wig). To evaluate the alignments, we first calculated the aligned read length by adding the number of M and I characters in the CIGAR string[13]. Based on this, the mapping identity was defined as (1-NM/aligned_reads)*100, where NM is the edit distance reported taken from minimap2. Read basecalling and mapping metrics can be found in Supplementary Table 1.

Gene expression analysis

For transcript abundance estimation we applied featureCounts (Rsubread 1.32.4) allowing that a read can be assigned to more than one feature (allowMultiOverlap = TRUE) and applying the setting for long reads (isLongRead = TRUE)[42]. Calculations were performed based on the genome coordinates of genomic feature types (tRNA, rRNA, protein-coding genes). For the abundance comparison to Illumina-sequencing, we applied a regularized log transformation from the DESeq2 package that transforms counts on a log2 scale, normalizing for the library size and minimizing differences between samples with small counts[43] (raw count data for TEX samples in Supplementary Table 2).

Poly(A) tail analysis

Poly(A) tail length was estimated by nanopolish following the recommended workflow (Version 0.10.2, https://nanopolish.readthedocs.io/en/latest/quickstart_polya.html)[44].

Detection of transcriptional units and annotation of transcription start sites and transcription termination sites

The definition of transcriptional units (TU) and our strategy to detect and annotate them was based on a recent study that re-defined the bioinformatical search for transcriptional units (TU)[29]. The TU annotation was performed in a two-step process in the following way: First, TU clusters were defined by collapsing all reads that overlap and fulfill certain criteria that are commented on extensively in the available code for this study (https://github.com/felixgrunberger/Native_RNAseq_Microbes). In short, reads were filtered out that did not align protein-coding genes (CDS) or tRNAs, had a mapping identity below 80%, were spliced, were shorter than 50% of the gene body and did not cover either the 5´ or the 3´ untranslated region. The remaining overlapping reads were collapsed in a strand-specific manner and merged.

Finally, the collapsed reads that represent the TU cluster, were split according to the coverage drop at the 3´ region of a gene. This was achieved by calculating the sequencing depth in a window of 20 nt upstream and downstream of the corresponding TTS and applying a deliberately low

threshold of 1.5x (higher coverage upstream compared to downstream, see transcriptional unit table in Supplementary Table 5).

TSS were predicted by calculating the median start position of all reads that map to one gene and cover the 5´ part of a CDS. To address the 3´ coverage bias and the underrepresentation of reads that map to the 5´end and also for the 12 missing nucleotides at the TSS in general, all reads starting at least 20 nt downstream of the annotated gene start were included. To not exclude too many reads, the position of TTS were predicted similarly, by also including reads that have end positions starting from 20 nt upstream of a gene end (TSS table in Supplementary Table 3, TTS table in Supplementary Table 4).

For the analysis of prokaryotic promoter elements, the sequences from 58 bases to 11 bases upstream of all uncorrected TSS that are starting a TU were extracted, to identify relevant motifs using MEME with default options except for a custom background file, calculated from intergenic sequences of the respective organism[45]. The analysis of terminator sequences was performed comparably by extracting all TTS that are located at the end of a TU and searching for terminators in a sequence window from -45 to +45 from the TTS using MEME and the custom background model. Heatmap analysis of motif positioning was performed by importing MEME FASTA information into R. Metaplots of the nucleotide enrichment analysis (compare[26,46]) were calculated by comparing the genomic sequences surrounding a TTS in a window from -45 to 45 to randomly selected intergenic positions (subsampled, n = 10000).

Modified base detection

The performance of two different approaches (Tombo vs. error and basecalling properties) for the detection of modified bases was evaluated:

(1) We used Tombo (Version 1.5, https://nanoporetech.github.io/tombo) to identify modified bases based on a comparison to a theoretical distribution (*de novo* model) and based on the comparison to a reference data set (sample-compare model)[47]. Briefly, for Fig. 6a reads mapping to 16S rRNA were preprocessed, resquiggled and the raw signal plotted at a specific genomic coordinate using the respective plotting command (tombo plot genome_locations). In addition, the probability of modified bases was calculated using the detect_modification de_novo command. For Fig. 6b the signals were calculated for both samples (wildtype and deletion mutant) and compared using the *control-fast5-basedirs* and *overplot Boxplot* option. For Fig 6c a reference data set was

created by sorting the reads mapping to the 16S rRNA based on the position of the read start (before or after gene start), thereby dividing the data set in reads that belong to mature and unprocessed 16S rRNAs. The two read sets were used as input for the sample-compare method. The two different models (*de novo* and sample-compare using non-mature 16S reads as a reference set) were applied to the complete 16S, base-modification-probabilities extracted from the tombo output and plotted using a custom R script.

(2) For calculating the frequency of correct, deleted and wrong nucleotides at a genomic position the function pileup from the Rsamtools package was used (https://rdrr.io/bioc/Rsamtools/). Plots were generated using custom R scripts. The results were compared to known modification sites in 16S rRNA for *Escherichia coli*[48] and *Haloferax volcanii*[49]. Given the high degree of conservation of the modifying enzymes in archaea and because there was no reference set available for *Pyrococcus furiosus*, the results were compared to the five modified positions that are described for *H. volcanii* (910, 1352, 1432, 1450/1451 correspond to 938, 1379,1469,1486/1487 in *P. furiosus*).

Public data

In addition to the in-house generated data, we made use of other published sequencing data sets and data repositories that are described in the following.

Transcriptional start sites

For all three model organisms, global transcriptional start sites were mapped recently using differential RNA sequencing[27,30,31]. Position data were extracted from the Supplementary data of the publications and compared with the TSS described in the ONT data sets given that a start site was found in both data sets.

Transcriptional termination sites

So far there are no transcription termination data sets available for *H. volcanii* and *P. furiosus.* The 3´ UTR lengths of the *E. coli* ONT set were compared to TTS predicted based on the Term-Seq method [26]

304

305 *Transcriptional units*

306 The widely used database DOOR2[50] was used to compare the TU annotation for both archaeal sets.
307 For *E. coli* a more recent, but also purely bioinformatical prediction, served as a reference set[29].

308

309 *Gene expression comparison*

310 For *P. furiosus* gene abundances from ONT data were compared to fragmented RNA sequencing
311 data of mixed growth conditions (conditions, library, sequencing, mapping described in[31]), by
312 applying a regularized log transformation as described earlier[43]. For *H. volcanii* comparison, raw
313 reads of a mixed RNA sequencing were extracted from the SRA (SRR7811297)[28] trimmed using
314 trimmomatic[51], (leading:20, trailing:20, slidingwindow:4:20, minlen:12), mapped to the reference
315 genome using bowtie2 (-N 0, -L 26)[52], converted to sorted bam files using samtools[41] and compared
316 to ONT data as described for *P. furiosus.* Illumina RNA sequencing data for *E. coli* were also
317 extracted from the NCBI (SRP056485, 37°C LB), analysed as described for *H. volcanii* Illumina data
318 and also compared to the ONT reference data.

Results

Library preparation for Nanopore native RNA sequencing of bacterial and archaeal transcriptomes

We performed whole-genome native RNA sequencing using the Nanopore sequencing technology (referred to as Nanopore native RNA sequencing in this work) that can be applied to any prokaryotic organism. The key steps of the library preparation are shown in Fig. 1a: after enzymatic polyadenylation, the RNA is reverse transcribed to increase sequencing throughput (recommended by ONT)[16]. Subsequently, a motor protein is added that feeds the RNA into the Nanopore. Following this workflow, native RNAs from prokaryotic organisms can be sequenced. Depending on the necessary sequencing depth, the libraries were barcoded using poreplex (https://github.com/hyeshik/poreplex), since this is not yet supported by the official kits and protocols from Oxford Nanopore. To identify primary and processed sites in both mRNA and rRNA, we partially digested 5´-monophosphorylated transcripts using the terminator exonuclease (TEX) and compared them to non-treated samples (see Supplementary Fig. 1a). In contrast to many Illumina sequencing-based approaches that make use of a specialised library preparation design to tackle a well-defined question[8], we evaluated the potential of native RNA sequencing to analyse multiple transcriptomic features simultaneously including the identification of *cis*-regulatory elements that govern transcription, the analysis of operon structures and transcriptional boundaries, rRNA processing and 16S rRNA modification patterns (see Supplementary Fig. 1b).

Sequencing yield and quality control of raw Nanopore reads

Native RNA sequencing was performed for three prokaryotic organisms: the bacterial model organism *Escherichia coli*, the halophilic archaeon *Haloferax volcanii* and the hyperthermophilic archaeon *Pyrococcus furiosus*. In order to show that native RNA sequencing can be applied to a wide variety of prokaryotic organisms, we specifically chose (i) organisms from the bacterial and archaeal domain of life with *P. furiosus* and *H. volcanii* belonging to the Euryarchaeota, (ii) organisms that are classified as mesophilic (*E. coli*, *H. volcanii*), hyperthermophilic (*P. furiosus*), or halophilic organism (*H. volcanii*) and (iii) organisms that differ significantly in their GC-content (*E. coli*: 50.8%[53], *H. volcanii*: 65%[39], *P. furiosus*: 40.8%[31]). The prepared libraries were sequenced on a

MinION device and reads were collected over 48 hours on R9.4 flow cells (see Supplementary Fig. 2a). The total number of reads in the TEX-treated libraries were sufficient to achieve good coverage of the transcriptome (*E.* coli: 9.2x, *P. furiosus*: 15.0x, *H. volcanii*: 10.3x, see Supplementary Table 1) to perform transcriptional unit annotation and accurate determination of transcript boundaries. Before mapping the reads to the reference genomes, the quality of the sequencing runs was evaluated based on raw read length distribution and quality of reads estimated by Guppy (see Supplementary Fig. 2b,c). To verify that no problems occurred during sequencing or library preparation a poly-adenylated spike-in control (yeast enolase) was used. The control showed a uniform length distribution (median lengths between 1212 and 1306 nucleotides) and a very good read quality (median quality between 10.8 and 12.2) in all samples, therefore, excluding any bias (see Supplementary Fig. 2b,c, Supplementary Table 1). Due to large differences in the composition of the RNAs used to train the basecaller and the sequenced reads, that contain mostly ribosomal RNAs known to harbor base modifications, the overall quality is lower than for the spike-in control in every data set[54]. To prevent useful reads from being discarded, we decided to also include reads that fall below the standard quality score threshold of 7 and continued with the mapping of the reads to the reference genomes[13].

Analysis of mapped reads

An advantage of the long-read Nanopore sequencing technique is that native RNA strands can be sequenced directly as near full-length transcripts[55]. This is also reflected in the sequenced data sets as aligned lengths up to 7864 nt can be observed (see *H. volcanii* NOTEX sample with a median read length of 1390 nt, Supplementary Fig. 3c). The majority of reads from *E. coli* and *H. volcanii* samples mapped to the 23S rRNA (Fig. 1b, see Supplementary Fig. 3a). In general, the read identity of CDS-mapping reads is higher than for rRNA mapping reads, but lower than the spike-in control (see Supplementary Fig. 3b,c). It is noteworthy, that a minimum aligned length of about 100 nt appears in the data sets (see Supplementary Fig. 3d) when using the native RNA sequencing protocol and minimap2 as mapping tool[16,40,56]. Unaligned reads had a median read length of 191 nt, in contrast to 572 nt for aligned reads (all data sets combined) suggesting that short reads could not be aligned properly. As small RNAs, CRIPSR-RNAs or tRNAs fall below this threshold, we

376 excluded these RNAs from further analysis. While short transcripts are a problem, longer RNAs are
377 sequenced without loss in quality (see Supplementary Fig. 3e). As the raw read quality correlates
378 with the mapping identity of the reads, problems during sequencing can be live-monitored in
379 MinKNOW and the run can be canceled allowing the loading of a new library (see Supplementary
380 Fig. 3f). Since the subsequent analysis of transcriptional units is heavily dependent on the integrity
381 of the data, we verified the data integrity in the next steps. Sources for bias could emerge (i) from
382 the poly(A)-tailing efficiency that might differ between transcripts and (ii) from the loss of certain
383 transcripts during library preparation or sequencing problems. Comparing the initial amount of
384 RNA with the reverse transcribed DNA, we observed that the enzymatic addition of a 3´ poly(A)-
385 tail is not very efficient (~10%). In general, the addition of poly(A)$_{20}$ (length of the reverse
386 transcription adapter) is sufficient to allow for the annealing of the poly(T)-adapter required for
387 reverse transcription and sequencing. This goes in line with the shortest median length we
388 observed for the 5S rRNA (see *E. coli* TEX sample) (see Supplementary Fig. 4). For most of the
389 transcripts, a tail with 50 to 100 nt was detected. The overall correlation of transcript abundances
390 calculated from sequencing data using Nanopore or Illumina technology was very high given that
391 only mixed conditions could be compared in the case of *H. volcanii* and *P. furiosus* (see
392 Supplementary Fig. 5a,b,c, transcript abundance data for TEX samples in Supplementary Table 2).
393

394 **Mapping of transcriptional boundaries**

395 *Transcription start sites*

396 Transcription start site (TSS) and transcription termination site (TTS) detection was based on the
397 determination of transcriptional units (TU) (compare material and methods section)[29]. In total, we
398 found substantial overlap between TTS detected by Illumina d(ifferential) RNA-seq[2–4] and
399 Nanopore native RNA-seq (Fig. 2a)(positions of TSS derived from ONT TEX-treated samples in
400 Supplementary Table 3). The portion of ONT only and Illumina only is mostly caused by the
401 different algorithms used and the limited sequencing depth in the ONT data sets. Strikingly, the
402 median 5´ untranslated region (UTR) lengths were very similar when data from ONT native and
403 Illumina-based RNA sequencing were compared (*E. coli*: 68 ONT vs. 62 Illumina; *P. furiosus*: 23
404 ONT vs. 13 Illumina; *H. volcanii*: 1 ONT vs. 0 Illumina). Please note that TSS-mapping based on

Nanopore native RNA-seq data must be corrected by 12 nucleotides (Fig. 2b). It has been observed previously that about 12 nt are missing at the 5'-end of the sequenced RNAs. This observation can be explained by a lack of control of the RNA translocation speed after the motor protein falls off the 5´ end of the RNA (see Supplementary Figure 6a)[56,57]. Promoter analysis confirmed the presence of well-known sequence motifs of bacterial and archaeal promoters[27,31,58]. This includes the TATA-box and B-recognition element (BRE) characteristic for archaeal promoters and the -10 element in bacterial promoters (Fig. 2d). The -35 element in *E. coli* has been previously shown to be less enriched compared to the -10 site[5], which might explain why this element cannot be detected in the Nanopore data set. To analyse TSSs in more detail, we compared the 5´ UTR lengths for all genes with predicted TSS in ONT and Illumina data sets (see Supplementary Figure 6b,c,d). The overall correlation between the two techniques was very high (see Supplementary Fig. 6). As expected, the correlation improves with increasing sequencing depth for a gene (>5 reads). While TEX-treatment is a common way of predicting TSS in Illumina sequencing, we observed that it is not necessary for ONT data as very similar TSS are found in both TEX and NOTEX data sets (ρ = 0.86) (see Supplementary Figure 6e).

Transcription termination sites

Native RNA reads are sequenced in the 3´ to 5´ direction, which is a major advantage in the detection of termination sites as any bias introduced after polyadenylation can be excluded. Currently, no reference data sets for the archaeal organisms were available, the distribution of 3´ untranslated regions (3´UTRs) in *E. coli* ONT data closely resembles the data from a previous Term-seq study[26]. Untranslated regions at the 3´end of an annotated transcript are longer and more uniformly distributed in *E. coli* than in the two archaeal species (Fig. 2c). In total, 1321 TTS in *E. coli*, 856 in *P. furiosus* and 1461 in *H. volcanii* were analysed (positions of TTS in TEX-treated samples in Supplementary Table 4). A meta-analysis of all TTS surrounding regions revealed different sequence-dependent termination mechanisms that were confirmed using motif scanning (Fig. 2e, see Supplementary Fig. 7a-d). Our data suggest that transcription in *P. furiosus* is terminated by a double-stretch of Uridines that are distributed over a length of 22 nt, a finding that is in line with the terminator sequences detected by Term-Seq in *S. acidocaldarius*[46] and similar to

the $U_{(8)}$ sequence in *Thermococcus kodakarensis* by an *in vivo* reporter assays[59]. The termination motif found in *H. volcanii* is a $(U)_{4}$-sequence and located right before the TTS (see Supplementary Fig. 7d). The motif locations for both *Haloferax* and *Pyrococcus* ONT sets suggest that accurate TTS detection of transcripts terminated by poly(U) stretches is currently not possible. We observed that homopolymer sequences are removed during trimming of the reads, which leads to TTS positions that are positioned upstream of the poly(U) signal. Analysing individual transcripts in *H. volcanii* and *P. furiosus*, we found that a single transcript can exhibit diverse 3' ends. This is true for the Pilin transcript in *H. volcanii* and the Alba transcript in *P. furiosus*, respectively (Supplementary Fig. 8). Both genes are highly expressed and some transcripts carry extended 3' UTRs. While the majority of transcripts are terminating at the first poly(U) stretch, a subset of transcripts is substantially longer and terminates at a second poly(U) termination signal (Supplementary Figure 8). Interestingly, homogeneous short poly(U) signals are found both at the canonical termination site and the termination site of the elongated 3'UTR in the case of the Pilin transcript in *H. volcanii.* In contrast, the Alba gene transcripts in *P. furiosus* are terminated at both sites via long consecutive poly(U) stretches. Surveying the heterogeneity of the transcripts with an extended 3'UTR, we found a heterogeneous distribution in the length of the transcripts, which possibly indicates a step-wise trimming of the 3'UTR to yield the mature RNAs.

As observed for *E. coli* termination sequences, Cytosines are enriched over Guanosine adjacent to the TTS in *P. furiosus* (Fig. 2e). Termination motifs detected in the *E. coli* data set cover both intrinsic (poly(U)) and Rho-dependent termination signatures (see Supplementary Fig. 7a,b)[26]. We classified the stem-loop forming motif as Rho-dependent signatures, as these were not enriched in the poly(U) terminated genes. However, the stem-loops could potentially also represent processing or pause sites.

Annotation of large transcriptional units

Long-read sequencing of native full-length RNAs has the potential to improve and facilitate genome-wide transcriptional unit (TU) annotation, which can be visually explored in a genome browser coverage track (Fig. 3a). For whole-genome analysis, the annotation strategy was based on two major observations: First, during RNA preparation, RNA processing or degradation can

463 occur, which limits the probability of sequencing an RNA in its native form as the percentage of
464 full-length transcripts decreases with expected gene size (see Supplementary Fig. 9a). Secondly,
465 we detected a decrease in coverage from the 3´ to 5´end of the RNA in all RNA classes except for
466 the spike-in control (see Supplementary Fig. 9b), which is a limitation reported in the
467 literature[13,60,61]. Therefore, we assume that not Nanopore sequencing but library preparation
468 causes this problem. Based on this information, we developed a strategy that first collapses all
469 overlapping reads and then splits them according to a significant coverage drop on the 3´ ends
470 (annotation of TUs based on this strategy in Supplementary Table 5). We compared the results to
471 database annotations and found that most of the differences are either caused by the low
472 sequencing depth or by single-unit operons that have been collapsed and are now two-unit
473 operons in the ONT data sets (see Supplementary Fig. 10a,b)[29,50]. Even though limited read
474 availability is a concern in all data sets, many large operons were detected for all organisms (see
475 Supplementary Fig. 10c). In case of limited bioinformatical resources, TUs can be explored visually
476 in a genome browser, which is mostly not possible for Illumina reads (Fig. 3, see Supplementary
477 Fig. 11,12). It further allows a quantitative analysis of individual transcripts in relation to other
478 elements of the TU and performs much better than pure bioinformatical prediction or molecular
479 biology methods (RT-PCR) as shown for the archaellum operon in *P. furiosus* (Fig. 3)[50,62]. Here, it
480 was possible to (i) detect multiple transcription units forming this cluster, (ii) confirm
481 transcriptional start sites and (iii) to confirm that flaB0, the protein that is referred to as the major
482 archaellin in *P. furiosus* [2,6], is transcribed in large excess over the other archaellum genes. The
483 largest TU cluster in *H. volcanii* consists of 25 ribosomal protein genes. Based on the native RNA-
484 seq data, the analysis shows that this operon is split into two transcription units. This shows that
485 the ONT native RNA sequencing method provides the opportunity to correctly annotate
486 transcriptional units thereby outperforming the bioinformatics-only prediction as well as the
487 visual inspection of Illumina coverage (see Supplementary Fig. 11). Besides, we confirmed the
488 complex transcription pattern of the major ribosomal protein gene cluster in *E. coli* that stretches
489 over more than 10 kB, including the accurate determination of TSS and TTS and a putative cleavage
490 site in the secY gene (see Supplementary Fig. 12) [63].

Detection and confirmation of rRNA processing in *E. coli*

Next, we aimed to analyse the multi-step rRNA processing pathway which is the major RNA maturation pathway in any prokaryotic cell. We first focus on the *E. coli* data set as the processing of bacterial rRNAs is well characterized[64–66]. Ribosomal RNA in *E. coli* is transcribed from 7 independent rDNA operons, containing the mature rRNAs (16S, 23S and 5S rRNAs) and some tRNAs which are intersped by RNA spacers elements[67]. In agreement with a previous study, transcription of *rrnC* from two promoters (transcription start sites at -293 and -175) was detected accurately (Fig. 4a,b)[68].

The rRNA maturation process, which requires the action of well-defined ribonuclease activities, culminates by the formation of stoichiometric amounts of mature 16S, 23S, and 5S rRNAs[64–66]. Unexpectedly, the sequencing efficiency of mature 16S rRNA was lower than the 23S rRNA (Fig. 4a). Moreover, although we could properly detect RNA of similar size or longer (see above), the short-lived full rDNA operon transcript detected in RNase III deficient strain[69], is not observed using our experimental set-up. In contrast, the downstream pre-rRNAs which are generated by the action of RNAse III were detected (Fig.4). Among these intermediates, the 17S pre-rRNA with 115 additional nt at the 5´ end and 33 nt at the 3´ end of the 16S rRNA, was identified (Fig. 4b). Additional 16S rRNA processing sites (RNase E: -66, RNase G: 0 5´mature) are also detected (Fig. 4 b,c). Recent studies were suggesting that the 3´end of the mature 16S rRNA is generated by 3´-5´ ribonucleases activity and/or endonucleolytic cleavage[70,71]. In addition to the mature 16S rRNA 3´end, we could observe reads of decreasing length spanning from +33 nt (RNaseIII) to 0 (related to 16S rRNA 3´end), that could account for 3´-5´ exonuclease processing activity (or degradation during sample preparation). Together we could accurately identify all the known processing sites at nucleotide resolution in wildtype *E. coli*.

Insights into archaeal ribosomal RNA processing

In comparison to bacteria or eukaryotes, ribosomal RNA processing in archaea is still poorly understood[70,72–74]. Our current knowledge suggests that the primary polycistronic rRNA precursor contains processing stems formed by the 5´-leader and 3´-trailer sequences surrounding the 16S and 23S rRNAs[72,73,75,76]. In Euryarchaeota, the 16S and 23S rRNAs are additionally separated by the

520 presence of an internal tRNA. In most archaea, the 16S and 23S rRNA processing stems contain a
521 bulge-helix-bulge motif which is, in the context of intron-containing tRNA, recognized by the
522 splicing endonuclease endA[75–77]. Similar to intron-containing tRNA maturation, processing at the
523 bulge-helix-bulge motifs results in the ligation of the resulting extremities, thereby generating the
524 archaeal specific circular pre-16S and circular pre-23S rRNAs[73,76,78,79]. The molecular mechanisms
525 by which the circular pre-rRNA intermediates are further processed into linear mature rRNA are
526 unknown[73,76].
527 By analysing single reads, we aimed to confirm and expand our knowledge on the poorly
528 characterized multi-step ribosomal maturation process in *P. furiosus* and *H. volcanii* (Fig. 5a,b,c,
529 see Supplementary Fig. 13a,b)[49,73,75,80]. Due to a more limited number of reads in the multiplexed
530 *H. volcanii* samples, we have mostly focused on the *P. furiosus* dataset. As expected, almost all reads
531 are categorized as fully maturated transcripts of the single 16S/23S rRNA cluster that do not
532 contain an extended 5′ or 3′ region (see Supplementary Fig. 13c). Surprisingly, 5′ and 3′end
533 positions did not precisely match the availablle annotations at NCBI
534 (https://www.ncbi.nlm.nih.gov/genome/) or the archaeal genome browser (AGB,
535 http://archaea.ucsc.edu). In contrast, these mature rRNA extremities did match our independent
536 experimental validations by primer extension analysis of the 5′ends of the 16S and 23S rRNAs of
537 *H. volcanii* (Supplementary Fig. 15f). In addition, the majority of the expected processing sites,
538 notably at the predicted or known bulge-helix-bulge motifs[78], were faithfully observed in *P.*
539 *furiosus* and *H. volcanii* (see Supplementary Fig. 14,15).
540 Despite the high sequencing depth of the (pre-)rRNA, we did not detect a full-length precursor
541 consisting of the 16S leading-16S-tRNA-23S- 23S trailing elements in *P. furiosus and H. volcanii,*
542 suggesting that early rRNA processing occurs rapidly in these cells. The remaining reads were
543 grouped according to their 5′ leading and 3′ trailing lengths into five additional categories that are
544 less abundant (with less than 1600 counts in each group): (1) a processing variant that entails the
545 16S rRNA leading/trailing sequence-tRNA-23S rRNA leading/trailing sequence, (2) probably
546 fragmented full-length precursor rRNAs, (3) 16S rRNA leading/trailing sequence-tRNA, (4) 16S
547 rRNA leading/trailing sequence-tRNA-23S rRNA and (5) tRNA-23S rRNA (Fig. 5, see
548 Supplementary Fig. 13). The early 16S rRNA leading/trailing sequence-tRNA-23S rRNA trailing
549 sequence precursor (1) or 16S rRNA leading/trailing sequence (3) generated by cleavage and

reciprocal ligation of the pre-16S and pre-23S rRNAs at the predicted bulge-helix-bulge motifs were detected very accurately, and are reminiscent of previous observations[76] (Fig. 5, see Supplementary Fig. 13d). Given the number of reads, the direction of ONT sequencing from 3´ to 5´ and the accurate mapping, it is unlikely that the additional putative rRNA precursors (variants 4, 5) carrying the leading sequence in combination with 23S rRNA and tRNA-23S rRNA are all arising from experimental artifact (Fig. 5a,b) (see also Discussion). This processing variation could have been overlooked in the past in the analysed archaea, as these are only minor fractions of the transcripts present in the cell or are specific to *P. furiosus*. Shortened 23S rRNA reads (variants 4, 5) and the overall visible 3´ to 5´ coverage drop can most likely be explained by fragmentation and degradation during library preparation, secondary structures that prevent efficient loading into the nanopore or truncation of RNA reads caused by physical artifacts that have been described before and include problems during RNA translocation or disruptions of the ionic current signal[13,52,56]. Furthermore, it cannot be excluded that non-mature reads are categorized as fragmented reads and will be assigned as mature 16S or 23S rRNA.

Highly enriched positions that differ from the annotation of the mature rRNAs and cannot be assigned to a processing site may represent a wrong annotation of the rRNA locus or currently unknown processing sites in the two archaeal species[31,39]. To provide further examples that show the potential to describe maturation and processing events for prokaryotic rRNA, we sequenced an *H. volcanii* wildtype strain grown under low salt conditions (see Supplementary Fig. 15b,c)[28]. Under these conditions, a 16S rRNA precursor with an extended 5´UTR (position -121 – 5´bulge) appears that is only enriched in this condition. Quantification by comparison with the NOTEX wildtype set confirms the previous detection of this rRNA variant in a gel-electrophoretic analysis of the rRNA (see Supplementary Fig. 15b)[28].

Towards mapping of RNA base modifications

More than 160 types of modified bases have been described in RNAs so far[48]. In contrast to other sequencing techniques, Nanopore-based sequencing offers the possibility to detect base modifications directly, as these modifications lead to an electric current signal that differs from the expected theoretical distribution obtained by the unmodified nucleotides sequence[14,15,20,81]. Based

on this approach, we benchmarked the potential to detect known modification sites in 16S rRNA. Before expanding the analysis to the full-length mature rRNA, we focused on the dimethylation ($m^6{}_2A$) introduced by the enzyme KsgA at position 1450/1451 in *H. volcanii*[49,82]. Using the *de novo* model in Tombo, a suite of tools for the identification of modified nucleotides from nanopore sequencing data (updated version from Stoiber et al.[47]), the calculated probability of a modification was very high for the stretch of Guanosines adjacent to position 1450 (Fig. 6a). Mapping to single-nucleotide resolution is difficult as more than one base contributes to the actual electric current signal in the nanopore[83]. In the next step, a comparison of a wildtype sample to a deletion mutant of KsgA confirmed the absence of the dimethylation in the deletion strain and confirmed that the $m^6{}_2A$ modifications and not the homopolymer-G-stretch are responsible for the altered current signal (Fig. 6b). The analysis further revealed a reduced signal variability at non-modified positions between the two samples in comparison to the theoretical distribution, which leads to less false positives in the statistical analysis. As the *de novo* model has a higher error rate compared to the sample-compare model, we generated two read sets by sorting full-length 16S rRNA transcripts according to their 5´ leading length to mature (n: 2744) and pre-rRNAs (n: 102) in the *Haloferax* wildtype data set (Fig. 6c). The calculated base modification probabilities and electric current signal distributions for this model are very similar to the wildtype/Δ*ksgA* set, indicating that the two $m^6{}_2A$ modifications were not quantitatively introduced in the pre-16S-rRNA reads (Fig. 6d). This finding is in line with a proposed "quality control" function of KsgA during late biogenesis of the small ribosomal subunit[84–86] and further highlights the potential to detect the introduction of base modifications at a different stage of rRNA maturation.

Next, we applied the described *de novo* and sample-compare model to the full 16S rRNA of all three organisms and compared the results to positions known from literature[48,49]. Applying a probability threshold of 0.5, we found that 60% (*de novo*) and 1.89% (sample-compare, 28 bases) of the bases in the 16S rRNA are modified in *H. volcanii* (Fig. 6e), which emphasizes the better performance of the sample-compare model by the vastly reduced number of false-positivly detected RNA modifications. 10 of the 28 predicted bases (please note that approximately five bases contribute to the signal in the nanopore and that a base single modification therefore influences the current signal at the surrounding bases as well) can be assigned to known base modifications at positions 910, 1430 and 1450/1451[49]. Using the same threshold for *E. coli* (*de novo*: 50.5%, sample-compare:

609 0.1%, Fig. 6f) and *P. furiosus* (*de novo*: 64.9 %, sample-compare: 17.7 %, Fig. 6g), better accuracy
610 of the sample-compare model could be confirmed. Moreover, these data support the long-standing
611 hypothesis that hyperthermophilic organisms might stabilize their rRNAs by a higher degree of
612 RNA modifications[87,88].
613 It is noteworthy that the predicted large number of modifications lead to a loss in read quality. This
614 is for example seen when analysing the *Pyrococcus* data set, as the quality score is also based on
615 the comparison of detected to theoretical read signal (compare Supplementary Fig. 3c). To look at
616 this observation in more detail, we compared the results from the raw squiggle Tombo approach
617 that is not dependent on mapping of the reads to the mapping properties of the reads. This
618 methodology is based on an approach that was recently used to identify m^6A RNA modifications
619 from native RNA reads with an accuracy of ~90%[14]. After mapping, the frequency for the
620 respective nucleotide at every position in the 16S rRNA to be correct, wrong or deleted was
621 calculated. Homopolymer stretches in the RNA increase the likelihood that the correct number of
622 nucleotides is not correctly identified and this, in turn, leads to an increased number of deleted
623 nucleotides, which is also visible in the *Haloferax* data sets at the G-stretch adjacent to the $m^6{}_2A$
624 modifications (Supplementary Fig, 16a,b). The overall frequency of wrongly assigned and deleted
625 nucleotides was much higher in *P. furiosus* (16.4%) than in *E. coli* (10.7%) and *H. volcanii* (10.4%)
626 (Supplementary Fig. 16c). Interrogating the full-length 16S rRNA, we found that miscalled
627 positions are equally distributed among the whole transcript (Supplementary Fig. 16d,e,f).
628 Statistical analysis of miscall-frequencies revealed a significant difference between already known
629 modification sites and non-modified positions in *Haloferax* and *Escherichia* suggesting that even a
630 very simple model can detect RNA base modifications when wrongly assigned bases considered.
631 The situation is different for *Pyrococcus*: Native Nanopore data suggest that the rRNA is heavily
632 modified. However, the number, exact positions and chemical identity of these modifications are
633 currently unknown. Hence, an analysis of how often base modifications lead to wrongly identified
634 nucleotides in the 16S rRNA sequence cannot be performed (Supplementary Fig. 16c,g).
635
636

Discussion

Performing whole-genome native RNA-seq study in prokaryotes provided us with a wealth of information on transcriptional and post-transcriptional processes in *E. coli* and the archaeal model organisms *H. volcanii* and *P. furiosus*. Here, we will mostly discuss new biological insights that emerged from our study. Additionally, we will reflect on the advantages and disadvantages of Nanopore native RNA-seq.

Insights into transcriptional processes

Bacterial and archaeal transcription is an intensely studied molecular process and the mechanisms of basal transcription are well understood[89]. Native RNA sequencing allowed us to retrieve accurate information of transcript boundaries on both 5´ and 3´ ends. Our data show that 3' UTRs length distributions are comparable between *E. coli*, *P. furiosus* and *H. volcanii* with the majority of mRNAs showing a length between 30 -70 nt. Similar to bacteria, archaea encode large numbers of small non-coding RNAs[90]. However, many regulatory events that involve the regulation via small RNAs take place at bacterial 5' UTRs[91]. We and others found that 5' UTRs are significantly shorter in archaea supporting the idea that post-transcriptional regulation is mediated via the 3' rather than the 5' UTR in archaea[92]. Additionally, we determined transcription termination sites, which are well analysed for bacterial species but only a few studies focused on archaeal termination mechanisms, especially on the genome-wide level. In both archaeal species, poly(U) stretches were overrepresented at termination sites agreeing well with termination sequences found in *Sulfolobus* and *Methanosarcina*[46]. Interestingly, the majority of TTS found in *Pyrococcus* transcripts is composed of two U-stretches with at least five consecutive uridine bases while a subclass of *Haloferax* transcripts is almost exclusively terminated by a single U-stretch with four uridine bases. It has been shown that a five base U-stretch is sufficient to induce termination *in vitro* [59,93,94]. Similar observations were described in a recent study by Berkemer et al, which identified a poly(U)$_4$ stretch to be the termination signal in intergenic regions[95]. However, the current data set suggests that this short termination signal might be a specific feature for a subclass of *Haloferax* transcripts resembling the poly(U) termination motif found in *E. coli*. All other archaeal organisms (*P. furiosus*, *M. mazei*, *S. acidocaldarius*) investigated so far terminate transcription at multiple

665 consecutive poly(U) stretches. Possibly, *Haloferax* relies on additional termination signals or yet
666 unknown termination factors. A putative candidate is archaeal CPSF1 (aCPSF1), a recently
667 described archaeal termination factor[96] that is widespread in archaea. aCPSF1 acts as ribonuclease
668 that cleaves transcripts after a poly(U) stretch to trim transcripts and facilitates transcription
669 termination in *Methanococcus maripaludis.* The arising 3' UTR isoforms were detected using Term-
670 seq analysis[96]. We also observed heterogeneity in the case of the Pilin and Alba transcripts,
671 respectively, that are distinguished by varying lengths of the 3' UTR suggesting that aCPSF1 might
672 trim a subset of genes in *H. volcanii* and *P. furiosus*. It is noteworthy that 3' UTR isoforms were also
673 detected in Term-seq studies with *Sulfolobus* and *Methanosarcina*[46]. However, in contrast to the
674 Pilin and Alba transcripts, the 3' UTR isoforms arise from termination at different sites of a single
675 continues poly(U) stretch suggesting that the isoforms arise from a stochastic termination process
676 of the RNA polymerases at an extended poly(U) stretch at the end of the gene. Taken together,
677 these data demonstrate that a variety of termination mechanisms (that can even co-occur in the
678 same cell) can be found in archaea ranging from stochastic intrinsic termination at extended
679 poly(U) stretches (*Pyrococcus, Sulfolobus, Methanosarcina*), to abrupt termination at short uridine
680 tracts (*H. volcanii*) and factor-dependent termination that results in trimming of the 3'UTR (*H.*
681 *volcanii, P. furiosus, M. maripaludis*).

682 In the context of transcription, the long and overlapping native RNA reads helped us to analyse the
683 transcriptional landscape at multigene operons. More specifically, we focused on the archaellum
684 operon as the transcription unit assignment remained ambiguous so far[62]. In contrast to
685 bioinformatical and Illumina RNA-seq-based predictions and attempts to unravel the TU via
686 primer extension experiments, we found that the archaellum operon in *P. furiosus* is transcribed
687 in multiple units with highly diverse abundances. The *fla*B0 gene encodes the major flagellin
688 protein that forms the filament of the archaellum and therefore, the organism has to produce this
689 protein in large quantities as apparent from the expression level[62]. Interestingly, FlaD is expressed
690 at comparably high levels as well as supporting the idea that FlaD is a major constituent of the
691 archaellum in *P. furiosus*. It has been speculated that FlaD forms the cytosolic ring of the archaellum
692 that anchors the filament in the outer membrane [97]. The identity and functional role of FlaD are,
693 however, not known so far.

694

695 **Insights into rRNA processing in archaea**

696 In this study, we have assessed the suitability of native RNA sequencing to obtain information on
697 the rRNA maturation pathway of different prokaryotes. Ribosomal RNA maturation proceeds via
698 the coordinated and defined order of ribonucleases action (exonucleolytic and/or endonucleolytic
699 cleavages) which generate pre-rRNA intermediates with defined premature rRNA
700 sequences[66,74,98,99]. The establishment of faithfull rRNA maturation maps in model organisms, like
701 *E. coli*, *S. cerevisiae* or human cell culture has required numerous analyses over the past decades
702 [66,74,98,99], and remains a technical challenge. Therefore, methodologies that might accelerate the
703 systematic analysis of rRNA maturation pathways across the tree of life, thereby enabling to
704 unravel the diversity of rRNA maturation strategies need to be established. Beyond the
705 identification of processing sites, the order of the processing events which can be, in part, deduced
706 from co-occurrence analysis of the 5´and 3´extremities is of biological relevance [66,74,98,99]. Whereas
707 we could confirm and extend our general view on the rRNA maturation pathway in archaea, the
708 3´-5´ processivity of Nanopore native RNA sequencing observed for rRNA impedes the accurate
709 analysis of pre-rRNA extremities co-segregation (see Fig. 5 and Supplementary Fig. 9b).
710 Nevertheless, we could confirm the presence of processing sites and pre-rRNA intermediates in
711 the different organisms analysed, with the exception of the archaea specific circular-pre-rRNA
712 intermediates[73,76,78,79], which escape analysis due to the lack of 5´/3´ extremities. Among the
713 identified pre-rRNA intermediates, precursor (4) observed in *P. furiosus*, which include ligation at
714 the bhb motif of the upstream region of the 16S leader and downstream region of the 16S trailer
715 sequences continuous to the downstream tRNA/23S sequences is of particular interest. The
716 presence of this intermediate suggests that the 16S rRNA bulge-helix-bulge processing occurs
717 prior to internal tRNA and 23S rRNA maturation. This observation is also in agreement with our
718 recent *cis*-acting element analysis performed in *H. volcanii* [73,76,79]. In fact, based on this analysis we
719 have proposed a model by which 16S rRNA maturation proceeds and is required for the
720 downstream maturation of the internal tRNA and 23S rRNA. Moreover, we have hypothesized that
721 ligation of the 16S rRNA leader/trailer resulting from the 16S rRNA bulge-helix-bulge maturation
722 process generates a putative new pre-rRNA intermediate which we could observe for the first time
723 in *Pyrococcus* using native RNA sequencing. In addition, the presence of intermediates (1) and (3)

support the idea that the maturation of the co-transcribed internal tRNA is inefficient or inhibited and may preferentially occur after processing of the 16S and 23S rRNA bulge-helix-bulge [suggested in[76,78]].

The biological relevance of the additionally identified intermediates for rRNA maturation is yet unclear and will need to be appraised in the light of additional functional analysis. In conclusion, despite some intrinsic limitations, direct RNA sequencing can be a useful tool to approach intricated maturation pathway like rRNA maturation, and expand our understanding of RNA maturation in prokaryotes.

Towards the mapping of rRNA modification patterns

RNA modifications have been described already in the 50-60s, and has gained significant attention over the last years, under the generic term of epitranscriptome[100–102]. The high-throughput analysis of these post-transcriptional modifications remains challenging and mostly relies on indirect methods, like primer extension stops analysis or chemical recoding/derivation strategies[103,104]. Native RNA sequencing may fill an important gap to systematically analyse RNA modifications on a genome-wide scale. However, global strategies enabling the faithful determination of RNA modification identity and position needs to be developed. Several recent focused analyses have explored different strategies to evaluate the capacity of ONT to accurately detect RNA modifications (e.g. m6A)[14,15,55,105,106]. In this study, we focused on the 16S rRNA which is known to carry different types of RNA modifications which are introduced at different positions and different stages of the small ribosomal subunit maturation[73,107]. Our study provides evidence that RNA modification detection benefits of the use of unmodified/hypo-modified references (in agreement with recent studies[15,106]). In this study, we have used a sample-compare approach analysing pre-rRNAs, which are expected to contain incomplete modification patterns, in comparison to mature rRNAs, which are expected to harbor a completed modification pattern. To validate our approach we have first focus on the almost universally conserved KsgA-dependent dimethylations of the 16S rRNA[15,82,85]. Using our sample-compare strategy and KsgA deletion strain we could unambiguously provide *in vivo* evidence that the archaeal KsgA-dependent methylations of the 16S rRNA are completed at a late stage of the small ribosomal subunit biogenesis. This result is in agreement with previous studies done in yeast and bacteria[84–86].

Moreover, our sample-compare approach also suggests an increased amount of rRNA modifications in the hyperthermophile *P. furiosus,* and a decrease amount of predicted rRNA modifications in halophile *H. volcanii* in comparison to *E. coli*. These differential modification patterns across archaea are in good agreement with previous studies and may reflect adaptation to the environmental conditions that these extremophilic archaea encounter[49,87,88]. Unfortunately, the exact nature of these modifications can not be unveiled yet. To facilitate their high-throughput identification, future studies will require to develop and train algorithms improving the identification confidence of diverse RNA modifications.

Benefits and limitations of Nanopore-based native RNA sequencing

Taken together, a key advantage of the native RNA-seq approach is that multiple features can be addressed at once distinguishing the technique from the Illumina sequencing technology or biochemical assays. ONT sequencing does not require large scale equipment and is a fast method. Moreover, the method does not necessitate a reverse transcription step or PCR amplification thereby avoiding biases introduced by these enzymes. Due to the limitations of the sequencing read analysis platform, ONT sequencing does not accurately detect small RNAs yet. Additional limitations of the native RNA-seq technique are currently (i) the high amount of input RNA required (2-5 µg) to reach good coverage of the transcriptome without rRNA depletion, (ii) the 3' bias during RNA sequencing (iii) limited throughput and (iv) limited possibilities for multiplexing. Although ONT sequencing has a comparably low sequencing accuracy, this did not pose a limitation for our analysis. Due to the extraordinary read length and the sensitivity to base modifications, ONT-based native RNA-seq can provide valuable insights into (r)RNA processing, (r)RNA modifications patterns and the transcription of large operons. Strikingly, ONT-based sequencing is a *bona fide* single-molecule method and hence molecular heterogeneity in the transcriptome can be analysed so that even minor RNA populations can be detected that are inevitably lost in ensemble sequencing approaches.

782 **Data availability**

783 Sequencing data sets (gzipped raw FAST5 files) will be deposited in the Sequence Read Archive
784 (SRA).

785

786 **Code availability**

787 A detailed documentation and code of all essential analysis steps (used tools and custom Rscripts)
788 are available from https://github.com/felixgrunberger/Native_RNAseq_Microbes.

789

790 **Author contributions**

791 F.G. established the nanopore workflow and performed all the bioinformatic analysis. F.G., R.K.,
792 M.J., R.R. and A.B. performed RNA extractions. M.F. helped to optimize the RNA treatment protocol.
793 F.G. carried out library preparations and performed sequencing. F.G., R.K., M.J. carried out *H.*
794 *volcanii* wildtype/Δ*ksgA* library preparations and sequencing. R.K. and S.F.-C. generated the KsgA
795 deletion strain. R.K. performed primer extension analysis. F.G., S.F.-C. and D.G. designed the study,
796 analysed and interpreted the data, and wrote the manuscript with the input of all authors. J.S., W.H.,
797 S.F.-C. and D.G. supervised the experiments. S.F.-C. and D.G. initiated and supervised the project.

798

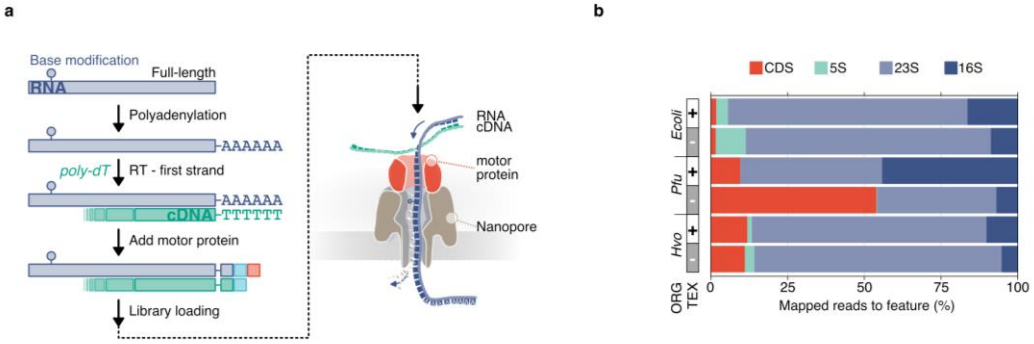
799 ## Acknowledgements

800 We gratefully acknowledge financial support by the Deutsche Forschungsgemeinschaft within the
801 collaborative research center framework (CRC/SFB960) “RNP biogenesis: assembly of ribosomes
802 and non-ribosomal RNPs and control of their function” [SFB960-TP7 to D.G.] [SFB960-TP-B13 to
803 S.F.-C.]. The work was also supported by the DFG through grant So264/21 to J.S.

804

805

806 **Figure and figure legends**



807

808 **Figure 1** | **Nanopore-based native RNA sequencing of prokaryotes. a**, Key steps of library preparation: (1) native RNA
809 is polyadenylated, which allows library preparation using the direct RNA kit from Oxford Nanopore and sequencing
810 on a MinION device. (2) 3´ ligation is performed to add the an adapter carrying the motor-protein (red square),
811 which unzips the RNA-cDNA hybrid and pulls the RNA through the Nanopore (detailed description see
812 Supplementary Fig. 1a). **b**, Data sets for three prokaryotic model organisms (Ecoli: *Escherichia coli*, Pfu: *Pyrococcus*
813 *furiosus*, Hvo: *Haloferax volcanii*) with (+) and without (-) Terminator 5´-Phosphate-Dependent Exonuclease (TEX)
814 treatment were collected and mapped to their respective reference genome. Transcript abundances of genomic
815 features (protein coding genes (CDS): red, 5S rRNA: green, 16S rRNA: purple, 23S rRNA: light-purple) were estimated
816 using featurecounts[42].

817

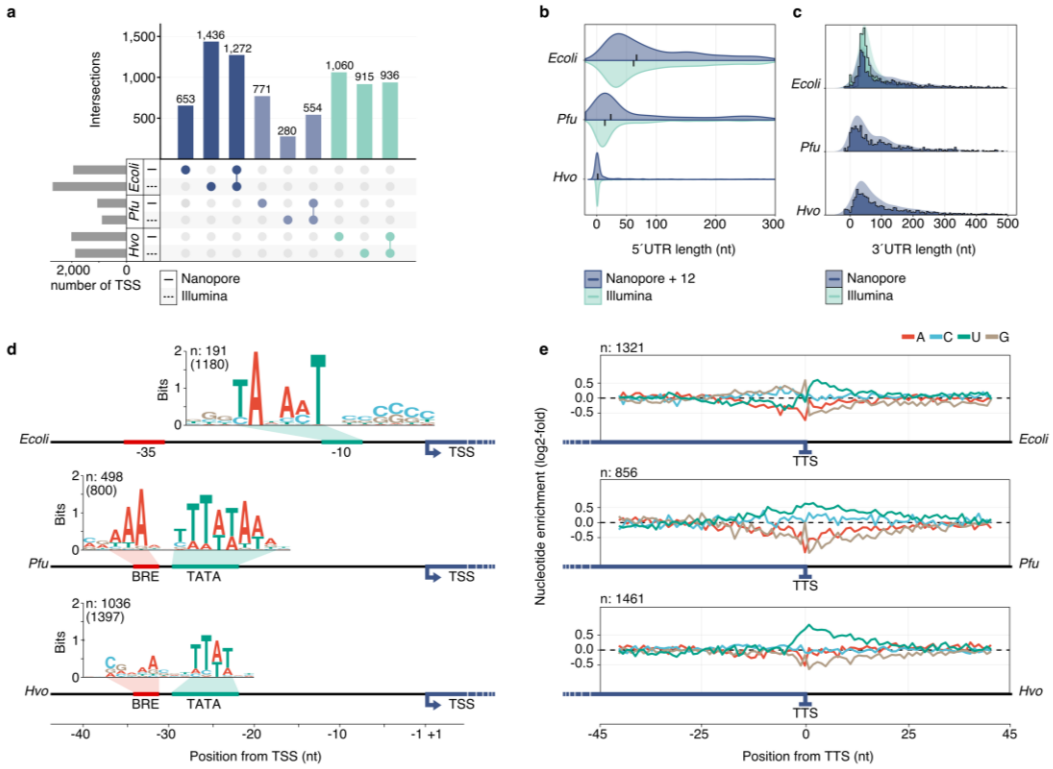


Figure 2: Detection of transcript boundaries. a, Primary transcription start sites (TSS) were predicted based on Nanopore reads and compared to Illumina d(ifferential) RNA-Seq data from published data sets for *E. coli* [30], *P. furiosus* [31] and *H. volcanii*[27]. The total number of all genes with a detected TSS is shown as grey barplots and results from the sum of Nanopore-only predicted TSS and the intersection to the Illumina data. **b**, Position of TSS is corrected for 12 nucleotides to calculate the length of 5´ untranslated regions (UTR) in the Nanopore data sets (purple). 5´ UTRs are compared to d(ifferential) RNA-Seq Illumina data sets (light-green). Median values are indicated by a black bar inside the distribution (compare Supplementary Fig. 6). **c**, Length of 3´ UTRs is based on the prediction of transcription termination sites (TTS) and the comparison to annotated gene ends. Distribution of lengths is shown for Nanopore data sets (purple) and compared to a Term-Seq *E. coli* Illumina data set (light-green)[108]. **d**, MEME analysis [45] of extracted sequences upstream of Nanopore-predicted TSS reveals bacterial (position -10) and archaeal-specific promoter elements (BRE: B-recognition element, recognized by transcription factor B, TATA: TATA-box bound by the TATA-binding protein TBP), therefore validating the positions of predicted TSS. **e**, Nucleotide enrichment meta analysis was carried out by comparing the genomic sequences surrounding the TTS (-45 to +45) to randomly selected intergenic positions of the respective organism (n: 10000) (Terminator motifs in Supplementary Fig. 7).

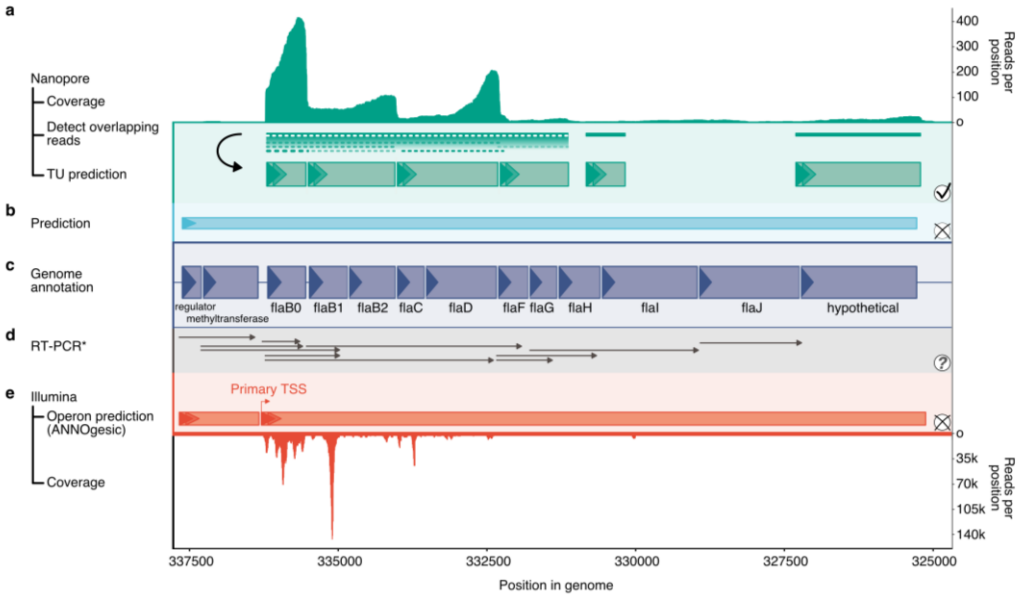


Figure 3 | Transcription unit (TU) annotation of the archaellum-operon in *P. furiosus*. **a**, Coverage of Nanopore reads is shown in the top panel. TU prediction is performed by detection and linkage of overlapping reads and splitting them according to a 3′ drop in coverage (see Supplementary Fig. 8). Predicted TUs are drawn with green boxes according to scale. **b**, Comparison to bioinformatical prediction using the DOOR2 database[50]. **c**, Genome annotation with abbreviated gene names, boxed drawn to scale and strand indicated by triangles[31]. **d**, Comparison to results from published RT-PCR experiments[62]. All transcripts detected are drawn by arrows. **e**, Operon prediction based on mixed Illumina-Seq (coverage in lower panel) and predicted by ANNOgesic[31,109]. The primary transcription start site (TSS) of the large transcriptional unit is highlighted.

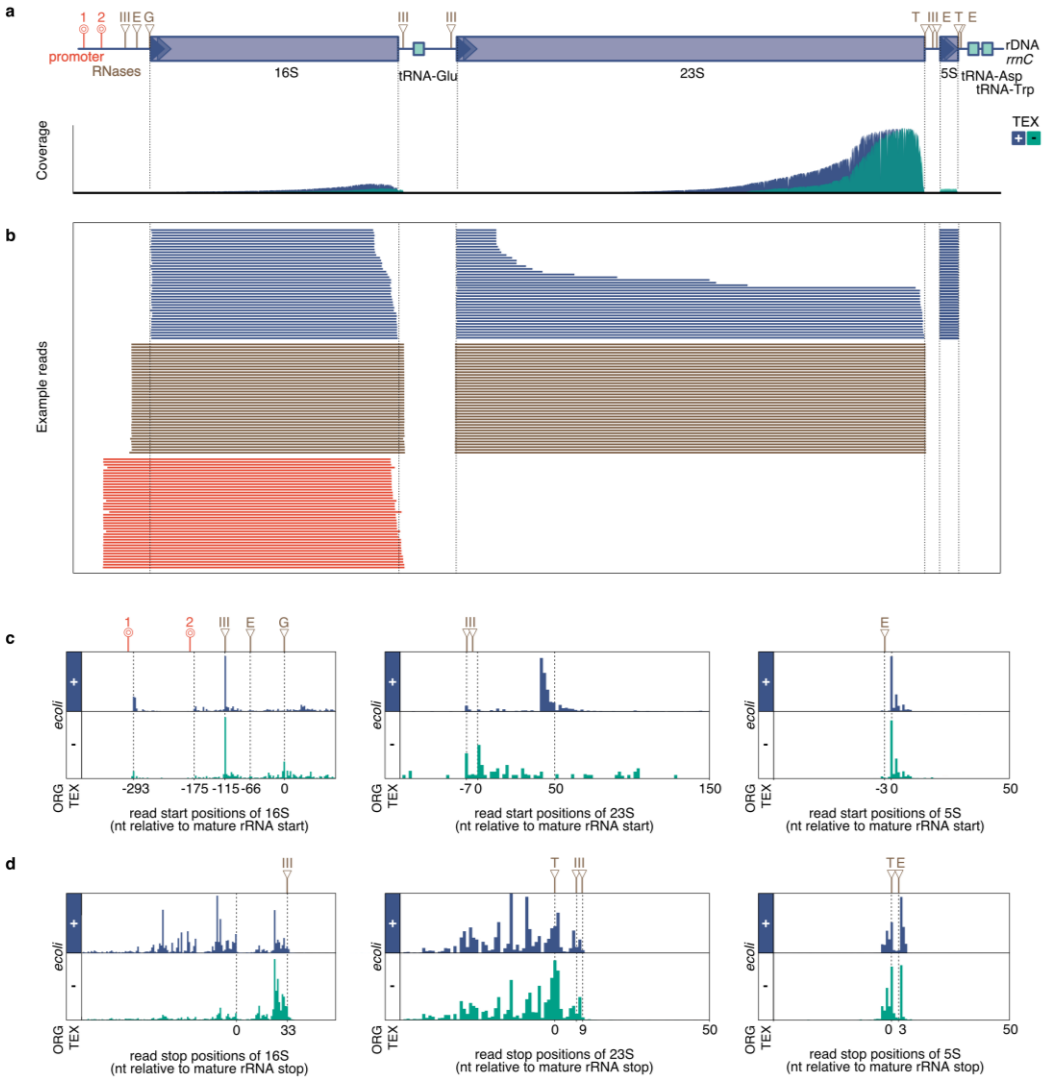


Figure 4 | Detection and confirmation of rRNA processing sites in *E. coli*. **a**, Transcription of the rDNA locus (rrnC) is starting from two promoters (transcription start sites at -293 and -175)[49]. Precursor RNAs are cleaved by RNases (brown) at depicted positions [64,65,73,80]. Nanopore coverage is shown for TEX (+, blue) and NOTEX (-, blue) samples. **b**, Single-read analysis of precursor and mature rRNAs. Reads were filtered according to read start and stop positions and are color-coded by mature rRNAs (blue), RNaseIII-processed precursors (brown) and precursor rRNAs starting from the TSS (red). For each category the 40 longest reads were selected and are shown here as single lines. **c**, Histograms of read start positions and **d**, read end positions of 16S, 23S and 5S rRNA relative to annotated boundaries of mature rRNAs of TEX (+) and NOTEX (-) samples. Cleavage sites, transcription start sites and transcription boundaries are indicated by dashed lines.

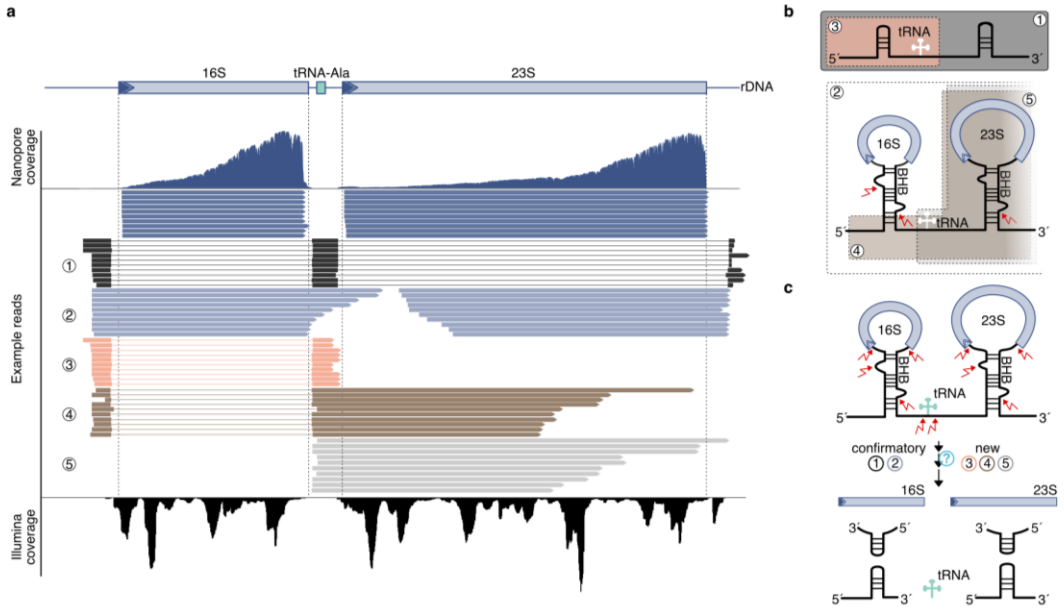


Figure 5 | Ribosomal RNA processing variation in *P. furiosus*. **a**, Nanopore coverage of the rDNA locus in *P. furiosus* containing the tRNA-Ala. Single reads were filtered according to their 5′ and 3′ lengths (couting from 16S and 23S rRNA annotation) and are shown in the example read track from a snapshot taken with the integrated genome browser (IGV)[110]. Besides mature 16S rRNA and 23S rRNA, the remaining reads were classified as (1) a processing variant that entails leading sequence-tRNA-trailing sequence, (2) probably fragmented full-length precursor rRNAs, (3) leading sequence-tRNA, (4) leading sequence-tRNA-23S rRNA and (5) tRNA-23S rRNA (quantification shown in Supplementary Fig. 13c). Illumina coverage from mixed RNA-Seq is shown in the bottom panel[31]. **b**, Read classes are shown as processing sketches and color-coded as on the left. **c**, Maturation steps of the rRNA processing pathway as proposed for *Archaeoglobus fulgidus*[49] with cleavage sites indicated by red arrows (full proposed pathway in Supplementary Fig. 13a). Processing variants found in *P. furiosus* (4, 5) potentially represent intermediate steps that refine the current rRNA maturation model in archaea (Supplementary Fig. 13b).

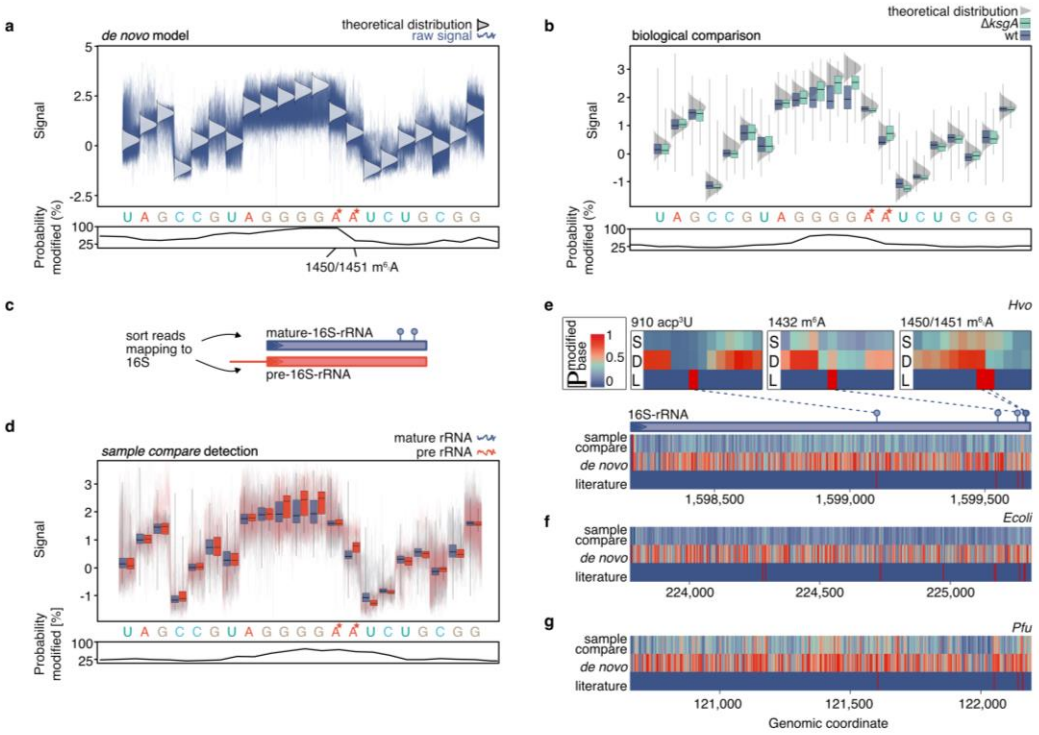


Figure 6 | Detection of modified bases in 16S rRNA based on raw signal comparisons. a, Raw signal of reads (blue squiggles) mapping to 16S rRNA in *H. volcanii* are compared to the theoretical distribution of native non-modified RNA (grey distribution) using the de novo detection model in tombo in the upper track[47]. The $m^6{}_2A$ modification at position 1450/1451 (from 16S start) is indicated by an asterisks in the sequence track. The probability of each base to be modified (in %) is calculated and shown in the lower panel for the selected sequence. **b**, Position-specific boxplot comparison of signals from sequences surrounding the $m^6{}_2A$ modification in *H. volcanii* wildtype (blue) and the ΔKsgA mutant. The theoretical distribution of read signal is indicated by a grey distribution curve for every base. The probability is computed based on the comparison of the two samples. **c**, Strategy for generating a comparison data set: Reads mapping to the 16S rRNA are categorized according to their read start position from the mature annotated rRNA as mature or pre-16S-rRNAs. **d**, Sample comparison model as shown in panel b based on mature (blue) and pre-rRNAs (red). **e**, Heatmap for the probability of a base to be modified (blue, light-green, red) shown for the mature 16S rRNA. Probabilities were calculated bases on a de novo (D) model, a sample comparison model (S) using the strategy described in panel c and plotted for already known modification sites (L)[49]. Three of those modifcations in *H. volcanii* are shown enlarged in the upper panel. Data are summarized for *H. volcanii* , **f**, *E. coli* [48] and **g**, *P. furiosus*.

References

1. Levy, S. E. & Myers, R. M. Advancements in Next-Generation Sequencing. *Annual review of genomics and human genetics* **17,** 95–115; 10.1146/annurev-genom-083115-022413 (2016).
2. Escobar-Zepeda, A., Vera-Ponce de León, A. & Sanchez-Flores, A. The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics. *Frontiers in Genetics* **6,** 348; 10.3389/fgene.2015.00348 (2015).
3. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics* **10,** 57–63; 10.1038/nrg2484 (2009).
4. Hör, J., Gorski, S. A. & Vogel, J. Bacterial RNA Biology on a Genome Scale. *Molecular cell* **70,** 785–799; 10.1016/j.molcel.2017.12.023 (2018).
5. Croucher, N. J. & Thomson, N. R. Studying bacterial transcriptomes using RNA-seq. *Current opinion in microbiology* **13,** 619–624; 10.1016/j.mib.2010.09.009 (2010).
6. Nowrousian, M. Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. *Eukaryotic cell* **9,** 1300–1310; 10.1128/EC.00123-10 (2010).
7. Saliba, A.-E., C Santos, S. & Vogel, J. New RNA-seq approaches for the study of bacterial pathogens. *Current opinion in microbiology* **35,** 78–87; 10.1016/j.mib.2017.01.001 (2017).
8. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nature Reviews Genetics*; 10.1038/s41576-019-0150-2 (2019).
9. Tilgner, H. *et al.* Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nature Biotechnology* **33,** 736–742; 10.1038/nbt.3242 (2015).
10. Byrne, A., Cole, C., Volden, R. & Vollmers, C. Realizing the potential of full-length transcriptome sequencing. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **374,** 20190097; 10.1098/rstb.2019.0097 (2019).
11. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science (New York, N.Y.)* **323,** 133–138; 10.1126/science.1162986 (2009).
12. Mikheyev, A. S. & Tin, M. M. Y. A first look at the Oxford Nanopore MinION sequencer. *Molecular ecology resources* **14,** 1097–1102; 10.1111/1755-0998.12324 (2014).
13. Soneson, C. *et al.* A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. *Nature Communications* **10,** 1–14; 10.1038/s41467-019-11272-z (2019).
14. Liu, H. *et al.* Accurate detection of m6A RNA modifications in native RNA sequences. *Nature Communications,* 1–9; 10.1101/525741 (2019).
15. Smith, A. M., Jain, M., Mulroney, L., Garalde, D. R. & Akeson, M. Reading canonical and modified nucleobases in 16S ribosomal RNA using nanopore native RNA sequencing. *PLOS ONE* **14,** 1–15; 10.1371/journal.pone.0216709 (2019).
16. Workman, R. E. *et al.* Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nature Methods* **16,** 1297–1305; 10.1038/s41592-019-0617-2 (2019).
17. Viehweger A, Krautwurst S, Lamkiewicz K, Madhugiri R, Ziebuhr J, Hölzer M, Marz M. Nanopore direct RNA sequencing reveals modification in full-length coronavirus genomes. *bioRxiv Genomics,* 1–15; 10.1101/483693 (2018).
18. Keller, M. W. *et al.* Direct RNA Sequencing of the Coding Complete Influenza A Virus Genome. *Scientific Reports* **8,** 1–8; 10.1038/s41598-018-32615-8 (2018).
19. Boldogkői, Z., Moldován, N., Balázs, Z., Snyder, M. & Tombácz, D. Long-Read Sequencing – A Powerful Tool in Viral Transcriptome Research. *Trends in Microbiology* **27,** 578–592; 10.1016/j.tim.2019.01.010 (2019).
20. Viehweger, A. *et al.* Direct RNA nanopore sequencing of full-length coron-avirus genomes provides novel insights into structural variants and enables modification analysis. *bioRxiv,* 483693; 10.1101/483693 (2019).
21. Tombácz, D. *et al.* Multiple Long-Read Sequencing Survey of Herpes Simplex Virus Dynamic Transcriptome. *Frontiers in Genetics* **10,** 834; 10.3389/fgene.2019.00834 (2019).

22. Zhao, L. *et al.* Analysis of Transcriptome and Epitranscriptome in Plants Using PacBio Iso-Seq and Nanopore-Based Direct RNA Sequencing. *Frontiers in Genetics* **10,** 253; 10.3389/fgene.2019.00253 (2019).
23. Bayega, A. *et al. Transcriptome landscape of the developing olive fruit fly embryo delineated by Oxford Nanopore long-read RNA-Seq* (2018).
24. Byrne, A. *et al.* Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nature Communications* **8**; 10.1038/ncomms16027 (2017).
25. Rahimi, K., Venø, M. T., Dupont, D. M. & Kjems, J. *Nanopore sequencing of full-length circRNAs in human and mouse brains reveals circRNA-specific exon usage and intron retention* (2019).
26. Dar, D. & Sorek, R. High-resolution RNA 3-ends mapping of bacterial Rho-dependent transcripts. *Nucleic Acids Research* **46,** 6797–6805; 10.1093/nar/gky274 (2018).
27. 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,** 1–19; 10.1186/s12864-016-2920-y (2016).
28. Laass, S. *et al.* Characterization of the transcriptome of Haloferax volcanii , grown under four different conditions , with mixed RNA-Seq. *PLOS ONE* **14,** 1–24; 10.1371/journal.pone.0215986 (2019).
29. Mao, X. *et al.* Revisiting operons: An analysis of the landscape of transcriptional units in E. coli. *BMC Bioinformatics*; 10.1186/s12859-015-0805-8 (2015).
30. Thomason, M. K. *et al.* Global transcriptional start site mapping using differential RNA sequencing reveals novel antisense RNAs in Escherichia coli. *Journal of Bacteriology* **197,** 18–28; 10.1128/JB.02096-14 (2015).
31. Grünberger, F. *et al.* Next Generation DNA-Seq and Differential RNA-Seq Allow Re-annotation of the Pyrococcus furiosus DSM 3638 Genome and Provide Insights Into Archaeal Antisense Transcription. *Frontiers in Microbiology*; 10.3389/fmicb.2019.01603 (2019).
32. Stetter, K. O., König, H. & Stackebrandt, E. Pyrodictium gen. nov., a New Genus of Submarine Disc-Shaped Sulphur Reducing Archaebacteria Growing Optimally at 105°C. *Systematic and Applied Microbiology*; 10.1016/S0723-2020(83)80011-3 (1983).
33. Allers, T. & Mevarech, M. Archaeal genetics - the third way. *Nature Reviews Genetics* **6,** 58–73; 10.1038/nrg1504 (2005).
34. Knüppel, R. *et al.* Insights into the evolutionary conserved regulation of Rio ATPase activity. *Nucleic Acids Research* **46,** 1441–1456; 10.1093/nar/gkx1236 (2018).
35. Chomczynski, P. & Sacchi, N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Analytical biochemistry* **162,** 156–159; 10.1006/abio.1987.9999 (1987).
36. Yan, B., Boitano, M., Clark, T. A. & Ettwiller, L. SMRT-Cappable-seq reveals complex operon variants in bacteria. *Nature Communications*; 10.1038/s41467-018-05997-6 (2018).
37. Weirather, J. L. *et al.* Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* **6,** 100; 10.12688/f1000research.10571.2 (2017).
38. Riley, M. *et al.* Escherichia coli K-12: a cooperatively developed annotation snapshot--2005. *Nucleic Acids Research* **34,** 1–9; 10.1093/nar/gkj405 (2006).
39. Hartman, A. L. *et al.* The complete genome sequence of Haloferax volcanii DS2, a model archaeon. *PLOS ONE* **5,** e9605; 10.1371/journal.pone.0009605 (2010).
40. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*; 10.1093/bioinformatics/bty191 (2018).
41. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079; 10.1093/bioinformatics/btp352 (2009).
42. Liao, Y., Smyth, G. K. & Shi, W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research*; 10.1093/nar/gkz114 (2019).
43. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15,** 550; 10.1186/s13059-014-0550-8 (2014).
44. Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods* **12,** 733–735; 10.1038/nmeth.3444 (2015).
45. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research* **37,** W202-8; 10.1093/nar/gkp335 (2009).

46. Dar, D., Prasse, D., Schmitz, R. A. & Sorek, R. Widespread formation of alternative 3' UTR isoforms via transcription termination in archaea. *Nature Microbiology* **1,** 16143; 10.1038/nmicrobiol.2016.143 (2016).
47. Stoiber, M. *et al. De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing* (2016).
48. Boccaletto, P. *et al.* MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Research* **46,** D303-D307; 10.1093/nar/gkx1030 (2018).
49. Grosjean, H., Gaspin, C., Marck, C., Decatur, W. A. & Crécy-Lagard, V. de. RNomics and Modomics in the halophilic archaea Haloferax volcanii: Identification of RNA modification genes. *BMC Genomics* **9,** 1–26; 10.1186/1471-2164-9-470 (2008).
50. Mao, X. *et al.* DOOR 2.0: Presenting operons and their functions through dynamic and integrated views. *Nucleic Acids Research* **42**; 10.1093/nar/gkt1048 (2014).
51. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120; 10.1093/bioinformatics/btu170 (2014).
52. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359; 10.1038/nmeth.1923 (2012).
53. Bohlin, J., Eldholm, V., Pettersson, J. H. O., Brynildsrud, O. & Snipen, L. The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes. *BMC Genomics* **18,** 151; 10.1186/s12864-017-3543-7 (2017).
54. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology* **20,** 1–10; 10.1186/s13059-019-1727-y (2019).
55. Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nature Methods* **15,** 201–206; 10.1038/nmeth.4577 (2018).
56. Workman, R. E. *et al.* Nanopore native RNA sequencing of a human poly(A) transcriptome. *bioRxiv*; 10.1101/459529 (2018).
57. Parker, M. T. *et al.* Nanopore direct RNA sequencing maps an Arabidopsis N6 methyladenosine epitranscriptome. *bioRxiv*; 10.1101/706002 (2019).
58. Ju, X., Li, D. & Liu, S. Full-length RNA profiling reveals pervasive bidirectional transcription terminators in bacteria. *Nature Microbiology*; 10.1038/s41564-019-0500-z (2019).
59. Santangelo, T. J., Cubonová, L.'u., Skinner, K. M. & Reeve, J. N. Archaeal intrinsic transcription termination in vivo. *Journal of Bacteriology* **191,** 7102–7108; 10.1128/JB.00982-09 (2009).
60. Smith, A. M. *et al.* Reading canonical and modified nucleotides in 16S ribosomal RNA using nanopore direct RNA sequencing. *bioRxiv*; 10.1101/132274 (2017).
61. Wongsurawat, T. *et al.* Rapid sequencing of multiple RNA viruses in their native form. *Frontiers in Microbiology* **10,** 1–8; 10.3389/fmicb.2019.00260 (2019).
62. Näther-Schindler, D. J., Schopf, S., Bellack, A., Rachel, R. & Wirth, R. Pyrococcus furiosus flagella: biochemical and transcriptional analyses identify the newly detected flaB0 gene to encode the major flagellin. *Frontiers in Microbiology* **5,** 695; 10.3389/fmicb.2014.00695 (2014).
63. Lioliou, E. *et al.* Global regulatory functions of the Staphylococcus aureus endoribonuclease III in gene expression. *PLoS Genetics* **8,** e1002782; 10.1371/journal.pgen.1002782 (2012).
64. Smith, B. A., Gupta, N., Denny, K. & Culver, G. M. Characterization of 16S rRNA Processing with Pre-30S Subunit Assembly Intermediates from E. coli. *Journal of Molecular Biology*; 10.1016/j.jmb.2018.04.009 (2018).
65. Shajani, Z., Sykes, M. T. & Williamson, J. R. Assembly of bacterial ribosomes. *Annual review of biochemistry* **80,** 501–526; 10.1146/annurev-biochem-062608-160432 (2011).
66. Bechhofer, D. H. & Deutscher, M. P. Bacterial ribonucleases and their roles in RNA metabolism. *Critical reviews in biochemistry and molecular biology* **54,** 242–300; 10.1080/10409238.2019.1651816 (2019).
67. Klappenbach, J. A., Saxman, P. R., Cole, J. R. & Schmidt, T. M. rrndb: the Ribosomal RNA Operon Copy Number Database. *Nucleic Acids Research* **29,** 181–184; 10.1093/nar/29.1.181 (2001).
68. Maeda, M., Shimada, T., Ishihama, A. & Semsey, S. Strength and Regulation of Seven rRNA Promoters in Escherichia coli. *PLOS ONE* **10,** e0144697; 10.1371/journal.pone.0144697 (2015).
69. Hofmann, S. & Miller, O. L. Visualization of ribosomal ribonucleic acid synthesis in a ribonuclease III-Deficient strain of Escherichia coli. *Journal of Bacteriology* **132,** 718–722 (1977).
70. Jacob, A. I., Köhrer, C., Davies, B. W., RajBhandary, U. L. & Walker, G. C. Conserved bacterial RNase YbeY plays key roles in 70S ribosome quality control and 16S rRNA maturation. *Molecular cell* **49,** 427–438; 10.1016/j.molcel.2012.11.025 (2013).

71. Sulthana, S. & Deutscher, M. P. Multiple exoribonucleases catalyze maturation of the 3' terminus of 16S ribosomal RNA (rRNA). *The Journal of biological chemistry* **288,** 12574–12579; 10.1074/jbc.C113.459172 (2013).
72. Yip, W. S. V., Vincent, N. G. & Baserga, S. J. Ribonucleoproteins in archaeal pre-rRNA processing and modification. *Archaea (Vancouver, B.C.)* **2013,** 614735; 10.1155/2013/614735 (2013).
73. Ferreira-Cerca, S. in *RNA Metabolism and Gene Expression in Archaea,* edited by B. Clouet-d'Orval (Springer International Publishing, Cham, 2017), pp. 129–158.
74. Henras, A. K., Plisson-Chastang, C., O'Donohue, M.-F., Chakraborty, A. & Gleizes, P.-E. An overview of pre-ribosomal RNA processing in eukaryotes. *Wiley interdisciplinary reviews. RNA* **6,** 225–242; 10.1002/wrna.1269 (2015).
75. Clouet-D'Orval, B. *et al.* Insights into RNA-processing pathways and associated RNA-degrading enzymes in Archaea. *FEMS Microbiology Reviews* **42,** 579–613; 10.1093/femsre/fuy016 (2018).
76. Tang, T. H. *et al.* RNomics in Archaea reveals a further link between splicing of archaeal introns and rRNA processing. *Nucleic Acids Research* **30,** 921–930; 10.1093/nar/30.4.921 (2002).
77. Russell, A. G., Ebhardt, H. & Dennis, P. P. Substrate requirements for a novel archaeal endonuclease that cleaves within the 5' external transcribed spacer of Sulfolobus acidocaldarius precursor rRNA. *Genetics* **152,** 1373–1385 (1999).
78. Jüttner, M. *et al.* A versatile cis-acting element reporter system to study the function, maturation and stability of ribosomal RNA mutants in archaea. *Nucleic Acids Research*; 10.1093/nar/gkz1156 (2019).
79. Danan, M., Schwartz, S., Edelheit, S. & Sorek, R. Transcriptome-wide discovery of circular RNAs in Archaea. *Nucleic Acids Research* **40,** 3131–3142; 10.1093/nar/gkr1009 (2012).
80. Ferreira-Cerca, S. *RNA Metabolism and Gene Expression in Archaea - Chapter6 - Life and Death of Ribosomes in Archaea* (2017).
81. Wongsurawat, T. *et al.* Decoding the Epitranscriptional Landscape from Native RNA Sequences. *bioRxiv* **17,** 487819; 10.1101/487819 (2018).
82. O'Farrell, H. C., Pulicherla, N., Desai, P. M. & Rife, J. P. Recognition of a complex substrate by the KsgA/Dim1 family of enzymes has been conserved throughout evolution. *RNA (New York, N.Y.)* **12,** 725–733; 10.1261/rna.2310406 (2006).
83. Rang, F. J., Kloosterman, W. P. & Ridder, J. de. From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy. *Genome Biology* **19,** 1–11; 10.1186/s13059-018-1462-9 (2018).
84. Strunk, B. S. *et al.* Ribosome assembly factors prevent premature translation initiation by 40S assembly intermediates. *Science (New York, N.Y.)* **333,** 1449–1453; 10.1126/science.1208245 (2011).
85. Xu, Z., O'Farrell, H. C., Rife, J. P. & Culver, G. M. A conserved rRNA methyltransferase regulates ribosome biogenesis. *Nature structural & molecular biology* **15,** 534–536; 10.1038/nsmb.1408 (2008).
86. Lafontaine, D. L., Preiss, T. & Tollervey, D. Yeast 18S rRNA dimethylase Dim1p: a quality control mechanism in ribosome synthesis? *Molecular and cellular biology* **18,** 2360–2370; 10.1128/mcb.18.4.2360 (1998).
87. Dennis, P. P., Tripp, V., Lui, L., Lowe, T. & Randau, L. C/D box sRNA-guided 2'-O-methylation patterns of archaeal rRNA molecules. *BMC Genomics*; 10.1186/s12864-015-1839-z (2015).
88. Gomes-Filho, J. V. & Randau, L. RNA stabilization in hyperthermophilic archaea. *Annals of the New York Academy of Sciences* **1447,** 88–96; 10.1111/nyas.14060 (2019).
89. Ebright, R. H., Werner, F. & Zhang, X. RNA Polymerase Reaches 60: Transcription Initiation, Elongation, Termination, and Regulation in Prokaryotes. *Journal of Molecular Biology* **431,** 3945–3946; 10.1016/j.jmb.2019.07.026 (2019).
90. Babski, J. *et al.* Small regulatory RNAs in Archaea. *RNA Biology* **11,** 484–493; 10.4161/rna.28452 (2014).
91. Oliva, G., Sahr, T. & Buchrieser, C. Small RNAs, 5' UTR elements and RNA-binding proteins in intracellular bacteria: impact on metabolism and virulence. *FEMS Microbiology Reviews* **39,** 331–349; 10.1093/femsre/fuv022 (2015).
92. Ren, G.-X., Guo, X.-P. & Sun, Y.-C. Regulatory 3' Untranslated Regions of Bacterial mRNAs. *Frontiers in Microbiology* **8,** 1276; 10.3389/fmicb.2017.01276 (2017).
93. Hirtreiter, A. *et al.* Spt4/5 stimulates transcription elongation through the RNA polymerase clamp coiled-coil motif. *Nucleic Acids Research* **38,** 4040–4051; 10.1093/nar/gkq135 (2010).

94. 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 Research* **38,** 585–596; 10.1093/nar/gkp928 (2010).
95. Berkemer, S. J. *et al. Identification of RNA 3' ends and termination sites in Haloferax volcanii* (2019).
96. Yue, L. *et al. aCPSF1 controlled archaeal transcription termination: a prototypical eukaryotic model* (2019).
97. Daum, B. *et al.* Structure and in situ organisation of the Pyrococcus furiosus archaellum machinery. *eLife* **6**; 10.7554/eLife.27470 (2017).
98. Deutscher, M. P. Twenty years of bacterial RNases and RNA processing: how we've matured. *RNA (New York, N.Y.)* **21,** 597–600; 10.1261/rna.049692.115 (2015).
99. Venema, J. & Tollervey, D. Processing of pre-ribosomal RNA in Saccharomyces cerevisiae. *Yeast (Chichester, England)* **11,** 1629–1650; 10.1002/yea.320111607 (1995).
100. LITTLEFIELD, J. W. & DUNN, D. B. Natural occurrence of thymine and three methylated adenine bases in several ribonucleic acids. *Nature* **181,** 254–255; 10.1038/181254a0 (1958).
101. SMITH, J. D. & DUNN, D. B. The occurrence of methylated guanines in ribonucleic acids from several sources. *The Biochemical journal* **72,** 294–301; 10.1042/bj0720294 (1959).
102. Li, S. & Mason, C. E. The pivotal regulatory landscape of RNA modifications. *Annual review of genomics and human genetics* **15,** 127–150; 10.1146/annurev-genom-090413-025405 (2014).
103. Schwartz, S. & Motorin, Y. Next-generation sequencing technologies for detection of modified nucleotides in RNAs. *RNA Biology* **14,** 1124–1137; 10.1080/15476286.2016.1251543 (2017).
104. Sas-Chen, A. & Schwartz, S. Misincorporation signatures for detecting modifications in mRNA: Not as simple as it sounds. *Methods (San Diego, Calif.)* **156,** 53–59; 10.1016/j.ymeth.2018.10.011 (2019).
105. Lorenz, D. A., Sathe, S., Einstein, J. M. & Yeo, G. W. Direct RNA sequencing enables m6A detection in endogenous transcript isoforms at base specific resolution. *RNA (New York, N.Y.)*; 10.1261/rna.072785.119 (2019).
106. Leger, A. *et al. RNA modifications detection by comparative Nanopore direct RNA sequencing* (2019).
107. Sloan, K. E. *et al.* Tuning the ribosome: The influence of rRNA modification on eukaryotic ribosome biogenesis and function. *RNA Biology* **14,** 1138–1152; 10.1080/15476286.2016.1259781 (2017).
108. Dar, D. & Sorek, R. Extensive reshaping of bacterial operons by programmed mRNA decay. *PLoS Genetics* **14,** 1–19; 10.1371/journal.pgen.1007354 (2018).
109. Yu, S.-H., Vogel, J. & Förstner, K. U. *ANNOgesic: A Swiss army knife for the RNA-Seq based annotation of bacterial/archaeal genomes* (2017).
110. Freese, N. H., Norris, D. C. & Loraine, A. E. Integrated genome browser: visual analytics platform for genomics. *Bioinformatics* **32,** 2089–2095; 10.1093/bioinformatics/btw069 (2016).