GENOME AND TRANSCRIPTOME ARCHITECTURE IN *PYROCOCCUS FURIOSUS*

DISSERTATION

ZUR ERLANGUNG DES DOKTORGRADES
DER NATURWISSENSCHAFTEN (DR. RER. NAT.)
DER FAKULTÄT FÜR BIOLOGIE UND VORKLINISCHE MEDIZIN
DER UNIVERSITÄT REGENSBURG

vorgelegt von

Felix Grünberger

aus Hutthurm

Im Jahr 2020

Das Promotionsgesuch wurde eingereicht am: **18.09.2020**

Die Arbeit wurde angeleitet von: **PD Dr. Winfried Hausner**

Unterschrift: **…….**

Felix Grünberger

References of Manuscripts

This thesis is composed of the following manuscripts:

1. **Grünberger, F.**, Reichelt, R., Bunk, B., Spröer, C., Overmann, J., Rachel, R., et al. (2019). Next Generation DNA-Seq and Differential RNA-Seq Allow Re-annotation of the *Pyrococcus furiosus* DSM 3638 Genome and Provide Insights Into Archaeal Antisense Transcription. *Front. Microbiol.* 10. doi:10.3389/fmicb.2019.01603

2. **Grünberger, F.**, Knüppel, R., Jüttner, M., Fenk, M., Borst, A., Reichelt, R., et al. (2020a). Exploring prokaryotic transcription, operon structures, rRNA maturation and modifications using Nanopore-based native RNA sequencing. *bioRxiv*, 2019.12.18.880849. doi:10.1101/2019.12.18.880849.

3. **Grünberger, F.,** Reichelt, R., Waege, I., Ned, V., Bronner, K., Kaljanac, M., et al. (2020b). CopR, a global regulator of transcription to maintain copper homeostasis in *Pyrococcus furiosus*. *bioRxiv*, 2020.08.14.251413. doi:10.1101/2020.08.14.251413.

Personal Contributions

The personal contribution from Felix Grünberger (**FG**) in the three manuscripts has been adapted from the author contributions section in the manuscripts:

1. RoR prepared the RNA from Pyrococcus. BB, CS, and JO performed the PacBio and Illumina sequencing and **FG** the nanopore sequencing. The bioinformatical analysis was carried out by **FG** and RoR. **FG**, RoR, DG, and WH wrote the manuscript. WH, RR, and DG coordinated and supervised the work. All authors approved the final version of the manuscript (Grünberger et al., 2019).

2. **FG** established the nanopore workflow and performed all the bioinformatic analysis. **FG**, R.K., M.J., R.R. and A.B. performed RNA extractions. M.F. helped to optimize the RNA treatment protocol. **FG** carried out library preparations and performed sequencing. **FG**, R.K., M.J. carried out *H. volcanii* wildtype/Δ*ksgA* library preparations and sequencing. M.F. and R.R. performed transcription assays. R.K. and S.F.-C. generated the KsgA deletion strain. R.K. performed primer extension analysis. **FG**, S.F.-C. and D.G. designed the study, analysed and interpreted the data, and wrote the manuscript with the input of all authors. J.S., W.H., S.F.-C. and D.G. supervised the experiments. S.F.-C. and D.G. initiated and supervised the project (Grünberger et al., 2020a).

3. **FG** did the DGE and the complete bioinformatic analysis, RR constructed the *Pyrococcus* deletion strain, IW did the *in vitro* transcription and the footprinting experiments. VN and LK performed the gel shift assays, KB the ChIP-seq experiments and the qPCR assays. MK, NW, ZE, GM and CZ did the negative-stain TEM imaging. **FG**, DG and WH wrote the manuscript and DG and WH coordinated and supervised the work. All authors agreed to the final version of the manuscript (Grünberger et al., 2020b).

Die vorliegende Arbeit wurde im Zeitraum von Februar 2016 bis September 2020 am Lehrstuhl für Mikrobiologie des Institutes für Biochemie, Genetik und Mikrobiologie der Fakultät für Biologie und Vorklinische Medizin der Universität Regensburg unter Anleitung von PD Dr. Winfried Hausner angefertigt.

Table of Contents

List of Main Figures

List of Supplementary Figures

List of Tables

List of Supplementary Tables

Abstract

Archaea nowadays are acknowledged for representing the second domain of life and for playing significant roles in the Earth´s biogeochemical cycles. Before their initial discovery by Carl Woese and colleagues in the late 1970s, Archaea have not been recognised and erroneously confused with look-alike Bacteria under the microscope for decades. Since their classification as the third primary "kingdom" in 1990, not only their position in the universal tree of life has changed, defining the archaeal ancestry of Eukaryotes. Also, the knowledge about their ecology, diversity, evolution and molecular principles has been extended tremendously. Notably, it has been revealed that on the molecular level, Archaea share remarkably striking characteristics with both Eukarya and Bacteria, with transcription as one of the prime examples. Here, we have primarily been interested in the genome and transcriptome architecture, the regulatory roles of transcription factors and post-transcriptional mechanisms in the hyperthermophilic model archaeon *Pyrococcus furiosus*.

To obtain the most accurate and informative background for further studies, we re-sequenced the culture collection strain DSM 3638 employing state-of-the-art hybrid Illumina and PacBio DNA sequencing and extensively expanded the annotation on the transcript level by using a differential RNA sequencing approach. Digestion of all non 5´-triphosphorylated transcripts by a Terminator-exonuclease allowed us to specifically enrich primary transcripts. The redefinition of the transcriptional landscape of *P. furiosus*

included the genome-wide detection of transcription start sites, promoter architectures, sense- and antisense-RNAs. Interestingly, we discovered bidirectional transcription from symmetric promoters as an extensive source of antisense transcription, which is presumably a widespread feature of archaeal transcription. Additionally, we could prove that despite the relatively high abundance of insertion sequences in the 2 Mbp genome, the handling of a lab culture for two years did not lead to genomic rearrangements. Although we did not specifically challenge the genomic integrity, this still suggests that the genome is more stable than previously anticipated, which is an essential prerequisite for the comparability and feasibility of future genome-wide studies in *P. furiosus*. For rapid and cost-efficient re-sequencing of archaeal strains, we established 3rd generation long-read Nanopore sequencing technology in the lab, which allowed us to sequence the lab strain with high consensus accuracy.

Next, we established a protocol for direct RNA sequencing in prokaryotes using the Nanopore technology, which is currently the only option for single-molecule sequencing of transcripts in their native context. The plethora of transcriptional and post-transcriptional events and features are usually tackled by short-read sequencing approaches that specifically have to be tailored to the respective research question by making adaptions to the library preparation protocol or by chemical treatment. In contrast, we evaluated the potential of native RNA sequencing to address multiple transcriptomic features simultaneously in a bacterial (*Escherichia coli*) and archaeal (*Haloferax volcanii*, *P. furiosus*) model organisms. Performing meta-data and single-molecule analysis we could (re-)annotate large transcriptional units and map transcription boundaries. Besides, we showed that long reads are a valuable tool for heterogeneous 3´-end detection and that diverse termination mechanisms occur in Archaea. Next, we used the single-molecule potential of Nanopore reads for the identification of previously known and unknown intermediates in the poorly understood rRNA maturation pathway in Archaea. Moreover, we were able to detect RNA base modifications in the form of systematic basecalling errors and shifts in the ionic current, which allowed us to follow the relative timely order of KsgA-dependent di-methylation and N^4-cytidine acetylation in mature and precursor 16S rRNAs in archaeal species.

Third, using the new reference genome of *P. furiosus*, we performed an integrative RNA-seq and ChIP-seq based approach to decipher the function of the transcriptional regulator CopR during copper detoxification in *P. furiosus.* To get a global view on the transcriptomic response and find components of the CopR-regulon, we performed differential gene expression analysis and ChIP-seq analysis after copper shock. We discovered that CopR, which is essential in copper detoxification, binds to the upstream regions of highly copper-induced genes, that all share a common palindromic motif. Additionally, negative-stain transmission electron microscopy and image analysis by 2D class averaging revealed that CopR binds to DNA in an octameric conformation similar to

other factors of the Lrp family. Finally, we proposed a model for allosteric regulation of CopR upon copper-binding and revealed different layers of copper detoxification in *P. furiosus*.

The findings of the studies that make up this thesis contribute to a deeper understanding of basic and regulatory principles of transcription in Archaea and update the genomic and transcriptomic landscape of *P. furiosus*. Also, the application of Nanopore-based native RNA sequencing not only represents a significant extension of the transcriptomic toolbox in prokaryotes but also provided us with a wealth of information, especially regarding transcriptional and post-transcriptional events during rRNA maturation.

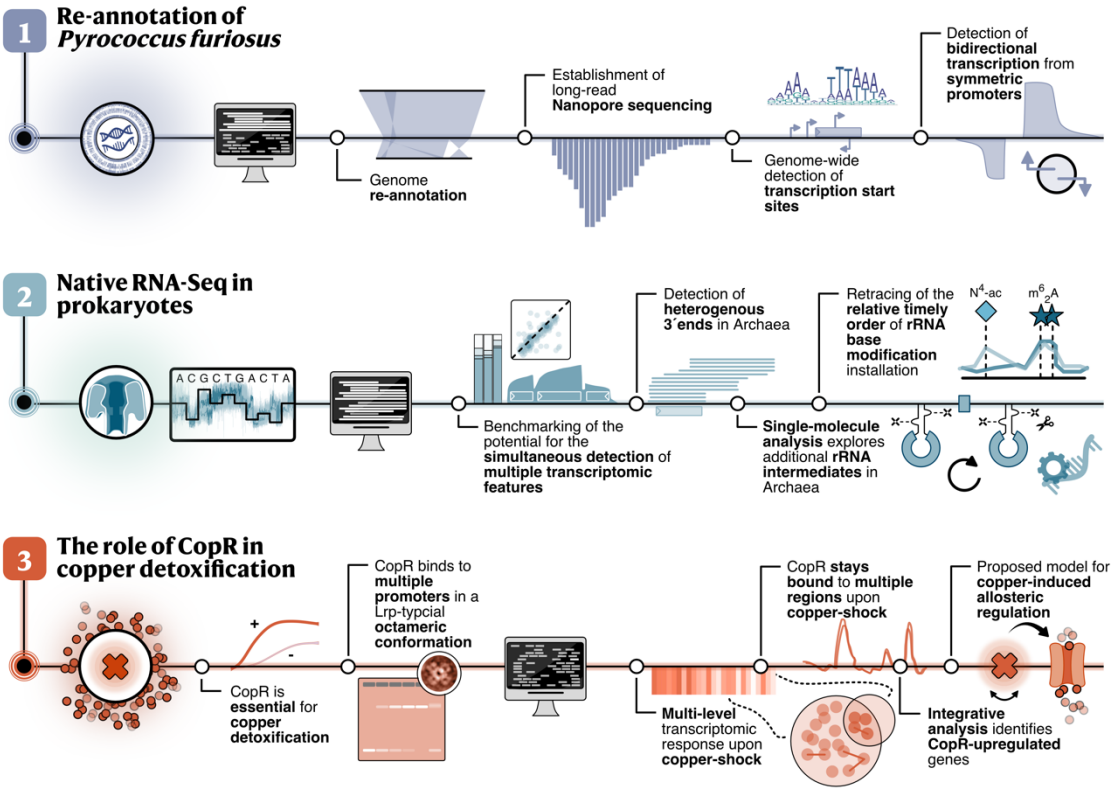


Figure 1 | Graphical abstract. This thesis comprises three publications, that deal with **(1)** the Re-annotation of *P. furiosus* (Grünberger et al., 2019), **(2)** the establishment, benchmarking and application of native RNA-seq in prokaryotes using Nanopore technology (Grünberger et al., 2020a) and **(3)** the essential role of the transcriptional regulator CopR during copper detoxification in *P. furiosus* (Grünberger et al., 2020b). The graphical summary visualises main results of the manuscripts.

CHAPTER I
General Introduction

Since the initial proposal of the archaeal domain more than 40 years ago, scientific discoveries in ecology, evolution and molecular insights into fundamental cellular processes have transformed the study of historically-speaking odd organisms into a vibrant research field. Recently, deep metagenomic sequencing of environmental samples combined with advanced phylogenetic modelling has reshaped the universal tree of life and provided clear evidence that Eukarya have emerged from within the Archaea (Zaremba-Niedzwiedzka et al., 2017; Williams et al., 2020). Throughout history, technological advancements, especially the exploration of metagenomic sequencing in a culture-independent way in the latest years, have been a driving force for the discovery of new archaeal species and phyla that ultimately led to the proposal of a two-domain tree. Interestingly, the close relationship between Archaea and Eukarya is also reflected on the molecular level and became evident very early by comparing the composition of the key machineries and the set of auxiliary proteins that facilitate transcription and translation (Zillig et al., 1979; Bell and Jackson, 1998; Hirata et al., 2008). However, that is just one side of the story, as all of the processes run in a bacterial-like gene-dense genomic environment (Koonin and Wolf, 2008; Kellner et al., 2018). Additionally, Archaea, that just like Bacteria are abundant in all kinds of ecological niches, can adapt to environmental changes on the transcriptional level with the help of dedicated transcription factors, that share structural and regulatory characteristics with their bacterial counterparts (Lemmens et al., 2019). Hence, on the molecular level, Archaea are chimaera representing both

eukaryal and bacterial features. Consequently, deciphering principles of archaeal transcription not only allows to address domain-specific questions but ultimately contributes to a better understanding of evolution.

To give a more in-depth background on the main topics that are covered in the articles, this introduction covers three parts: (1) Historical perspective of principles and advancements in three generations of sequencing technologies, (2) milestones in archaeal research, highlighting general genomic features and *P. furiosus* as a well-established hyperthermophilic model organism and (3) current understanding of basal and regulated transcription in Archaea. While the latter two chapters extend the transcription-themed questions that were addressed in the publications, the first chapter is more technical, providing a summary of how dramatic improvements over the years revolutionised the biological sciences. The tremendous increase in the generation of genomics data had and still has a strong impact, of course not exclusively but particularly on many aspects of archaeal research. In the years to come, it will be interesting to see whether the field will completely transition from 2nd generation short-read to 3rd generation long-read technologies and how this will affect genomics and transcriptomics.

1. Three generations of sequencing technologies

Since the discovery of DNA and its crucial role in determining genetic inheritance, it has always been an important goal to decipher the sequence of these life-forming building blocks. The process of determining the nucleic acid sequence is nowadays referred to as DNA sequencing. In 2017, DNA sequencing celebrated its 40th anniversary (reviewed in Shendure et al., 2017). Despite its relatively short developmental phase, this technology has already undergone some fundamental conceptual changes. Based on their shared key ideas and their timely appearance, the different approaches have been classified into three successive generations: (i) 1st generation Sanger sequencing, (ii) 2nd generation massively parallel sequencing and (iii) 3rd generation long-read sequencing (Goodwin et al., 2016; van Dijk et al., 2018).

While initial efforts have been very laborious and expensive, especially the Human Genome Project (HGP) from 1990 to 2001 pushed the development and continuous improvement of sequencing technologies (Lander et al., 2001). Ultimately, this led to a dramatic cost-reduction, a throughput-increase from only a few to billions of bases and democratisation of sequencing by bringing the possibilities of this technology to many labs around the scientific globe. Meanwhile, genomic information has shaped our understanding of many different aspects of biology from evolution to diseases. During the last decade, researches have started to expand the usability of DNA sequencing to the messenger molecule RNA and established different protocols for transcriptome-wide investigations (reviewed in Stark et al., 2019). Given the contributions in all fields of biology, it is suspected that it will only be a matter of time until sequencing will be acknowledged in the same way as the invention of the microscope back in the 16th century (Shendure et al., 2017).

1.1. 1st generation: Sanger & Maxam-Gilbert

1977 was the year of breakthroughs in DNA sequencing. In that year, two studies from Walter Gilbert´s and Frederic Sanger´s lab sharing remarkably similar ideas were independently published (Maxam and Gilbert, 1977; Sanger et al., 1977b). Short after, in 1980, both have been awarded a Nobel prize in Chemistry, which highlights the tremendous impact of their work already at that time. The determination of the DNA sequence in this first generation is either based on the chemical treatment of DNA (Gilbert) or on the incorporation of dideoxynucleotides (Sanger). After using nucleotide-selective conditions, the reaction products were size separated on ultrathin polyacrylamide gels and the sequence estimated from the relative position of the bases. Using that approach, it was now possible to sequence DNAs with a length of about 100 bases (reviewed in Heather and Chain, 2016).

Gilbert´s chemical cleavage method was an adaption of earlier protocols that have been developed in the Sanger lab to sequence tRNAs and proteins in the late 1960s but relied on a vast amount of starting material (Maxam and Gilbert, 1977). They improved

the sensitivity of the method by detecting chemical breaks at terminally radioactively labelled DNAs in a two-step process: (i) Introduction of base-specific (C, T+C, G, A+G) chemical modifications that weaken the respective base and (ii) removing this base from the phosphate backbone. However, this method was still very time-consuming and relied on high concentrations of radioactive isotopes. Improving both of these aspects, Sanger sequencing soon became far more prevalent and would dominate the market for the next 30 years (Shendure et al., 2017). For the chain-terminating sequencing, Sanger and his co-workers significantly improved their plus and minus method, that had already been applied at that time to solve the genome of the 5.2 kb bacteriophage φX174 (Sanger et al., 1977a). Using an *in vitro* system, an elongating DNA polymerase selectively incorporates a dideoxynucleotide (dd) instead of a deoxynucleotide (d) during replication of a DNA template. Lacking the 3´-hydroxyl group, the dd´s cannot be further extended, and the reaction terminates at that position (Figure 2). In the early years of that technique, four parallel sequencing libraries were made by adding each of the four dideoxynucleotides in a separate reaction (Sanger et al., 1977b). As for Gilbert´s approach, the detection was also performed by size-separation of the products using polyacrylamide gel electrophoresis (PAGE) and visualisation using autoradiography (Maniatis et al., 1975; Sanger and Coulson, 1978).

Based on these concepts, the process of DNA sequencing necessitates (i) the construction of a sequencing library and (ii) the actual determination of the bases (Figure 2). For the next years, the main challenge was to further develop and refine both aspects to improve the two most important metrics for DNA sequencing, costs and throughput. This would ultimately allow the assembly of the larger genomes of more complex organisms. To achieve this goal, one had to overcome the limitations of the short read-length of up to 300 base pairs (bp). Therefore, the genomic DNA was mechanically sheared and the resulting fragments of about 1 to 10 kb ligated into a vector (Staden, 1979; Messing et al., 1981). Randomly sequencing of the clones ensured that overlapping parts of the genome are sequenced, which helped during the assembly process. After its development, scientists at first used this so-called shotgun-approach for small genomes, like the bacteriophage λ in 1982, but then successfully applied to increasingly complex species (Sanger et al., 1982)(Figure 2). The following years brought new inventions and concepts that improved every single step of DNA sequencing. Two of the most important innovations that contributed to the semi-automation of the whole process were: (i) The replacement of radioactive labels by fluorescent labels, which enabled a one-pot reaction and (ii) a new way to detect the fluorometric signals by capillary electrophoresis (Prober et al., 1987; Zhang et al., 1995). Based on these improvements, Applied Biosciences introduced the first version of an automated DNA sequencer in 1986, that was capable of sequencing 5 kb per day (Smith et al., 1986)(Figure 2).

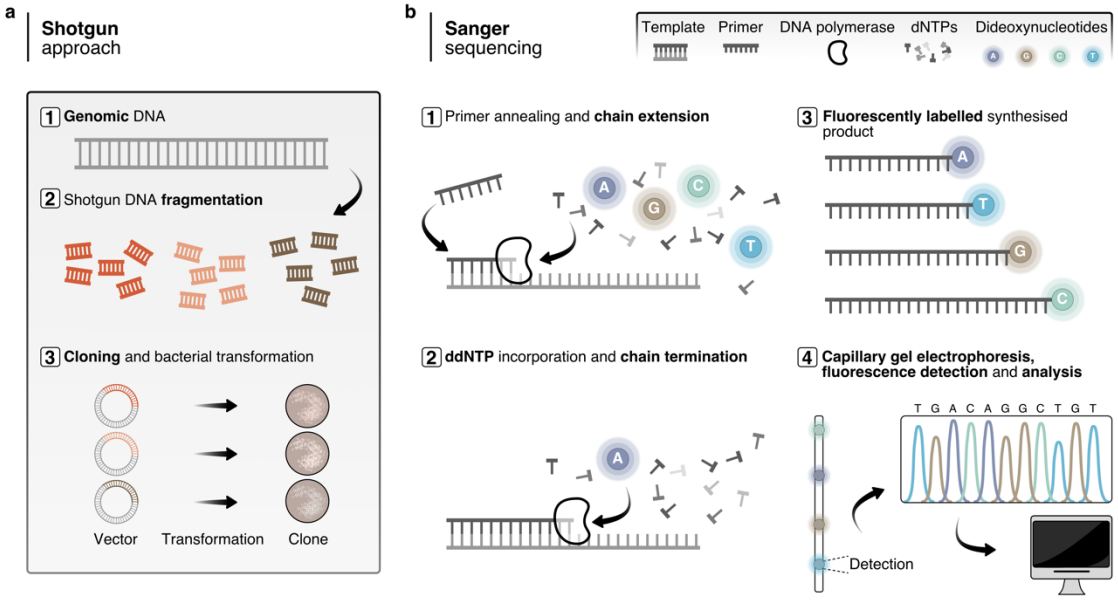


Figure 2 | 1st generation shotgun Sanger sequencing. a, During shotgun sequencing, DNA is first randomly fragmented into numerous small segments, that are subsequently cloned and amplified in appropriate vectors (Sanger et al., 1982). **b**, Improved protocol of Sanger sequencing using fluorescent labels and detection by capillary electrophoresis, which was used in the first automated DNA sequencers (Smith et al., 1986). The incorporation of a dideoxynucleotide by the DNA polymerase leads to chain termination (Sanger et al., 1977b). The size of the end-labelled product is detected by capillary electrophoresis and bases can be assigned based on the fluorescent signal.

In 1990, these improvements led to the launch of the Human Genome Project (HGP) by the U. S. government. Their vision was to determine the 3.3 billion bp of the human genome within the next 15 years, which was achieved ahead of its schedule with an initial draft published in 2001 and declared completed in 2003 with estimated costs of $3 billion (Craig Venter et al., 2001; Lander et al., 2001). On the way to the first human draft genome, different cloning strategies and technical improvements, like the use of magnetic beads for DNA purification, were tested in pilot projects (Deangelis et al., 1995). This led to the release of the first fully sequenced, 1.8 Mbp long genome of a free-living organism, *Haemophilus influenzae*, in 1995 (Fleischmann et al., 1995). Shortly after, a whole-genome shotgun (WGS) approach was also applied to the first archaeal organism, *Methanocaldococcus jannaschii* (Bult et al., 1996).

In the years that followed, many genomes of prokaryotic and more complex eukaryotic model organisms were released (Myers et al., 2000). Although costs had already dropped a lot since the beginning of the first sequencing era, it was still costly and the long-term goal using routine genomics as a diagnostic tool not anywhere near a reality yet. Consequently, the next big step was to dramatically reduce the costs by four orders of magnitude, enabling the sequencing of a human genome for $1,000 or less (Schloss, 2008). This project, started in 2004, was hoped to be achieved within ten years and brought one major innovation: Massively parallel sequencing.

1.2. 2nd generation: Massively parallel sequencing

In contrast to the three-decade-long monopoly of Sanger sequencing in the 1st generation, the 2nd generation started with many different ideas (reviewed in Goodwin et al., 2016). Despite their distinct characteristics, all "next-generation" sequencing approaches share specific attributes that allow massively parallel sequencing of DNA molecules: (i) Amplification of a target DNA is not achieved by cloning and transformation, but by the preparation of an *in vitro* library, (ii) millions of reactions are performed in parallel, (iii) direct detection of the signal (reviewed in van Dijk et al., 2014, 2018).

For the construction of a library and therefore the amplification of the template, DNA fragments are first ligated to platform-specific sequencing adapters and immobilised on a two-dimensional surface (Figure 3). Illumina, for instance, utilises a so-called "bridge-amplification", that tightly clusters copies of each template (Mitra, 1999; Adessi, 2000). A second approach that is used for pyrosequencing performs amplification by PCR in emulsion droplets (Margulies et al., 2005). Alternatively, circularised fragments can be copied by a rolling-circle amplification, which leads to many head-to-tail copies of one fragment (Drmanac et al., 2010). The resulting copies self-assemble in clonal DNA "nanoballs" (DNB) that are attached to the patterned array of a flow cell (Porreca, 2010). This technology is currently applied in the DNBseq™ platforms of the Beijing Genomics Institute (BGI) (Xu et al., 2019).

After amplification of the template, most of the platforms follow a sequence-by-synthesis (SBS) approach. In a cycle-wise manner, the incorporation of one nucleotide at a time can be monitored by different methods. During 2005 and 2011, three main strategies were used on various platforms:

Similar to Illumina sequencing, pyrosequencing involves the stepwise addition of each deoxynucleotide in a picolitre-sized reaction volume. The detection is based on a light-emitting chain reaction after pyrophosphate gets released when a dNTP is incorporated (Rothberg and Leamon, 2008). This set-up was used in the first NGS instrument, commercialised by 454 (now Roche) in 2005, and was able to produce reads in the length of 400 to 500 bp (Margulies et al., 2005). Although the strategy of Ion torrent sequencing is remarkably similar, it does not rely on the optical detection of incorporated nucleotides. Instead, semiconductor technology detects the emission of protons (Pennisi, 2010; Rusk, 2011).

The second approach, called sequencing by oligo ligation detection (SOLiD), was commercialised in 2007 by applied biosystems (now life technologies) (Valouev et al., 2008). SOLiD utilises the specificity of DNA ligases to attach fluorescently labelled oligonucleotide probes to an anchor sequence (Ju et al., 2006). Despite its very high accuracy (>99.99%), the short read length of initially 35 bps was a significant shortcoming.

The third technology that has been developed by Solexa (now Illumina) includes the stepwise addition of fluorescently labelled deoxynucleotides by a DNA polymerase (Mitra

et al., 2003). To ensure that only one nucleotide is incorporated at a time, each deoxynucleotide carries a reversible terminator. These systems make use of heavily engineered high-fidelity DNA polymerase to increase the sequencing throughput dramatically. After dNTP incorporation, the fluorescent dye is imaged multiple times, depending on channel chemistry (Figure 3). Following imaging, the terminating group and the fluorescent label are removed, and extension continues in the next cycle (Ruparel et al., 2005; Seo et al., 2005).

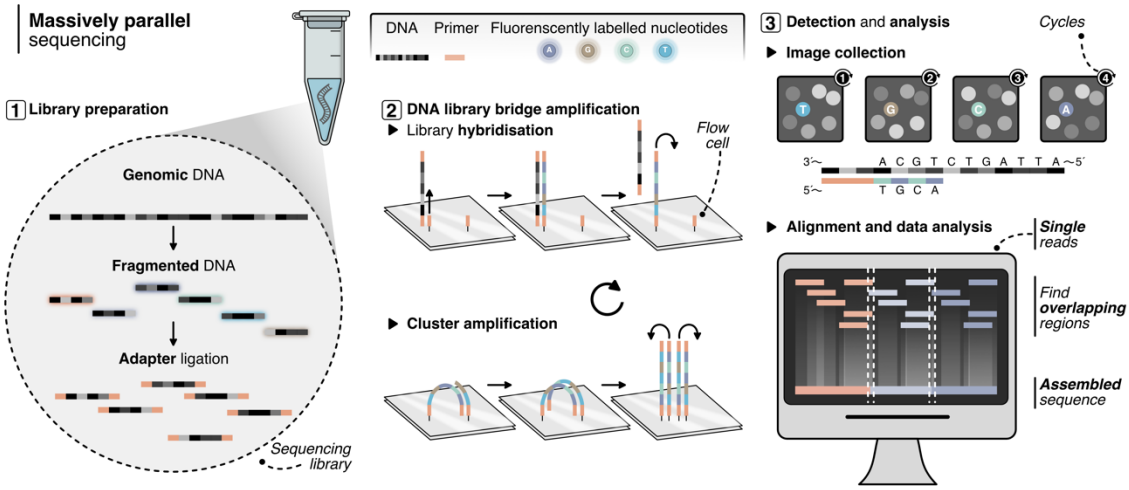


Figure 3 | 2nd generation massively parallel sequencing using Illumina technology. (1) Amplification of target DNA is performed by the preparation of an *in vitro* library after fragmentation of the DNA and adapter ligation. Sequencing is performed by a sequencing-by-synthesis approach (van Dijk et al., 2014). **(2)** Therefore, the library is first hybridised to complementary Illumina adapters on a flow cell and clusters are generated by bridge amplification. In each cluster, one fluorescently labelled deoxynucleotide with a reversible terminator is added by a polymerase one at a time (Mitra et al., 2003; Ruparel et al., 2005; Seo et al., 2005). **(3)** Images of the flow cell are collected during each cycle. After the image(s) are collected, the label and the terminating group are removed, and extension continues in the next cycle. For genome sequencing, short read sequences are scanned for overlapping regions and assembled to larger continuous sequences.

While different flavours of NGS technologies coexisted over many years, today undoubtedly Illumina dominates the market. With faster CCD cameras, higher-fidelity polymerases and more tightly packed patterned nanowells, the company managed to improve every step of their technology, which led to an unpredicted acceleration of data production. In 2014, the introduction of the HiSeq X Ten series, that was capable of sequencing 1.8 Terabases per run, broke the long-standing "$1000 genome" barrier (Check Hayden, 2014b, 2014a). Currently, Illumina prices a human genome with 30x standard coverage at about $600. At the 2020 "Advances in Genome Biology and Technology" conference, MGI announced that using a highly-parallel system and a revolutionary new antibody technology, they managed to drop costs to $100 per human genome (Drmanac et al., 2020).

However, one major limitation for some applications, that all 2nd generation technologies share is the short read length on all of their platforms. For Illumina and pyrosequencing, the main reason that impedes longer reads is the so-called phasing error (Fuller et al., 2009; Schirmer et al., 2015): In every cycle, there is a low probability that the terminator cap of a deoxynucleotide is not removed correctly. Hence, the sequence lags behind the rest of the cluster, which is termed post-phasing. Additionally, a pre-phasing error can occur, caused by inefficient flushing of the flowcell, which leads to the incorporation of two (or more) nucleotides in one cycle. During the SBS approach, this error adds up, and the clusters get out of sync (Pfeiffer et al., 2018). Ultimately, phasing leads to a polluted fluorescence signal, and a decreasing quality score over increasing length limits the maximum read length to about 300 bp (Schirmer et al., 2015).

1.3. 3rd generation: Long-read sequencing

With more and more genomes assembled using second-generation platforms, two significant concerns were raised: First, the aforementioned short reads are insufficient to assemble genomes containing highly repetitive regions (Salzberg and Yorke, 2005). Secondly, GC-rich parts are inefficiently PCR-amplified and are underrepresented in the therefore biased libraries (Dohm et al., 2008; Kozarewa et al., 2009; Chen et al., 2013). In parallel with the evolution of short-read platforms, different companies and labs worked on solutions to overcome the limitations of 2nd generation technologies. In 2009, the release of the first single-molecule fluorescent sequencer from Helicos Biosciences marked the transition from the 2nd generation to the 3rd generation of DNA sequencing (Gupta, 2008). The company used an implementation on the well-established Illumina protocol but without any bridge-amplification, facilitating a single-molecule readout (Pushkarev et al., 2009; Thompson and Steinmann, 2010). Inevitably, this dramatically decreased the throughput and was still not yet a solution to the read length-limiting phasing problem.

The first company that commercialised a revolutionary new approach was Pacific Biosciences (PacBIO) with the release of the PacBIO RS in 2011 that for the first time fulfilled all three key features of the third generation sequencing technology: (i) Single-molecule detection of (ii) long reads (iii) in real-time (Eid et al., 2009). Therefore, DNA sequencing is performed on a chip that contains zero-mode waveguides (ZMWs) with a single DNA polymerase and a template fixed to the bottom of each well (Levene et al., 2003) (Figure 4a). To ensure efficient loading into ZMWs and make use of the complementary strand information, a SMRTbell™ adapter is ligated to each template, that forms a hairpin cap at both ends. After the incorporation of one fluorescent nucleotide at a time, the tag is cleaved off and diffuses into the zeptoliter-big observation volume of the ZMW field. The limited reaction volume ensures that only a single molecule is detected at a time (Buermans and den Dunnen, 2014; van Dijk et al., 2014). However, because of a low signal-to-noise ratio and a minimum dwell time of wrong nucleotides, the error rate of

this process is quite high (13-15 %) (van Dijk et al., 2018). Although the company managed to strongly improve the mean read length from 1.5 kilobases (kb) in the beginning to more than 15 kb now and also increased the total throughput by more than 100-fold, the randomly distributed error rate is still an issue. To overcome this, PacBIO established two different sequencing modes: A circular consensus mode (HiFi mode), offering highly accurate reads (>99.999% accuracy) by sequencing up to 15 kb long circular templates and a continuous long read (CLR) protocol for reads larger than 50 kb (Wenger et al., 2019). The HiFi mode has a sequencing accuracy comparable to the Illumina platform at the cost of reduced throughput. Over the years, PacBIO technology has sharply improved on multiple aspects by changing the sequencing chemistries and by increasing the number of ZMW holes. However, ultralong reads on PacBIO platforms are not expected to become a reality soon, given the processivity limitations of the DNA polymerases that are currently available (van Dijk et al., 2018).

A second approach that is getting more and more attention is Nanopore sequencing. Surprisingly, from conceptualisation in the 1980s to the first Nanopore sequencer, it almost took 25 years (reviewed in Deamer et al., 2016). This long initial developmental phase was caused by many technical difficulties setting up a radically different sequencing platform. The basic idea is that the current-induced translocation of nucleotides through a Nanopore alters the electric signal in a base-specific way (Branton et al., 2008). Nanopores are tiny hole-forming proteins that are embedded into a membrane and provide the only way how ions can get from the *cis* to the *trans* side of a closed electrolyte-buffered chamber. By applying a constant voltage (*trans* side positive, *cis* side negative), the ionic current drives negatively charged single-stranded RNA (ssRNA) or DNA (ssDNA) through the pore (Figure 4b). For this, a patch-clamp amplifier supplies voltage and measures ionic alterations at the same time.

While most of the sequencing technologies from all generations adapted on polymerase-mediated sequencing approaches and could profit from the learning curve of each other, Nanopore sequencing had to start from scratch. One of the first breakthroughs came out of Nanopore-pioneer David Deamers lab in 1996. By applying an electric potential of about 120 mV, they managed to push a homopolymer of polyuridylic acid (poly U) through the channel-forming α-hemolysin from *Staphylococcus aureus* (Kasianowicz et al., 1996). The migration caused a current blockage that they used to estimate the length of the polymer.

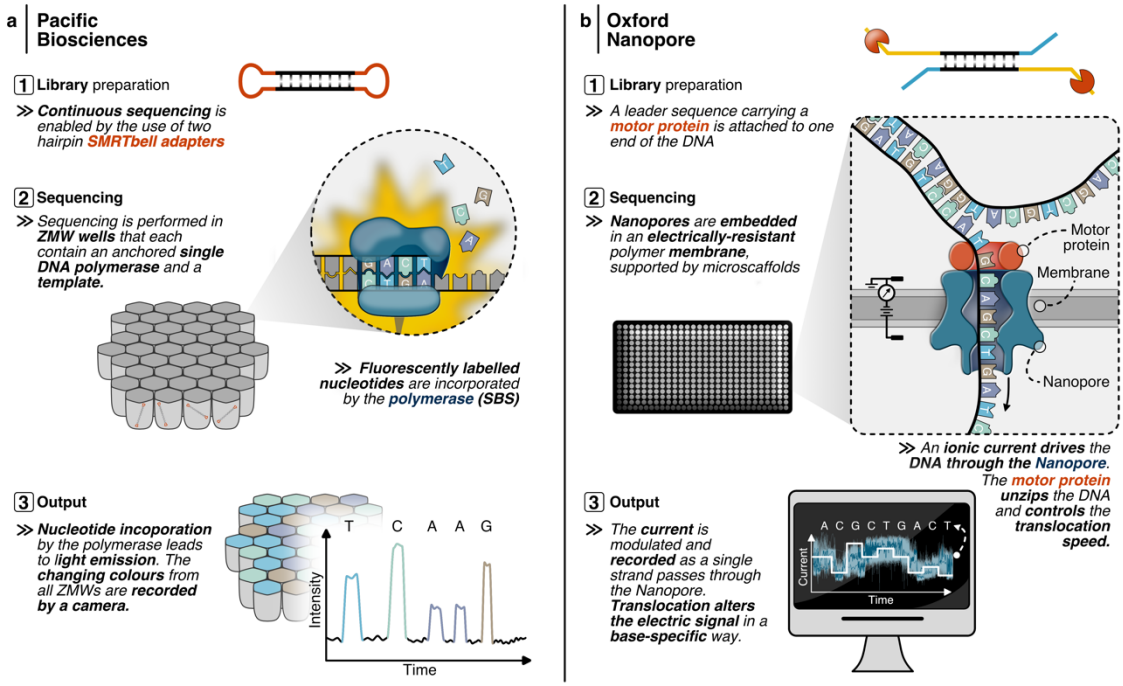


Figure 4 | 3rd generation long-read sequencing technologies. a, The development of PacBio sequencing marks the transition to 3rd generation technology, allowing long-read single-molecule sequencing in real-time. Sequencing is performed by sequencing-by-synthesis in millions of zero-mode waveguides in parallel that each contain a template, DNA polymerase and a template fixed to the bottom of a well (Eid et al., 2009). Light emission after nucleotide incorporation is recorded by a camera. **b**, Nanopore sequencing is based on changes in the ionic current when nucleotides translocate through a membrane-embedded pore-forming protein from the *cis* side to the *trans* side in a closed electrolyte-buffered chamber. During library preparation, a motor-protein carrying adapter is added, which unzips double strands, controls the speed and drives the nucleotides through a heavily engineered Nanopore. Raw current signals are recorded and basecalled to obtain the nucleotide sequence (reviewed in Deamer et al., 2016).

With the crystal structure of α-hemolysin solved in 1996, it was then possible to understand the molecular basis of translocation and blockage, which also helped in designing new experiments (Song et al., 1996). The heptameric pore has a vestibule diameter of 2.6 nm, therefore preventing the entry of dsDNA and a limiting aperture of 1.5 nm in the neck of the channel. A few years later, this mushroom-like structure enabled the first Nanopore-based discrimination of purines and pyrimidines based on the comparison of the blockage amplitudes (Akeson et al., 1999). However, a time-resolved differentiation of single nucleotides was not possible at that time, because under the applied voltage, the speed of the rate of translocation was too high (Meller et al., 2000). Finding a way to slow this process down constituted a significant problem. After a decade long search, a phage polymerase, called φ29, was discovered, that could hop on DNA, control the DNAs movement against the voltage bias through the pore in a stable way, and work under the sequencing conditions (Cherf et al., 2012; Manrao et al., 2012). Using this φ29 DNAP in combination with the α-hemolysin proved to be not successful because the stem of the pore

protein was too long and also contained negatively charged residues possibly interfering with the nucleotides. Therefore, the DNAP was combined with the funnel-like MspA pore, that had been shown to be a good candidate for current-based discrimination of different bases (Manrao et al., 2011). In a milestone paper published in 2012, the MspA/φ29 approach allowed sequencing of six different DNA templates (42 to 53 nucleotides long) at single-nucleotide resolution (Manrao et al., 2012).

Since that time, the fundamental concepts, to use a speed-controlling motor protein, and a structurally advantageous pore-forming protein, have not changed. In fact, they have been optimised and commercialised by 2005 founded British company Oxford Nanopore Technologies (ONT). After abandoning an "exonuclease-sequencing" approach, that works by sequentially cleaving off single nucleotides close to the Nanopore, but is limited to less than 80 bases, ONT entirely focused on the "strand sequencing" approach (Clarke et al., 2009; Reiner et al., 2012). At the Advances in Genome Biology and Technology conference in 2012, the company not only announced the alpha version of the MinION, a handheld DNA sequencer using Nanopore technology, but also the successful *de novo* sequencing of the lambda phage with reads spanning the whole genome of more than 48 kb (Brown et al., 2015). After technical difficulties in the chip design, an early access program (MAP) was initiated, in which the MinION devices were given to multiple labs (Jain et al., 2016). By continually improving highly engineered pores and motor proteins, ONT managed to increase the average identity (proportion of matches in a mapped sequence) from 66% (R6 pore, June 2014), to 85% (R7 pore, December 2014), and to >90% in the R9 pore versions (Jain et al., 2015, 2017; Laver et al., 2015). The R9 pore is a nonameric derivative of the Curlin sigma s-dependent growth gene (CsgG) from *Escherichia coli* which runs at about 250 bases per second (Remaut et al., 2014; Brown and Clarke, 2016). The latest release in pore chemistry, R10, enables a better resolution of homopolymer regions by a longer barrel and dual reader head, which improved consensus accuracy as high as Q50 (99.999%) (Karst et al., 2020, press release ONT 13th January 2020). Despite the tremendous improvements in the accuracy, high error rates, especially compared to Illumina sequencing, are still a matter of debate in the academic community. Improving the accuracy using a circular consensus approach, as performed by PacBIO, is not possible, as the template has to migrate through the pore. Recently, ONT introduced a $1D^2$ library preparation kit, that enables sequencing of both strands without physical ligation, which improved the accuracy to 97% compared to 90% 1D assay (press release ONT).

The high error rate in Nanopore sequencing data can be traced back to two different problems: First, inherent sequencing errors mainly due to a low signal-to-noise ratio. This is caused by the similar electrical field of the nucleotides, the fact that more than one base contributes to a signal change, the stochastic motion of the motor protein and homopolymer errors that cannot be distinguished (Laszlo et al., 2014; Bowden et al., 2019). The second problem comes from errors that are introduced in the basecalling process from

raw data into a DNA sequence (Rang et al., 2018; Wick et al., 2019). New pores and motor proteins are tested continuously to decrease inherent sequencing errors. Additionally, new algorithms and different concepts are developed and tested for better data interpretation (Wick et al., 2017).

Nevertheless, it depends on the application, whether sequencing errors constitute a problem, or if the long-read sequencing capability outweighs this limitation. Excitingly, Nanopore sequencing allows that DNA molecules of any length can be sequenced (Jain et al., 2018; Tyler et al., 2018). Conceptually, read length is not limited by ONT sequencing itself, rather than by the preparation of the DNA. Whale-fishing for ultra-long reads led to the rapid development of specialised protocols with a current record of a 2.44 Mb fusion read (Jain et al., 2018; Payne et al., 2019). Although the unlimited sequencing length is maybe the main bonus of Nanopore sequencing, other aspects like the availability of rapid field-suitable protocols, the omission of PCR steps and a meanwhile increased throughput are also of great advantage and likely to improve in the future (Mongan et al., 2020).

From a historical perspective, sequencing technologies have evolved continuously over the last 40 years (for excellent reviews see Goodwin et al., 2016; Shendure et al., 2017). With a dramatic increase in throughput, sequencing is now possible at reasonable costs, which enabled ground-breaking discoveries in all fields of biology. One of the initial ideas, to use genomics in precision medicine has become a reality. Additionally, the technologies have successfully been applied to monitor viral outbreaks, emphasising their role as a genomic surveillance system (Quick et al., 2016, 2017; Mongan et al., 2020). Also, during the current COVID-19 pandemic outbreak second- and third-generation sequencing technologies have already contributed critical findings to fight the pandemic and provided a way for cost-efficient rapid mass-testing (Huang et al., 2020; Zhou et al., 2020; Zhu et al., 2020). Recently, the FDA (U.S. Food and Drug Administration) approved the first COVID-19 diagnostic test utilising Illumina technology, which will not only help in the detection of SARS-CoV-2 from individuals but also can be used to track mutation rates. Additionally, ONT for the first time has developed a highly scalable diagnostic assay, called lamPORE, for the detection of SARS-CoV-2.

Besides the crucial role in the current situation, the generation of large-scale data, complemented with additional analysis, has transformed archaeal research and allowed exciting insights into their diversity and evolution.

2. Archaea: Evolution, model organisms & genomes

2.1. Shaping the tree of life

The universal tree of life has been revisited twice within the last 45 years. After the proposal of a three-domain model consisting of Eukarya, Bacteria and Archaea in 1990, this classification into three major lineages became widely accepted and found its way into the textbooks (Woese et al., 1990). Meanwhile, this view has again been replaced by a (still highly debated) two-domain topology, with Eukarya emerging from within the Archaea (Williams et al., 2020).

Before that time, Archaea have been misclassified as Bacteria for many years, mainly caused by their morphological similarity under the microscope and other shared physical and metabolic features. This changed in 1977 when Carl Woese and George Fox introduced a new concept for phylogenetic taxonomy that was based upon 16S ribosomal RNA sequences. Therein Archaea (initially designated as "Archaebacteria") were defined as the third domain of life (Balch et al., 1977; Woese and Fox, 1977). In the following years, more in-depth phylogenetic analysis was limited by the number of available sequences in the databases (Figure 5). However, the representatives of the archaeal domain known in the early 1980s all had very odd lifestyles, inhabiting strictly anoxic (methanogens), hypersaline (halophiles) and geothermally heated (hyperthermophiles) environments (Albers, 2016; Forterre and Fagan, 2016). Given their evolutionary distance to Bacteria and their preference for extreme habitats, Archaea were initially considered as an ancient form of living. This opinion was revised thanks to the development of cultivation-independent techniques, which started in 1992. During the following years, these approaches revealed that Archaea make up a large part of marine and terrestrial ecosystems and therefore also play significant roles in "nonextreme" habitats (DeLong, 1992, 1998; Fuhrman and McCallum, 1992). The phylogenetic diversity was further increased with the discovery of other lifestyles, like the nanosized *Nanoarchaeum equitans*, that lives in an ectosymbiotic partnership with its host cell *Ignicoccus hospitalis* (Huber et al., 2002). Genomic sequencing revealed that *N. equitans* has a minimal, reduced genome with only 0.49 Mb, resulting in metabolic limitations and therefore a growth-dependency on its host (Huber et al., 2002; Waters et al., 2003).

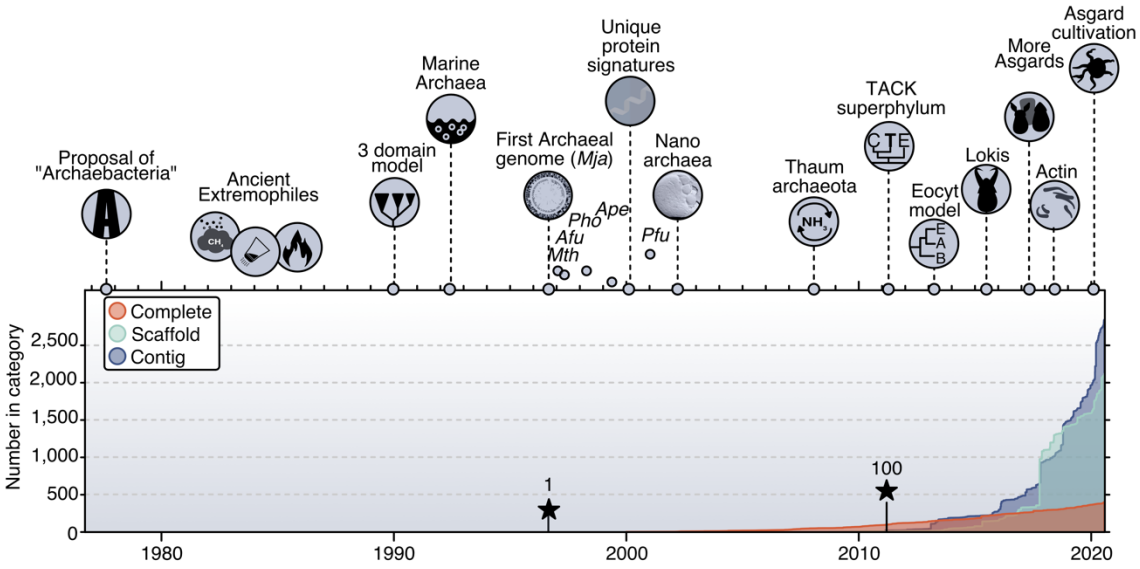


Figure 5 | History of Archaea in a genomics context. Graphical summary of selected events in more than 40 years of archaeal research. Insights in their ecology, diversity and evolution were dramatically influenced by the availability of genomic sequences (bottom panel). The total number of archaeal complete genomes (red), scaffolds (light-green) and genomic contigs (purple) have been retrieved from the NCBI assembly database (Kitts et al., 2016). The increase in scaffolds and contigs from 2016 on can be traced back to the development of deep metagenomic sequencing (Scholz et al., 2016).

Although all prokaryotes share similar genome sizes, which allowed whole-genome sequencing at reasonable costs already in the 1990s, the number of complete archaeal genomes lagged behind the total number of sequenced Bacteria (Kellner et al., 2018). Despite that limitation, a signature set of archaeal protein clusters could be identified by analysing the genomes of nine available archaeal representatives (Graham, 2000). Almost 15% of the encoded proteins were exclusively found in archaeal species, which further validated the hypothesis at that time that Archaea comprise a lineage that diverged early in evolution (Graham, 2000). All representatives of that lineage had been grouped into one of the two major phyla, the Euryarchaeota and the Crenarchaeota, based on the RNA component of their small ribosomal subunit (Woese et al., 1990). In 2008, Thaumarchaeota were proposed as a third archaeal phylum, by combining two major findings (Brochier-Armanet et al., 2008): First, the identification of mesophilic ammonium-oxidizers and whole-genome sequencing of one of the representatives, *Cenarchaeum symbiosum*, that is remarkably distant from the other phyla (Hallam et al., 2006). Second, a more sophisticated phylogenetic method based on concatenated ribosomal proteins, that improved the resolution of the tree (Matte-Tailliez et al., 2002). Later, Thaumarchaeota were also identified as the significant archaeal players of the human skin microbiome, possibly contributing to skin health by influencing the pH (Probst et al., 2013; Moissl-Eichinger et al., 2017). Archaeal communities are not only present on the human skin, but also in the gastrointestinal and respiratory tract of the human body (Koskinen et al., 2017). Although methanogenic Archaea have been associated with different human diseases, no pathogenic

species has been identified so far, and the role of Archaea in human health remains yet to be determined (Bang and Schmitz, 2015; Lurie-Weinberger and Gophna, 2015).

Excitingly, the availability of archaeal genomes, however incomplete, exploded in the mid-2010s due to the development of large-scale metagenomic sequencing (Figure 5). The results obtained from these new opportunities rekindled the conflicting question about the role of Archaea in the evolution of the first eukaryotic cell. In the classical three-domain tree of life, eukaryotes share a common ancestor with Archaea, but arise independently and not from within the Archaea, which in contrast is the idea of the eocyte hypothesis (Woese et al., 1990; Rivera and Lake, 1992). Over time, the latter idea became supported by new phylogenetic methods and the occurrence of eukaryotic signature proteins (ESPs) in the proposed 'TACK' superphylum (Guy and Ettema, 2011; Williams et al., 2013).

However, the nature of a putative archaeal ancestor remained hidden until the discovery of a novel deeply rooting phylum, called Lokiarchaeota, in 2015 (Spang et al., 2015). This new group sampled from a site at the Arctic Mid-Ocean Spreading Ridge, called Loki´s castle, forms a monophyletic group with Eukaryotes and therefore strongly supported the eocyte hypothesis (Pedersen et al., 2010; Jaeschke et al., 2012). In the following years, closely related metagenome-assembled genomes (MAGs), all named after north gods (Heimdall, Thor, Odin, Hel) were revealed and grouped in the Asgard superphylum (Seitz et al., 2016, 2019; Zaremba-Niedzwiedzka et al., 2017). Based on genomic analysis, an unusually high number of ESPs, among them proteins that may have a cytoskeletal function or are essential for membrane remodelling, were found in all Asgards (Eme and Ettema, 2018). Although the biochemical and structural analysis of Loki- and Odin-profilin suggested that Asgard Archaea have a profilin-regulated actin cytoskeleton, it was not clear whether all of the ESPs found in Archaea have the same function as their eukaryotic counterparts (Akıl and Robinson, 2018; Eme and Ettema, 2018).

Nevertheless, a more in-depth characterisation of the most likely eukaryotic ancestor was mainly limited by the absence of any cultivated or at least imaged Asgards. Excitingly, in 2020, Imachi et al. reported that in the course of a 12-year-long study, they finally managed to culture an Asgard archaeon (Imachi et al., 2020). The extremely slow-growing organism, which they propose to name *Prometheoarchaeum syntrophicum*, was found in deep marine sediments and lives in a syntrophic association with microbial partners. Despite the absence of intracellular organelle-like structures, the organism has a unique morphology and forms long protrusions, that could help with the engulfment of a symbiont (Imachi et al., 2020; Schleper and Sousa, 2020). This scenario provides yet another model for eukaryogenesis, that is difficult to prove. Nevertheless, what seems to be accepted in the scientific community is the two-domain topology of the tree of life. This now has been reasserted multiple times by different sophisticated phylogenetic techniques, and also by the discovery of the Asgards, which are currently the closest relatives to Eukarya (Williams et al., 2020).

2.2. Microbe profile: *Pyrococcus furiosus*

In a historical context, hyperthermophilic Archaea once have been placed near the root of the phylogenetic tree, therefore representing the closest relatives to the last universal common ancestor (LUCA) (Pace, 1991; Stetter, 2006). Later it was shown that the phylogenetic models at that time, which were solely based on the 16S rRNA, were heavily flawed by the high rRNA-GC content in this group of microorganisms (Forterre, 1996). However, it was not only evolutionary aspects that made them so interesting, but also their extreme lifestyle and the presence of thermostable enzymes that could be used in industrial processes (Ebaid et al., 2019). Hyperthermophiles per definition are microbes with an optimal growth temperature of above 80°C, that inhabit various habitats, from marine hydrothermal vents to hot springs and are also found from Bacteria to Archaea (Urbieta et al., 2015). One group of organisms that belongs to the Euryarchaeota and can survive the anaerobic, boiling environment of hydrothermal vents is called Thermococcales (Price et al., 2015). The order consists of three genera, *Pyrococcus*, *Thermococcus*, and *Paleococcus*, which amongst other things can be distinguished by their optimal growth temperatures of 100°C, 85°C and 65°C, respectively (Zillig et al., 1983; Fiala and Stetter, 1986; Takai et al., 2000). Despite their harsh growth temperature under anaerobic conditions, they can grow very rapidly to high cell densities on simple organic media (Leigh et al., 2011). Given the feasibility of biochemical studies, the early release of complete genomes and the development of genetic systems, some of the representatives in the order of Thermococcales are favoured model organisms, to study different aspects of archaeal biology.

One of the best-described hyperthermophiles is *Pyrococcus furiosus*, the "furious fireball" (Kengen, 2017). This fast swimmer was isolated in 1986 from geothermally heated marine sediments at the beach of Porto di Levante at the island of Vulcano in Italy (Fiala and Stetter, 1986). It is an obligately anaerobic, hyperthermophilic organism with an optimal growth temperature of about 100°C (Fiala and Stetter, 1986). With a rapid doubling time of ~37 min and the tolerance of handling small amounts of oxygen, *P. furiosus* can grow to high cell densities under laboratory conditions. *P. furiosus* owes his name to its motility by a bundle of Type IV pilin-like archaella (Näther et al., 2006; Daum et al., 2017). However, archaella are not only used for swimming, but also to form dense cell-cell connections by building cable-like structures, and can also adhere to solid surfaces, which allows them to colonize black smokers (Näther et al., 2006; Wirth et al., 2018). *Pyrococcus* has a heterotrophic metabolism and grows on a variety of carbon sources, like starch, maltose, chitin and pyruvate. Depending on the absence or presence of S^0 as an electron acceptor, hydrogen or hydrogen sulfide, organic acids and CO_2 are produced as the main fermentation products (Adams et al., 2001; Koning et al., 2002; Gao et al., 2003; Lee et al., 2006). The discovery of novel metabolic enzymes in *P. furiosus* led to the proposal of a modified Embden-Meyerhof pathway that uses ferredoxin instead of NAD as

the final electron acceptor and does not require a phosphorylation step (Kengen et al., 1994; Mukund and Adams, 1995). Interestingly, ATP is generated by a primitive type of respiration, that involves a single membrane-bound hydrogenase that uses ferredoxin to build up a proton gradient and therefore couples hydrogen production to energy conservation (Sapra et al., 2003).

These studies created the basis for future metabolic engineering, which became possible after the development of a genetic system (Waege et al., 2010; Lipscomb et al., 2011). Meanwhile, it is possible to intervene in metabolic pathways to ultimately produce products of interest for biotechnology in an advantageous continuous high-temperature fermentation process, which lowers cooling costs and the risk for contaminations (Zeldes et al., 2015). The combination of heterologous expression of target genes and knockout of other enzymes was for example used for the production of ethanol and butanol (Basen et al., 2014; Keller et al., 2015, 2017, other examples reviewed in Kengen, 2017). However, not only the metabolomic flexibility but also highly thermostable enzymes of *P. furiosus* are of outstanding interest for biotechnology (Cabrera and Blamey, 2018). The most prominent example is the DNA polymerase of *P. furiosus*, that was isolated and described in 1991, and since then has been used as the key enzyme in polymerase chain reactions (Lundberg et al., 1991). Unlike the traditional *Taq* (*Thermus aquaticus*) polymerase, the *Pfu* polymerase possesses an additional 3´ to 5´ exonuclease activity, providing it with a proofreading ability (Lundberg et al., 1991; Cline, 1996). This activity results in an error rate that is approximately 8-fold lower compared to the *Taq* polymerase and is advantageous in applications that require high fidelity (McInerney et al., 2014). However, because higher accuracy is achieved at the cost of processivity, commercialized *Pfu* polymerases have been fused to a DNA-binding domain and are now commonly used as a standard application in polymerase chain reactions.

2.3. Genomic architecture

Many of the discoveries in *P. furiosus* were driven by the exploration of its genomic content, providing an extensive source for further studies on the biochemical and molecular level. The original genome version released in 2001 had a size of 1.9 Mbp and contained 2065 open reading frames (ORFs) (Robb et al., 2001). Meanwhile, the genome annotation of the initial assembly has been updated multiple times, but the numbers roughly stayed the same (Poole et al., 2005; Haft et al., 2018). In general, these numbers reflect the common gene dense genome architecture that is shared by Archaea and Bacteria (Koonin and Wolf, 2008; Kellner et al., 2018). Interestingly, this architecture was not shaped by similar molecular processes, but by the adaption to shared environmental niches during evolution (Brooks et al., 2011).

As a consequence of tightly packed genomes, intergenic regions are rather small. This effect is caused by the so-called deletion bias and a selection against non-functional regions

during evolution (Sela et al., 2016). Accordingly, the rate of deletion is assumed to be always higher than the rate of the acquisition of new genes, which leads to a streamlining of prokaryotic genomes (Kuo and Ochman, 2009). However, there are two remarkable differences between the prokaryotic domains: First, there is a trend towards larger intergenic regions with increased genome size in Archaea (Kellner et al., 2018) (Figure 6a). This correlation is caused by a weaker purifying selection in bigger archaeal genomes compared to Bacteria, which leads to an enrichment of selfish DNA (Lyu et al., 2017). Secondly, the variation in the genome sizes in Bacteria is much more considerable (100x) than in archaeal genomes (10x) (Figure 6a). Although the lower number of available archaeal genomes may to some extent introduce bias, it is far more likely that genome-size variation is caused by a higher degree of reductive genome evolution for bacterial parasites and symbionts (Kellner et al., 2018; Zhu et al., 2019b). Another result of the streamlining of prokaryotic genomes is the operonisation of a substantial fraction of genes (Lynch, 2006; Rocha, 2008; Koonin, 2009).

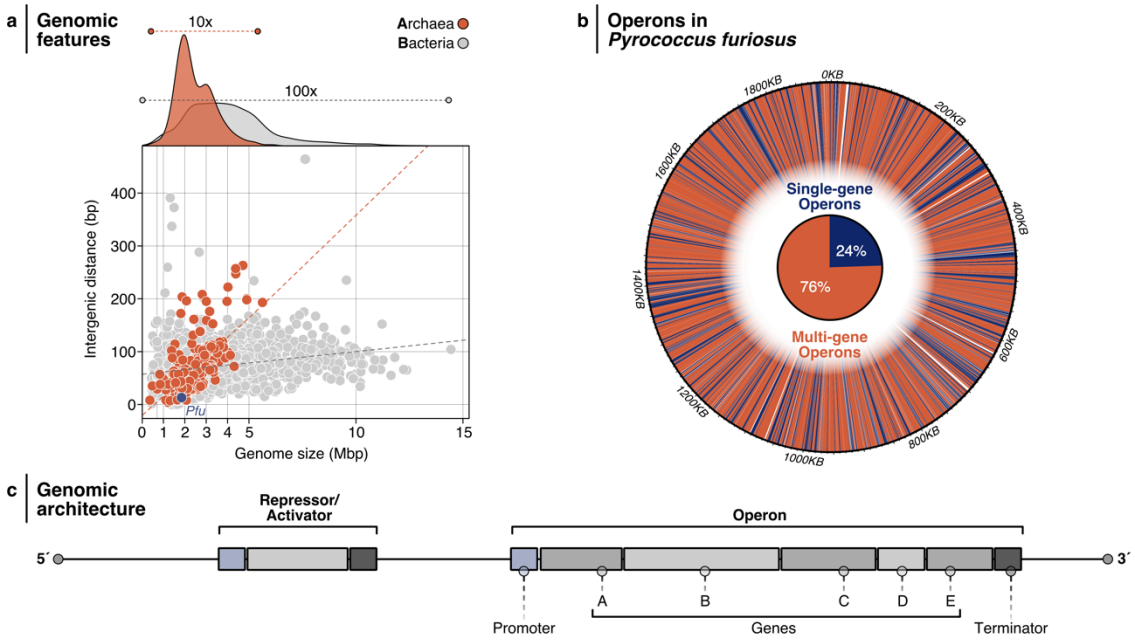


Figure 6 | Genomic features and operon organisation. a, Correlation between genome size (x-axis, in million base pairs) and the intergenic distance (y-axis, in base pairs) in archaeal (red, n = 246) and bacterial (grey, n = 2346) species. Higher correlation in Archaea is indicated by the dashed red line. Variation in genome size is shown as a density plot in the upper panel. **b**, Circular representation of operonisation in *Pyrococcus furiosus*. 76% of the genes are arranged in multi-gene operons (red) containing at least two genes. **c**, Schematic of features in the genomic architecture of gene-dense prokaryotic genomes that frequently contain multi-gene operons under the control of a single promoter and terminator. Transcription is fine-tuned with the help of repressors/activators, which genes are mostly encoded in upstream orientation of the operon.

In most of the archaeal and bacterial species, many genes are transcribed from a common promoter, which results in the formation of polycistronic transcripts (Figure 6b,c). In quantitative terms, 470 multi-gene operons have been predicted in *P. furiosus*, which constitutes a large proportion of the genome (Tran et al., 2007) (Figure 6b,c). Operons improve the fitness of organisms that reorganise their genomes frequently in their natural habitat by large-scale deletion events, the activity of insertion sequences (IS) or horizontal gene transfer (HGT) (Yoon et al., 2011). Because these events happen more or less randomly during evolution, the composition and order of operons are poorly conserved, even in closely related organisms (Baliga et al., 2004). However, the total degree of operonisation is remarkably similar between phylogenetic classes, presumably shaped by natural selection constraints (Yoon et al., 2011). Interestingly, this hypothesis is supported by a study that links operon stability to thermo-adaption, as operon disruption would lead to a significant loss of fitness in hyperthermophiles (Glansdorff, 1999; Yoon et al., 2011). Understanding higher-level genome organisation is critical for the analysis of gene-regulatory aspects, as a single regulatory gene can efficiently modulate the mRNA levels of operons.

Throughout evolution, archaeal genomes have been shaped under the premise of the functioning of the transcriptional apparatus, its interplay with regulatory elements on the DNA, transcription factors and the translation machinery. Although the influence of operonisation on RNA levels and differential expression of transcriptional units are only poorly understood in Archaea, the concepts of basal and regulated archaeal transcription are well-studied, established the basis for further studies and are addressed in the next paragraph.

3. The mosaic nature of archaeal transcription

The central dogma in molecular biology initially described the one-sided flow of information from DNA to RNA to protein (Crick, 1970). While the simplicity of this model is meanwhile refuted, the essential molecular processes behind, namely replication, transcription and translation, are still the same in all living cells. Considering their close evolutionary relationship, it is not surprising that Archaea have remarkably similar molecular machineries compared to Eukarya (Eme et al., 2017; Fouqueau et al., 2018). In contrast, the archaeal genomic landscape is very different from Eukarya, and like bacterial genomes has been streamlined during evolution (Kellner et al., 2018). Regarding transcription from DNA to RNA, Archaea use a simplified eukaryotic basal transcription machinery set in a bacterial-like genomic architecture, which represents some special requirements. Moreover, the fine-tuning of expression is performed by gene-specific transcription factors (TF), that structurally resemble the TFs from Bacteria (Lemmens et al., 2019).

This fascinating architecture of basal and regulated transcription relies on the interplay of DNA signatures and multiple proteins, which play crucial roles in one of the most fundamental cellular processes.

3.1. Basal transcription

3.1.1. The RNAP

Long before the close relationship between Archaea and Eukarya was revealed based on data from metagenomic sequencing and new phylogenetic models, striking similarities were already observed in the early 1980s by comparing the structures of DNA-dependent RNA polymerases (RNAP) (Zillig et al., 1979; Huet et al., 1983; Williams et al., 2020). All cellular RNAPs share a common core, comprising of five universally conserved subunits, that includes the catalytic centre and is essential for transcription. However, additional subunits are only present in Archaea and Eukarya, that help during assembly or interact with DNA, RNA and proteins (Werner and Grohmann, 2011). In contrast to the set of eukaryotic RNAPs, that specialised in transcribing rRNAs (RNAP I), mRNAs (RNAP II) and tRNAs (RNAP III), Archaea and Bacteria use a single RNAP for transcription (Decker and Hinton, 2013). Nevertheless, not only the RNAP subunits but most features in archaeal transcription are strikingly similar to the eukaryotic RNAP II system, including general transcription factors and DNA sequence elements. In general, transcription from DNA to RNA is achieved in a tightly coordinated iterative multi-step process, which consists of initiation, elongation and termination.

3.1.2. Initiation

Initiation of transcription in Archaea starts with the recognition of promoter elements by the general transcription factors TBP (TATA-binding protein) and TFB (Transcription factor B) (Blombach et al., 2019) (Figure 7). These factors bind to promoter elements, called BRE (Transcription factor B recognition element) and TATA-box (named after its TATA consensus sequence), that are positioned upstream of the transcription start site (TSS) of a gene (Littlefield et al., 1999; Smollett et al., 2017a). In many archaeal phyla, binding of TBP induces a bending of the DNA double-strand of about 90°, which is further stabilised by TFB (Gietl et al., 2014). Furthermore, TFB determines the orientation of this complex and recruits the RNAP by multiple interactions to form the preinitiation complex (PIC) (Bell et al., 1999b; Dexl et al., 2018). Initially, the PIC is in a closed conformation (CC) and not able to productively synthesise RNAs. The conformational change to a productive open complex (OC) is facilitated by TFE (Transcription factor E) (Blombach et al., 2015) (Figure 7). Additionally, DNA melting during open complex formation is enhanced in some Archaea by a motif-unspecific, yet AT-rich DNA signature, called initially melted region (IMR) (Fouqueau et al., 2018). After the template strand is positioned in the active centre, the RNAP preferentially starts transcription precisely at a pyrimidine/purine dinucleotide initiator element (INR) located 22 to 27 base pairs downstream from the TATA box (Hausner et al., 1991; Blombach et al., 2016; Smollett et al., 2017b). However, because of the tight interactions to TBP and TFB, the RNAP is constrained from leaving the promoter (Spitalny and Thomm, 2003; Fouqueau et al., 2013). This state leads to abortive cycles of synthesis accompanied by the release of very short RNA products (3-9 nt) (Goldman et al., 2009). During that process, the RNAP remains stationary and pulls in the downstream DNA template, which most likely accumulates inside the enzyme and leads to DNA scrunching, similar to what has been observed in Bacteria (Revyakin et al., 2006).

3.1.3. Elongation

To enter the productive phase of transcription, the RNAP first has to be released from TBP and TFB. This involves interactions of the DNA-RNA hybrid and displacement of TFB domains, that lead to a destabilisation of TFB, which in turn induces the promoter escape and marks the transition to the elongation phase (Fouqueau et al., 2013; Dexl et al., 2018). During this transition, TFE is swapped by the only RNAP-associated universally conserved transcription factor Spt4/5, which further stabilises the elongation complex and improves the processivity (Hirtreiter et al., 2010a; Grohmann et al., 2011; Smollett et al., 2017b) (Figure 7). This function is required as the RNAP is frequently disturbed during

elongation, either by protein barriers that build a roadblock, by pausing sites in a sequence-specific context or by DNA lesions (Sanders et al., 2019).

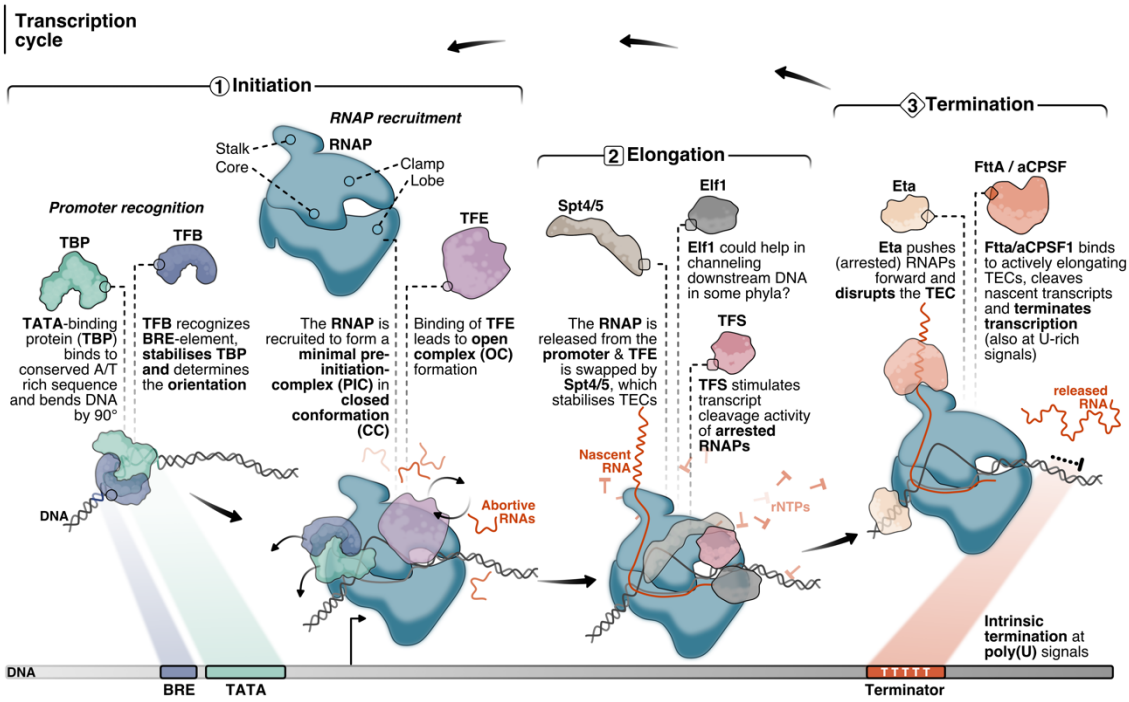


Figure 7 | The archaeal transcription cycle. Sequence elements and general transcription factors involved during the three-step transcription cycle in Archaea (reviewed in Blombach et al., 2019, Fouqueau et al., 2018). **(1)** After promoter recognition by TBP and TFB, the RNAP is recruited to form an initial pre-initiation complex in closed conformation, not able to productively transcribe RNAs. In the next step, TFE first stimulates open complex formation and **(2)** is then swapped by Spt4/5, which stabilises the actively transcribing elongation complex. **(3)** Transcription is terminated intrinsically by poly(T) sequences on the DNA or in a factor-dependent way by FttA/aCPSF1 (Maier and Marchfelder, 2019; Sanders et al., 2020; Yue et al., 2020).

In addition to Spt4/5, TFS (transcript cleavage factor S) helps to accelerate transcription through archaeal histones, that represent one of the most common protein barriers (Sanders et al., 2019) (Figure 7). Like all the initiation factors, TFS is also homologous to a factor from the eukaryotic Pol II system, namely TFIIS. Mechanistically TFS rescues stalled and backtracked transcription complexes by enhancing the intrinsic cleavage activity of the RNAP and is critical for the transcription in a chromatin context, while the Spt4/5 complex stabilises a closed clamp configuration and therefore increases RNAP processivity (Hausner et al., 2000; Fouqueau et al., 2017). Structural insights from the yeast transcription system have revealed the presence of yet another factor, called elongation factor 1 (Elf1), that interacts directly with the RNAP in close vicinity to Spt4/5 and could play a role in channelling DNA (Ehara et al., 2017) (Figure 7). Genome-wide profiling data of the eukaryotic Pol II machinery revealed an elongation-typical occupancy of Elf1, emphasising the importance of this factor on a global level (Mayer et al., 2010).

However, in Archaea, Elf1 is only present in the TACK and Asgard superphylum, and its role remains speculative (Daniels et al., 2009; Spang et al., 2015).

3.1.4. Termination

Challenged by different obstacles, like roadblocks or DNA lesions, the transcription elongation complex (TEC) has to be highly stable to prevent the RNAP from falling off the DNA template and achieve effective full-length RNA production (Gehring and Santangelo, 2017). Nevertheless, at the end of genes transcription has to be terminated very efficiently, to minimise collisions of converging RNAPs and conflicts with replication (McGlynn et al., 2012; Maier and Marchfelder, 2019).

Termination can either occur factor-independent at intrinsic signals or factor-dependent with the help of additional proteins that disrupt the TEC (Blombach et al., 2019). Intrinsic termination occurs at poly(U) signals *in vitro* and *in vivo* (Santangelo and Reeve, 2006; Santangelo et al., 2009; Maier and Marchfelder, 2019) (Figure 7). Bacteria follow the same strategy; however, poly(U) stretches are usually preceded by stem-loop structures, which are not essential in Archaea (Dar et al., 2016; Ray-Soni et al., 2016). Interestingly, this varies depending on the environment, as high temperatures impede the formation of short secondary structures. While stem-loops do not precede terminators in *S. solfataricus*, secondary structures have been shown to play a role in the mesophilic *Haloferax volcanii* (Dar et al., 2016; Berkemer et al., 2020a). Mechanistically, the poly(U) sequence leads to pausing of the RNAP caused by a weakened DNA-RNA hybrid, which leads to an invasion of the growing RNA strand into the DNA cleft of the TEC and finally to a dissociation of the RNAP (Sugimoto et al., 1995; Epshtein et al., 2007). While in most *in vitro* assays and also in *H. volcanii* short poly(U)$_5$ sequences are sufficient, *Methanosarcina mazei* and *Saccharolobus solfataricus* both predominantly have poly(U) stretches that are up to 20 nt long (Santangelo et al., 2009; Dar et al., 2016; Berkemer et al., 2020a). The necessity of long poly(U)s for efficient transcription termination is not yet understood and seems unintuitive, as the RNA-DNA hybrid is only up to 9 nt long. It will eventually turn out if this feature helps in slowing down actively transcribing TECs or facilitates interactions with additional termination factors.

Rho mediates factor-dependent transcription termination in Bacteria (Peters et al., 2011). This factor binds to C-rich, but G-poor *rut* (Rho-utilization) sequences on the nascent mRNA and translocates in 5´to 3´direction. When Rho reaches the RNAP, that is enriched at a pausing site up to 100 bases downstream of the *rut* sequence, it can unwind the RNA-DNA duplex and release the nascent transcript (Banerjee et al., 2006; Peters et al., 2011; Mitra et al., 2017). Interestingly, Rho has been shown to effectively terminate transcription of the archaeal transcription machinery *in vitro*, which emphasises the sensitivity of Archaea towards factor-dependent termination (Santangelo and Reeve, 2006).

The presence of an archaeal termination factor remained an enigma for many years until recently two proteins have been described that terminate transcription in *T. kodakarensis* (Walker et al., 2017; Sanders et al., 2020):

Eta (euryarchaeal termination activity) was the first factor that was shown to disrupt the TEC in an ATP-dependent way, by pushing the RNAP forward, which results in the release of the nascent RNA. However, there are multiple indications that the non-essential factor is not involved in general transcription termination but more likely stimulates the release of stalled or arrested TECs (Chamieh et al., 2016; Walker et al., 2017) (Figure 7). In comparison, the archaeal ribonuclease aCPSF1, designated as FttA (Factor that terminates transcription in Archaea) in *Thermococcus kodakarensis*, is a *bona fide* archaeal transcription termination factor and universally conserved in all archaeal genomes (Sanders et al., 2020; Yue et al., 2020). FttA can bind to transcripts of actively elongating TECs and mediate cleavage in a Rho-like manner, although both proteins are unrelated (Figure 7). Similar to the stimulating role of NusG for Rho-dependent termination in bacteria, FttA activity is coupled to Spt4/5 and additionally enhanced by transient interactions with the stalk domain. In summary, FttA is able the bind TECs, cleave and release RNA and use its 5´to 3´exonuclease activity to further degrade the nascent RNA from the 3´end (Sanders et al., 2020). Interestingly, depletion of *Mmp-aCPSF1* in *Methanococcus maripaludis* caused a genome-wide transcription read-through. This also applies to uridine-rich terminators, where the endoribonuclease activity of aCPSF1 is essential for termination *in vivo* (Yue et al., 2020).

Given the wide distribution of poly(U) signatures, the ubiquitous distribution of aCPSF1 among archaeal phyla and the genome-wide termination defects of *Mmp-aCPSF1* depleted cells, it is tempting to speculate about a critical interplay of factor-dependent and intrinsic termination mechanisms. Ultimately, more biochemical and omics data from diverse phyla will be needed to get a clearer view of transcription termination in Archaea.

3.2. Specific regulation by bacterial-like TFs

3.2.1. General characteristics

All organisms rely on the tight regulation of the transcription machinery, which is achieved on the basal level by the interplay between general transcription factors, elements on the DNA and the nascent RNA (Chen et al., 2018; Blombach et al., 2019; Mejía-Almonte et al., 2020) (Figure 7). However, this is not sufficient to rapidly adapt to changing environmental or metabolic conditions. Hence, an extensive amount of fine-tuning is necessary, which in Bacteria and Archaea (amongst other mechanisms described in the discussion) can be achieved by gene-specific transcription factors (TFs) (Browning and Busby, 2016; Karr et al., 2017). Binding of the TFs to DNA sequences neighbouring

promoter elements increases or represses the transcription rate mainly by two scenarios: Activators bind upstream of the promoter and stabilise the PIC formation, by stimulating the recruitment of TBP and TFB. In contrast, repressors overlap or bind downstream of the promoter and impede PIC formation by sterically blocking the binding of general transcription factors or the recruitment of the RNAP (Peeters et al., 2013) (Figure 8).

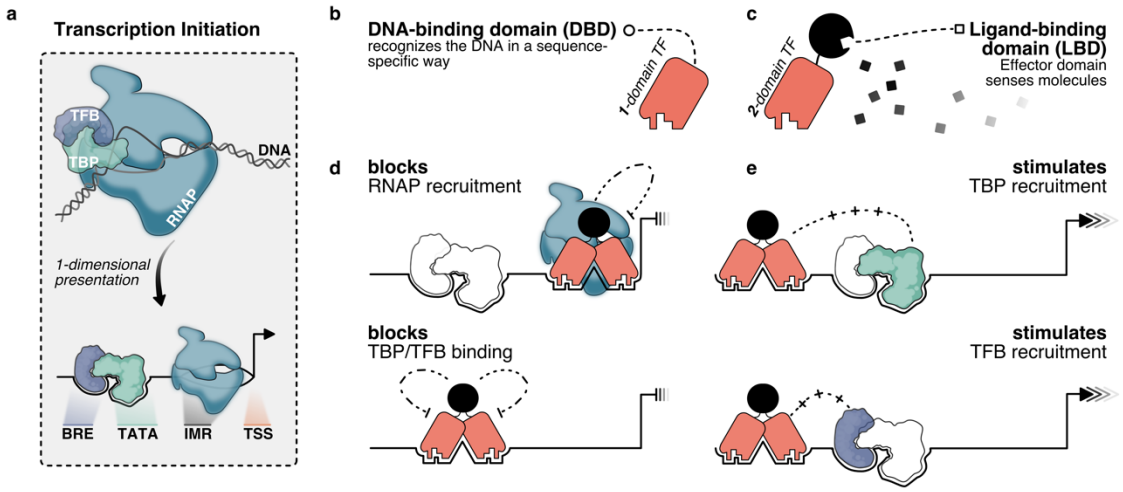


Figure 8 | Basal mechanisms of transcription factor mediated regulation in Archaea. a, One-dimensional representation of sequence elements (BRE, TATA, IMR, TSS) and proteins (TFB, purple; TBP, green; RNAP, blue) that are important during transcription initiation. **b**, Transcription factors (TFs) can recognise sequence motifs with the help of a DNA-binding domain (DBD. **c**, While one-domain TFs only have a DBD, two-domain TFs additionally have a ligand-binding domain (LBD), which sense intracellular signals. **d**, Transcription repression can be achieved by TFs that sterically block TBP, TFB or RNAP. **e**, In contrast, activators stimulate transcription by increasing the rate of TBP or TFB recruitment (reviewed in Karr et al., 2017, Peeters et al. 2013).

While these models of regulation appear very simple, other more complicated strategies have been discovered, that involve DNA bending, structural rearrangements, dual activities and histone-like functions, that dramatically expand TF functionality, flexibility and complexity (Karr et al., 2017). To modulate their binding behaviour, TFs can detect effector molecules, which is achieved by two fundamentally different mechanisms. Membrane-bound kinases, that are part of a two-component system (TCS), sense extracellular signals and transmit the signal via phosphorylation to a response regulator (Stock et al., 2000; Wuichet et al., 2010). One-component systems (OCS) are far more common in Archaea. They consist of a single regulator that has two domains, a DNA-binding domain (DBD) and a ligand-binding domain (LBD), which senses intracellular signals (Figure 8b,c). The DBD binds to the major groove of the DNA and is therefore limited in the possible structural folds it can adapt. More than 80% of all archaeal TFs have a helix-turn-helix motif (HTH), that interacts with the DNA by its recognition helix (Aravind and Koonin, 1999; Kyrpides and Ouzounis, 1999; Charoensawan et al., 2010; Lemmens et al., 2019). In two-domain TFs, specificity for diverse signals is achieved by

the structural flexibility of the LBDs, which is also used to classify TFs into families (Perez-Rueda et al., 2018).

3.2.2. Evolution & TF families

Despite the evolutionary distance between Archaea and Bacteria, they have a similar repertoire of TF families (Lemmens et al., 2019). This striking resemblance either suggests that the last universal common ancestor (LUCA) already contained many of the TFs or points to frequent horizontal gene transfer events throughout evolution (Aravind and Koonin, 1999; Ashby, 2006; Nelson-Sathi et al., 2015). Interestingly, Eukarya are equipped with many novel classes of TFs, assuming that transition to multicellularity and transcription in a chromatin landscape most likely represent the significant driving forces for development (Lambert et al., 2018; de Mendoza and Sebé-Pedrós, 2019). In comparison to Bacteria, the relative number of TFs per genome size (5% vs 8-10%) and also the average protein size of TFs (179 vs 236 amino acids) is smaller in Archaea. However, the total number of TFs correlates significantly with the genome sizes in all prokaryotes. This indicates that organisms living in complex environmental niches, in general, have higher coding potential, hence bigger genomes and necessitate diverse options for regulation (Perez-Rueda and Janga, 2010). Accordingly, Crenarchaeota have smaller genomes and less TFs, as they occupy less-challenging habitats as Euryarchaeota. Although they are phylogenetically quite distant, both share a similar set of TF families, supporting the theory that many of the TF families were already present early in evolution (Figure 9a). The higher total number of TFs in Archaea with bigger genomes is therefore not caused by diversification of TF families but reflects frequent gene duplications events (Plaisier et al., 2014; Denis et al., 2018; Perez-Rueda et al., 2018).

Dissecting the distribution of TF families across different superphyla, the Lrp/AsnC, MarR, TrmB, SinR and ArsR families are among the most common (Perez-Rueda et al., 2018; Lemmens et al., 2019). The repertoire of TFs in *Pyrococcus furiosus* highlights the diversity of domain architectures, that can occur in one organism: The majority of the 86 TFs is not composed of the two-domain DBD-LBD architecture, but consists of one (45%) or three (14%) structural domains (Denis et al., 2018). Interestingly, the family distribution resembles more the one from the TACK superphylum than from the Euryarchaeota, which could point to a common evolutionary development sharing extreme, but less complex environments (Figure 9b).

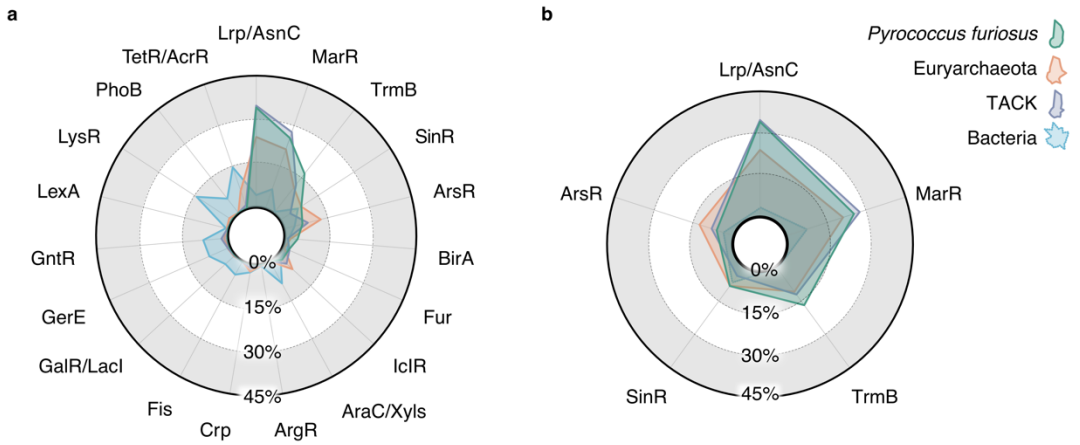


Figure 9 | Distribution of transcription factor families in prokaryotes. a, Relative distribution of transcription factor families that are shared in the repertoire of Bacteria and Archaea. TFs and associated families have been retrieved from Perez-Rueda et al., 2018, and compared to the total number of TFs. **b**, Comparison of the five most abundant TF families in *P. furiosus* (green) to the distribution in Euryarchaeota (orange), the TACK phylum (purple) and Bacteria (blue).

3.2.3. Regulatory Mechanisms

Although protein folds and domain architectures provide useful family information, they do not allow a *de novo* prediction regarding the type of regulation or the biological process and the TF regulon. In general, the mode of action is specified by the location of the TF-binding site, which is upstream of the archaeal promoter for activators and downstream or overlapping for repressors (Karr et al., 2017).

Mechanistically, stimulation of transcription is achieved by binding of a TF adjacent to the BRE element and subsequent recruitment of TBP or TFB by direct protein-protein interactions (Peeters et al., 2013). A higher number of basal transcription factors leads to an increased formation of stable PICs at a time and ultimately to a higher transcription rate, especially at weak promoters, that differ from the consensus sequence. Stimulation by TBP recruitment is achieved by Ptr2 from *Methanocaldococcus jannaschii*, which belongs to the Lrp family of transcriptional regulators and binds to a palindromic site via its N-terminal HTH (Ouhammouch et al., 2003, 2005). This recognition strategy is the most common among archaeal TFs and correlates with dimerisation of the proteins and a two-fold symmetry. After binding, the C-terminal Ptr2 effector domain directly interacts and recruits TBP, hence accelerating the rate-limiting step at Ptr2-dependent promoters (Pritchett et al., 2009). Complementary to transcription stimulation by TBP recruitment to weak TATA boxes, TFB can be recruited to weak BRE elements. Currently, this mechanism is only described for TFB-RF1 from *P. furiosus* and the *araS* promoter from *Sulfolobus solfataricus* (Peng et al., 2009; Ochs et al., 2012; Reichelt et al., 2018b).

In contrast to the activation mechanisms, repression can be achieved by either blocking the basal transcription factors or the RNAP through steric hindrance. While Lrs14 binding sites from *S. solfataricus* overlap with the basal elements of its promoter, MDR1

from *Archaeoglobus fulgidus* or Phr from *P. furiosus* bind further downstream and prevent stable recruitment of the RNAP (Bell et al., 1999a; Napoli et al., 1999; Bell and Jackson, 2000; Vierke et al., 2003). These differences may be attributed to temporal aspects of de-repression. In contrast to a heat-shock (Phr) or metal-deficiency response (Mdr1), the negative autoregulation mechanism of Lrs14 does not rely on a fast regulation.

In addition to mechanistically straightforward ways for positive or negative regulation that apply to many transcription factors and regulatory networks, other far more complicated types of regulation co-exist in Archaea: For instance, some TFs are not limited to a single function but act as dual regulators either at different promoters or at the same promoter after structural reorientation. Activation or repression by the global sugar regulator TrmBL1 from *P. furiosus* depends on the location of the TGM (Thermococcales glycolytic motif) downstream or upstream of the promoter (Lee et al., 2007; Reichelt et al., 2016) (Figure 10a). The sulfur-response regulator SurR also uses this dual mechanism in *P. furiosus* (Lipscomb et al., 2009). The TF belongs to the Trmb family and is only present in the Thermococcales family (Schut et al., 2013; Kim et al., 2016). It targets both up- and downregulated genes, as a primary response to S^0 (Lipscomb et al., 2017). Interestingly, SurR contains a CxxC motif adjacent to its HTH, that acts as an S^0-inducable redox switch (Yang et al., 2010). Oxidation of the cysteine residues leads to a conformational change, therefore a diminished DNA binding affinity and explains the crucial role in the metabolic switch from H_2 to H_2S (Figure 10b).

Furthermore, some factors modulate their control function in a concentration-dependent manner. LrpB from *S. solfataricus* has three target boxes on its promoter that are filled up with increasing TF concentrations (Peeters et al., 2004, 2006) (Figure 10c). In this context, autorepression is achieved by a conformational change after all three boxes are occupied, while in the ground state, that is limited to two bound boxes, *lrpB* is auto-induced (Peeters et al., 2013).

In contrast to gene-specific TFs that are only moderately expressed, the Thermococcales architectural protein TrmBL2 is highly abundant in the cell and plays a role in genome organisation, next to its function as a global repressor for 6.5% of the genes in *T. kodakarensis* (Maruyama et al., 2011; Efremov et al., 2015). The one-domain protein that lacks a C-terminal LBD has been shown to form a thick filament by covering dsDNA, additionally stabilises DNA secondary structures and may help to maintain chromosomal DNA stability at high temperatures (Efremov et al., 2015; Wierer et al., 2016) (Figure 10d).

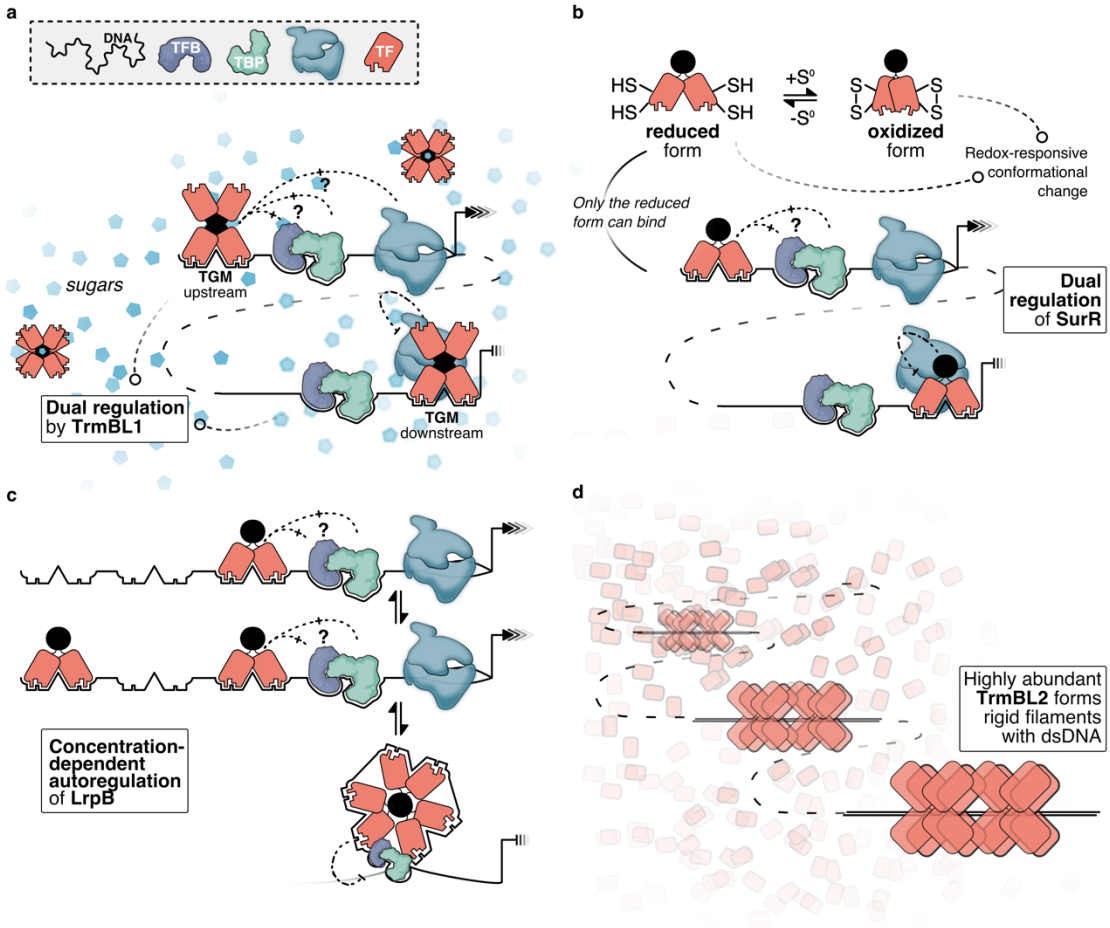


Figure 10 | Complex TF-mediated regulatory mechanisms. a, Regulation by the global sugar regulator TrmBL1 depends on the location of the TGM-recognition motif upstream (activation) or downstream (repression) of the promoter (Lee et al., 2007; Reichelt et al., 2016). **b**, SurR is a sulfur-response regulator that can only bind to DNA in its reduced form. Activation of repression of certain genes is achieved by a similar dual regulation mechanisms as described for TrmBL1 (Lipscomb et al., 2009; Yang et al., 2010). **c**, LrpB has multiple recognition boxes on the DNA and regulates transcription in a concentration-dependent way (Peeters et al., 2004, 2006, 2013). **d**, In contrast to gene-specific TFs, TrmBL2 is highly abundant in the cells, forms rigid filaments with DNA and has been shown to play a role in chromosomal organisation (Maruyama et al., 2011; Efremov et al., 2015; Wierer et al., 2016).

In conclusion, archaeal TFs offer diverse possibilities to modulate transcription that help the organisms to react to changing environmental conditions rapidly. It is astonishing that although species of both prokaryotic domains co-exist in many ecological niches, archaeal genomes harbour a substantially lower number of TFs and have a smaller family repertoire compared to Bacteria. Nevertheless, as this still seems to be sufficient for efficient gene regulation, it is estimated that additional regulation layers exist waiting to be discovered (Lemmens et al., 2019). Currently, the analysis of multi-layer principles is primarily limited by the lack of system-level data, like genome-wide TF profiling or gene expression studies of TF knockouts. Combined with detailed biochemical characterisation,

these studies ultimately would allow for interrogating global regulatory networks in Archaea (Martinez-Pastor et al., 2017).

3.3. Genome-wide profiling of transcription

For many years, it was only possible to study the mechanisms of archaeal transcription by *in vitro* approaches. This not only applies to basal transcription but also the fine-tuning process mediated by gene-specific TFs. Although the analysis in most of the cases was limited to single loci - like tRNAs, rRNAs or highly expressed mRNAs - valuable insights concerning their function were made, thanks to the biochemical tractability of the systems. Nevertheless, inferring general statements on aspects like TF occupancy at different promoters or the rules behind sequence-specific recruitment, was nearly impossible.

Meanwhile, the study of microbial genomes has been revolutionised by assays employing second and third-generation sequencing technologies. Not only the genomics field profited tremendously by the possibility to perform cost-efficient deep metagenomic sequencing, but it also led to a variety of applications in the transcriptomics field (Tringe and Rubin, 2005; Siezen et al., 2010; Sorek and Cossart, 2010; Stark et al., 2019).

Transcript-focused profiling started more than ten years ago with the development of microarrays, followed by modified library preparation protocols that indirectly allowed sequencing of the RNA using NGS technologies (Stark et al., 2019). Since then, these methods have been used to explore the transcriptome architecture of archaeal model organisms, but mostly to study differential gene expression under environmentally relevant conditions (Sorek and Cossart, 2010). In this context, the genome-wide study of transcription start sites (TSS) by differential RNA-seq, and the elucidation of termination sites by Term-seq have helped to improve the standard genome annotation and revealed new insights into 5´- or 3´-guided post-transcriptional regulation, and the role of antisense (asRNAs) and other noncoding RNAs (ncRNAs) (Sharma et al., 2010; Sharma and Vogel, 2014; Bischler et al., 2015; Dar et al., 2016; Berkemer et al., 2020a).

A major improvement for the *in vivo* study of TFs in a global context was the development of a technique called ChIP-seq (Chromatin Immunoprecipitation followed by next-generation sequencing) (Johnson et al., 2007; Wilbanks et al., 2012). The protocol includes crosslinking of the TF to the DNA, DNA-shearing and enrichment of the TF-DNA complex by immunoprecipitation (Furey, 2012). Following this approach, the occupancies of the basal transcription machinery of *M. jannaschii* and many gene-specific archaeal TFs have been mapped (Reichelt et al., 2016; Karr et al., 2017; Smollett et al., 2017b).

Genome-wide studies have become a powerful tool to study transcription on multiple levels. For a long time, RNA molecules have only been considered as a simple intermediate between DNA and proteins that can be used as a counter for the quantification of gene expression. Although this remains one of the major applications, technological advances

have led to a paradigm shift and identified RNA as an astonishingly complex molecule with high regulatory potential (Stark et al., 2019). Considering that archaeal research is generally lagging behind bacterial and eukaryotic systems, more extensive integration of multi-omics data and dedicated *in vitro* approaches will help to disentangle previously unknown regulatory principles and ultimately allow to draw evolutionary conclusions.

4. Scope of this thesis

The scope of this thesis was to investigate features of basal and regulated transcription in the hyperthermophilic model organism *Pyrococcus furiosus* and to develop wet lab and bioinformatical protocols for native RNA sequencing using Nanopore technology in prokaryotes, which can be used for the analysis of general transcriptomic and post-transcriptional features.

To that end, we first aimed to set a framework for future global studies by investigating the genome and transcriptome architecture of *P. furiosus*. After re-sequencing the genome using a state-of-the-art hybrid Illumina/PacBio approach, we updated and expanded the annotation with additional features derived from a differential RNA sequencing approach. Additionally, to test the genomic stability that is challenged by multiple IS elements, we established third generation Nanopore sequencing and could confirm the integrity of a two-year-old lab strain.

Secondly, we explored prokaryotic transcriptomes by setting up native RNA sequencing, which allowed us to map multiple transcriptomic and post-transcriptional features simultaneously. Taking advantage of the single-molecule technique, we traced back and expanded the multi-step rRNA maturation process in Archaea and correlated intermediates with selected rRNA modifications.

In the third part of the dissertation, we elucidated the essential role of the transcription factor CopR in copper detoxification performing integrative RNA-seq, and ChIP-seq complemented with in-depth biochemical and structural analysis.

CHAPTER II
Publications

The "Publications" part of this dissertation comprises three articles. One manuscript has already been published and the other two are available as preprints on bioRxiv and have been submitted to peer-reviewed journals. The PhD candidate, Felix Grünberger, has authored every manuscript herein as a first author. In the following part, published or submitted text versions (bioRxiv version) of the manuscripts are printed, including main figures and tables. Supplementary tables and figures are attached in the Appendix of the dissertation.

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

Felix Grünberger[1], Robert Reichelt[1], Boyke Bunk[2], Cathrin Spröer[2], Jörg Overmann[2,3], Reinhard Rachel[1], Dina Grohmann[1], Winfried Hausner[1] *

[1]Institute of Microbiology and Archaea Center, University of Regensburg, Regensburg, Germany
[2]Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures, 38124 Braunschweig, Germany
[3]Microbiology, Braunschweig University of Technology, Braunschweig, Germany

*This paper is dedicated to the memory of our colleague and friend, Prof. Dr. Reinhard Wirth, who recently passed away. Reinhard has identified the LS and the BBR strain of *Pyrococcus furiosus* which exhibit different amounts of flagella and unusual cell morphology.

Correspondence: Winfried Hausner. E-mail: Winfried.Hausner@ur.de

Keywords: archaea, *Pyrococcus*, RNA sequencing, Nanopore sequencing, PacBio sequencing, bidirectional transcription, antisense transcription

Publication information:
Frontiers in Microbiology (2019)
Received: 22 March 2019; Accepted: 26 June 2019, Published: 12 July 2019
Link: https://doi.org/10.3389/fmicb.2019.01603

1. Abstract

Pyrococcus furiosus DSM 3638 is a model organism for hyperthermophilic archaea with an optimal growth temperature near 100 °C. The genome was sequenced about 18 years ago. However, some publications suggest that in contrast to other *Pyrococcus* species, the genome of *P. furiosus* DSM 3638 is prone to genomic rearrangements. Therefore, we re-sequenced the genome using third generation sequencing techniques. The new *de novo* assembled genome is 1,889,914 bp in size and exhibits high sequence identity to the published sequence. However, two major deviations were detected: (1) The genome is 18,342 bp smaller than the NCBI reference genome due to a recently described deletion. (2) The region between PF0349 and PF0388 is inverted most likely due an assembly problem for the original sequence. In addition, numerous minor variations, ranging from single nucleotide exchanges, deletions or insertions were identified. The total number of insertion sequence (IS) elements is also reduced from 30 to 24 in the new sequence. Re-sequencing of a two-year-old "lab culture" using Nanopore sequencing confirmed the overall stability of the *P. furiosus* DSM 3638 genome even under normal lab conditions without taking any special care.

To improve genome annotation, the updated DNA sequence was combined with an RNA sequencing approach. Here, RNAs from eight different growth conditions were pooled to increase the number of detected transcripts. Furthermore, a differential RNA-Seq approach was employed for the identification of transcription start sites (TSSs). In total, 2515 TSSs were detected and classified into 834 primary (pTSS), 797 antisense (aTSS), 739 internal and 145 secondary TSSs. Our analysis of the upstream regions revealed a well conserved archaeal promoter structure. Interrogation of the distances between pTSSs and aTSSs revealed a significant number of antisense transcripts, which are a result of bidirectional transcription from the same TATA box. This mechanism of antisense transcript production could be further confirmed by *in vitro* transcription experiments. We assume that bidirectional transcription gives rise to non-functional antisense RNAs and that this is a widespread phenomenon in archaea due to the architecture of the TATA element and the symmetric structure of the TATA-binding protein.

2. Introduction

Pyrococcus furiosus was isolated from geothermally heated marine sediments taken from the beach of Porto di Levante, Vulcano Island, Italy (Fiala and Stetter, 1986). It is a strictly anaerobic heterotroph, growing on maltose, starch, pyruvate, peptone and complex organic substrates. When carbohydrates are used as energy source, acetate, carbon dioxide and hydrogen are the major fermentation products (Schäfer and Schönheit, 1992; Kengen et al., 1994). In the presence of peptides, elemental sulfur is required for efficient growth and hydrogen sulfide is generated as end product. With a doubling time of only 37 minutes at the optimal growth temperature of 100 °C, *P. furiosus* has developed to one of the best studied hyperthermophilic organisms. The first published genome sequence of *P. furiosus* DSM 3638 revealed a GC content of 40.8 % and a genome size of 1.91 Mb encoding 2,225 genes and 2,122 proteins (Robb et al., 2001).

However, some publications suggest that the *P. furiosus* genome is susceptible to genomic rearrangements in comparison to the related *Pyrococcus* species *P. abyssi* and *P. horikoshii* (DiRuggiero et al., 2000; Brügger et al., 2002; Zivanovic et al., 2002). Genome comparison suggests that transposition events, most frequently induced by insertion sequence (IS) elements, are the major driving force for such genome variations in the *P. furiosus* genome. No full-length IS elements were identified in the genomes of the other two *Pyrococcus* species (Zivanovic et al., 2002). IS elements are short DNA sequences with a typical length between 700 to 2500 bp (Siguier et al., 2014). They contain an open reading frame (ORF) encoding a transposase, which is usually flanked by inverted repeats and promote translocation of DNA segments within and between genomes. A study analyzing the IS elements of a collection of *Pyrococcus* strains isolated from the original habitat, Vulcano Island, suggested that these elements play an important role for genetic drift in the diversification of a geographically isolated population of *P. furiosus* (Escobar-Páramo et al., 2005). Furthermore, the identification of an almost identical 16 kb region transposable region between *P. furiosus* and *Thermococcus litoralis* with only 153 nucleotide differences, indicates that these mobile elements are also involved in horizontal gene transfer (DiRuggiero et al., 2000). This region belongs to one of six highly variable chromosomal regions, which were previously identified by comparative genome hybridization using *P. furiosus* and seven *Pyrococcus* isolates from Vulcano Island (White et al., 2008). The 16 kb transposable region harbors genes encoding an ABC transport system for maltose and trehalose and is only present in *P. furiosus*, but absent from all other Vulcano isolates. This is also true for *P. woesei*, which was isolated one year later from the same habitat (Zillig et al., 1987). The physiology of *P. woesei* seems to be very similar to that of *P. furiosus* and the rRNA operons of these strains have identical sequences (Kanoksilapatham et al., 2004). Although the complete genome sequence of *P. woesei* is not available, hybridization of genomic sequence from *P. woesei* to a DNA

microarray of *P. furiosus* revealed the presence of additional genes in two clusters in *P. furiosus* (Hamilton-Brehm et al., 2005). One of these clusters is the 16 kb transposable region involved in the maltose metabolism. It is interesting to note that a ChIP-Seq approach from our group recently revealed the deletion of this 16 kb fragment also in *P. furiosus* and an additional southern blot analysis with a new strain from the German Collection of Microorganisms and Cell Cultures (DSMZ, Braunschweig, Germany) confirmed a rapid deletion of the fragment even with growth on maltose (Reichelt et al., 2016). Altogether, these findings support the previous suggestion to rename *P. woesei* as *P. furiosus* subsp. *woesei* (Kanoksilapatham et al., 2004).

Genome variability of *P. furiosus* strains can be observed in the natural habitat but also in strains cultivated in the laboratory. Several years of re-cultivation from stocks stored at 4°C resulted in the emergence of at least two additional *P. furiosus* strains, LS and BBR, in the Archaea Center at the University of Regensburg (Näther-Schindler et al., 2014). Both strains differ in comparison to the deposited type strain *P. furiosus* DSM 3638 in the degree of flagellation and cell morphology (Daum et al., 2017). A similar observation concerning the occurrence of a lab strain was made in Michael Adams group (University of Georgia, Athens). This lab strain exhibits an extended exponential growth phase and atypical cell aggregation behavior (Lewis et al., 2015). The genome sequence of the mutant showed 145 genes with one or more insertions, deletions or substitutions. The data clearly demonstrate that *P. furiosus* has most likely due to the presence of IS elements an inherent property for efficient genome rearrangements. This facilitates the selection of special mutants under a proper selection pressure, but also stimulates the nonspecific accumulation of different mutations within the genome. The best example for this process is the development of the genetically tractable *P. furiosus* strain COM1 (Bridger et al., 2012). The genome sequence of this strain is 1,571 bp longer than the type strain and contains numerous chromosomal rearrangements, deletions, and single base changes, which lead to the inactivation of 20 genes and to amino acid sequence variations of another 102 gene products. These changes impact various cellular functions including a riboflavin requirement for growth. The alignment of the COM1 genome sequence with the published *P. furiosus* genome revealed major inversions, but an additional analysis of the chromosomal orientations of the original DSMZ strain (ordered in October 2010) by PCR showed that some of this major inversions are also present in the original DSMZ strain (Bridger et al., 2012).

The *P. furiosus* genome sequence was published 18 years ago, but sequencing technologies and bioinformatic analysis have been revolutionized during the last ten years. The introduction of massively parallel sequencing led to a significant reduction of sequencing costs. However, these so-called second-generation sequencing techniques produce only short reads, which impedes the assembly of the complete genome as repetitive regions cannot be resolved (Verma et al., 2017). Meanwhile, third-generation sequencing

techniques have also entered the market. These systems act directly on the native DNA without the requirement for PCR amplification and show a significant increase in read length, which facilitates complete genome assemblies.

The discrepancy between the published *P. furiosus* genome sequence (Robb et al., 2001) and the detected deviations in the genome from a recently ordered *P. furiosus* strain DSM 3638 from the DSMZ (Bridger et al., 2012; Reichelt et al., 2016) and the fact that many groups make use of the originally described *P. furiosus* strain, encouraged us to re-sequence the type strain *P. furiosus*. To address the problem concerning the observed genome rearrangements and to allow for complete genome assembly we used a hybrid approach of third generation long-read PacBio sequencing complemented with highly accurate short-read Illumina sequencing (Rhoads and Au, 2015). We amended the DNA-sequencing approach with differential RNA sequencing data, to generate a high-resolution annotation using the ANNOgesic pipeline (Sharma et al., 2010; Yu et al., 2018). Last but not least, to gain insight into the genomic variability of continuously cultivated lab strains and to investigate, if it is possible to maintain genome stability by avoiding strong selection pressure during cultivation, we re-sequenced *P. furiosus* again after two years of cultivation employing the recently developed Nanopore sequencing technology (Loman et al., 2015). Our results indicate a quite stable genome even with the strain cultivated for two years in the lab and differential RNA-Seq data revealed that bidirectional transcription is a significant source for archaeal antisense transcripts.

3. Material and Methods

3.1. Strains and growth conditions

P. furiosus strain DSM 3638 was stored as freeze-dried culture at 12°C in the dark at the *Deutsche Sammlung von Mikroorganismen und Zellkulturen* (DSMZ) in Braunschweig, Germany. For the isolation of DNA for combined PacBio and Illumina sequencing, cells were grown anaerobically in 20 ml SME medium supplemented with 0.1 % yeast extract and 0.1 % starch at 95 °C to late-exponential phase (Fiala and Stetter, 1986).

For Nanopore sequencing, a culture was obtained from the DSMZ in 2015 and after growth in SME media the strain was stored in the gas phase of liquid nitrogen at the archaea center in Regensburg for one year. After that, the strain was recultivated and handled in the lab for about two years with numerous inoculations, to simulate storage and daily-life usage conditions of many labs with a focus on microbiology. We assume that during these two years the culture was about -roughly estimated- thirty times transferred into fresh medium. In between, liquid cultures were stored at room temperature or 4 °C. For the isolation of DNA cells were grown anaerobically in 40 ml SME medium supplemented with 40 mM pyruvate, 0.1 % peptone and 0.1 % yeast extract at 85 °C to mid-exponential phase.

For RNA sequencing *P. furiosus* cells were grown under eight different conditions to maximize the number of different transcripts in the genome: Cells were grown anaerobically in 20 ml SME medium supplemented with 0.1 % starch, 0.1 % peptone and 0.1 % yeast extract at 95 °C to early- ($1x10^7$ cells/ml), mid-exponential ($5x10^7$ cells/ml) or late-exponential ($1x10^8$ cells/ml) phase (conditions 1, 2 and 3). In addition, cells were grown in 20 ml SME medium supplemented with 0.1 % starch, 0.1 % peptone and 0.1 % yeast extract at 95 °C to late-exponential phase, further incubated at 4 °C for 1 h (condition 4; cold shock) or 110 °C for 15 min (condition 5; heat shock). Moreover, cells were grown anaerobically in 20 ml SME medium supplemented with 0.1 % starch, 0.1 % peptone and 0.1 % yeast extract at 75 °C to late-exponential phase (condition 6; cold adaption). Furthermore, cells were grown anaerobically in 20 ml SME medium supplemented with 0.1 % yeast extract and 0.1 % starch (condition 7; glycolytic growth) or 40 mM pyruvate (condition 8; gluconeogenic growth) at 95 °C to late-exponential phase. Cells were harvested by centrifugation at 3,939 x g for 45 min at 4 °C, cell pellets were frozen in liquid nitrogen and stored at – 80 °C until used for the isolation of DNA or RNA.

3.2. DNA Isolation

Genomic DNA was isolated using ReliaPrep™ gDNA Tissue Miniprep System (Promega) according to the instructions of the manufacturer. Quantity and quality were analyzed using Nanodrop One, Qubit dsDNA HS assay kit (both from Thermo Fisher

Scientific) and agarose gel electrophoresis. For Nanopore sequencing size distribution was checked using pulsed field gel electrophoresis on a CHEF-DR®III system (Bio-rad).

3.3. RNA Extraction

P. furiosus total RNA was purified using the peqGOLD TriFast™ reagent (VWR). 20 ml cell culture was pelleted, and cells were lysed by addition of 1 ml TriFast followed by rigorous shaking for 15 min. After adding 0.2 ml 2 M sodium acetate pH 4.0 RNA was isolated according to the manufacturer instructions. Contaminating DNA was removed via the TURBO DNA-free™ Kit (Thermo Fisher Scientific, Waltham, MA, United States). The integrity of the total RNA was assessed via a Bioanalyzer (Agilent) run using the RNA 6000 Pico Kit (Agilent) and purified RNA was stored at −80.

3.4. PacBio Library Preparation and Sequencing

SMRTbell™ template library was prepared according to the instructions from Pacific Biosciences, Menlo Park, CA, United States, following the Procedure and Checklist – Greater Than 10 kb Template Preparation. Briefly, for preparation of 15 kb libraries 8 µg genomic DNA was sheared using g-tubes™ from Covaris, Woburn, MA, United States according to the manufacturer's instructions. DNA was end-repaired and ligated overnight to hairpin adapters applying components from the DNA/Polymerase Binding Kit P6 from Pacific Biosciences, Menlo Park, CA, United States. Reactions were carried out according to the manufacturer's instructions. BluePippin™ Size-Selection to greater than 7 kb was performed according to the manufacturer's instructions (Sage Science, Beverly, MA, United States). Conditions for annealing of sequencing primers and binding of polymerase to purified SMRTbell™ template were assessed with the Calculator in RS Remote, Pacific Biosciences, Menlo Park, CA, United States. SMRT sequencing was carried out on the PacBio *RSII* (Pacific Biosciences, Menlo Park, CA, United States) taking one 240-min movie on two SMRT Cells.

3.5. Nanopore sequencing (MinION)

3.5.1. Genome assembly, error correction, and annotation

SMRT Cell data was assembled using the "RS_HGAP_Assembly.3" protocol included in SMRT Portal version 2.3.0 using default parameters. The assembly revealed a single circular chromosome. The chromosome was circularized, particularly artificial redundancies at the ends of the assembled contig were removed and adjusted to *cdc6* as the first gene. Error-correction was performed by a mapping of paired-end reads of 2x100 bp generated on an Illumina HiSeq 2500 onto finished genomes using BWA (Li and Durbin, 2010) with subsequent variant and consensus calling using VarScan (Koboldt et al., 2012). A consensus concordance of QV60 could be confirmed. Finally, an annotation was carried

out using NCBI prokaryotic genome annotation pipeline (Tatusova et al., 2016). The genome sequence was deposited in NCBI GenBank under Accession Number CP023154.

3.5.2. Basecalling, *de novo* assembly, polishing and evaluation

For MinION data analysis raw reads in fast5 data format were base-called and de-multiplexed using Albacore 2.3.4. In a first step a *de novo* genome assembly was done using canu 1.8 (genomeSize=1.9m, minReadLength=500, minOverlapLength=100) (Koren et al., 2016), before improving the consensus sequence in a second step with *nanopolish* 0.11 (min-candidate-frequency=0.1) (Simpson et al., 2017). The chromosome was circularized, artificial redundancies at the ends of the assembled contig removed and adjusted to *cdc6* as the first gene (compare 2.5). To determine the identity of the *de novo* assembly to the hybrid PacBio-Illumina approach, statistics from dnadiff (MUMmer version 3) were calculated and visualized using R package *genoPlotR* (Kurtz et al., 2004; Guy et al., 2011). Read length and nucleotide frequencies were analyzed using the statistical program R with ggplot2 (R Development Core Team, 2011; Wickham, 2016). Code is available at: https://github.com/felixgrunberger/pyrococcus_reannotation.

3.6. Illumina sequencing (RNA-Seq)

RNAs purified from cells grown under eight different growth conditions were pooled equally and submitted for library preparation and sequencing to the Core Unit Systems Medicine (SysMed) at the University Würzburg, Germany. Three different libraries were prepared to fulfill the requirements for usage in the ANNOgesic pipeline: fragmented, unfragmented with terminator exonuclease treatment (+TEX) and unfragmented without TEX-treatment (-TEX). For the fragmented sample, RNA was fragmented for 2 min at 94 °C using the NEBNext Magnesium RNA Fragmentation Module. Afterwards RNA was treated with T4 Polynucleotide Kinase (PNK) without ATP for 6 h at 37 °C and 1 h at 37 °C with 2 mM ATP and fresh T4 PNK. After overnight ethanol precipitation, 5' triphosphates were removed using RNA 5' Pyrophosphohydrolase (RppH) for 1 h at 37 °C. RNA was again precipitated and resuspended in 6 µl H_2O. The two samples (+/- TEX) for the transcription start site detection were either treated with TEX (+ TEX) or with H_2O as a mock control (- TEX) for 30 min at 37 °C. Afterwards both samples were treated with RppH for 1 h at 37 °C before they were precipitated, and the RNA was resuspended in 6 µl H_2O. After the pre-treatment, all three libraries were prepared using the NEBNext® Multiplex Small RNA Library Prep Kit for Illumina according to the manufacturer's protocol with small modifications. The first linker ligation was performed for 18 h at 16 °C and libraries were amplified with 12 PCR cycles with an extended elongation time of 75 s. Libraries were pooled in a 2:1:1 ratio (fragmented: + TEX : - TEX) and sequenced on an Illumina NextSeq 500 high-output single-end 75 nt run.

3.7. Trimming and mapping of RNA-Seq reads

Illumina reads in FASTQ format were quality/length/adapter trimmed using *trimmomatic* (v.0.36) in single-end-mode allowing for a minimum length of 12 nt and a cut-off Phred score of 20, calculated in a sliding window of 4 bases (Bolger et al., 2014). Next, reads were mapped using STAR aligner (v.2.5.3) to the new assembled genome of *P. furiosus* (Dobin et al., 2013). Mapping statistics (input, filtered, uniquely aligned reads) can be found in the Supplementary Table 3. To use ANNOgesic for RNA-based annotation of *P. furiosus*, strand-specific coverage files in wiggle format of all three sequencing data sets were generated (Yu et al., 2018). As recommended, reads were additionally mapped with segemehl 0.2.0 to detect circular RNAs within the ANNOgesic pipeline (Otto et al., 2014).

3.8. Reference genome annotation using ANNOgesic

ANNOgesic is a recently published pipeline that predicts transcriptome-wide features based on a combination of differential and fragmented RNA sequencing (Yu et al., 2018). Amongst others, it is built on TSSpredator, using adaptive parameter optimization, which simplifies and improves detection of transcription start sites (TSS) (Dugar et al., 2013). Following subcommands were executed in the provided Docker image of ANNOgesic to improve annotation of *P. furiosus* (basic parameters if not stated otherwise): optimize_tss_ps (with 50 manually detected TSS as a reference, 4000 iterations), tss_ps (with optimized parameters from previous step), transcript, terminator, utr, operon, srna, sorf, circrna (cut-off supported reads: 200), promoter, crispr. Features in gff file format were combined with the merge_features command and added to the gff file from DNA sequencing and assembly (Supplementary Table 5).

3.9. RNA-Seq data analysis

Data analysis of output files from ANNOgesic was done using the R/Bioconductor environment (R Development Core Team, 2011). Scripts for analysis were uploaded to https://github.com/felixgrunberger/pyrococcus_reannotation.

3.9.1. Detection of promoter elements

For the detection of common archaeal promoter elements, a position weight matrix (PWM) was calculated from sequences 50 bases upstream to 10 bases downstream of all available TSS. The resulting motif was visualized in R using ggseqlogo (Wagih, 2017).

The sequences 51 bases upstream of every TSS were extracted to identify the best ranking promoter motif for each TSS category using MEME with default options except “search given strand only” (Bailey et al., 2009). Motifs and position tables were further analyzed using ggseqlogo and ggplot2. For internal TSS a repetitive sequence coming from CRISPR regions gave the best motif but was excluded from further analysis.

Length of 5`UTRs of pTSS and sTSS was already calculated in the ANNOgesic pipeline and visualized using ggplot2. Internal and antisense TSS positions relative to a gene were sorted in three windows: 300 bp upstream, 300 bp downstream and in between annotated genes. Positions between start and stop site were scaled according to gene length.

To find a motif for possible bidirectional promoters we filtered all primary TSS that had strong TEX signal on the antisense strand (more than 40% of the reads from -400 to +400 in the region 100 bp upstream of pTSS) and calculated a motif using MEME (default options).

3.9.2. Coverage plots

We generated average coverage profile plots to check for the enrichment of TSS in the TEX data set and to validate the TSSpredator classification. The R package CoverageView was used to calculate the coverage for each sequencing data set from 400 bp upstream and downstream relative to a TSS in a window of 10 bp (Lowy, 2017). Every position for a single TSS was scaled proportionally, before calculating mean values for plotting.

Coverage plots were also generated for the analysis of putative bidirectional promoters. A similar protocol as mentioned above was used. We split the data set into two groups (head-to-head and head-to-tail) considering the orientation of the upstream gene. From this data set we also calculated the intergene distance for annotated genes.

3.9.3. Antisense enrichment around IS elements

IS elements for the available *Pyrococcus* assemblies were identified using ISEScan 1.6 (Xie and Tang, 2017). Genomic positions of these elements were extracted to scan for antisense TSS nearby. The relative position was calculated in a window 100 bp upstream from the start of an IS element, 100 bp downstream from the end and in between scaled according to length. The aTSS and IS elements used for the analysis to create Figure 18b are listed in Supplementary Table 6.

3.10. *In vitro* analysis of bidirectional transcription

For the analysis of bidirectional transcription *in vitro* transcription reactions were assembled in 100 µl reaction volumes (40 mM Na-HEPES, pH 7.3, 0.1 mM EDTA, 0.1 mg/ml BSA, 2.5 mM $MgCl_2$, 250 mM KCl): 8.8 nM DNA was combined with 190 nM TBP, 105 nM TFB, 108 nM TFE and 10.5 nM RNA polymerase. The 349 bp DNA fragment was prepared by PCR with the primers 5´-gaaagggcgaaccagttagattgaacg and 5´-tgttgggcttctccccaagctgag using genomic DNA as template. Transcription was initiated by the addition of NTPs to a final concentration of 100 µM and reactions were incubated at 80 °C for 10 min. Reactions were stopped by extraction with one volume phenol/chloroform/isoamyl alcohol, RNA was precipitated with ethanol and resuspended

in 20 µl H_2O. For the analysis of the transcripts primer extension experiments were carried out using labeled primers in sense or antisense direction. 10 µl *in vitro* RNA were combined with 125 nM of the corresponding primer in a total volume of 15 µl. After RNA denaturing at 70 °C for 5 min, primer annealing was performed at 0 °C for 5 min. Primer extension was started by the addition of 5 µl reverse transcription mixture containing 50 units of M-MLV RNase H minus (Promega) and 1 mM dNTPs. After incubation at 48 °C for 30 min, cDNA products were purified by ethanol precipitation, resuspended in 10 µl formamide buffer and analyzed on a 20% denaturing polyacrylamide gel. The DNA fragments were visualized with a ChemiDoc MP imaging system (Biorad).

4. Results and Discussion

4.1. Strategy for genome re-annotation of *Pyrococcus furiosus*

In order to address the described questions concerning the stability of the *Pyrococcus* genome, we employed a combination of DNA and RNA sequencing techniques to generate an updated version of the *P. furiosus* genome (Figure 11). We used the current gold standard in genome assembly approaches, a combination of long read PacBio sequencing and short read Illumina sequencing, to obtain a highly accurate reference genome of QV60 (<1 error per Mbp) for further analysis. Differential RNA sequencing was used to map primary transcription start sites (TSS) and to improve genome annotation with the recently developed ANNOgesic pipeline (Yu et al., 2018). In order to test whether *P. furiosus*' genome is subject to genome instability upon long-term cultivation of the strain, we re-sequenced a two-year-old lab culture of *P. furiosus* employing the Nanopore sequencing technique.

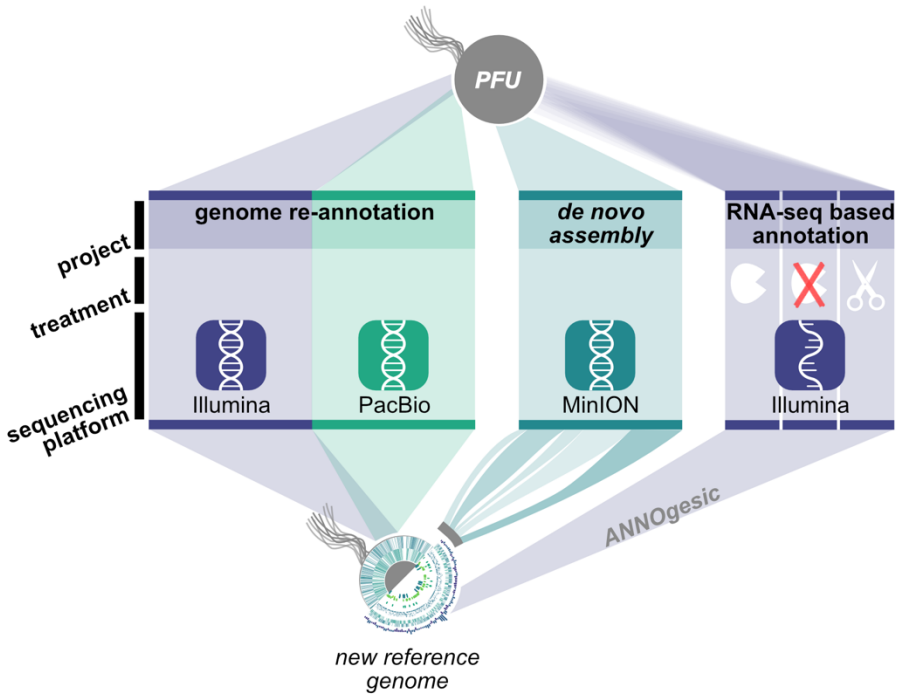


Figure 11 | Outline for the study. To build the new reference genome of *P. furiosus* DSM 3638 we used a hybrid PacBio-Illumina approach. After 2 years of subcultivation genome stability of the same strain was tested using Nanopore MinION sequencing and *de novo* assembly. Genome annotation was improved with an RNA-Seq based approach of eight mixed growth conditions to cover a broad range of transcripts. Three different RNA treatments (terminator-exonuclease treated, not-treated, fragmented) were used to map transcription start sites and additional features using the ANNOgesic pipeline (Yu et al., 2018).

4.2. Genome comparison

4.2.1. A new reference genome with two major deviations

Based on PacBio sequencing data that provided a 194-fold coverage of the genome, the *P. furiosus* type strain DSM 3638 genome was assembled *de novo* to a single contig sequence, which was error-corrected by Illumina data. The comparison with the current NCBI reference sequence (NC_003413) revealed that the new genome sequence (CP023154) is strongly syntenous to the published sequence (Figure 12, upper part).

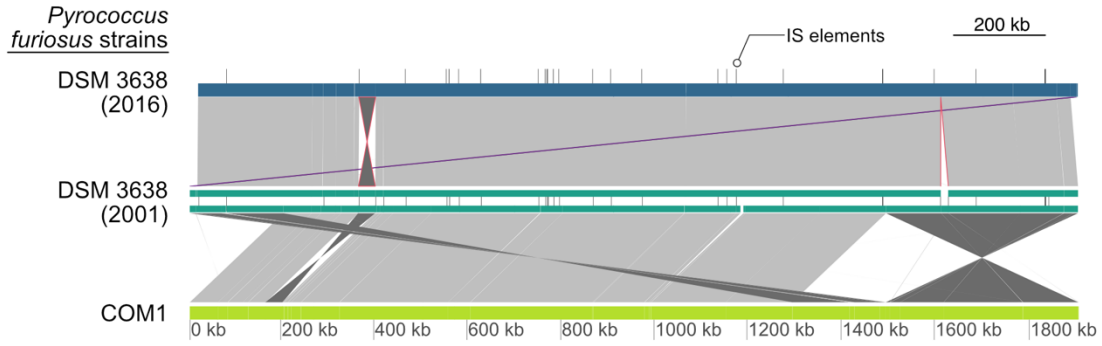


Figure 12 | Global pairwise comparison of the genome organization of the new and the original *P. furiosus* DNA sequence together with the COM1 strain (Bridger et al., 2012). Whole genome alignments were calculated with dnadiff and visualized using GenoPlotR (Kurtz et al., 2004; Guy et al., 2011). Each genome is shown as colored blocks (blue, green, light green), whereas IS elements of DSM 3638 are indicated by vertical lines. Direct matches with high similarity between genomes are colored in light grey and inversions in dark grey. The inversion of the fragment from PFDSM3638_01715 to PFDSM3638_01910 and the deletion of the 17,075 bp fragment are indicated in red and the adjusted annotation start to *cdc6* as the first gene is highlighted in purple. Additional minor variations are below the resolution of the presented map. In contrast to the large genomic rearrangements between the COM1 strain and the type strain of *P. furiosus*, there are only minor differences between the old and the new DNA sequence of *P. furiosus* DSM 3638.

However, we found two major variations:

(1) The fragment encoding the genes PFDSM3638_01715 to PFDSM3638_01910 is inverted in comparison to the corresponding region PF_RS01790 (PF0349) to PF_RS01990 (PF0388) in the original reference sequence NC_003413. This inversion represents one of numerous described chromosomal rearrangements identified for the *P. furiosus* strain COM1 (Figure 12, lower part) (Bridger et al., 2012). But in contrast to the other differences this deviation was also found in the sequencing data of a newly ordered type strain from the DSMZ by the same group (Bridger et al., 2012). This indicates that this inversion is most likely caused by an assembly problem of the original sequence (Robb et al., 2001).

(2) The region from 1,613,139 to 1,630,214 is deleted in the re-sequenced genome. The deletion of this region has also been identified previously by a ChIP-Seq approach (Reichelt et al., 2016). It belongs to a highly variable chromosomal region, which is flanked by two IS elements and proposed as an example for a recent transposon mediated gene

transfer between *P. furiosus* and *Thermococcus litoralis* (DiRuggiero et al., 2000). The fragment encodes a trehalose/maltose-specific-ABC-transporter. Even growth on maltose could not prevent this deletion (Reichelt et al., 2016).

Due to the above-mentioned deletion and a reduced number of IS elements, the complete sequence consists of 1,889,914 bp, which is 18,342 bp smaller than the NCBI reference genome (Figure 12, upper part, Table 1). The GC content and the total number of genes are also slightly reduced due to the deletion events. Using the NCBI prokaryotic genome annotation pipeline (Tatusova et al., 2016), the new genome sequence harbors 2,035 genes of which 1,982 encode proteins, whereas the residual 68 genes transcribe four rRNA genes (one 23S rRNA, one 16S rRNA and two 5S rRNAs), 46 tRNAs and 18 additional ncRNAs. The values are very similar in comparison to the annotations from the other two assemblies (Table 1). Moreover, the number of pseudogenes decreased in the new assembly from 74 to 53, which indicates that the re-sequencing allowed for correction of frameshifted genes now being present correctly annotated in full-length. All three genomes harbor seven CRISPR arrays, but the number of spacers in the CRISPR array 6 differs between the new and old *P. furiosus* type strain DSM 3638 assemblies (36 vs. 21). This might be due to assembly problems of repetitive sequences in the initial sequencing in 2002. In addition, numerous minor variations were identified including single nucleotide exchanges (causing silent or missense mutations), frameshift insertions or deletions and deletions or insertions of complete genes (summarized in Supplementary Table 1). Some of these variations were already reported by previous studies. For example, the *flaB0* gene was discovered in an earlier study, which encodes the major flagellin of the *P. furiosus* archaeellum apparatus (Näther-Schindler et al., 2014). A comparison of all coding sequences from the annotation based on the new genome assembly (CP023154) with the annotation based on the genome assembly (NC_003413) from 2002 is shown in the Supplementary Table 2.

Table 1 | Genome comparison of the re-sequenced Pyrococcus furiosus DSM 3638 together with the first published NCBI reference sequence (NC_003413) and P. furiosus COM1 (NC_018092).

	DSM 3638_2016	**DSM 3638_2001**	**COM_1**
NCBI GenBank Accession	CP023154	NC_003413	NC_018092
Genome length (bp)[1]	1889914	1908256	1909827
GC content [%]	40.75	40.77	40.79
Genes (total)[1]	2,035	2,053	2,066
Genes (coding)[1]	1,982	1,979	2,001
Genes (RNA)[1]	68	68	67
complete rRNAs[1]	2, 1, 1 (5S, 16S, 23S)	2, 1, 1 (5S, 16S, 23S)	2, 1, 1 (5S, 16S, 23S)
tRNAs[1]	46	46	46
ncRNAs[1]	18	18	17
Pseudo Genes (total)[1]	53	74	65
CRISPR Arrays[1,2]	7	7	7
CRISPR1 Spacer[2]	51	51	51
CRISPR2 Spacer[2]	21	20	21
CRISPR3 Spacer[2]	23	22	23
CRISPR4 Spacer[2]	30	30	30
CRISPR5 Spacer[2]	45	45	44
CRISPR6 Spacer[2]	36	21	36
CRISPR7 Spacer[2]	11	11	11
Total no. of IS elements[3]	24	30	40
IS200/IS605[3]	1	1	1
IS6[3]	17	23	33
IS982[3]	5	5	5
new[3]	1	1	1

[1] NCBI Prokaryotic Genome Annotation Pipeline (Tatusova et al., 2016)

[2] CRISPRFinder (Grissa et al., 2007)

[3] ISEScan (Xie and Tang, 2017)

4.2.2. Nanopore sequencing confirms genome stability

Re-sequencing of the type strain, which was stored for more than fifteen years under optimal conditions at the DSMZ, indicated indeed a very stable genome over the years. But is this also true for a strain handled in the lab which is repeatedly inoculated and stored over time in liquid cultures. To answer this question, we re-sequenced the "lab culture" about two years after we performed the Illumina/PacBio sequencing employing this time the recently developed Nanopore sequencing technique. During these two years the culture was about -roughly estimated- thirty times transferred into fresh medium. A total of 397,582 reads were accumulated of which 328,862 (82.7 %) had a mean Phred-based quality score (qscore albacore) equal or better than 7.0 representing ~308-fold

genome coverage. The median Phred quality score for all reads used for further assembly steps was 8.82.

We performed a *de novo* assembly of the genome based on the Nanopore sequencing data. First, we generated a draft genome using Canu. Subsequently, we improved the consensus sequence with Nanopolish (Koren et al., 2016; Simpson et al., 2017). In the first step, we were able to reach a closed assembly with 1 contig (1,891,829 bp) with no genomic rearrangements observed compared to the PacBio/Illumina reference sequence. In general, the genome was only 0.1% assembled larger with an identity of 99.42 % compared to the reference genome (Table 2). After polishing, the sequence identity further improved to 99.92 %. Most of the additional base pairs (+0.35 %) can be explained by insertions throughout the genome with a slight preference for additional As and Ts (A: 27 %, T: 27% of all insertions).

Table 2 | Nanopore sequencing is suitable for generating a high identity genome *de novo* in comparison with hybrid Illumina/PacBio data.

Assembly	Total Bases	No. of contigs	GC content	% Identity (1-to-1 dnadiff mummer)
Illumina/PacBio	1889914	1	40.75	100
Nanopore raw (Canu)	1891829	1	40.69	99.42
Nanopore polished (Nanopolish)	1896610	1	40.75	99.92

One major advantage of the Nanopore technique is the significantly increased read length in comparison to Illumina sequencing, which facilitates genome assembly and helps to identify genome rearrangements. The usage of a column-based DNA purification protocol led to a fragmentation of DNA, which can be observed in the length distribution of sequenced reads (median length: 1,160 bp; longest read: 31,965 bp) (Figure 13a).

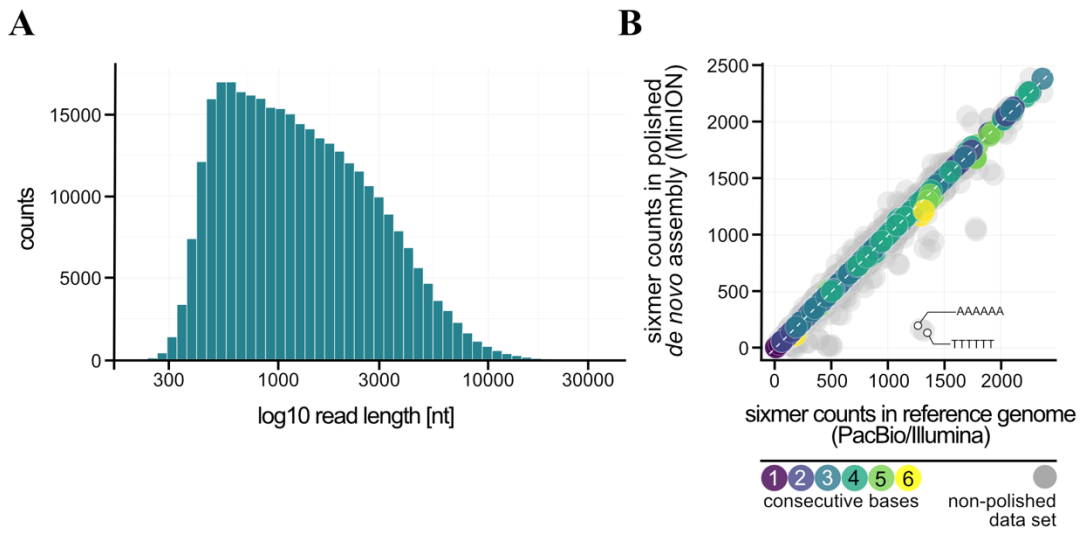


Figure 13 | Nanopore sequencing of *P. furiosus*. **a**, Fragmentation of the DNA due to the DNA purification protocol used can be ascertained from the read length distribution after sequencing (median: 1,160 bp). **b**, Sixmer-comparison of PacBio/Illumina assembly to Nanopore *de novo* assembly (polished: colored-, not-polished: grey) shows known drawbacks of Nanopore-sequencing with limited resolution of homopolymers that requires bioinformatic polishing (Simpson et al., 2017).

To build the new reference genome of *P. furiosus* DSM 3638 we used a hybrid PacBio-Illumina approach. After 2 years of subcultivation genome stability of the same strain was tested using Nanopore MinION sequencing and *de novo* assembly. Genome annotation was improved with an RNA-Seq based approach of eight mixed growth conditions to cover a broad range of transcripts. Three different RNA treatments (terminator-exonuclease treated, not-treated, fragmented) were used to map transcription start sites and additional features using the ANNOgesic pipeline (Yu et al., 2018).

Most of the errors in the non-polished assembly were in fact not random but can be explained when the counts of sixmers (all combinations of 6 nucleotides) in both assemblies are analyzed (Figure 13b). Bioinformatical polishing successfully reduced the differences in sixmer usage and lead to a very high sequence identity. All in all, we could not observe any large differences in genome organization after two years on/off cultivation and storage. Therefore, we conclude that the *Pyrococcus* genome is more stable than previously expected.

4.3. Analysis of the primary transcriptome

We combined the updated DNA sequence with an RNA sequencing approach to improve genome annotation with the recently developed ANNOgesic pipeline (Yu et al., 2018). To maximize the number of detected transcripts in the *P. furiosus* genome we pooled RNA preparations from eight different growth conditions. In one dataset we fragmented the RNA before generation of cDNA libraries to increase RNA coverage. For the

enrichment of primary transcription start sites (TSS), we employed the differential RNA-Seq (dRNA-Seq) approach, which uses a terminator exonuclease (TEX) treatment to degrade RNAs that exhibit a 5´ monophosphate that arise from nucleolytic degradation of primary transcripts but not RNAs with a 5´ triphosphate (Sharma et al., 2010). A TEX-untreated cDNA library served as a control, which includes in addition the 5´ ends of processed or degraded RNAs. Sequencing and mapping statistics can be found in Supplementary Table 3.

Transcription start sites were identified within the ANNOgesic pipeline using TSSpredator and classified into 834 primary (P), 797 antisense (A), 739 internal (I) and 145 secondary (S) transcripts according to their position relative to the next gene (Figure 14a,b) (Dugar et al., 2013; Yu et al., 2018). After using an iterative optimization process in the parameter selection module of the newly developed TSSpredator (Yu et al., 2018), the total number of TSS is similar to previously published archaeal primary transcriptome sets considering different genome sizes (Jäger et al., 2009, 2014; Babski et al., 2016; Cho et al., 2017; Smollett et al., 2017b). As a result of the densely packed genome, some of these identified transcripts belong to more than one category. For example, 212 TSS were categorized both as pTSS and aTSS that arise from head-to-head oriented genes.

Using Operon-mapper we analyzed in more detail how transcription of the 2035 identified genes is organized (Taboada et al., 2018). The program recognized 953 transcription units, which consist of 501 single genes and 452 operon structures, which contain the residual 1534 genes. One half (760) is organized in operons with two or three genes and the other half (774) is located in more complex operons with four or more genes (Supplementary Table 4). A comparison of the pTSS with the 953 transcription units revealed that almost 70% (571) of the identified primary transcripts match perfectly with the identified transcription units. It has to be emphasized that the *in silico* based operon prediction is purely based on intergenic distances and functional relationships between the genes (Taboada et al., 2018). Using a multiple conditions approach, we were able to use high-resolution transcriptomic data to improve current annotation by predicting operons this time based on TSSs, transcripts and genes within the ANNOgesic pipeline (Yu et al., 2018). The total number of transcription units decreased to 693 with the great majority (473, 68.25%) being identical to the Operon-mapper predicted units. The transcriptomic-based detection of transcription units reflects the true biological organization rather than an artefact from annotation, because we were able to detect every single gene in the fragmented RNA-Seq dataset (minimum number of reads per gene: 69).

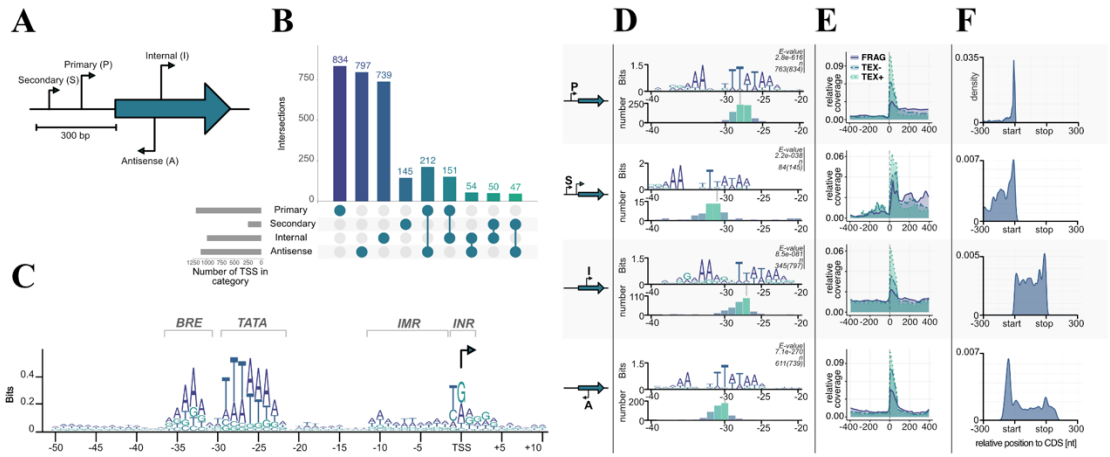


Figure 14 | Transcription start site (TSS) classification. **a**, Primary (P), Secondary (S), Internal (I), and Antisense (A) TSSs are classified by TSSpredator according to their position relative to the next gene (Dugar et al., 2013). **b**, Number of TSS identified in each category after using adaptive parameter optimization from ANNOgesic (Yu et al., 2018). **c**, Known archaeal promoter elements can be detected by visualizing a PWM calculated from all TSS from −50 to +10 bases to a start site. **d**, Promoter motifs for the individual TSS categories identified by MEME search of all upstream sequences (−50, +1). The *e*-values and the number of sequences contributing to each motif are shown on top of each panel. The panels at the bottom of each category shows a histogram aligned to the third T of the corresponding TATA boxes. **e**, Relative coverage of all reads in three sequencing datasets fragmented (FRAG, purple), terminator-exonuclease treated (TEX, blue) and TEX-non-treated-control (NOTEX, green) are plotted in relative position to a TSS in a window of −400 to +400 to confirm the output of the classification algorithm. **f**, Position of all TSS according to next gene with normalized length in a window of −300 to annotation-start and +300 to annotation-end (on gene level) are plotted. Primary TSS have predominantly short 5′ UTR lengths with a median of 13 nt, secondary TSS have larger 5′ UTRs and internal TSS are equally distributed across the corresponding gene length. Antisense TSS seem to be enriched in gene-flanking regions.

We also analyzed in more detail the promoter structure of these transcription units. In general, most archaeal promoters consist of three conserved parts, TFB recognition element (BRE), TATA box and a pyrimidine/purine dinucleotide (INR) at the -1/+1 position of the TSS (Hausner et al., 1991; Soppa, 1999; Van De Werken et al., 2006). A position weight matrix (PWM) calculated in a region -50 to +10 from all TSS confirmed the presence of a highly conserved promoter structure with consensus sequences for the BRE -36(RRAAA)-32, the TATA box -30(WTTTAAAW)-23, and the INR -1(YR)+1 (Figure 14c). It is also possible to identify the initially melted region from -11 to -2 which facilitates open complex formation due to accumulation of AT sequences. The identified promoter for *P. furiosus* fits well with published data of related organisms, e. g. *P. abyssi*, *Thermococcus kodakarensis* or *Thermococcus onnurineus* NA1 (Toffano-Nioche et al., 2013; Jäger et al., 2014; Cho et al., 2017). This is also an additional indication that the TEX treatment was successful and all the identified TSS indeed represent initiation start points of the RNA polymerase. To answer the question if different promoter structures are used

in the case of secondary, internal or antisense transcripts, the sequences from -50 to +1 of all identified TSS were also individually analyzed using *MEME* for identification of the best fitting motif (Figure 14d) (Bailey et al., 2009). All motifs exhibit typical BRE and TATA box sequences, but with reduced conservation for secondary, internal and antisense TSS. Furthermore, the location of the TATA box in relation to the TSS is slightly different. In the case of pTSS the last conserved adenine nucleotide of the TATA box is 23 bp upstream of the TSS, which is in perfect agreement with the consensus sequence of all TSS (Figure 14c) and with published data (Toffano-Nioche et al., 2013; Jäger et al., 2014; Cho et al., 2017). In contrast, the position for sTSS is at -27, for iTSS at -22 and for aTSS at -25. From previous *in vitro* transcription experiments it is known, that these distances still enable transcription, but most likely with a reduced efficiency at least for the sTSS and aTSS (Hausner et al., 1991). To further validate the identification of different TSS classes, we analyzed the distribution of reads in a window from -400 to +400 bp of all annotated TSSs (Figure 14e). As expected, we observed the highest enrichment for the TEX dataset in all TSS classes, confirming a successful enzymatical treatment and downstream bioinformatical analysis. To exclude any further bias and gain insights into possible different regulation mechanisms of the four TSS groups, we plotted the positions of all TSS 300 bp upstream and downstream of the corresponding coding sequences (Figure 14f). Most of the transcripts initiate in close proximity to the coding sequence (pTSS, median: 13 nt, mean: 49.17 nt), indicated by the strong peak of the pTSS near the start codon in contrast to sTSS that exhibit significantly longer 5´ UTR sequences. iTSS are equally distributed over the whole gene length with one prominent peak at the end of the coding sequence. Due to the high gene density in *P. furiosus* it is possible that some of these iTSS represent pTSS of downstream genes. This is also true for the strong peak upstream of the coding sequence within the aTSS. In general, the distribution of 5` untranslated regions (5`UTR) is similar to the data published in the *Thermococcales* (Jäger et al., 2014; Cho et al., 2017), but quite different from predominantly leaderless transcripts in *Haloferax volcanii* (Babski et al., 2016) and long untranslated leader regions in *Methanocaldococcus jannaschii* (Smollett et al., 2017b) (Supplementary Figure 1).

4.4. Characterization of bidirectional transcription in the context of aTSS

It is interesting to note the high number of aTSS which is in agreement with data from other archaeal organisms (Babski et al., 2016). The function of these transcripts is not known, but we assume that at least some of them are most likely nonsense transcripts, which arise from symmetric promoter sequences. A closer look to the consensus sequence (Figure 14c) revealed an almost symmetric TATA box with TTTAAA as the most conserved structure and 3 bp upstream AAA and three bp downstream TTT, although less conserved. TBP is known to bind symmetrically to the TATA box and the orientation of

transcription is determined by the binding of TFB (Cox et al., 1997; Bell et al., 1999c; Werner and Grohmann, 2011). Therefore, it is possible that some of these antisense transcripts are initiated by opposed TBP binding, which in turn results in two transcripts on opposite strands. An inspection of individual pTSSs using the IGV browser (Robinson et al., 2011) reveals strong signal counts on the antisense strand in short distance upstream of promoter elements for many transcripts. Most striking examples with head-to-tail orientation of the neighboring genes are shown in Figure 15.

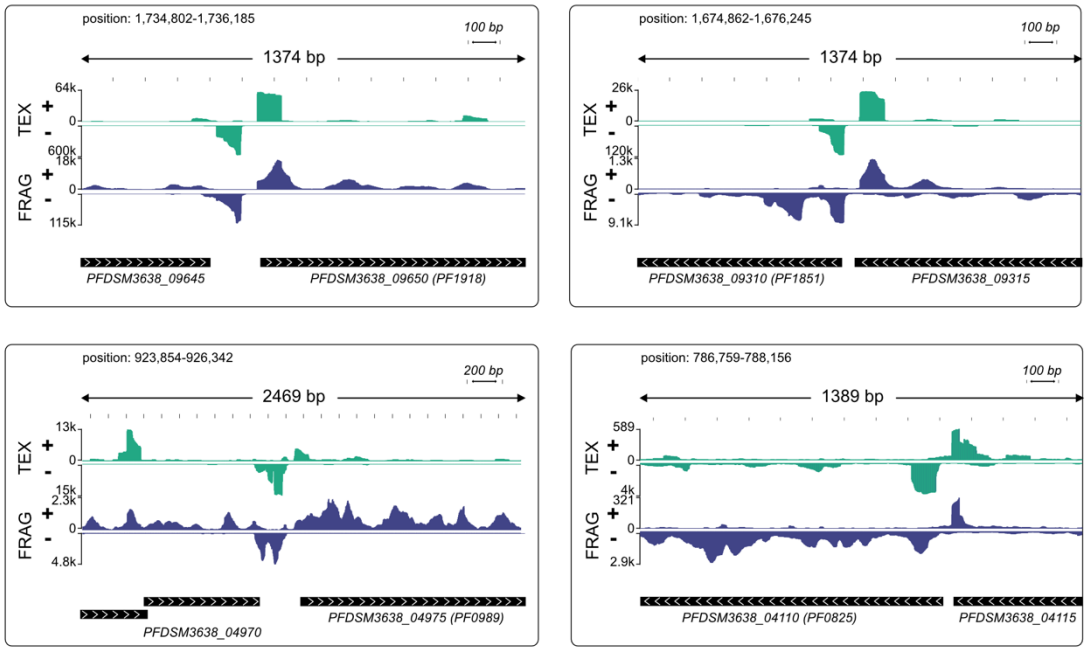


Figure 15 | Modified IGV snapshots from head-to-tail genes with an antisense TSS in close proximity to a pTSS. The exact positions of each fragment within the genome are indicated. The genes are annotated with the new genome locus tags and with old locus tags in brackets. RNA coverage on both strands is autoscaled to fit the window, with sequencing depth indicated on the *y*-axis, TEX-treated RNA in green and fragmented RNA in blue.

To investigate the occurrence of these bidirectional transcription reads in more detail we plotted the read density for all genes with a detected pTSS from TSSpredator on sense and antisense strand (n = 834). To exclude bias from genes in head-to-head orientation with actual pTSS on negative strand we split our dataset according to gene orientation (head-to-head: 388, head-to-tail: 442). This analysis showed a strong antisense peak for head-to-tail orientated genes starting about 50 bp upstream of pTSS (Figure 16a). More than 10 % of the 442 genes (49) with this orientation had very strong TEX signals in the region up to 100 bp upstream of the pTSS (more than 40 % of the reads from -400 to 400 region). In the case of head-to-head orientated genes, we also observed an enrichment of TEX signals on the antisense strand 50 bp upstream of a pTSS (Figure 16b). The expanded signal distribution is most likely caused by overlapping signals of aTTS and corresponding pTTS of upstream located genes.

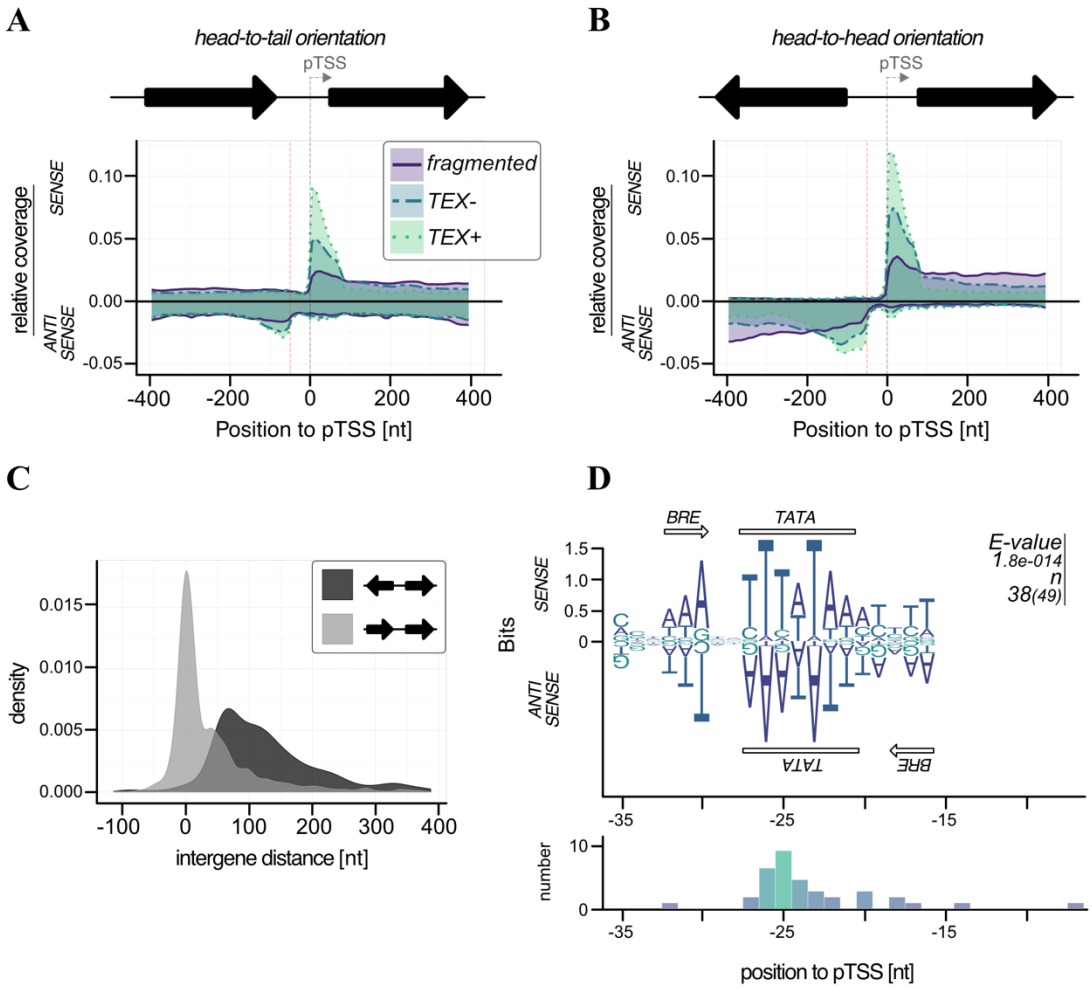


Figure 16 | Bidirectional transcription in *P. furiosus*. Genes were sorted in **a**, head-to-tail or **b**, head-to-head oriented groups and relative coverage plots (compare Figure 4d) were calculated for both sense and antisense strand. The distance of 50 b upstream of the pTSS is indicated by a red dotted line. **c**, The intergene distance for both groups is based on gene annotation, head-to-tail orientation are shown in light gray, head-to-head orientation in dark gray. **d**, MEME motif search for promoter regions with strong antisense signals resulted in a bidirectional BRE-TATA-BRE motif, that is shown on both strands. The *e*-values and the number of sequences contributing to the motif are shown on the right of the panel. This distance to pTSS is shown in the lower panel.

A more detailed analysis of the intergenic region for head-to-head genes confirmed the short intergene distances (median 117 bp) which impedes any possibility to discriminate between both signals (Figure 16c). The strong accumulation of these antisense transcripts in a distance of approximately 50 bp upstream of the pTSS most likely indicates a shared TATA element for the primary and the corresponding antisense transcript. In this case, we expect an additional BRE element downstream of the TATA element for TFB recruitment in antisense direction. To circumvent the problem with head-to-head orientated genes we only analyzed promoter sequences with head-to-tail orientation. In fact, about 78 % of these promoter regions (38) exhibited a bidirectional BRE-TATA-BRE motif located in the middle between a pTSS (position 0) and an aTSS (position -50). The BRE

on the antisense strand is less prominent than on the sense strand but can still be detected (Figure 16d).

To get additional evidence for antisense transcription induced by bidirectional promoter sequences, we analyzed the promoter of *PF1918* (PFDSM3638_09650, Figure 15 upper left) in more detail using *in vitro* transcription. A detailed sequence analysis of the upstream region confirmed the presence of a bidirectional promoter (Figure 17a) and the gene upstream of *PF1918* is located in head-to-tail orientation (Figure 17b). To distinguish between sense and antisense transcripts *in vitro,* the RNA was analyzed by primer extension experiments (Figure 17c).

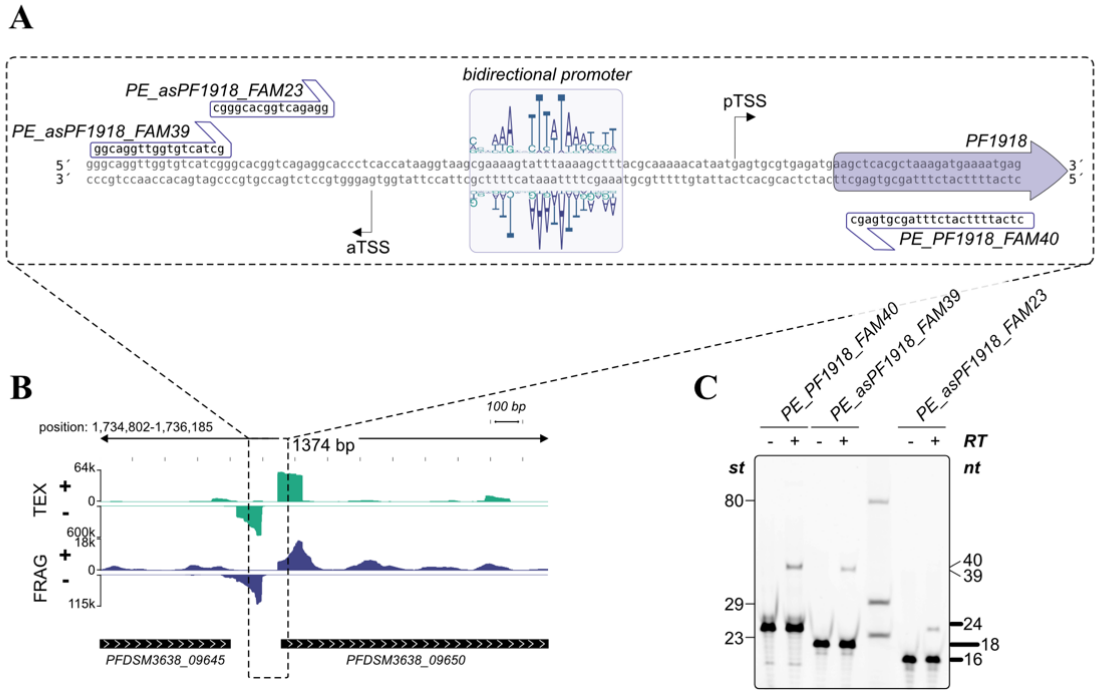


Figure 17 | Validation of bidirectional transcripts. **a**, The promoter region of PFDSM3638_09650 (old locus tag PF1918) with the bidirectional motif and the corresponding sequences used for primer extension are specified. **b**, Zoom out of the promoter region. RNA coverage on both strands is indicated. **c**, Primer extension analysis of *in vitro* synthesized RNA. The presence (+) or absence (−) of Reverse Transcriptase and the used primers are shown on top of each lane. Lengths of marker fragments (st) are shown on the left, primer signals in bold on the right and expected length of primer extension signals also on the right side.

A comparison of the 40-nucleotide sense and the 39-nucleotide antisense signal revealed that this bidirectional promoter produces the main transcript as well as the antisense transcript in almost similar amount. The distance between both TSS is 49 bp, which clearly indicates that both transcripts originate from the same TATA element. This is the first *in vitro* evidence in archaea that some of the numerous antisense transcripts can be induced by bidirectional transcription. It is possible that the AT-rich promoter sequence in combination with the low GC content of *P. furiosus* increases the frequency of

bidirectional promoters, but we assume that the symmetrical binding of the archaeal TBP to the TATA element (Cox et al., 1997) is especially prone to antisense transcription from bidirectional promoter sequences. This is in line with recent findings in eukaryotes indicating the promoter regions are intrinsically bidirectional and are shaped by evolution to bias transcription towards coding versus noncoding RNAs (Jin et al., 2017; Xu et al., 2009). Furthermore, divergent transcription is a mechanistic feature that does not imply a function for these transcripts. Transcriptional noise as a main result of antisense transcription seems to be also common in bacteria in particular in combination with a high AT content (Lloréns-Rico et al., 2016).

4.5. IS elements and a potential regulation by antisense transcripts

As already mentioned in the introduction, *P. furiosus* seems to prone to genomic rearrangements most likely due to an increased number of IS elements, which has been known for a long time as driving force for genomic reorganizations (Brügger et al., 2002; DiRuggiero et al., 2000; Zivanovic et al., 2002; Sapienza et al., 1982). Using ISEScan 1.6 we identified 24 IS elements in the new assembled genome (Figure 18a), whereas the number of IS elements in the "old" strain is 30 and 40 in the COM1 strain. In detail, the number of the IS6 family type of transposable elements is reduced (Table 1). IS6-mediated gene rearrangements have been already described in the early 2002`s as the first genome sequences of different *Pyrococcus* species became available (Zivanovic et al., 2002). Therefore, we assume that the decreased number of IS elements in the new sequenced DSM 3638 strain is a decisive point that ensures genome integrity.

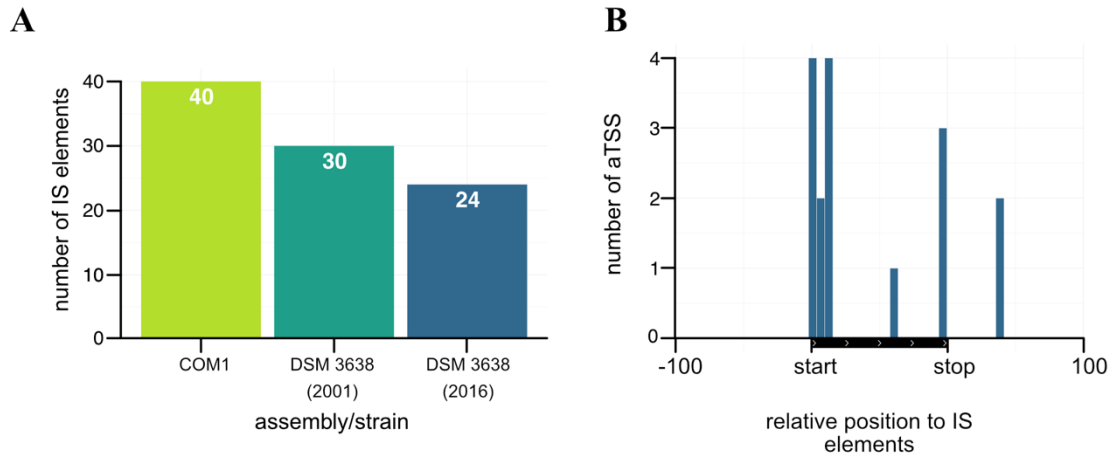


Figure 18 | Accumulation of antisense transcripts in the neighborhood of IS elements. **a**, Comparison of the number of IS elements in the COM1 strain (Bridger et al., 2012) together with the two sequences of the *P. furiosus* type strain. **b**, Position of aTSSs relative to the transposon coding sequence of IS elements.

Furthermore, there is increasing evidence that the activity of IS elements could be suppressed by corresponding antisense RNAs in particular under stress conditions (Ellis and Haniford, 2016). As our RNA library pool also contained stress conditions like heat or

cold shock and the accumulation of antisense transcripts associated with transposase-encoding genes has also been observed in other archaea (Tang et al., 2002, 2005; Jäger et al., 2009; Straub et al., 2009; Wurtzel et al., 2010; Yoon et al., 2011; Bernick et al., 2012; Heyer et al., 2012; Su et al., 2013; Toffano-Nioche et al., 2013) we mapped aTSS to the relative position of IS elements (Figure 18b). This analysis revealed an increased number of antisense transcripts, which in most cases overlap with the start of the open reading frame of the transposase. Therefore, it seems feasible that antisense transcripts of IS elements might also play a role in gene silencing in *P. furiosus* to avoid genome instability.

5. Conclusion

This study provides an updated genome assembly of *P. furiosus* using a combination of long-read PacBio sequencing and short-read Illumina sequencing. The new genome is 18,342 bp smaller than the NCBI reference from 2001 mainly due to a recently described deletion (Reichelt et al., 2016), but the overall structure is still almost identical to the published sequence of *P. furiosus*. The stability of the *P. furiosus* genome was confirmed by re-sequencing of a "lab culture" two years after initial sequencing of the strain. Our data demonstrate that it is possible to ensure genome stability in "lab cultures" by avoiding strong selection pressure, even with a strain which was assumed highly susceptible for genome rearrangements (DiRuggiero et al., 2000; Brügger et al., 2002; Zivanovic et al., 2002).

The updated DNA sequence in combination with RNA sequencing enabled us to improve genome annotation using the recently developed pipeline ANNOgesic. We included additional features, such as operon structures, TSSs and terminator sequences as well as noncoding or circRNAs to provide a comprehensive dataset of the genome features of the *P. furiosus* type strain DSM 3638 for future research (Supplementary Figure 2 and Supplementary Table 5).

Data Availability

Raw sequencing data has been submitted to the NCBI Sequence Read Archive (BioProject: PRJNA382684, BioSample: SAMN06711904). Code, raw figures, and data used during the bioinformatical analysis were uploaded to https://github.com/felixgrunberger/pyrococcus_reannotation.

Author Contributions

RoR prepared the RNA from *Pyrococcus*. BB, CS, and JO performed the PacBio and Illumina sequencing and FG the nanopore sequencing. The bioinformatical analysis was carried out by FG and RoR. FG, RoR, DG, and WH wrote the manuscript. WH, RR, and DG coordinated and supervised the work. All authors approved the final version of the manuscript.

Funding

This work was supported by the Institute of Microbiology and Archaea Center of the University of Regensburg, the SFB960, and by the German Research Foundation (DFG) with the funding program Open Access Publishing. The CU SysMed was supported by the IZKF at the University of Würzburg (project Z-6).

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgements

The authors thank Renate Richau from the University of Regensburg and Simone Severitt and Nicole Heyer from the DSMZ for excellent technical assistance. The authors thank Esther Schüler (DSMZ, Department Microorganisms) for providing biomass and Annett Bellack for providing genomic DNA from *Pyrococcus furiosus*.

Supplementary Material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.01603/full#supplementary-material

Exploring prokaryotic transcription, operon structures, rRNA maturation and modifications using Nanopore-based native RNA sequencing

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

Publication information:
bioRxiv (2020)
Received 1st version: 19 Dec 2019;
Received 2nd version: 29 May 2020;
Link: https://doi.org/10.1101/2019.12.18.880849

1. Abstract

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 (re-)annotate large transcriptional units and boundaries. Our analysis of transcription termination sites suggests that diverse termination mechanisms are in place in archaea. Moreover, we shed additional light on the poorly understood rRNA processing pathway in Archaea. 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 methylation sites and mapping of N^4-acetylcytosines modifications in rRNAs. Notably, we were able to follow the relative timely order of the installation of these modifications in the rRNA processing pathway.

2. Introduction

In the last decade, next-generation sequencing (NGS) technologies (Levy and Myers, 2016) revolutionized the field of microbiology (Escobar-Zepeda et al., 2015), 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 (Wang et al., 2009; Hör et al., 2018) that shaped our understanding of the transcriptional landscape (Croucher and Thomson, 2010; Nowrousian, 2010) and RNA-mediated regulatory processes in prokaryotes (Saliba et al., 2017). RNA sequencing (RNA-seq) technologies can be categorized according to their platform-dependent read lengths and necessity of a reverse transcription and amplification step to generate cDNA (Stark et al., 2019). 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 (Tilgner et al., 2015; Byrne et al., 2019). 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 (Eid et al., 2009; Mikheyev and Tin, 2014). 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 (Eid et al., 2009). 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 (Figure 19a). 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 (Soneson et al., 2019). Direct RNA sequencing based on the PacBio platform has also been realised but requires a customised sequencing workflow using a reverse transcriptase in the sequencing hotspot instead of a standard DNA polymerase (Vilfan et al., 2013). Native RNA-seq holds the capacity to sequence full-length transcripts and first attempts have been made to use ONT sequencing to identify RNA base modifications (e.g. methylations (Liu et al., 2019; Smith et al., 2019)). ONT sequencing is a *bona fide* single-molecule technique and hence offers the possibility to detect molecular heterogeneity in a transcriptome (Workman et al., 2019). Recently, the technology was exploited to sequence viral RNA genomes (Keller et al., 2018; Boldogkői et al., 2019; Viehweger et al., 2019; Taiaroa et al., 2020) to gain insights into viral and eukaryotic transcriptomes (Bayega et al., 2018; Boldogkői et al., 2019; Tombácz et al.,

2019; Zhao et al., 2019) and to detect RNA isoforms in eukaryotes (Byrne et al., 2017; Rahimi et al., 2019). 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 transcriptomes to make them amenable for ONT´s direct RNA sequencing kit. In the first part, we evaluated the applicability of the ONT native RNA sequencing approach to survey transcriptomic features in prokaryotes and discuss weaknesses and strengths of this method. To this end, we assessed the accuracy and reliability of native RNA-seq in comparison to published Illumina-based sequencing studies of bacterial (*Escherichia coli*) and archaeal (*Haloferax volcanii*, *Pyrococcus furiosus*) model organisms (Mao et al., 2015; Thomason et al., 2015; Babski et al., 2016; Dar and Sorek, 2018; Grünberger et al., 2019; Laass et al., 2019). The transcriptomic analysis included determination of transcript boundaries, providing, among others, insights into termination mechanisms in archaea. We moreover tested the applicability of the ONT-based native RNA sequencing approach i) to identify transcription units, (ii) to analyze pre-ribosomal RNA processing pathways and iii) to identify base modifications in (pre-)rRNAs. Despite, intrinsic limitations of the ONT-platform, we 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. As RNA modifications are retained when sequencing native RNAs, we explored the possibility to trace a selection of rRNA modifications in prokaryotes. Moreover, we provide data that position the relative timely order of the KsgA-dependent methylation and acetylation of rRNAs in archaea. Together, our comparative analysis suggests that rRNA modifications are more abundant in an hyperthermophilic organism.

3. Material and Methods

3.1. 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 (Stetter et al., 1983) 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 (Allers and Mevarech, 2005). 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 (Knüppel et al., 2018).

3.2. 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 total RNA was purified as described previously (Grünberger et al., 2019). In short, 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 total RNA was purified using the RNeasy kit (Qiagen) according to the manufacturer´s instructions. Alternatively, total RNA was isolated according to the method described by Chomczynski and Sacchi (CHOMZYNSKI, 1987), including a DNA-removal step with RNase-free DNase I (Thermo Fisher 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 by performing 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.

3.3. 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., 2020). 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.

3.4. *In vitro* transcription assays

RNA polymerase from *P. furiosus* cells and recombinant TBP and TFB were purified as described previously (Hausner et al., 1996; Kostrewa et al., 2009; Waege et al., 2010). The gene encoding histone A1 (*hpyA1*) as well as the native promoter and terminator regions was used as template for transcription reactions as described in (Spitalny and Thomm, 2008).

Run-off transcription assays (Ochs et al., 2012; Dexl et al., 2018) were carried out in a 25-µl reaction volume containing the following buffer: 40 mM HEPES (pH 7.5), 2.5 mM $MgCl_2$, 0.125 mM EDTA, 0.25 M KCl, 20 µg/ml BSA supplied with 100 µM ATP, 100 µM GTP, 100 µM CTP, 2 µM UTP, 0.037 MBq [α-^{32}P]-UTP (Hartmann Analytics) with 8.5 nM *hpy1A* template DNA, 10.5 nM RNAP, 85 nM TBP and 52 nM TFB. Reactions were incubated at 80°C or 90°C for 10 min. The radiolabeled products were extracted with phenol/chloroform and transcription products were separated on a 8%TBE-Urea (7M)-PAGE. The gel was transferred and fixed to a Whatman chromatography paper.

Gels with radioactive samples were exposed to an Imaging Plate for autoradiography. Signals derived from radiolabeled RNA transcripts were detected with FUJIFILM FLA 7000 PhosphoImager (Fuji) and analysed with Image Lab™ Software (Biorad).

3.5. 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 were 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 (Yan et al., 2018). 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).

3.6. 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 Figure 3a). 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 7. The efficiency of poly(A)-tailing was low. However, this could be compensated with a higher amount of input RNA. We added the control RNA (RCS, yeast enolase, provided in the SQK-RNA001 kit) 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.

3.7. Data analysis

3.7.1. 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 (Weirather et al., 2017; Soneson et al., 2019)), 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 Figure 4, Supplementary Table 7).

3.7.2. Mapping of reads and quantification

Files were mapped to reference genomes from *Escherichia coli* K12 MG1655 (GenBank: U00096.2) (Riley et al., 2006), *Haloferax volcanii* (NCBI Reference Sequence NC_013967) (Hartman et al., 2010) and *Pyrococcus furiosus* DSM3638 (Grünberger et al., 2019) using minimap2 (Release 2.17-r941, https://github.com/lh3/minimap2) (Li, 2018). 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 (Li et al., 2009). 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 (Soneson et al., 2019). 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 7. Transcriptome coverage was estimated by dividing the total number of CDS-mapping reads by the sum of all CDS genomic regions.

3.7.3. 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) (Liao et al., 2019). 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 to a log2 scale, normalizing for the library size and minimizing differences

between samples with small counts (Love et al., 2014a) (raw count data for TEX samples in Supplementary Table 8).

3.7.4. 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) (Loman et al., 2015).

3.7.5. 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) (Mao et al., 2015). 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 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 11).

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 9, TTS table in Supplementary Table 10).

For the analysis of prokaryotic promoter elements, the sequences 46 basepairs upstream of the corrected transcription start site were analysed to identify relevant motifs using MEME with default options except for a custom background file, calculated from intergenic sequences of the respective organism (Bailey et al., 2009).

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 (upstream) to +45 (downstream) 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 (Dar et al., 2016; Dar and Sorek, 2018)) 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). Next, the log_2-fold enrichment was calculated and plotted as in Figure 20e and Supplementary Figure 9.

RNA structural stability was predicted by folding the 45 nt long DNA upstream of the TTS using the RNAfold software from the Vienna RNA package (Lorenz et al., 2011). The results were compared to randomly selected intergenic positions of the respective organism (size = 45 nt, n = 10000) and to published TTS positions derived from Term-Seq data (Dar et al., 2016; Berkemer et al., 2020b).

Additionaly, accuracy of TTS prediction was analysed by comparing the 3´UTRs in *H. volcanii for* genes, that were detected in both Term-Seq and Nanopore data (TEX set *H. volcanii* was used for this analysis (Berkemer et al., 2020b)). The strength of the association between the two variables was investigated by calculating Pearson´s correlation coefficient.

3.7.6. Detection of rRNA processing sites and classification of rRNA intermediates

Processing site detection in bacteria and archaea was done by enrichment analysis of start and end positions of reads mapping to the relevant rRNA region. Next, co-occurence analysis in *E. coli* was performed by (i) categorizing reads according to enriched and literature-expected 5´ positions, (ii) selecting all reads that start within +/-1 from the relevant 5´ position and (iii) analysing the respective read ends. Note that non-circular reads were 5´extended by 12 nucleotides which corresponds to the actual transcript start. Exemplary reads of selected categories with enriched connected terminal positions were visualised in a genome browser-like view.

In addition to terminal enriched positions, read categories in archaea are based on the number of junctions that are detected (njunc argument, compare post-16S-bhb/pre-ligation and RNA chimera category in Supplementary Figure 16), and clipping properties of the alignments on the 5´ end of the reads (see circular RNA detection).

3.7.7. Circular RNA detection and confirmation

Circular reads were initially observed in a subset of reads, which end near/at the 5´ cleavage site of the bulge-helix-bulge (bhb), but are extensively left-clipped, which happens during mapping if the nucleotides further upstream do not match the 5´ leading, but the 3´ trailing region of the rRNA. Accuracy of 5´ and 3´ cleavage site detection using Nanopore reads was further evaluated by secondary structure prediction of the potential bulge-helix-bulge regions using RNAfold (Lorenz et al., 2011).

To investigate circular rRNA reads in more detail, a permuted linear sequence was created. This sequence contained 500 nt upstream of the annotated rRNA end to the predicted 3´cleavage site of the bhb site and was joined with the 5´ cleavage site of the bhb up to 500 nt downstream of the annotated rRNA start. Nanopore reads were re-mapped to the linear permuted sequence and again categorised by their 5´ ends and 3´ ends as circular (3´random breaks within the rRNA) or opened-circular (3´breaks at mature rRNA start, compare Supplementary Figure 17). Additionally, a shorter permuted sequence was created that included x-1 nt upstream and downstream of 3´-bhb cleavage and 5´-bhb cleavage, respectively, where x is depending on the available read length of the additional Illumina data sets used (*H. volcanii*: 100 nt, ; *P. furiosus*: 75 nt) (Babski et al., 2016; Grünberger et al., 2019). Illumina reads were also re-mapped to the permuted sequence using bowtie2, allowing for no mismatches (-D1 -N 0 -L32 -I S,1,0.50 --score-min C,0,0) and filtering out all reads that do not overlap the joined 3´-to-5´- bulge.

3.7.8. Modified base detection

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

(1) We used Tombo (Version 1.5.1, 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) (Stoiber et al., 2016). Briefly, for Figure 24f 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 Figure 24g the signals were calculated for both samples (wildtype and deletion mutant) and compared using the *control-fast5-basedirs* and *overplot Boxplot* option. For Figure 25b reference data sets were created by sorting the reads mapping to the 16S rRNA based on the pre-determined rRNA maturation categories. 5´-extended pre-rRNA were used in all cases as a background data set in the sample-compare approach. Probabilities were calculated for the sample-compare model for all read categories and plotted using custom R-scripts.

(2) For calculating the frequency of correct, deleted, inserted and wrong nucleotides at a genomic position pysamstats (https://github.com/alimanfoo/pysamstats) was used. Plots were generated using custom R scripts. The results were compared to known modification sites in 16S rRNA for *H. volcanii* (Grosjean et al., 2008) and *P. furiosus*. Note that the positions of modified RNA base modifications for *P. furiosus* are derived from a recently published study in *P. abyssi* (Coureux et al., 2020).

3.8. 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.

3.8.1. Transcriptional start sites

For all three model organisms, global transcriptional start sites were mapped recently using differential RNA sequencing (Thomason et al., 2015; Babski et al., 2016; Grünberger et al., 2019). 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.

3.8.2. Transcriptional termination sites

So far there is no transcription termination data set available for *P. furiosus*. The 3´ UTR lengths of the *E. coli* and *H. volcanii* ONT sets were compared to TTS predicted based on the Term-Seq method (Dar and Sorek, 2018; Berkemer et al., 2020a).

3.8.3. Transcriptional units

The widely used database DOOR2 (Mao et al., 2014) was used to compare the TU annotation for both archaeal sets. For *E. coli* a more recent, but also purely bioinformatical prediction, served as a reference set (Mao et al., 2015).

3.8.4. Gene expression comparison

For *P. furiosus* gene abundances from ONT data were compared to fragmented RNA sequencing data of mixed growth conditions (conditions, library, sequencing, mapping described in (Grünberger et al., 2019)), by applying a regularized log transformation as described earlier (Love et al., 2014a). For *H. volcanii* comparison, raw reads of a mixed RNA sequencing were extracted from the Sequence Read Archive SRA (SRR7811297) (Laass et al., 2019) trimmed using trimmomatic (Bolger et al., 2014), (leading:20, trailing:20, slidingwindow:4:20, minlen:12), mapped to the reference genome using bowtie2 (-N 0, -L 26) (Langmead and Salzberg, 2012), converted to sorted bam files using samtools (Li and Durbin, 2010) and compared to ONT data as described for *P. furiosus*. Illumina RNA sequencing data for *E. coli* were also extracted from the NCBI (SRP056485, 37°C LB), analysed as described for *H. volcanii* Illumina data and also compared to the ONT reference data.

3.8.5. Confirmation of circular rRNA precursors

For the confirmation of circular rRNA precursors we re-mapped Illumina reads to permuted rRNA sequences (see above). Illumina RNA sequencing data for *H. volcanii*

(SRR3623113) (Babski et al., 2016) and *P. furiosus* (SRR8767848) (Grünberger et al., 2019) were obtained from the SRA.

4. Results

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

ONT allows single-molecule sequencing of RNAs in their native form. However, at present, the direct RNA sequencing kit is designed to capture polyadenylated transcripts in the first step of library preparation. As prokaryotic RNAs are not polyadenylated, we first set up a workflow that allows whole-transcriptome native RNA sequencing using the Nanopore sequencing technology (referred to as Nanopore native RNA sequencing in this work) and that can be applied to any prokaryotic organism. The key steps of the library preparation are shown in Figure 19a: after enzymatic polyadenylation, the RNA is reverse transcribed to improve the performance by resolving secondary structures at the 3´ end (recommended by ONT) (Workman et al., 2019). Please note that, despite the synthesis of a cDNA strand during the reverse transcription step, the RNA strand and not the DNA strand is fed into the Nanopore by the motor protein. 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 discriminate primary from mature rRNAs, we used a terminator exonuclease (TEX) specifically targeting 5´-monophosphorylated ends of transcripts and compared them to non-treated samples (NOTEX, see Supplementary Figure 3a). The trimming effect of the exonuclease leads to the degradation of mature rRNAs and in turn to an enrichment of terminal positions in the non-treated samples, ultimately allowing the annotation of rRNA transcription start sites and mature rRNAs. In contrast to the experimental design of a differential RNA-seq approach, where TEX is used to detect primary transcripts in preferentially rRNA-depleted samples, we did not expect to see an effect on mRNAs, given the overall excess of rRNAs. In addition, as many Illumina sequencing-based approaches make use of a specialised library preparation design to tackle a well-defined question (Stark et al., 2019), 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 rRNA modification patterns (see Supplementary Figure 3b).

4.2. 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% (Bohlin et al., 2017), *H. volcanii*: 65% (Hartman et al., 2010), *P. furiosus*: 40.8% (Grünberger et al., 2019)). The prepared libraries were sequenced on a MinION device and reads were collected over 48 hours on R9.4 flow cells (see Supplementary Figure 4a). Although we did not deplete rRNAs, the total number of reads was still sufficient to also achieve good coverage of the mRNA transcriptome (*E.* coli: 9.2x, *P. furiosus*: 15.0x, *H. volcanii*: 10.3x, see Supplementary Table 8), which allowed us to perform transcriptional unit annotation and determination of transcript boundaries. Before mapping the reads to the reference genomes, the quality of the sequencing runs were evaluated based on raw read length distribution and quality of reads estimated by Guppy (see Supplementary Figure 4b,c). To verify that no problems occurred during sequencing or library preparation the poly-adenylated spike-in control (yeast enolase) provided in the ONT-RNA kit was used. The control showed a uniform length distribution (median lengths between 1212 and 1306 nucleotides) and a very good read quality (median quality as ascertained by the Phred score between 10.8 and 12.2) in all samples, therefore, excluding any bias during sequencing (see Supplementary Figure 4b,c, Supplementary Table 7). Lower quality in the original samples as compared to the spike-in control can be attributed to multiple reasons, including (i) compositional differences to the RNAs used to train the basecaller, and (ii) the fact that mostly ribosomal RNAs are sequenced in our samples that are known to harbor base modifications, which in turn may lead to a lower quality score especially in *P. furiosus* (Wick et al., 2019).

4.3. 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 (Garalde et al., 2018). This is also reflected in the sequenced data sets as aligned lengths up to 7864 nt can be observed (Figure 19c, Supplementary Figure 5c). As expected, the majority of reads from all samples mapped to ribosomal RNAs, whereby the 23S rRNA represents the largest proportion (Figure 19b, see Supplementary Figure 5a).

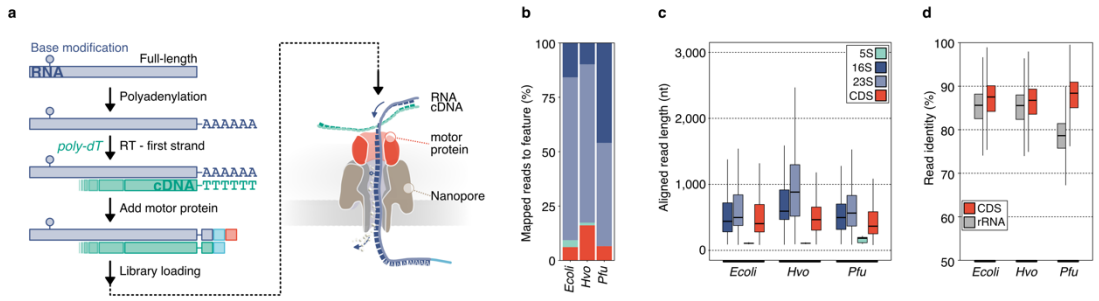


Figure 19 | Nanopore-based native RNA sequencing of prokaryotes. a, Key steps of library preparation: (1) native RNA is polyadenylated, which allows library preparation using the direct RNA kit from Oxford Nanopore and sequencing on a MinION device. (2) 3´ ligation is performed to add an adapter carrying the motor-protein (red square), which unzips the RNA-cDNA hybrid and pulls the RNA through the Nanopore (detailed description see Supplementary Figure 3a). **b**, Data sets for three prokaryotic model organisms (*Ecoli*: *Escherichia coli*, *Pfu*: *Pyrococcus furiosus*, *Hvo*: *Haloferax volcanii*) were collected and mapped to their respective reference genome. Transcript abundances of genomic features (protein coding genes (CDS): red, 5S rRNA: green, 16S rRNA: purple, 23S rRNA: light-purple) were estimated using featurecounts (Liao et al., 2019) (TEX-treated samples are shown as example in Figure 19). **c**, Aligned read lengths across different genomic features. **d**, Comparison of read identities between CDS (red) and rRNA (grey)-mapping reads.

In general, the read identity of CDS-mapping reads is higher than for rRNA mapping reads, but lower than the spike-in control (Figure 19d, see Supplementary Figure 5b,c). It is noteworthy, that accurate mapping of very short reads is currently not supported by the minimap2 mapping tool, which explains the 100 nt cut-off in our data sets (see Supplementary Figure 5d) (Garalde et al., 2018; Li, 2018; Workman et al., 2019). 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, CRISPR-RNAs or tRNAs fall below this threshold, we excluded these RNAs from further analysis. While short transcripts are problematic, longer RNAs can be sequenced and mapped accurately without loss in quality (see Supplementary Figure 5e). As the raw read quality correlates with the mapping identity of the reads, problems during sequencing can be live-monitored in MinKNOW and the run can be canceled allowing the loading of a new library (see Supplementary Figure 5f). Since the subsequent analysis of transcriptional units is heavily dependent on the integrity of the data, we verified the data integrity in the next steps. The addition of poly(A)$_{20}$ (length of the reverse transcription adapter) is sufficient to allow for the annealing of the poly(T)-adapter required for reverse transcription and sequencing. This goes in line with the shortest median length we observed for the 5S rRNA (see *E. coli* TEX sample) (see Supplementary Figure 6). For most of the transcripts, a poly(A) tail with 50 to 100 nt was detected. In addition, the overall correlation of transcript abundances calculated from sequencing data using Nanopore or Illumina technology was very high suggesting that a good coverage of the transcriptome

was achieved and that native RNA sequencing is not biased towards a subset of transcripts (see Supplementary Figure 7a,b,c, transcript abundance data in Supplementary Table 8).

4.4. Mapping of transcriptional boundaries

4.4.1. Transcription start sites

Transcription start site (TSS) and transcription termination site (TTS) detection was based on the determination of transcriptional units (TU) (compare material and methods section) (Mao et al., 2015). In total, we identified a comparably high number of TSS in ONT data sets compared to TSS detected by Illumina d(ifferential) RNA-seq (see Supplementary Figure 8a) (Thomason et al., 2015; Babski et al., 2016; Grünberger et al., 2019).Furthermore, the substantial overlap of genes with a predicted TSS in both technolgies (see Supplementary Figure 8a), allowed us to evaluate the accuracy of ONT TSS mapping (positions of TSS derived from ONT TEX-treated samples in Supplementary Table 9). For example, in case of *E. coli*, we could annotate the TSS for 1925 genes using the results of a published dRNA-seq study, and 1,272 TSS were detected by ONT native RNA sequencing. The portion of TSS identified only based on the ONT sequencing data (653 TSS) or Illumina sequencing data (1,436 TSS) is mostly caused by the different algorithms used and the limited sequencing depth in the ONT data sets. Strikingly, despite missing specific enrichment of primary transcripts, the median 5´ untranslated region (UTR) lengths were very similar when data from ONT native and Illumina-based RNA sequencing were compared (*E. coli*: 68 ONT vs. 62 Illumina; *P. furiosus*: 23 ONT vs. 13 Illumina; *H. volcanii*: 1 ONT vs. 0 Illumina, Figure 20c). Please note that TSS-mapping based on Nanopore native RNA-seq data must be corrected by 12 nucleotides (Figure 20a,c). 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 (Figure 20a) (Workman et al., 2019; Parker et al., 2020).

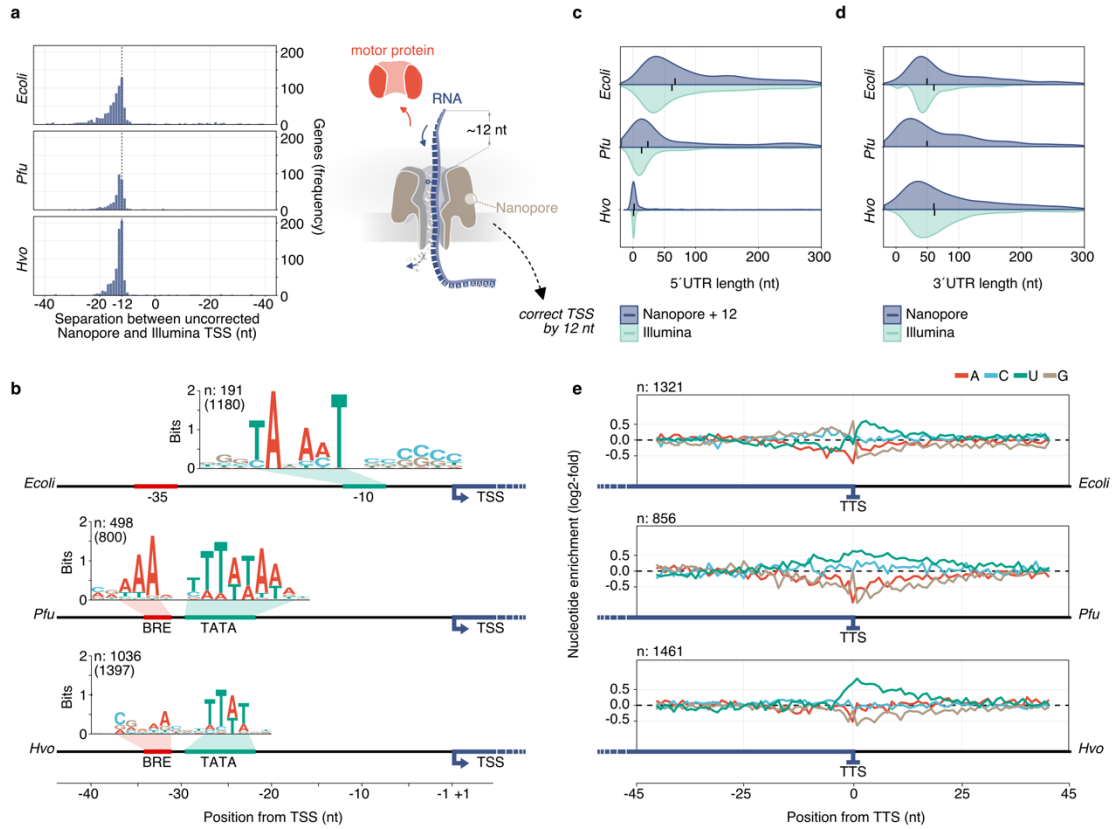


Figure 20 | Detection of transcript boundaries. a, Left panel: Separation between uncorrected Nanopore-predicted TSS and comparison to Illumina d(ifferential) RNA-Seq data from published data sets for *E. coli* (Dar and Sorek, 2018), *P. furiosus* (Grünberger et al., 2019) and *H. volcanii* (Babski et al., 2016). Right panel: The translocation speed of the last 12 nucleotides (nt) is not controlled, as the motor protein is falling off. Therefore, native RNA reads are shortened by ~12 nt. **b**, MEME analysis (Bailey et al., 2009) of extracted sequences upstream of Nanopore-predicted TSS reveals bacterial (position -10) and archaeal-specific promoter elements (BRE: B-recognition element, TATA: TATA-box recognized by transcription factor B), therefore validating the positions of predicted TSS. **c**, 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 Figure 8). **d**, 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 Term-Seq Illumina data from *E. coli* and *H. volcanii* (light-green) (Dar and Sorek, 2018; Berkemer et al., 2020a). **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 Figure 9).

Promoter analysis confirmed the presence of well-known sequence motifs of bacterial and archaeal promoters (Thomason et al., 2015; Babski et al., 2016; Grünberger et al., 2019). This includes the TATA-box and TFB-recognition element (BRE) characteristic for archaeal promoters and the -10 element in bacterial promoters (Figure 20d). The -35 element in *E. coli* has been previously shown to be less enriched compared to the -10 site (Yan et al., 2018), 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 8b,c,d). The overall correlation between the two techniques was high even though in some instances only a moderate correlation was found (see Supplementary Figure 8). 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 ($\rho = 0.86$) (see Supplementary Figure 8e).

4.4.2. Transcription termination sites

In prokaryotes, transcription termination is mediated either by a proteineous factor (Rho in bacteria (Mitra et al., 2017), CPSF in archaea (Sanders et al., 2020)) or intrinsic RNA sequences (bacteria: a GC-rich sequence that forms a stem-loop followed by a U-rich sequence (Ray-Soni et al., 2016), in archaea: poly(U) stretch (Dar et al., 2016)). 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. This approach opened up the opportunity to not only map termination sites but to also gain insights into 3' UTR lengths, for which no reference data sets for *P. furiosus* were available. The distribution of 3´ UTRs in *E. coli* and *H. volcanii* ONT data closely resembles the data from previous Illumina-based studies (Dar and Sorek, 2018; Berkemer et al., 2020a). Strikingly, the length of untranslated regions at the 3´end of annotated transcripts is very similar between the three prokaryotes (Figure 20d). In total, 1321 TTS in *E. coli*, 856 in *P. furiosus* and 1461 in *H. volcanii* were analysed (positions of TTS in Supplementary Table 10). A meta-analysis of all TTS surrounding regions revealed different sequence-dependent termination mechanisms that were confirmed using motif scanning and ΔG analysis (Figure 20e, see Supplementary Figure 9a-e). 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* (Dar et al., 2016) and similar to the $U_{(8)}$ sequence in *Thermococcus kodakarensis* determined by an *in vivo* reporter assay (Santangelo et al., 2009). The termination motif found in *H. volcanii* is a $(U)_4$-sequence and located right after the TTS (see Supplementary Figure 9d). In *P. furiosus*, the poly(U) is not preceded by a stem-loop structure, confirming that stem-loop structures do not play a role in hyperthermophilic organisms for general termination (see Supplementary Figure 9e) (Dar et al., 2016; Berkemer et al., 2020a). However, this is less clear in *H.* volcanii, where stem loops have been shown to terminate transcripts, although less efficiently (see Supplementary Figure 9e) (Berkemer et al., 2020a). 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

currently cannot be basecalled accurately, which leads to problems during mapping and ultimately to TTS positions that are positioned upstream of the poly(U) signal. This is also supported by a position-specific comparison of TTS in *H. volcanii* identified with Illumina and ONT reads (see Supplementary Figure 9,g). However, it was encouraging to see, that with increasing sequencing depth the correlation significantly improves (see Supplementary Figure 9). 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 Histone A1 transcript in *P. furiosus*, respectively (see Supplementary Figure 10). 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 terminate at subsequent poly(U) termination signals (see Supplementary Figure 10). 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.* The same applies to the termination of the histone and Alba gene transcripts in *P. furiosus* (see Supplementary Figure 10). The histone transcript has already been shown to terminate at four consecutive U-stretches (U1-U4) consisting of at least five U´s *in vitro* (Spitalny and Thomm, 2008). While we could confirm that the archaeal RNAP mainly terminates at the U1 site, the downstream sites seem to deviate from the three U_5 backup TTS *in vivo*. Instead, termination already occurred at two U_4 stretches, each upstream of U2 and U3, respectively (see Supplementary Figure 10d-f). Surveying the heterogeneity of the transcripts with an extended 3'UTR, we found a heterogeneous distribution in the length of the transcripts for both, the Histone and Alba mRNA. This pattern suggests that either a step-wise trimming of the 3'UTR occurs that eventually yields the mature RNAs or that the RNA polymerase reads through the first termination sequence, stochastically stops transcription after the first termination sequence and RNA polymerases that continued transcription beyond the first TTS terminate at one of the following TTS.

As observed for *E. coli* termination sequences, Cytosines are enriched over Guanosine adjacent to the TTS in *P. furiosus* (Figure 20e). Motifs detected in the *E. coli* data set correspond to intrinsic (poly(U)) termination signatures and REP sequences (Khemici and Carpousis, 2003; Liang et al., 2015) that can frequently be found in intergenic regions and cause transcription termination at Rho-dependent attenuators (see Supplementary Figure 9a,b) (Dar and Sorek, 2018). As expected, fold stability analysis of intrinsic *E. coli* terminators and hairpin-forming REP sequences, revealed secondary structures in both cases (Supplementary Figure 9e). However, the stem-loops could potentially also represent processing or pause sites.

4.5. 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 (Figure 21a).

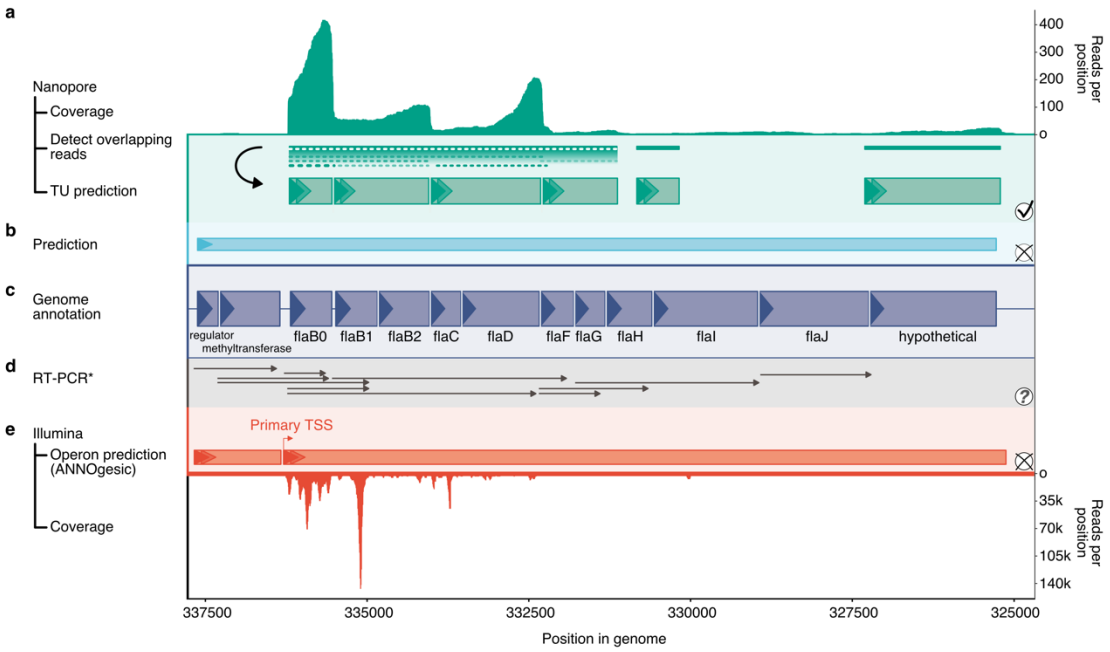


Figure 21 | Transcription unit (TU) annotation of the flagellum-operon in *P. furiosus*. **a**, Coverage of Nanopore reads is shown in the top panel. TU prediction is performed by detection and linkage of overlapping reads and splitting them according to a 3´ drop in coverage (see Supplementary Figure 10). Predicted TUs are drawn with green boxes according to scale. **b**, Comparison to bioinformatical prediction using the DOOR2 database (Mao et al., 2014). **c**, Genome annotation with abbreviated gene names, boxed drawn to scale and strand indicated by triangles (Grünberger et al., 2019). **d**, Comparison to results from published RT-PCR experiments (Näther-Schindler et al., 2014). All transcripts detected are drawn by arrows. **e**, Operon prediction based on mixed Illumina-Seq (coverage in lower panel) and predicted by ANNOgesic (Yu et al., 2018; Grünberger et al., 2019). The primary transcription start site (TSS) of the large transcriptional unit is highlighted.

For whole-genome analysis, the annotation strategy was based on two major observations: First, during RNA preparation, RNA processing or degradation can occur, which limits the probability of sequencing an RNA in its native form as the percentage of full-length transcripts decreases with expected gene size (see Supplementary Figure 11a). Secondly, we detected a decrease in coverage from the 3´ to 5´end of the RNA in all RNA classes except for the spike-in control (see Supplementary Figure 11b), which is a limitation reported in the literature (Smith et al., 2019; Soneson et al., 2019; Wongsurawat et al., 2019). Therefore, we assume that not Nanopore sequencing but library preparation causes this problem. Based on this information, we developed a strategy that first collapses all overlapping reads and then splits them according to a significant coverage drop on the 3´ ends (annotation of TUs based on this strategy in Supplementary Table 11). We compared

the results to database annotations and found that most of the differences are either caused by the low sequencing depth or by single-unit operons that have been collapsed and are now two-unit operons in the ONT data sets (see Supplementary Figure 12a,b) (Mao et al., 2014, 2015). Even though limited read availability is a concern in all data sets, many large operons were detected for all organisms (see Supplementary Figure 12c). In case of limited bioinformatical resources, TUs can be explored visually in a genome browser, which is mostly not possible for Illumina reads (Figure 21, see Supplementary Figure 13, Supplementary Figure 14). It further allows a quantitative analysis of individual transcripts in relation to other elements of the TU and performs much better than pure bioinformatical prediction or molecular biology methods (RT-PCR) as shown for the flagellum/archaellum operon in *P. furiosus* (Figure 21) (Mao et al., 2014; Näther-Schindler et al., 2014). Here, it was possible to (i) detect multiple transcription units forming this cluster, (ii) confirm transcriptional start sites and (iii) to confirm that flaB0, the protein that is referred to as the major archaellin in *P. furiosus* (Näther-Schindler et al., 2014; Grünberger et al., 2019), is transcribed in large excess over the other archaellum genes. The largest identified TU cluster in *H. volcanii* mainly consists of ribosomal protein genes. Based on the native RNA-seq data, the analysis suggests that this operon is split into two transcription units. This shows that the ONT native RNA sequencing method provides the opportunity to annotate transcriptional units thereby outperforming the bioinformatics-only prediction as well as the visual inspection of Illumina coverage (see Supplementary Figure 13). Besides, we confirmed the complex transcription pattern of the major ribosomal protein gene cluster in *E. coli* that stretches over more than 10 kb, including the accurate determination of TSS and TTS and a putative cleavage site in the secY gene (see Supplementary Figure 14) (Lioliou et al., 2012).

4.6. 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 (Shajani et al., 2011; Smith et al., 2018; Bechhofer and Deutscher, 2019). Ribosomal RNA in *E. coli* is transcribed from 7 independent rDNA operons encoding the mature rRNAs (16S, 23S and 5S rRNAs) and some tRNAs which are interspersed by RNA spacer elements (Klappenbach, 2001). In agreement with a previous study, transcription of *rrnC* from two promoters (transcription start sites at -293 and -175) was detected accurately in the TEX-treated sample, which is enriched in primary transcripts (Figure 22a,b) (Maeda et al., 2015).

The rRNA maturation process, which requires the action of well-defined endo- and exo-ribonuclease activities, culminates in the formation of stoichiometric amounts of mature 16S, 23S, and 5S rRNAs (Shajani et al., 2011; Smith et al., 2018; Bechhofer and Deutscher, 2019; Jain, 2020). Unexpectedly, the sequencing efficiency of mature 16S rRNA

was lower than the 23S rRNA (Figure 13b, see Supplementary Figure 15). The reasons for this apparent discrepancy is so far unclear.

To re-trace the multi-step rRNA maturation process, we performed a co-occurrence analysis of read start and read end positions. Strikingly, we could identify most of the known 5´-processing/intermediate sites at nucleotide resolution in wildtype *E. coli* (Figure 22b). Next, we categorized reads based on their experimentally verified and literature expected 5´ terminal positions and analysed 3´-enriched connected positions (Figure 22c). Considering the 3´-to-5´ sequencing strategy of Nanopore sequencing, this co-occurrence analysis allows the assignment of 3´ terminal positions and distinction to random 3´ degraded reads.

Although we could detect RNA of similar size or longer (see above) very well, the short-lived full rDNA operon transcript detected in RNase III deficient strain (Hofmann and Miller, 1977), is not observed using our experimental set-up. In contrast, the downstream known pre-rRNA intermediates, which are generated by the action of RNAse III were detected (Fig.4). Among these intermediates, the 17S pre-rRNA (115 additional nt at the 5´ end and 33 nt at the 3´ end of the 16S rRNA) and the P23S (7 additional nt at the 5´ end and 8 nt at the 3´ end of the 23S rRNA), were identified (Figure 22b,c,d). Final 5´ end maturation of the 16S rRNA mainly occurs before the 3´ end (Smith et al., 2018) by the action of additional ribonucleases (RNase E -66, RNase G -3, RNase AM 0 5´ mature (Jain, 2020)), which leads to an enrichment of reads that have extended 3´trailing regions compared to the mature position (Figure 22c,d). Together we could identify most of the known rRNA processing-intermediates/-sites at near-nucleotide resolution in wildtype *E. coli*. However, it should be noted that the current experimental set-up can be biased by 5´ and 3´ degradation events, prohibiting precise 3´ end mapping in some cases and causing difficulties to identify short-lived/low-aboundant pre-rRNA intermediates.

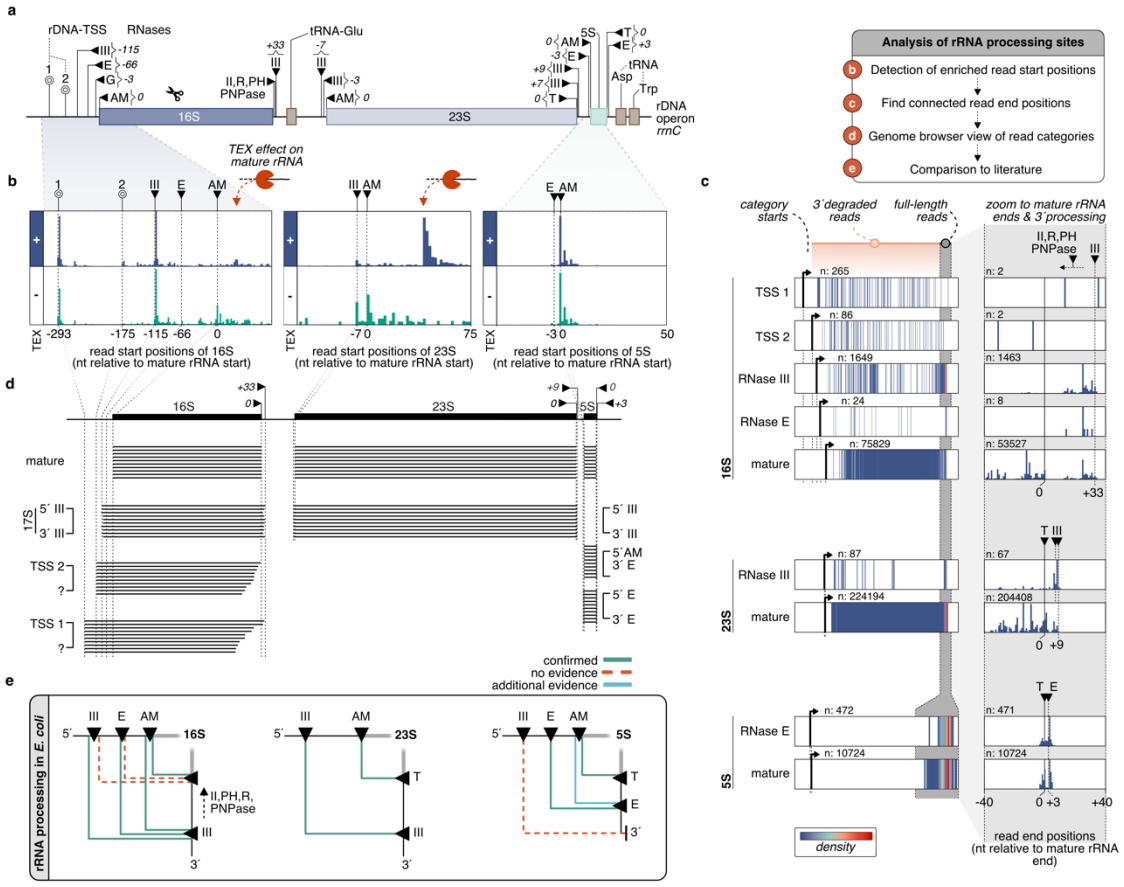


Figure 22 | Detection and confirmation of rRNA processing sites in *E. coli*. **a**, Transcription of the rDNA locus (rrnC) is starting from two promoters (transcription start sites at -293 and -175) (Maeda et al., 2015). Precursor RNAs are cleaved by RNases (black triangles) at depicted positions (Shajani et al., 2011; Ferreira-Cerca, 2017; Smith et al., 2018; Jain, 2020). **b**, Histograms of read start positions for 16S, 23S and 5S rRNA. Positions are relative to the annotated boundaries of mature rRNAs and shown for TEX (+, purple) and NOTEX (-, green) samples. **c**, Read start positions were used to classify categories and analyse the co-occurrence of read start to end positions. In the left panel, the color-coded density (low: blue, high: red) of read end positions from the category start position (black arrow) to the expected 3´ terminal area (grey area) is shown for the selected categories. While all lines outside of the grey area presumably represent 3´degraded reads, the full-length read end positions inside the shaded area have been analysed in more detail in the right panel (only TEX sample is shown). **d**, Based on the co-occurrence analysis of enriched read start and end positions, single reads were extracted and are visualised in a genome-browser view. **e**, The action of endo- and exonucleases (black triangles) is required for the maturation of rRNAs in E. coli. The multi-step maturation process leads to intermediates we could confirm (green lines) using Nanopore sequencing. While red dashed lines indicate intermediate pre-rRNA we cannot detect, blue lines indicate the presence of an additional intermediate.

4.7. Insights into archaeal ribosomal RNA processing

In comparison to bacteria or eukaryotes, ribosomal RNA processing in archaea is still poorly understood (Jacob et al., 2013; Yip et al., 2013; Ferreira-Cerca, 2017). Our current knowledge suggests that the primary polycistronic rRNA precursor contains two processing stems formed by the 5´ leader and 3´ trailer sequences surrounding the 16S and 23S rRNAs

(Tang et al., 2002; Yip et al., 2013; Ferreira-Cerca, 2017; Clouet-D'Orval et al., 2018). In Euryarchaeota, the 16S and 23S rRNAs are additionally separated by the presence of an internal tRNA. In most archaea, the 16S and 23S rRNA processing stems contain a bulge-helix-bulge (bhb) motif which is, in the context of intron-containing tRNA, recognized by the splicing endonuclease endA (Russell et al., 1999; Tang et al., 2002; Clouet-D'Orval et al., 2018). Similar to intron-containing tRNA maturation, processing at the bulge-helix-bulge motifs is followed by the covalent ligation of the resulting extremities, thereby generating the archaeal specific circular pre-16S and circular pre-23S rRNAs (Tang et al., 2002; Danan et al., 2012; Ferreira-Cerca, 2017; Jüttner et al., 2020). The exact molecular mechanisms by which the circular pre-rRNA intermediates are further processed into linear mature rRNAs remain to be fully characterized (Tang et al., 2002; Ferreira-Cerca, 2017; Qi et al., 2020).

Performing enrichment analysis of terminal positions, we aimed to confirm and expand our knowledge on the poorly characterized multi-step ribosomal maturation process in two evolutionary divergent archaea, *P. furiosus* and *H. volcanii* (Figure 23, see Supplementary Figure 16) (Grosjean et al., 2008; Ferreira-Cerca, 2017; Clouet-D'Orval et al., 2018). As expected, almost all reads are categorized as fully maturated transcripts of the single 16S/23S rRNA cluster that do not contain extended 5´ or 3´ spacer regions (Figure 23c). Surprisingly, and in contrast to our analysis performed in *E. coli* (Figure 22), some of the observed mature rRNAs 5´ positions did not precisely match the available annotations at NCBI (https://www.ncbi.nlm.nih.gov/genome/) or the archaeal genome browser (AGB, http://archaea.ucsc.edu), which are also showing discrepancies (summarized in Supplementary Figure 14). However, selected examination of the putative mature rRNA extremities obtained by ONT did match our independent experimental validations by primer extension analysis of the 5´ends of the 16S and 23S rRNAs of *H. volcanii* (Supplementary Figure 16d). These results, and those obtained for *E. coli*, suggest that the mature 5´ of the rRNAs determined by native RNA sequencing most probably represent the genuine mature rRNA extremities.

Despite the high sequencing depth of the (pre-)rRNA, we did not detect a full-length precursor consisting of the 16S leading-16S-tRNA-23S-23S trailing elements in *P. furiosus* and *H. volcanii,* suggesting that, similar to the *E. coli* situation (see above), very early rRNA processing events may occur rapidly in these cells. The remaining rRNA reads were grouped according to (i) their 5´ leading and 3´ trailing lengths, (ii) the number of junctions and (iii) clipping properties of the alignments into several additional categories that are overall less abundant than the mature rRNAs and may represent either rRNA processing intermediates or are RNA elements generated as a product of pre-rRNA processing. Among these putative pre-rRNA-related intermediates, some are common to both archaea analysed, whereas others are apparently only found in one or the other organism. The pre-rRNA-related intermediates were selected on the basis of abundance

and/or biological interpretability and/or prior characterization. The overall findings were used to extract an hypothetical rRNA maturation pathway in archaea which is summarized in Figure 5a. The rational for the selected pre-rRNA intermediates for *H. volcanii* and *P. furiosus* is described in more detail below and examplified in Figure 23b.

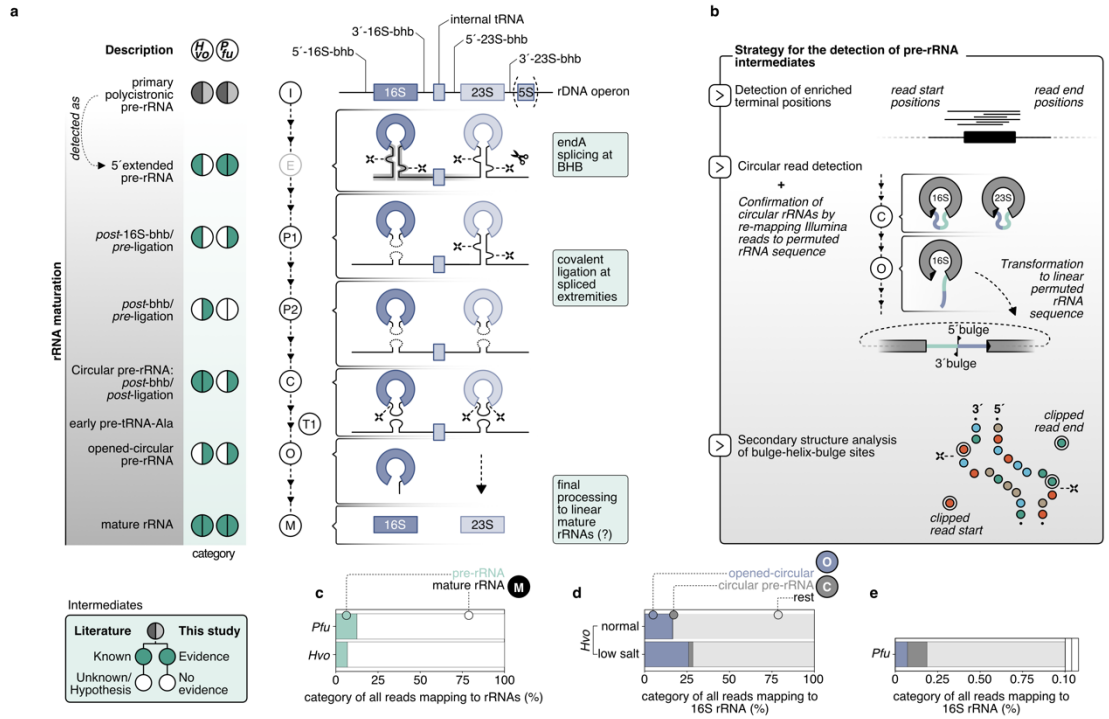


Figure 23 | Update of the archaeal rRNA processing model. a, The processing stems of the primary polycistronic pre-rRNA, formed by 5´-leader and 3´-trailer sequences, contain bulge-helix-bulge motifs that are recognized and cleaved by the endonuclease endA. This is followed by the covalent ligation of the resulting extremities, which leads to archaeal-specific circular pre-16S and pre-23S rRNAs. Further maturation steps are so far unkown. The multi-step maturation process was analysed based on the strategy depicted in b and compared to already known events. **b**, Strategy for the detection of pre-rRNA intermediates: Categories were first selected based on enriched terminal positions. Clipping abnormalities lead to the detection of circular reads that could be verified by re-mapping Nanopore and Illumina reads to a linear permuted rRNA sequence containing the joined 3´-to-5´bulge region. The exact position of the joined region was additionally verified by secondary structure analysis of the bulge-helix-bulge sites (Lorenz et al., 2011). **c**, Quantification of reads mapping to the mature rRNA (purple) and precursors (brown). **d**, Quantification of full-circular (green) and open-circular (orange) reads in *H. volcanii* wt and low-salt sample and in *P. furiosus*. The total number of these circular reads was compared to the number of reads mapping to the 16S rRNA.

In *H. volcanii*, we identified 3 classes of putative intermediates. In the first class (class P2) the pre-rRNA boundaries of these intermediates match the previously described bhb processing sites located within the 16S and 23S rRNA processing stems, respectively (see Supplementary Figure 16). These intermediates extend from the 5´ to 3´bulge cleavage sites, however, these extremities are not covalently ligated and may correspond to post-bhb cleavage/pre-ligation pre-rRNA intermediates. An exemplary verification of the 5´

boundary of the post-bhb cleavage/pre-ligation pre-23S rRNA analysed by primer extension is provided in Supplementary Figure 14d. The second class (class C) correspond to permuted reads covalently connecting the 5´ and 3´ bulge cleavage sites, and are likely observed as the result of random nicking of the circular pre-rRNA intermediates during sample preparations and are categorized as post-bhb/post-ligation pre-rRNA intermediates. We verified reads by re-mapping Nanopore and Illumina reads to a permuted RNA sequence that was designed by joining the 3´ bulge with the 5´ bulge to mimic the actual sequence of circular rRNAs (Figure 23, see Supplementary Figure 17). Similarly, we detected a third main class (class O), which corresponds to a putative pre-16S rRNA intermediate showing an immature 3´ end, which is extended by the typical permuted spacers sequence observed in the circular-pre-16S rRNA (Jüttner et al., 2020). This topology possibly results from linearization of the circular pre-16S rRNA intermediate at the mature 16S rRNA 5´end (opened-circular-pre-16S rRNA). This putative pre-rRNA intermediate is relatively abundant in *H. volcanii* (n: 1120, 15% of all reads mapping in the 16S rRNA region) and strikingly shows a non random 3´end extremity - matching with the linearization of circ-pre-rRNA at the mature 16S rRNA 5´end. In contrast, the resulting 5´end were rather heterogenous, probably due to degradation during sample preparation (see Supplementary Figure 17b).

To provide additional examples that show the potential to describe rRNA processing events in archaea, we sequenced an *H. volcanii* wildtype strain grown under low salt conditions known to accumulate large amounts of a longer 16S rRNA variant (see Supplementary Figure 16b,c) (Laass et al., 2019). Under these conditions, a 16S rRNA (precursor) with extended 5´ and 3´ UTRs (5´: -108 5´bulge, 3´: +70 3´bulge) appears that is enriched in this context. Quantification by comparison with the NOTEX wildtype set confirms the previous detection of this rRNA variant in a gel-electrophoretic analysis of total RNA (see Supplementary Figure 16b) (Laass et al., 2019). To reveal more details about the nature of these precursors, we re-mapped the reads to the permuted linear rRNA 16S sequence. We observed that the relative number of reads obtained for circ-pre-16S rRNA (class C) and especially of opened-circ-pre-16S rRNA (class O) obtained under low salt conditions were exceeding the ones from “normal” conditions, indicating that rRNA maturation and/or turnover is affected in this “stress” condition (Figure 23d). The functional relevance of these observations remain to be analysed.

In *P. furiosus*, for which we have obtained larger amounts of reads, we could define 4 categories of pre-rRNA-related intermediates (see Supplementary Figure 16k, ranked by their timely appearance): (1) Fragmented full-length precursor rRNAs (I), (2) 16S rRNA leading/trailing sequence-tRNA-23S rRNA, (3) a putative RNA chimera resulting from processing and RNA ligation activities that entails the 16S rRNA leading/trailing sequence-tRNA-23S rRNA leading/trailing sequence (T1), and (4) permuted 16S and 23S pre-rRNA intermediates (C) (Figure 23, see Supplementary Figure 16).

The putative permuted reads (4) are reminiscent of the reads typically observed for circular pre-rRNA in *H. volcanii* (class C) and may correspond to the covalent ligation of the 5´ and 3´ spacers generated by cleavages at the bulge-helix-bulge motifs within the processing stems (Figure 23 and (Tang et al., 2002; Danan et al., 2012; Jüttner et al., 2020)). To verify this hypothesis, we performed RNA structure prediction of the corresponding double stranded RNA regions (see Supplementary Figure 18). In agreement with the permuted reads, we could place the corresponding extremities within the bulge-helix-bulge motifs. However, the 23S processing stem does not adopt a canonical bhb motif, but forms an alternative structure similar to the one previously described for the 16S rRNA bhb motif in *S. acidocaldarius* (Durovic and Dennis, 1994; Russell et al., 1999). Others and we could previously demonstrate that this alternative structure is compatible with circular-pre-16S rRNA formation in *S. acidocaldarius* (Danan et al., 2012; Jüttner et al., 2020). Therefore, these permuted reads likely originate from random opening of the archaeal specific circular-pre-rRNA intermediates during sample preparation/sequencing (as observed for *H. volcanii*) and suggest that like various archaea analysed so far, circular pre-rRNA intermediates are also produced in *P. furiosus.* Although we could detect a similar total number of circular pre-rRNA transcripts for *H. volcanii* and *P. furiosus*, the proportion of this category with respect to all reads mapping to the 16S rRNA is very low, which might reflect the actual abundance (Figure 23e). However, considering the differing sequencing efficiencies of the 16S rRNAs absolute quantifications cannot be made.

The early RNA chimera 16S rRNA leading/trailing sequence-tRNA-23S rRNA trailing sequence precursor (T1) likely 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 (Tang et al., 2002) (Figure 23, see Supplementary Figure 16l). 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 precursor (P1) carrying the leading sequence in combination with tRNA-23S rRNA is arising from an experimental artifact (Figure 23, see Supplementary Figure 16). In fact, this variant is in good agreement with our recent *cis*-acting element analysis in *H. volcanii* (Jüttner et al., 2020) (see also Discussion).

Taken together, our analysis confirms and expands the number of putative pre-rRNA intermediates in archaea. Moreover, this extended framework provides an additional basis to facilitate further definition of common and specific principles of rRNA maturation in archaea.

4.8. Towards mapping of RNA base modifications

More than 160 types of modified bases have been described in RNAs so far (Boccaletto et al., 2018). 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 nucleotide sequence (Liu et al., 2019; Smith et al., 2019; Wongsurawat et al., 2019). In turn, these signal changes might also lead to differences in the basecalling profiles (e.g. systematic errors or a drop in basecalling quality). Approaches based on signal deviations or basecalling errors have already been applied to map RNA and DNA modifications in different organisms. However, accurate *de novo* RNA modification prediction with single-nucleotide resolution is still challenging as more than one base affects the current through the pore (see Figure 24a). In addition, current deviation is influenced by the type of modification and the surrounding sequence context (Rand et al., 2017).

Despite these limitations, we aimed to study rRNA modifications in archaea and explored different analysis strategies (Figure 24). Based on the approaches mentioned above, we benchmarked the potential to detect known and putative modification sites in the 16S rRNA. We focused first on the 16S rRNA modifications of *P. furiosus*, using the recently established 16S rRNA modification pattern in the close relative *Pyrococcus abyssi* (Coureux et al., 2020). This set includes 34 N^4-acetylcytidines, and 10 other modifications of diverse types (Coureux et al., 2020). Compared to a background set consisting of all other positions in the *P. furiosus* 16S rRNA, we observed that the surrounding sequencing context of all modified bases is significantly enriched in basecalling errors (Figure 24b) and also had a comparatively low mapping quality (Figure 24c). Depending on the type of modification (acetylation vs. diverse), these metrics looked very different across a sequence context from -5 to +5 from the exact position of the modified base (Figure 24d, see Supplementary Figure 19a,b). While N^4-acetylcytidines were mostly miscalled at their predicted position, the other diverse base modifications had various effects on all metrics (Figure 31b, see Supplementary Figure 19a,b). Using the position information derived from *P. abyssi*, our analysis suggest that the putative N^4-acetylcytidine modification leads to a wrong base assignment during basecalling (Figure 24d,e). In fact, this systematic non-random error was also reflected by the high proportion of a central T instead of C in the CCG context of the acetylation (see Supplementary Figure 19c,d).

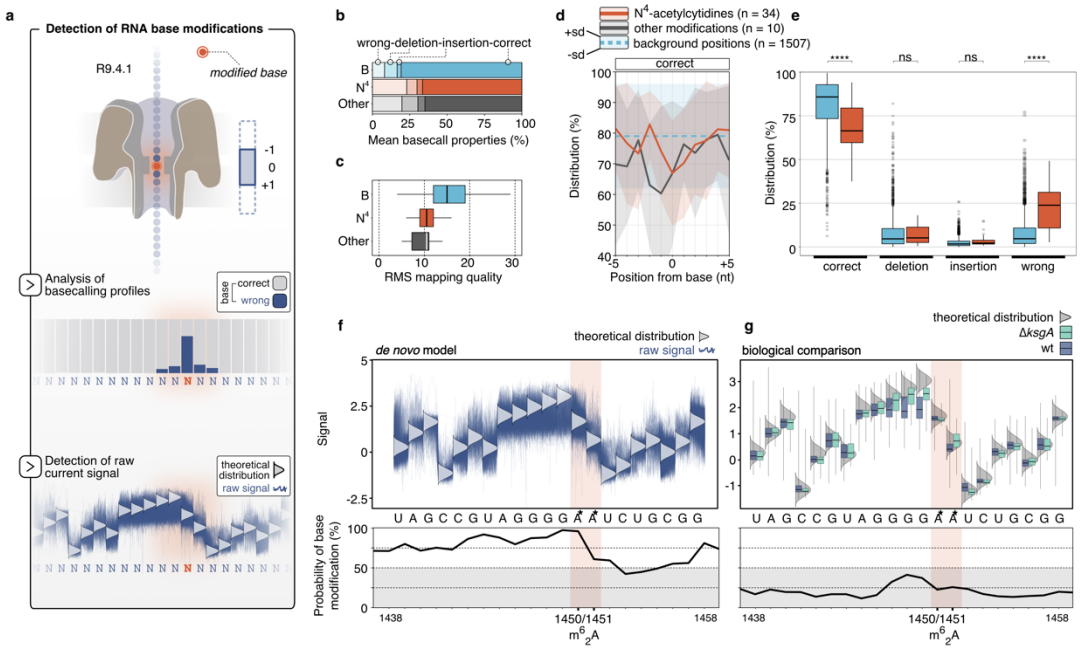


Figure 24 | Detection of RNA base modifications in archaeal 16S rRNA based on basecalling and raw signal profiles. a, In the R9.4.1 pore more than one base affects the current through the pore. RNA base modifications can be predicted as the modification might alter the raw signal which can in turn influence the basecalling profile. The performance of the two strategies was evaluated for potentially present N^4-acetylations in *P. furiosus* (**b-e**, basecalling properties) and the KsgA/Dim1-dependent dimethylation ($m^6{}_2A$) in *H. volcanii* (**f**, **g**, raw current signals). **b**, Analysis of mean basecall profiles of N^4-acetylated positions (red), diverse other modifications (grey) and all other positions of the 16S rRNA in *P. furiosus*. The properties (wrong, deletion, insertion, correct) are shown by different transparencies. **c**, The root mean square (RMS) mapping quality gives an estimation of the overall mapping quality and is shown for the defined position categories in b. **d**, The proportion of correct basecalls is shown in a window from -5 to +5 from the presumably modified/or background base. Shaded areas show the upper and lower standard deviation, while the lines show the mean values. **e**, Distribution of basecall properties for the 34 N^4-acetylcytidines. Statistical significance (p-values, T-test) is indicated by asterisks (p-value > 0.05: ns (not significant), p <= 0.0001: ****). **f**, 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 (Stoiber et al., 2016). The $m^6{}_2A$ modification at position 1450/1451 (from 16S start) is indicated by an asterisk 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. **g**, Position-specific boxplot comparison of signals from sequences surrounding the $m^6{}_2A$ modification in *H. volcanii* wildtype (blue) and the Δ*ksgA* 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.

To analyse whether these modifications are already established at early steps of rRNA maturation, we looked at the basecalling properties of 5´extended pre-rRNAs and compared them to mature 16S rRNA. Importantly, we did not observe significant basecalling errors in these selected precursor rRNAs (5´extended pre-16S RNA) indicating that the cytidine N^4-acetylation is not occurring early in the rRNA maturation pathway (see Supplementary Figure 19e) (see below for further details).

As the approach based on systematic errors gave us promising results for N^4-acetylations, but were less unambiguous for diverse modifications, we wanted to evaluate the potential to detect RNA base modification from raw signals using Tombo (Figure 24a, updated Tombo version from (Stoiber et al., 2016)). To this end, we first focused on the dimethylation ($m^6{}_2A$) introduced by the enzyme KsgA/Dim1 at position A1450/A1451 in *H. volcanii* (A1518/A1519 *E. coli* numbering) (O'Farrell et al., 2006; Grosjean et al., 2008). Using the *de novo* model in Tombo the calculated probability of a modification was very high for the stretch of Guanosines adjacent to position A1450 (Figure 24f). Mapping to single-nucleotide resolution is difficult as more than one base contributes to the actual electric current signal in the nanopore (Rang et al., 2018). In the next step, a comparison of a wildtype sample to a deletion mutant of archaeal KsgA/Dim1 homologue helped us to confirm that the current signal alteration in this particular region is dependent on the KsgA/Dim1 $m^6{}_2A$ modifications and not the homopolymer-G-stretch (Figure 24g). 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 and highlights the benefits of a background read model.

Assuming that early pre-rRNAs represent a state where base modifications are not yet quantitatively introduced, we used these reads as a background model to explore the potential to detect the introduction of base modifications at different stages of rRNA maturation in archaea. Therefore, we generated multiple sets by sorting reads according to the main classes of pre-rRNA intermediates described above for *H. volcanii* and *P. furiosus* (Figure 25a). For *H. volcanii* and *P. furiosus*, we compared 5´extended 16S rRNA, circular pre-16S rRNA, opened circular-pre-16S rRNA and mature 16S rRNA. We first focussed on analysing the $m^6{}_2A$ and N^4-acetylcytidine signatures across the putative different stage of rRNA maturation in archaea. To this end, we evaluated basecalling and raw signal profiles using the 5´ extended pre-rRNA as a background model as the former performed well for N^4-acetylcytidines, while the latter for $m^6{}_2A$ detection.

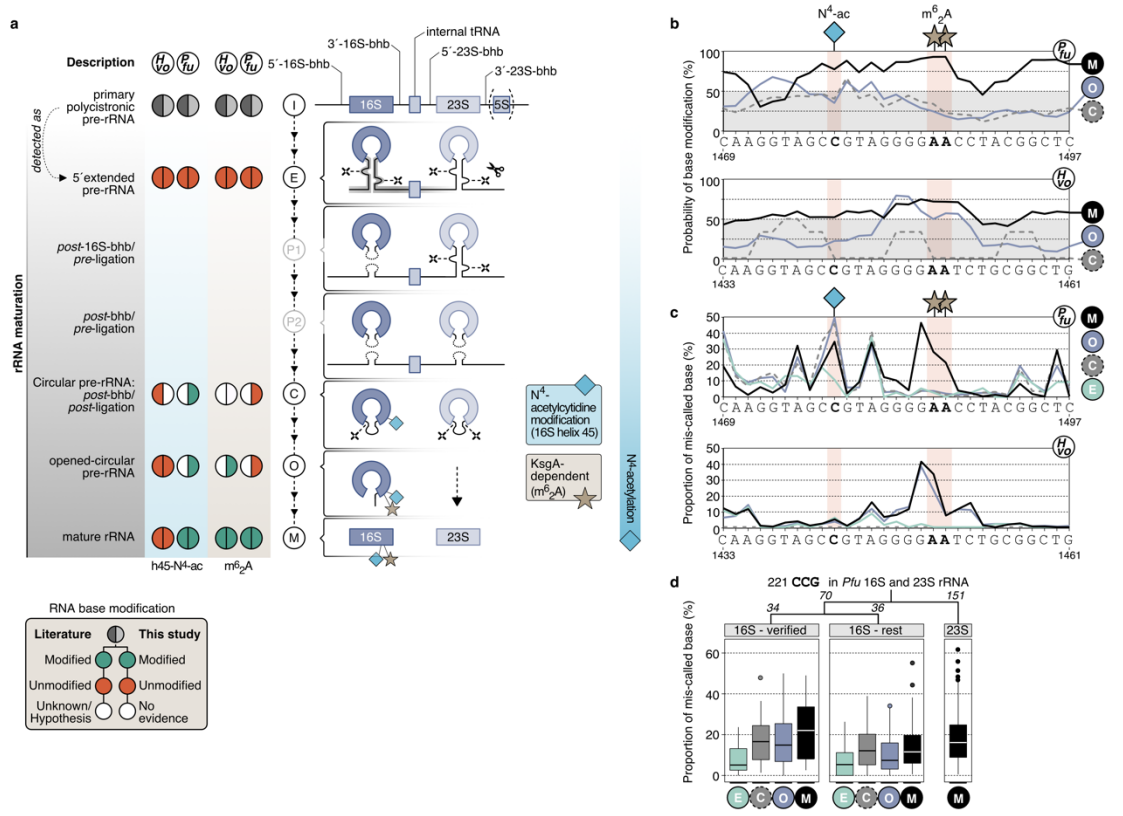


Figure 25 | Detection of RNA base modifications at different stages of rRNA maturation in archaea. a, During the maturation of ribosomal rRNAs base modifications are introduced at different time points. While the the KsgA/Dim1-dependent dimethlyation (m^6_2A) is proposed to function as a quality control during late biogenesis (mature and open-circular pre-rRNA), N^4-acetylations in *P. furiosus* seem to be added successively during rRNA maturation. **b**, Probability of base modifications calculated by the tombo sample-compare approach using 5´-extended pre-rRNAs as a background model for *P. furiosus* (upper panel) and *H. volcanii* (lower panel). The approach was applied to mature (M, black), open-circular (O, purple) and circular pre-rRNAs (C, grey). The illustration shows the sequence region of the 16S rRNA containing the N^4-acetylcytidine modification in helix 45 and the dimethylation. **c**, The basecalling profile of the same section was analyed during the maturation stages. The proportion of the respective base to be mis-called (category wrong) is shown for categories M, O, C and 5´-extended pre-rRNAs (E, green). **d**, Comparison of the proportion of mis-called based for all CCGs detected in the 16S and 23S rRNA in *P. furiosus*. 34 positions have been experimentally verified in *P. abyssi* and are potentially also present in *P. furiosus* (Coureux et al., 2020).

Basecalling anomaly in the m^6_2A region was detected within the mature 16S rRNA of both *P. furiosus* and *H. volcanii*, and could be confirmed by a high probability of base modification around this positions using the tombo model (Figure 25b,c, see Supplementary Figure 20). Interestingly, similar profiles were detected for the putative opened-circ-pre-16S rRNA in *H. volcanii*, but not in *P. furiosus*. This finding is in line with a proposed "quality control" function of KsgA/Dim1 during late biogenesis of the small ribosomal subunit (Lafontaine et al., 1998; Xu et al., 2008; Strunk et al., 2011). Similarly, we analysed N^4-acetylcytidine known to occur in the vicinity of the m^6_2A. However, this modification occurs prior to the KsgA/Dim1-dependent modification during eukaryotic ribosome

biogenesis (Ito et al., 2014a; Sharma et al., 2015; Iost et al., 2019; Sleiman and Dragon, 2019). In agreement with previous analysis (Grosjean et al., 2008), no apparent N^4-acetylcytidine modification was observed at the equivalent position in *H. volcanii* (Figure 25c). In contrast, an increase in base-calling errors at the expected position was observed in circular pre-rRNAs and mature 16S rRNA, but not within 5´extended pre-16S rRNA in *P. furiosus*. Therefore, and similarly in eukaryotes, N^4-acetylcytidine modification in helix 45 precede the KsgA/Dim1-dependent $m^6{}_2A$ modifications in *P. furiosus*. Note that the results for the circular pre-16S rRNA (grey, Figure 25) should be taken with care in *H. volcanii*, given the limited number of reads in this category (see Supplementary Figure 20). However, it is tempting to speculate that a different timing of events in *H. volcanii*, observed by the earlier $m^6{}_2A$ modifications, is caused by the absence of N^4-acetylations.

To further extend on the timely order of all N^4-acetylcytidine modification potentially present in *P. furiosus*, we analysed the basecalling profiles across different rRNA maturation stages (Figure 25d). In addition, we compared it to all other CCGs that are present in 16S and 23S rRNA as N^4-acetylation have been shown to be introduced in a CCG context in *P. abyssi* (Coureux et al., 2020). This analysis suggests that (i) N^4-acetylcytidine modifications may be also scattered across the 23S rRNA sequence, and (ii) these modifications are established in the course of pre-rRNA maturation (circ-pre-rRNA in Figure 25d).

Taken together, our analysis, suggests that despite the current limitations, ONT allows to discriminate (some) rRNA modifications across selected rRNA maturation events. Moreover, these data support the long-standing hypothesis that hyperthermophilic organisms might stabilize their rRNAs by a higher degree of RNA modifications (Dennis et al., 2015; Gomes-Filho et al., 2019).

5. Discussion

Performing whole-transcriptome 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.

5.1. Insights into transcriptional processes

Bacterial and archaeal transcription is an intensely studied molecular process and the mechanisms of basal transcription are well understood (Ebright et al., 2019). 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 a large number of small non-coding RNAs (Babski et al., 2014). However, many regulatory events that involve the regulation via small RNAs take place at bacterial 5' UTRs (Oliva et al., 2015). We and others found that 5' UTRs are significantly shorter in many archaea supporting the idea that post-transcriptional regulation is mediated via the 3' rather than the 5' UTR in these groups (Ren et al., 2017). 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 studied, poly(U) stretches were overrepresented at termination sites agreeing well with termination sequences found in *Sulfolobus* and *Methanosarcina* (Dar et al., 2016). 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* (Santangelo et al., 2009; Hirtreiter et al., 2010a, 2010b). 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 (Berkemer et al., 2020b). Notably, the *H. volcanii* genome is distinguised by a high GC content leading to a low propability for the occurence of U_5 stretches and hence, the transcription machinery might have adapted to recognise U_4 stretches as termination signal. 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 consecutive poly(U) stretches. Possibly, *Haloferax* relies on additional termination signals or yet unknown termination factors. A putative candidate is archaeal CPSF1 (aCPSF1, also known as FttA), a recently described archaeal termination factor (Yue et al., 2019; Sanders

et al., 2020) that is widespread in archaea. aCPSF1 acts as ribonuclease that was shown to cleave transcripts after a poly(U) stretch to trim transcripts and facilitates transcription termination in *Thermococcus kodakarensis* (Sanders et al., 2020) and *Methanococcus maripaludis* (Yue et al., 2019). The arising 3' UTR isoforms were detected using Term-seq analysis (Yue et al., 2019). We also observed heterogeneity in the case of the Pilin and histone transcripts, respectively, that are distinguished by varying lengths of the 3' UTR suggesting that aCPSF1 might trim a subset of genes in *H. volcanii* and *P. furiosus*. It is noteworthy that 3' UTR isoforms were also detected in Term-seq studies with *Sulfolobus* and *Methanosarcina* (Dar et al., 2016). However, in contrast to the Pilin, Alba and histone transcripts, the 3' UTR isoforms arise from termination at different sites of a single continues poly(U) stretch suggesting that the isoforms arise from a stochastic termination process of the RNA polymerases at an extended poly(U) stretch at the end of the gene. The gradual termination observed in this study might also be influenced by the coupling of transcription and translation. These genes are all highly expressed and translated. Hence, it seems plausible that the ribosome is efficiently coupled to the RNAP (French et al., 2007) (as observed in bacteria (O'Reilly et al., 2020; Webster et al., 2020)). Several studies in bacteria showed that the ribosome influences transcription (and *vice versa*) (Vogel and Jensen, 1994; Proshkin et al., 2010; Stevenson-Jones et al., 2020). The stochastic termination might therefore be a result of the uncoupling of the ribosome at the end of the mRNA potentially also inducing the dissociation of the transcription elongation complex. Taken together, these data suggest that a variety of termination mechanisms (that can even co-occur in the same cell) can be found in archaea ranging from stochastic intrinsic termination at extended poly(U) stretches (*Pyrococcus, Sulfolobus, Methanosarcina*), to abrupt termination at short uridine tracts (*H. volcanii*) and factor-dependent termination that results in trimming of the 3'UTR (*H. volcanii, P. furiosus, M. maripaludis, T. kodakarensis*).

In the context of transcription, the long and overlapping native RNA reads helped us to analyse the transcriptional landscape at multigene operons. More specifically, we focused on the archaeal flaggelum (archaellum) operon, encoding for the archaeal motility machinery (Albers and Jarrell, 2015), as the transcription unit assignment remained ambiguous so far (Näther-Schindler et al., 2014). In contrast to bioinformatical and Illumina RNA-seq-based predictions and attempts to unravel the TU via primer extension experiments, we found that the archaellum operon in *P. furiosus* is transcribed in multiple units with highly diverse abundances. The *fla*B0 gene encodes the major archaellin/flagellin protein that forms the filament of the archaellum and therefore, the organism has to produce this protein in large quantities as apparent from the expression level (Näther-Schindler et al., 2014). Interestingly, FlaD mRNA is expressed at comparably high levels as well supporting the idea that FlaD is a major constituent of the archaellum in *P. furiosus*. It has been speculated that FlaD forms the cytosolic ring of the archaellum that

anchors the filament in the outer membrane (Daum et al., 2017). The identity and functional role of FlaD are, however, not known so far.

5.2. Insights into rRNA processing in archaea

In this study, we have assessed the suitability of native RNA sequencing to obtain information on the rRNA maturation pathway of different prokaryotes. Ribosomal RNA processing proceeds via the coordinated and defined order of ribonucleases action (exonucleolytic and/or endonucleolytic cleavages) which generate pre-rRNA intermediates with defined premature rRNA sequences (Venema and Tollervey, 1995; Deutscher, 2015; Henras et al., 2015; Bechhofer and Deutscher, 2019). The establishment of faithfull rRNA maturation maps in model organisms, like *E. coli*, *S. cerevisiae* or human cell culture has required numerous analyses over the past decades (Venema and Tollervey, 1995; Deutscher, 2015; Henras et al., 2015; Bechhofer and Deutscher, 2019), and remains a technical challenge. Therefore, methodologies that might accelerate the systematic analysis of rRNA maturation pathways across the tree of life, thereby enabling to unravel the diversity of rRNA processing strategies need to be established. Beyond the identification of processing sites, the order of the processing events which can be, in part, deduced from co-occurrence analysis of the 5´ and 3´ extremities is of biological relevance (Venema and Tollervey, 1995; Deutscher, 2015; Henras et al., 2015; Bechhofer and Deutscher, 2019). Whereas we could confirm and extend our general view on the rRNA maturation pathway in archaea, the 3´-5´ processivity of Nanopore native RNA sequencing observed for rRNA and the potential RNA degradation during sample preparation impedes the accurate quantitative analysis of pre-rRNA extremities co-segregation (see Figure 23 and Supplementary Figure 11b). Nevertheless, we could, in most of the cases, confirm and expand the presence of pre-rRNA intermediates and processing sites in the different organisms analysed, including the archaeal specific circular-pre-rRNA intermediates (Tang et al., 2002; Danan et al., 2012; Ferreira-Cerca, 2017; Jüttner et al., 2020; Qi et al., 2020) (see discussion below). Together our findings are summarized into an updated archaeal rRNA processing model described in figure 5 and are discussed below.

The full length theoretical primary rRNA transcript was not identified in any of the archaeal organisms analysed. Similarly, this primary rRNA is generally difficult to observe in wildtype *E. coli* ((Nikolaev et al., 1973; Hofmann and Miller, 1977) and this work). Collectively, these observations suggest that short-lived and/or low abundant pre-rRNA intermediates escape the detection capacity of the current experimental set-up. Accordingly, it is also difficult to infer differences in rRNA processing features between different (archaeal) organisms by virtue of observed pre-rRNA intermediates absence/presence pattern. In fact, these differences may also be related to organism-specific changes in pre-rRNA intermediates relative levels, which will depend on the sum of the reaction kinetics of the different maturation steps in a given condition.

Among the identified pre-rRNA intermediates, the *post*-16S-bhb/*pre*-ligation precursor (P1), which is observed in *P. furiosus* and includes ligation at the bhb motif of the upstream region of the 16S leader and downstream region of the 16S trailer sequences and continues to the downstream tRNA/23S sequences, is of particular interest (see Supplementary Figure 16k). The presence of this ligation event suggests that the 16S rRNA bulge-helix-bulge processing occurs prior to internal tRNA and 23S rRNA maturation. Although, this ligation event was not identified by ONT in *H. volcanii*, this observation is in agreement with our recent functional *cis*-acting element analysis performed in *H. volcanii* (Tang et al., 2002; Ferreira-Cerca, 2017; Jüttner et al., 2020). In fact, based on this previous analysis we have proposed a model by which 16S rRNA maturation proceeds and is required for the downstream maturation of the internal tRNA and 23S rRNA. Moreover, we have hypothesized that ligation of the 16S rRNA leader/trailer resulting from the 16S rRNA bulge-helix-bulge maturation process generates a putative new pre-rRNA intermediate for which the corresponding ligation event could be observed in *Pyrococcus furiosus* using native RNA sequencing (Jüttner et al., 2020). In addition, the presence of an RNA chimera containing the leading/trailing/tRNA parts (T1) (*post*-bhb/*post*-ligation) 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 which liberate the circular pre-16S and pre-23S rRNAs (C) (suggested in (Tang et al., 2002; Jüttner et al., 2020), and this work). The presence of circular pre-16S and pre-23S rRNAs and their processing sites could be verified and established in *H. volcanii* and *P. furiosus*, respectively (Figure 23, see Supplementary Figure 17, Supplementary Figure 18). Recently, we determined the functional requirement of the bulge-helix-bulge motifs for the formation of circ-pre-rRNAs in *H. volcanii*. Moreover, in analogy to intron containing-tRNA splicing, the rRNA bhb motifs are presumably cleaved by the tRNA splicing endonuclease (endA) prior to covalent circularization (Clouet-D'Orval et al., 2018; Jüttner et al., 2020; Qi et al., 2020). Although intact circular RNA cannot be directly sequenced by ONT, we noticed the presence of permuted transcript in *H. volcanii* corresponding to the ligation events previously identified for circ-pre-rRNAs in *H. volcanii* (Jüttner et al., 2020). Most of these permuted reads were also showing random and heterogenous 5´and 3´ends thereby suggesting that these pre-rRNAs were likely the result of randomly nicking of circular pre-rRNA intermediates during sample preparation (see Supplementary Figure 17). Noteworthy, similar permuted reads were observed in *P. furiosus*, for which the presence of circular-pre-rRNA intermediates is not established thus far. Whereas, the observed ligation could be accurately mapped to the predicted 16S bhb motif, the 23S bhb motif could not be accurately predicted (data not shown). However, our manual inspection suggests that the permuted reads extremities match to an imperfect, presumably less stable, bhb motif within the 23S processing stem (see Supplementary Figure 18b). This property is reminiscent to the "aberrant" 16S bhb motif used for circular-pre-16S rRNA formation

in *S. acidocaldarius* (Durovic and Dennis, 1994; Russell et al., 1999; Danan et al., 2012; Jüttner et al., 2020). Whether these structural features are stabilized by additional factors or enable a certain degree of regulation during the rRNA maturation process in the cellular context is unknown.

In addition to the circular pre-rRNAs, we observed pre-rRNA intermediates cleaved at the bhb motifs but not yet ligated into circular pre-rRNA in *H. volcanii (post*-bhb*/pre*-ligation pre-rRNAs*)* (P2). Whereas, the presence of this intermediate processing step is theoretically expected, they were only detectable in *H. volcanii* (see Figure 23a, see Supplementary Figure 16), suggesting that the maturation kinetics or stability of these pre-rRNA intermediates varies among these organisms.

How the circular pre-rRNAs are further processed into linear mature rRNA is not well understood. Based on our current knowledge, several non-mutually exclusive hypothesis can be drawn: (i) opening of the circular-pre-rRNA within the ligated spacer region and subsequent maturation of the 5´and 3´end; (ii) opening of the circular pre-rRNA by first maturation of the 5´end mature rRNA followed by 3´end maturation; or (iii) opening of the circular pre-rRNA by first maturation of the 3´end mature rRNA followed by 5´end maturation. A category of putative 16S pre-rRNA intermediates observed in *H. volcanii*, may provide some indications how linearization of the circular pre-16S rRNA is achieved. In fact, this particular intermediate was extended in its 3´end by the presence of the ligated 5´and 3´spacers normally observed in the circular pre-16S rRNAs and this 3´extension consistently ended just prior to the 16S 5´mature ends. This particular configuration is suggestive of 5´end maturation of circular-pre-16S rRNA prior to final 3´end maturation, thereby generating opened-circular pre-16S intermediates (O) (see Supplementary Figure 17b). Although the majority of opened-circular pre-16S rRNAs is degraded from its 5´end, we detected a subset representing the theoretical full length (see Supplementary Figure 17b,c). Additional properties of this putative intermediate is in agreement with its positioning during rRNA maturation (see below discussion on rRNA modifications) and with the prevalence of 16S rRNA 5´maturation prior to its 3´end previously observed in bacteria and eukaryotes (Henras et al., 2015; Bechhofer and Deutscher, 2019). Overall, future functional charaterization of the *cis*- and *trans*-acting elements required for pre-rRNAs maturation will be necessary to further refine our view on archaeal rRNA processing.

In conclusion, despite some intrinsic limitations, we provide evidence that direct RNA sequencing technologies can be a useful tool to approach intricated maturation pathway like rRNA maturation, and expand our understanding of RNA maturation in prokaryotes.

5.3. Towards the mapping of rRNA modification patterns

RNA modifications have been described already in the 50-60s, and have gained significant attention over the last years, under the generic term of the epitranscriptome (Littlefield and Dunn, 1958; Smith and Dunn, 1959; Li and Mason, 2014). The high-throughput analysis of these post-transcriptional modifications remains challenging and mostly relies on indirect methods, like primer extension stops analysis and/or chemical recoding/derivation strategies (Schwartz and Motorin, 2017; Sas-Chen and Schwartz, 2019). 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 analyses have explored different strategies to evaluate the capacity of ONT to accurately detect RNA modifications (e.g. m^6A) (Garalde et al., 2018; Leger et al., 2019; Liu et al., 2019; Smith et al., 2019; Lorenz et al., 2020).

RNA modifications can lead to electric current signals varying from the theoretical signal expected for the unmodified canonical ribonucleotides. These properties can be harnessed, on the one hand, to predict RNA modification probability by comparing theoretical and experimental electric current signal distribution, and on the other hand, variation of electric signals may increase the rate of basecalling errors. In both cases, the comparison of the profiles to a background distribution of non-modified nucleotides leads to a significant reduction of false-positives. We evaluated the potential to use early rRNA precursors, which are expected to contain incomplete modification patterns, as a background model and applied this strategy to analyse different stages of rRNA maturation.

To validate our approach, we first focussed on two types of modification occurring in close proximity in helix 45 of the 16S/18S rRNA, but at distinct stages of rRNA maturation, namely the almost universally conserved KsgA-dependent dimethylations ($m^6{}_2A$) and the less conserved Kre33/Nat10-dependent N^4-cytidine acetylation (O'Farrell et al., 2006; Grosjean et al., 2008; Xu et al., 2008; Ebersberger et al., 2014; Smith et al., 2019). By analysing basecalling profiles and raw signals in wt 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 in both *H. volcanii* and *P. furiosus*, and may predominantly occur after linearization of the circular-pre-16S rRNA. In contrast, helix 45 N^4-cytidine acetylation, which is absent in *H. volcanii*, appears to be added at the circular-pre-16S rRNA stage, prior to completion of the KsgA-dependent modifications in *P. furiosus* (Figure 25). These results are in good agreement with previous studies done in eukaryotes and bacteria (Lafontaine et al., 1998; Xu et al., 2008; Strunk et al., 2011; Ito et al., 2014b, 2014a; Sharma et al., 2015, 2017; Sleiman and Dragon, 2019). Moreover, expanding 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 (Grosjean et al., 2008; Dennis et al., 2015; Gomes-Filho et al., 2019). Recently, it has been shown that *P. abyssi* 16S rRNA is heavily acetylated at CCG motifs (Coureux et al., 2020). Our analysis suggests that N^4-acetylcytidine modifications (i) increases the rate of basecalling errors (e.g. C>T) at the expected modified residue, (ii) are distributed across the 16S and 23S rRNA sequences in *P. furiosus*, and (iii) are successively added during rRNA maturation to reach "completeness" in the mature rRNAs. Future studies will be necessary to decipher, how widespread this type of modification is among archaea, and to evaluate their contribution for ribosomal subunit biogenesis and function in the cellular context.

Whereas ONT may facilitate RNA modification analysis in general, the exact chemical nature of these modifications can not be unveiled without prior knowledge and remain a challenging task which greatly benefits of the use of unmodified/hypo-modified references (in agreement with recent studies (Leger et al., 2019; Smith et al., 2019)). To facilitate high-throughput identification of RNA (DNA) modifications, future studies will required to develop and train algorithms improving the *de novo* identification confidence of diverse RNA/DNA modifications.

5.4. 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 need for a enzymatic poly-adenylation step of non polyA+ RNA, (iii) the 3' bias during RNA sequencing (iv) limited throughput and (v) 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 modification 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.

Data availability

Raw sequencing data sets (gzipped raw FAST5 files) have been deposited in the Sequence Read Archive (SRA) and are available under project accession number PRJNA632538.

Code availability

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

Author contributions

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

Acknowledgements

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

CopR, a global regulator of transcription to maintain copper homeostasis in *Pyrococcus furiosus*

Felix Grünberger[1], Robert Reichelt[1], Ingrid Waege[1], Verena Ned[1], Korbinian Bronner[1], Marcell Kaljanac[2], Nina Weber[1], Zubeir El Ahmad[1], Lena Knauss[1], M. Gregor Madej[2], Christine Ziegler[2], Dina Grohmann[1], Winfried Hausner[1*]

[1]Institute of Microbiology and Archaea Centre, University of Regensburg, Universitätsstraße 31, 93053 Regensburg, Germany

[2]Department of Structural Biology, Institute of Biophysics and Physical Biochemistry, University of Regensburg, Universitätsstraße 31, 93053 Regensburg, Germany

*Correspondence: Winfried Hausner. E-mail: Winfried.Hausner@ur.de

Keywords: Archaea, transcription, Pyrococcus, CopR, copper-regulation

1. Abstract

Although copper is in many cases an essential micronutrient for cellular life, higher concentrations are toxic. Therefore, all living cells have developed strategies to maintain copper homeostasis. In this manuscript, we have analysed the transcriptome-wide response of *Pyrococcus furiosus* to increased copper concentrations and described the essential role of the putative copper-sensing metalloregulator CopR in the detoxification process.

To this end, we employed biochemical and biophysical methods to characterise the role of CopR. Additionally, a *copR* knockout strain revealed an amplified sensitivity in comparison to the parental strain towards increased copper levels, which designates an essential role of CopR for copper homeostasis. To learn more about the CopR-regulated gene network, we performed differential gene expression and ChIP-seq analysis under normal and 20 µM copper-shock conditions. By integrating the transcriptome and genome-wide binding data, we found that CopR binds to the upstream regions of many copper-induced genes. Negative-stain transmission electron microscopy and 2D class averaging revealed an octameric assembly formed from a tetramer of dimers for CopR, similar to published crystal structures from the Lrp family. In conclusion, we propose a model for CopR-regulated transcription and highlight the complex regulatory network that enables *Pyrococcus* to respond to increased copper concentrations.

2. Introduction

The archaeal transcription system combines strategies and regulatory mechanisms known from eukaryotic as well as from bacterial species (Werner and Grohmann, 2011; Peeters et al., 2013). Archaea rely on a single RNA polymerase that synthesises all RNA species in the cell and is highly homologous to the eukaryotic RNA polymerase II (Grohmann and Werner, 2011). The presence of general transcription initiation factors (TATA box binding protein, Transcription factor B, Transcription factor E) and defined promotor elements (B recognition element, TATA box, initially melted region, initiation site) stresses the close relationship to eukaryotes especially for transcription initiation (Blombach et al., 2016). In contrast, the fine-tuning process of gene expression is mainly achieved by bacterial-like transcriptional regulators (Lemmens et al., 2019). Positive or negative regulation is mediated by the binding of these transcription factors (TFs) to promoter regions of specific genes.

The genome of the hyperthermophilic euryarchaeon *Pyrococcus furiosus* contains a total number of 86 putative DNA-binding TFs. However, the exact function of most of these factors, which represent about 4 % of all open reading frames (ORFs), is unknown (Denis et al., 2018). In an attempt to close that knowledge gap, functional and structural aspects of some of these TFs have been analysed over the last two decades. While the regulation of sugar or sulfur metabolism and other changing environmental conditions have been studied in detail, the underlying mechanisms to maintain metal homeostasis are only poorly understood (Vierke et al., 2003; Lipscomb et al., 2009; Yang et al., 2010; Gindner et al., 2014; Karr, 2014).

Playing an essential role in the cycling of elements, Archaea not only have to transform and make use of a variety of metals but also have to withstand elevated levels in the respective habitat (Bini, 2010). For many organisms, copper (Cu) is one of the essential trace elements used as a cofactor in a variety of proteins. These are mainly involved in electron transfers due to the ability to undergo redox changes from the reduced form Cu^{+} to the oxidised Cu^{2+}. Despite its essential role, high intracellular concentrations are toxic for prokaryotic and eukaryotic cells. Copper catalyses the conversion of H_2O_2 to hydroxyl radicals via the Fenton reaction, which leads to oxidative damage of nucleic acids, proteins and lipids (Gunther et al., 1995; Pham et al., 2013). Cu^{+} also is a strong soft metal and can attack and destroy iron-sulfur proteins either by direct interaction or by blocking iron-sulfur cluster biogenesis (Macomber and Imlay, 2009; Tan et al., 2017). To prevent cellular damage, all cells have developed various copper detoxification strategies. In prokaryotes, this is mainly achieved by active export of copper ions and in rarer cases by sequestration or exclusion (Bini, 2010; Martínez-Bussenius et al., 2017).

The two subfamilies of ATPases, P_{1B-1} (CopA) and P_{1B-3} (CopB), are the key players in cellular copper export. CopA transports Cu^{+}, and CopB is proposed to transport Cu^{2+} (Mana-Capelli et al., 2003; Meloni et al., 2014). To elucidate the mechanism of the

exporting enzymes, the structures of homologous archaeal Cu-transporting ATPases CopA and CopB were studied in the hyperthermophilic *Archaeoglobus fulgidus*. The two enzymes seem to have different affinities for Cu^{+} and Cu^{2+} (Mandal et al., 2002; Tsuda and Toyoshima, 2009; Agarwal et al., 2010). Recent data, however, suggest that both subclasses, P_{1B-1} and P_{1B-3}, have to be assigned as Cu^{+} transporters, which is consistent with the presence of only Cu^{+} in the reducing environment of the cytoplasm (Purohit et al., 2018). Furthermore, a corresponding metallochaperone of the CopZ family is capable of reducing Cu^{2+} to Cu^{+} and is most likely involved in the transport of the reduced ion to CopA (Sazinsky et al., 2007).

Many Archaea use the metallochaperone CopM, which contains a TRASH- instead of a heavy-metal-associated (HMA)-domain of the CopZ family (Ettema et al., 2006). TRASH is a novel domain that has been proposed to be uniquely involved in metal-binding in sensors, transporters and trafficking proteins in prokaryotes (Ettema et al., 2003). In addition to the specific binding of copper by chaperons, copper can also be buffered by small peptides like GSH and other reducing agents, to prevent cellular damage (Rensing and McDevitt, 2013).

In several Archaea, the Cu-transporting ATPase and the copper chaperone are arranged in a conserved copper resistance gene cluster (*cop*), which also contains an additional gene, encoding for a DNA-binding transcriptional regulator. In previous studies, PF0739 has been bioinformatically predicted to be the copper-dependent regulator CopR in *P. furiosus* (Ettema et al., 2006; Villafane et al., 2011; Hong et al., 2019). Based on biochemical data, *in vitro* analysis and growth experiments using knockout strains, CopR was proposed to play opposing regulatory roles in different Archaea: While in *Thermococcus onnurineus* the transcriptional regulator (TON_0836) represses *copA*, both transporter and chaperone are activated in *Saccharolobus solfataricus* (SSO2652) (Villafane et al., 2011; Hong et al., 2019).

Here, we have characterised the metal-sensing transcriptional regulator CopR in *P. furiosus*. First, we described the influence of different metal ions on the DNA-binding ability of CopR to the shared *copR/copA* promoter and analysed the growth of parental and *copR* knockout strains under increasing copper levels. We further performed a differential gene expression analysis (DGE) and chromatin immunoprecipitation with high-throughput sequencing (ChIP-seq) under normal and copper-shock conditions to elucidate the CopR-regulated gene network in *P. furiosus*. Integrating the genome-wide results with a more in-depth functional and structural characterisation, we propose that CopR acts as a dual regulator to maintain copper homeostasis.

3. Material and Methods

3.1. Strains, Plasmids and Primers

All strains, plasmids and primers used in the study are listed in Supplementary Table 12.

3.2. Construction of the *copR* deletion strain

For the construction of *P. furiosus* parental strain MURPf52 and Δ*copR* strain MURPf74 a modified genetic system was developed for *P. furiosus* DSM3638 allowing markerless disruption of genes onto the chromosome. This system is based on selection via agmatine-auxotrophy and counter selection via 6-methylpurine as described for *P. furiosus* COM1 strain and *T. kodakarensis* (Santangelo et al., 2010; Lipscomb et al., 2011).

First, for disruption of the *Pyrococcus pdaD* gene (PF1623; arginine decarboxylase gene) via a double-crossover event plasmid pMUR264 was constructed according to Kreuzer et al. (Kreuzer et al., 2013). The first fusion PCR product containing upstream and downstream regions flanking the *Pf pdaD* gene encoding a Pyruvoyl-dependent arginine decarboxylase was created using the following two primer pairs: (Pf1622_fP_AscI/Pf1622_rP) and (Pf1624_fP/Pf1624_rP_NotI). The second fusion PCR product consisted of a two-gene resistance cassette which was needed for the selection-counter-selection system. The resistance cassette contained a *gdh* promoter, the *hmgCoA* reductase from *T. kodakarensis*, the region coding for the *xgprt* (PF1950, Xanthine-guanine phosphoribosyltransferase) and the histone A1 terminator sequence of *P. furiosus* (Waege et al., 2010). The first part was amplified using the primers: (SimV_NotI_F/SimV_Rv). For the second part, the primer pair: (Pf1624_fP_fus_2/Pf1624_rP_SbfI_N) was used. Both PCR products were combined with single-overlap extension PCR, ligated using a NotI restriction site and inserted into a modified pUC19 vector (Kreuzer et al., 2013) using AscI and SbfI restriction sites.

However, all attempts to markerless delete the *Pf pdaD* gene using this construct were not successful and thus pMUR264 was modified to allow gene disruption via a single crossover event. To remove the second homologous downstream region, plasmid pMUR264 was amplified using the primer pair: (pUC19_SbfI_F/Pf1950_SbfI_R). The resulting PCR product was digested by SbfI and ligated. This plasmid was denoted as pMUR242 and used for transformation of *P. furiosus* as described (Waege et al., 2010; Kreuzer et al., 2013). To obtain the markerless double mutant MUR37Pf, circular plasmid DNA of pMUR242 and strain MURPf27 (Kreuzer et al., 2013) were used and the corresponding transformants were selected with 10 µM simvastatin in SME-starch liquid medium supplemented with 8 mM agmatine sulfate at 85°C for 48 h. Pure cultures of the intermediate mutant MUR37Pf_i were obtained by plating the cells on solidified medium in the presence of 10 µM simvastatin and 8 mM agmatine sulfate. The integration of the

plasmid into the genome by single cross-over was verified by analyzing corresponding PCR products.

Cultures of the correct intermediate mutant were washed with medium under anaerobic conditions to remove the simvastatin. In detail, 1.5 ml of a grown culture were centrifuged in an anaerobic chamber for 4 min at 6,000 g and resuspended in fresh culture medium without simvastatin. This procedure was repeated three times. For the counter selection, the cultures were grown in the presence of 50 µM 6-methylpurine and 8 mM agmatine sulfate to induce a second homologous recombination step to recycle the selection marker and to eliminate integrated plasmid sequences. Pure cultures were obtained by single cell isolation using an optical tweezer (Huber et al., 1995). Cultures had to be grown in the presence of 8 mM agmatine sulfate and 8 mM Inosine and Guanine (I+G). The genotype of the final mutant was confirmed by PCR and Southern blot experiments.

For markerless disruption of the *Pf copR* gene (PF0739), plasmid pMUR527 was constructed. First a modified resistance cassette had to be designed, which was needed for the selection-counter-selection system. The two-gene resistance cassette contained the *Pf pdaD* gene including the promoter and terminator region sequence of *P. furiosus* which was amplified using the primers: (PF1623F_Pr_BHI/PF1623R_Term). For the second part, the three primer pairs: (F_PF1950_P_F_Fu/R_PF1950_Prom, F_PF1950_Fs_P/PF1950_R and F_PF1950_Fus_T/PF1950_T_R_BHI) were used to amplify the promoter, coding and terminator region of the *Pf* xgprt gene. The four PCR products were combined with single-overlap extension PCR and subcloned into pUC19 vector via the SmaI restriction site. In the next step it was cloned via NotI and SbfI restriction sites into plasmid pMUR47 (Kreuzer et al., 2013). The upstream and downstream flanking regions of the Pf copR gene were amplified using the primer pairs: (0739upAscIFW/0739up2RW and 0739dofus2FW/0739doNotIRW). Both PCR products were combined with single-overlap extension PCR and cloned into modified pMUR47 vector via the AscI and NotI restriction sites. The resulting construct (pMUR527) was verified by DNA sequencing.

Circular plasmid DNA and strain MURPf37 were used for transformation and selection was carried out in SME-starch liquid medium without agmatine sulfate and I+G at 85°C for 12 h. Pure cultures of the intermediate mutant MUR65Pf_i were obtained by plating the cells on solidified medium. The integration of the plasmid into the genome by single cross-over was verified by analyzing corresponding PCR products.

For the counter selection cells were plated on solidified medium containing 50 µM 6-methylpurine and 8 mM agmatine sulfate to induce a second homologous recombination step to recycle the selection marker and to eliminate integrated plasmid sequences. The genotype of the final mutant (MUR65Pf) was confirmed by PCR and cells had to be grown in the presence of 8 mM agmatine sulfate and 8 mM I+G.

To restore wild type growth properties (growth without agmatine sulfate and I+G) plasmid pMUR310 was created. The newly designed two-gene resistance cassette was amplified from the pUC19 subclone using the primer pair: (pYS_PF1623F_GA/pYS_PF1950R_GA). This PCR product was cloned into PCR-amplified (PF1623_pYSF_GA/PF1950_pYSR_GA) pYS3 plasmid (Waege et al., 2010) using NEB Gibson Assembly® Cloning Kit. Correctness of the construct was tested by Sanger sequencing. 1 µg of the circular plasmid was transformed into MURPf37 and MURPf65 as described (Waege et al., 2010; Kreuzer et al., 2013). Selection was carried out in 1/2 SME liquid medium without agmatine sulfate and I+G at 85°C for 12 h. Pure cultures of the mutant MUR52Pf and MURPf74 were obtained by plating the cells on solidified medium. Plasmid uptake was verified by re-transformation into *E. coli* and DNA sequencing of purified plasmids. Final mutants could be grown without agmatine sulfate and I+G supplementation.

3.3. RNAP, TBP, TFB

For in vitro transcription assays and EMSA analysis, we used RNAP purified from *P. furiosus* cells and recombinant TBP and TFB as described previously (Waege et al., 2010; Ochs et al., 2012; Reichelt et al., 2018b).

3.4. CopR, CopRΔTRASH, CopRΔHIS and CopRΔHISΔTRASH

3.4.1. Cloning and expression

The gene sequence of PF0739 was amplified from genomic DNA of *P. furiosus* using primers with additional BamHI and NdeI restriction recognition sites. The PCR product was cloned into vector pET-30b (NEB) using the respective restriction sites. PF0739 protein variants lacking potential metal-sensing domains (ΔTRASH, ΔHIS, ΔHISΔTRASH) were based on the full-length plasmid version and ligated after amplification using one phosphorylated primer, respectively (see Supplementary Table 12). Subsequently, the constructs were transformed into *E. coli* DH5-α for amplification and grown on Kanamycin (50 µg /ml) supplemented LB media. Next, the constructs were transformed into *E. coli* BL21 STAR™ (DE3) expression strain and grown on Kanamycin (50 µg /ml) supplemented LB medium at 37°C. Protein expression was induced by addition of 0.5 mM IPTG to the cell culture medium at an OD_{600} of about 0.6. Cultures were further cultivated at 18°C overnight, before harvesting the cells by centrifugation at 10,000 g for 10 min at 4°C. Cells were stored at -80°C until protein purification.

3.4.2. Cell disruption and pre-purification

For the purification of PF0739 and PF0739 variants, cells were first resuspended in 50 ml low salt buffer containing 40 mM HEPES (pH 7.5), 80 mM ammonium sulfate, 1 mM EDTA, 10% glycerol (w/v) and a protease inhibitor tablet (Roche). The cell lysis was

done by sonification on ice, whereby breakage efficiency was monitored at a light microscope. After cell disruption, DNase I (Roche) was added and all cultures incubated at 37°C for 1 hour. In the next step, the lysate was centrifuged at 48,000 g for 20 min at 4°C and the supernatant was transferred to a new tube. A pre-purification step was carried out by applying a heat treatment of 90°C for 15 min and subsequent centrifugation at 48,000 g for 20 min at 4°C.

3.4.3. Affinity and size exclusion chromatography

The supernatant containing the protein of interest was filtered and loaded onto a 5-ml HiTrap™ Heparin HP column equilibrated with low salt buffer. Next, the protein was eluted by gradually increasing the buffer concentration of the high salt buffer (compare low salt, but 1 M ammonium sulfate). Fractions containing PF0739 (checked on SDS-PAGE) were further purified using size exclusion chromatography by pooling and concentrating of the relevant fractions and loading onto a 24-ml HiLoad™ 10/300 GL Superdex™ 200 column pre-equilibrated with low salt buffer. This column was also used to study multimerization of PF0739 (data not shown).

3.5. Growth experiments using an optical device

P. furiosus was cultivated under anaerobic conditions in 40 ml ½ SME medium supplemented with 0.1 % yeast extract, 0.1 % peptone and 40 mM pyruvate at 95°C, as described previously (Fiala and Stetter, 1986; Waege et al., 2010). For growth comparison experiments, the medium was supplemented with different $CuSO_4$ concentrations (compare Figure 27) and each condition for MURPf52 (parental strain) and MURPf74 (ΔcopR strain) was recorded in biological triplicates during 48 hours of incubation by measuring the turbidity changes in situ using a photodiode and a LED with 850 nm as light source. The recorded values were converted to cell/ml by using a calibration curve with known cell concentrations, calculated in a Thoma counting chamber (0.02-mm depth; Marienfeld, Lauda-Königshofen, Germany) using phase-contrast microscopy.

3.6. Differential gene expression analysis

3.6.1. Growth conditions and RNA isolation

P. furiosus parental strain MURPf52 was grown in standard medium at 95°C to late-exponential phase. After reaching a cell density of $1x10^8$, cells were either shocked by adding 20 µM $CuSO_4$ (copper-shock) or left untreated (control) and incubated for 30 minutes. The experiment was performed in biological triplicates.

Total RNA was isolated using the Monarch RNA purification Kit (NEB), including the recommended genomic DNA removal by on-column DNase treatment, according to the instructions of the manufacturer. Quantity, quality and integrity were measured using

Nanodrop One, Qubit RNA HS assay kit (Thermo Fisher Scientific) and the Prokaryote total RNA Nano Kit on a Bioanalyzer to measure RIN values (Agilent).

3.6.2. Library preparation and sequencing

Library preparation and RNA-seq were carried out as described in the Illumina TruSeq Stranded mRNA Sample Preparation Guide, the Illumina HiSeq 1000 System User Guide (Illumina, Inc., San Diego, CA, USA), and the KAPA Library Quantification Kit - Illumina/ABI Prism User Guide (Kapa Biosystems, Inc., Woburn, MA, USA). In brief, 100 ng of total RNA from *P. furiosus* was fragmented to an average insert size of 200-400 bases using divalent cations under elevated temperature (94°C for 4 minutes), omitting the mRNA purification step with poly-T oligo-attached magnetic beads. Next, the cleaved RNA fragments were reverse transcribed into first strand cDNA using reverse transcriptase and random hexamer primers. Actinomycin D was added to improve strand specificity by preventing spurious DNA-dependent synthesis. Blunt-ended second strand cDNA was synthesized using DNA Polymerase I, RNase H and dUTP nucleotides. The incorporation of dUTP, in place of dTTP, quenched the second strand synthesis during the later PCR amplification, because the polymerase does not incorporate past this nucleotide. The resulting cDNA fragments were adenylated at the 3' ends, the indexing adapters were ligated and subsequently specific cDNA libraries were created by PCR enrichment. The libraries were quantified using the KAPA SYBR FAST ABI Prism Library Quantification Kit. Equimolar amounts of each library were used for cluster generation on the cBot with the Illumina TruSeq SR Cluster Kit v3. The sequencing run was performed on a HiSeq 1000 instrument using the indexed, 50 cycles single-read (SR) protocol and the TruSeq SBS v3 Reagents according to the Illumina HiSeq 1000 System User Guide. Image analysis and base calling resulted in .bcl files, which were converted into FASTQ files with the bcl2fastq v2.18 software.

Library preparation and RNA-seq were performed at the service facility "KFB - Center of Excellence for Fluorescent Bioanalytics" (Regensburg, Germany; www.kfb-regensburg.de).

3.6.3. Data analysis using the DESeq2 pipeline

For differential gene expression analysis, Illumina reads in FASTQ format were quality/length/adapter trimmed using trimmomatic (v. 0.36) in single-end-mode (Bolger et al., 2014). Therefore, we allowed for a minimum length of 12 bases and a cut-off Phred score of 20, calculated in a sliding window of 4 bases. Next, we used the STAR aligner (v. 2.5.4) to map the reads to a recently published updated version of the P. furiosus genome (Dobin et al., 2013; Grünberger et al., 2019). Mapping statistics are included in the Supplementary Table 13. The sorted BAM files were then used to generate count tables using featureCounts (Liao et al., 2019). Differential gene expression analysis was performed

using the DESeq2 pipeline (Love et al., 2014b). Furthermore, we used the apeglm method for effect size shrinkage and calculation of fold changes (Zhu et al., 2019a). All steps of the analysis, including the generation of plots were performed using R and can be found at www.github.com/felixgrunberger/CopR (R Foundation for Statistical Computing., 2018).

Enrichment analysis of archaeal cluster of orthologous genes (arCOGs) was performed by extracting the gene specific arCOG information from the arCOG database (ftp://ftp.ncbi.nih.gov/pub/wolf/COGs/arCOG/) and calculation using the goseq package, which allowed for custom genome sets (Young et al., 2010; Makarova et al., 2015).

3.6.4. Confirmation of data using RT-qPCR

RT-qPCR reactions were performed similar as described previously (Reichelt et al., 2018b). In short, total isolated RNA was reverse transcribed using the ProtoScript® II First Strand cDNA Synthesis Kit (NEB), according to the manufacturer´s instructions and using a random primer mix (Promega). The reactions were assembled in triplicates using the qPCRBio SyGreen Mix Lo-Rox Kit (PCRBiosystems) with reverse transcribed cDNA from the first step in a 1:10 dilution, including a control reaction that lacked the reverse transcriptase (-RT) and a no template control (NTC). RT-qPCR reactions were run on a Rotor-Gene Q cycler (Qiagen) in a three-step protocol: 95°C – 10′ for one cycle; 95°C – 30″, 58°C – 30″, 72°C – 30″ for 40 cycles. Data evaluation was done using the corresponding Rotor-Gene Q software package (Qiagen). Relative expression levels were calculated using the delta-delta Ct method ($2^{-\Delta\Delta Ct}$), by comparing the Ct values from biological triplicates of the gene of interest to a house-keeping gene pf0256 (Spt5). The applicability of pf0256 as a calibrator was evaluated before (Reichelt et al., 2018b).

3.7. ChIP-seq analysis

3.7.1. Immunoprecipitation

We used an adaption of a ChIP-seq protocol that was established previously for P. furiosus (Reichelt et al., 2016). *P. furiosus* cells were grown under anaerobic conditions in serum bottles containing 40 ml ½ SME medium at 95°C as described earlier. After the cells reached a density of $2x10^8$, formaldehyde was injected into the flask to a final concentration of 0.1 % (v/v). After 60 seconds the crosslinking reaction was stopped by addition of glycine to a final concentration of 15 mM (v/v). For the copper-treated samples, the cells were shocked with 20 µM $CuSO_4$ for five minutes before the crosslink reaction was induced.

Cell disruption and DNA fragmentation was performed in one step via sonication for 25 minutes using the ultrasonic homogenizer Sonopuls HD 2070 (Bandelin, Berlin, Germany) until an average fragment length of 250 to 400 bp. The insoluble particles were removed by centrifugation. For determination of the DNA concentration and fragment length, 1 volume of crude cell extract was mixed with 4 volumes of ChIP elution buffer (10 mM Tris, pH 8.0, 1% (w/v) SDS, 0.1 mM EGTA) and incubated over night at 65°C. After

RNase treatment, the DNA was purified using the NucleoSpin® Gel and PCR Clean-up Kit (Macherey-Nagel). The DNA concentration was determined using the Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific) and the fragment length by agarose gel electrophoresis.

For immunoprecipitation (IP) 100 µl Protein G beads (Dynabeads, Invitrogen) were coupled to 3920 µg serum antibodies against CopR (PF0739), according to manufacturer´s instructions. Polyclonal antibodies were produced by Davids Biotechnology (Regensburg, Germany) from recombinantly expressed and purified CopR.

100 µl of antibody-beads complexes were mixed with 900 µl of *P. furiosus* crude extract adjusted to a DNA concentration of 4.44 ng/µl (4 µg DNA/sample) in PBST. The samples were incubated with rotation for two hours at room temperature. The immunoprecipitated samples were placed on a magnet, the supernatant was discarded, and the bead-pellet was washed 2x with low salt buffer (50 mM HEPES, pH 7.4, 150 mM NaCl, 1 mM EDTA, 0.1% (w/v) SDS, 0.1% (w/v) Deoxycholic acid, 1% (v/v) Triton X-100), 1x with high salt buffer (50 mM HEPES, pH 7.4, 500 mM NaCl, 1 mM EDTA, 0.1% (w/v) SDS, 0.1% (w/v) Deoxycholic acid, 1% (v/v) Triton X-100), 1x with ChIP wash buffer (10 mM Tris, pH 8.0, 250 mM LiCl, 1 mM EDTA, 0.5% (v/v) Nonidet P-40, 0.5% (w/v) Deoxycholic acid) and 1x with TE buffer (10 mM Tris, pH 8.0, 0.1 mM EDTA, 20 % (v/v) Methanol, 25 mM Tris, 192 mM Glycine) (Aparicio et al., 2005). Each washing step was done with 1 ml buffer by rotation for one minute. To elute the immuno-bound DNA from the beads, the bead-pellet was resuspended in 25 µl ChIP elution buffer, transferred to a PCR cup and incubated for 10 minutes at 65°C. The cup was placed on a magnet, the supernatant was transferred to a new cup, the bead-pellet was resuspended in 25 µl TE buffer supplemented with 0.67% SDS (v/v) and incubated for 10 minutes at room temperature. Afterwards, both eluates were combined in one PCR cup. For the input sample, 400 ng DNA of *P. furiosus* crude extract was mixed 1:4 with ChIP elution buffer.

Eluted complexes and input samples were incubated overnight at 65°C to reverse the crosslink. After the incubation, the samples were treated with RNase A (0.1 mg/ml final concentration) for 15 minutes at 37°C and Proteinase K (0,2 mg/ml final concentration) for 15 minutes at 65°C. ChIP-DNA and input DNA were purified using the NucleoSpin® Gel and PCR Clean-up Kit (Macherey-Nagel). The DNA concentration of the input DNA was determined using the Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific).

3.7.2. Library preparation and sequencing

Library preparations were done using the NEBNext® UltraTM II DNA Library Prep Kit for Illumina® with the NEBNext® Multiplex Oligos for Illumina® Index Primers Set 2 and 3 according to the manufacturer's protocol. Library quantification was done with the NEBNext® Library Quant Kit for Illumina® according to manufacturer's instructions. Before sequencing, the libraries were pooled in equimolar ratios. The library pool was

quantified with the KAPA SYBR FAST ABI Prism Library Quantification Kit (Kapa Biosystems, Inc., Woburn, MA, USA) and used for cluster generation on the cBot with the Illumina TruSeq SR Cluster Kit v3. Sequencing was performed on a HiSeq 1000 instrument controlled by the HiSeq Control Software (HCS) 2.2.38, using the indexed, 50 cycles single-read (SR) protocol and the TruSeq SBS v3 Reagents according to the Illumina HiSeq 1000 System User Guide. Image analysis and base calling were done by the Real Time Analysis Software (RTA) 1.18.61. The resulting .bcl files were converted into FASTQ files with the CASAVA Software 1.8.2. Sequencing was performed at the service facility "KFB - Center of Excellence for Fluorescent Bioanalytics" (Regensburg, Germany).

3.7.3. Analysis of ChIP-seq data

FASTQ files were quality/length/adapter trimmed with trimmomatic (v. 0.36) in single-end-mode using a minimum-length of 40 bp, a cut-off Phred score of 20 (Bolger et al., 2014). Reads were mapped to the P. furiosus genome using Bowtie 2 (v. 2.2.3) with default settings (Langmead and Salzberg, 2012). SAM files were converted to sorted BAM files using samtools and extended towards the 3´direction by their fragment-size to better represent the precise protein-DNA interaction (Li et al., 2009; Leleu et al., 2010). Position-specific enrichments were calculated by extracting the mean counts of biological triplicates from bed files for the ChIP-sample, comparison to the input files and taking the $\log_2$.

3.7.4. Confirmation of data using quantitative real-time PCR (RT-qPCR)

qPCR primer pairs were designed using the Primer3 software package and quality assessed. qPCR reactions were assembled as technical triplicates using 2x qPCRBioSyGreen Mix separate Rox kit in a total volume of 10 µl. Primers were added to a final concentration of 0.3 µM. 6 µl of the master mix were mixed with 4 µl of template DNA or H_2O_{DEPC} as NTC. The specificity of the PCR product was verified by melting curve analysis.

qPCR reactions were run on the Rotor Gene Q cycler with a three-step PCR program described in 2.5.4. Replicates with a deviation >0.5 were excluded from the analysis. The fold enrichment was again calculated according to the delta-delta Ct method.

3.8. In vitro assays

3.8.1. Electrophoretic mobility shift assay (EMSA)

DNA templates were obtained from genomic DNA by PCR amplification with the corresponding primer pairs (see Supplementary Table 12). One of the two primers was labelled at the 5´-end with a fluorescent dye. 20 nM DNA was assembled in a 15 µl reaction volume containing: 50 ng competitor DNA (Hind-III-digested λ DNA), 670 µM DTT, 20 µg/ml BSA, 6.7 % glycerol, 40 mM HEPES (pH 7.4), 80 mM $(NH_4)_2SO_4$ and various amounts of proteins and metals, as described in the results part. The reactions were

incubated for 5 minutes at 70°C and analysed using a non-denaturing 5 % polyacrylamide gel. After electrophoresis, the DNA fragments were visualized with a Fujifilm FLA-5000 fluorescence imager.

3.8.2. DNase I footprinting

DNase I footprinting was performed as previously described (Ochs et al., 2012). In short, the DNA template containing the promoter regions of pf0739 and pf0740 was obtained from genomic DNA by PCR amplification (see Supplementary Table 12). HEX-labelled primers were used in two separate reactions for strand-specific labelling. 4.4 nM template DNA was assembled in a 15 µl reaction volume containing: 40 mM Na-HEPES (pH 7.5), 125 mM NaCl, 0.1 mM EDTA, 0.1 mg/ml BSA, 1 mM DTT, 0.5 mM $MgCl_2$ and 2.8 µM CopR, 1 µM TBP and 0.8 µM TFB according to Figure 5. After incubation for 20 minutes at 70°C, 0.05 units of DNase I (Fermentas) was added and incubated for another minute at 70°C. The reaction was terminated by the addition of 5 µl 95 % formamide and incubation for 3 minutes at 95°C. The DNA was precipitated with ethanol and resuspended in 2-4 µl formamide buffer. A DNA sequencing ladder was generated using a DNA cycle Sequencing Kit (Jena Bioscience) according to manufacturer´s instructions. Samples were loaded onto a 4.5 % denaturing polyacrylamide gel and analysed using an ABI 377 DNA sequencer.

3.8.3. In vitro transcription assay

In vitro transcription assays were performed similar as described previously (Ochs et al., 2012; Reichelt et al., 2018b). To this end, template DNAs containing the promoter regions of the respective gene were amplified from genomic DNA or plasmid pUC19/gdh by PCR amplification (see Supplementary Table 12). The reactions were assembled in a total volume of 25 µl containing: 2.5 nM template, 5 nM RNAP, 30 nM TFB, 95 nM TBP, 40 mM Na-HEPES (pH 7.4), 250 mM KCl, 2.5 mM $MgCl_2$, 5 % (v/v) glycerol, 0.1 mM EDTA, 0.1 mg/ml BSA, 40 µM GTP, 40 µM ATP, 40 µM CTP, 2 µM UTP, and 0.15 MBq (110 TBq/mmol) [α-32P]-UTP, if not indicated otherwise (see Figure 5). After incubation for 10 minutes at 80°C, the RNA transcripts were extracted by phenol/chloroform, denatured in formamide buffer for 3 minutes at 95°C and separated on a denaturing 6 % polyacrylamide gel. Finally, the transcripts were visualised using a fluorescence imager (FLA-5000, Fuji, Japan).

3.9. Negative-stain transmission electron microscopy and image analysis by 2D class averaging

For transmission electron microscopy (TEM), protein solutions with a concentration of 220 ng/µl were chosen. Samples were negatively stained with a solution containing 2 % (w/v) uranyl acetate (UAc) in presence of 0.005 % n-dodecyl β-D-Maltopyranosid (DDM).

A carbon film coated grid - 400 Square Mesh (Plano GmbH), was incubated with 3 µL of protein sample for 45 seconds. Excess stain was blotted off using filter paper and samples washed using 3 µl of a 2 % UAc solution. The blotting and washing procedures were repeated and the sample finally air-dried and stored at room temperature.

Negative-stained grids were imaged on a TEM JEOL-2100F (200 kV) equipped with a 4k x 4k F416 camera with CMOS chip/detector, TVIPS at a 50K magnification (0.211 nm/pixel) with a defocus range from -0.5 to -1.4 µm. The contrast transfer function for a total of 32 micrographs was determined with CTFfind4 (Rohou and Grigorieff, 2015). Subsequently, the particles were extracted with a mask of 180 Å and processed in RELION 3.0 (Zivanov et al., 2018) to yield the 2D class averages.

4. Results

4.1. *Pyrococcus copR* is part of the conserved archaeal *cop* gene cluster

A conserved *cop* resistance gene cluster plays a critical role in copper homeostasis in Archaea. The gene cluster has been identified in various archaeal species using comparative genomics (Ettema et al., 2006). This *cop* cluster consists of a copper-exporting P_{1B}-ATPase CopA (*Ferroplasma acidarmanus* Fer1: CopB), a transcriptional regulator CopR (*Saccharolobus solfataricus* P2: CopT, *F. acidarmanus* Fer1: CopY) and occasionally the metallochaperone CopT (*S. solfataricus* P2: CopM, *F. acidarmanus* Fer1: CopZ) (Figure 26) (Baker-Austin et al., 2005; Ettema et al., 2006; Villafane et al., 2009; Hong et al., 2019).

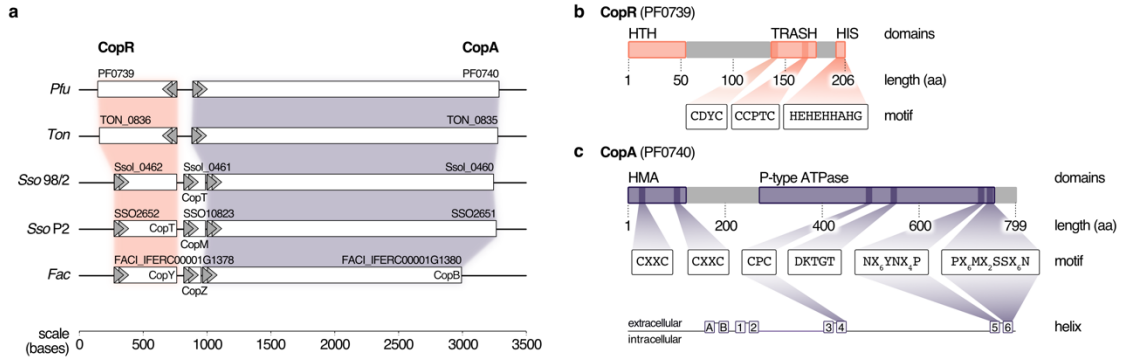


Figure 26 | PF0739 (CopR) is part of the conserved archaeal cop cluster in *Pyrococcus furiosus*. **a**, Copper regulation in Archaea is achieved by a highly conserved *cop* gene cluster consisting of a transcriptional regulator (CopR/CopT/CopY), a transporter (CopA/CopB) and optionally a chaperone (CopT/CopM/CopZ). The organisation of the *cop* cluster in *P. furiosus* was compared to *cop* clusters of other archaeal organisms (*Ton = Thermococcus onnurineus*, *Sso96/2 = Saccharolobus solfataricus 96/2, SsoP2 = Saccharolobus solfataricus P2, Fac = Ferroplasma acidarmanus*) (Baker-Austin et al., 2005; Ettema et al., 2006; Villafane et al., 2011; Hong et al., 2019). Genes are drawn to scale; directionality is indicated by arrows **b**, Schematic representation of the transcriptional regulator (CopR) encoded by *pf0739* in *P. furiosus*. The regulator consists of a N-terminal helix-turn-helix (HTH) domain that mediates DNA binding, and a metal-sensing TRASH domain (Ettema et al., 2003). The C-terminal Histidine-rich sequence (HIS) is only found in *Thermococcales*. Further sequence features of the domains are highlighted. **c**, Bioinformatical analysis of conserved amino acids in transmembrane helices 4, 5 and 6 classify PF0740 as the copper exporter CopA.

In *Pyrococcus furiosus*, the transcriptional regulator CopR is encoded by the gene *pf0739*, which is in divergent orientation to *copA* (*pf0740*). It consists of an N-terminal helix-turn-helix domain and a C-terminal metal-sensing TRASH domain together with a Histidine-rich region (HIS), the latter one is only present in the order *Thermococcales* (Figure 26b) (Ettema et al., 2003). To unravel more details about the function of CopR in *P. furiosus*, we expressed the wild type protein together with three mutants in *Escherichia coli* and tested the proteins for DNA binding using gel-shift assays. Here, the intergenic region between *copR* and *copA* served as target DNA. Binding of CopR to the target region

with increasing protein concentrations resulted in a specific protein-DNA shift (Supplementary Figure 21). The mutated variants, lacking the putative metal-binding domains, TRASH, HIS or both, showed a very similar DNA binding affinity for all variants, indicating that DNA binding is primarily mediated by the HTH domain whereas the metal-binding domains are dispensable for DNA binding.

To determine both selectivity and sensitivity towards the recognized metal of the CopR/CopA system in *P. furiosus*, we performed two types of experiments: First, motif analysis of the heavy metal-binding domains (HMBDs) classified PF0740 as a copper-exporting ATPase of type 1B (Figure 26c) (Argüello, 2003; Sitsel et al., 2015). Secondly, increasing concentrations of different metal ions ($AgNO_3$, $CuSO_4$, $FeCl_3$ and, $CoCl_2$) were supplemented in the binding assays of CopR. All ions reduced the binding affinity, but the most potent effect was observed for copper and silver ions (Supplementary Figure 22a). In contrast to the copper-induced release in the EMSA analysis using the full-length CopR, the effect was slightly reduced in the case of a CopR HIS mutant and significantly reduced for a CopR TRASH mutant (Supplementary Figure 22b,c,d). It is interesting to note that lower metal concentrations resulted in a smear with reduced mobility and higher levels in an increasing amount of released DNA. The minimal concentration used in the binding assays that resulted in CopR release was 12.5 µM $CuSO_4$, which is very similar to the detection range of CopR in *T. onnurineus* (79 % amino acid sequence identity) (Hong et al., 2019). Taken together, these results indicate that the CopR/CopA system is involved in copper regulation in *P. furiosus*.

4.2. Deletion of *copR* transcriptional regulator leads to a copper-sensitive phenotype

To learn about the importance of CopR for copper-detoxification in *P. furiosus*, a *copR* deletion mutant was constructed, using an established genetic system in this hyperthermophilic organism (Waege et al., 2010; Kreuzer et al., 2013). Growth analysis of this deletion mutant (MURPf74) in comparison to the parental strain (MURPf52) was performed using increasing amounts of copper (0 to 100 µM). Starting at sub-lethal concentrations, we observed prolonged lag phases and reduced cell densities in both strains (Figure 27).

While the collected curves were almost identical for 0 and 10 µM $CuSO_4$, higher copper concentrations caused a significant effect on the growth of the knockout strain. In contrast to the parental strain, the growth of the *copR*-disrupted strain was almost completely abolished in the presence of 100 µM copper. This finding supports the idea that copper homeostasis is a tightly controlled system with a sensitivity in the µM-range and indicates that CopR acts as a transcriptional activator of the copper-exporting ATPase *pf0740* in *P. furiosus*.

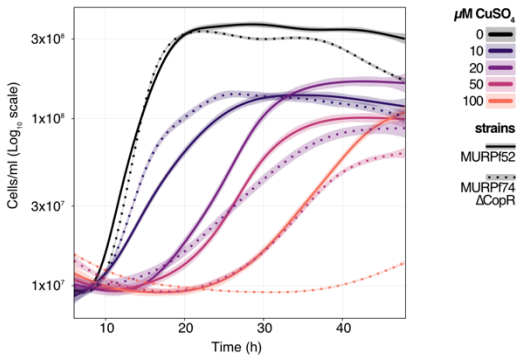


Figure 27 | Growth analysis of the *P. furiosus* parental strain (MURPf52) and CopR-knockout strain (MURPf74) in the presence of $CuSO_4$. Triplicates of 40 ml cultures were analysed in standard medium at 95°C, supplemented with 0 to 100 µM $CuSO_4$ (indicated by colour scale). Growth was recorded during 48 hours of cultivation of the parental strain (solid lanes) and the *copR* deletion strain (dashed lines). Each curve represents the fitted line of three independent experiment, with the shaded area displaying the confidence interval (0.99).

4.3. Characteristics of the *P. furiosus* transcriptome in response to a copper-shock

To investigate the role of CopR in the copper regulation network in *P. furiosus*, we applied an integrative approach, combining differential gene expression (DGE) analysis and genome-wide binding analysis by ChIP-seq. For DGE, we cultivated the parental strain (MURPf52) until the middle of the log phase, shocked the cells with 20 µM $CuSO_4$ for 20 minutes and isolated the RNA for next-generation sequencing. PCA analysis confirmed that indeed the copper-shock (and not handling of the biological replicates) caused most of the variance in the experimental setup (Figure 28a). Hence, we were able to compare the transcriptomic changes primarily due to copper shock and not as a result of a general stress response or cell death. By analysing the transcript abundances, we could confirm the essential role of *copA* in removing excess ions from the cell, observing a 70-fold up-regulation of the mRNA levels after copper treatment (Figure 28b). Altogether, 34 genes were up-regulated, but not a single gene was significantly down-regulated by more than 2-fold (Figure 28c,d). RT-qPCR experiments verified the up-regulation of two of the most prominent genes (PF0740, PF0738.1n) (Supplementary Figure 23). Notably, *pf0727* is among the most up-regulated genes (105-fold). Based on the domain annotation and strong induction upon copper treatment, PF0727 is most likely the missing chaperone in the *cop* cluster in *P. furiosus*. Due to the presence of an HMA domain instead of a TRASH domain the protein belongs to the CopZ and not to the CopM family. A closer look at the clusters of archaeal orthologous genes (arCOGs) revealed that most of the up-regulated genes belong to the groups O (posttranslational modification, protein turnover, chaperones), S (function unknown) and P (inorganic ion transport and metabolism) (Figure 28e,f) (Makarova et al., 2015). The group of the hypothetical genes consists of eight candidates, including *pf0738.1n*, which exhibits the most substantial up-regulation (290-fold).

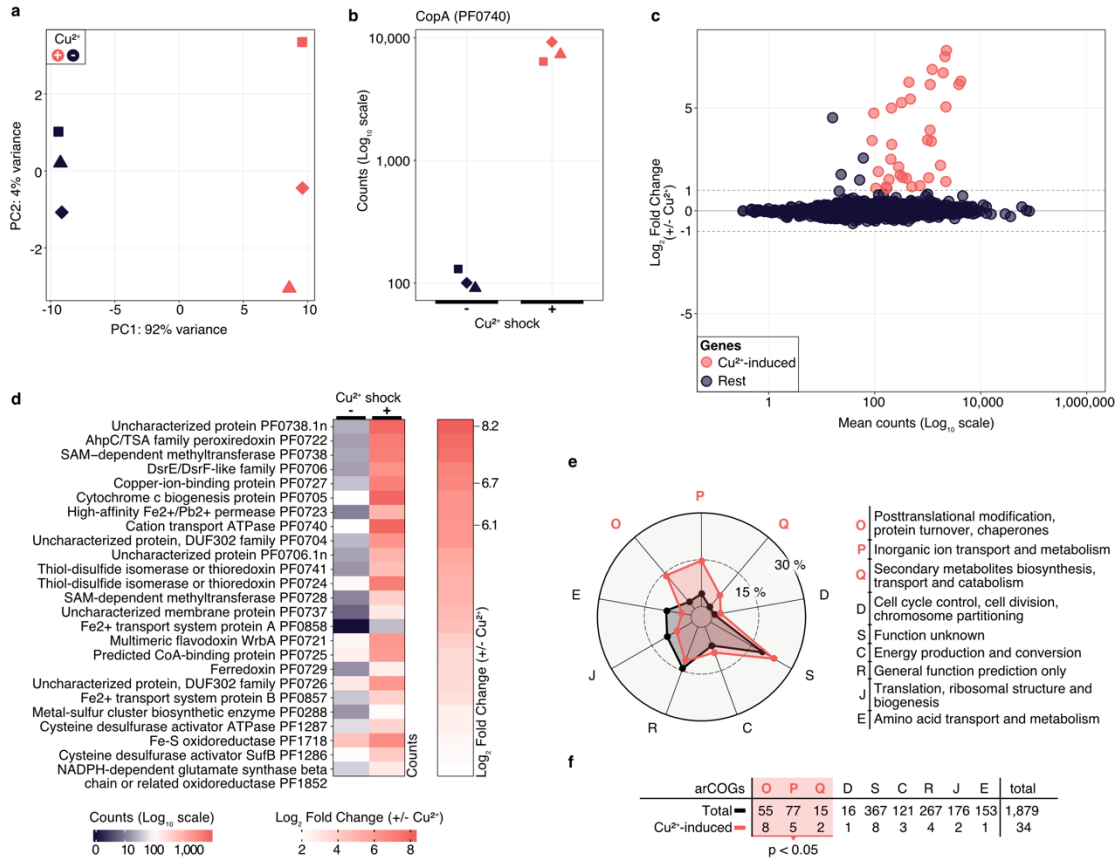


Figure 28 | Differential gene expression analysis of *P. furiosus* after 20 min copper shock with 20 µM $CuSO_4$. **a**, Principal component analysis of variance stabilized transformed RNA-Seq read-counts of normal conditions (dark blue) and Cu^{2+} shock conditions (red) shows that most of the variance in the experimental setup is caused by the treatment of the cells. Replicates are indicated by different shapes. **b**, Comparison of raw RNA-Seq read-counts mapping to PF0740 shows a 6.14 Log_2 Fold Change after Cu^{2+} shock (adjusted p value = 1.80e-173). **c**, MA plot showing the distribution of mean read counts (log_{10} scale) against log_2 fold change. 34 genes that are more than 2-fold up-regulated with an adjusted p-value <0.05 are highlighted in red. **d**, Heatmaps for mean read-counts for control and Cu^{2+} shock condition are shown for the 25 most significant regulated genes. **e**, Enrichment analysis of archaeal clusters of orthologous genes (arCOGs) found in the 34 significantly up-regulated genes (compare panel c). Contribution of each category is calculated in percentage and compared to the total background set with significantly up-regulated categories marked in orange. **f**, Enrichment analysis is based on the comparison of the total number of genes found in an arCOG category (Total) and the number of genes that are significantly up-regulated (>2 fold).

4.4. Integrative RNA-seq and ChIP-seq identifies CopR targets

It is interesting to note that the 14 most up-regulated genes are located within a 28 kb region of the genome. To answer the question if the transcriptional activation of these genes upon intoxication is connected to the binding of CopR, we performed a ChIP-seq experiment with and without copper shock (20 µM $CuSO_4$). The results from these experiments demonstrated a very similar CopR occupancy independent of the copper treatment (Figure 29a). This finding is in agreement with the EMSA analysis, where a

considerable amount of the transcriptional regulator remained bound to the DNA after addition of 25 µM $CuSO_4$ (comparable amount as used in the *in vivo* experiments, see Supplementary Figure 22). Furthermore, the binding pattern of CopR overlaps with the upstream regions of up-regulated genes or operons under both conditions, which confirms specific binding of CopR, as well as the possible role in transcriptional activation.

To elucidate the sequence specificity of CopR binding and regulation, we compared the nucleotide content of the CopR-regulated promoter regions to a background set consisting of 763 sequences that contributed to a recently published consensus motif in *P. furiosus* (Grünberger et al., 2019) (Figure 29b). This archaeal-typical promoter motif is characterized by elements that facilitate transcription by the recruitment of the basal transcription factors TFB (BRE element) and TBP (TATA box) and melting of the region initially upstream of the transcription start site (TSS). In comparison to the *Pyrococcus* consensus motif, the promoter sequences of the up-regulated genes showed some minor deviations in the BRE element and the conserved A(T) at position -10 that contributes to the promoter strength (Torarinsson et al., 2005), but the most striking difference was a C-enriched TATA box (Figure 29b). Further upstream of the promoter sequence we identified a TC-rich and AG-rich signature from -35 to -50 that also differed from the consensus sequence content. A motif enrichment analysis using MEME identified a semi-palindromic-like motif with the minimal palindromic consensus sequence TTNNCAWWWTGNNAA, which is located at almost all CopR-regulated promoters directly upstream of the BRE element (8 of 9 with an annotated TSS) (Figure 29c,d). Scanning of this motif in the promoter region of all known TSSs further confirmed the specificity, as all of the motif occurrences were bound and up-regulated by CopR (8 of 8 total hits, q-value < 0.01). To validate the findings of the ChIP-seq experiments, we performed gel-shift assays and ChIP-qPCR, that both confirmed specific binding of CopR to multiple genomic regions (Figure 29e) and enrichment under both conditions (Supplementary Figure 24b).

Based on the genome-wide binding pattern and in combination with the results from the DGE analysis, we propose a currently undescribed regulating role of CopR on a global level to maintain copper homeostasis in *P. furiosus*.

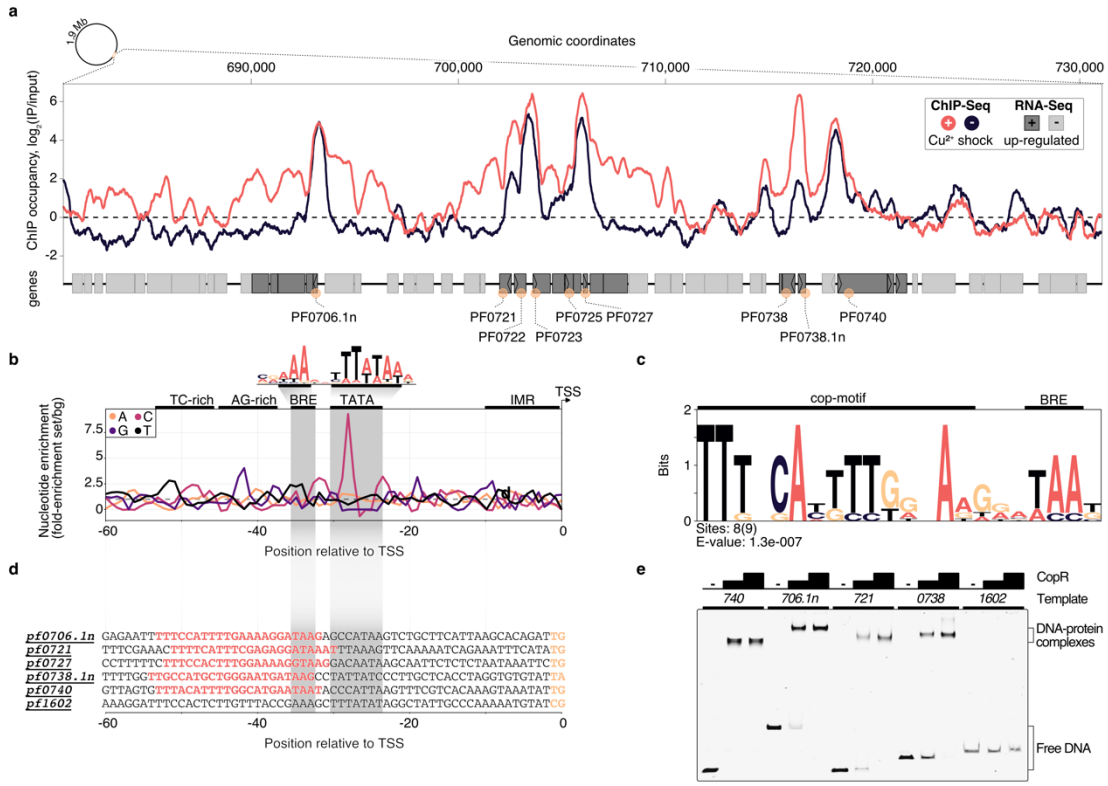


Figure 29 | ChIP-seq and integration with differential gene expression data identifies CopR as a global regulator of copper homeostasis in *Pyrococcus furiosus.* **a**, ChIP occupancy of CopR zoomed to genomic region ~680,000 to 730,000, which contains the 14 most up-regulated genes from the differential gene expression analysis (compare Figure 28). ChIP-seq curves were generated for Cu^{2+} shocked (red) and untreated (dark blue) samples by comparing the IPs to input samples (mean values of triplicates are shown). Genome annotation is shown at the bottom according to scale with significantly up-regulated genes (adjusted p value < 0.05, Log_2 fold change +/- Cu^{2+} > 1) colored in dark grey. Up-regulated genes that are bound by CopR under both conditions are highlighted with orange circles. An example of an unbound genomic region is shown in Supplementary Figure 24a). **b**, Nucleotide enrichment analysis of upstream regions dependent on the transcription start sites comparing the sequences of selected genes (orange circles, n = 9) with the nucleotide content of sequences contributing to the consensus motif of primary transcripts in *P. furiosus*. The consensus motif, which consist of a B recognition element (BRE) and a TATA box, is highlighted above the enrichment plot (Grünberger et al., 2019). **c**, MEME motif analysis of selected genes. A semi-palindromic motif positioned directly upstream of the BRE element was found. **d**, Promoter sequences of genes that were further analysed by EMSA or ChIP-qPCR (compareSupplementary Figure 24b). **e**, EMSA analysis (20 nM DNA, 0/200/400 nM protein) of selected promoter regions confirms specific binding of CopR to multiple regions, whereas no binding to a control promoter (*pf1602, gdh*) could be observed.

4.5. CopR activates transcription *in vitro*

To verify the stimulating role of CopR, we performed *in vitro* transcription experiments with a DNA template that allows simultaneous transcription of the divergently orientated *copR* and *copA* genes (Figure 30a,b). In the absence of CopR, the main transcript originated from the own strong promoter, and *copA* was only weakly transcribed.

However, with increasing concentrations of CopR, transcriptional output increased for the *copA* gene and *copR* transcription was significantly reduced (Figure 30b). In contrast, it did not affect a control template, lacking the CopR binding site (Figure 30a), which clearly indicates that CopR is responsible for both, activation of *copA* and repression of *copR*. The reason for the observed *in vitro* CopR-induced activation is not known, since the stimulating effect of *copA in vivo* was observed only under the presence of copper ions. Attempts to increase the *copA* stimulating effect by adding $CuSO_4$ or $AgNO_3$ failed. The presence of such ions inhibited the transcription in general (data not shown).

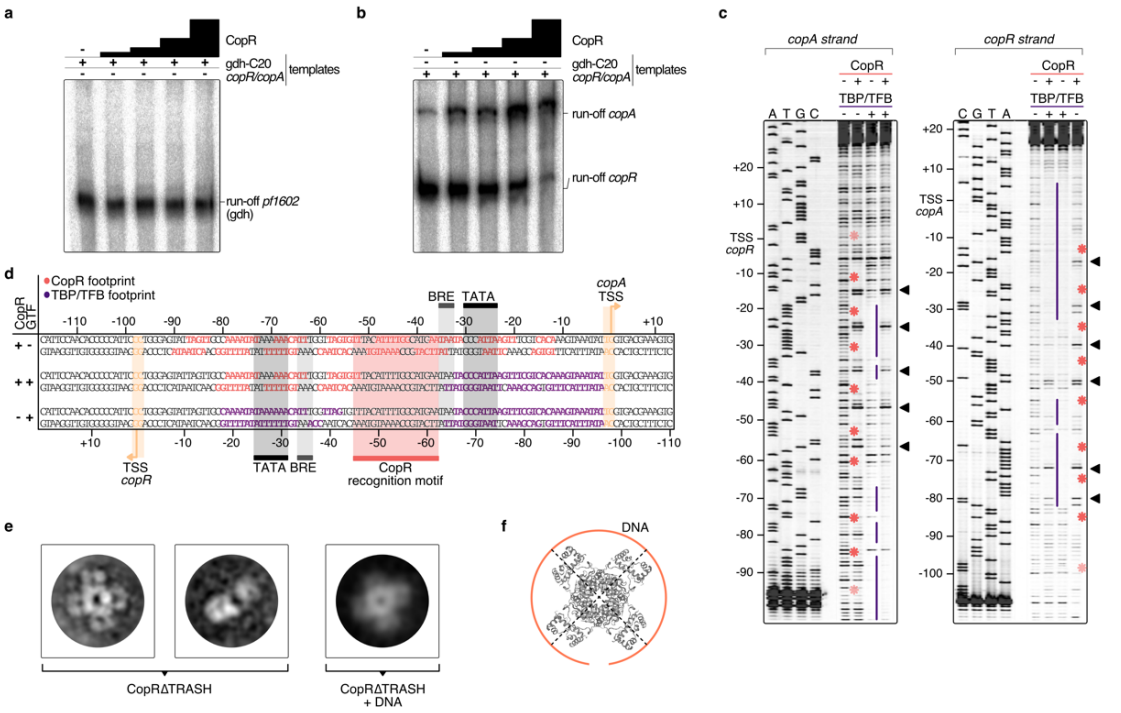


Figure 30 | Mechanistic and structural characterisation of CopR. a, Influence of CopR on *in vitro* transcription**.** 2 nM of the *gdh* and **b,** the *copR/copA* templates were transcribed in the presence of increasing concentrations of CopR (0.3, 0.6, 1.2, 2.3 µM). **c**, DNase I footprint on the *copR/copA* template in the presence of CopR, TBP/TFB and all three components. TBP/TFB-protected regions (purple lines), CopR-protected regions (red asterisks) and hypersensitive sites (black triangles) are highlighted. **d**, Summary of the protected regions determined in the *in vitro* DNAse I footprinting assay. Promoter elements, transcript boundaries and the semi-palindromic CopR recognition motif are highlighted. **e**, Representative 2D classes of CopRΔTRASH in the absence and presence of a *copR/copA* DNA template reveal an octameric assembly formed by a tetramer of dimers in both states. **f**, Putative model of CopR bound to DNA. The octameric cartoon structure (PDB: 1I1G) represents LrpA from *P. furiosus*, that forms a similar structure (Leonard et al., 2001).

For additional information about the mechanism of activation, we performed DNase I footprinting experiments at the *copR/copA* promoter (Figure 30c,d). The binding of CopR revealed an extended binding pattern in this region consisting of eight strong single footprints in the central region of the fragment and some additional weaker footprints towards the border of the fragment. Most of the strong signals are separated by

hypersensitive sites separated by approximately one helical turn (Figure 30c). The correlation of the binding motif obtained from the ChIP-seq experiment with the footprint pattern revealed that the motif is located between the divergent TBP/TFB binding sites of *copR* and *copA.* CopR footprint signals are positioned in the centre of the motif and nearby upstream and downstream. These signals are separated by hypersensitive sites, which touch only one or two bases of the beginning or the end of the consensus sequence. Simultaneous presence of the two basal transcription factors, TBP and TFB, and CopR in the reaction (Figure 30d, lane 2), revealed that under these conditions only CopR is in contact with DNA in the region of the *copR* promoter. This finding is in agreement with the identified repression of *copR* in the *in vitro* transcription experiments. In contrast, the position of the CopR footprint upstream of the TBP/TFB footprint that is still present at the *copA* promoter enables CopR-mediated stimulation of *copA*.

4.6. Towards a structural view of CopR

CopR is the crucial player in the copper-triggered differential regulation of genes that supports detoxification of the cell. At the same time, however, it is not apparent how activation is achieved considering that CopR always seems to be bound to the respective promoter regions. Therefore, we aimed to elucidate the structural properties of CopR. To this end, we developed a purification protocol for the isolation of highly pure CopR. Based on the comparison of CopR to standard calibration proteins, elution profiles of size exclusion chromatography runs (Superdex 200) indicated an octameric conformation (data not shown). As the elution profile revealed an increased symmetric peak for CopRΔTRASH in comparison to the wild type protein, we have selected CopRΔTRASH for further analysis.

Negative-stain TEM imaging confirmed an extremely high monodisperse fraction of the CopRΔTRASH mutant. Finally, 2D classification of more than 50,000 particles, gave rise to the assumption that CopR forms an octameric assembly oriented in a cruciform-like structure (Figure 30e). Due to the high similarity of the structure to LrpA and F11 (Leonard et al., 2001; Yokoyama et al., 2007), we assume a similar outside orientation of the helix-turn-helix domain of the four CopR dimers that positions DNA at the outside of the dimer. Furthermore, CopR complexed with DNA of the *pf706.1n* promoter did not alter the overall octameric structure of the protein (Figure 30e). With a diameter of 16 nm, a DNA fragment of about 150 bp would be necessary to completely wrap the DNA around the protein (Figure 30f) (Leonard et al., 2001).

5. Discussion

From the combination of *in vitro* based approaches, genetic manipulation and the integration of DGE and ChIP-seq data used in this study, we conclude that CopR from *P. furiosus* is a copper-sensing global regulator of transcription and essential during copper detoxification.

Firstly, the importance of CopR for maintaining copper homeostasis became apparent by growth experiments. While we could show that a *P. furiosus* parental strain can adapt to µM-concentrations of Cu^{2+} (up to 100 µM were tested), the growth defect on a CopR-knockout strain was significant. We assume that limiting amounts of essential components necessary for maintaining copper homeostasis are responsible for the observed growth phenotypes. The most obvious component is the copper efflux pump CopA, which showed a 70-fold increase at the RNA level under copper-shock conditions. *In vitro* experiments confirmed that enhanced *copA* transcription is mediated by CopR-induced activation. The *substantial copA* enrichement is also in agreement with the observed CopR-binding to the upstream region of *copA* as indicated by EMSA and DNase I footprinting analysis. In summary, we assume that CopR is responsible for sensing copper concentrations and transcriptional activation of the corresponding genes necessary to maintain copper homeostasis.

Our findings for CopR are in agreement with data from *S. solfataricus* strain 98/2 and *Halobacterium salinarum* as the corresponding knockout mutants also revealed CopR as a positive regulator for *copA* transcription (Kaur et al., 2006; Villafane et al., 2011; Darnell et al., 2017). In contrast, recent data from the closely related *T. onnurineus* NA1 suggested CopR as a repressor for autoregulation and *copA* transcription (Hong et al., 2019). This fundamental discrepancy for *copA* transcription is remarkable in the light of the identical organization of the operons and 79% sequence identity between both proteins instead of 30% to the *Saccharolobus* CopR. A comparison in more detail revealed that in *Pyrococcus* -in contrast to *Thermococcus*- low µM $CuSO_4$ concentrations do not lead to a full release of the protein from the DNA but result in a low mobility complex indicating a conformational change of the CopR-DNA complex. These findings are also in line with *in vitro* DNA-binding studies from *S. solfataricus* strain P2, which also suggest a partial rearrangement of CopR-binding in the presence of copper instead of dissociation (Ettema et al., 2006).

An additional difference between both organisms is the quaternary structure of CopR in solution: In the case of *Thermococcus* a tetrameric structure was determined using size exclusion chromatography analysis (Hong et al., 2019) and for *Pyrococcus* an octameric complex was found by negative-stain TEM imaging and gel filtration experiments. In order to explain these differences, it is tempting to speculate that the presence of a His_6 tag in the N-terminal region of *Thermococcus* CopR is responsible for the observed different

quaternary structure of the protein. Furthermore, it is possible that the presence of this tag also contributes to an increased sensitivity to copper towards dissociation from the DNA instead of allowing a conformational switch necessary for transcriptional activation. To verify the different regulation mechanism of CopR in *T. onnurineus* NA1, additional information about the susceptibility of the mentioned *copR* deletion strain (Hong et al., 2019) to increasing copper concentrations or the behaviour of CopR without a His_6 tag would be helpful.

Despite the inconsistency of the function of CopR as repressor or activator for *copA* transcription between *Thermococcus* and *Pyrococcus*, the ChIP-seq data collected in this study demonstrate an almost identical DNA binding pattern of CopR independent of the presence or absence of copper ions. This finding also points to a required structural rearrangement of CopR on the DNA to activate transcription in the presence of copper ions. Such behaviour would be very similar to the function of the copper-sensing transcription factor CueR from *Escherichia coli* (Philips et al., 2015). CueR can activate transcription by controlling open complex formation, while it is continuously bound to DNA (Martell et al., 2015). This mechanism may permit more rapid responses to environmental changes. In an evolutionary context, the high copper-toxicity could have been a driving force for the independent development of this regulatory mechanism in different domains. Nevertheless, there is no sequence similarity between both proteins, CueR belongs to the predominant bacterial MerR family of regulators (PROSITE documentation PDOC00477), and CopR belongs to the Lrp/AsnC family (PDOC00520). The latter one is a rather old family of prokaryotic transcriptional regulators and very common in Archaea (Peeters and Charlier, 2010). Crystal structures of several archaeal Lrp members indicate a highly conserved octameric structure with an N-terminal winged HTH motif for DNA binding and a C-terminal domain necessary for oligomerization and effector binding (Leonard et al., 2001; Yokoyama et al., 2007; Kumarevel et al., 2008). Our TEM imaging data of negative-stained CopR also revealed a tetrameric assembly of dimers with most likely the DNA wrapped around the protein. This finding is in line with published structures of DNA-protein complexes FL-11 and Grp (Yokoyama et al., 2007; Kumarevel et al., 2008; Yamada et al., 2009) which further confirmed an accumulated occurrence of an octameric assembly within the Lrp family. The identification of extended DNA binding regions and hypersensitive sites in footprinting experiments is also in agreement with the assumed wrapping of the DNA around the octamer (Ouhammouch, 2001; Liu et al., 2014). Therefore, the behaviour of CopR is very similar to published data of the Lrp family with the difference that CopR uses copper ions as effector instead of molecules of the amino acid metabolism. However, the ability to use a variety of effector molecules is pervasive for the archaeal subfamily. A detailed analysis of eight Lrp/AsnC paralogs in *Halobacterium salinarum* revealed that these proteins are involved in regulating

genes in response to copper or oxidative stress, changes in K^+ or NAD^+ concentrations or modified growth conditions (Plaisier et al., 2014).

The Lrp/AsnC family is not only involved in the regulative response to a wide range of different physiological conditions but also employ different mechanisms of transcriptional regulation. Repression by preventing the recruitment of the RNAP is demonstrated for LrpA from *P. furiosus* (Dahlke, 2002) and activation by stimulating the binding of TBP is shown for Ptr2 in *Methancaldococcus jannaschii* (Ouhammouch et al., 2003). Besides, a dual regulator mechanism has been shown for Ss-LrpB from *Saccharolobus solfataricus*, which activates transcription at low factor-concentrations, whereas at high concentrations, transcription is repressed (Peeters et al., 2009).

Based on our results, we conclude that CopR also has a dual function as repressor and activator, but the situation seems to differ from the LrpB from *S. solfataricus*. *In vitro* transcription experiments in the absence of CopR indicate a strong *copR* promoter without the necessity for further activation. However, *in vivo*, CopR remains always bound to the *copR* promoter region, which blocks TBP/TFB recruitment, represses transcription of its gene and only allows some basal expression. This goes in line with the DGE data, which indicate low-level expression of *copR* independent of copper ions and high-level expression of *copA* in the presence of copper. Similar results about these differences in the expression rates were also described in *Saccharolobus solfataricus* P2 (Ettema et al., 2006), *Saccharolobus solfataricus* 98/2 (Villafane et al., 2009), *Sulfolobus metallicus* (Orell et al., 2013), *Ferroplasma acidarmanus* Fer1 (Baker-Austin et al., 2005) and *Halobacterium salinarum* (Kaur et al., 2006).

To interpret our results and to integrate these data with the knowledge gained for the Lrp family in general, we suggest the following regulation mechanism for divergent transcription (Figure 31a): Under normal growth conditions, CopR binds to multiple binding sites with about 150 bp of DNA wrapped around each octamer. Each dimer of the octamer is in "direct contact" with a weakly conserved DNA sequence as indicated by motif analysis of the ChIP-seq data and the footprinting experiments. We assume an increased affinity to binding sequences located directly upstream of promoter sequences and cooperative binding, which seems to be a common feature of Lrp molecules (Peeters et al., 2004; Chen et al., 2005), to weaker signals stimulated by the octameric structure of CopR downstream of the promoter. An additional contact downstream of the TATA box in combination with transcriptional activation is already described for Ptr2 and BarR (Ouhammouch et al., 2005; Liu et al., 2014). For transcriptional activation, we suggest an allosteric regulation mechanism where the binding of effector molecules (most likely Cu^+) is sensed by the TRASH domain alone or in combination with the HIS stretch at the C-terminal end. The role of the TRASH domain in copper binding is indicated by the decreased metal sensitivity of the ΔTRASH mutant in the gel shift assays and was also previously demonstrated by mutational analysis in *T. onnurineus* NA1 (Hong et al., 2019).

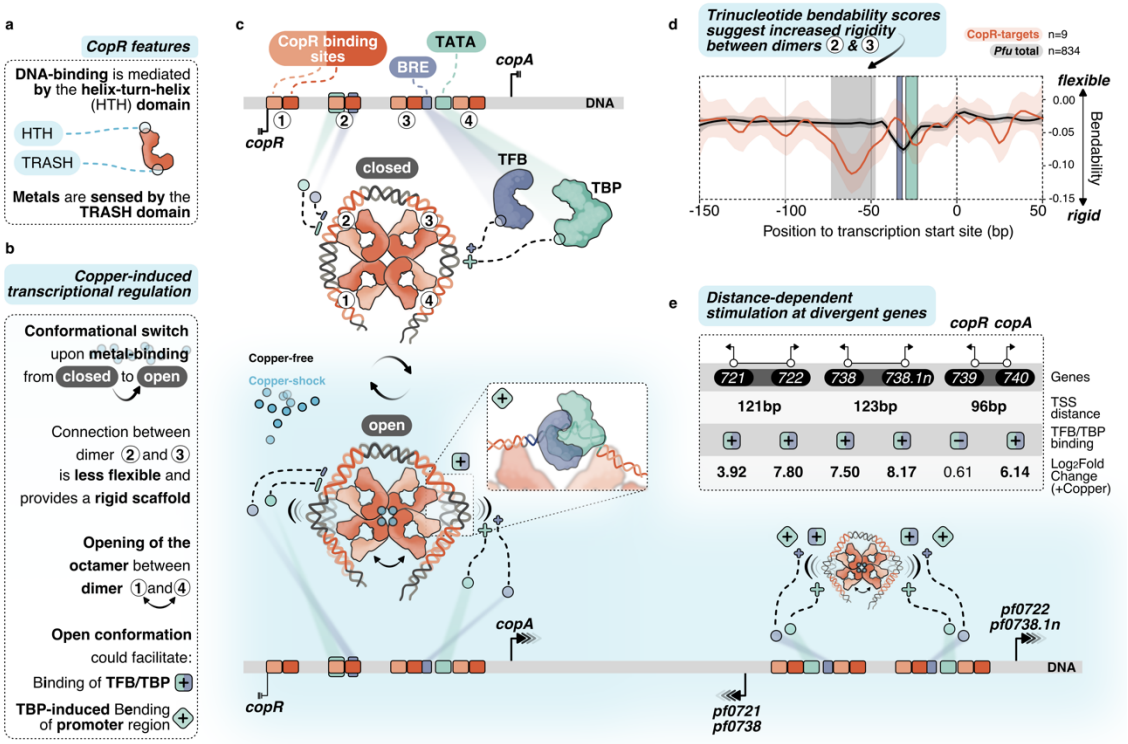


Figure 31 | Putative models of CopR regulation and copper detoxification in *P. furiosus*. **a**, Domain architecture of CopR (monomer shown schematically in red) highlighting the DNA-binding HTH domain and the metal-sensing TRASH domain. **b**, Binding of copper presumably triggers a conformational switch from a closed state to a complex, which is opened between dimer 1 and 4. In combination with the rigid connection between dimers 2 and 3 it is possible that this conformational switch could facilitate TBP/TFB binding (+ square) and/or TBP-induced bending to the corresponding promoter regions (+ turned square). **c**, CopR-regulated promoter regions include CopR-binding sites (orange/red) and the archaeal-specific promoter elements BRE (recruits TFB, purple) and the TATA box (bound by TBP, light-green). Transcription start sites (TSS) are indicated by vertical lines. Transcription is either repressed (3 lines) or stimulated (arrows) under copper-shock conditions (lower panel, light-blue). The octameric CopR assembly in open conformation allows bending of critical promoter regions and facilitates binding of TBP/TFB depending on the distance of the divergent TSS: While CopR prevents binding of general transcription factors to the *copR* promoter, TBP/TFB can bind to the second promoter (*copA*). Simultaneous binding of CopR and two sets of GTFs stimulates transcription in both directions for *pf0722/pf0271* and *pf0738/pf0738.1n*. **d**, Major groove bendability of selected promoters revealed increased rigidity between dimers 2 and 3. Major groove bendability of promoter sequences was estimated based on trinucleotide scales derived from DNase-I cutting frequencies (Brukner et al., 1995; Meysman et al., 2014). More negative values are the result of less cutting by DNase-I and indicate that the DNA is not bend towards the major groove and is therefore less flexible. Trinucleotides were extracted from CopR-target promoters (n=9) and all available promoters defined in *P. furiosus* previously (Grünberger et al., 2019) from -150 bp to +50 from the TSS. Each line represents the smoothed conditional mean with confidence intervals (0.95) displayed as shaded areas. Boxes represent area between dimers 2 and 3 (grey), BRE (purple) and TATA box (light-green). **e**, Summary of the TSS-distance dependent stimulation of divergent CopR-regulated transcripts.

Due to the binding of the metal, we expect a conformational switch resulting in the opening of the quaternary structure of the octamer similar to the structure of Lrp from *Escherichia coli* or FL11 with bound arginine from *Pyrococcus* OT3 (de los Rios and Perona, 2007; Yamada et al., 2009). We assume that opening between dimer 1 and 4 is preferred due to increased flexibility at this position as these dimers are not directly linked with the wrapped DNA (Figure 31b,c). In contrast, there is a direct connection from dimer 1 to dimer 2, 3 and 4, which most likely provides a more rigid scaffold. The bendability of CopR-regulated promoters in comparison to the total set of promoters in *P. furiosus* (Grünberger et al., 2019) was estimated by comparing trinucleotide scores (Brukner et al., 1995; Meysman et al., 2014) and clearly shows a less flexible region between dimer 2 and 3 (Figure 31d). The movement of dimer 1 towards 2 and dimer 4 towards 3, initiated by the opening of the octamer, may reduce torsional stress on the DNA between these corresponding pairs of dimers, which could either facilitate the accessibility of TBP and TFB to the corresponding promoter sequences or enable TBP-induced bendability or is involved in both (Figure 31b,c). Torsional stress on the DNA as limitation for binding of TBP was already demonstrated by single molecule FRET experiments in *Methanocaldococcus jannaschii* (Nickels et al., 2016).

Interestingly, provided a divergent gene organization and an appropriate distance between the two TSSs, one CopR octamer can promote transcriptional activation of two separate gene clusters (Figure 31e). Simultaneous stimulation by CopR requires a distance of about 122 bp between the two TSSs (*pf0721/pf0722*; *pf0738/pf0738*.1n). In contrast, the reduced distance of 96 bp in the case of the *copR/copA* gene cluster leads to repression of the *copR* gene independent of the presence of copper ions, most likely due to binding interference of dimer 3 with the *copR* promoter. Additionally, there is also the possibility of activation at only one side of the octamer, whereas the position of the other side is located at the end of another gene (*pf0726/pf0727*).

Besides mechanistic details of CopR transcriptional regulation, our data also allow insights into a general copper-specific transcriptomic response (Figure 32). To avoid a transcriptomic response due to cell-death rather than a metal-specific response, we applied a moderate copper shock using 20 µM $CuSO_4$ in all *in vivo* experiments. However, under these conditions, we already saw a strong phenotype of the CopR knockout mutant, which emphasizes the essential role of CopR. Using this “semi-toxic” concentration in the DGE analysis, we found 34 strongly up-regulated genes (> 2-fold) upon copper-shock (Supplementary Table 14). The transcriptional pattern under coper shock conditions is comparable to other metal stress transcriptomic responses in prokaryotes and eukaryotes, and apart from metal-specific genes also includes non-metal related genes that cooperatively contribute to metal resistance (Bini, 2010; Lagorce et al., 2012).

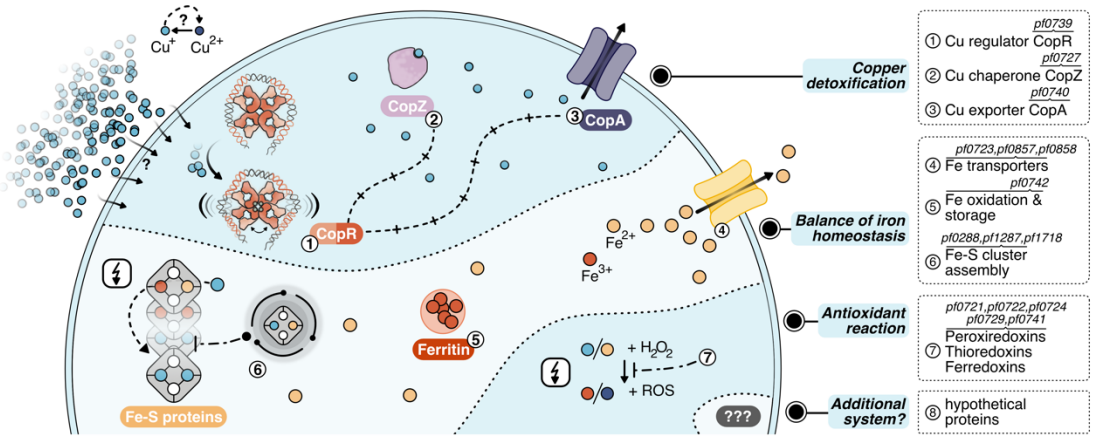


Figure 32 | Layers of copper detoxification in *P. furiosus*. In a primary response, CopR (1,orange-red) senses excessive amounts of Cu^{+} ions (light-blue), that entered the cell by a currently unknown mechanism or diffusion and activates the transcription of the copper-chaperone CopZ (2,light-purple) and the copper-exporter CopA (3,purple). After CopZ potentially delivers the Cu^{+} ions to the transporter, they are exported out of the cell. The main toxic effect of copper ions is caused by replacing iron ions in iron-sulfur clusters. Therefore, the induction of Fe-S cluster assembly proteins, Fe-transporters and ferritin helps to re-balance iron homeostasis (4-6). Additionally, antioxidant enzymes prevent the induction of the Fenton reaction (7), which otherwise causes toxicity by the production of reactive oxidative species (ROS).

For the primary detoxification mechanism, *Pyrococcus* relies on the induction of the copper efflux pump CopA and the metallochaperone CopZ (PF0727), which directly interact with the copper ions. An additional cluster of genes deals with i) iron homeostasis involving several transporters to pump iron ions (PF0723, PF0857, PF0858) (Zhu et al., 2013), ii) ferritin (PF0742) which combines oxidation of Fe^{2+} to Fe^{3+} together with storage of the oxidized iron inside the protein cavity (Honarmand Ebrahimi et al., 2015) and iii) Fe-S cluster assembly proteins (PF0288, PF1286, PF1287, PF1718). This collection of genes fit well into the recently emerging concept that the primary toxic effect of copper is the replacement of iron in iron-sulfur cluster proteins and not the conversion of H_2O_2 to hydroxyl radicals (Lagorce et al., 2012; Tan et al., 2017). Therefore, the induction of these proteins helps to re-balance displaced iron ions and to avoid inactivation of iron-sulfur proteins. Besides, antioxidant enzymes as peroxiredoxins, thioredoxins or ferredoxins (PF0721, PF0722, PF0724, PF0729, PF0741) can also assist in preventing the induction of the Fenton reaction by the released Fe^{2+} or Cu^{+} ions. This is in line with the finding that some of these enzymes are also induced after exposure to hydrogen peroxide (Strand et al., 2010). Since the constitutively expressed superoxide reductase also produces hydrogen peroxide (Thorgersen et al., 2012; Khatibi et al., 2017), it is most likely that the induction of these antioxidant enzymes successfully inhibits the production of hydroxyl radicals via the Fenton reaction under this low dose of copper. In consequence, there is almost no induction of genes dealing with general stress response or DNA repair mechanisms. In contrast, a copper shock in *Metallosphaera sedula* induced a mixed gene population of

metal-specific and also generic responses indicating that the conditions used have had much more potent effects concerning viability in comparison to our setup (Wheaton et al., 2016).

In this context, it is interesting to note that *Metallosphaera* uses an additional mechanism for copper resistance: sequestration with inorganic polyphosphate to facilitate export from the cytoplasm (Rivero et al., 2018). Such a mechanism is also described for *S. solfataricus*, as a mutant strain -unable to accumulate polyphosphate- showed an increased copper sensitivity in spite of *copA* up-regulation (Soto et al., 2019). Based on our data, there is no indication that a comparable system is implemented in *Pyrococcus*. Nevertheless, the induction of eight hypothetical proteins opens up the possibility that different sequestration systems or an additional mechanism for copper detoxification exist.

Data Availability

Raw sequence data have been uploaded to the SRA and are available under project accession number PRJNA603674.

Code Availability

The R scripts detailing the analysis can be found in the corresponding Github repository under www.github.com/felixgrunberger/CopR.

Author Contributions

FG did the DGE and the complete bioinformatic analysis, RR constructed the *Pyrococcus* deletion strain, IW did the *in vitro* transcription and the footprinting experiments. VN and LK performed the gel shift assays, KB the ChIP-seq experiments and the qPCR assays. MK, NW, ZE, GM and CZ did the negative-stain TEM imaging. FG, DG and WH wrote the manuscript and DG and WH coordinated and supervised the work. All authors agreed to the final version of the manuscript.

Funding

This work was supported by the Institute of Microbiology and Archaea Center of the University of Regensburg, the SFB960 and by the German Research Foundation (DFG) with the funding program Open Access Publishing.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments

The authors thank Renate Richau and Wolfgang Forster for excellent technical assistance, Dominik Strobel for constructing a CopR expression clone, Markus Schick and Christine Lindenthal for initial work in the setup of the genetic system and Thomas Kopp from the electronic workshop from the University of Regensburg for the setup of the optical device to measure cell density.

CHAPTER III
General Discussion

1. Comprehensive summary

This thesis comprises three articles that contribute to the current knowledge of general and regulated archaeal transcription, mainly using the hyperthermophilic archaeon *Pyrococcus furiosus* as a model organism (Grünberger et al., 2019, 2020b). Additionally, the application of Nanopore sequencing for *de novo* genome sequencing and native RNA sequencing is a notable expansion of the current spectrum of genome-wide methods in prokaryotes (Grünberger et al., 2020a). It provides data sets, wet lab and bioinformatical protocols that are open-source and can be used to dissect different transcriptional and post-transcriptional aspects in prokaryotes.

Genome stability under laboratory conditions is one of the most critical pre-requisitions for the reproducibility and informative of gene regulatory studies in a genome-wide context. We challenged the prevailing opinion that *P. furiosus* has a highly unstable genome by re-sequencing a lab culture using Nanopore technology, therefore simulating normal work conditions in many microbiological labs, and comparing it to a new reference genome obtained from hybrid Illumina-PacBio DNA sequencing. Thereby, no genomic rearrangements were observed in the *P. furiosus* genome, that is accordingly more stable

than previously anticipated at least under “normal” conditions. Also, long-read Nanopore sequencing proofed as a valuable tool for the *de novo* sequencing of small archaeal genomes by ending up with a reasonably high identity of 99.92% in the assembly compared to the gold standard protocol. To provide a framework for future genome-wide studies, we explored the transcriptome of *P. furiosus* by using a differential RNA sequencing approach, that allowed us to add multiple transcriptomic features to the current annotation specifically by enriching for primary transcripts. We found that the majority of mRNAs, antisense RNAs and intergenic RNAs are preceded by Archaea-typical promoter elements, which goes in line with other genome-wide studies in the order *Thermococcales*. Interestingly, we could not only confirm that antisense transcription is widespread in Archaea but also discovered bidirectional promoters as a possible source for this phenomenon and suggested a potential role of asRNAs in IS element silencing. Although it is known since a long time that the overall symmetry of the TATA box necessitates the direction information of the BRE element, bidirectional promoters with two BRE elements on opposite strands have not been described before in Archaea. However, functional consequences and distribution in other archaeal species remain to be elucidated.

After successfully establishing workflows for Nanopore DNA sequencing, we expanded this approach and applied the technology for single-molecule sequencing of RNA in its native context. Therefore, we used the protocols provided from Oxford Nanopore on polyA-tailed and non-rRNA-depleted RNAs from the prokaryotic model organisms *E. coli, H. volcanii and P. furiosus*. Although native RNA sequencing is still in its infancy, sequencing throughput and quality are already good enough to address multiple transcriptome-wide features at once, including transcript boundaries, operon structures, overall RNA levels and post-transcriptional processing. Taking advantage of the long-read sequencing method, we found that some transcripts in Archaea have heterogeneous 3´ends, pointing towards more diverse termination mechanisms than previously expected.

Notably, we were able to retrace and expand on the insufficiently described multi-step rRNA maturation process in Archaea. Ribosomal RNA processing includes the dedicated action of ribonucleases and modifying enzymes in a defined timely order. Performing co-occurrence analysis of terminal positions from coherent long reads, we identified multiple different precursors, including the Archaea-specific circular pre-rRNAs. Our data not only suggest that internal tRNA and 23S rRNA processing is preceded by 16S pre-rRNA cleavage at the bulge-helix-bulge by the splicing endonuclease EndA, but also that the spacer region is ligated afterwards and creates a circular pre-rRNA intermediate, which is subsequently opened in the next steps of maturation.

Finally, we were able to correlate the consecutive steps of archaeal rRNA maturation with their respective RNA modification status. By analysing systematic basecalling errors and raw ionic current signals, we provided *in vivo* evidence of selected RNA modifications. While dimethylation by the universally conserved KsgA is completed at late stages of small

ribosomal subunit biogenesis in *H. volcanii* and *P. furiosus*, h45 N^4-cytidine acetylation is absent in *H. volcanii* and appears to be already introduced in earlier steps in *P. furiosus*. Furthermore, expanding the modification models to the complete 16S rRNA confirmed the recently detected widespread acetylation at CCG motifs, hence supporting the critical role of stabilising RNA modifications in hyperthermophilic Archaea.

Excitingly, the detection of potentially new rRNA precursors and base modifications in the hyperthermophilic *P. furiosus* highlighted the possibilities of single-molecule long-read sequencing of native RNAs. We believe that while the current limitations of Nanopore native RNA sequencing will likely improve very soon, the technology in its current state is already a highly valuable addition to the prokaryotic transcriptome toolbox.

The highly accurate re-annotation of a stable *P. furiosus* genome and the extensive annotation of transcriptomic features facilitated an integrated genome-wide analysis of the copper-sensitive transcriptional regulator CopR (PF0739). Initially, this factor was coincidentally identified from a pull-down RNAP fraction, probably owed to its His-tag like C-terminus. CopR is part of a conserved cop-cluster, that is present in many Archaea and regulates transcription of a copper-exporting ATPase CopA (PF0740), that is in divergent orientation. Using *in vitro* binding assays and growth experiments with a CopR-knockout strain, we could show that cells lacking CopR are more sensitive to low µM concentrations of copper ions. Applying mild copper shock conditions, we performed a differential gene expression (DGE) analysis of normal and shock conditions and genome-wide binding analysis of CopR under the same conditions. The aim was to (i) analyse the transcriptome-wide copper-shock response of *P. furiosus* and (ii) identify targets that are regulated by CopR.

We found that the copper shock triggered the induction of 34 strongly up-regulated genes (> 2 fold), which we clustered in (i) primary detoxification mechanisms by the efflux pump CopA and the chaperone CopZ, (ii) genes balancing of iron homeostasis, (iii) antioxidant reactions to prevent the induction of the radical producing Fenton reaction, and iv) a set of hypothetical genes with currently unknown function.

Integrating these results with the CopR occupancy data, we identified 12 CopR regulated gene clusters. Most of them share a semi-palindromic recognition sequence and a rigid region upstream of the TATA box. Most interestingly, CopR stays bound under normal and copper shock conditions to the promoters in a Lrp-like octameric conformation. We propose an allosteric model of CopR regulation that includes the conformational change after copper binding, which could help in the recruitment and accessibility of general transcription factors or the essential TBP-guided bending of the promoter.

In summary, the presented work has contributed new tools and provided new insights to transcriptional and post-transcriptional mechanisms in Archaea in general and more specifically in the hyperthermophilic *P. furiosus*. The knowledge of transcriptional regulatory aspects for many years has been shaped extensively by dedicated *in vitro* studies.

Only recently, the development of genome-wide applications enabled a look at transcription from a different angle. While many features of archaeal transcription could be confirmed or extended to a global scale, entirely new or previously neglected questions were posed. These include the widespread use of asRNAs, the coupling of translation and transcription and its effect on termination, the function of modified RNA bases, regulation of transcription in a histone-context, and many more characteristics of archaeal transcription. In the future, integrated structural and biochemical analysis, in combination with large-scale omics data, will be critical to dissect the multiple levels of gene regulation in Archaea.

2. Dissecting archaeal transcription

Archaeal transcription is shaped by a plethora of different features on multiple levels. Many of the basal regulatory mechanisms are encoded in a sequence-, position- or distance-specific way and rely on the interaction of basal and gene-specific transcription factors with the RNAP. Additionally, the increasing amount of identified post-transcriptional events provide fine-tuning opportunities through diverse strategies. Together, all of these mechanisms contribute to the transcriptional output and allow fast reactions to perturbated environmental conditions, a critical prerequisite for microorganisms. Only in recent years, global *in vivo* sequencing studies broadened the knowledge of transcription in a genome-wide context and complemented, but in many ways also challenged formal concepts that have been widely accepted in the community. These concepts were in many cases the outcome of dedicated *in vitro* approaches using model promoters, e.g. from tRNAs, rRNAs, viral or highly-expressed genes, and reconstituted transcription systems (Blombach et al., 2019). However, it is not possible to account for unknown factors, which can limit and bias the analysis in an unpredictable way. Besides, a meta-analysis of multiple genes instead of single-gene characterisation can result in a more balanced picture of shared regulatory principles. Accordingly, global analysis has the power to put the lessons learned *in vitro* into a genome-wide context and to further untangle common features that are shared between archaeal phyla.

2.1. The uniformity of archaeal promoter elements

In comparison to Bacteria and Eukarya, transcription initiation in Archaea is simplistic and reduced in its diversity of factors as well as promoter elements (Smollett et al., 2017a; Blombach et al., 2019). Archaeal promoters only contain three core elements, the BRE, TATA and INR (Figure 33a). In contrast, multiple additional up- but especially downstream elements, like MTE (motif ten element), DPE (downstream core promoter element), DCE (downstream core element), DTIE (downstream transcription initiation element) amongst others, can be found in the eukaryotic RNAP II transcription system (Vo Ngoc et al., 2017, 2019). Unlike this diversity of promoter subtypes that can be linked to diverse biological functions in Eukarya, fundamental differences in core elements have not been observed in Archaea (Smollett et al., 2017a; Vo Ngoc et al., 2017). Genome-wide TSS mapping in many archaeal model organisms revealed a remarkable uniformity of promoter structures across different phyla (Jäger et al., 2009, 2014; Wurtzel et al., 2010; Babski et al., 2016; Cho et al., 2017; Smollett et al., 2017b; Grünberger et al., 2019) (Figure 33a). The only outliers in this context seem to be halophilic organisms, like *H. volcanii*, that have a shortened TATA box and a hardly detectable BRE element, which we could confirm using the Nanopore-based sequencing approach (Babski et al., 2016; Grünberger et al., 2020a). However, it remains unclear, if the halo-archaeal promoter structure is in any way associated with the presence of multiple TBP and TFB paralogues or if they have

a different function (Baliga et al., 2000). Intriguingly, competition for the weak BRE by multiple TFBs but interaction with only a single TBP has been shown in *Halobacterium* NRC-1, which reminds of the well characterised alternate sigma factor utilisation in Bacteria (Wade et al., 2006; Facciotti et al., 2007).

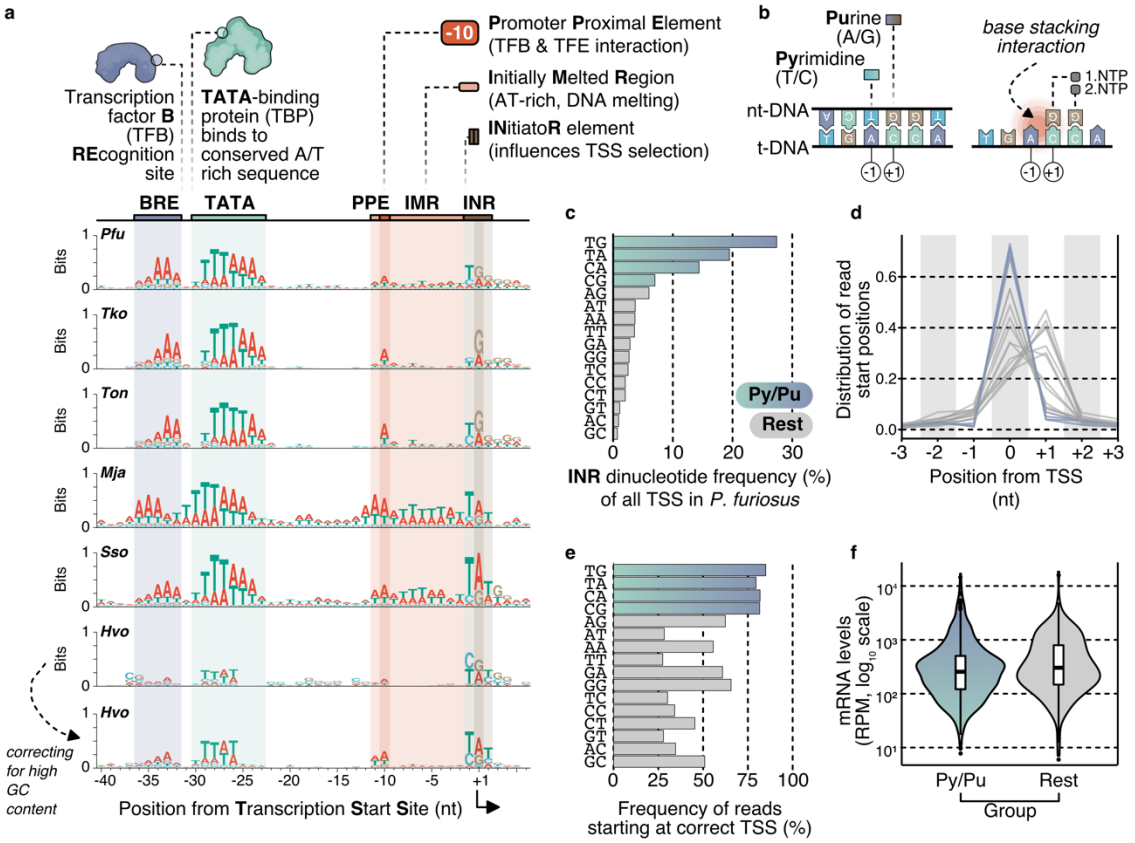


Figure 33 | Promoter elements and start site selection. a, Consensus motifs of promoters from selected archaeal species between -40 and +5 from the transcription start site (TSS). Functional important sequence elements are highlighted in purple (BRE), green (TATA box), red (PPE), orange (IMR) and brown (INR). Primary TSSs have been extracted from publications describing the primary transcriptome of the respective organism (*Pfu* = *Pyrococcus furiosus* (Grünberger et al., 2019), *Tko* = *Thermococcus kodakarensis* (Jäger et al., 2014), *Ton* = *Thermococcus onnurineus* (Cho et al., 2017), *Mja* = *Methanocaldococcus jannaschii* (Smollett et al., 2017b), *Sso* = *Sulfolobus solfataricus* (Wurtzel et al., 2010), *Hvo* = *Haloferax volcanii* (Babski et al., 2016). Additionally, the promoter motif of primary transcripts in *H. volcanii* is shown after correcting for the genomic GC content. **b**, PIC stabilisation by base stacking interactions formed between a pyrimidine/purine dinucleotide at the -1/+1 position, which affects promoter strength and accuracy of start site selection (Hausner et al., 1991; Ao et al., 2013; Basu et al., 2014; Smollett et al., 2017b). **c**, Frequency of INR dinucleotides derived from all (primary, secondary, internal, antisense) TSS in *P. furiosus* (Grünberger et al., 2019). Pyrimidine/Purine (Py/Pu) combinations are highlighted in blue-green. **d**, Accuracy of start site selection for Py/Pu and all other (grey, rest) dinucleotides is estimated by analysing the distribution of read start positions in close vicinity to the predicted TSSs. **e**, Relative quantification of reads start counts between -3 and +3. **f**, No statistical relevant difference is observed between the mRNA levels of transcripts starting with a Py/Pu dinucleotide and the rest of the transcripts in *P. furiosus*.

Indeed, promoter motif comparisons of TFB paralogues revealed slight deviations in the BRE elements, although it is unclear whether this is sufficient for discrimination (Seitzer et al., 2012; Smollett et al., 2017b). Opposingly, the information content of all moderate BRE and TATA elements might be acceptable for factor recognition in the GC-rich halophilic genomes, which is supported after GC-correcting the *H. volcanii* motif (Figure 33a). In a functional context, specialised sigma-like promoters have not been observed in Archaea for different transcript classes, which have been assigned based on their relative gene orientation as primary, secondary, internal or antisense (Jäger et al., 2014; Grünberger et al., 2019). Within the Thermococcales, the BRE and TATA elements in all classes resemble one another very much and are clearly detectable, while the INR motif is absent in *T. kodakarensis* (Toffano-Nioche et al., 2013; Jäger et al., 2014; Cho et al., 2017; Grünberger et al., 2019).

The pyrimidine/purine dinucleotide at the -1/+1 position plays a role in DNA strand stabilisation within the PIC in Bacteria by forming base-stacking interactions, which affects promoter strength and precise TSS selection *in vivo* and *in vitro* (Hausner et al., 1991; Ao et al., 2013; Basu et al., 2014; Smollett et al., 2017b; Yu et al., 2017) (Figure 33b). In *P. furiosus*, the vast majority of INRs independent of the transcription class consist of a pyrimidine/purine dinucleotide (T/G, T/A, C/A, C/G), which triggers transcription to start from the correct position in about 80% of the cases (Figure 33c,d,e). However, re-analysing the read start positions also revealed that transcripts starting with another dinucleotide are characterised by heterogeneous 5´ ends, which has already been shown in *M. jannaschii* (Smollett et al., 2017b) (Figure 33c). Start site selection is especially bad for TSS starting with a pyrimidine, indicating functional consequences provided that step-size based annotation by ANNOgesic is correct (Yu et al., 2018) (Figure 33e). Although the INR contributes to the promoter strength, there is no statistically relevant difference in the mRNA levels of the genes that start transcription at a pyrimidine/purine compared to the rest (Figure 33f). It is interesting to note that dinucleotide selection for a precise transcription start is an important promoter feature not only in Archaea but also in Bacteria and Eukaryotes, caused by the universal RNAP active site architecture (Shultzaberger et al., 2007; Werner and Grohmann, 2011; Basu et al., 2014).

In addition to the three core elements which are known to contribute to promoter strength, there is an additional loosely defined element called IMR that ranges from about -12 to +2 relative to the TSS. This region is characterized by an enriched AT content, which facilitates strand separation of the DNA during OC formation based on less stable base pair interactions. The upstream edge of the IMR has a special status, is more conserved in many organisms and was therefore called PPE (promoter proximal element). This element is centred approximately at the -10 position, which is in good agreement with the border of the transcription bubble (Spitalny and Thomm, 2003; Nagy et al., 2015). Consequently, footprinting data highlighted the importance of this region for TFEα/β,

which triggers strand opening in archaea (Blombach et al., 2015). Additionally, the PPE helps in recruiting TFB by direct interactions and has been shown to increase transcription (Peng et al., 2009; Gehring et al., 2016). For *P. furiosus* the *in vivo* TSS data, that reveal a particularly A/T rich signal at the -10 position correlate very well with the strongest footprint observed for TFB at this region (Micorescu et al., 2008). While the PPE was first described as a feature limited to leaderless RNAs, it is also prevalent in leadered transcripts in our data (Torarinsson et al., 2005; Grünberger et al., 2019).

It is well established, that all promoter elements contribute to the promoter strength, which conceptually is defined as an idealised sequence. This comprises of elements that maximise the initial recruitment of TBP and TFB, facilitate strand separation and open complex formation. However, promoter strength does only modestly or not at all correlate with mRNA levels *in vivo*, highlighting the not yet fully understood impact of gene-specific regulators, post-transcriptional mechanisms, transcription-translation coupling and occupancy by histone-like proteins for transcriptional regulation in Archaea (Figure 34a) (French et al., 2007; Clouet-D'Orval et al., 2018; Blombach et al., 2019; Gomes-Filho et al., 2019). Interestingly, genome-wide occupancy data for all basal TFs and the RNAP in *M. jannaschii* revealed that PIC formation is critical for transcription. However, TBP and TFB levels did not correlate with promoter strength (Smollett et al., 2017b).

Re-evaluating promoter signatures and expression data from *P. furiosus* confirm the presence of more complex transcriptional regulatory mechanisms. The comparison to the RNA levels of mixed growth conditions presumably allows obtaining a more balanced picture as they cover multiple potential regulatory states (Laass et al., 2019). Despite the overall low correlation, splitting the genes up in divergent and non-divergent transcripts influences the correlation. Divergent transcripts seem to be higher regulated than the rest (Figure 34a). In an evolutionary context, this could be an efficient way to use a single TF for the transcriptional control of multiple targets, which we could show for CopR (Grünberger et al., 2020b). After copper-shock, CopR stimulates transcription dependent on the distance between the TSS of divergent genes. Therefore, a minimum distance of about 120 bases is sufficient to allow binding of the CopR complex in between two adjacent promoters, which allows simultaneous stimulation of both transcripts. This finding goes in line with the weak promoter scores for PF0740, PF0706.1n, PF0727 and PF0738, that are all highly upregulated after copper shock (Figure 34a,b).

Nevertheless, small intergenic distances allow not only efficient activation but also repression, which makes them a hotspot for transcriptional regulation. It is tempting to speculate that this is reflected in the absent correlation of promoter strength to the expression level for divergent genes with small intergenic regions (Figure 34a). However, the features contributing to promoter strength need more balanced data sets that would allow taking the regulatory mechanisms of multiple TFs and effects on the transcriptome into consideration.

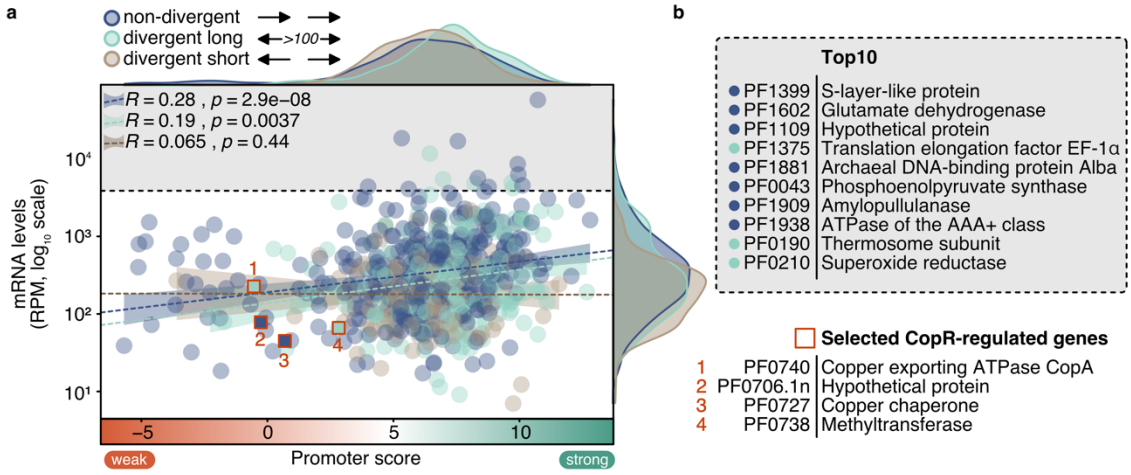


Figure 34 | Correlation of promoter strength and mRNA levels in *P. furiosus*. **a**, Promoter scores have been calculated for all primary transcripts in *P. furiosus* using MOODS and compared to the mRNA levels in reads per million of the mixed RNA-sequencing (Korhonen et al., 2009; Grünberger et al., 2019). Note that scores bigger than 6 are considered a good promoter. Transcripts have been grouped according to their relative orientation to the next TSS and the distance to non-divergent (purple), divergent-long (light-green, distance between 2 adjacent annotated gene start positions >100 bases) and divergent-short (brown). Pearson´s correlation R measuring the linear dependency between the two variables is given for each set. **b**, List of the 10 most abundant mRNAs in *P. furiosus* in the mixed RNA-seq conditions (Grünberger et al., 2019). Additionally, selected CopR-regulated genes are shown and can be assigned according to the numbers to the plot on the left (Grünberger et al., 2020b).

2.2. Bidirectional promoters

A particular case of divergent promoters that we identified for the first time in Archaea are bidirectional promoters that originate transcripts in both directions from a shared TATA box and two mirrored BRE elements (Grünberger et al., 2019). Transcriptional polarity in Archaea is determined by the interaction of TFB with the BRE element and the orientation to the TATA box (Bell et al., 1999c). Mechanistically this is fundamentally different from eukaryotic promoters, where bidirectionality can be stimulated by the positioning of activator or enhancer binding sites and not by the intrinsic symmetry of the core promoter (Haberle and Stark, 2018). This configuration often results in antisense RNAs that arise thousands of bases away from the sense transcripts and are harder to assign than in Archaea. Although more than 80% of eukaryotic promoters are characterised by bidirectional transcription, nearly all of the asRNAs are only the by-product of RNAP II transcription and functionally irrelevant (Jin et al., 2017). However, other consequences that have been proposed in Eukaryotes might also play a role in the selection and evolution of bidirectional promoters in Archaea. For instance, widespread divergent transcription is a hallmark of active promoters, by creating and maintaining a nucleosome-depleted region, that allows subsequent regulation and stimulation of transcription (Core et al., 2008; Seila et al., 2008; Wu and Sharp, 2013). Besides, it was

shown that recently acquired DNA is enriched in bidirectional promoters, which could be one of the fundamental driving forces for new gene formation by shaping asRNAs in a functional mRNA context (Wu and Sharp, 2013; Jin et al., 2017). In conclusion, this indicates that directionality is an evolutionary process in Eukaryotes, that involves the selection of DNA binding sites and TFs.

Interestingly, a recently published pre-print described the evolutionary and functional causes of bidirectional promoters in prokaryotes (Warman et al., 2020). While the key ideas, namely that intrinsic symmetry of TATA boxes in Archaea and -10 elements in Bacteria cause bidirectional transcription, go in line with our findings in *P. furiosus*, they additionally proposed that bidirectional promoters are significantly enriched in horizontally acquired genes. Hence, the directionality of promoters is not only shaped during evolution in Eukarya but also prokaryotes. Strikingly, their analysis revealed that 19% of all detected TSS in *E. coli* show bidirectional signatures, starting transcription at opposing strands separated by 18 bps. Reanalysing available TSS maps of different species, strong bidirectional transcription was also revealed in proteobacteria, actinobacteria, firmicutes and the archaeal organisms *T. kodakarensis* and *H. volcanii*. They conclude that two scenarios lead to bidirectional transcription in Archaea: Predominantly, 52 bp-spaced divergent TSSs arise from symmetric TATA boxes, which we also observed in *P. furiosus*. Second and less frequently, transcription can start at the CG dinucleotide of a BRE element (5´-CGAAA-3´), if a second TATA box is positioned downstream of the first one (Warman et al., 2020). Based on these data, bidirectional transcription in Archaea follows sequence- and position-specific rules that can be used to re-evaluate the TSS map from *P. furiosus* in a more quantitative context and to compare it to the readcount-based approach in our paper (Grünberger et al., 2019) (Figure 35).

Therefore, the distances of TSSs on the top DNA strand to the nearest TSS on the bottom strand was first calculated and compared to the distribution in *T. kodakarensis* (Figure 35a,b). As expected, most of the TSSs partners are close to each other in both organisms and also peak at -52 bp in *P. furiosus* (Figure 35c). To extract the bidirectional promoters, all TSSs that have a divergent TSS between -60 and -30 were filtered out as suggested in the pre-print. Accordingly, 190 promoters (10% of all assigned TSSs) could be identified, that have a consensus sequence, which resembles the bidirectional promoter we already obtained in the re-annotation publication (Figure 35d,e). Additionally, this approach allowed the assignment of promoters to different TSS classes (Figure 35f). Most of these transcripts are primary transcripts that have a divergent primary or antisense partner. One additional promoter element that is missing on the bottom strand promoter is the INR, which leads to less precise start site selection and many reads originating from positions -50 and -51 on the negative strand (Figure 35e,g). Although one could assume that a second BRE element has a negative influence on sense transcription by concurring

TFBs or PICs, there was no significant difference on the mRNA levels to genes preceded by a consensus promoter (Figure 35h).

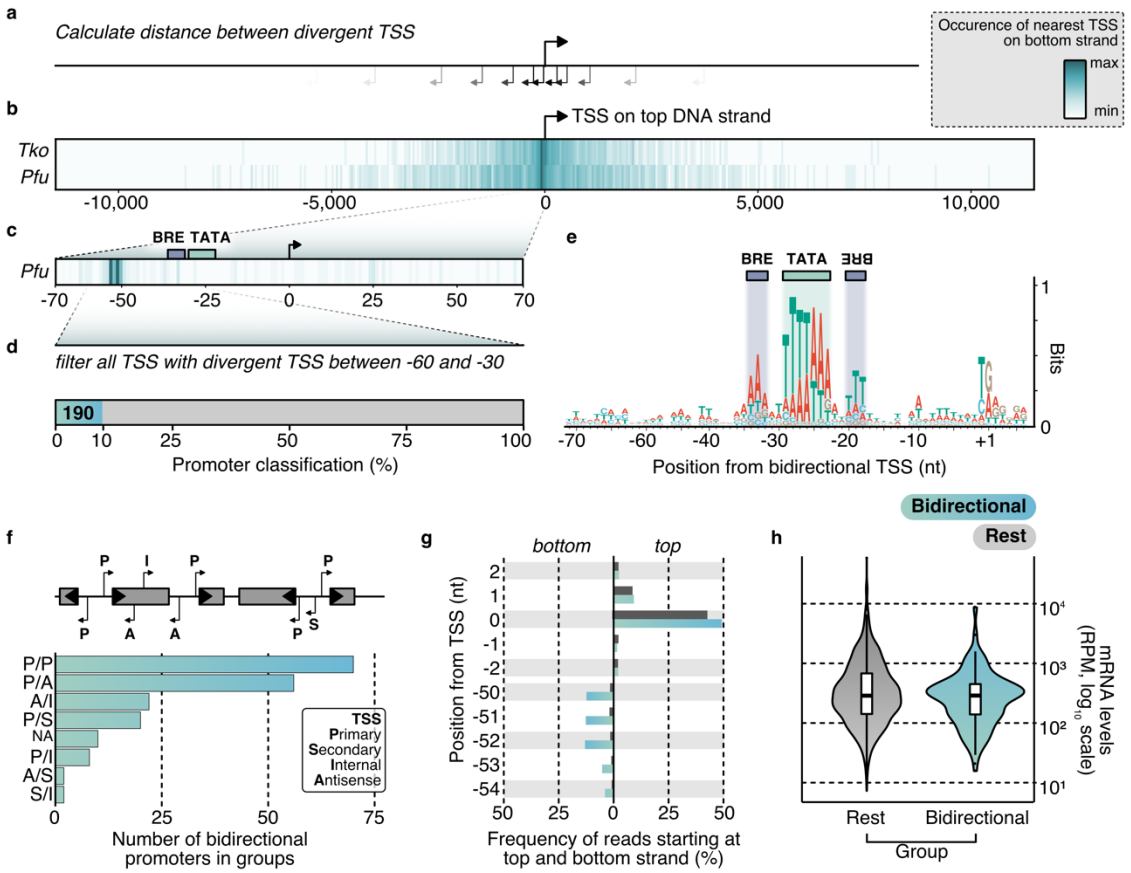


Figure 35 | Re-evaluation of bidirectional promoters in *P. furiosus*. **a**, Based on the method described in Warman et al., 2020, first the distance between divergent TSS was calculated for transcripts in *T. kodakarensis* and *P. furiosus* (Jäger et al., 2014; Grünberger et al., 2019). **b**, The occurrence of nearest TSS on bottom strand to the fixed top strand TSS (position 0) is shown as a heatmap ranging from -10000 to 10000 bases depending on the TSS. **c**, A **z**oom into region -70 to 70 reveals that most of the top strand TSS have a corresponding divergent bottom strand TSS at about -52 in *P. furiosus* and most likely arise from a bidirectional promoter. **d**, Quantification of bidirectional promoters (all TSS with a divergent TSS partner at a distance between -60 and -30). **e**, Consensus motif of the 190 bidirectional promoters reveals a mirrored BRE element. **f**, Classification of the TSS pairs of bidirectional promoters. **g**, Quantification of read start positions on top and bottom strand depending from a TSS (position 0). The colors refer to all TSS that are transcribed from a bidirectional promoter (green-blue) and all other TSS (grey). **h**, Abundance of mRNAs (reads per million) under mixed growth conditions that are transcribed from bidirectional (green-blue) or other (grey) promoters (Grünberger et al., 2019).

In addition to the read-based approach, Warman et al. introduced a symmetry-score model that only depends on the sequence information and allows to classify the promoters based on their intrinsic symmetry (Warman et al., 2020). Following this approach, an archaea-specific version of this model was implemented to identify bidirectional promoters *de novo* for multiple archaeal model organisms (Figure 36). Therefore, the detection is based on a high symmetry of the top and bottom strand between -36 bp and -18 bp, which

includes the BRE-TATA-BRE elements. By comparing PWMs for both strands over sliding windows, a symmetry-score was calculated and used to classify the promoters as bidirectional or not. The highest symmetry was achieved at position -26, which bisects the TATA box. This position is enriched in the bidirectional promoters assigned by the comparison of divergent TSS, which validates the model (Figure 36a,b). This approach was applied to *de novo* detect bidirectional promoters from TSS maps of 5 archaeal species, that all showed a BRE-TATA-BRE signature in about 10% of the cases (Figure 36c). It should be noted that the *M. jannaschii* promoter seems to be shifted two bases upstream, which is also reflected in the position of the second BRE element (Figure 36d). Given the high GC content and the less defined BRE in *H. volcanii,* the model logically failed and would need additional adjusting for the genomic GC content. For *P. furiosus* the *de novo* symmetry detection (n = 104) and the divergent TSS-based detection (n = 90) only have a 40% overlap.

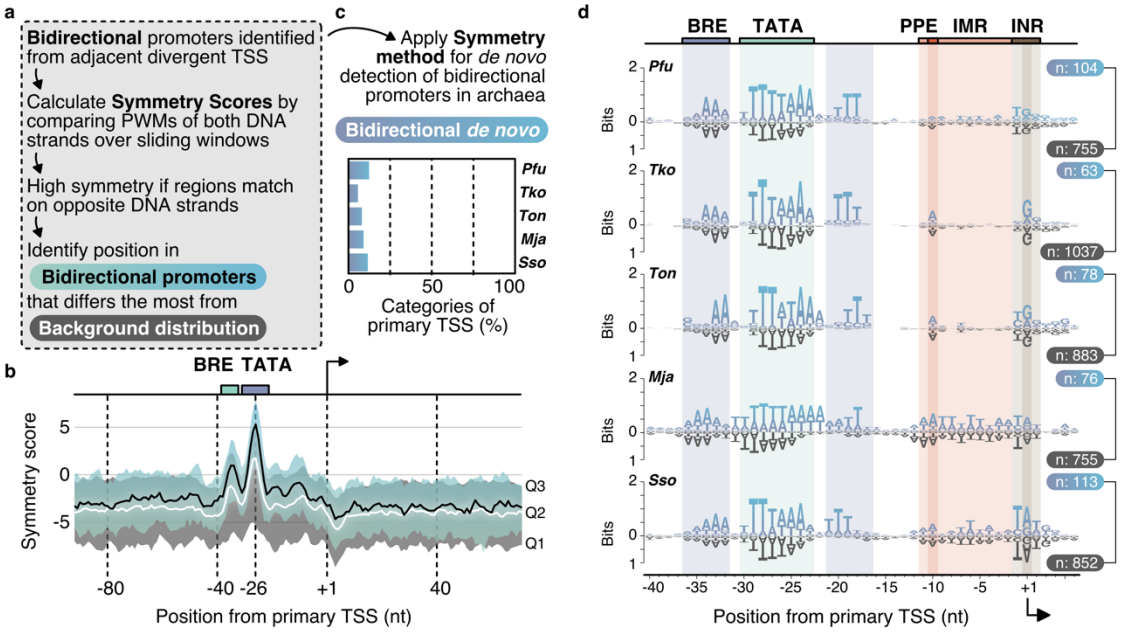


Figure 36 | Symmetry-score based quantification of bidirectional promoters. a, Rational behind the symmetry-score based detection of bidirectional promoters in prokaryotic organisms as defined in Warman et al., 2020. **b**, Comparison of the calculated symmetry-scores for bidirectional promoters identified form adjacent divergent TSS (n = 190, green-blue, compare Figure 35) and the background distribution of all other promoters. Median is shown as a solid line, quantiles as shaded areas. **c**, The symmetry-score model was applied for the *de novo* detection of bidirectional promoters in selected archaeal species (Wurtzel et al., 2010; Jäger et al., 2014; Cho et al., 2017; Smollett et al., 2017b; Grünberger et al., 2019). The plot shows the relative number of primary TSS that have a bidirectional promoter. **d**, Consensus motifs of *de novo* detected bidirectional promoters (purple-blue) and all other primary TSS (grey).

Although it is reasonably robust to predict bidirectional promoters *de novo*, the limited overlay between the motif-based and TSS-based detection method indicates that a subset of the expected asRNA output of bidirectional regulators cannot be detected. The

functional and regulatory principles yet have to be explored. Nevertheless, the reanalysis can serve as an initial appraisal for future in-depth characterisation of bidirectional transcription in Archaea.

In summary, the directionality of transcription is not apparent, and although there are fundamental differences in the promoter landscapes of prokaryotes and eukaryotes, the evolutionary ground state seems to be bidirectional. Another important point is that *E. coli* cells lacking a histone-like nucleoid structuring protein (H-NS) have an increased number of bidirectional transcripts, which raises the question about the role of histone-like proteins in Archaea (Singh et al., 2014; Warman et al., 2020). However, it remains to be determined if the arising antisense transcripts are nonsense transcripts without any function and are terminated by chance or if they fulfil a currently unknown function.

2.3. Re-visiting archaeal terminators

Transcription termination is defined as the release of the RNA from the elongation complex and can occur via factor-dependent or independent mechanisms in Archaea (Maier and Marchfelder, 2019; Sanders et al., 2020). Despite the recent identification of FttA, which Santangelo and colleagues entitled as the "last missing piece of the archaeal transcription cycle", the current picture of termination is somewhat puzzling (Sanders et al., 2020). Using long-read native RNA sequencing, we not only identified TTS in *H. volcanii* and *P. furiosus* but also provided a single-molecule method that allows the detection of heterogeneous 3´ ends (Grünberger et al., 2020a). Our data confirmed recent Term-seq results from *Sulfolobus acidocaldarius, Methanosarcina mazei* and *H. volcanii* and also numerous *in vitro* studies that all conclude that during intrinsic termination the archaeal RNAP is sensitive to poly(T) signals of varying lengths independent of secondary structures (Dar et al., 2016; Maier and Marchfelder, 2019; Berkemer et al., 2020a). This is in contrast to intrinsic termination in Bacteria that relies on the formation of a GC-rich RNA-hairpin 8 to 9 nt from the 3´ end, that effects conformational changes and ultimately destabilises the complex (Gusarov and Nudler, 1999; Roberts, 2019).

Interestingly, multipartite mechanisms seem to be favoured in all forms of life for effective termination of stable elongation complexes that synthesise long transcripts, except for Archaea. Their strategy resembles the termination of eukaryotic RNAP III after the transcription of short tRNAs and 5S rRNAs, that is solely triggered by short poly(T) sequences and independent of other *cis*-elements or *trans*-acting factors (Arimbasseri et al., 2013b). Specifically, weak interactions of the template strand and the RNA, but also interactions of the non-template strand with the RNAP are required for efficient transcript release. The inevitable control sequence for the 17-subunit RNAP III, containing the subunits C53/37/11, is a stretch of 5 or more T residues, where T3 and T4 are required for the formation of a metastable pre-termination complex, and T5 is the final signal for transcript release (Maraia et al., 2015). This is distinct from the RNAP III core mechanism,

that is independent of C53/37/11 and requires 8 to 9 T´s for efficient termination, which is only rarely the case in the genomes of Eukaryotes (Iben and Maraia, 2012; Arimbasseri and Maraia, 2013).

The defined sequence context of archaeal intrinsic termination varies among species, and especially *in vitro* seems to be heavily affected by multiple experimental parameters (Maier and Marchfelder, 2019). In some cases, like the first detected archaeal termination signal for the tRNA-Val from *Methanococcus vannielii,* only four consecutive T´s are necessary and sufficient for termination, although with the help of an upstream sequence (Thomm et al., 1994). However, *in vitro* studies more commonly revealed that single T-stretches between five and eight residues terminate transcription effectively in *P. furiosus* (T_5), *M. thermoautotrophicus* (T_6), *M. jannaschii* (T_7) and *T. kodakarensis* (T_8) without the help of any other signals (reviewed in Maier and Marchfelder, 2019). On a global level, this is not only in contrast to the *in vivo* detected longer, less T-strict termination motifs, but also to the occurrence of consecutive T stretches in the downstream region of genes in different archaeal taxonomic classes (Figure 37a). Comparing the proportion of T_3 to T_8 sequences in a GC- and position-specific context, it is obvious (i) that downstream regions (200 nt from stop codon) are enriched in poly(T)s and (ii) that there is a strong correlation to the genomic GC content (Figure 37a). For all of the groups, a T_8 sequence, used very efficiently *in vivo* in *T. kodakarensis*, is very rare (Santangelo et al., 2009). However, in this example, termination already occurred at positions 4,5 and 6 within the T stretch and could even be partly replaced by a $(TA)_4$ sequence, which highlights that there is some degree of flexibility, allowing termination in GC-rich organisms. For instance, halophilic organisms like *H. volcanii* only have five or more consecutive T residues downstream of 13% of the genes, which explains the significantly shorter consensus sequence that has been found in Term-seq and native RNA-seq studies (Berkemer et al., 2020a) (Figure 37a,b).

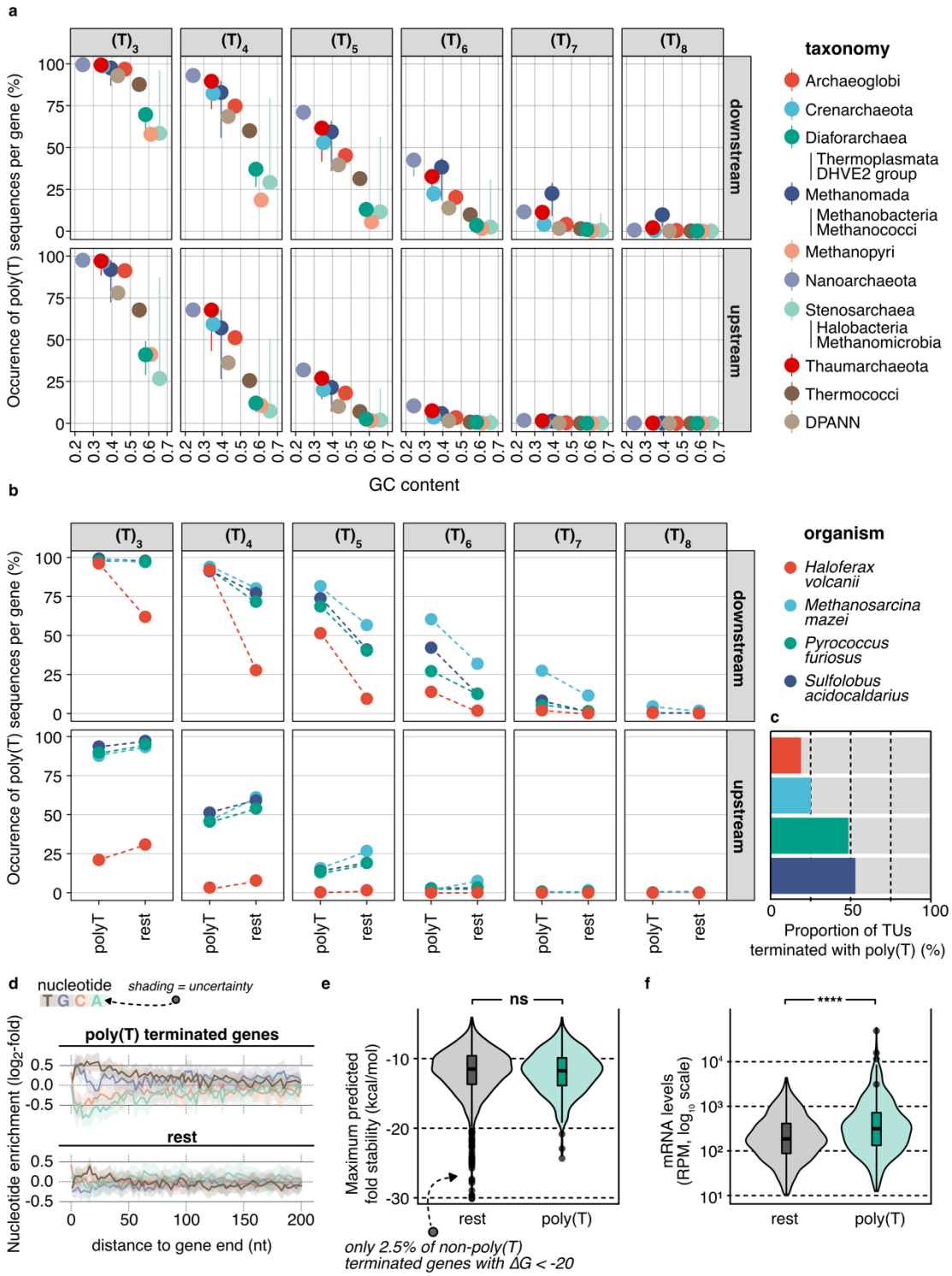


Figure 37 | Re-evaluation of archaeal poly(T)-based transcription termination. a, Occurence of poly(T)$_n$ (n = 3-8) stretches upstream (200 bases upstream) and downstream (200 bases downstream) from annotated gene ends in different archaeal classes. Points represent mean of the respective phylogenetic class with standard deviations indicated by vertical lines. **b**, Occurrence of poly(T) sequences in archaeal organisms with available genome-wide termination data. Terminations sites classified as intrinsically terminated by T-stretches have been extracted from the publications and compared to the rest (Dar et al., 2016; Berkemer et al., 2020b; Grünberger et al., 2020a). **c**, Absolute quantification of poly(T) terminated transcriptional units. **d**, Nucleotide enrichment analysis of poly(T) terminated genes (derived from Nanopore data) and the rest in *P.*

furiosus. **e**, Secondary structure analysis calculated by RNAFold in a 44 base sliding window (Lorenz et al., 2011). For every gene only the maximum predicted fold stability is given. Statistical testing was performed using a t-test (ns = not significant). **f**, Comparison of mRNA levels derived from mixed RNA short read sequencing for poly(T) (green) and other terminated (grey) genes (t-test statistic, ****: p value < 0.0001).

Similar to the recognition of weak haloarchaeal BRE elements of TFB during transcription initiation, the RNAP might also be sensitive towards the genomic GC content, which is however speculative at the moment given the conserved core architecture of the archaeal RNAP (Werner and Grohmann, 2011). Hence, this immediately raises the question of what fraction of the genes *in vivo* is terminated in a purely factor-independent way. Unfortunately, quantitative aspects and a more in-depth discussion considering the interplay with poly(T) terminators is missing in the recent publication of FttA, the archaeal ribonuclease aCPSF1 from *T. kodakarensis* (Sanders et al., 2020). Interestingly, aCPSF1 from *Methanococcus maripaludis* was recently characterised using Term-seq of a *Mmp-aCPSF1* depleted strain (Yue et al., 2020). They could show that depletion of aCPSF1 leads to an increased readthrough also at poly(T) terminators, which links intrinsic to factor-dependent termination in Archaea and resembles the eukaryotic RNAP II termination mode (Hill et al., 2019). The identification of aCPSF1 as the essential archaeal termination factor will open up new opportunities to look at archaeal transcription termination and help to explain 3´ heterogeneity in Archaea.

On a genome-wide scale, it is interesting to see a drop from T_4 to T_5 in all analysed species, that have available TTS data, which might represent the minimum requirement for a conserved mechanism (Figure 37a,b). In order to reconsider possible mechanistic differences between poly(T) and non-poly(T) terminated genes in *P. furiosus*, the nucleotide content, fold stability and mRNA levels were analysed (Figure 37d,e,f). The proposed Rho-like C-over-G enrichment for FttA activity, could not be seen in either one of the data sets, which could be owed to the limitation to downstream sequences or an averaging effect (Figure 37d). Also, secondary structures do not play a role at all or only very rarely, confirming the consensus of many *in vitro* and *in vivo* studies in Archaea (Washio et al., 1998; Unniraman, 2002; Dar et al., 2016) (Figure 37e). Analogous to *E. coli*, genes with intrinsic termination mechanisms in *P. furiosus* had a higher expression value than the rest, which was already suggested to be an indicator for an additional level of gene regulation (Dar and Sorek, 2018) (Figure 37f).

One possibility that allows the synthesis of large amounts of poly(T) terminated transcripts is the direct and very efficient transition of the RNAP from termination to initiation on the same template by moving in the 1D direction (Arimbasseri et al., 2013a). This is the critical mechanism for the eukaryotic RNAP III, possibly facilitated by an unusual loose fit of the active centre around the DNA-RNA hybrid, and also has been suggested for the termination of the archaeal histone *hpyA1* from *P. furiosus* (Spitalny and

Thomm, 2008; Hoffmann et al., 2015). In the latter case, termination is achieved at multiple T-stretches at 90°C but not at 80°C (Spitalny and Thomm, 2008). The intriguing point is, that after termination the RNAP can reinitiate on the same template, which however heavily depends on sequence elements that balance the rate of pausing at poly(T) sites and downstream GC-rich regions that lower the translocation rate (Spitalny and Thomm, 2008).

Future research will show if this is a particular case or if reinitiation plays a more general role in archaeal termination. Unfortunately, the timely order of the termination process, from the release of the RNA, over structural rearrangements of the RNAP and final disassembly, is currently unknown in Archaea and only allows speculations regarding RNAP recycling. Interestingly, a single-molecule FRET study of intrinsic bacterial termination revealed that in 94% of the cases, the RNA is released first, which creates a previously unknown post-terminational stage for the RNAP, termed recycling. After the release of the product, the RNAP stays bound on the template and diffuses along in 1D direction, until it either initiates at the next downstream promoter or reinitiates at the original promoter in upstream direction (Kang et al., 2020). In either case, this RNAP III-like strategy for the synthesis of highly expressed mRNAs in Archaea and Bacteria seems to disagree with the current model of transcriptional and translational coupling in prokaryotes, although this recently has been challenged by a study in *B. subtilis* that showed that these molecular processes are disjointed (French et al., 2007; McGary and Nudler, 2013; Johnson et al., 2020). However, in *E. coli* coupling is necessary to prevent premature termination, as the ribosomal protein S10 can bind to the NusG KOW domain and block the recruitment of Rho (Burmann et al., 2010; Proshkin et al., 2010; Kohler et al., 2017). In Archaea, FttA termination efficiency has also been shown to be dependent on the association of the NusG homologue Spt4/5 to the RNAP, which indicates a similar Rho-like mechanism that allows immediate termination once the expressome complex reaches the stop codon and is uncoupled.

Unfortunately, the endonuclease enzyme activity of FttA in Archaea produces heterogenous 3´ends that complicate the analysis (Sanders et al., 2020). In contrast to many *in vitro* studies, Term-seq data revealed that termination stops at a precise position in Archaea and Bacteria (Dar et al., 2016; Dar and Sorek, 2018). In this context, current results of short-read technologies should be taken with care and might not reflect the actual mechanism as library preparation and especially bioinformatical analysis can introduce errors that favour precise 3´ ends.

Based on our results, we believe that diverse termination mechanisms occur in Archaea that might depend on the stochastic termination at multiple consecutive T-stretches. Performing single-read analysis of the long Nanopore reads, we were able to analyse long 3´ UTRs of selected genes and additionally confirmed the *in vitro* predicted usage of multiple T-stretches for the termination of one gene *in vivo* (Spitalny and Thomm, 2008; Grünberger et al., 2020a). However, the current 3´ end accuracy is limited by

basecalling and mapping problems of long homopolymer regions that are clipped off, which is likely to be improved very soon.

In conclusion, Nanopore sequencing provides a single-molecule method allowing for the detection of underrepresented sub-transcripts, and heterogeneous 3´ ends, that will be helpful for the detection and re-evaluation of termination sites in prokaryotes in the future. The recent identification of FttA provides a new possibility to consider the interplay of poly(T) sequences and termination factors and could potentially help to understand the role of imperfect terminators, that appear multiple times downstream of a gene in Archaea and results in heterogenous 3´ends (Koide et al., 2009). Some results also indicate that termination has a bigger role in maintaining high levels of gene expression than previously estimated. In that context, imperfect poly(T) sequences and downstream elements have been shown to influence RNAP activity, with dramatic impacts on mRNA levels once this balance is disturbed (Spitalny and Thomm, 2008).

3. Regulation beyond basal transcription

During evolution, organisms have developed vital mechanisms to adapt to changing environmental conditions rapidly (López-Maury et al., 2008). To this end, gene expression in prokaryotes can be regulated at multiple levels, e.g. by the action of dedicated TFs and post-transcriptional events. While TFs usually turn genes on or off by determining the initiation rate of transcription, processing events fine-tune mRNA expression and are crucial for maturation and functioning of other RNA classes (Peeters et al., 2013; Clouet-D'Orval et al., 2018). In the manuscripts that are part of this dissertation, we have addressed both of these aspects, on the one hand by describing the essential role of CopR for copper detoxification in *P. furiosus* and by interrogating the transcriptome architecture in Archaea in general and on the other hand by analysing the multi-step rRNA maturation pathway in Archaea and other features of post-transcriptional regulation (Grünberger et al., 2019, 2020a, 2020b).

3.1. Pre-transcriptional regulation

The reaction of microbes to environmental fluctuations is balanced at different temporal scales and accordingly can be divided into short-term and long-term adaptions (Brooks et al., 2011; Zorraquino et al., 2017). One of the primary examples for transient changes is represented by the exposure to toxic compounds which demands a rapid reaction of the cell in the milliseconds to minutes range (Holmqvist and Wagner, 2017). In contrast, prolonged changes, e.g. in nutrient availability, may have more dramatic effects by leading to a stream-lining of the genome by gene loss (Koonin and Wolf, 2010).

In our manuscript, we investigated the short-term response (20 minutes) of *P. furiosus* to elevated copper concentrations and identified the metalloregulator CopR as the essential TF during this process in *P. furiosus* (Grünberger et al., 2020b). By integrating RNA-seq and ChIP-seq data, we found that CopR remains bound to target promoters under normal and copper-shock (20 µM) conditions. Additional 2D structural analysis and size exclusion chromatography revealed an octameric conformation in both forms. Together, these findings suggest an allosteric mechanism triggered by availability of copper ions with CopR constantly bound to the DNA, which is only rarely the case for metal-specific regulators (Peeters and Charlier, 2010; Lemmens et al., 2019; Baksh and Zamble, 2020). Most commonly, activation or repression is achieved depending on the binding state and the relative position of the recognition motif towards basal promoter elements (Peeters et al., 2013).

In contrast, we believe that CopR in *P. furiosus* activates transcription upon copper binding possibly by altering the topology of the promoter. Among bacterial TFs, members of the MerR family respond with a similar mechanism to environmental stimuli like drugs, chemical agents and most interestingly also metals (Hobman et al., 2005; Baksh and Zamble, 2020). For instance, the Cu(I)-responsive regulator CueR of *E. coli* forms a

homodimer and coordinates metals by a binding-loop at the dimerisation helix (Changela et al., 2003). Metal-binding leads to a restructuring of the loop, which pulls the DNA-binding domains closer together and ultimately distorts the DNA structure (Philips et al., 2015; Sameach et al., 2017). This allosteric mechanism enables repression in the metal-unbound state and activation after a structural switch by twisting the DNA, which promotes binding of the RNAP and stimulates transcription (Martell et al., 2015; Sameach et al., 2017).

In general, metal detoxification requires a rapid response of the respective TF to stimulate the expression of detoxifying components, like copper exporters (Bini, 2010; Zorraquino et al., 2017). However, after metal homeostasis is restored, transcription has to be turned off again as fast as possible not to waste energy. Considering the temporal dynamics, especially the unbinding of the regulator is a critical step to allow the functional switch. This is evident from the fact that copper ions are not released from the binding pocket due to their tight interactions. Hence, the activated DNA-bound regulator has to be replaced by the metal-free version and vice versa (Changela et al., 2003; Foster et al., 2014). While our data do not have the temporal resolution to investigate the dynamics of the binding, this has been analysed in detail for CueR (Joshi et al., 2012; Chen et al., 2015). Indeed, multiple mechanisms facilitate the swap between copper-free and copper-bound CueR, are advantageous for copper regulation in *E. coli* and may also play a role during CopR regulation in Archaea:

First, using single-molecule Förster resonance energy transfer (FRET) measurements, it was shown that CueR flips its binding orientation at the recognition site, which might improve dynamic protein-protein interactions with the RNAP (Joshi et al., 2012). Considering the activation at 120 bp-spaced divergent promoters, it remains speculative if CopR has a similar flipping-mechanism or if the octameric conformation allows stimulation in both directions, which we currently think is more likely (Grünberger et al., 2020b).

In contrast to CopR, CueR binds to nonspecific DNA, which accelerates scanning along the DNA for its recognition site due to dimensionality reduction (Joshi et al., 2012). This concept of facilitated diffusion has been studied in detail for LacR in *E. coli*. Accordingly, nonspecific binding enables sliding in 1D and allows the TF to finds its target considerably faster than with diffusion-search in 3D (Hammar et al., 2012; Suter, 2020). Search efficiency in 1D is further affected by the properties of the TF and the recognition element (Suter, 2020).

Interestingly, FRET measurements also showed, that free CueR proteins enable the dissociation of DNA-bound CueR in a concentration-dependent way (Joshi et al., 2012). Considering the archaeal nucleoid organisation, it can be assumed that protein-assisted displacement would also be beneficial for CopR and facilitate rapid reactions upon copper shock. Archaeal genomes are either occupied by a number of eukaryotic-like histones (Euryarchaeota) or bacterial-like nucleoid-associated proteins (Crenarchaeota), that

compete with other DNA-binding proteins like TFs (Peeters et al., 2015). Hence, the constant binding of copper-free and copper-bound CopR would prevent promoter occlusion by any other factors.

In addition to the protein-facilitated dissociation of CueR, single-molecule tracking *in vivo* suggested that the unbinding rate is also influenced by the chromosomal organisation and therefore links transcriptional regulation to the cell cycle (Chen et al., 2015). In particular, at less condensed bacterial chromosomes, which are a sign of cellular stress, the copper-bound activating CueR remained longer on the DNA leading to an extended transcriptional activation. However, it is unclear to what extent this can be interpolated to the higher chromosomal organisation of stressed euryarchaeal genomes.

In conclusion, the discussed mechanisms add another layer of complexity at the pre-transcriptional level that significantly influences the temporal dynamics of the biological systems. Higher-resolved structural and genome-wide binding data will be necessary to reveal which of these principles hold true for the maintenance of copper homeostasis by CopR in *P. furiosus*.

3.2. Post-transcriptional regulation & RNA stabilisation

Although transcription is the primary regulatory point in gene expression, post-transcriptional events add to the increasing functionality and complexity of RNAs (Evguenieva-Hackenberg and Klug, 2011). There is growing evidence that these mechanisms significantly contribute to the tolerance and adaption to unfavourable environmental conditions, which is essential for prokaryotes (Picard et al., 2009; Tollerson and Ibba, 2020). Additionally, RNA processing steps provide critical checkpoints to yield fully functional transcripts. In contrast to Eukaryotes, RNA maturation in prokaryotes is in particular (but not exclusively) important for tRNAs and rRNAs and requires the tight coordination between RNA structures, chemical base modifications and mostly terminal cleavage by additional enzymes on a defined timescale (Ferreira-Cerca, 2017; Clouet-D'Orval et al., 2018).

In general, each step in the life cycle of an RNA constitutes an opportunity for fine-tuning the regulatory process. One of the most important aspects, especially for mRNAs, is the balance between the initial transcription rate and RNA decay, which ultimately affects translation (Stoecklin and Mühlemann, 2013; Durand et al., 2015; Mohanty and Kushner, 2016). Interestingly, the half-lives of some archaeal mRNAs are significantly longer than observed in the bacterial counterparts and range from 2 to 20 min in *Sulfolobus* to up to 80 min in *H. mediterranei* compared to a median half-live of only about 5.4 min in *E. coli* with most stable transcripts not exceeding 15 min (Bini et al., 2002; Jäger et al., 2002; Andersson et al., 2006; Nouaille et al., 2017). However, considering the similarity of essential molecular processes, it is not immediately obvious why some archaeal transcripts have longer half-lives (Clouet-D'Orval et al., 2018). To get a better look at the complexity

of RNA stability ultimately more data are required, to unravel the influence of ribonucleases, but also secondary structures, 5´- UTR lengths, small regulatory RNAs and potentially the binding of archaeal Sm-like proteins (Evguenieva-Hackenberg and Klug, 2011; Babski et al., 2014; Cao et al., 2014; Märtens et al., 2017). In comparison, eukaryotic mRNAs have substantially longer half-lives, as they require even more complex co-regulated and coordinated steps, including nuclear export, splicing, 5´capping, 3´polyadenylation, and RNA modifications (Tian and Manley, 2016; Zhao et al., 2016; Manning and Cooper, 2017).

Notably, deep sequencing studies of prokaryotic total RNAs and the development of adapted library preparation protocols to capture primary or processed transcripts have improved transcriptome annotations, allowed the identification of transcript cleavage sites and revealed an unprecedented high number of small noncoding RNAs (Bernick et al., 2012; Babski et al., 2014). To enrich for primary transcripts carrying a 5´-triphosphate, we used a protocol developed by Sharma et al. which is based on the use of a Terminator Exonuclease (Sharma et al., 2010; Sharma and Vogel, 2014). Following this dRNA-seq approach on mixed RNAs, we were not only able to map primary TSS with a single-nucleotide resolution but also detected a large number of antisense (797) and internal (739) TSS (Grünberger et al., 2019). This finding is in line with previously published datasets from *H. volcanii* (1851 pTSS, 1244 aTSS) and other archaeal species (Jäger et al., 2014; Babski et al., 2016).

Moreover, we showed that antisense RNAs are predominantly transcribed from an archaeal consensus promoter (83%) and are significantly enriched at the 5´end of transposase open reading frames, possibly downregulating these elements and thereby preventing genomic rearrangements during stress conditions, which were part of the mixed RNA pool (Grünberger et al., 2019). Transposon-encoded asRNAs have already been identified in many archaeal species, including *S. solfataricus*, *T. kodakarensis*, *M. mazei* and *H. volcanii* (Tang et al., 2005; Jäger et al., 2009, 2014; Straub et al., 2009; Wurtzel et al., 2010; Märtens et al., 2013), and typically inhibit translation directly by complementary base-pairing with the mRNA (Arini et al., 1997; Ellis et al., 2015; Ellis and Haniford, 2016).

However, despite their functionality in transposon-silencing, the global post-transcriptional impact of asRNAs in *P. furiosus* and other Archaea remains elusive. Performing dRNA-seq and *in vitro* analysis, we could show that asRNAs can be the bi-product of bidirectional transcription from symmetric promoters, which according to the promoter re-analysis based on symmetry-scores, is a widespread phenomenon (compare Figure 35, Figure 36). Future work will reveal what proportion of asRNAs are nonsense RNAs that are rapidly degraded and how many indeed have a meaningful biological function. Importantly, transcriptome information from bulk deep sequencing studies need to be critically considered as a snapshot of all processes that are going on in multiple cells at the same time, including active transcription, as well as processing in different cell states.

Hence, the development of additional screening methods in combination with knocking down RNA-binding or processing enzymes allows a more in-depth analysis of post-transcriptional events in Archaea (Clouet-D'Orval et al., 2018). For this purpose, many different approaches have already been published, and transcriptome analysis of single cells will unquestionably be the essential tool to reveal heterogeneity at the transcriptional and post-transcriptional level (Blattman et al., 2020; Imdahl et al., 2020). While to date, single-cell RNA sequencing has not been applied to Archaea for technical reasons, deep sequencing data have been used frequently to identify the regulatory roles of small RNAs. Recently, a modified version of the dRNA-seq protocol was applied to obtain a genome-wide processing map in *Methanolobus psychrophilus* L15 (Qi et al., 2017). This approach revealed enrichment of processing sites in intergenic and 5´UTR regions of polycistronic operons encoding ribosomal proteins, suggesting that post-transcriptional mechanisms mainly regulate protein synthesis in this context (Qi et al., 2017). Furthermore, ncRNAs are key players for nitrogen-regulation in *M. mazei* and many other metabolic but also stress responses in Archaea (Straub et al., 2009; Marchfelder et al., 2012; Babski et al., 2014; Buddeweg et al., 2018). Interestingly, based on our differential gene expression analysis, not a single ncRNA in *P. furiosus* seems to be involved in copper detoxification as none of the 413 annotated non-coding transcripts is significantly up- or down-regulated (data not shown in Grünberger et al., 2020b).

With heat as a constant stressor, hyperthermophilic organisms have to ensure efficient regulation at elevated temperatures. Therefore, they have developed multiple strategies to protect DNAs and also RNAs, which are otherwise thermolabile, from denaturation and premature degradation (Grogan, 1998; Gomes-Filho et al., 2019). As the additional hydrogen bond in GC base pairs immediately provides higher stability, increasing the GC content offers the most straightforward stabilisation strategy. However, after sequencing the first complete archaeal genomes, it was somewhat unexpected to see that the genomes of many hyperthermophilic organisms are rather AT-rich (Daniel and Cowan, 2000) (Figure 38). Nevertheless, already at that time, it was well-established that various mechanisms are used to stabilise chromosomal DNA, like positive supercoiling by a reverse gyrase, histones, increased intracellular ionic concentrations and high polyamine content (Kikuchi and Asai, 1984; Marguet and Forterre, 1998; Higashibata et al., 1999).

In contrast, the mechanisms of RNA stabilisation are very different and based on the ability to form secondary structures, circularisation events, RNA modifications and RNA-binding proteins (Gomes-Filho et al., 2019). Noncoding RNAs in hyperthermophiles frequently form duplex stretches that are favoured by GC base pairs and provide an increased hydrogen bond stability. Indeed, there is a strong correlation between the GC content of rRNAs, tRNAs, additional ncRNA classes, like ribonuclease P (RNase P) and signal recognition particle (SRP), and the optimal growth temperature (Galtier and Lobry, 1997; Hurst and Merchant, 2001; Nakashima et al., 2003) (Figure 39). Comparing the GC

content of major RNA classes from the hyperthermophilic organisms *P. furiosus* and *M. jannaschii* to the mesophilic *H. volcanii* supports this concept (Babski et al., 2016; Smollett et al., 2017b; Grünberger et al., 2019; Laass et al., 2019) (Figure 39). Notably, all structured RNA classes deviate with more than 20% from the global average (40.7% GC) in *P. furiosus*, which is also true for *M. jannaschii*, but not for the different RNA classes in *H. volcanii*. Although C/D box small nucleolar RNAs (snoRNAs), comprehensively annotated by our dRNA-seq approach in *P. furiosus*, also form stem-loops, their intramolecular structures are less conserved, more flexible, thus less GC-rich and rely on the interaction with protein partners (Gaspin et al., 2000; Klein et al., 2002; Rozhdestvensky et al., 2003; Grünberger et al., 2019).

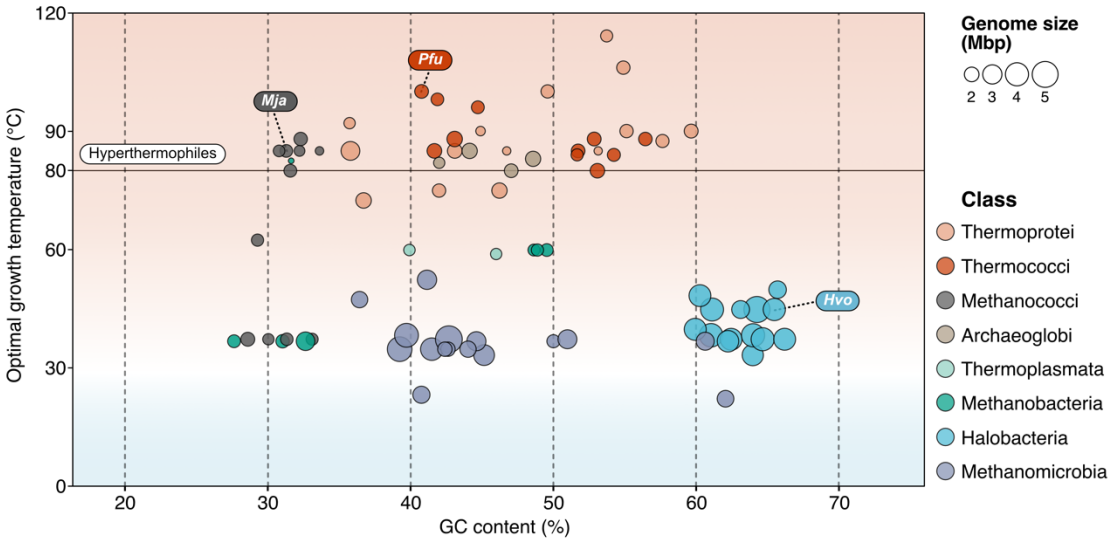


Figure 38 | GC content is not correlated to the optimal growth temperature in Archaea. Comparison of the genomic GC content (in %) to the optimal growth temperature (°C) of 80 archaeal species. Genome sizes are additionally indicated by circle sizes. Class membership is color-coded on the right.

In general, the majority of noncoding RNAs, which have been identified in transcriptome studies in hyperthermophilic Archaea (excluding the functional classes already described) do not display high GC contents (Toffano-Nioche et al., 2013; Jäger et al., 2014; Cho et al., 2017; Smollett et al., 2017b; Grünberger et al., 2019). To protect them from degradation in the natural habitat, some ncRNAs are stabilised by RNA-binding proteins (RBPs), that specifically recognise structural or sequence-specific RNA motifs. These include the interaction between Sm-like proteins and U- (or A-) rich tracts, k-turn binding by archaeal L7Ae proteins and the stabilisation of CRISPR RNAs by Cas proteins, amongst others (Rozhdestvensky et al., 2003; Murina and Nikulin, 2011; Lilley, 2012; Reichelt et al., 2018a).

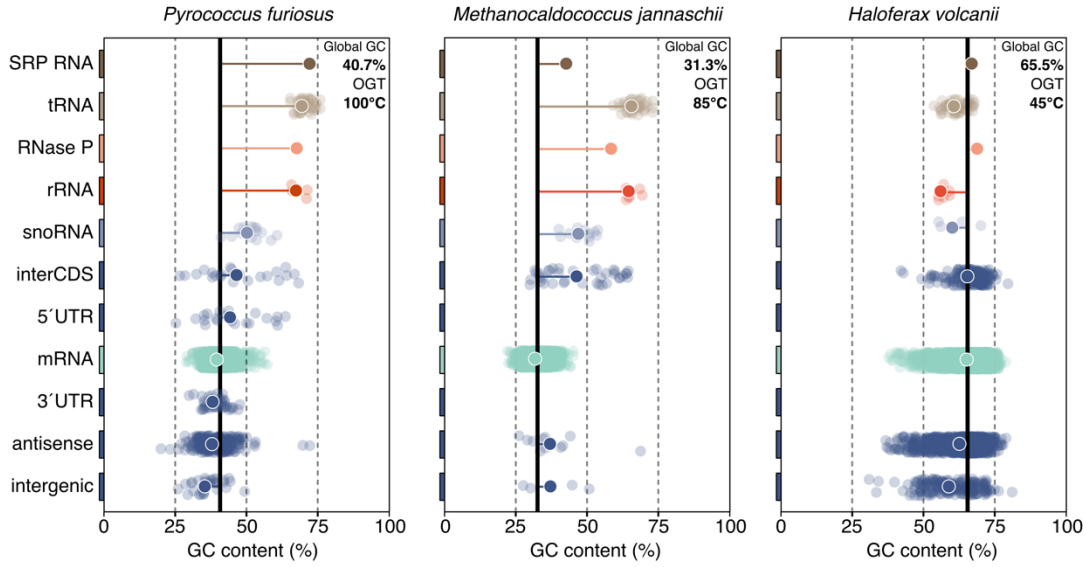


Figure 39 | GC content from selected RNA classes in different archaeal species. The different RNA classes have been assigned based on transcriptomic studies in the archaeal organisms (Babski et al., 2016; Smollett et al., 2017b; Grünberger et al., 2019; Laass et al., 2019). Solid points with white borders show the mean GC content of the RNA class, while individual transcripts are plotted as transparent points. Distance from the mean RNA class GC content to the mean genomic GC content is indicated by horizontal lines. The mean global GC content and the optimal growth temperature (OGT) are listed in the upper right corner of each plot.

Ribosomal RNAs of hyperthermophilic Archaea are not only a prime example for the thermo-adaptive role of RNA base modifications but display a variety of mechanisms that are equally important for their stability and functionality (Ferreira-Cerca, 2017; Sas-Chen et al., 2020). Indeed, most of the principles apply to all archaeal rRNAs independent of the respective growth temperature. Using Nanopore sequencing, we analysed the ribosomal maturation pathway in Archaea, which starts with the transcription of the primary polycistronic rRNA precursor and includes Archaea-specific circularisation and processing events to finally yield mature rRNAs that are incorporated in the ribosome (Tang et al., 2002; Yip et al., 2013; Ferreira-Cerca, 2017; Clouet-D'Orval et al., 2018; Grünberger et al., 2020a; Jüttner et al., 2020; Qi et al., 2020). The circular pre-rRNA intermediates are generated after cleavage of the primary precursor at the bulge-helix-bulge (BHB) containing processing stems by the splicing endonuclease endA, followed by the covalent ligation of the resulting ends (Russell et al., 1999; Tang et al., 2002; Ferreira-Cerca, 2017; Clouet-D'Orval et al., 2018; Jüttner et al., 2020). Although our data did not allow a quantitative analysis of the circular precursors, most likely due to problems sequencing highly-structured RNAs, we could show for the first time that they are also present in *P. furiosus* (Grünberger et al., 2020a). Based on the current knowledge, circularisation is an efficient way to (i) provide protection against nuclease activities, (ii) stabilise the rRNA scaffold, (iii) decrease distance between terminal ends and ultimately facilitate early steps of ribosome biogenesis (Ferreira-Cerca, 2017). Circularisation events are not limited to

rRNAs, but also are advantageous for other RNAs and have been observed in a permuted SRP RNA version of *Thermoproteus tenax* (Plagens et al., 2015). This RNA lacks the helix H1 forming sequences and forms a BHB structure that is further processed by the tRNA splicing machinery and results in a thermo-stable circular SRP RNA (Brown et al., 1993; Plagens et al., 2015; Gomes-Filho et al., 2018). Future studies, including more diverse organisms, will show if circularisation is truly a conserved intermediate step during archaeal ribosomal biogenesis and to what extent it additionally contributes to survival under challenging environmental conditions.

In contrast to the archaeal-specific circular rRNA precursors, post-transcriptional rRNA modifications are present across all domains of life (Lafontaine et al., 1998; Piekna-Przybylska et al., 2008; Ferreira-Cerca, 2017; Sloan et al., 2017). Nanopore sequencing of native rRNAs provided a unique opportunity to identify and sort the consecutive rRNA intermediate states in *H. volcanii* and *P. furiosus* and simultaneously record the base modification profiles (Grünberger et al., 2020a). Focussing on the universally conserved KsgA-dependent dimethylation and the less conserved Kre33/Nat10-dependent N^4-cytidine acetylation, allowed us to confirm the general hypothesis that modifications are introduced during ribosome biogenesis (O'Farrell et al., 2006; Grosjean et al., 2008; Xu et al., 2008; Ebersberger et al., 2014; Smith et al., 2019; Grünberger et al., 2020a). Moreover, our data are in line with the prediction that halophilic Archaea have a reduced set of modifications that are all clustered in the active centre of the ribosome (Decatur and Fournier, 2002; Grosjean et al., 2008; Ferreira-Cerca, 2017). Although many rRNA modifications still lack an in-depth functional analysis, this distribution suggests that these have a stabilising effect and may fine-tune translation (Ferreira-Cerca, 2017; Sloan et al., 2017). Remarkably, the number of base modifications in hyperthermophilic Archaea is dramatically increased and supposed to be a general feature of thermoadaption (Dennis et al., 2015; Gomes-Filho et al., 2019; Coureux et al., 2020; Sas-Chen et al., 2020).

Recently, N^4-cytidine acetylation at CCG motifs has been shown to be prevalent in the 30S ribosomal initiation complex of *P. abyssi,* which is supported by the higher amount of systematic basecalling errors and shifts in the raw current in our Nanopore data of the closely related *P. furiosus* (Coureux et al., 2020; Grünberger et al., 2020a). Notably, ac^4C-seq, developed in the Schwartz lab, revealed that acetylations are highly abundant not only in rRNAs but also tRNAs, ncRNAs and mRNAs (Sas-Chen et al., 2020). Additionally, Sas-Chen et al. provided evidence for the thermoadaptive role of RNA acetylations in the archaeal order Thermococcales and suggest that ac^4C catalysis by the acetyltransferase may be statistical, hence not affecting every single motif. Re-analysis of the *P. furiosus* Nanopore data in a single-molecule context using a better *in vitro* transcribed background model could help to address this feature, resolve the minimum percentage of CCG motifs that have to be modified to stabilise the ribosome and challenge the current deterministic concept of RNA modifying enzymes (Gomes-Filho et al., 2019; Sas-Chen et al., 2020).

To conclude, post-transcriptional regulatory mechanisms are increasingly gaining attention and are nowadays acknowledged for critically contributing to the survival of organisms. Especially hyperthermophilic Archaea are a prime example for diverse RNA stabilisation strategies to protect otherwise heat-labile RNAs (Gomes-Filho et al., 2019). Nanopore sequencing of native RNAs will be a valuable single-molecule-sensitive addition to the prokaryotic transcriptomic toolbox that can be tailored to address different post-transcriptional features and detect hitherto unknown heterogeneity.

4. Conclusion: Perspectives

The chimeric setup of fundamental cellular processes in Archaea allows a unique opportunity to study key molecular events in a simplified genomic context. Despite their discovery more than 40 years ago, the archaeal field still significantly lacks behind research in Bacteria and Eukaryotes in many ways. While our studies could contribute to a deeper knowledge concerning genome and transcriptome architecture, general and regulatory aspects of transcription and post-transcriptional mechanisms in Archaea in general and specifically in *P. furiosus*, they also brought up new conceptual ideas, that can be addressed in the future. Currently, some fundamental questions are limited by the absence of techniques, that have already been applied to bacterial and eukaryotic organisms, but yet wait to be established in Archaea. Notably, single-cell RNA sequencing recently applied to capture the transcriptomes of individual bacterial cells, will revolutionise transcriptomic analysis and enable to address dynamic cell-to-cell variability (Blattman et al., 2020; Imdahl et al., 2020). Innovations in high-throughput technologies and adaptions to current sequencing library preparation methods will hopefully allow the genome-wide measurement of RNA secondary structures, the identification of hitherto unknown RNA-binding partners and reveal the regulatory roles of base modifications on both DNA and RNA. Considering that in 2020 the bacterial transcription and translation field has been reshaped by in-depth analysis of the supramolecular expressome complex, formed by the ribosome and the RNAP, it will be one of the most exciting and important issues to decipher the occurrence and the structural and functional consequences of transcription-translation coupling in Archaea (Johnson et al., 2020; O'Reilly et al., 2020; Wang et al., 2020; Webster et al., 2020). Additionally, the revival of cultivation-based approaches, possibly supported by the pre-screening of metabolic pathways and nutritional requirements using metagenomics/metatranscriptomics/metaproteomics, will allow addressing evolutionary aspects of critical cellular processes.

Zusammenfassung

Archaeen werden aktuell als zweite Domäne des Lebens und als wichtiger Bestandteil aller biogeochemischen Abläufe der Erde anerkannt. Vor ihrer Entdeckung durch Carl Woese und Kollegen in den späten 1970er Jahren, wurden sie allerdings nicht wahrgenommen und fälschlicherweise für Bakterien gehalten, da diese meist anhand mikroskopischer Merkmale nicht voneinander zu unterscheiden sind. Seit ihrer Klassifizierung als drittes primäres „Königreich“ im Jahr 1990 hat sich nicht nur ihre Stellung im universellen Baum des Lebens geändert und die archaische Abstammung der Eukaryoten definiert. Auch das Wissen über ihre Ökologie, Diversität, Evolution und Molekularbiologie wurde enorm erweitert und vertieft. Bemerkenswerterweise teilen Archaeen auf molekularer Ebene erstaunlich viele Charakteristika mit Eukaryoten und Bakterien, wobei der Vorgang der Transkription als Musterbeispiel gilt.

Diese Arbeit beschäftigt sich vor allem mit der grundlegenden Architektur des Genoms und Transkriptoms, der regulatorischen Rolle von Transkriptionsfaktoren und post-transkriptionellen Mechanismen im hyperthermophilen Modellorganismus *Pyrococcus furiosus*. Um für zukünftige Studien eine möglichst exakte und informationsreiche Grundlage zu schaffen, wurde zunächst das Genom des Typ Stammes DSM 3638 durch einen Hybridansatz aus Illumina und PacBIO Sequenzierung bestimmt. Außerdem wurde die Annotation auf dem Transkript-Level durch einen differentiellen RNA Sequenzieransatz erweitert. Der Verdau aller Transkripte ohne 5´-Triphosphat durch eine

Terminatorexonuklease erlaubte uns das primäre Transkriptom von *P. furiosus* neu zu definieren, inklusive der genomweiten Bestimmung aller Transkriptionsstarts, Promoter-Architekturen, und der Beschreibung von sense und antisense RNAs. Interessanterweise konnten wir feststellen, dass symmetrische Promotoren eine häufige Ursache für die bidirektionale Transkription von sense und antisense RNAs sind, was ein genereller Mechanismus in Archaeen zu sein scheint. Außerdem hat sich gezeigt, dass ein normaler Umgang mit einer Laborkultur über einen Zeitraum von zwei Jahren nicht zu einer Umgruppierung von Genen führt, obwohl es sehr viele IS-Elemente im 2 Millionen Basenpaare großen Genom von *P. furiosus* gibt. Obwohl wir die Integrität nicht in besonderer Weise herausgefordert haben, bedeutet dies dennoch, dass das Genom stabiler ist als bisher angenommen, was eine Grundvoraussetzung für die Vergleichbarkeit, Reproduzierbarkeit und damit die Durchführbarkeit von zukünftigen genomweiten Studien in *Pyrococcus* darstellt. Für eine schnelle und kosteneffiziente Re-Sequenzierung von archaeellen Stämmen wurde die Nanopore Technologie im Labor etabliert. Dies ist eine Sequenziermethode der dritten Generation, die es uns, dank der Eigenschaft sehr lange zusammenhängende Sequenzstücke generieren zu können, ermöglicht hat, den Laborstamm mit hoher Konsensus-Genauigkeit zu sequenzieren.

Darauf aufbauend wurde ein Protokoll für die direkte Sequenzierung von RNA in Prokaryoten basierend auf der Nanopore Technologie entwickelt, was aktuell noch die einzige Möglichkeit darstellt, Einzelmolekülsequenzierung im nativen Kontext durchzuführen. Die Fülle an transkriptionellen und posttranskriptionellen Ereignissen und Merkmalen auf genomweiter Ebene wird normalerweise durch Illumina Sequenzierungsansätze bearbeitet. Je nach Fragestellung gibt es dabei Variationen und Adaptionen in der Vorbereitung der Sequenzierbibliothek oder es werden im Vorlauf chemische Behandlung an der RNA vorgenommen. Im Gegensatz dazu haben wir das Potenzial der nativen RNA Sequenzierung evaluiert, gleichzeitig verschiedene transkriptionelle Merkmale in einem Bakterium (*Escherichia coli*) und in Archaeen (*Haloferax volcanii*, *P. furiosus*) zu adressieren. Durch die Analyse von Meta- und Einzelmoleküldaten waren wir in der Lage lange Transkriptionseinheiten neu zu annotieren und Transkriptgrenzen akkurat zu definieren. Außerdem konnten wir zeigen, dass lange Sequenzierstücke ein wertvolles Werkzeug für die Erkennung heterogener Terminationsstellen sind und dass in Archaeen verschiedene Termiationsmechanismen auftreten. Insbesondere haben wir das Einzelmolekül-Potenzial dazu genutzt, um durch die Beschreibung zusätzlicher Intermediate den bisher nur unzureichend verstandenen Reifungsprozess der ribosomalen RNA in Archaeen nachzuvollziehen. Darüber hinaus waren wir in der Lage, RNA-Basenmodifikationen in Form systematischer Basecalling-Fehler und Verschiebungen im aufgezeichneten Profil des Ionenstroms nachzuweisen, wodurch wir die relative zeitliche Abfolge der KsgA-abhängigen Dimethylierung und N^4-Acetylierung in reifer und Vorläufer-16S-rRNA nachverfolgen konnten.

Als drittes Projekt, wurde das neue Referenzgenom als Grundlage dafür genutzt, um in einem integrativen RNA-seq- und ChIP-seq-basierten Ansatz die Funktion des Transkriptionsregulators CopR während der Kupferentgiftung in *P. furiosus* zu entschlüsseln. Um einen globalen Überblick über die transkriptomische Reaktion zu erhalten und Komponenten des CopR-Regulons zu identifizieren, führten wir nach Kupferschock eine differentielle Genexpressionsanalyse und eine ChIP-seq-Analyse durch. Wir haben festgestellt, dass CopR, das für die Kupferentgiftung essentiell ist, an die stromaufwärts gelegenen Regionen stark kupferinduzierter Gene bindet, die alle durch ein gemeinsames palindromisches Motiv gekennzeichnet sind. Zusätzlich zeigte die Transmissionselektronenmikroskopie von negativ-gefärbten Präparaten und anschließende Bildanalyse durch Mittelung der 2D-Klassen, dass CopR, ähnlich wie andere Faktoren der Lrp-Familie, in einer oktameren Konformation an die DNA bindet. Basierend auf den Ergebnissen schlugen wir ein Modell für die allosterische Regulation von CopR nach Kupferbindung vor und beschrieben die verschiedenen Stadien der Kupferentgiftung bei *P. furiosus.*

Die Ergebnisse der Manuskripte, aus denen diese Arbeit besteht, tragen zu einem tieferen Verständnis der grundlegenden und regulatorischen Prinzipien der Transkription in Archaeen bei und liefern zudem ein Update der genomischen und transkriptomischen Landschaft von *P. furiosus.* Außerdem stellt die Anwendung der auf Nanoporen basierenden nativen RNA-Sequenzierung nicht nur eine bedeutende Erweiterung des Methodenspektrums für das Transkriptionsfeld dar, sondern lieferte uns auch eine Fülle von Informationen, insbesondere zu transkriptionellen und posttranskriptionellen Ereignissen während der rRNA-Reifung.

Bibliography

Adams, M. W. W., Holden, J. F., Menon, A. L., Schut, G. J., Grunden, A. M., Hou, C., et al. (2001). Key role for sulfur in peptide metabolism and in regulation of three hydrogenases in the hyperthermophilic archaeon Pyrococcus furiosus. *J. Bacteriol.* doi:10.1128/JB.183.2.716-724.2001.

Adessi, C. (2000). Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res.* doi:10.1093/nar/28.20.e87.

Agarwal, S., Hong, D., Desai, N. K., Sazinsky, M. H., Argüello, J. M., Rosenzweig, A. C., et al. (2010). Structure and interactions of the C-terminal metal binding domain of Archaeoglobus fulgidus CopA. *Proteins Struct. Funct. Bioinforma.* 78, 2450–2458. doi:10.1002/prot.22753.

Akeson, M., Branton, D., Kasianowicz, J. J., Brandin, E., and Deamer, D. W. (1999). Microsecond time-scale discrimination among polycytidylic acid, polyadenylic acid, and polyuridylic acid as homopolymers or as segments within single RNA molecules. *Biophys. J.* doi:10.1016/S0006-3495(99)77153-5.

Akıl, C., and Robinson, R. C. (2018). Genomes of Asgard archaea encode profilins that regulate actin. *Nature*. doi:10.1038/s41586-018-0548-6.

Albers, S.-V. (2016). Extremophiles: Life at the deep end. *Nature* 538, 457–457. doi:10.1038/538457a.

Albers, S.-V. V., and Jarrell, K. F. (2015). The archaellum: how archaea swim. *Front. Microbiol.* 6. doi:10.3389/fmicb.2015.00023.

Allers, T., and Mevarech, M. (2005). Archaeal genetics — the third way. *Nat. Rev. Genet.* 6, 58–73. doi:10.1038/nrg1504.

Andersson, A. F., Lundgren, M., Eriksson, S., Rosenlund, M., Bernander, R., and Nilsson, P. (2006). Global analysis of mRNA stability in the archaeon Sulfolobus. *Genome Biol.* doi:10.1186/gb-2006-7-10-r99.

Ao, X., Li, Y., Wang, F., Feng, M., Lin, Y., Zhao, S., et al. (2013). The Sulfolobus initiator element is an important contributor to promoter strength. *J. Bacteriol.* doi:10.1128/JB.00768-13.

Aparicio, O., Geisberg, J. V., Sekinger, E., Yang, A., Moqtaderi, Z., and Struhl, K. (2005). Chromatin Immunoprecipitation for Determining the Association of Proteins with Specific Genomic Sequences In Vivo. *Curr. Protoc. Mol. Biol.* 69. doi:10.1002/0471142727.mb2103s69.

Aravind, L., and Koonin, E. V (1999). DNA-binding proteins and evolution of transcription regulation in the archaea. *Nucleic Acids Res.* 27, 4658–70. doi:10.1093/nar/27.23.4658.

Argüello, J. M. (2003). Identification of Ion-Selectivity Determinants in Heavy-Metal Transport P1B-type ATPases. *J. Membr. Biol.* 195, 93–108. doi:10.1007/s00232-003-2048-2.

Arimbasseri, A. G., and Maraia, R. J. (2013). Distinguishing Core and Holoenzyme Mechanisms of Transcription Termination by RNA Polymerase III. *Mol. Cell. Biol.* doi:10.1128/mcb.01733-12.

Arimbasseri, A. G., Rijal, K., and Maraia, R. J. (2013a). Comparative overview of RNA polymerase II and III transcription cycles, with focus on RNA polymerase III termination and reinitiation. *Transcription*. doi:10.4161/trns.27369.

Arimbasseri, A. G., Rijal, K., and Maraia, R. J. (2013b). Transcription termination by the eukaryotic RNA polymerase III. *Biochim. Biophys. Acta - Gene Regul. Mech.* doi:10.1016/j.bbagrm.2012.10.006.

Arini, A., Keller, M. P., and Arber, W. (1997). An antisense RNA in IS30 regulates the translational expression of the transposase. *Biol. Chem.* doi:10.1515/bchm.1997.378.12.1421.

Ashby, M. K. (2006). Distribution, structure and diversity of "bacterial" genes encoding two-component proteins in the Euryarchaeota. *Archaea.* doi:10.1155/2006/562404.

Babski, J., Haas, K. A., Näther-Schindler, D., Pfeiffer, F., Förstner, K. U., Hammelmann, M., et al. (2016). Genome-wide identification of transcriptional start sites in the haloarchaeon Haloferax volcanii based on differential RNA-Seq (dRNA-Seq). *BMC Genomics* 17, 629. doi:10.1186/s12864-016-2920-y.

Babski, J., Maier, L.-K., Heyer, R., Jaschinski, K., Prasse, D., Jäger, D., et al. (2014). Small regulatory RNAs in Archaea. *RNA Biol.* 11, 484–493. doi:10.4161/rna.28452.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208. doi:10.1093/nar/gkp335.

Baker-Austin, C., Dopson, M., Wexler, M., Sawers, R. G., and Bond, P. L. (2005). Molecular insight into extreme copper resistance in the extremophilic archaeon "Ferroplasma acidarmanus" Fer1. *Microbiology* 151, 2637–2646. doi:10.1099/mic.0.28076-0.

Baksh, K. A., and Zamble, D. B. (2020). Allosteric control of metal-responsive transcriptional regulators in bacteria. *J. Biol. Chem.* doi:10.1074/jbc.REV119.011444.

Balch, W. E., Magrum, L. J., Fox, G. E., Wolfe, R. S., and Woese, C. R. (1977). An ancient divergence among the bacteria. *J. Mol. Evol.* 9, 305–311. doi:10.1007/BF01796092.

Baliga, N. S., Bonneau, R., Facciotti, M. T., Pan, M., Glusman, G., Deutsch, E. W., et al. (2004). Genome sequence of Haloarcula marismortui: A halophilic archaeon from the Dead Sea. *Genome Res.* doi:10.1101/gr.2700304.

Baliga, N. S., Goo, Y. A., Ng, W. V., Hood, L., Daniels, C. J., and DasSarma, S. (2000). Is gene expression in Halobacterium NRC-1 regulated by multiple TBP and TFB transcription factors? *Mol. Microbiol.* doi:10.1046/j.1365-2958.2000.01916.x.

Banerjee, S., Chalissery, J., Bandey, I., and Sen, R. (2006). Rho-dependent transcription termination: More questions than answers. *J. Microbiol.*

Bang, C., and Schmitz, R. A. (2015). Archaea associated with human surfaces: Not to be underestimated. *FEMS Microbiol. Rev.* doi:10.1093/femsre/fuv010.

Basen, M., Schut, G. J., Nguyen, D. M., Lipscomb, G. L., Benn, R. A., Prybol, C. J., et al. (2014). Single gene insertion drives bioalcohol production by a thermophilic archaeon. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.1413789111.

Basu, R. S., Warner, B. A., Molodtsov, V., Pupov, D., Esyunina, D., Fernández-Tornero, C., et al. (2014). Structural Basis of Transcription Initiation by Bacterial RNA Polymerase Holoenzyme. *J. Biol. Chem.* 289, 24549–24559. doi:10.1074/jbc.M114.584037.

Bayega, A., Oikonomopoulos, S., Zorbas, E., Wang, Y. C., Gregoriou, M.-E., Tsoumani, K. T., et al. (2018). Transcriptome landscape of the developing olive fruit fly embryo delineated by Oxford Nanopore long-read RNA-Seq. *bioRxiv*, 478172. doi:10.1101/478172.

Bechhofer, D. H., and Deutscher, M. P. (2019). Bacterial ribonucleases and their roles in RNA metabolism. *Crit. Rev. Biochem. Mol. Biol.* 54, 242–300. doi:10.1080/10409238.2019.1651816.

Bell, S. D., Cairns, S. S., Robson, R. L., and Jackson, S. P. (1999a). Transcriptional regulation of an archaeal operon in vivo and in vitro. *Mol. Cell* 4, 971–982. doi:10.1016/S1097-2765(00)80226-9.

Bell, S. D., and Jackson, S. P. (1998). Transcription and translation in Archaea: A mosaic of eukaryal and bacterial features. *Trends Microbiol.* 6, 222–227. doi:10.1016/S0966-842X(98)01281-5.

Bell, S. D., and Jackson, S. P. (2000). Mechanism of autoregulation by an archaeal transcriptional repressor. *J. Biol. Chem.* doi:10.1074/jbc.M005422200.

Bell, S. D., Kosa, P. L., Sigler, P. B., and Jackson, S. P. (1999b). Orientation of the transcription preinitiation complex in archaea. *Proc. Natl. Acad. Sci. U. S. A.* 96, 13662–7. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=24121&tool=pmcentrez&rendertype=abstract [Accessed December 16, 2015].

Bell, S. D., Kosa, P. L., Sigler, P. B., and Jackson, S. P. (1999c). Orientation of the transcription preinitiation complex in Archaea. *Proc. Natl. Acad. Sci.* 96, 13662–13667. doi:10.1073/PNAS.96.24.13662.

Berkemer, S. J., Maier, L.-K., Amman, F., Bernhart, S. H., Wörtz, J., Märkle, P., et al. (2020a). Identification of RNA 3´ ends and termination sites in Haloferax volcanii. *RNA Biol.*, 1–14. doi:10.1080/15476286.2020.1723328.

Berkemer, S. J., Maier, L.-K., Amman, F., Bernhart, S. H., Wörtz, J., Märkle, P., et al. (2020b). Identification of RNA 3´ ends and termination sites in Haloferax volcanii. *RNA Biol.* 17, 663–676. doi:10.1080/15476286.2020.1723328.

Bernick, D. L., Dennis, P. P., Lui, L. M., and Lowe, T. M. (2012). Diversity of antisense and other non-coding RNAs in archaea revealed by comparative small RNA sequencing in four Pyrobaculum species. *Front. Microbiol.* 3, 1–18. doi:10.3389/fmicb.2012.00231.

Bini, E. (2010). Archaeal transformation of metals in the environment. *FEMS Microbiol. Ecol.* 73, 1–16. doi:10.1111/j.1574-6941.2010.00876.x.

Bini, E., Dikshit, V., Dirksen, K., Drozda, M., and Blum, P. (2002). Stability of mRNA in the hyperthermophilic archaeon Sulfolobus solfataricus. *RNA*. doi:10.1017/S1355838202021052.

Bischler, T., Tan, H. S., Nieselt, K., and Sharma, C. M. (2015). Differential RNA-seq (dRNA-seq) for annotation of transcriptional start sites and small RNAs in Helicobacter pylori. *Methods* 86, 89–101. doi:10.1016/j.ymeth.2015.06.012.

Blattman, S. B., Jiang, W., Oikonomou, P., and Tavazoie, S. (2020). Prokaryotic single-cell RNA sequencing by in situ combinatorial indexing. *Nat. Microbiol.* doi:10.1038/s41564-020-0729-6.

Blombach, F., Matelska, D., Fouqueau, T., Cackett, G., and Werner, F. (2019). Key Concepts and Challenges in Archaeal Transcription. *J. Mol. Biol.* 431, 4184–4201. doi:10.1016/j.jmb.2019.06.020.

Blombach, F., Salvadori, E., Fouqueau, T., Yan, J., Reimann, J., Sheppard, C., et al. (2015). Archaeal TFEα/β is a hybrid of TFIIE and the RNA polymerase III subcomplex hRPC62/39. *Elife* 4, 1–23. doi:10.7554/eLife.08378.

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

Boccaletto, P., MacHnicka, M. A., Purta, E., Piątkowski, P., Bagiński, B., Wirecki, T. K., et al. (2018). MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res.* 46, D303–D307. doi:10.1093/nar/gkx1030.

Bohlin, J., Eldholm, V., Pettersson, J. H. O., Brynildsrud, O., and Snipen, L. (2017). The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes. *BMC Genomics* 18, 151. doi:10.1186/s12864-017-3543-7.

Boldogkői, Z., Moldován, N., Balázs, Z., Snyder, M., and Tombácz, D. (2019). Long-Read Sequencing – A Powerful Tool in Viral Transcriptome Research. *Trends Microbiol.* 27, 578–592. doi:10.1016/j.tim.2019.01.010.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*. doi:10.1093/bioinformatics/btu170.

Bowden, R., Davies, R. W., Heger, A., Pagnamenta, A. T., de Cesare, M., Oikkonen, L. E., et al. (2019). Sequencing of human genomes with nanopore technology. *Nat. Commun.* doi:10.1038/s41467-019-09637-5.

Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., et al. (2008). The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* doi:10.1038/nbt.1495.

Bridger, S. L., Andrew Lancaster, W., Poole, F. L., Schut, G. J., and Adams, M. W. W. (2012). Genome sequencing of a genetically tractable Pyrococcus furiosus strain reveals a highly dynamic genome. *J. Bacteriol.* 194, 4097–4106. doi:10.1128/JB.00439-12.

Brochier-Armanet, C., Boussau, B., Gribaldo, S., and Forterre, P. (2008). Mesophilic crenarchaeota: Proposal for a third archaeal phylum, the Thaumarchaeota. *Nat. Rev. Microbiol.* doi:10.1038/nrmicro1852.

Brooks, A. N., Turkarslan, S., Beer, K. D., Yin Lo, F., and Baliga, N. S. (2011). Adaptation of cells to new environments. *Wiley Interdiscip. Rev. Syst. Biol. Med.* doi:10.1002/wsbm.136.

Brown, C., Brown, and Clive (2015). AGBT 2012 Presentation (Oxford Nanopore Technologies). *F1000Research*. doi:10.7490/F1000RESEARCH.1110935.1.

Brown, C. G., and Clarke, J. (2016). Nanopore development at Oxford Nanopore. *Nat. Biotechnol.* doi:10.1038/nbt.3622.

Brown, J. W., Haas, E. S., and Pace, N. R. (1993). Characterization of ribonuclease P RNAs from thermophilic bacteria. *Nucleic Acids Res.* doi:10.1093/nar/21.3.671.

Browning, D. F., and Busby, S. J. W. (2016). Local and global regulation of transcription initiation in bacteria. *Nat. Rev. Microbiol.* doi:10.1038/nrmicro.2016.103.

Brügger, K., Redder, P., She, Q., Confalonieri, F., Zivanovic, Y., and Garrett, R. A. (2002). Mobile elements in archaeal genomes. *FEMS Microbiol. Lett.* doi:10.1016/S0378-1097(01)00504-3.

Brukner, I., Sánchez, R., Suck, D., and Pongor, S. (1995). Sequence-dependent bending propensity of DNA as revealed by DNase I: Parameters for trinucleotides. *EMBO J.* doi:10.1002/j.1460-2075.1995.tb07169.x.

Buddeweg, A., Sharma, K., Urlaub, H., and Schmitz, R. A. (2018). sRNA41 affects ribosome binding sites within polycistronic mRNAs in Methanosarcina mazei Gö1. *Mol. Microbiol.* doi:10.1111/mmi.13900.

Buermans, H. P. J. P. J., and den Dunnen, J. T. T. (2014). Next generation sequencing technology: Advances and applications. *Biochim. Biophys. Acta - Mol. Basis Dis.* 1842, 1932–1941. doi:10.1016/j.bbadis.2014.06.015.

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., et al. (1996). Complete Genome Sequence of the Methanogenic Archaeon, Methanococcus jannaschii. *Science (80-.).* 273, 1058–1073. doi:10.1126/science.273.5278.1058.

Burmann, B. M., Schweimer, K., Luo, X., Wahl, M. C., Stitt, B. L., Gottesman, M. E., et al. (2010). A NusE:NusG complex links transcription and translation. *Science (80-.).* doi:10.1126/science.1184953.

Byrne, A., Beaudin, A. E., Olsen, H. E., Jain, M., Cole, C., Palmer, T., et al. (2017). Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* 8, 16027. doi:10.1038/ncomms16027.

Byrne, A., Cole, C., Volden, R., and Vollmers, C. (2019). Realizing the potential of full-length transcriptome sequencing. *Philos. Trans. R. Soc. B Biol. Sci.* 374, 20190097. doi:10.1098/rstb.2019.0097.

Cabrera, M. Á., and Blamey, J. M. (2018). Biotechnological applications of archaeal enzymes from extreme environments. *Biol. Res.* doi:10.1186/s40659-018-0186-3.

Cao, Y., Li, J., Jiang, N., and Dong, X. (2014). Mechanism for stabilizing mRNAs involved in methanol-dependent methanogenesis of cold-adaptive Methanosarcina mazei zm-15. *Appl. Environ. Microbiol.* doi:10.1128/AEM.03495-13.

Chamieh, H., Ibrahim, H., and Kozah, J. (2016). Genome-wide identification of SF1 and SF2 helicases from archaea. *Gene* 576, 214–228. doi:10.1016/j.gene.2015.10.007.

Changela, A., Chen, K., Xue, Y., Holschen, J., Outten, C. E., O'Halloran, T. V., et al. (2003). Molecular basis of metal-ion selectivity and zeptomolar sensitivity by CueR. *Science (80-.).* doi:10.1126/science.1085950.

Charoensawan, V., Wilson, D., and Teichmann, S. A. (2010). Genomic repertoires of DNA-binding transcription factors across the tree of life. *Nucleic Acids Res.* doi:10.1093/nar/gkq617.

Check Hayden, E. (2014a). Is the $1,000 genome for real? *Nature.* doi:10.1038/nature.2014.14530.

Check Hayden, E. (2014b). Technology: The $1,000 genome. *Nature* 507, 294–295. doi:10.1038/507294a.

Chen, F. X., Smith, E. R., and Shilatifard, A. (2018). Born to run: Control of transcription elongation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* doi:10.1038/s41580-018-0010-5.

Chen, S., Iannolo, M., and Calvo, J. M. (2005). Cooperative Binding of the Leucine-Responsive Regulatory Protein (Lrp) to DNA. *J. Mol. Biol.* 345, 251–264. doi:10.1016/j.jmb.2004.10.047.

Chen, T. Y., Santiago, A. G., Jung, W., Krzeminski, L., Yang, F., Martell, D. J., et al. (2015). Concentration-and chromosome-organization-dependent regulator unbinding from DNA for transcription regulation in living cells. *Nat. Commun.* doi:10.1038/ncomms8445.

Chen, Y. C., Liu, T., Yu, C. H., Chiang, T. Y., and Hwang, C. C. (2013). Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly. *PLoS One.* doi:10.1371/journal.pone.0062856.

Cherf, G. M., Lieberman, K. R., Rashid, H., Lam, C. E., Karplus, K., and Akeson, M. (2012). Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. *Nat. Biotechnol.* doi:10.1038/nbt.2147.

Cho, S., Kim, M.-S., Jeong, Y., Lee, B.-R., Lee, J.-H., Kang, S. G., et al. (2017). Genome-wide primary transcriptome analysis of H2-producing archaeon Thermococcus onnurineus NA1. *Sci. Rep.* 7, 43044. doi:10.1038/srep43044.

CHOMZYNSKI, P. (1987). Single-Step Method of RNA Isolation by Acid Guanidinium Thiocyanate–Phenol–Chloroform Extraction. *Anal. Biochem.* 162, 156–159. doi:10.1006/abio.1987.9999.

Clarke, J., Wu, H. C., Jayasinghe, L., Patel, A., Reid, S., and Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* doi:10.1038/nnano.2009.12.

Cline, J. (1996). PCR fidelity of pfu DNA polymerase and other thermostable DNA polymerases. *Nucleic Acids Res.* doi:10.1093/nar/24.18.3546.

Clouet-D'Orval, B., Batista, M., Bouvier, M., Quentin, Y., Fichant, G., Marchfelder, A., et al. (2018). Insights into RNA-processing pathways and associated RNA-degrading enzymes in Archaea. *FEMS Microbiol. Rev.* 42, 579–613. doi:10.1093/femsre/fuy016.

Core, L. J., Waterfall, J. J., and Lis, J. T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science (80-.).* doi:10.1126/science.1162228.

Coureux, P.-D. D., Lazennec-Schurdevin, C., Bourcier, S., Mechulam, Y., and Schmitt, E. (2020). Cryo-EM study of an archaeal 30S initiation complex gives insights into evolution of translation initiation. *Commun. Biol.* 3, 58. doi:10.1038/s42003-020-0780-0.

Cox, J. M., Hayward, M. M., Sanchez, J. F., Gegnas, L. D., van der Zee, S., Dennis, J. H., et al. (1997). Bidirectional binding of the TATA box binding protein to the TATA box. *Proc. Natl. Acad. Sci.* doi:10.1073/pnas.94.25.13475.

Craig Venter, J., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., et al. (2001). The sequence of the human genome. *Science (80-.).* doi:10.1126/science.1058040.

Crick, F. (1970). Central dogma of molecular biology. *Nature.* doi:10.1038/227561a0.

Croucher, N. J., and Thomson, N. R. (2010). Studying bacterial transcriptomes using RNA-seq. *Curr. Opin. Microbiol.* 13, 619–624. doi:10.1016/j.mib.2010.09.009.

Dahlke, I. (2002). A Pyrococcus homolog of the leucine-responsive regulatory protein, LrpA, inhibits transcription by abrogating RNA polymerase recruitment. *Nucleic Acids Res.* 30, 701–710. doi:10.1093/nar/30.3.701.

Danan, M., Schwartz, S., Edelheit, S., and Sorek, R. (2012). Transcriptome-wide discovery of circular RNAs in Archaea. *Nucleic Acids Res.* 40, 3131–3142. doi:10.1093/nar/gkr1009.

Daniel, R. M., and Cowan, D. A. (2000). Biomolecular stability and life at high temperatures. *Cell. Mol. Life Sci.* doi:10.1007/PL00000688.

Daniels, J. P., Kelly, S., Wickstead, B., and Gull, K. (2009). Identification of a crenarchaeal orthologue of Elf1: Implications for chromatin and transcription in Archaea. *Biol. Direct.* doi:10.1186/1745-6150-4-24.

Dar, D., Prasse, D., Schmitz, R. A., and Sorek, R. (2016). Widespread formation of alternative 3′ UTR isoforms via transcription termination in archaea. *Nat. Microbiol.* 1, 16143. doi:10.1038/nmicrobiol.2016.143.

Dar, D., and Sorek, R. (2018). High-resolution RNA 3-ends mapping of bacterial Rho-dependent transcripts. *Nucleic Acids Res.* 46, 6797–6805. doi:10.1093/nar/gky274.

Darnell, C. L., Tonner, P. D., Gulli, J. G., Schmidler, S. C., and Schmid, A. K. (2017). Systematic Discovery of Archaeal Transcription Factor Functions in Regulatory Networks through Quantitative Phenotyping Analysis. *mSystems.* doi:10.1128/msystems.00032-17.

Daum, B., Vonck, J., Bellack, A., Chaudhury, P., Reichelt, R., Albers, S.-V. V., et al. (2017). Structure and in situ organisation of the Pyrococcus furiosus archaellum machinery. *Elife* 6. doi:10.7554/eLife.27470.

de los Rios, S., and Perona, J. J. (2007). Structure of the Escherichia coli Leucine-responsive Regulatory Protein Lrp Reveals a Novel Octameric Assembly. *J. Mol. Biol.* 366, 1589–1602. doi:10.1016/j.jmb.2006.12.032.

de Mendoza, A., and Sebé-Pedrós, A. (2019). Origin and evolution of eukaryotic transcription factors. *Curr. Opin. Genet. Dev.* 58–59, 25–32. doi:10.1016/j.gde.2019.07.010.

Deamer, D., Akeson, M., and Branton, D. (2016). Three decades of nanopore sequencing. *Nat. Biotechnol.* 34, 518–524. doi:10.1038/nbt.3423.

Deangelis, M. M., Wang, D. G., and Hawkins, T. L. (1995). Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Res.* doi:10.1093/nar/23.22.4742.

Decatur, W. A., and Fournier, M. J. (2002). rRNA modifications and ribosome function. *Trends Biochem. Sci.* doi:10.1016/S0968-0004(02)02109-6.

Decker, K. B., and Hinton, D. M. (2013). Transcription Regulation at the Core: Similarities Among Bacterial, Archaeal, and Eukaryotic RNA Polymerases. *Annu. Rev. Microbiol.* 67, 113–139. doi:10.1146/annurev-micro-092412-155756.

DeLong, E. (1998). Archaeal means and extremes. *Science (80-.).* doi:10.1126/science.280.5363.542.

DeLong, E. F. (1992). Archaea in coastal marine environments. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.89.12.5685.

Denis, A., Martínez-Núñez, M. A., Tenorio-Salgado, S., and Perez-Rueda, E. (2018). Dissecting the Repertoire of DNA-Binding Transcription Factors of the Archaeon Pyrococcus furiosus DSM 3638. *Life* 8, 40. doi:10.3390/life8040040.

Dennis, P. P., Tripp, V., Lui, L., Lowe, T., and Randau, L. (2015). C/D box sRNA-guided 2'-O-methylation patterns of archaeal rRNA molecules. *BMC Genomics.* doi:10.1186/s12864-015-1839-z.

Deutscher, M. P. (2015). Twenty years of bacterial RNases and RNA processing: how we've matured. *RNA* 21, 597–600. doi:10.1261/rna.049692.115.

Dexl, S., Reichelt, R., Kraatz, K., Schulz, S., Grohmann, D., Bartlett, M., et al. (2018). Displacement of the transcription factor B reader domain during transcription initiation. *Nucleic Acids Res.* 46, 10066–10081. doi:10.1093/nar/gky699.

DiRuggiero, J., Dunn, D., Maeder, D. L., Holley-Shanks, R., Chatard, J., Horlacher, R., et al. (2000). Evidence of recent lateral gene transfer among hyperthermophilic archaea. *Mol Microbiol.* doi:10.1063/1.1497710.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi:10.1093/bioinformatics/bts635.

Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* doi:10.1093/nar/gkn425.

Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., et al. (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science (80-.).* doi:10.1126/science.1181498.

Drmanac, S., Callow, M., Chen, L., Zhou, P., Eckhardt, L., Xu, C., et al. (2020). CoolMPS^{TM}: Advanced massively parallel sequencing using antibodies specific to each natural nucleobase. *bioRxiv*, 2020.02.19.953307. doi:10.1101/2020.02.19.953307.

Dugar, G., Herbig, A., Förstner, K. U., Heidrich, N., Reinhardt, R., Nieselt, K., et al. (2013). High-Resolution Transcriptome Maps Reveal Strain-Specific Regulatory Features of Multiple Campylobacter jejuni Isolates. *PLoS Genet.* 9, e1003495. doi:10.1371/journal.pgen.1003495.

Durand, S., Tomasini, A., Braun, F., Condon, C., and Romby, P. (2015). sRNA and mRNA turnover in gram-positive bacteria. *FEMS Microbiol. Rev.* doi:10.1093/femsre/fuv007.

Durovic, P., and Dennis, P. P. (1994). Separate pathways for excision and processing of 16S and 23S rRNA from the primary rRNA operon transcript from the hyperthermophilic archaebacterium Sulfolobus acldocaldarius : similarities to eukaryotic rRNA processing. *Mol. Microbiol.* 13, 229–242. doi:10.1111/j.1365-2958.1994.tb00418.x.

Ebaid, R., Wang, H., Sha, C., Abomohra, A. E. F., and Shao, W. (2019). Recent trends in hyperthermophilic enzymes production and future perspectives for biofuel industry: A critical review. *J. Clean. Prod.* doi:10.1016/j.jclepro.2019.117925.

Ebersberger, I., Simm, S., Leisegang, M. S., Schmitzberger, P., Mirus, O., von Haeseler, A., et al. (2014). The evolution of the ribosome biogenesis pathway from a yeast perspective. *Nucleic Acids Res.* 42, 1509–1523. doi:10.1093/nar/gkt1137.

Ebright, R. H., Werner, F., and Zhang, X. (2019). RNA Polymerase Reaches 60: Transcription Initiation, Elongation, Termination, and Regulation in Prokaryotes. *J. Mol. Biol.* 431, 3945–3946. doi:10.1016/j.jmb.2019.07.026.

Efremov, A. K., Qu, Y., Maruyama, H., Lim, C. J., Takeyasu, K., and Yan, J. (2015). Transcriptional repressor TrmBL2 from Thermococcus kodakarensis forms filamentous nucleoprotein structures and competes with histones for DNA binding in a salt- and DNA supercoiling-dependent manner. *J. Biol. Chem.* doi:10.1074/jbc.M114.626705.

Ehara, H., Yokoyama, T., Shigematsu, H., Yokoyama, S., Shirouzu, M., and Sekine, S. I. (2017). Structure of the complete elongation complex of RNA polymerase II with basal factors. *Science (80-.).* doi:10.1126/science.aan8552.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science (80-.).* 323, 133–138. doi:10.1126/science.1162986.

Ellis, M. J., and Haniford, D. B. (2016). Riboregulation of bacterial and archaeal transposition. *Wiley Interdiscip. Rev. RNA.* doi:10.1002/wrna.1341.

Ellis, M. J., Trussler, R. S., and Haniford, D. B. (2015). A cis-encoded sRNA, Hfq and mRNA secondary structure act independently to suppress IS200 transposition. *Nucleic Acids Res.* doi:10.1093/nar/gkv584.

Eme, L., and Ettema, T. J. G. (2018). The eukaryotic ancestor shapes up. *Nature.* doi:10.1038/d41586-018-06868-2.

Eme, L., Spang, A., Lombard, J., Stairs, C. W., and Ettema, T. J. G. (2017). Archaea and the origin of eukaryotes. *Nat. Rev. Microbiol.* doi:10.1038/nrmicro.2017.133.

Epshtein, V., Cardinale, C. J., Ruckenstein, A. E., Borukhov, S., and Nudler, E. (2007). An Allosteric Path to Transcription

Termination. *Mol. Cell.* doi:10.1016/j.molcel.2007.10.011.

Escobar-Páramo, P., Ghosh, S., and DiRuggiero, J. (2005). Evidence for Genetic Drift in the Diversification of a Geographically Isolated Population of the Hyperthermophilic Archaeon Pyrococcus. *Mol. Biol. Evol.* 22, 2297–2303. doi:10.1093/molbev/msi227.

Escobar-Zepeda, A., Vera-Ponce de León, A., and Sanchez-Flores, A. (2015). The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics. *Front. Genet.* 6. doi:10.3389/fgene.2015.00348.

Ettema, T. J. G., Brinkman, A. B., Lamers, P. P., Kornet, N. G., de Vos, W. M., and van der Oost, J. (2006). Molecular characterization of a conserved archaeal copper resistance (cop) gene cluster and its copper-responsive regulator in Sulfolobus solfataricus P2. *Microbiology* 152, 1969–1979. doi:10.1099/mic.0.28724-0.

Ettema, T. J. G., Huynen, M. A., De Vos, W. M., and Van Der Oost, J. (2003). TRASH: A novel metal-binding domain predicted to be involved in heavy-metal sensing, trafficking and resistance. *Trends Biochem. Sci.* 28, 170–173. doi:10.1016/S0968-0004(03)00037-9.

Evguenieva-Hackenberg, E., and Klug, G. (2011). New aspects of RNA processing in prokaryotes. *Curr. Opin. Microbiol.* doi:10.1016/j.mib.2011.07.025.

Facciotti, M. T., Reiss, D. J., Pan, M., Kaur, A., Vuthoori, M., Bonneau, R., et al. (2007). General transcription factor specified global gene regulation in archaea. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.0611663104.

Ferreira-Cerca, S. (2017). *RNA Metabolism and Gene Expression in Archaea - Chapter6 - Life and Death of Ribosomes in Archaea.* doi:10.1007/978-3-319-65795-0.

Fiala, G., and Stetter, K. O. (1986). Pyrococcus furiosus sp. nov. represents a novel genus of marine heterotrophic archaebacteria growing optimally at 100°C. *Arch. Microbiol.* 145, 56–61. doi:10.3747/pdi.2011.00058.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., et al. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science (80-.).* doi:10.1126/science.7542800.

Forterre, P. (1996). A hot topic: The origin of hyperthermophiles. *Cell.* doi:10.1016/S0092-8674(00)81262-3.

Forterre, P., and Fagan, T. L. (2016). *Microbes from Hell.* doi:10.7208/chicago/9780226265964.001.0001.

Foster, A. W., Osman, D., and Robinson, N. J. (2014). Metal preferences and metallation. *J. Biol. Chem.* 289, 28095–28103. doi:10.1074/jbc.R114.588145.

Fouqueau, T., Blombach, F., Cackett, G., Carty, A. E., Matelska, D. M., Ofer, S., et al. (2018). The cutting edge of archaeal transcription. *Emerg. Top. Life Sci.* 2, 517–533. doi:10.1042/etls20180014.

Fouqueau, T., Blombach, F., Hartman, R., Cheung, A. C. M., Young, M. J., and Werner, F. (2017). The transcript cleavage factor paralogue TFS4 is a potent RNA polymerase inhibitor. *Nat. Commun.* doi:10.1038/s41467-017-02081-3.

Fouqueau, T., Zeller, M. E., Cheung, A. C., Cramer, P., and Thomm, M. (2013). The RNA polymerase trigger loop functions in all three phases of the transcription cycle. *Nucleic Acids Res.* 41, 7048–7059. doi:10.1093/nar/gkt433.

French, S. L., Santangelo, T. J., Beyer, A. L., and Reeve, J. N. (2007). Transcription and Translation are Coupled in Archaea. *Mol. Biol. Evol.* 24, 893–895. doi:10.1093/molbev/msm007.

Fuhrman, J. A., and McCallum, K. (1992). Novel major archaebacterial group from marine plankton. *Nature.* doi:10.1038/356148a0.

Fuller, C. W., Middendorf, L. R., Benner, S. A., Church, G. M., Harris, T., Huang, X., et al. (2009). The challenges of sequencing by synthesis. *Nat. Biotechnol.* doi:10.1038/nbt.1585.

Furey, T. S. (2012). ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions. *Nat. Rev. Genet.* doi:10.1038/nrg3306.

Galtier, N., and Lobry, J. R. (1997). Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J. Mol. Evol.* doi:10.1007/PL00006186.

Gao, J., Bauer, M. W., Shockley, K. R., Pysz, M. A., and Kelly, R. M. (2003). Growth of hyperthermophilic archaeon Pyrococcus furiosus on chitin involves two family 18 chitinases. *Appl. Environ. Microbiol.* doi:10.1128/AEM.69.6.3119-3128.2003.

Garalde, D. R., Snell, E. A., Jachimowicz, D., Sipos, B., Lloyd, J. H., Bruce, M., et al. (2018). Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* 15, 201–206. doi:10.1038/nmeth.4577.

Gaspin, C., Cavaillé, J., Erauso, G., and Bachellerie, J. P. (2000). Archaeal homologs of eukaryotic methylation guide small nucleolar RNAs: Lessons from the Pyrococcus genomes. *J. Mol. Biol.* doi:10.1006/jmbi.2000.3593.

Gehring, A. M., and Santangelo, T. J. (2017). Archaeal RNA polymerase arrests transcription at DNA lesions. *Transcription.* doi:10.1080/21541264.2017.1324941.

Gehring, A. M., Walker, J. E., and Santangelo, T. J. (2016). Transcription regulation in archaea. *J. Bacteriol.* 198, JB.00255-16. doi:10.1128/JB.00255-16.

Gietl, A., Holzmeister, P., Blombach, F., Schulz, S., Von Voithenberg, L. V., Lamb, D. C., et al. (2014). Eukaryotic and archaeal TBP and TFB/TF(II)B follow different promoter DNA bending pathways. *Nucleic Acids Res.* doi:10.1093/nar/gku273.

Gindner, A., Hausner, W., and Thomm, M. (2014). The TrmB family: a versatile group of transcriptional regulators in Archaea. *Extremophiles* 18, 925–936. doi:10.1007/s00792-014-0677-2.

Glansdorff, N. (1999). On the origin of operons and their possible role in evolution toward thermophily. *J. Mol. Evol.* doi:10.1007/PL00006566.

Goldman, S. R., Ebright, R. H., and Nickels, B. E. (2009). Direct detection of abortive RNA transcripts in vivo. *Science (80-.).* doi:10.1126/science.1169237.

Gomes-Filho, J. V., Daume, M., and Randau, L. (2018). Unique Archaeal Small RNAs. *Annu. Rev. Genet.* 52, 465–487. doi:10.1146/annurev-genet-120417-031300.

Gomes-Filho, J. V., Randau, L., Gomes-Filho, J. V., and Randau, L. (2019). RNA stabilization in hyperthermophilic archaea. *Ann. N. Y. Acad. Sci.* 1447, 88–96. doi:10.1111/nyas.14060.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351. doi:10.1038/nrg.2016.49.

Graham, D. E. (2000). An archaeal genomic signature. *Proc. Natl. Acad. Sci.* doi:10.1073/pnas.050564797.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007). CRISPRFinder: A web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.* doi:10.1093/nar/gkm360.

Grogan, D. W. (1998). Hyperthermophiles and the problem of DNA instability. *Mol. Microbiol.* doi:10.1046/j.1365-2958.1998.00853.x.

Grohmann, D., Nagy, J., Chakraborty, A., Klose, D., Fielden, D., Ebright, R. H. H., et al. (2011). The Initiation Factor TFE and the Elongation Factor Spt4/5 Compete for the RNAP Clamp during Transcription Initiation and Elongation. *Mol. Cell* 43, 263–274. doi:10.1016/j.molcel.2011.05.030.

Grohmann, D., and Werner, F. (2011). Recent advances in the understanding of archaeal transcription. *Curr. Opin. Microbiol.* 14, 328–334. doi:10.1016/j.mib.2011.04.012.

Grosjean, H., Gaspin, C., Marck, C., Decatur, W. A., and de Crécy-Lagard, V. (2008). RNomics and Modomics in the halophilic archaea Haloferax volcanii: identification of RNA modification genes. *BMC Genomics* 9, 470. doi:10.1186/1471-2164-9-470.

Grünberger, F., Knüppel, R., Jüttner, M., Fenk, M., Borst, A., Reichelt, R., et al. (2020a). Exploring prokaryotic transcription, operon structures, rRNA maturation and modifications using Nanopore-based native RNA sequencing. *bioRxiv*, 2019.12.18.880849. doi:10.1101/2019.12.18.880849.

Grünberger, F., Reichelt, R., Bunk, B., Spröer, C., Overmann, J., Rachel, R., et al. (2019). Next Generation DNA-Seq and Differential RNA-Seq Allow Re-annotation of the Pyrococcus furiosus DSM 3638 Genome and Provide Insights Into Archaeal Antisense Transcription. *Front. Microbiol.* 10. doi:10.3389/fmicb.2019.01603.

Grünberger, F., Reichelt, R., Waege, I., Ned, V., Bronner, K., Kaljanac, M., et al. (2020b). CopR, a global regulator of transcription to maintain copper homeostasis in Pyrococcus furiosus. *bioRxiv*, 2020.08.14.251413. doi:10.1101/2020.08.14.251413.

Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). Circlize implements and enhances circular visualization in R. *Bioinformatics*. doi:10.1093/bioinformatics/btu393.

Gunther, M. R., Hanna, P. M., Mason, R. P., and Cohen, M. S. (1995). Hydroxyl radical formation from cuprous ion and hydrogen peroxide: a spin-trapping study. *Arch. Biochem. Biophys.* 316, 515–522. doi:10.1006/abbi.1995.1068.

Gupta, P. K. (2008). Single-molecule DNA sequencing technologies for future genomics research. *Trends Biotechnol.* doi:10.1016/j.tibtech.2008.07.003.

Gusarov, I., and Nudler, E. (1999). The mechanism of intrinsic transcription termination. *Mol. Cell* 3, 495–504. doi:10.1016/S1097-2765(00)80477-3.

Guy, L., and Ettema, T. J. G. (2011). The archaeal "TACK" superphylum and the origin of eukaryotes. *Trends Microbiol.* doi:10.1016/j.tim.2011.09.002.

Guy, L., Kultima, J. R., Andersson, S. G. E., and Quackenbush, J. (2011). GenoPlotR: comparative gene and genome visualization in R. in *Bioinformatics* doi:10.1093/bioinformatics/btq413.

Haberle, V., and Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. *Nat. Rev. Mol. Cell Biol.* doi:10.1038/s41580-018-0028-8.

Haft, D. H., DiCuccio, M., Badretdin, A., Brover, V., Chetvernin, V., O'Neill, K., et al. (2018). RefSeq: An update on prokaryotic genome annotation and curation. *Nucleic Acids Res.* doi:10.1093/nar/gkx1068.

Hallam, S. J., Konstantinidis, K. T., Putnam, N., Schleper, C., Watanabe, Y. I., Sugahara, J., et al. (2006). Genomic analysis of the uncultivated marine crenarchaeote Cenarchaeum symbiosum. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.0608549103.

Hamilton-Brehm, S. D., Schut, G. J., and Adams, M. W. W. (2005). Metabolic and evolutionary relationships among Pyrococcus species: Genetic exchange within a hydrothermal vent environment. *J. Bacteriol.* doi:10.1128/JB.187.21.7492-7499.2005.

Hammar, P., Leroy, P., Mahmutovic, A., Marklund, E. G., Berg, O. G., and Elf, J. (2012). The lac repressor displays facilitated diffusion in living cells. *Science (80-.).* doi:10.1126/science.1221648.

Hartman, A. L., Norais, C., Badger, J. H., Delmas, S., Haldenby, S., Madupu, R., et al. (2010). The Complete Genome Sequence of Haloferax volcanii DS2, a Model Archaeon. *PLoS One* 5, e9605. doi:10.1371/journal.pone.0009605.

Hausner, W., Frey, G., and Thomm, M. (1991). Control regions of an archaeal gene. A TATA box and an initiator element promote cell-free transcription of the tRNAVal gene of Methanococcus vannielii. *J. Mol. Biol.* doi:10.1016/0022-2836(91)90492-O.

Hausner, W., Lange, U., and Musfeldt, M. (2000). Transcription factor S, a cleavage induction factor of the archaeal RNA polymerase. *J. Biol. Chem.* 275, 12393–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10777522 [Accessed December 21, 2015].

Hausner, W., Wettach, J., Hethke, C., and Thomm, M. (1996). Two Transcription Factors Related with the Eucaryal Transcription Factors TATA-binding Protein and Transcription Factor IIB Direct Promoter Recognition by an

Archaeal RNA Polymerase. *J. Biol. Chem.* 271, 30144–30148. doi:10.1074/jbc.271.47.30144.

Heather, J. M., and Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*. doi:10.1016/j.ygeno.2015.11.003.

Henras, A. K., Plisson-Chastang, C., O'Donohue, M.-F., Chakraborty, A., and Gleizes, P.-E. (2015). An overview of pre-ribosomal RNA processing in eukaryotes. *Wiley Interdiscip. Rev. RNA* 6, 225–242. doi:10.1002/wrna.1269.

Heyer, R., Dörr, M., Jellen-Ritter, A., Späth, B., Babski, J., Jaschinski, K., et al. (2012). High throughput sequencing reveals a plethora of small RNAs including tRNA derived fragments in Haloferax volcanii. *RNA Biol.* 9, 1011–8. doi:10.4161/rna.20826.

Higashibata, H., Fujiwara, S., Takagi, M., and Imanaka, T. (1999). Analysis of DNA compaction profile and intracellular contents of archaeal histones from Pyrococcus kodakaraensis KOD1. *Biochem. Biophys. Res. Commun.* doi:10.1006/bbrc.1999.0533.

Hill, C. H., Boreikaitė, V., Kumar, A., Casañal, A., Kubík, P., Degliesposti, G., et al. (2019). Activation of the Endonuclease that Defines mRNA 3′ Ends Requires Incorporation into an 8-Subunit Core Cleavage and Polyadenylation Factor Complex. *Mol. Cell*. doi:10.1016/j.molcel.2018.12.023.

Hirata, A., Klein, B. J., and Murakami, K. S. (2008). The X-ray crystal structure of RNA polymerase from Archaea. *Nature*. doi:10.1038/nature06530.

Hirtreiter, A., Damsma, G. E., Cheung, A. C. M., Klose, D., Grohmann, D., Vojnic, E., et al. (2010a). Spt4/5 stimulates transcription elongation through the RNA polymerase clamp coiled-coil motif. *Nucleic Acids Res.* 38, 4040–4051. doi:10.1093/nar/gkq135.

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

Hobman, J. L., Wilkie, J., and Brown, N. L. (2005). A design for life: Prokaryotic metal-binding MerR family regulators. in *BioMetals* doi:10.1007/s10534-005-3717-7.

Hoffmann, N. A., Jakobi, A. J., Moreno-Morcillo, M., Glatt, S., Kosinski, J., Hagen, W. J. H. H., et al. (2015). Molecular structures of unbound and transcribing RNA polymerase III. *Nature* 528, 231–236. doi:10.1038/nature16143.

Hofmann, S., and Miller, O. L. (1977). Visualization of ribosomal ribonucleic acid synthesis in a ribonuclease III-Deficient strain of Escherichia coli. *J. Bacteriol.* 132, 718–22. doi:10.1128/JB.132.2.718-722.1977.

Holmqvist, E., and Wagner, G. H. (2017). Impact of bacterial sRNAs in stress responses. *Biochem. Soc. Trans.* doi:10.1042/BST20160363.

Honarmand Ebrahimi, K., Hagedoorn, P. L., and Hagen, W. R. (2015). Unity in the biochemistry of the iron-storage proteins ferritin and bacterioferritin. *Chem. Rev.* doi:10.1021/cr5004908.

Hong, S. K., Jeong, J., Sung, M. K., Kang, G., Kim, M., Ran, A., et al. (2019). Characterization of the copper-sensing transcriptional regulator CopR from the hyperthermophilic archeaon Thermococcus onnurineus NA1. *BioMetals* 2, 1–15.

Hör, J., Gorski, S. A., and Vogel, J. (2018). Bacterial RNA Biology on a Genome Scale. *Mol. Cell* 70, 785–799. doi:10.1016/j.molcel.2017.12.023.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*. doi:10.1016/S0140-6736(20)30183-5.

Huber, H., Hohn, M. J., Rachel, R., Fuchs, T., Wimmer, V. C., and Stetter, K. O. (2002). A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature*. doi:10.1038/417063a.

Huber, R., Burggraf, S., Mayer, T., Barns, S. M., Rossnagel, P., and Stetter, K. O. (1995). Isolation of a hyperthermophilic archaeum predicted by in situ RNA analysis. *Nature*. doi:10.1038/376057a0.

Huet, J., Schnabel, R., Sentenac, A., and Zillig, W. (1983). Archaebacteria and eukaryotes possess DNA-dependent RNA polymerases of a common type. *EMBO J.* 2, 1291–1294. doi:10.1002/j.1460-2075.1983.tb01583.x.

Hurst, L. D., and Merchant, A. R. (2001). High guanine-cytosine content is not an adaptation to high temperature: A comparative analysis amongst prokaryotes. *Proc. R. Soc. B Biol. Sci.* doi:10.1098/rspb.2000.1397.

Iben, J. R., and Maraia, R. J. (2012). tRNAomics: tRNA gene copy number variation and codon use provide bioinformatic evidence of a new anticodon:codon wobble pair in a eukaryote. *RNA*. doi:10.1261/rna.032151.111.

Imachi, H., Nobu, M. K., Nakahara, N., Morono, Y., Ogawara, M., Takaki, Y., et al. (2020). Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* 577. doi:10.1038/s41586-019-1916-6.

Imdahl, F., Vafadarnejad, E., Homberger, C., Saliba, A.-E., and Vogel, J. (2020). Single-cell RNA-sequencing reports growth-condition-specific global transcriptomes of individual bacteria. *Nat. Microbiol.* doi:10.1038/s41564-020-0774-1.

Iost, I., Chabas, S., and Darfeuille, F. (2019). Maturation of atypical ribosomal RNA precursors in Helicobacter pylori. *Nucleic Acids Res.* 47, 5906–5921. doi:10.1093/nar/gkz258.

Ito, S., Akamatsu, Y., Noma, A., Kimura, S., Miyauchi, K., Ikeuchi, Y., et al. (2014a). A Single Acetylation of 18 S rRNA Is Essential for Biogenesis of the Small Ribosomal Subunit in Saccharomyces cerevisiae. *J. Biol. Chem.* 289, 26201–26212. doi:10.1074/jbc.M114.593996.

Ito, S., Horikawa, S., Suzuki, T., Kawauchi, H., Tanaka, Y., Suzuki, T., et al. (2014b). Human NAT10 Is an ATP-dependent RNA Acetyltransferase Responsible for N 4 -Acetylcytidine Formation in 18 S Ribosomal RNA (rRNA). *J. Biol. Chem.* 289, 35724–35730. doi:10.1074/jbc.C114.602698.

Jacob, A. I., Köhrer, C., Davies, B. W., RajBhandary, U. L., and Walker, G. C. (2013). Conserved Bacterial RNase YbeY Plays Key Roles in 70S Ribosome Quality Control and 16S rRNA Maturation. *Mol. Cell* 49, 427–438.

doi:10.1016/j.molcel.2012.11.025.

Jaeschke, A., Jørgensen, S. L., Bernasconi, S. M., Pedersen, R. B., Thorseth, I. H., and Früh-Green, G. L. (2012). Microbial diversity of Loki's Castle black smokers at the Arctic Mid-Ocean Ridge. *Geobiology*. doi:10.1111/gbi.12009.

Jäger, A., Samorski, R., Pfeifer, F., and Klug, G. (2002). Individual gvp transcript segments in Haloferax mediterranei exhibit varying half-lives, which are differentially affected by salt concentration and growth phase. *Nucleic Acids Res.* doi:10.1093/nar/gkf699.

Jäger, D., Förstner, K. U., Sharma, C. M., Santangelo, T. J., and Reeve, J. N. (2014). Primary transcriptome map of the hyperthermophilic archaeon Thermococcus kodakarensis. *BMC Genomics* 15, 684. doi:10.1186/1471-2164-15-684.

Jäger, D., Sharma, C. M., Thomsen, J., Ehlers, C., Vogel, J., and Schmitz, R. A. (2009). Deep sequencing analysis of the Methanosarcina mazei Go1 transcriptome in response to nitrogen availability. *Proc. Natl. Acad. Sci.* 106, 21878–21882. doi:10.1073/pnas.0909051106.

Jain, C. (2020). RNase AM, a 5′ to 3′ exonuclease, matures the 5′ end of all three ribosomal RNAs in E. coli. *Nucleic Acids Res.* 48, 5616–5623. doi:10.1093/nar/gkaa260.

Jain, M., Fiddes, I., Miga, K. H., Olsen, H. E., Paten, B., and Akeson, M. (2015). Improved data analysis for the MinION nanopore sequencer Miten. *Nat. Methods* 12, 351–356. doi:10.1007/s00261-015-0542-5.

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* doi:10.1038/nbt.4060.

Jain, M., Olsen, H. E., Paten, B., and Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* 17, 256. doi:10.1186/s13059-016-1122-x.

Jain, M., Tyson, J. R., Loose, M., Ip, C. L. C., Eccles, D. A., O'Grady, J., et al. (2017). MinION Analysis and Reference Consortium: Phase 2 data release and analysis of R9.0 chemistry. *F1000Research*. doi:10.12688/f1000research.11354.1.

Jin, Y., Eser, U., Struhl, K., and Churchman, L. S. (2017). The Ground State and Evolution of Promoter Region Directionality. *Cell*, 1–10. doi:10.1016/j.cell.2017.07.006.

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science (80-.).* doi:10.1126/science.1141319.

Johnson, G. E., Lalanne, J.-B., Peters, M. L., and Li, G.-W. (2020). Functionally uncoupled transcription–translation in Bacillus subtilis. *Nature*. doi:10.1038/s41586-020-2638-5.

Joshi, C. P., Panda, D., Martella, D. J., Andoy, N. M., Chen, T. Y., Gaballa, A., et al. (2012). Direct substitution and assisted dissociation pathways for turning off transcription by a MerR-family metalloregulator. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.1208508109.

Ju, J., Kim, D. H., Bi, L., Meng, Q., Bai, X., Li, Z., et al. (2006). Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.0609513103.

Jüttner, M., Weiß, M., Ostheimer, N., Reglin, C., Kern, M., Knüppel, R., et al. (2020). A versatile cis-acting element reporter system to study the function, maturation and stability of ribosomal RNA mutants in archaea. *Nucleic Acids Res.* 48, 2073–2090. doi:10.1093/nar/gkz1156.

Kang, W., Ha, K. S., Uhm, H., Park, K., Lee, J. Y., Hohng, S., et al. (2020). Transcription reinitiation by recycling RNA polymerase that diffuses on DNA after releasing terminated RNA. *Nat. Commun.* doi:10.1038/s41467-019-14200-3.

Kanoksilapatham, W., González, J. M., Maeder, D. L., Diruggiero, J., and Robb, F. T. (2004). A proposal to rename the hyperthermophile Pyrococcus woesei as Pyrococcus furiosus subsp. woesei. *Archaea*. doi:10.1155/2004/513563.

Karr, E. A. (2014). "Transcription Regulation in the Third Domain," in *Advances in Applied Microbiology* (Elsevier Inc.), 101–133. doi:10.1016/B978-0-12-800259-9.00003-2.

Karr, E. A., Isom, C. E., Trinh, V., and Peeters, E. (2017). "Transcription Factor-Mediated Gene Regulation in Archaea," in *RNA Metabolism and Gene Expression in Archaea* Nucleic Acids and Molecular Biology., ed. B. Clouet-d'Orval (Cham: Springer International Publishing), 27–69. doi:10.1007/978-3-319-65795-0_2.

Karst, S. M., Ziels, R. M., Kirkegaard, R. H., Sørensen, E. A., McDonald, D., Zhu, Q., et al. (2020). Enabling high-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *bioRxiv*, 645903. doi:10.1101/645903.

Kasianowicz, J. J., Brandin, E., Branton, D., and Deamer, D. W. (1996). Characterization of individual polynucleotide molecules using a membrane channel. in *Proceedings of the National Academy of Sciences of the United States of America* doi:10.1073/pnas.93.24.13770.

Kaur, A., Pan, M., Meislin, M., Facciotti, M. T., El-gewely, R., and Baliga, N. S. (2006). A systems view of haloarchaeal strategies to withstand stress from transition metals A systems view of haloarchaeal strategies to withstand stress from transition metals. *Genome Res.* 16, 841–854. doi:10.1101/gr.5189606.

Keller, M. W., Lipscomb, G. L., Loder, A. J., Schut, G. J., Kelly, R. M., and Adams, M. W. W. (2015). A hybrid synthetic pathway for butanol production by a hyperthermophilic microbe. *Metab. Eng.* doi:10.1016/j.ymben.2014.11.004.

Keller, M. W., Lipscomb, G. L., Nguyen, D. M., Crowley, A. T., Schut, G. J., Scott, I., et al. (2017). Ethanol production by the hyperthermophilic archaeon Pyrococcus furiosus by expression of bacterial bifunctional alcohol dehydrogenases. *Microb. Biotechnol.* doi:10.1111/1751-7915.12486.

Keller, M. W., Rambo-Martin, B. L., Wilson, M. M., Ridenour, C. A., Shepard, S. S., Stark, T. J., et al. (2018). Direct RNA Sequencing of the Coding Complete Influenza A Virus Genome. *Sci. Rep.* 8, 14408. doi:10.1038/s41598-018-32615-8.

Kellner, S., Spang, A., Offre, P., Szöllősi, G. J., Petitjean, C., and Williams, T. A. (2018). Genome size evolution in the Archaea. *Emerg. Top. Life Sci.* 2, 595–605. doi:10.1042/etls20180021.

Kengen, S. W. M. (2017). 'Pyrococcus furiosus, 30 years on.' *Microb. Biotechnol.* 10, 1441–1444. doi:10.1111/1751-7915.12695.

Kengen, S. W. M. M., De Bok, F. A. M. M., Van Loo, N. D., Dijkema, C., Stams, A. J. M. M., and De Vos, W. M. (1994). Evidence for the operation of a novel Embden-Meyerhof pathway that involves ADP-dependent kinases during sugar fermentation by Pyrococcus furiosus. *J. Biol. Chem.*

Khatibi, P. A., Chou, C. J., Loder, A. J., Zurawski, J. V., Adams, M. W. W., and Kelly, R. M. (2017). Impact of growth mode, phase, and rate on the metabolic state of the extremely thermophilic archaeon Pyrococcus furiosus. *Biotechnol. Bioeng.* doi:10.1002/bit.26408.

Khemici, V., and Carpousis, A. J. (2003). The RNA degradosome and poly(A) polymerase of Escherichia coli are required in vivo for the degradation of small mRNA decay intermediates containing REP-stabilizers. *Mol. Microbiol.* 51, 777–790. doi:10.1046/j.1365-2958.2003.03862.x.

Kikuchi, A., and Asai, K. (1984). Reverse gyrase - A topoisomerase which introduces positive superhelical turns into DNA. *Nature*. doi:10.1038/309677a0.

Kim, M., Park, S., and Lee, S. J. (2016). Global transcriptional regulator TrmB family members in prokaryotes. *J. Microbiol.* doi:10.1007/s12275-016-6362-7.

Kitts, P. A., Church, D. M., Thibaud-Nissen, F., Choi, J., Hem, V., Sapojnikov, V., et al. (2016). Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res.* 44, D73–D80. doi:10.1093/nar/gkv1226.

Klappenbach, J. A. (2001). rrndb: the Ribosomal RNA Operon Copy Number Database. *Nucleic Acids Res.* 29, 181–184. doi:10.1093/nar/29.1.181.

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

Knüppel, R., Christensen, R. H., Gray, F. C., Esser, D., Strauß, D., Medenbach, J., et al. (2018). Insights into the evolutionary conserved regulation of Rio ATPase activity. *Nucleic Acids Res.* 46, 1441–1456. doi:10.1093/nar/gkx1236.

Knüppel, R., Fenk, M., Jüttner, M., and Ferreira-Cerca, S. (2020). "In vivo RNA chemical footprinting analysis in archaea," in *Methods in Molecular Biology* doi:10.1007/978-1-0716-0231-7_12.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., et al. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* doi:10.1101/gr.129684.111.

Kohler, R., Mooney, R. A., Mills, D. J., Landick, R., and Cramer, P. (2017). Architecture of a transcribing-translating expressome. *Science (80-.).* doi:10.1126/science.aal3059.

Koide, T., Reiss, D. J., Bare, J. C., Pang, W. L., Facciotti, M. T., Schmid, A. K., et al. (2009). Prevalence of transcription promoters within archaeal operons and coding sequences. *Mol. Syst. Biol.* doi:10.1038/msb.2009.42.

Koning, S. M., Konings, W. N., and Driessen, A. J. M. (2002). Biochemical evidence for the presence of two α-glucoside ABC-transport systems in the hyperthermophilic archaeon Pyrococcus furiosus. *Archaea*. doi:10.1155/2002/529610.

Koonin, E. V. (2009). Evolution of genome architecture. *Int. J. Biochem. Cell Biol.* doi:10.1016/j.biocel.2008.09.015.

Koonin, E. V., and Wolf, Y. I. (2008). Genomics of bacteria and archaea: The emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.* doi:10.1093/nar/gkn668.

Koonin, E. V., and Wolf, Y. I. (2010). Constraints and plasticity in genome and molecular-phenome evolution. *Nat. Rev. Genet.* doi:10.1038/nrg2810.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2016). Canu : scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.*, 1–35. doi:10.1101/gr.215087.116.Freely.

Korhonen, J., Martinmäki, P., Pizzi, C., Rastas, P., and Ukkonen, E. (2009). MOODS: Fast search for position weight matrix matches in DNA sequences. *Bioinformatics* 25, 3181–3182. doi:10.1093/bioinformatics/btp554.

Koskinen, K., Pausan, M. R., Perras, A. K., Beck, M., Bang, C., Mora, M., et al. (2017). First insights into the diverse human archaeome: Specific detection of Archaea in the gastrointestinal tract, lung, and nose and on skin. *MBio.* doi:10.1128/mBio.00824-17.

Kostrewa, D., Zeller, M. E., Armache, K.-J., Seizl, M., Leike, K., Thomm, M., et al. (2009). RNA polymerase II–TFIIB structure and mechanism of transcription initiation. *Nature* 462, 323–330. doi:10.1038/nature08548.

Kozarewa, I., Ning, Z., Quail, M. A., Sanders, M. J., Berriman, M., and Turner, D. J. (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat. Methods.* doi:10.1038/nmeth.1311.

Kreuzer, M., Schmutzler, K., Waege, I., Thomm, M., and Hausner, W. (2013). Genetic engineering of Pyrococcus furiosus to use chitin as a carbon source. *BMC Biotechnol.* doi:10.1186/1472-6750-13-9.

Kumarevel, T., Nakano, N., Ponnuraj, K., Gopinath, S. C. B., Sakamoto, K., Shinkai, A., et al. (2008). Crystal structure of glutamine receptor protein from Sulfolobus tokodaii strain 7 in complex with its effector l -glutamine: implications of effector binding in molecular association and DNA binding. *Nucleic Acids Res.* 36, 4808–4820. doi:10.1093/nar/gkn456.

Kuo, C.-H., and Ochman, H. (2009). Deletional Bias across the Three Domains of Life. *Genome Biol. Evol.* doi:10.1093/gbe/evp016.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., et al. (2004). Versatile and open software for comparing large genomes. *Genome Biol.* doi:10.1186/gb-2004-5-2-r12.

Kyrpides, N. C., and Ouzounis, C. A. (1999). Transcription in Archaea. *Proc. Natl. Acad. Sci. U. S. A.* 96, 8545–8550. doi:10.1073/pnas.96.15.8545.

Laass, S., Monzon, V. A., Kliemt, J., Hammelmann, M., Pfeiffer, F., Förstner, K. U., et al. (2019). Characterization of the

transcriptome of Haloferax volcanii, grown under four different conditions, with mixed RNA-Seq. *PLoS One* 14, e0215986. doi:10.1371/journal.pone.0215986.

Lafontaine, D. L. J., Preiss, T., and Tollervey, D. (1998). Yeast 18S rRNA Dimethylase Dim1p: a Quality Control Mechanism in Ribosome Synthesis? *Mol. Cell. Biol.* 18, 2360–2370. doi:10.1128/MCB.18.4.2360.

Lagorce, A., Fourçans, A., Dutertre, M., Bouyssiere, B., Zivanovic, Y., and Confalonieri, F. (2012). Genome-wide transcriptional response of the Archaeon Thermococcus gammatolerans to Cadmium. *PLoS One* 7. doi:10.1371/journal.pone.0041935.

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., et al. (2018). The Human Transcription Factors. *Cell*. doi:10.1016/j.cell.2018.01.029.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*. doi:10.1038/35057062.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi:10.1038/nmeth.1923.

Laszlo, A. H., Derrington, I. M., Ross, B. C., Brinkerhoff, H., Adey, A., Nova, I. C., et al. (2014). Decoding long nanopore sequencing reads of natural DNA. *Nat. Biotechnol.* doi:10.1038/nbt.2950.

Laver, T., Harrison, J., O'Neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., et al. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol. Detect. Quantif.* doi:10.1016/j.bdq.2015.02.001.

Lee, H. S., Shockley, K. R., Schut, G. J., Conners, S. B., Montero, C. I., Johnson, M. R., et al. (2006). Transcriptional and biochemical analysis of starch metabolism in the hyperthermophilic archaeon Pyrococcus furiosus. *J. Bacteriol.* doi:10.1128/JB.188.6.2115-2125.2006.

Lee, S. J., Surma, M., Seitz, S., Hausner, W., Thomm, M., and Boos, W. (2007). Characterization of the TrmB-like protein, PF0124, a TGM-recognizing global transcriptional regulator of the hyperthermophilic archaeon Pyrococcus furiosus. *Mol. Microbiol.* 65, 305–318. doi:10.1111/j.1365-2958.2007.05780.x.

Leger, A., Amaral, P. P., Pandolfini, L., Capitanchik, C., Capraro, F., Barbieri, I., et al. (2019). RNA modifications detection by comparative Nanopore direct RNA sequencing. *bioRxiv*, 843136. doi:10.1101/843136.

Leigh, J. A., Albers, S. V., Atomi, H., and Allers, T. (2011). Model organisms for genetics in the domain Archaea: Methanogens, halophiles, Thermococcales and Sulfolobales. *FEMS Microbiol. Rev.* doi:10.1111/j.1574-6976.2011.00265.x.

Leleu, M., Lefebvre, G., and Rougemont, J. (2010). Processing and analyzing ChIP-seq data: From short reads to regulatory interactions. *Brief. Funct. Genomics*. doi:10.1093/bfgp/elq022.

Lemmens, L., Maklad, H. R., Bervoets, I., and Peeters, E. (2019). Transcription Regulators in Archaea: Homologies and Differences with Bacterial Regulators. *J. Mol. Biol.* 431, 4132–4146. doi:10.1016/j.jmb.2019.05.045.

Leonard, P. M., Smits, S. H. J., Sedelnikova, S. E., Brinkman, A. B., De Vos, W. M., Van Der Oost, J., et al. (2001). Crystal structure of the Lrp-like transcriptional regulator from the archaeon Pyrococcus furiosus. *EMBO J.* 20, 990–997. doi:10.1093/emboj/20.5.990.

Levene, H. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G., and Webb, W. W. (2003). Zero-mode waveguides for single-molecule analysis at high concentrations. *Science (80-.).* doi:10.1126/science.1079700.

Levy, S. E., and Myers, R. M. (2016). Advancements in Next-Generation Sequencing. *Annu. Rev. Genomics Hum. Genet.* 17, 95–115. doi:10.1146/annurev-genom-083115-022413.

Lewis, D. L., Notey, J. S., Chandrayan, S. K., Loder, A. J., Lipscomb, G. L., Adams, M. W. W., et al. (2015). A mutant ('lab strain') of the hyperthermophilic archaeon Pyrococcus furiosus, lacking flagella, has unusual growth physiology. *Extremophiles*. doi:10.1007/s00792-014-0712-3.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi:10.1093/bioinformatics/bty191.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. doi:10.1093/bioinformatics/btp698.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352.

Li, S., and Mason, C. E. (2014). The Pivotal Regulatory Landscape of RNA Modifications. *Annu. Rev. Genomics Hum. Genet.* 15, 127–150. doi:10.1146/annurev-genom-090413-025405.

Liang, W., Rudd, K. E., and Deutscher, M. P. (2015). A Role for REP Sequences in Regulating Translation. *Mol. Cell* 58, 431–439. doi:10.1016/j.molcel.2015.03.019.

Liao, Y., Smyth, G. K., and Shi, W. (2019). The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.* 47, e47–e47. doi:10.1093/nar/gkz114.

Lilley, D. M. J. (2012). The structure and folding of kink turns in RNA. *Wiley Interdiscip. Rev. RNA* 3, 797–805. doi:10.1002/wrna.1136.

Lioliou, E., Sharma, C. M., Caldelari, I., Helfer, A.-C. C., Fechter, P., Vandenesch, F., et al. (2012). Global Regulatory Functions of the Staphylococcus aureus Endoribonuclease III in Gene Expression. *PLoS Genet.* 8, e1002782. doi:10.1371/journal.pgen.1002782.

Lipscomb, G. L., Keese, A. M., Cowart, D. M., Schut, G. J., Thomm, M., Adams, M. W. W., et al. (2009). SurR: A transcriptional activator and repressor controlling hydrogen and elemental sulphur metabolism in Pyrococcus furiosus. *Mol. Microbiol.* doi:10.1111/j.1365-2958.2008.06525.x.

Lipscomb, G. L., Stirrett, K., Schut, G. J., Yang, F., Jenney, F. E., Scott, R. A., et al. (2011). Natural competence in the hyperthermophilic archaeon Pyrococcus furiosus facilitates genetic manipulation: Construction of markerless deletions of genes encoding the two cytoplasmic hydrogenases. *Appl. Environ. Microbiol.* doi:10.1128/AEM.02624-10.

Littlefield, J. W., and Dunn, D. B. (1958). Natural occurrence of thymine and three methylated adenine bases in several ribonucleic acids. *Nature*. doi:10.1038/181254a0.

Littlefield, O., Korkhin, Y., and Sigler, P. B. (1999). The structural basis for the oriented assembly of a TBP/TFB/promoter complex. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.96.24.13668.

Liu, H., Begik, O., Lucas, M. C., Ramirez, J. M., Mason, C. E., Wiener, D., et al. (2019). Accurate detection of m6A RNA modifications in native RNA sequences. *Nat. Commun.*, 1–9. doi:10.1038/s41467-019-11713-9.

Liu, H., Orell, A., Maes, D., van Wolferen, M., Ann-Christin, L., Bernander, R., et al. (2014). BarR, an Lrp-type transcription factor in S ulfolobus acidocaldarius , regulates an aminotransferase gene in a β-alanine responsive manner. *Mol. Microbiol.* 92, 625–639. doi:10.1111/mmi.12583.

Lloréns-Rico, V., Cano, J., Kamminga, T., Gil, R., Latorre, A., Chen, W.-H., et al. (2016). Bacterial antisense RNAs are mainly the product of transcriptional noise. *Sci. Adv.* 2, 1–10. doi:10.1126/sciadv.1501363.

Loman, N. J., Quick, J., and Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* 12, 733–735. doi:10.1038/nmeth.3444.

López-Maury, L., Marguerat, S., and Bähler, J. (2008). Tuning gene expression to changing environments: From rapid responses to evolutionary adaptation. *Nat. Rev. Genet.* doi:10.1038/nrg2398.

Lorenz, D. A., Sathe, S., Einstein, J. M., and Yeo, G. W. (2020). Direct RNA sequencing enables m 6 A detection in endogenous transcript isoforms at base-specific resolution. *RNA* 26, 19–28. doi:10.1261/rna.072785.119.

Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., et al. (2011). ViennaRNA Package 2.0. *Algorithms Mol. Biol.* 6, 26. doi:10.1186/1748-7188-6-26.

Love, M. I., Anders, S., and Huber, W. (2014a). *Differential analysis of count data - the DESeq2 package.* doi:110.1186/s13059-014-0550-8.

Love, M. I., Huber, W., and Anders, S. (2014b). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550. doi:10.1186/s13059-014-0550-8.

Lowy, E. (2017). CoverageView: Coverage visualization package for R.

Lundberg, K. S., Shoemaker, D. D., Adams, M. W. W., Short, J. M., Sorge, J. A., and Mathur, E. J. (1991). High-fidelity amplification using a thermostable DNA polymerase isolated from Pyrococcus furiosus. *Gene.* doi:10.1016/0378-1119(91)90480-Y.

Lurie-Weinberger, M. N., and Gophna, U. (2015). Archaea in and on the Human Body: Health Implications and Future Directions. *PLoS Pathog.* doi:10.1371/journal.ppat.1004833.

Lynch, M. (2006). Streamlining and Simplification of Microbial Genome Architecture. *Annu. Rev. Microbiol.* doi:10.1146/annurev.micro.60.080805.142300.

Lyu, Z., Li, Z.-G., He, F., and Zhang, Z. (2017). An Important Role for Purifying Selection in Archaeal Genome Evolution. *mSystems.* doi:10.1128/msystems.00112-17.

Macomber, L., and Imlay, J. A. (2009). The iron-sulfur clusters of dehydratases are primary intracellular targets of copper toxicity. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.0812808106.

Maeda, M., Shimada, T., and Ishihama, A. (2015). Strength and Regulation of Seven rRNA Promoters in Escherichia coli. *PLoS One* 10, e0144697. doi:10.1371/journal.pone.0144697.

Maier, L. K., and Marchfelder, A. (2019). It's all about the T: Transcription termination in Archaea. *Biochem. Soc. Trans.* 47, 461–468. doi:10.1042/BST20180557.

Makarova, K., Wolf, Y., and Koonin, E. (2015). Archaeal Clusters of Orthologous Genes (arCOGs): An Update and Application for Analysis of Shared Features between Thermococcales, Methanococcales, and Methanobacteriales. *Life* 5, 818–840. doi:10.3390/life5010818.

Mana-Capelli, S., Mandal, A. K., and Argüello, J. M. (2003). Archaeoglobus fulgidus CopB is a thermophilic Cu2+-ATPase: Functional role of its histidine-rich N-terminal metal binding domain. *J. Biol. Chem.* doi:10.1074/jbc.M306907200.

Mandal, A. K., Cheung, W. D., and Argüello, J. M. (2002). Characterization of a thermophilic P-type Ag+/Cu+-ATPase from the extremophile Archaeoglobus fulgidus. *J. Biol. Chem.* doi:10.1074/jbc.M109964200.

Maniatis, T., Jeffrey, A., and van deSande, H. (1975). Chain Length Determination of Small Doubleand Single-Stranded DNA Molecules by Polyacrylamide Gel Electrophoresis. *Biochemistry.* doi:10.1021/bi00688a010.

Manning, K. S., and Cooper, T. A. (2017). The roles of RNA processing in translating genotype to phenotype. *Nat. Rev. Mol. Cell Biol.* doi:10.1038/nrm.2016.139.

Manrao, E. A., Derrington, I. M., Laszlo, A. H., Langford, K. W., Hopper, M. K., Gillgren, N., et al. (2012). Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nat. Biotechnol.* doi:10.1038/nbt.2171.

Manrao, E. A., Derrington, I. M., Pavlenok, M., Niederweis, M., and Gundlach, J. H. (2011). Nucleotide discrimination with DNA immobilized in the MSPA nanopore. *PLoS One.* doi:10.1371/journal.pone.0025723.

Mao, X., Ma, Q., Liu, B., Chen, X., Zhang, H., and Xu, Y. (2015). Revisiting operons: an analysis of the landscape of transcriptional units in E. coli. *BMC Bioinformatics* 16, 356. doi:10.1186/s12859-015-0805-8.

Mao, X., Ma, Q., Zhou, C., Chen, X., Zhang, H., Yang, J., et al. (2014). DOOR 2.0: presenting operons and their functions through dynamic and integrated views. *Nucleic Acids Res.* 42, D654–D659. doi:10.1093/nar/gkt1048.

Maraia, R. J., Arimbasseri, A. G., and Maraia, R. J. (2015). Mechanism of Transcription Termination by RNA Polymerase III Utilizes a Non-template Strand Sequence-Specific Signal Element. *Mol. Cell* 58, 1124–1132. doi:10.1016/j.molcel.2015.04.002.

Marchfelder, A., Fischer, S., Brendel, J., Stoll, B., Maier, L. K., Jäger, D., et al. (2012). Small RNAs for defence and regulation in archaea. *Extremophiles*. doi:10.1007/s00792-012-0469-5.

Marguet, E., and Forterre, P. (1998). Protection of DNA by salts against thermodegradation at temperatures typical for hyperthermophiles. *Extremophiles*. doi:10.1007/s007920050050.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. doi:10.1038/nature03959.

Martell, D. J., Joshi, C. P., Gaballa, A., Santiago, A. G., Chen, T.-Y., Jung, W., et al. (2015). Metalloregulator CueR biases RNA polymerase's kinetic sampling of dead-end or open complex to repress or activate transcription. *Proc. Natl. Acad. Sci.* 112, 13467–13472. doi:10.1073/pnas.1515231112.

Märtens, B., Manoharadas, S., Hasenöhrl, D., Manica, A., and Bläsi, U. (2013). Antisense regulation by transposon-derived RNAs in the hyperthermophilic archaeon Sulfolobus solfataricus. *EMBO Rep.* doi:10.1038/embor.2013.47.

Märtens, B., Sharma, K., Urlaub, H., and Bläsi, U. (2017). The SmAP2 RNA binding motif in the 3′UTR affects mRNA stability in the crenarchaeum Sulfolobus solfataricus. *Nucleic Acids Res.* doi:10.1093/nar/gkx581.

Martínez-Bussenius, C., Navarro, C. A., and Jerez, C. A. (2017). Microbial copper resistance: importance in biohydrometallurgy. *Microb. Biotechnol.* doi:10.1111/1751-7915.12450.

Martinez-Pastor, M., Tonner, P. D., Darnell, C. L., and Schmid, A. K. (2017). Transcriptional Regulation in Archaea : From Individual Genes to Global Regulatory Networks. *Annu. Rev. Genet.* doi:10.1146/annurev-genet-120116-023413.

Maruyama, H., Shin, M., Oda, T., Matsumi, R., Ohniwa, R. L., Itoh, T., et al. (2011). Histone and TK0471/TrmBL2 form a novel heterogeneous genome architecture in the hyperthermophilic archaeon Thermococcus kodakarensis. *Mol. Biol. Cell* 22, 386–398. doi:10.1091/mbc.e10-08-0668.

Matte-Tailliez, O., Brochier, C., Forterre, P., and Philippe, H. (2002). Archaeal phylogeny based on ribosomal proteins. *Mol. Biol. Evol.* doi:10.1093/oxfordjournals.molbev.a004122.

Maxam, A. M., and Gilbert, W. (1977). A new method for sequencing DNA. *Proc. Natl. Acad. Sci.* 74, 560–564. doi:10.1073/pnas.74.2.560.

Mayer, A., Lidschreiber, M., Siebert, M., Leike, K., Söding, J., and Cramer, P. (2010). Uniform transitions of the general RNA polymerase II transcription complex. *Nat. Struct. Mol. Biol.* doi:10.1038/nsmb.1903.

McGary, K., and Nudler, E. (2013). RNA polymerase and the ribosome: The close relationship. *Curr. Opin. Microbiol.* doi:10.1016/j.mib.2013.01.010.

McGlynn, P., Savery, N. J., and Dillingham, M. S. (2012). The conflict between DNA replication and transcription. *Mol. Microbiol.* doi:10.1111/j.1365-2958.2012.08102.x.

McInerney, P., Adams, P., and Hadi, M. Z. (2014). Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase. *Mol. Biol. Int.* doi:10.1155/2014/287430.

Mejía-Almonte, C., Busby, S. J. W., Wade, J. T., van Helden, J., Arkin, A. P., Stormo, G. D., et al. (2020). Redefining fundamental concepts of transcription initiation in bacteria. *Nat. Rev. Genet.* doi:10.1038/s41576-020-0254-8.

Meller, A., Nivon, L., Brandin, E., Golovchenko, J., and Branton, D. (2000). Rapid nanopore discrimination between single polynucleotide molecules. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.97.3.1079.

Meloni, G., Zhang, L., and Rees, D. C. (2014). Transmembrane type-2-like Cu2+ site in the P1B-3-type ATPase CopB: Implications for metal selectivity. *ACS Chem. Biol.* doi:10.1021/cb400603t.

Messing, J., Crea, R., and Seeburg, P. H. (1981). A system for shotgun DNA sequencing. *Nucleic Acids Res.* doi:10.1093/nar/9.2.309.

Meysman, P., Collado-Vides, J., Morett, E., Viola, R., Engelen, K., and Laukens, K. (2014). Structural properties of prokaryotic promoter regions correlate with functional features. *PLoS One* 9. doi:10.1371/journal.pone.0088717.

Micorescu, M., Grünberg, S., Franke, A., Cramer, P., Thomm, M., and Bartlett, M. (2008). Archaeal transcription: Function of an alternative transcription factor B from Pyrococcus furiosus. *J. Bacteriol.* doi:10.1128/JB.01498-07.

Mikheyev, A. S., and Tin, M. M. Y. (2014). A first look at the Oxford Nanopore MinION sequencer. *Mol. Ecol. Resour.* 14, 1097–1102. doi:10.1111/1755-0998.12324.

Mitra, P., Ghosh, G., Hafeezunnisa, M., and Sen, R. (2017). Rho Protein: Roles and Mechanisms. *Annu. Rev. Microbiol.* 71, 687–709. doi:10.1146/annurev-micro-030117-020432.

Mitra, R. (1999). In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Res.* doi:10.1093/nar/27.24.e34.

Mitra, R. D., Shendure, J., Olejnik, J., Krzymanska-Olejnik, E., and Church, G. M. (2003). Fluorescent in situ sequencing on polymerase colonies. *Anal. Biochem.* doi:10.1016/S0003-2697(03)00291-4.

Mohanty, B. K., and Kushner, S. R. (2016). Regulation of mRNA Decay in Bacteria. *Annu. Rev. Microbiol.* doi:10.1146/annurev-micro-091014-104515.

Moissl-Eichinger, C., Probst, A. J., Birarda, G., Auerbach, A., Koskinen, K., Wolf, P., et al. (2017). Human age and skin physiology shape diversity and abundance of Archaea on skin. *Sci. Rep.* doi:10.1038/s41598-017-04197-4.

Mongan, A. E., Tuda, J. S. B., and Runtuwene, L. R. (2020). Portable sequencer in the fight against infectious disease. *J. Hum. Genet.* doi:10.1038/s10038-019-0675-4.

Mukund, S., and Adams, M. W. W. (1995). Glyceraldehyde-3-phosphate ferredoxin oxidoreductase, a novel tungsten-containing enzyme with a potential glycolytic role in the hyperthermophilic archaeon Pyrococcus furiosus. *J. Biol. Chem.* doi:10.1074/jbc.270.15.8389.

Murina, V. N., and Nikulin, A. D. (2011). RNA-binding Sm-like proteins of bacteria and archaea. Similarity and difference in structure and function. *Biochem.* 76, 1434–1449. doi:10.1134/S0006297911130050.

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., et al. (2000). A whole-genome assembly of Drosophila. *Science (80-.).* doi:10.1126/science.287.5461.2196.

Nagy, J., Grohmann, D., Cheung, A. C. M., Schulz, S., Smollett, K., Werner, F., et al. (2015). Complete architecture of the archaeal RNA polymerase open complex from single-molecule FRET and NPS. *Nat. Commun.* 6, 6161. doi:10.1038/ncomms7161.

Nakashima, H., Fukuchi, S., and Nishikawa, K. (2003). Compositional changes in RNA, DNA and proteins for bacterial adaptation to higher and lower temperatures. *J. Biochem.* doi:10.1093/jb/mvg067.

Napoli, A., Van Der Oost, J., Sensen, C. W., Charlebois, R. L., Rossi, M., and Ciaramella, M. (1999). An Lrp-like protein of the hyperthermophilic archaeon Sulfolobus solfataricus which binds to its own promoter. *J. Bacteriol.* doi:10.1128/jb.181.5.1474-1480.1999.

Näther-Schindler, D. J., Schopf, S., Bellack, A., Rachel, R., and Wirth, R. (2014). Pyrococcus furiosus flagella: biochemical and transcriptional analyses identify the newly detected flaB0 gene to encode the major flagellin. *Front. Microbiol.* 5. doi:10.3389/fmicb.2014.00695.

Näther, D. J., Rachel, R., Wanner, G., and Wirth, R. (2006). Flagella of Pyrococcus furiosus: Multifunctional organelles, made for swimming, adhesion to various surfaces, and cell-cell contacts. *J. Bacteriol.* doi:10.1128/JB.00527-06.

Nelson-Sathi, S., Sousa, F. L., Roettger, M., Lozada-Chávez, N., Thiergart, T., Janssen, A., et al. (2015). Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature.* doi:10.1038/nature13805.

Nickels, P. C., Wünsch, B., Holzmeister, P., Bae, W., Kneer, L. M., Grohmann, D., et al. (2016). Molecular force spectroscopy with a DNA origami-based nanoscopic force clamp. *Science (80-.).* doi:10.1126/science.aah5974.

Nikolaev, N., Silengo, L., and Schlessinger, D. (1973). Synthesis of a Large Precursor to Ribosomal RNA in a Mutant of Escherichia coli. *Proc. Natl. Acad. Sci.* 70, 3361–3365. doi:10.1073/pnas.70.12.3361.

Nouaille, S., Mondeil, S., Finoux, A. L., Moulis, C., Girbal, L., and Cocaign-Bousquet, M. (2017). The stability of an mRNA is influenced by its concentration: A potential physical mechanism to regulate gene expression. *Nucleic Acids Res.* doi:10.1093/nar/gkx781.

Nowrousian, M. (2010). Next-Generation Sequencing Techniques for Eukaryotic Microorganisms: Sequencing-Based Solutions to Biological Problems. *Eukaryot. Cell* 9, 1300–1310. doi:10.1128/EC.00123-10.

O'Farrell, H. C., Pulicherla, N., Desai, P. M., and Rife, J. P. (2006). Recognition of a complex substrate by the KsgA/Dim1 family of enzymes has been conserved throughout evolution. *RNA* 12, 725–733. doi:10.1261/rna.2310406.

O'Reilly, F. J., Xue, L., Graziadei, A., Sinn, L., Lenz, S., Tegunov, D., et al. (2020). In-cell architecture of an actively transcribing-translating expressome. *Science (80-.).* 369, 554–557. doi:10.1126/science.abb3758.

Ochs, S. M., Thumann, S., Richau, R., Weirauch, M. T., Lowe, T. M., Thomm, M., et al. (2012). Activation of Archaeal Transcription Mediated by Recruitment of Transcription Factor B. *J. Biol. Chem.* 287, 18863–18871. doi:10.1074/jbc.M112.365742.

Oliva, G., Sahr, T., and Buchrieser, C. (2015). Small RNAs, 5′ UTR elements and RNA-binding proteins in intracellular bacteria: impact on metabolism and virulence. *FEMS Microbiol. Rev.* 39, 331–349. doi:10.1093/femsre/fuv022.

Orell, A., Remonsellez, F., Arancibia, R., and Jerez, C. A. (2013). Molecular characterization of copper and cadmium resistance determinants in the biomining thermoacidophilic archaeon sulfolobus metallicus. *Archaea* 2013. doi:10.1155/2013/289236.

Otto, C., Stadler, P. F., and Hoffmann, S. (2014). Lacking alignments? The next-generation sequencing mapper segemehl revisited. *Bioinformatics* 30, 1837–1843. doi:10.1093/bioinformatics/btu146.

Ouhammouch, M. (2001). A thermostable platform for transcriptional regulation: the DNA-binding properties of two Lrp homologs from the hyperthermophilic archaeon Methanococcus jannaschii. *EMBO J.* 20, 146–156. doi:10.1093/emboj/20.1.146.

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

Ouhammouch, M., Langham, G. E., Hausner, W., Simpson, A. J., El-Sayed, N. M. A., and Geiduschek, E. P. (2005). Promoter architecture and response to a positive regulator of archaeal transcription. *Mol. Microbiol.* 56, 625–637. doi:10.1111/j.1365-2958.2005.04563.x.

Pace, N. R. (1991). Origin of life-facing up to the physical setting. *Cell.* doi:10.1016/0092-8674(91)90082-A.

Parker, M. T., Knop, K., Sherwood, A. V, Schurch, N. J., Mackinnon, K., Gould, P. D., et al. (2020). Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m6A modification. *Elife* 9. doi:10.7554/eLife.49658.

Payne, A., Holmes, N., Rakyan, V., and Loose, M. (2019). Bulkvis: A graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics.* doi:10.1093/bioinformatics/bty841.

Pedersen, R. B., Rapp, H. T., Thorseth, I. H., Lilley, M. D., Barriga, F. J. A. S., Baumberger, T., et al. (2010). Discovery of a black smoker vent field and vent fauna at the Arctic Mid-Ocean Ridge. *Nat. Commun.* doi:10.1038/ncomms1124.

Peeters, E., Albers, S.-V., Vassart, A., Driessen, A. J. M., and Charlier, D. (2009). Ss-LrpB, a transcriptional regulator from Sulfolobus solfataricus , regulates a gene cluster with a pyruvate ferredoxin oxidoreductase-encoding operon and permease genes. *Mol. Microbiol.* 71, 972–988. doi:10.1111/j.1365-2958.2008.06578.x.

Peeters, E., and Charlier, D. (2010). The Lrp Family of Transcription Regulators in Archaea. *Archaea* 2010, 1–10. doi:10.1155/2010/750457.

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

Peeters, E., Peixeiro, N., and Sezonov, G. (2013). Cis-regulatory logic in archaeal transcription. in *Biochemical Society Transactions*, 326–31. doi:10.1042/BST20120312.

Peeters, E., Thia-Toong, T.-L., Gigot, D., Maes, D., and Charlier, D. (2004). Ss-LrpB, a novel Lrp-like regulator of Sulfolobus solfataricus P2, binds cooperatively to three conserved targets in its own control region. *Mol. Microbiol.* 54, 321–336. doi:10.1111/j.1365-2958.2004.04274.x.

Peeters, E., Willaert, R., Maes, D., and Charlier, D. (2006). Ss-LrpB from Sulfolobus solfataricus condenses about 100 base pairs of its own operator DNA into globular nucleoprotein complexes. *J. Biol. Chem.* doi:10.1074/jbc.M600383200.

Peng, N., Xia, Q., Chen, Z., Liang, Y. X., and She, Q. (2009). An upstream activation element exerting differential transcriptional activation on an archaeal promoter. *Mol. Microbiol.* doi:10.1111/j.1365-2958.2009.06908.x.

Pennisi, E. (2010). Semiconductors inspire new sequencing technologies. *Science (80-.).* doi:10.1126/science.327.5970.1190.

Perez-Rueda, E., Hernandez-Guerrero, R., Martinez-Nuñez, M. A., Armenta-Medina, D., Sanchez, I., and Ibarra, J. A. (2018). Abundance, diversity and domain architecture variability in prokaryotic DNA-binding transcription factors. *PLoS One.* doi:10.1371/journal.pone.0195332.

Perez-Rueda, E., and Janga, S. C. (2010). Identification and genomic analysis of transcription factors in archaeal genomes exemplifies their functional architecture and evolutionary origin. *Mol. Biol. Evol.* 27, 1449–1459. doi:10.1093/molbev/msq033.

Peters, J. M., Vangeloff, A. D., and Landick, R. (2011). Bacterial transcription terminators: The RNA 3′-end chronicles. *J. Mol. Biol.* doi:10.1016/j.jmb.2011.03.036.

Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M., Schultze, J. L., et al. (2018). Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci. Rep.* doi:10.1038/s41598-018-29325-6.

Pham, A. N., Xing, G., Miller, C. J., and Waite, T. D. (2013). Fenton-like copper redox chemistry revisited: Hydrogen peroxide and superoxide mediation of copper-catalyzed oxidant production. *J. Catal.* doi:10.1016/j.jcat.2013.01.025.

Philips, S. J., Canalizo-Hernandez, M., Yildirim, I., Schatz, G. C., Mondragon, A., and O'Halloran, T. V. (2015). Allosteric transcriptional regulation via changes in the overall topology of the core promoter. *Science (80-.).* 349, 877–881. doi:10.1126/science.aaa9809.

Picard, F., Dressaire, C., Girbal, L., and Cocaign-Bousquet, M. (2009). Examination of post-transcriptional regulations in prokaryotes by integrative biology. *Comptes Rendus - Biol.* doi:10.1016/j.crvi.2009.09.005.

Piekna-Przybylska, D. D., Decatur, W. A., and Fournier, M. J. (2008). The 3D rRNA modification maps database: With interactive tools for ribosome analysis. *Nucleic Acids Res.* doi:10.1093/nar/gkm855.

Plagens, A., Daume, M., Wiegel, J., and Randau, L. (2015). Circularization restores signal recognition particle RNA functionality in Thermoproteus. *Elife.* doi:10.7554/elife.11623.

Plaisier, C. L., Lo, F.-Y., Ashworth, J., Brooks, A. N., Beer, K. D., Kaur, A., et al. (2014). Evolution of context dependent regulation by expansion of feast/famine regulatory proteins. *BMC Syst. Biol.* 8, 122. doi:10.1186/s12918-014-0122-2.

Poole, F. L., Gerwe, B. A., Hopkins, R. C., Schut, G. J., Weinberg, M. V, Jenney, F. E., et al. (2005). Defining genes in the genome of the hyperthermophilic archaeon Pyrococcus furiosus: implications for all microbial genomes. *J. Bacteriol.* 187, 7325–32. doi:10.1128/JB.187.21.7325-7332.2005.

Porreca, G. J. (2010). Genome sequencing on nanoballs. *Nat. Biotechnol.* doi:10.1038/nbt0110-43.

Price, M. T., Fullerton, H., and Moyer, C. L. (2015). Biogeography and evolution of Thermococcus isolates from hydrothermal vent systems of the Pacific. *Front. Microbiol.* doi:10.3389/fmicb.2015.00968.

Pritchett, M. A., Wilkinson, S. P., Geiduschek, E. P., and Ouhammouch, M. (2009). Hybrid Ptr2-like activators of archaeal transcription. *Mol. Microbiol.* doi:10.1111/j.1365-2958.2009.06884.x.

Prober, J. M., Trainor, G. L., Dam, R. J., Hobbs, F. W., Robertson, C. W., Zagursky, R. J., et al. (1987). A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science (80-.).* doi:10.1126/science.2443975.

Probst, A. J., Auerbach, A. K., and Moissl-Eichinger, C. (2013). Archaea on Human Skin. *PLoS One.* doi:10.1371/journal.pone.0065388.

Proshkin, S., Rachid Rahmouni, A., Mironov, A., Nudler, E., Rahmouni, A. R., Mironov, A., et al. (2010). Cooperation Between Translating Ribosomes and RNA Polymerase in Transcription Elongation. *Science (80-.).* 328, 504–508. doi:10.1126/science.1184939.

Purohit, R., Ross, M. O., Batelu, S., Kusowski, A., Stemmler, T. L., Hoffman, B. M., et al. (2018). Cu+-specific CopB transporter: Revising p1B-type ATPase classification. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.1721783115.

Pushkarev, D., Neff, N. F., and Quake, S. R. (2009). Single-molecule sequencing of an individual human genome. *Nat. Biotechnol.* doi:10.1038/nbt.1561.

Qi, L., Li, J., Jia, J., Yue, L., and Dong, X. (2020). Comprehensive analysis of the pre-ribosomal RNA maturation pathway in a methanoarchaeon exposes the conserved circularization and linearization mode in archaea. *RNA Biol.*, 1–15.

doi:10.1080/15476286.2020.1771946.

Qi, L., Yue, L., Feng, D., Qi, F., Li, J., and Dong, X. (2017). Genome-wide mRNA processing in methanogenic archaea reveals post-transcriptional regulation of ribosomal protein synthesis. *Nucleic Acids Res.* doi:10.1093/nar/gkx454.

Quick, J., Grubaugh, N. D., Pullan, S. T., Claro, I. M., Smith, A. D., Gangavarapu, K., et al. (2017). Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat. Protoc.* doi:10.1038/nprot.2017.066.

Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., et al. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature.* doi:10.1038/nature16996.

R Development Core Team, R. (2011). *R: A Language and Environment for Statistical Computing.* doi:10.1007/978-3-540-74686-7.

R Foundation for Statistical Computing. (2018). *R: a Language and Environment for Statistical Computing.*

Rahimi, K., Venø, M. T., Dupont, D. M., and Kjems, J. (2019). Nanopore sequencing of full-length circRNAs in human and mouse brains reveals circRNA-specific exon usage and intron retention. *bioRxiv*, 567164. doi:10.1101/567164.

Rand, A. C., Jain, M., Eizenga, J. M., Musselman-Brown, A., Olsen, H. E., Akeson, M., et al. (2017). Mapping DNA methylation with high-throughput nanopore sequencing. *Nat. Methods* 14, 411–413. doi:10.1038/nmeth.4189.

Rang, F. J., Kloosterman, W. P., and de Ridder, J. (2018). From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* 19, 90. doi:10.1186/s13059-018-1462-9.

Ray-Soni, A., Bellecourt, M. J., and Landick, R. (2016). Mechanisms of Bacterial Transcription Termination: All Good Things Must End. *Annu. Rev. Biochem.* 85, 319–347. doi:10.1146/annurev-biochem-060815-014844.

Reichelt, R., Gindner, A., Thomm, M., and Hausner, W. (2016). Genome-wide binding analysis of the transcriptional regulator TrmBL1 in Pyrococcus furiosus. *BMC Genomics* 17, 40. doi:10.1186/s12864-015-2360-0.

Reichelt, R., Grohmann, D., and Willkomm, S. (2018a). A journey through the evolutionary diversification of archaeal Lsm and Hfq proteins*. *Emerg. Top. Life Sci.* doi:10.1042/etls20180034.

Reichelt, R., Ruperti, K. M. A. A., Kreuzer, M., Dexl, S., Thomm, M., and Hausner, W. (2018b). The transcriptional regulator TFB-RF1 activates transcription of a putative ABC transporter in Pyrococcus furiosus. *Front. Microbiol.* 9. doi:10.3389/fmicb.2018.00838.

Reiner, J. E., Balijepalli, A., Robertson, J. W. F., Drown, B. S., Burden, D. L., and Kasianowicz, J. J. (2012). The effects of diffusion on an exonucleasenanopore-based DNA sequencing engine. *J. Chem. Phys.* doi:10.1063/1.4766363.

Remaut, H., Jayasinghe, L., Howorka, S., Wallace, J., Clarke, J., Hambley, R., et al. (2014). *Mutant CsgG Pores.*

Ren, G.-X., Guo, X.-P., and Sun, Y.-C. (2017). Regulatory 3′ Untranslated Regions of Bacterial mRNAs. *Front. Microbiol.* 8. doi:10.3389/fmicb.2017.01276.

Rensing, C., and McDevitt, S. F. (2013). "The Copper Metallome in Prokaryotic Cells," in doi:10.1007/978-94-007-5561-1_12.

Revyakin, A., Liu, C., Ebright, R. H., and Strick, T. R. (2006). Abortive initiation and productive initiation by RNA polymerase involve DNA scrunching. *Science (80-.).* doi:10.1126/science.1131398.

Rhoads, A., and Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics Bioinforma.* doi:10.1016/j.gpb.2015.08.002.

Riley, M., Abe, T., Arnaud, M. B., Berlyn, M. K. B., Blattner, F. R., Chaudhuri, R. R., et al. (2006). Escherichia coli K-12: a cooperatively developed annotation snapshot--2005. *Nucleic Acids Res.* 34, 1–9. doi:10.1093/nar/gkj405.

Rivera, M. C., and Lake, J. A. (1992). Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *Science (80-.).* doi:10.1126/science.1621096.

Rivero, M., Torres-Paris, C., Muñoz, R., Cabrera, R., Navarro, C. A., and Jerez, C. A. (2018). Inorganic Polyphosphate, Exopolyphosphatase, and Pho84 -Like Transporters May Be Involved in Copper Resistance in Metallosphaera sedula DSM 5348 T. *Archaea* 2018, 1–12. doi:10.1155/2018/5251061.

Robb, F. T., Maeder, D. L., Brown, J. R., DiRuggiero, J., Stump, M. D., Yeh, R. K., et al. (2001). Genomic sequence of hyperthermophile, Pyrococcus furiosus: implications for physiology and enzymology. *Methods Enzymol.* 330, 134–57. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11210495 [Accessed December 11, 2015].

Roberts, J. W. (2019). Mechanisms of Bacterial Transcription Termination. *J. Mol. Biol.* doi:10.1016/j.jmb.2019.04.003.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative genomics viewer. *Nat. Biotechnol.* doi:10.1038/nbt.1754.

Rocha, E. P. C. (2008). The Organization of the Bacterial Genome. *Annu. Rev. Genet.* doi:10.1146/annurev.genet.42.110807.091653.

Rohou, A., and Grigorieff, N. (2015). CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* doi:10.1016/j.jsb.2015.08.008.

Rothberg, J. M., and Leamon, J. H. (2008). The development and impact of 454 sequencing. *Nat. Biotechnol.* doi:10.1038/nbt1485.

Rozhdestvensky, T. S., Tang, T. H., Tchirkova, I. V., Brosius, J., Bachellerie, J. P., and Hüttenhofer, A. (2003). Binding of L7Ae protein to the K-turn of archaeal snoRNAs: A shared RNA binding motif for C/D and H/ACA box snoRNAs in Archaea. *Nucleic Acids Res.* doi:10.1093/nar/gkg175.

Ruparel, H., Bi, L., Li, Z., Bai, X., Kim, D. H., Turro, N. J., et al. (2005). Design and synthesis of a 3′-O-allyl photocleavable fluorescent nucleotide as a reversible terminator for DNA sequencing by synthesis. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.0501962102.

Rusk, N. (2011). Torrents of sequence. *Nat. Methods*. doi:10.1038/nmeth.f.330.

Russell, A. G., Ebhardt, H., and Dennis, P. P. (1999). Substrate requirements for a novel archaeal endonuclease that cleaves within the 5' external transcribed spacer of Sulfolobus acidocaldarius precursor rRNA. *Genetics* 152, 1373–85. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10430568.

Saliba, A.-E., C Santos, S., and Vogel, J. (2017). New RNA-seq approaches for the study of bacterial pathogens. *Curr. Opin. Microbiol.* 35, 78–87. doi:10.1016/j.mib.2017.01.001.

Salzberg, S. L., and Yorke, J. A. (2005). Beware of mis-assembled genomes. *Bioinformatics*. doi:10.1093/bioinformatics/bti769.

Sameach, H., Narunsky, A., Azoulay-Ginsburg, S., Gevorkyan-Aiapetov, L., Zehavi, Y., Moskovitz, Y., et al. (2017). Structural and Dynamics Characterization of the MerR Family Metalloregulator CueR in its Repression and Activation States. *Structure*. doi:10.1016/j.str.2017.05.004.

Sanders, T. J., Lammers, M., Marshall, C. J., Walker, J. E., Lynch, E. R., and Santangelo, T. J. (2019). TFS and Spt4/5 accelerate transcription through archaeal histone-based chromatin. *Mol. Microbiol.* doi:10.1111/mmi.14191.

Sanders, T. J., Wenck, B. R., Selan, J. N., Barker, M. P., Trimmer, S. A., Walker, J. E., et al. (2020). FttA is a CPSF73 homologue that terminates transcription in Archaea. *Nat. Microbiol.* 5, 545–553. doi:10.1038/s41564-020-0667-3.

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., et al. (1977a). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265, 687–695. doi:10.1038/265687a0.

Sanger, F., and Coulson, A. R. (1978). The use of thin acrylamide gels for DNA sequencing. *FEBS Lett.* doi:10.1016/0014-5793(78)80145-8.

Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F., and Petersen, G. B. (1982). Nucleotide sequence of bacteriophage λ DNA. *J. Mol. Biol.* doi:10.1016/0022-2836(82)90546-0.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977b). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.74.12.5463.

Santangelo, T. J., Čuboňová, L., and Reeve, J. N. (2010). Thermococcus kodakarensis Genetics: Tk1827-encoded β-glycosidase, new positive-selection protocol, and targeted and repetitive deletion technology. *Appl. Environ. Microbiol.* doi:10.1128/AEM.02497-09.

Santangelo, T. J., Cubonová, L., Skinner, K. M., Reeve, J. N., Cubonová, L., Skinner, K. M., et al. (2009). Archaeal Intrinsic Transcription Termination In Vivo. *J. Bacteriol.* 191, 7102–7108. doi:10.1128/JB.00982-09.

Santangelo, T. J., and Reeve, J. N. (2006). Archaeal RNA polymerase is sensitive to intrinsic termination directed by transcribed and remote sequences. *J. Mol. Biol.* 355, 196–210. doi:10.1016/j.jmb.2005.10.062.

Sapienza, C., Rose, M. R., and Doolittle, W. F. (1982). High-frequency genomic rearrangements involving archaebacterial repeat sequence elements. *Nature*. doi:10.1038/299182a0.

Sapra, R., Bagramyan, K., and Adams, M. W. W. (2003). A simple energy-conserving system: Proton reduction coupled to proton translocation. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.1331436100.

Sas-Chen, A., and Schwartz, S. (2019). Misincorporation signatures for detecting modifications in mRNA: Not as simple as it sounds. *Methods* 156, 53–59. doi:10.1016/j.ymeth.2018.10.011.

Sas-Chen, A., Thomas, J. M., Matzov, D., Taoka, M., Nance, K. D., Nir, R., et al. (2020). Dynamic RNA acetylation revealed by quantitative cross-evolutionary mapping. *Nature*. doi:10.1038/s41586-020-2418-2.

Sazinsky, M. H., LeMoine, B., Orofino, M., Davydov, R., Bencze, K. Z., Stemmler, T. L., et al. (2007). Characterization and structure of a Zn2+ and [2Fe-2S]-containing copper chaperone from Archaeoglobus fulgidus. *J. Biol. Chem.* doi:10.1074/jbc.M703311200.

Schäfer, T., and Schönheit, P. (1992). Maltose fermentation to acetate, CO2and H2in the anaerobic hyperthermophilic archaeon Pyrococcus furiosus: evidence for the operation of a novel sugar fermentation pathway. *Arch. Microbiol.* doi:10.1007/BF00290815.

Schirmer, M., Ijaz, U. Z., D'Amore, R., Hall, N., Sloan, W. T., and Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.* doi:10.1093/nar/gku1341.

Schleper, C., and Sousa, F. L. (2020). Meet the relatives of our cellular ancestor. *Nature*. doi:10.1038/d41586-020-00039-y.

Schloss, J. A. (2008). How to get genomes at one ten-thousandth the cost. *Nat. Biotechnol.* doi:10.1038/nbt1008-1113.

Scholz, M., Ward, D. V., Pasolli, E., Tolio, T., Zolfo, M., Asnicar, F., et al. (2016). Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat. Methods*. doi:10.1038/nmeth.3802.

Schut, G. J., Boyd, E. S., Peters, J. W., and Adams, M. W. W. (2013). The modular respiratory complexes involved in hydrogen and sulfur metabolism by heterotrophic hyperthermophilic archaea and their evolutionary implications. *FEMS Microbiol. Rev.* doi:10.1111/j.1574-6976.2012.00346.x.

Schwartz, S., and Motorin, Y. (2017). Next-generation sequencing technologies for detection of modified nucleotides in RNAs. *RNA Biol.* 14, 1124–1137. doi:10.1080/15476286.2016.1251543.

Seila, A. C., Calabrese, J. M., Levine, S. S., Yeo, G. W., Rahl, P. B., Flynn, R. A., et al. (2008). Divergent transcription from active promoters. *Science (80-.).* doi:10.1126/science.1162253.

Seitz, K. W., Dombrowski, N., Eme, L., Spang, A., Lombard, J., Sieber, J. R., et al. (2019). Asgard archaea capable of anaerobic hydrocarbon cycling. *Nat. Commun.* doi:10.1038/s41467-019-09364-x.

Seitz, K. W., Lazar, C. S., Hinrichs, K. U., Teske, A. P., and Baker, B. J. (2016). Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. *ISME J.* doi:10.1038/ismej.2015.233.

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

Sela, I., Wolf, Y. I., and Koonin, E. V. (2016). Theory of prokaryotic genome evolution. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.1614083113.

Seo, T. S., Bai, X., Kim, D. H., Meng, Q., Shi, S., Ruparel, H., et al. (2005). Four-color DNA sequencing by synthesis on a chip using photocleavable flourescent nucleotides. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.0501965102.

Shajani, Z., Sykes, M. T., and Williamson, J. R. (2011). Assembly of Bacterial Ribosomes. *Annu. Rev. Biochem.* 80, 501–526. doi:10.1146/annurev-biochem-062608-160432.

Sharma, C. M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiß, S., Sittka, A., et al. (2010). The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature*. doi:10.1038/nature08756.

Sharma, C. M., and Vogel, J. (2014). Differential RNA-seq: The approach behind and the biological insight gained. *Curr. Opin. Microbiol.* 19, 97–105. doi:10.1016/j.mib.2014.06.010.

Sharma, S., Langhendries, J.-L., Watzinger, P., Kötter, P., Entian, K.-D., and Lafontaine, D. L. J. (2015). Yeast Kre33 and human NAT10 are conserved 18S rRNA cytosine acetyltransferases that modify tRNAs assisted by the adaptor Tan1/THUMPD1. *Nucleic Acids Res.* 43, 2242–2258. doi:10.1093/nar/gkv075.

Sharma, S., Yang, J., van Nues, R., Watzinger, P., Kötter, P., Lafontaine, D. L. J., et al. (2017). Specialized box C/D snoRNPs act as antisense guides to target RNA base acetylation. *PLOS Genet.* 13, e1006804. doi:10.1371/journal.pgen.1006804.

Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., et al. (2017). DNA sequencing at 40: Past, present and future. *Nature* 550. doi:10.1038/nature24286.

Shultzaberger, R. K., Chen, Z., Lewis, K. A., and Schneider, T. D. (2007). Anatomy of Escherichia coli σ70 promoters. *Nucleic Acids Res.* doi:10.1093/nar/gkl956.

Siezen, R. J., Wilson, G., and Todt, T. (2010). Prokaryotic whole-transcriptome analysis: deep sequencing and tiling arrays. *Microb. Biotechnol.* 3, 125–130. doi:10.1111/j.1751-7915.2010.00166.x.

Siguier, P., Gourbeyre, E., and Chandler, M. (2014). Bacterial insertion sequences: Their genomic impact and diversity. *FEMS Microbiol. Rev.* 38, 865–891. doi:10.1111/1574-6976.12067.

Simpson, J. T., Workman, R. E., Zuzarte, P. C., David, M., Dursi, L. J., and Timp, W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* 14, 407–410. doi:10.1038/nmeth.4184.

Singh, S. S., Singh, N., Bonocora, R. P., Fitzgerald, D. M., Wade, J. T., and Grainger, D. C. (2014). Widespread suppression of intragenic transcription initiation by H-NS. *Genes Dev.* doi:10.1101/gad.234336.113.

Sitsel, O., Grønberg, C., Autzen, H. E., Wang, K., Meloni, G., Nissen, P., et al. (2015). Structure and Function of Cu(I)- and Zn(II)-ATPases. *Biochemistry* 54, 5673–5683. doi:10.1021/acs.biochem.5b00512.

Sleiman, S., and Dragon, F. (2019). Recent Advances on the Structure and Function of RNA Acetyltransferase Kre33/NAT10. *Cells* 8, 1035. doi:10.3390/cells8091035.

Sloan, K. E., Warda, A. S., Sharma, S., Entian, K. D., Lafontaine, D. L. J., and Bohnsack, M. T. (2017). Tuning the ribosome: The influence of rRNA modification on eukaryotic ribosome biogenesis and function. *RNA Biol.* 14, 1138–1152. doi:10.1080/15476286.2016.1259781.

Smith, A. M., Jain, M., Mulroney, L., Garalde, D. R., and Akeson, M. (2019). Reading canonical and modified nucleobases in 16S ribosomal RNA using nanopore native RNA sequencing. *PLoS One* 14, e0216709. doi:10.1371/journal.pone.0216709.

Smith, B. A., Gupta, N., Denny, K., and Culver, G. M. (2018). Characterization of 16S rRNA Processing with Pre-30S Subunit Assembly Intermediates from E. coli. *J. Mol. Biol.* 430, 1745–1759. doi:10.1016/j.jmb.2018.04.009.

Smith, J. D., and Dunn, D. B. (1959). The occurrence of methylated guanines in ribonucleic acids from several sources. *Biochem. J.* 72, 294–301. doi:10.1042/bj0720294.

Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., et al. (1986). Fluorescence detection in automated DNA sequence analysis. *Nature*. doi:10.1038/321674a0.

Smollett, K., Blombach, F., Fouqueau, T., and Werner, F. (2017a). "A Global Characterisation of the Archaeal Transcription Machinery," in *RNA Metabolism and Gene Expression in Archaea. Nucleic Acids and Molecular Biology*, 1–26. doi:10.1007/978-3-319-65795-0_1.

Smollett, K., Blombach, F., Reichelt, R., Thomm, M., and Werner, F. (2017b). A global analysis of transcription reveals two modes of Spt4/5 recruitment to archaeal RNA polymerase. *Nat. Microbiol.* 2, 17021. doi:10.1038/nmicrobiol.2017.21.

Soneson, C., Yao, Y., Bratus-neuenschwander, A., Patrignani, A., Robinson, M. D., and Hussain, S. (2019). A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. *Nat. Commun.* 10, 1–14. doi:10.1038/s41467-019-11272-z.

Song, L., Hobaugh, M. R., Shustak, C., Cheley, S., Bayley, H., and Gouaux, J. E. (1996). Structure of staphylococcal α-hemolysin, a heptameric transmembrane pore. *Science (80-.).* doi:10.1126/science.274.5294.1859.

Soppa, J. (1999). Transcription initiation in Archaea: facts, factors and future aspects. *Mol. Microbiol.* doi:10.1046/j.1365-2958.1999.01273.x.

Sorek, R., and Cossart, P. (2010). Prokaryotic transcriptomics: A new view on regulation, physiology and pathogenicity. *Nat. Rev. Genet.* 11, 9–16. doi:10.1038/nrg2695.

Soto, D. F., Recalde, A., Orell, A., Albers, S.-V. V., Paradela, A., Navarro, C. A., et al. (2019). Global effect of the lack of inorganic polyphosphate in the extremophilic archaeon Sulfolobus solfataricus: A proteomic approach. *J. Proteomics*

191, 143–152. doi:10.1016/j.jprot.2018.02.024.

Spang, A., Saw, J. H., Jørgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., et al. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* 521, 173–179. doi:10.1038/nature14447.

Spitalny, P., and Thomm, M. (2003). Analysis of the open region and of DNA-protein contacts of archaeal RNA polymerase transcription complexes during transition from initiation to elongation. *J. Biol. Chem.* 278, 30497–30505. doi:10.1074/jbc.M303633200.

Spitalny, P., and Thomm, M. (2008). A polymerase III-like reinitiation mechanism is operating in regulation of histone expression in archaea. *Mol. Microbiol.* 67, 958–970. doi:10.1111/j.1365-2958.2007.06084.x.

Staden, R. (1979). A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res.* doi:10.1093/nar/6.7.2601.

Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years. *Nat. Rev. Genet.* doi:10.1038/s41576-019-0150-2.

Stetter, K. O. (2006). Hyperthermophiles in the history of life. in *Philosophical Transactions of the Royal Society B: Biological Sciences* doi:10.1098/rstb.2006.1907.

Stetter, K. O., König, H., and Stackebrandt, E. (1983). Pyrodictium gen. nov., a New Genus of Submarine Disc-Shaped Sulphur Reducing Archaebacteria Growing Optimally at 105°C. *Syst. Appl. Microbiol.* 4, 535–551. doi:10.1016/S0723-2020(83)80011-3.

Stevenson-Jones, F., Woodgate, J., Castro-Roa, D., and Zenkin, N. (2020). Ribosome reactivates transcription by physically pushing RNA polymerase out of transcription arrest. *Proc. Natl. Acad. Sci.* 117, 8462–8467. doi:10.1073/pnas.1919985117.

Stock, A. M., Robinson, V. L., and Goudreau, P. N. (2000). Two-Component Signal Transduction. *Annu. Rev. Biochem.* doi:10.1146/annurev.biochem.69.1.183.

Stoecklin, G., and Mühlemann, O. (2013). RNA decay mechanisms: Specificity through diversity. *Biochim. Biophys. Acta - Gene Regul. Mech.* doi:10.1016/j.bbagrm.2013.04.002.

Stoiber, M. H., Quick, J., Egan, R., Lee, J. E., Celniker, S. E., Neely, R., et al. (2016). De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. *bioRxiv*, 094672. doi:10.1101/094672.

Strand, K. R., Sun, C., Li, T., Jenney, F. E., Schut, G. J., and Adams, M. W. W. (2010). Oxidative stress protection and the repair response to hydrogen peroxide in the hyperthermophilic archaeon Pyrococcus furiosus and in related species. *Arch. Microbiol.* 192, 447–459. doi:10.1007/s00203-010-0570-z.

Straub, J., Brenneis, M., Jellen-Ritter, A., Heyer, R., Soppa, J., and Marchfelder, A. (2009). Small RNAs in haloarchaea: Identification, differential expression and biological function. *RNA Biol.* doi:10.4161/rna.6.3.8357.

Strunk, B. S., Loucks, C. R., Su, M., Vashisth, H., Cheng, S., Schilling, J., et al. (2011). Ribosome Assembly Factors Prevent Premature Translation Initiation by 40S Assembly Intermediates. *Science (80-.).* 333, 1449–1453. doi:10.1126/science.1208245.

Su, A. A. H., Tripp, V., and Randau, L. (2013). RNA-Seq analyses reveal the order of tRNA processing events and the maturation of C/D box and CRISPR RNAs in the hyperthermophile Methanopyrus kandleri. *Nucleic Acids Res.* 41, 6250–6258. doi:10.1093/nar/gkt317.

Sugimoto, N., Nakano, S. ichi, Katoh, M., Matsumura, A., Nakamuta, H., Ohmichi, T., et al. (1995). Thermodynamic Parameters To Predict Stability of RNA/DNA Hybrid Duplexes. *Biochemistry.* doi:10.1021/bi00035a029.

Suter, D. M. (2020). Transcription Factors and DNA Play Hide and Seek. *Trends Cell Biol.* doi:10.1016/j.tcb.2020.03.003.

Taboada, B., Estrada, K., Ciria, R., and Merino, E. (2018). Operon-mapper: a web server for precise operon identification in bacterial and archaeal genomes. *Bioinformatics.* doi:10.1093/bioinformatics/bty496.

Taiaroa, G., Rawlinson, D., Featherstone, L., Pitt, M., Caly, L., Druce, J., et al. (2020). Direct RNA sequencing and early evolution of SARS-CoV-2. *bioRxiv*, 2020.03.05.976167. doi:10.1101/2020.03.05.976167.

Takai, K., Sugai, A., Itoh, T., and Horikoshi, K. (2000). Palaeococcus ferrophilus gen. nov., sp. nov., a barophilic, hyperthermophilic archaeon from a deep-sea hydrothermal vent chimney. *Int. J. Syst. Evol. Microbiol.* doi:10.1099/00207713-50-2-489.

Tan, G., Yang, J., Li, T., Zhao, J., Sun, S., Li, X., et al. (2017). Anaerobic copper toxicity and iron-sulfur cluster biogenesis in Escherichia coli. *Appl. Environ. Microbiol.* doi:10.1128/AEM.00867-17.

Tang, T.-H., Bachellerie, J.-P., Rozhdestvensky, T., Bortolin, M.-L., Huber, H., Drungowski, M., et al. (2002). Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus. *Proc. Natl. Acad. Sci.* doi:10.1073/pnas.112047299.

Tang, T. H., Polacek, N., Zywicki, M., Huber, H., Brugger, K., Garrett, R., et al. (2005). Identification of novel non-coding RNAs as potential antisense regulators in the archaeon Sulfolobus solfataricus. *Mol. Microbiol.* doi:10.1111/j.1365-2958.2004.04428.x.

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., et al. (2016). NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* 44, 6614–24. doi:10.1093/nar/gkw569.

Thomason, M. K., Bischler, T., Eisenbart, S. K., Förstner, K. U., Zhang, A., Herbig, A., et al. (2015). Global Transcriptional Start Site Mapping Using Differential RNA Sequencing Reveals Novel Antisense RNAs in Escherichia coli. *J. Bacteriol.* 197, 18–28. doi:10.1128/JB.02096-14.

Thomm, M., Hausner, W., and Hethke, C. (1994). Transcription Factors and Termination of Transcriptionin Methanococcus. *Syst. Appl. Microbiol. = Syst. Appl. Microbiol.* Available at: http://epub.uni-

regensburg.de/11002/1/ubr04680_ocr.pdf [Accessed January 20, 2016].

Thompson, J. F., and Steinmann, K. E. (2010). Single molecule sequencing with a HeliScope genetic analysis system. *Curr. Protoc. Mol. Biol.* doi:10.1002/0471142727.mb0710s92.

Thorgersen, M. P., Stirrett, K., Scott, R. A., and Adams, M. W. W. (2012). Mechanism of oxygen detoxification by the surprisingly oxygen-tolerant hyperthermophilic archaeon, Pyrococcus furiosus. *Proc. Natl. Acad. Sci.* 109, 18547–18552. doi:10.1073/pnas.1208605109.

Tian, B., and Manley, J. L. (2016). Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.* doi:10.1038/nrm.2016.116.

Tilgner, H., Jahanbani, F., Blauwkamp, T., Moshrefi, A., Jaeger, E., Chen, F., et al. (2015). Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat. Biotechnol.* 33, 736–742. doi:10.1038/nbt.3242.

Toffano-Nioche, C., Ott, A., Crozat, E., Nguyen, A. N., Zytnicki, M., Leclerc, F., et al. (2013). RNA at 92 °C: The non-coding transcriptome of the hyperthermophilic archaeon Pyrococcus abyssi. *RNA Biol.* 10, 1211–1220. doi:10.4161/rna.25567.

Tollerson, R., and Ibba, M. (2020). Translational regulation of environmental adaptation in bacteria. *J. Biol. Chem.* 295, 10434–10445. doi:10.1074/jbc.REV120.012742.

Tombácz, D., Moldován, N., Balázs, Z., Gulyás, G., Csabai, Z., Boldogkői, M., et al. (2019). Multiple Long-Read Sequencing Survey of Herpes Simplex Virus Dynamic Transcriptome. *Front. Genet.* 10. doi:10.3389/fgene.2019.00834.

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

Tran, T. T., Dam, P., Su, Z., Poole, F. L., Adams, M. W. W., Zhou, G. T., et al. (2007). Operon prediction in Pyrococcus furiosus. *Nucleic Acids Res.* doi:10.1093/nar/gkl974.

Tringe, S. G., and Rubin, E. M. (2005). Metagenomics: DNA sequencing of environmental samples. *Nat. Rev. Genet. 6* 6, 805–14. doi:10.1038/nrg1709.

Tsuda, T., and Toyoshima, C. (2009). Nucleotide recognition by CopA, a Cu -transporting P-type ATPase. *EMBO J.* doi:10.1038/emboj.2009.143.

Tyler, A. D., Mataseje, L., Urfano, C. J., Schmidt, L., Antonation, K. S., Mulvey, M. R., et al. (2018). Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. *Sci. Rep.* doi:10.1038/s41598-018-29334-5.

Unniraman, S. (2002). Conserved economics of transcription termination in eubacteria. *Nucleic Acids Res.* doi:10.1093/nar/30.3.675.

Urbieta, M. S., Donati, E. R., Chan, K. G., Shahar, S., Sin, L. L., and Goh, K. M. (2015). Thermophiles in the genomic era: Biodiversity, science, and applications. *Biotechnol. Adv.* doi:10.1016/j.biotechadv.2015.04.007.

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., et al. (2008). A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res.* doi:10.1101/gr.076463.108.

Van De Werken, H. J. G., Verhees, C. H., Akerboom, J., De Vos, W. M., and Van Der Oost, J. (2006). Identification of a glycolytic regulon in the archaea Pyrococcus and Thermococcus. *FEMS Microbiol. Lett.* 260, 69–76. doi:10.1111/j.1574-6968.2006.00292.x.

van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends Genet.* 30, 418–426. doi:10.1016/j.tig.2014.07.001.

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018). The Third Revolution in Sequencing Technology. *Trends Genet.* 34, 666–681. doi:10.1016/j.tig.2018.05.008.

Venema, J., and Tollervey, D. (1995). Processing of pre-ribosomal RNA in Saccharomyces cerevisiae. *Yeast* 11, 1629–1650. doi:10.1002/yea.320111607.

Verma, M., Kulshrestha, S., and Puri, A. (2017). "Genome Sequencing," in *Bioinformatics: Volume I: Data, Sequence Analysis, and Evolution*, ed. J. M. Keith (New York, NY: Springer New York), 3–33. doi:10.1007/978-1-4939-6622-6_1.

Viehweger, A., Krautwurst, S., Lamkiewicz, K., Madhugiri, R., Ziebuhr, J., Hölzer, M., et al. (2019). Direct RNA nanopore sequencing of full-length coron-avirus genomes provides novel insights into structural variants and enables modification analysis. *Genome Res.*, 483693. doi:10.1101/483693.

Vierke, G., Engelmann, A., Hebbeln, C., and Thomm, M. (2003). A novel archaeal transcriptional regulator of heat shock response. *J. Biol. Chem.* 278, 18–26. doi:10.1074/jbc.M209250200.

Vilfan, I. D., Tsai, Y.-C., Clark, T. A., Wegener, J., Dai, Q., Yi, C., et al. (2013). Analysis of RNA base modification and structural rearrangement by single-molecule real-time detection of reverse transcription. *J. Nanobiotechnology* 11, 8. doi:10.1186/1477-3155-11-8.

Villafane, A., Voskoboynik, Y., Cuebas, M., Ruhl, I., and Bini, E. (2009). Response to excess copper in the hyperthermophile Sulfolobus solfataricus strain 98/2. *Biochem. Biophys. Res. Commun.*, 67–71. doi:10.1126/scisignal.2001449.Engineering.

Villafane, A., Voskoboynik, Y., Ruhl, I., Sannino, D., Maezato, Y., Blum, P., et al. (2011). CopR of Sulfolobus solfataricus represents a novel class of archaeal-specific copper-responsive activators of transcription. *Microbiology* 157, 2808–2817. doi:10.1099/mic.0.051862-0.

Vo Ngoc, L., Kassavetis, G. A., and Kadonaga, J. T. (2019). The RNA Polymerase II Core Promoter in Drosophila. *Genetics* 212, 13–24. doi:10.1534/genetics.119.302021.

Vo Ngoc, L., Wang, Y. L., Kassavetis, G. A., and Kadonaga, J. T. (2017). The punctilious RNA polymerase II core promoter. *Genes Dev.* doi:10.1101/gad.303149.117.

Vogel, U., and Jensen, K. F. (1994). The RNA chain elongation rate in Escherichia coli depends on the growth rate. *J. Bacteriol.* 176, 2807–2813. doi:10.1128/JB.176.10.2807-2813.1994.

Wade, J. T., Roa, D. C., Grainger, D. C., Hurd, D., Busby, S. J. W., Struhl, K., et al. (2006). Extensive functional overlap between σ factors in Escherichia coli. *Nat. Struct. Mol. Biol.* doi:10.1038/nsmb1130.

Waege, I., Schmid, G., Thumann, S., Thomm, M., and Hausner, W. (2010). Shuttle Vector-Based Transformation System for Pyrococcus furiosus. *Appl. Environ. Microbiol.* 76, 3308–3313. doi:10.1128/AEM.01951-09.

Wagih, O. (2017). Ggseqlogo: A versatile R package for drawing sequence logos. *Bioinformatics*. doi:10.1093/bioinformatics/btx469.

Walker, J. E., Luyties, O., and Santangelo, T. J. (2017). Factor-dependent archaeal transcription termination. *Proc. Natl. Acad. Sci. U. S. A.* 114, E6767–E6773. doi:10.1073/pnas.1704028114.

Wang, C., Molodtsov, V., Firlar, E., Kaelber, J. T., Blaha, G., Su, M., et al. (2020). Structural basis of transcription-translation coupling. *Science (80-.).*, eabb5317. doi:10.1126/science.abb5317.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi:10.1038/nrg2484.

Warman, E., Forrest, D., Wade, J. T., and Grainger, D. C. (2020). Widespread divergent transcription from prokaryotic promoters. *bioRxiv*, 2020.01.31.928960. doi:10.1101/2020.01.31.928960.

Washio, T., Sasayama, J., and Tomita, M. (1998). Analysis of complete genomes suggests that many prokaryotes do not rely on hairpin formation in transcription termination. *Nucleic Acids Res.* doi:10.1093/nar/26.23.5456.

Waters, E., Hohn, M. J., Ahel, I., Graham, D. E., Adams, M. D., Barnstead, M., et al. (2003). The genome of Nanoarchaeum equitans: Insights into early archaeal evolution and derived parasitism. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.1735403100.

Webster, M. W., Takacs, M., Zhu, C., Vidmar, V., Eduljee, A., Abdelkareem, M., et al. (2020). Structural basis of transcription-translation coupling and collision in bacteria. *Science (80-.).*, eabb5036. doi:10.1126/science.abb5036.

Weirather, J. L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J. J., et al. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* 6, 100. doi:10.12688/f1000research.10571.2.

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. C., Hall, R. J., Concepcion, G. T., et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* doi:10.1038/s41587-019-0217-9.

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

Wheaton, G. H., Mukherjee, A., and Kelly, R. M. (2016). Transcriptomes of the extremely thermoacidophilic archaeon Metallosphaera Sedula exposed to metal “shock” reveal generic and specific metal responses. *Appl. Environ. Microbiol.* doi:10.1128/AEM.01176-16.

White, J. R., Escobar-Paramo, P., Mongodin, E. F., Nelson, K. E., and DiRuggiero, J. (2008). Extensive genome rearrangements and multiple horizontal gene transfers in a population of Pyrococcus isolates from Vulcano Island, Italy. *Appl. Environ. Microbiol.* doi:10.1128/AEM.01024-08.

Wick, R. R., Judd, L. M., and Holt, K. E. (2017). Comparison Of Oxford Nanopore Basecalling Tools. *Doi.Org.* doi:10.5281/ZENODO.1043612.

Wick, R. R., Judd, L. M., and Holt, K. E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* 20, 129. doi:10.1186/s13059-019-1727-y.

Wickham, H. (2016). ggplot2: Elegant Graphic for Data Analysis. *Springer*. doi:10.1007/978-0-387-98141-3.

Wierer, S., Daldrop, P., Din Ahmad, M. U., Boos, W., Drescher, M., Welte, W., et al. (2016). TrmBL2 from Pyrococcus furiosus interacts both with double-stranded and single-stranded DNA. *PLoS One*. doi:10.1371/journal.pone.0156098.

Wilbanks, E. G., Larsen, D. J., Neches, R. Y., Yao, A. I., Wu, C.-Y., Kjolby, R. A. S., et al. (2012). A workflow for genome-wide mapping of archaeal transcription factors with ChIP-seq. *Nucleic Acids Res.* 40, e74–e74. doi:10.1093/nar/gks063.

Williams, T. A., Cox, C. J., Foster, P. G., Szöllősi, G. J., and Embley, T. M. (2020). Phylogenomics provides robust support for a two-domains tree of life. *Nat. Ecol. Evol.* doi:10.1038/s41559-019-1040-x.

Williams, T. A., Foster, P. G., Cox, C. J., and Embley, T. M. (2013). An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* 504, 231–6. doi:10.1038/nature12779.

Wirth, R., Luckner, M., and Wanner, G. (2018). Validation of a hypothesis: Colonization of black smokers by hyperthermophilic microorganisms. *Front. Microbiol.* doi:10.3389/fmicb.2018.00524.

Woese, C. R., and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. U. S. A.* 74, 5088–5090. doi:10.1073/pnas.74.11.5088.

Woese, C. R., Kandler, O., and Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U. S. A.* 87, 4576–4579. doi:10.1073/pnas.87.12.4576.

Wongsurawat, T., Jenjaroenpun, P., Taylor, M. K., Lee, J., Tolardo, A. L., Parvathareddy, J., et al. (2019). Rapid

Sequencing of Multiple RNA Viruses in Their Native Form. *Front. Microbiol.* 10, 1–8. doi:10.3389/fmicb.2019.00260.

Workman, R. E., Tang, A. D., Tang, P. S., Jain, M., Tyson, J. R., Razaghi, R., et al. (2019). Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat. Methods* 16, 1297–1305. doi:10.1038/s41592-019-0617-2.

Wu, X., and Sharp, P. A. (2013). XDivergent transcription: A driving force for new gene origination? *Cell*. doi:10.1016/j.cell.2013.10.048.

Wuichet, K., Cantwell, B. J., and Zhulin, I. B. (2010). Evolution and phyletic distribution of two-component signal transduction systems. *Curr. Opin. Microbiol.* doi:10.1016/j.mib.2009.12.011.

Wurtzel, O., Sapra, R., Chen, F., Zhu, Y., Simmons, B. A., and Sorek, R. (2010). A single-base resolution map of an archaeal transcriptome. *Genome Res.* 20, 133–41. doi:10.1101/gr.100396.109.

Xie, Z., and Tang, H. (2017). ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics*. doi:10.1093/bioinformatics/btx433.

Xu, Y., Lin, Z., Tang, C., Tang, Y., Cai, Y., Zhong, H., et al. (2019). A new massively parallel nanoball sequencing platform for whole exome research. *BMC Bioinformatics* 20, 153. doi:10.1186/s12859-019-2751-3.

Xu, Z., O'Farrell, H. C., Rife, J. P., and Culver, G. M. (2008). A conserved rRNA methyltransferase regulates ribosome biogenesis. *Nat. Struct. Mol. Biol.* 15, 534–536. doi:10.1038/nsmb.1408.

Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Münster, S., Camblong, J., et al. (2009). Bidirectional promoters generate pervasive transcription in yeast. *Nature* 457, 1033–7. doi:10.1038/nature07728.

Yamada, M., Ishijima, S. A., and Suzuki, M. (2009). Interactions between the archaeal transcription repressor FL11 and its coregulators lysine and arginine. *Proteins Struct. Funct. Bioinforma.* 74, 520–525. doi:10.1002/prot.22269.

Yan, B., Boitano, M., Clark, T. A., and Ettwiller, L. (2018). SMRT-Cappable-seq reveals complex operon variants in bacteria. *Nat. Commun.* 9, 3676. doi:10.1038/s41467-018-05997-6.

Yang, H., Lipscomb, G. L., Keese, A. M., Schut, G. J., Thomm, M., Adams, M. W. W., et al. (2010). SurR regulates hydrogen production in Pyrococcus furiosus by a sulfur-dependent redox switch. *Mol. Microbiol.* doi:10.1111/j.1365-2958.2010.07275.x.

Yip, W. S. V., Vincent, N. G., and Baserga, S. J. (2013). Ribonucleoproteins in Archaeal Pre-rRNA Processing and Modification. *Archaea* 2013, 1–14. doi:10.1155/2013/614735.

Yokoyama, K., Ishijima, S. A., Koike, H., Kurihara, C., Shimowasa, A., Kabasawa, M., et al. (2007). Feast/Famine Regulation by Transcription Factor FL11 for the Survival of the Hyperthermophilic Archaeon Pyrococcus OT3. *Structure*. doi:10.1016/j.str.2007.10.015.

Yoon, S. H., Reiss, D. J., Bare, J. C., Tenenbaum, D., Pan, M., Slagel, J., et al. (2011). Parallel evolution of transcriptome architecture during genome reorganization. *Genome Res.* 21, 1892–1904. doi:10.1101/gr.122218.111.

Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* 11, R14. doi:10.1186/gb-2010-11-2-r14.

Yu, L., Winkelman, J. T., Pukhrambam, C., Strick, T. R., Nickels, B. E., and Ebright, R. H. (2017). The mechanism of variability in transcription start site selection. *Elife*. doi:10.7554/eLife.32038.

Yu, S.-H. H., Vogel, J., and Förstner, K. U. (2018). ANNOgesic: A Swiss army knife for the RNA-Seq based annotation of bacterial/archaeal genomes. *Gigascience*, 1–11. doi:10.1101/143081.

Yue, L., Li, J., Zhang, B., Qi, L., Li, Z., Zhao, F., et al. (2020). The conserved ribonuclease aCPSF1 triggers genome-wide transcription termination of Archaea via a 3′-end cleavage mode. *Nucleic Acids Res.* doi:10.1093/nar/gkaa702.

Yue, L., Li, J., Zhang, B., Qi, L., Zhao, F., Li, L., et al. (2019). aCPSF1 controlled archaeal transcription termination: a prototypical eukaryotic model. *bioRxiv*, 843821. doi:10.1101/843821.

Zaremba-Niedzwiedzka, K., Caceres, E. F., Saw, J. H., Bäckström, Di., Juzokaite, L., Vancaester, E., et al. (2017). Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*. doi:10.1038/nature21031.

Zeldes, B. M., Keller, M. W., Loder, A. J., Straub, C. T., Adams, M. W. W., and Kelly, R. M. (2015). Extremely thermophilic microorganisms as metabolic engineering platforms for production of fuels and industrial chemicals. *Front. Microbiol.* doi:10.3389/fmicb.2015.01209.

Zhang, J. Z., Fang, Y., Hou, J. Y., Ren, H. J., Jiang, R., Roos, P., et al. (1995). Use of Non-Cross-Linked Polyacrylamide for Four-Color DNA Sequencing by Capillary Electrophoresis Separation of Fragments up to 640 Bases in Length in Two Hours. *Anal. Chem.* doi:10.1021/ac00120a026.

Zhao, B. S., Roundtree, I. A., and He, C. (2016). Post-transcriptional gene regulation by mRNA modifications. *Nat. Rev. Mol. Cell Biol.* doi:10.1038/nrm.2016.132.

Zhao, L., Zhang, H., Kohnen, M. V., Prasad, K. V. S. K., Gu, L., and Reddy, A. S. N. (2019). Analysis of Transcriptome and Epitranscriptome in Plants Using PacBio Iso-Seq and Nanopore-Based Direct RNA Sequencing. *Front. Genet.* 10. doi:10.3389/fgene.2019.00253.

Zhou, P., Yang, X. Lou, Wang, X. G., Hu, B., Zhang, L., Zhang, W., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. doi:10.1038/s41586-020-2012-7.

Zhu, A., Ibrahim, J. G., and Love, M. I. (2019a). Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics* 35, 2084–2092. doi:10.1093/bioinformatics/bty895.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., et al. (2020). A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* doi:10.1056/NEJMoa2001017.

Zhu, Q., Mai, U., Pfeiffer, W., Janssen, S., Asnicar, F., Sanders, J. G., et al. (2019b). Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.* 10. doi:10.1038/s41467-019-

13443-4.

Zhu, Y., Kumar, S., Menon, A. L., Scott, R. A., and Adams, M. W. W. W. (2013). Regulation of iron metabolism by pyrococcus furiosus. *J. Bacteriol.* 195, 2400–2407. doi:10.1128/JB.02280-12.

Zillig, W., Holz, I., Janekovic, D., Schäfer, W., and Reiter, W. D. (1983). The Archaebacterium Thermococcus celer Represents, a Novel Genus within the Thermophilic Branch of the Archaebacteria. *Syst. Appl. Microbiol.* doi:10.1016/S0723-2020(83)80036-8.

Zillig, W., Holz, I., Klenk, H. P., Trent, J., Wunderl, S., Janekovic, D., et al. (1987). Pyrococcus woesei, sp. nov., an ultra-thermophilic marine archaebacterium, representing a novel order, Thermococcales. *Syst. Appl. Microbiol.* doi:10.1016/S0723-2020(87)80057-7.

Zillig, W., Stetter, K. O., and Janeković, D. (1979). DNA-dependent RNA polymerase from the archaebacterium Sulfolobus acidocaldarius. *Eur. J. Biochem.* 96, 597–604. doi:10.1111/j.1432-1033.1979.tb13074.x.

Zivanov, J., Nakane, T., Forsberg, B. O., Kimanius, D., Hagen, W. J. H., Lindahl, E., et al. (2018). New tools for automated high-resolution cryo-EM structure determination in RELION-3. *Elife.* doi:10.7554/eLife.42166.

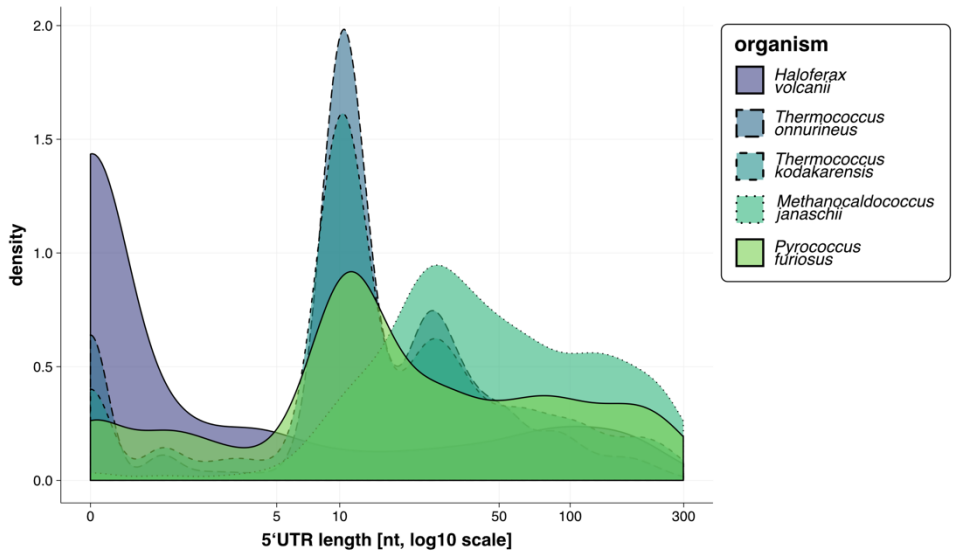
Zivanovic, Y., Lopez, P., Philippe, H., and Forterre, P. (2002). Pyrococcus genome comparison evidences chromosome shuffling-driven evolution. *Nucleic Acids Res.* doi:10.1093/nar/30.9.1902.

Zorraquino, V., Kim, M., Rai, N., and Tagkopoulos, I. (2017). The genetic and transcriptional basis of short and long term adaptation across multiple stresses in Escherichia coli. *Mol. Biol. Evol.* doi:10.1093/molbev/msw269.

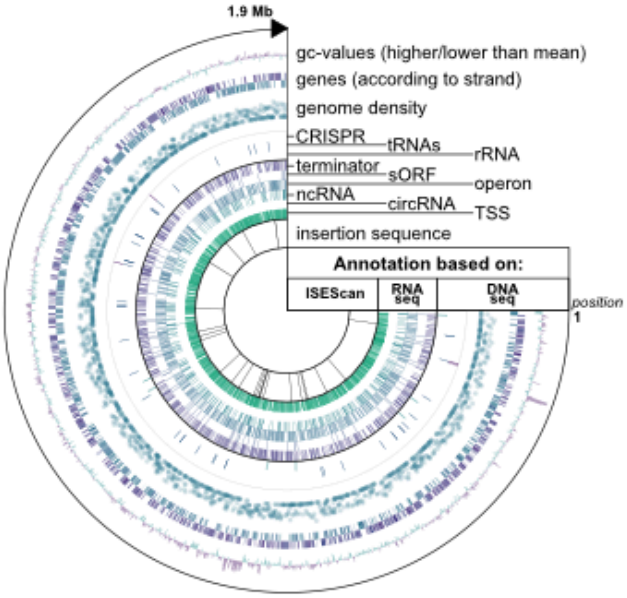
Appendix

The Appendix includes additional figures and tables that are referenced in the three manuscripts.

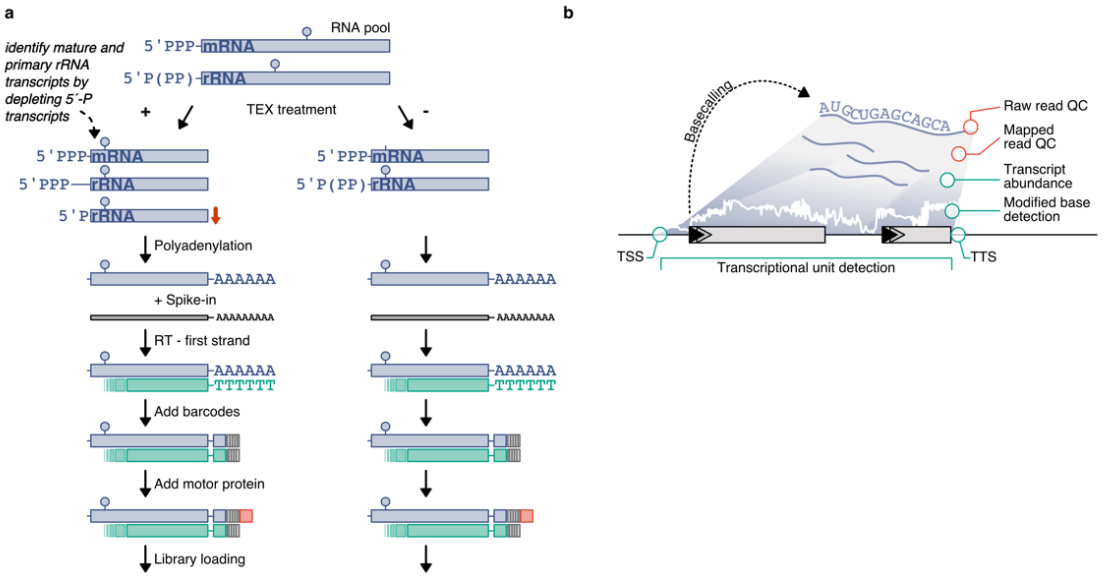
Supplementary Figures



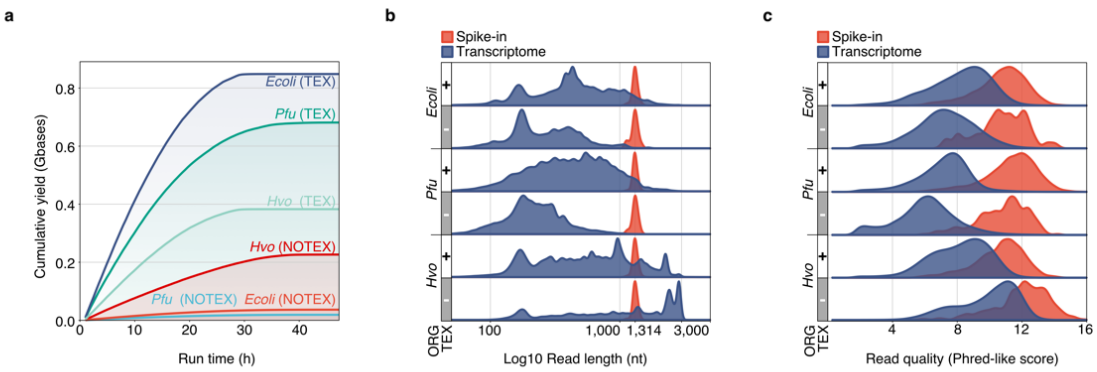
**Supplementary Figure 1 | Analysis of 5`UTRs in five archaeal model organisms.** The distribution of 5`UTR lengths is shown on a log10 scale between 1 and 300 nucleotides. Data for generating this plot were downloaded from the supplementary files of cited publications and re-analyzed (Jäger et al., 2014; Babski et al., 2016; Cho et al., 2017; Smollett et al., 2017b).



Supplementary Figure 2 | Summary of features added to new annotation of *Pyrococcus furiosus* DSM 3638 using ISelement scanning, DNA- and RNA-sequencing. Circular representation of whole genome showing from outside to inside 1) GC content higher/below average GC of 40.75%, 2) genes on positive/negative strand, 3) rainfall plot (location of events on x-axis, distance between consecutive events on y-axis), 4) terminators, 5) ncRNAs, 6) TSSs, 7) tRNAs, 8) SRP RNAs, 9) snoRNAs, 10) rRNAs, 11) CRISPR regions, 12) pseudogenes and 13) IS elements (Xie and Tang, 2017; Yu et al., 2018). Visualization was done using the circlize package in R (Gu et al., 2014).

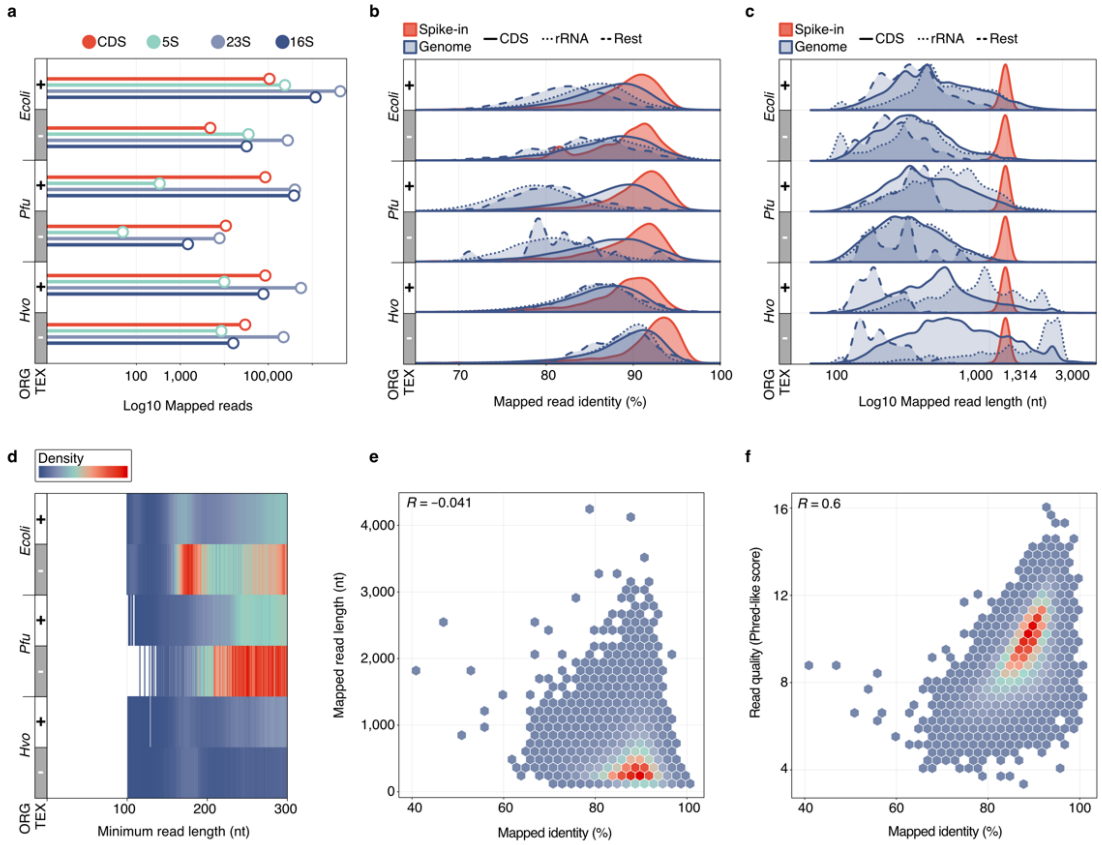


Supplementary Figure 3 | Extended workflow and objectives of Nanopore-based native RNA sequencing. a, Partial digestion of RNAs that are not 5´-triphosphorylated (e.g. tRNAs, rRNAs) is done using a Terminator 5´ -Phosphate-Dependent Exonuclease (TEX), allowing for the analysis of both mRNAs and rRNAs. After enzymatic poly(A)-addition, library preparation is performed according to a modified SQK-RNA001 protocol. The spike-in enolase is added and the RNA is reverse transcribed using the RTA adapter (green poly(T)). Depending on the necessary sequencing depth, custom barcodes (poreplex) can be added in the next step that replace the 3´-RMX adapter (lined grey square). A further 3´- ligation is performed to add the motor-protein carrying adapter (red square). Libraries are then loaded on a R9.4 flowcell and sequenced for 48 hours on a MinION device. **b**, After demultiplexing and basecalling of raw reads, quality control is performed on raw reads and mapped reads (red) to detect problems during library preparation or sequencing. Multiple transcriptomic features highlighted in green, including transcriptional start sites (TSS), termination sites (TTS), transcript abundance, modified bases and composition of transcriptional units can be adressed simultanously.

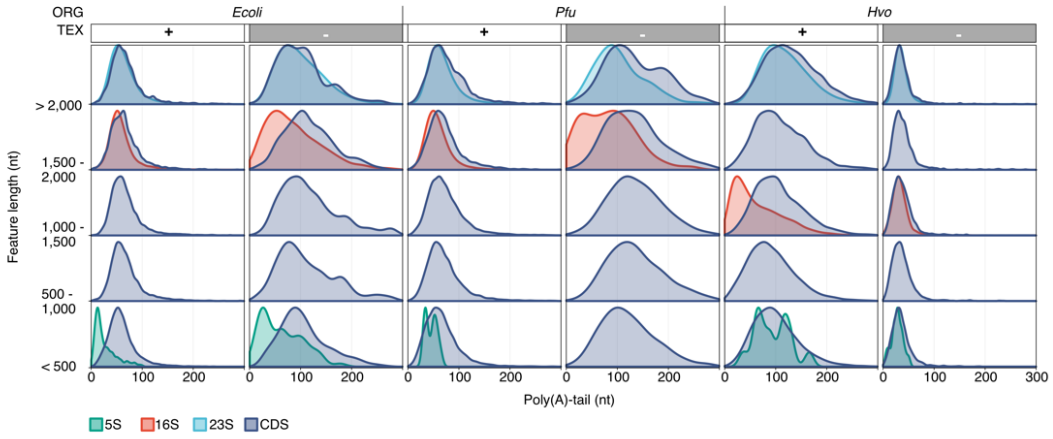


Supplementary Figure 4 | Raw read analysis. a, Libraries are loaded on R9.4 flow cells and reads are collected over 48 hours using the recommended MinKNOW script. FAST5 files are then demultiplexed (poreplex) and converted to FASTQ files during guppy basecalling. Differences in the plotted cumulative yield derive mostly from different multiplexing strategies. **b**, Raw read length

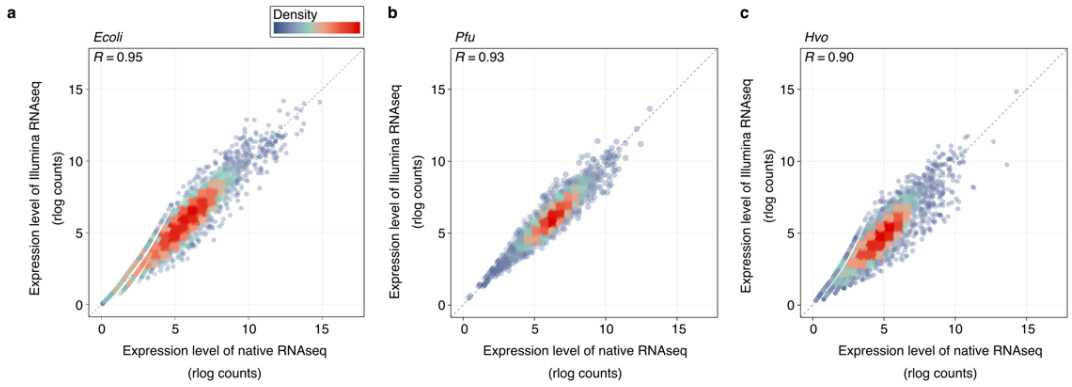
($\log_{10}$ scale) and **c**, raw read quality (Phred-like score estimated by guppy) is compared between samples for transcriptome-derived reads (purple) and the spike-in control (red).



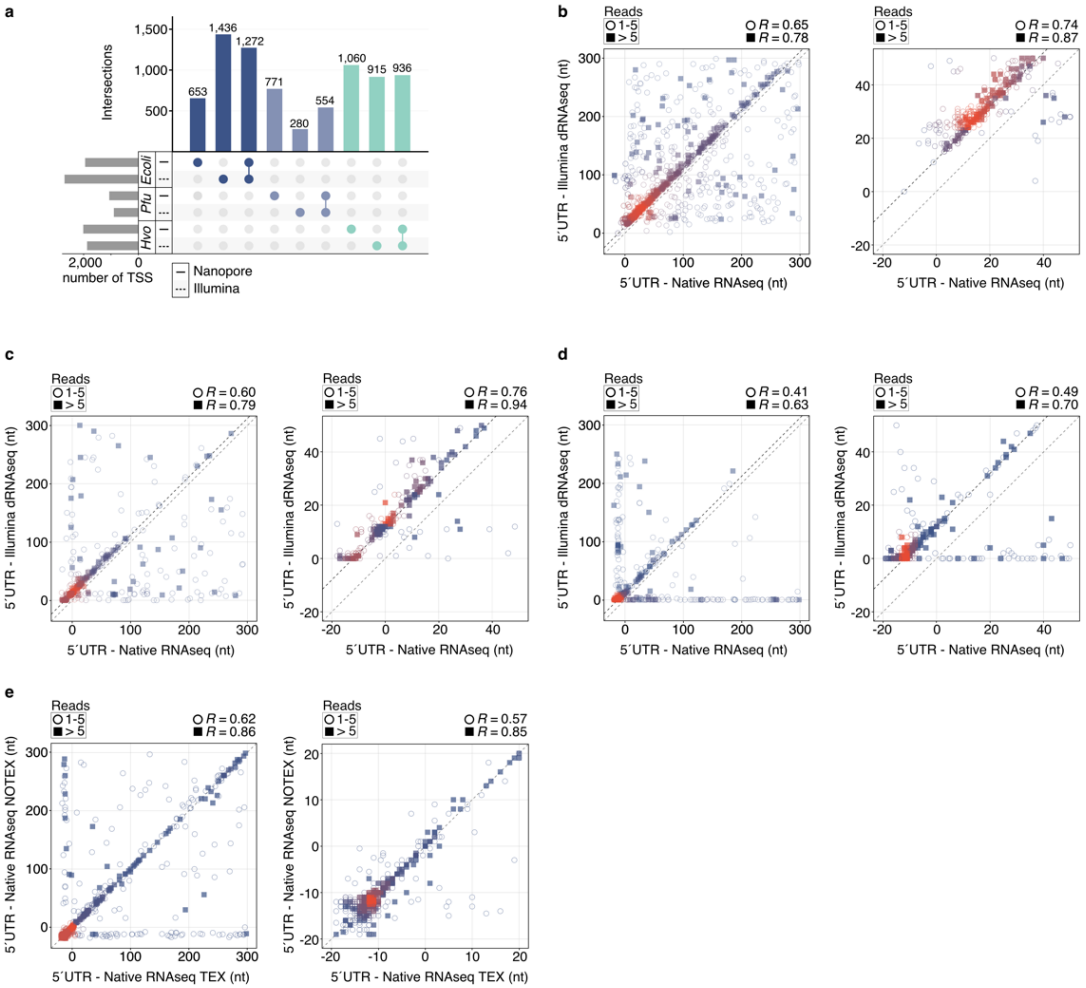
Supplementary Figure 5 | Mapped read analysis. a, Number of mapped reads (Log_{10} scale) color-coded for different features (protein coding genes (CDS): red, 5S rRNA: green, 16S rRNA: purple, 23S rRNA: light-purple, compare Figure 19b). **b**, Mapped read identity is defined as (1-NM/aligned_reads)*100, where NM is the edit distance reported taken from minimap2 (Li, 2018). Distribution of read identity (in %) and **c**, mapped read length (in Log_{10} scale) are shown for spike-in control (enolase yeast, red) and genomic features (purple). Different feature types are highlighted with solid line (CDS), dotted line (rRNA), long-dashed line (rest) and are based on the featurecounts classification (Liao et al., 2019). **d**, Limitations in the minimum read length that can be mapped (in nts) is shown for all datasets. Density of reads is indicated by a color-scale from darkblue (few reads) to red (many reads). **e**, Mapped identity (%) is compared to mapped read lengths (nt) for the same sample (Spearman´s rho: -0.041). **f**, Raw read quality correlates with calculated mapped identity (Spearman´s rho = 0.6). Density is color-coded as described before. TEX treated *P. furiosus* set (CDS mapping reads) is shown as a representative example.



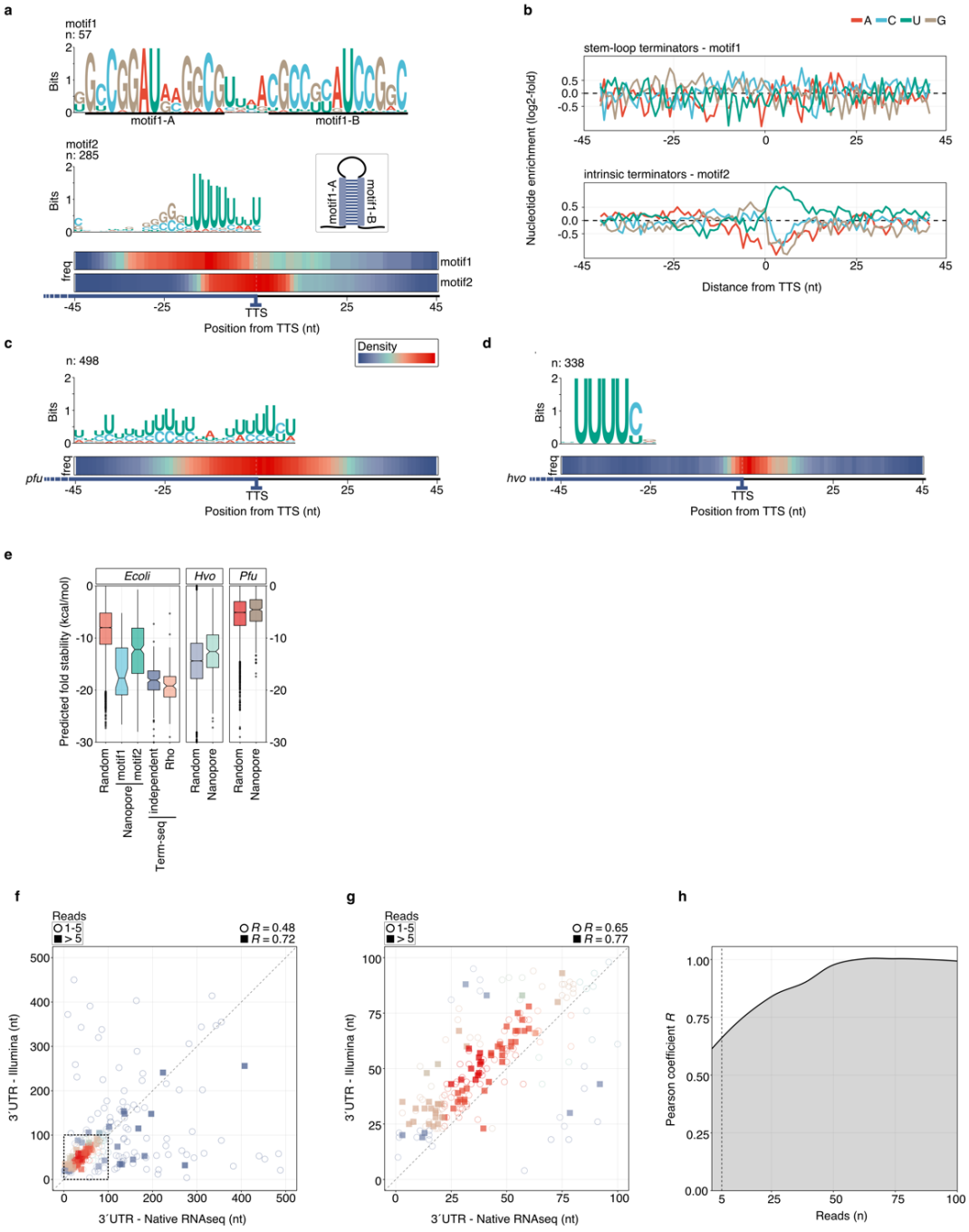
Supplementary Figure 6 | Poly(A)-tailing efficiency. Distribution of poly(A)-tails (nt) estimated by nanopolish (Loman et al., 2015) is shown for color-coded transcript groups that were separated based on their region lengths (protein coding genes (CDS): purple, 5S: green, 16S: red, 23S: blue).



Supplementary Figure 7 | Correlation of transcript abundance levels between Nanopore native RNA-Seq and Illumina RNA-Seq. Transcript counts are calculated by featurecounts and rlog transformed using the DESeq2 package in R to account for differences in sample sizes (Love et al., 2014a; Liao et al., 2019). Each gene is represented by a point colored by density from low (darkblue) to high (red) to address for overplotting. Correlation is calculated by Spearman´s rho and shown in the upper left. Publicly available data sets generated by Illumina sequencing (see detailed description in Material and Methods) were used to compare ONT samples in **a**, *E. coli* (SRP056485), **b**, *P. furiosus* (Grünberger et al., 2019) and **c**, *H. volcanii* (SRR7811297)(Laass et al., 2019).

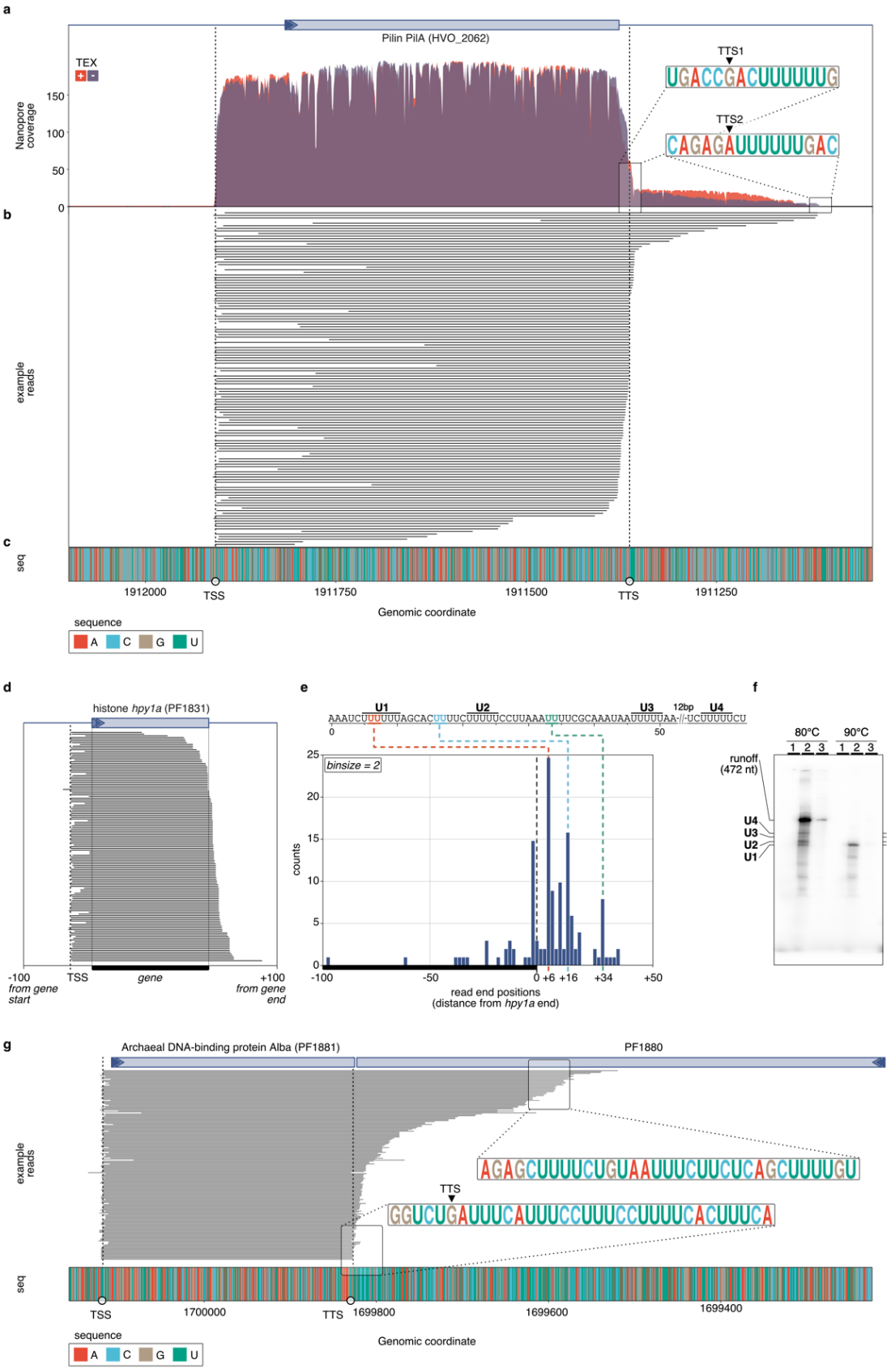


Supplementary Figure 8 | Transcription start site (TSS) analysis. a, Primary transcription start sites (TSS) were predicted based on Nanopore reads and compared to Illumina d(ifferential) RNA-Seq data from published data sets for *E. coli* (Thomason et al., 2015), *P. furiosus* (Grünberger et al., 2019) and *H. volcanii* (Babski et al., 2016). 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. Panel **b-e**: Correlation analysis between different 5´UTR lenghts with every dot representing one gene, the number of mapped reads indicated by an empty circle (1-5) or filled square (> 5) and density of data shown by color scale from darkblue (low) to high (red). Data are shown for *E. coli* in **b**, *P. furiosus* in **c**, *H. volcanii* in **d**, and TEX and NOTEX set of *H. volcanii* in **e**. Spearman´s rho correlation is shown for all comparisons.



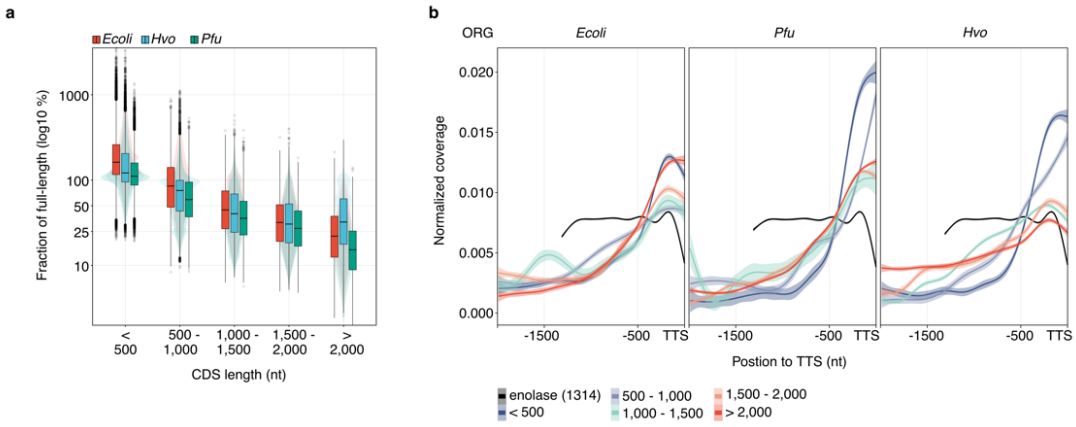
Supplementary Figure 9 | Transcription termination site (TTS) analysis. **a**, MEME analysis (Bailey et al., 2009) of TTS in *E. coli* identifies significantly enriched motifs in *E. coli* (stem-loop forming motif1 represent REP sites, intrinsic poly(U)-containing motif2) (Dar and Sorek, 2018). Enriched motifs and position of motifs in the scanned region from -45 to +45 from the TTS is shown as sequence pattern and heatmap. **b**, Nucleotide enrichment analysis of terminating regions for motif1 and motif2. Enrichment was calculated by comparing the genomic sequences surrounding the TTS (-45 to +45) to randomly selected intergenic positions (n: 10000). **c**, MEME output (enriched motifs and position-heatmap) of termination site motif scanning in *P. furiosus* and **d**, *H. volcanii*. **e**, Predicted fold stability of selected termination sequences (45 bases upstream of TTS) and comparison to random generated sequences. Fold stability was calculated by RNAfold(Lorenz et al., 2011). **f**, Comparison of 3´UTRs in *H. volcanii* predicted by native RNA sequencing (x-axis)

and a short-read Term-Seq approach (Berkemer et al., 2020a). Note that every dot represents one gene, the number of mapped reads is indicated by an empty circle (1-5) or filled square (> 5) and density of data is shown by color scale from darkblue (low) to high (red). **g**, Zoom to the majority of 3´UTRs that have 3´UTRs < 100 nts. Spearman´s rho correlation is shown for all comparisons. **h**, Calculated Spearman´s rho coefficient was calculated depending on the number of reads that mapped to a CDS.

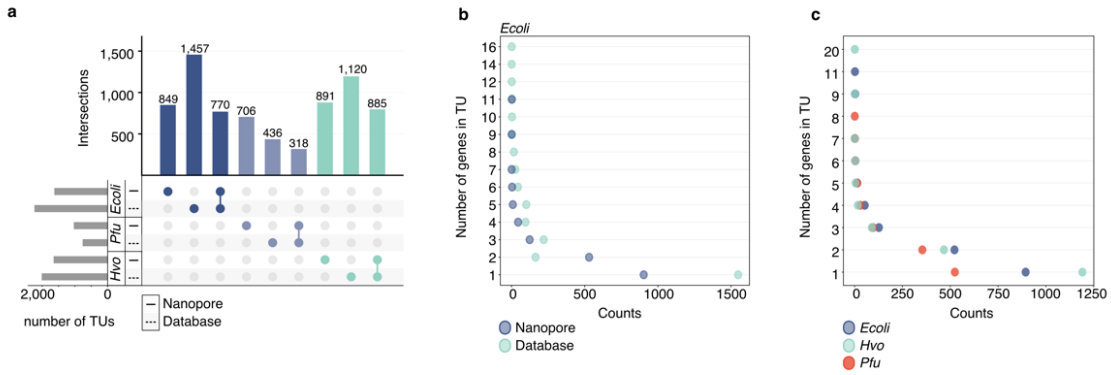


Supplementary Figure 10 | Analysis of termination events for the Pilin PilA gene (HVO_2062) in *Haloferax volcanii* and the hsitone *hpy1a* (PF1831) gene in *P. furiosus.*

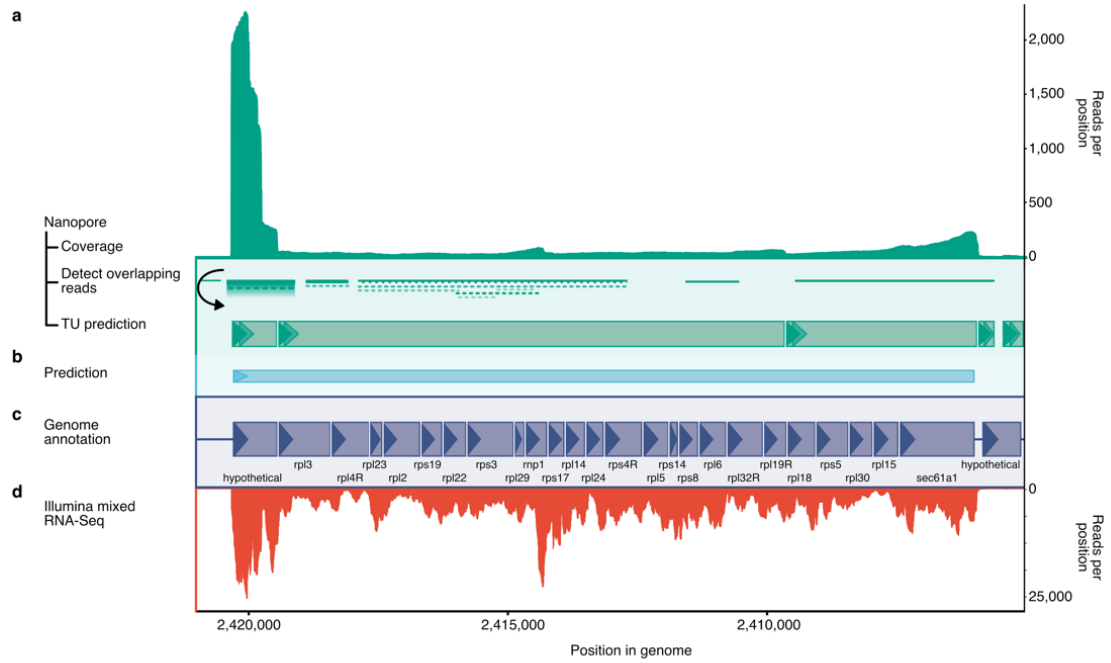
a, Nanopore coverage calculated for both TEX (red) and NOTEX (purple) sample in *Haloferax volcanii*. Coverage drops results from mismatches in the mapped reads. The TU-based estimated transcript boundaries (TSS, TTS) are indicated by vertical dashed lines. Termination specific poly(U) sequences are highlighted in the 3´ UTR region. **b**, IGV snapshot of example reads highlighting different 3´ UTR variants. **c**, Sequence of the plotted region visualized by color-coding the different nucleotides (red: A, blue: C, brown: G, green: U). **d**, Single read track of reads mapping to the *hpy1a* gene. **e**, The 3' read end positions of reads in d were extracted and are shown in a histogram view. Note that one bar reflects the added up counts of 2 positions. In the upper panel the sequence starting from the annotated gene end is shown, with U_5 sequences highlighted with U1-U4 (Spitalny and Thomm, 2008). Enriched Nanopore read ends are color-coded. **f**, *In vitro* transcription of the *hpy1a* template displayed at 80°C and 90°C using 2mM ATP, 2mM GTP and (1) no CTP, 0.04 mM UTP, (2) 2 mM CTP, 0.04 mM UTP, (3) 2 mM CTP, 2 mM UTP (see Material & Methods). Lenghts of transcripts are depicted on the left. **g**, IGV snapshot of Alba-mapping reads in *P. furiosus*. Two consecutive terminating U-stretches and the color-coded sequence are highlighted.



Supplementary Figure 11 | Prerequisites for transcriptional unit (TU) detection. a, The fraction of full-length transcripts is compared between protein-coding gene (CDS) classes, that are sorted by gene length in all three TEX-treated samples. **b**, Coverage metaplots aligned to the transcription termination site (TTS) of the 100 most transcribed genes in all samples, show a 3´ to 5´ drop in all native RNA classes (colored by gene length), except for the enolase control (black).

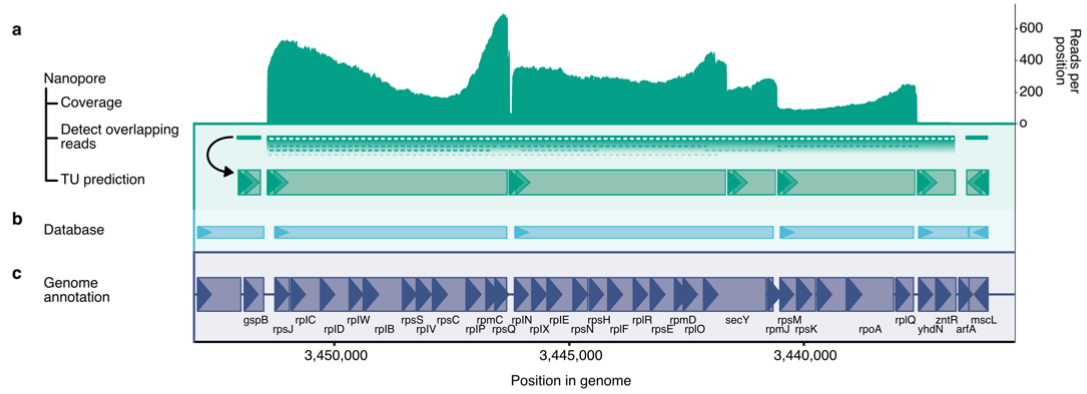


Supplementary Figure 12 | Detection of transcriptional units (TU) in three prokaryotic model organisms. **a**, The number and composition of TUs is compared between Nanopore-detected TUs and operon annotations retrieved from databases for *E. coli* (Mao et al., 2015), *P. furiosus* (DOOR2 database) and *H. volcanii* (DOOR2) (Mao et al., 2014). The total number of TUs is indicated by grey bars and results from the sum of Nanopore-only predicted TUs and the intersection to the database data. **b**, Comparison of the number of genes in transcriptional units in *E. coli* Nanopore data (purple) and the database set (light-green) (Mao et al., 2015). **c**, Number of genes in a TU in all three datasets. TU prediction is shown only for TEX-treated samples.

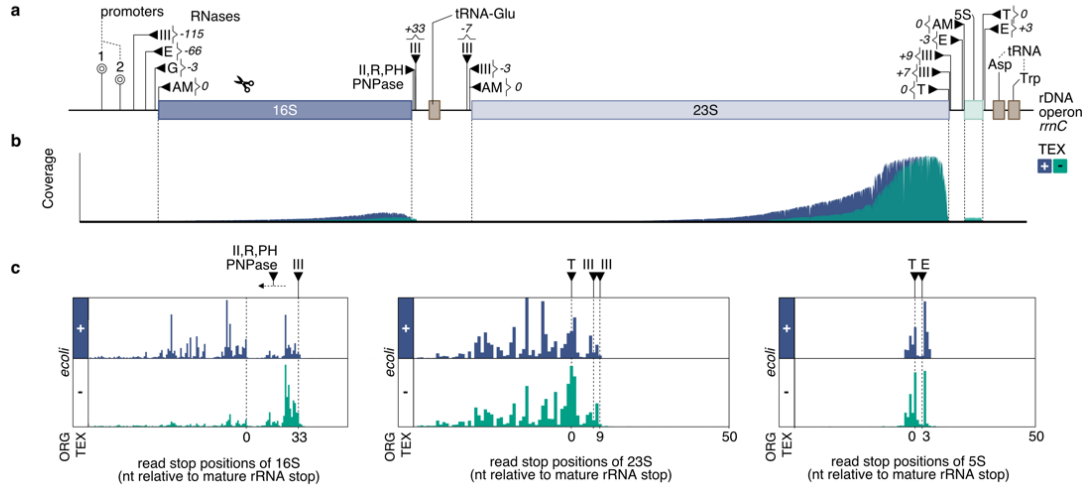


Supplementary Figure 13 | Large transcriptional unit (TU) annotation of a large ribosomal-protein-containing operon in *Haloferax volcanii*. **a**, Coverage of Nanopore reads is shown in the top panel. TU prediction is performed by detection and linkage of overlapping reads and splitting them according to a 3´ drop in coverage (compare Supplementary Figure 10b). Predicted TUs are drawn with green boxes according to scale. **b**, Comparison to bioinformatical prediction using the DOOR2 database (Mao et al., 2014). **c**, Genome annotation with abbreviated

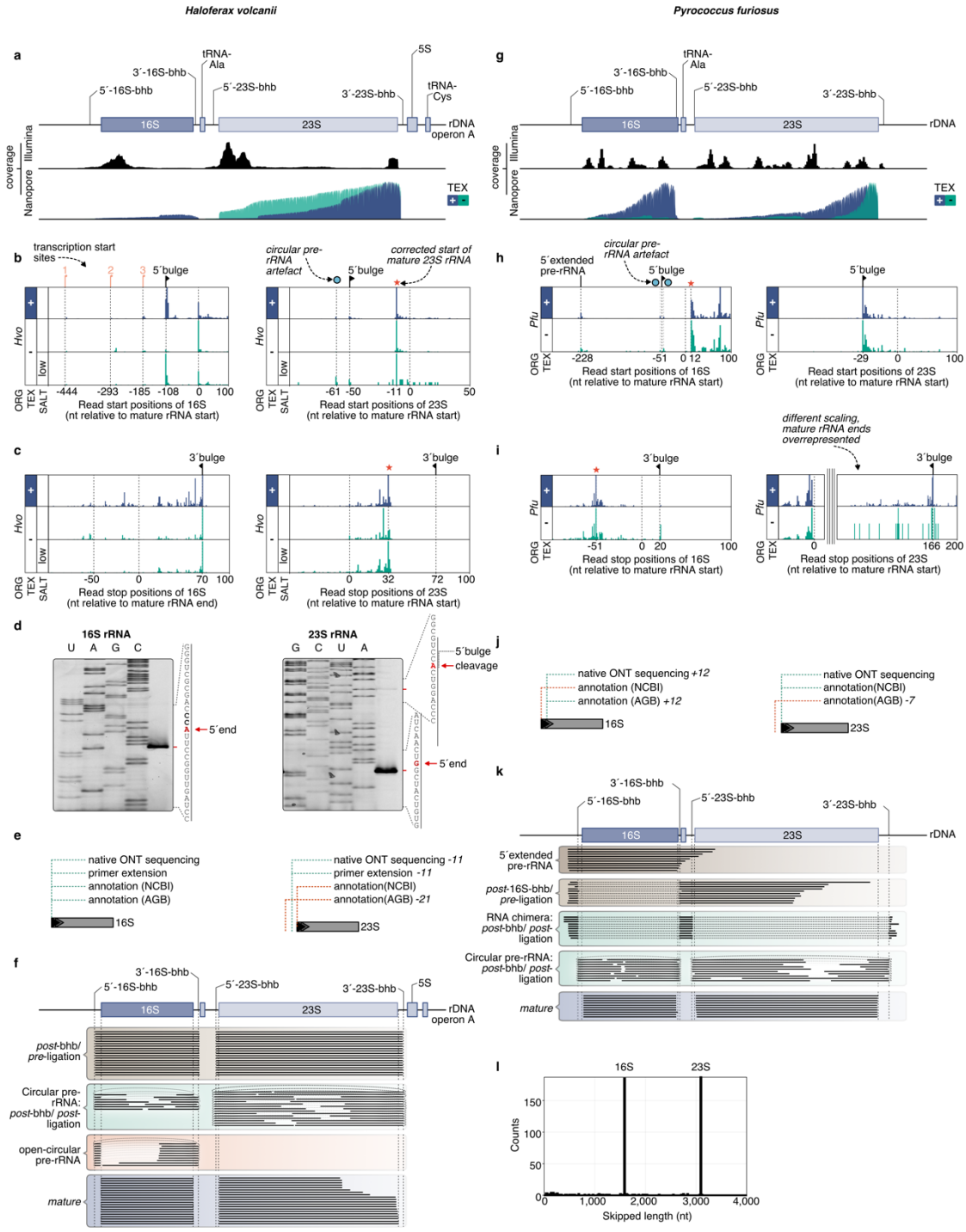
gene names, boxes drawn to scale and strand indicated by triangles (Hartman et al., 2010). **d**, Coverage detected using Illumina sequencing of a mixed RNA sample (Laass et al., 2019).



Supplementary Figure 14 | Transcriptional unit (TU) annotation of the large ribosomal-protein-containing operon in *Escherichia coli*. **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 (compare Supplementary Figure 10b). **b**, Comparison to published analysis of TUs in *E. coli* based on a bioinformatical approach (Mao et al., 2015). **c**, Genome annotation with abbreviated gene names, boxes drawn to scale and strand indicated by triangles (Riley et al., 2006).

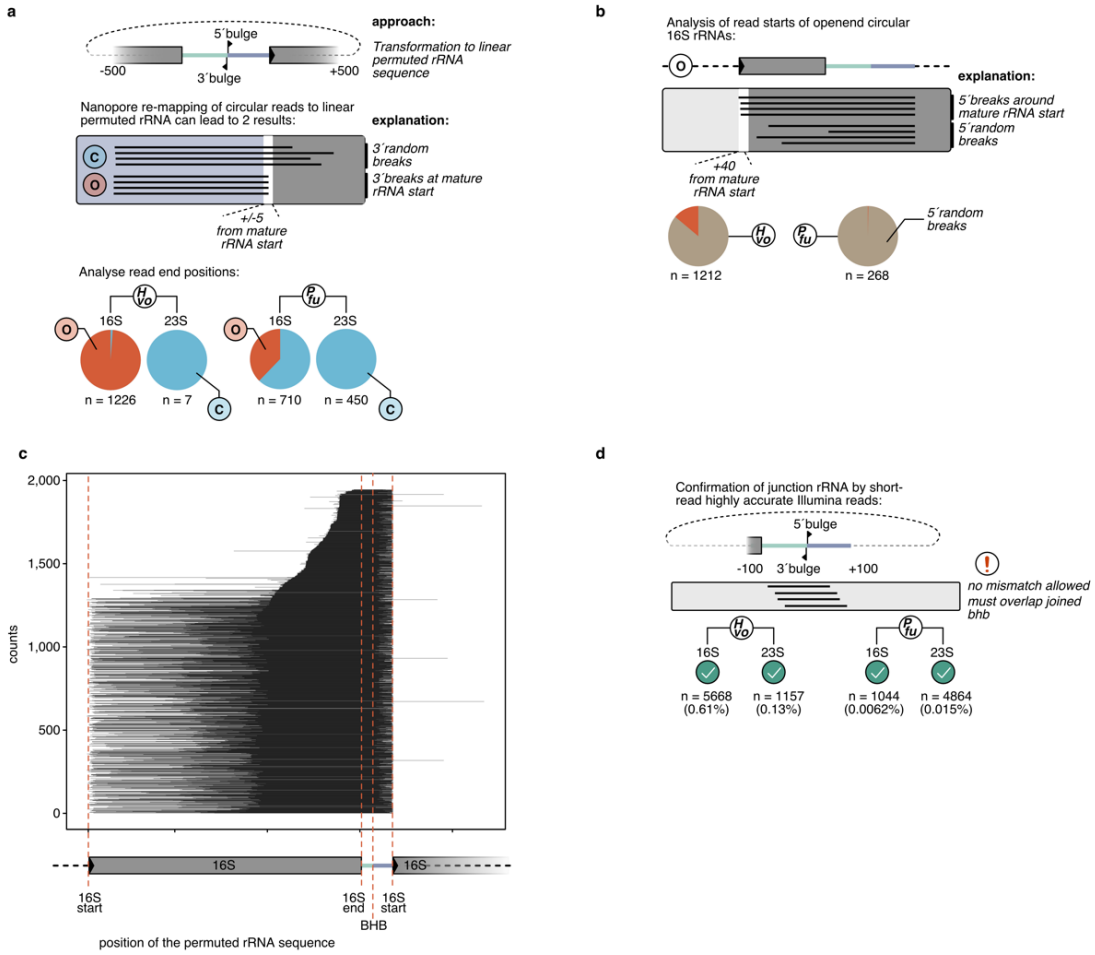


Supplementary Figure 15 | Detection of ribosomal RNA processing sites in *E. coli*. **a**, Transcription of the rDNA locus (rrnC) is starting from two promoters (transcription start sites at -293 and -175) (Maeda et al., 2015). Precursor RNAs are cleaved by RNases (black triangles) at depicted positions (Shajani et al., 2011; Ferreira-Cerca, 2017; Smith et al., 2018; Jain, 2020). **b**, Histograms of read end positions for 16S, 23S and 5S rRNA. Positions are relative to the annotated boundaries of mature rRNAs and shown for TEX (+, purple) and NOTEX (-, green) samples.

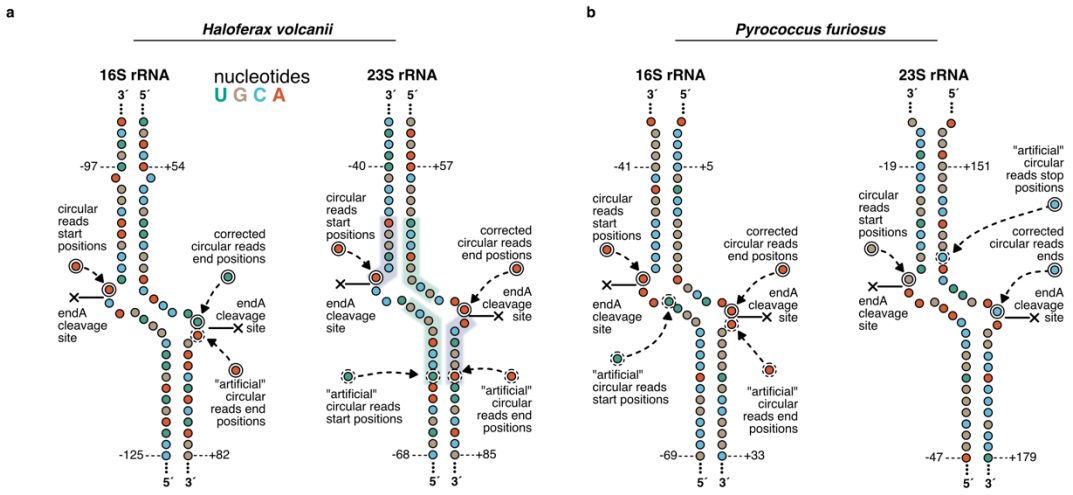


Supplementary Figure 16 | rRNA processing site detection in *H. volcanii* and *P. furiosus*. **a/g**, Schematic of the rDNA locus operon A in *H. volcanii* and the only rDNA locus in *P. furiosus*. Nanopore coverage tracks are shown for TEX (+, blue) and NOTEX (-, green) samples and compared to short-read Illumina coverages. **b/h**, Histograms of read start positions and **c/i**, read end positions of 16S and 23S rRNA relative to annotated boundaries of mature rRNAs of TEX (+) and NOTEX (-) samples. Cleavage sites are indicated by black triangles. Asterisks mark sites of mature rRNA that are potentially not annotated correctly. **d**, Mapping of mature 16S and 23S rRNA 5´ends by primer extension. Primer extension were performed with the indicated primers as described in Material and Methods. From the comparison to a sequencing ladder, positions of mature

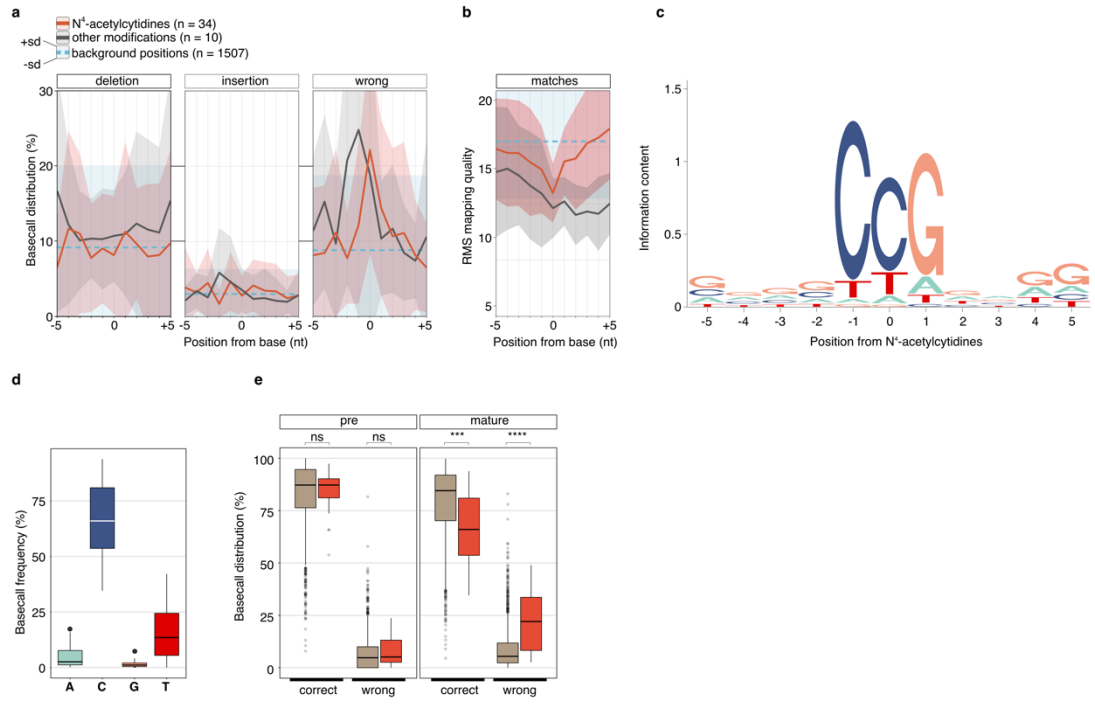
16S rRNA and 23S rRNA and the 5´bulge of the 23S rRNA could be assigned (red arrows). **e/j** Annotation of the 16S and 23S rRNA start based on NCBI (https://www.ncbi.nlm.nih.gov/genome/) and archaeal genome browser (AGB, http://archaea.ucsc.edu) versions, compared to positions enriched in native ONT sequencing and experimentally verified by primer extension. **f/k** Single-read tracks of read categories derived from (i) co-occurrence analysis of read start and stop positions, (ii) the number of junctions and (iii) clipping properties. **l**, Skipped length of the RNA chimera class verifies accuracy of "spliced" read detection.



Supplementary Figure 17 | Circular read detection of archaeal rRNA precurosors. a, Circular reads were confirmed by re-mapping Nanopore reads to a permuted rRNA sequence, containing the joined bulge-helix-bulge site. We categorized reads based on their 3´terminal positions as circular pre-RNA (C, random breaks) and open-circular pre-rRNAs (O, 3´end at mature rRNA start) and counted the number of reads fulfilling the criteria (*Pfu* TEX and *Hvo* NOTEX). **b**, Read start positions of open-circular pre-rRNAs. **c**, Circular reads were confirmed by mapping highly accurate Illumina reads to the permuted sequence, allowing for no mismatch and filtering out all reads that do not overlap the joined bulge-helix-bulge.

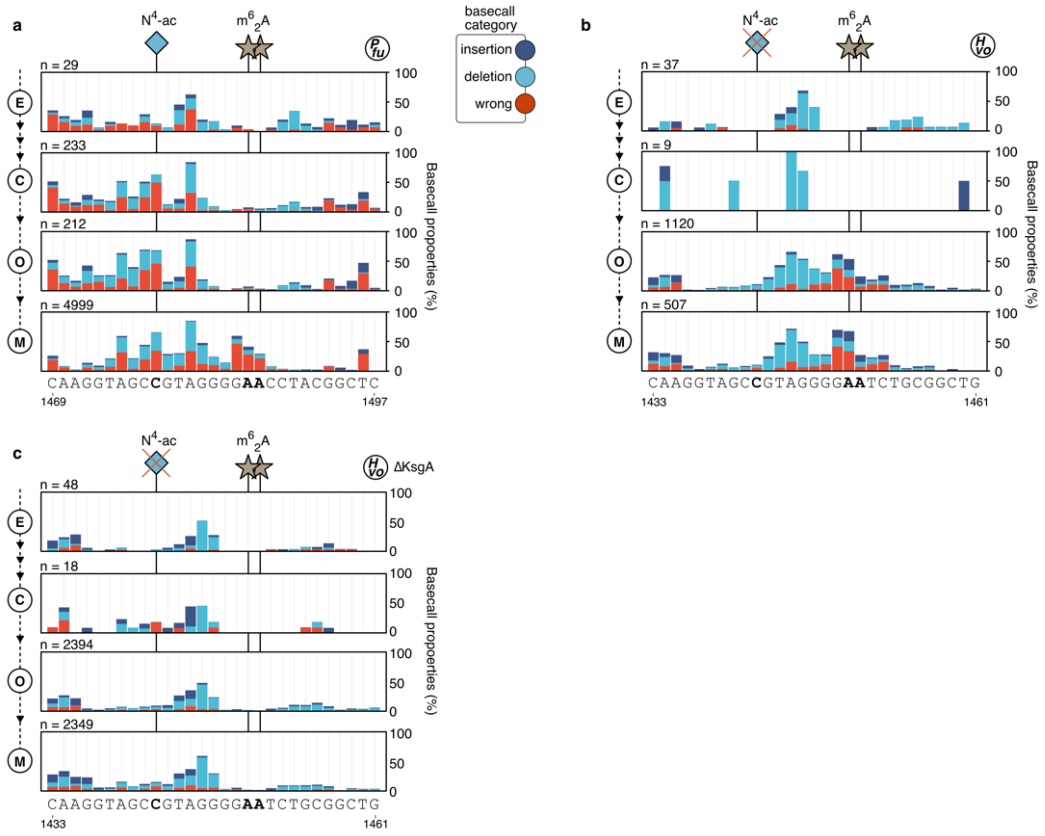


Supplementary Figure 18 | Secondary structure prediction of bulge-helix-bulges. **a**, Secondary structure predicted was performed by RNAfold from the Vienna RNA package (Lorenz et al., 2011). Artificial circular read end and start positions are caused by similarities in the 5´leading and 3´-trailing sequences, but are mostly accurately detected at the endA cleavage site in *H. volcanii* and **b**, *P. furiosus*.

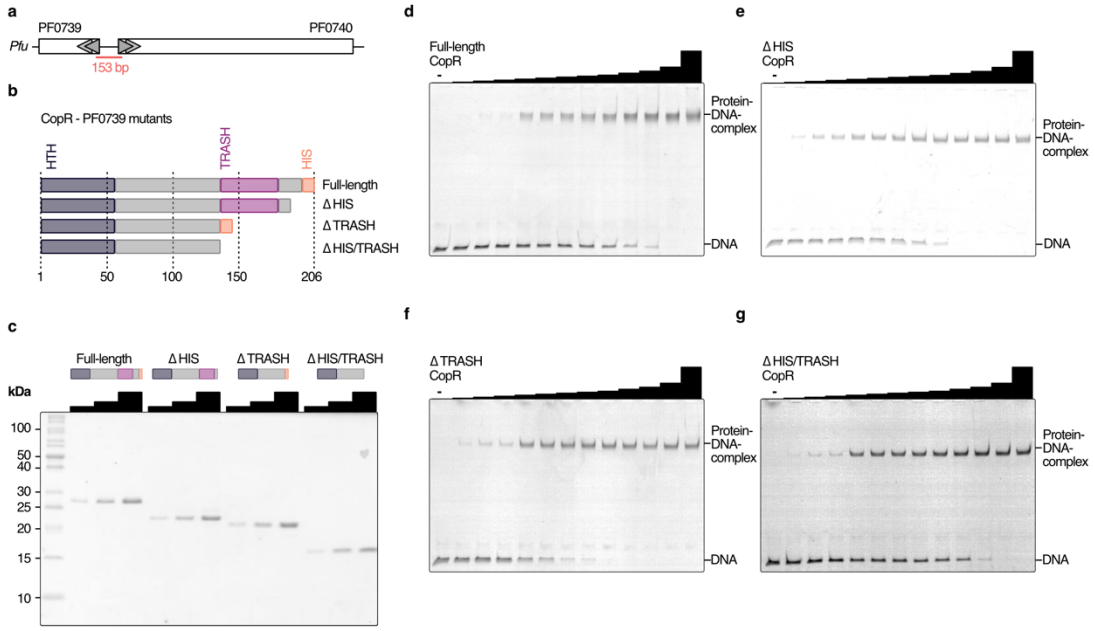


Supplementary Figure 19 | Detection of N^4-acetylcytidine modifications in *P. furiosus*. **a**, The proportion of deletions, insertions and wrong bases is shown in a window from -5 to +5 from the presumably modified or background base. Shaded areas show the upper and lower standard deviation, while the lines show the mean values of N^4-acetylated positions (red), diverse other modifications (grey) and all other positions of the 16S rRNA in *P. furiosus* using the recently established 16S rRNA modification pattern in the close relative *Pyrococcus abyssi* (Coureux et al., 2020). **b**, Root mean square (RMS) mapping quality in the surrounding sequence context of potentially modified bases. **c**, Sequence logo of actual sequenced bases averaged for all 34 potentially

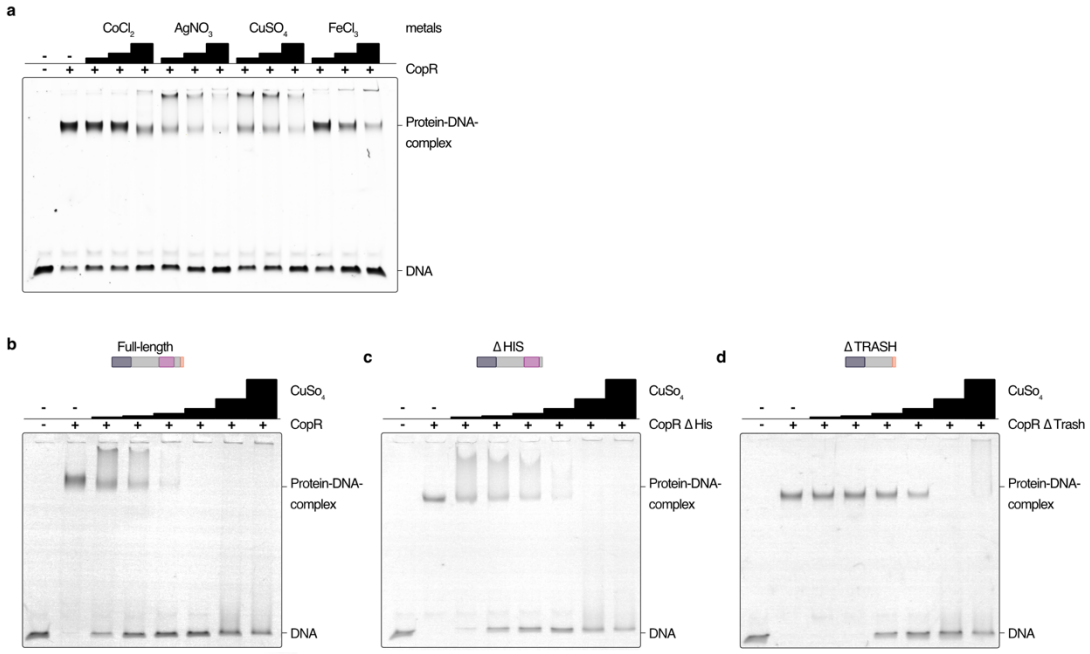
N^4-acetylated positions in the 16S rRNA. **d**, Frequency of each nucleotide to be basecalled in a N^4-acetylated CCG context. **e**, Comparison of the basecall properties of 5´-extended pre-rRNA and mature rRNAs. Statistical significance (p-values, T-test) is indicated by asterisks (p-value > 0.05: ns (not significant), p <= 0.0001: ****).



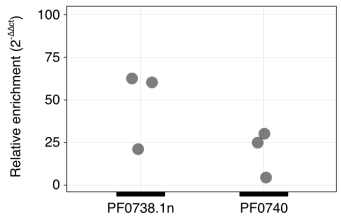
Supplementary Figure 20 | Detection of h45-N^4-acetylation and KsgA-dependent $m^6{}_2A$ modification. **a**, Basecall properties (insertion: dark-blue, deletion: light-blue, wrong: red) are shown for read subsets reflecting different stages of rRNA maturation (E: 5´-extended pre-rRNAs, C: circular pre-rRNAs, O: open-circular pre-rRNAs, M: mature rRNAs) in *P. furiosus*, **b**, *H. volcanii* wildtype and **c**, *H. volcanii* ΔKsgA.



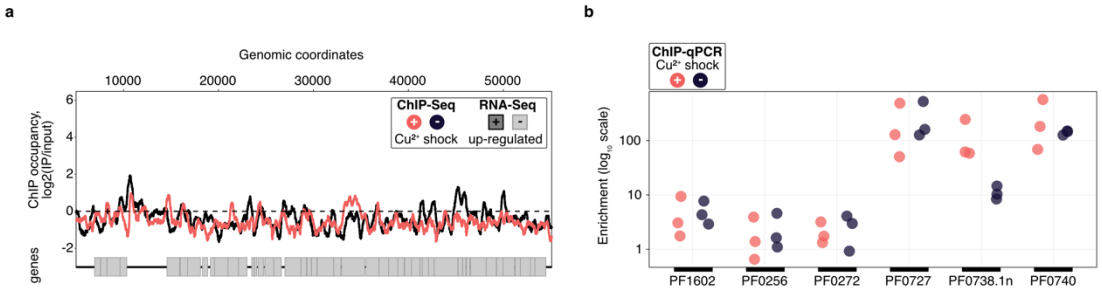
Supplementary Figure 21 | CopR (PF0739) binds to the promoter region of CopA (PF0740). a, Schematic representation of the *copR/copA* locus in *P. furiosus*. The template used for the EMSA analysis is highlighted in red and overlaps with both translation start sites. **b**, Schematic of mutants generated for the functional characterisation of CopR. DNA-binding helix-turn-helix (HTH), metal-sensing TRASH domain and additional C-terminal Histidine-rich sequence are highlighted in different colors. **c**, SDS-PAGE analysis of purified recombinant proteins. **d**, EMSA analysis were performed using 20 nM of DNA (153 b, see panel a) and increasing concentrations of recombinant protein (12.5, 25, 37.5, 50, 62.5, 75, 87.5, 100, 125, 150, 200, 400 nM).



Supplementary Figure 22 | Metal specificity of CopR and domain-deleted mutants in *P. furiosus* **a**, EMSA analysis was performed using 20 nM DNA (*copR/copA* promoter), 200 nM full-length protein and increasing concentrations (12.5, 25, 50 µM) of the respective metal ($CoCl_2$, $AgNO_3$, $CuSO_4$, $FeCl_3$). **b**, Influence of increasing $CuSO_4$ concentrations (50, 100, 200, 400, 800, 1600 µM) on the DNA-binding behavior of full-length CopR, **c**, CopRΔHIS and **d**, CopRΔTRASH using 20 nM DNA and 200 nM protein.



Supplementary Figure 23 | Relative enrichment of *pf0740* and *pf0738.1n* measured by RT-qPCR. Expression levels from biological triplicates (individual points are shown) were compared to a house-keeping gene *pf0256* (Spt5). Enrichment was calculated using ΔΔct method.



Supplementary Figure 24 | Confirmatory analysis of ChIP-seq results. a, Exemplary CopR-unbound 50-kb region of *P. furiosus* (compare Figure 29a). ChIP-seq curves were generated for Cu^{2+} shocked (red) and untreated (dark blue) samples by comparing the IPs to input samples (mean values of triplicates are shown). Genome annotation is shown at the bottom according to scale with significantly up-regulated genes (adjusted p value < 0.05, Log_2 fold change +/- $Cu^{2+} > 1$) colored in dark grey. **b**, ChIP-qPCR results of multiple ChIP-seq identified CopR-unbound (PF1602, PF0256, PF0272) and CopR-bound (PF0727, PF0738.1n, PF0740) regions. Data are shown as individual points of biological replicates for normal conditions (dark blue) and copper-treated cells (red). Values are calculated as fold enrichment over input sample.

Supplementary Tables:

To facilitate accessibilty of multi-column tables that are in a non-printable format, all supplementary tables have been uploaded to https://github.com/felixgrunberger/dissertation.
Additionally, supplementary tables can be found from the respective publisher´s site:
https://www.frontiersin.org/articles/10.3389/fmicb.2019.01603/full#supplementary-material (Grünberger et al., 2019)
https://www.biorxiv.org/content/10.1101/2019.12.18.880849v2.supplementary-material (Grünberger et al., 2020a)
https://www.biorxiv.org/content/10.1101/2020.08.14.251413v1.supplementary-material (Grünberger et al., 2020b)

Supplementary Table 1 | Details for mutations noted in the new genome assembly

Supplementary Table 2 | List of new locus tags

Supplementary Table 3 | RNA-seq and mapping statistics

Supplementary Table 4 | Operon-mapper output

Supplementary Table 5 | ANNOgesic output

Supplementary Table 6 | aTSS overlapping IS elements

Supplementary Table 7 | Run and reads statistics

Supplementary Table 8 | Read counts of ONT data sets calculated with featurecounts

Supplementary Table 9 | TSS estimated using Nanopore sequencing

Supplementary Table 10 | TTS estimated using Nanopore sequencing

Supplementary Table 11 | TUs determined using Nanopore sequencing

Supplementary Table 12 | Used strains, plasmids and primer sequences

Supplementary Table 13 | Illumina sequencing and mapping statistics

Supplementary Table 14 | DESeq2 output

Acknowledgements

In erster Linie möchte ich mich bei PD Dr. Winfried Hausner nicht nur für die Betreuung meiner Promotion bedanken, sondern auch für die Freiheit in der Ausarbeitung, die ich nicht für selbstverständlich halte und sehr zu schätzen weiß. Ein großer Dank gilt auch Prof. Dr. Dina Grohmann für die Möglichkeit an ihrem Lehrstuhl zu promovieren, die ansteckende Begeisterung, viele interessante Gespräche und die Ermöglichung der Umsetzung spannender Projekte. Dankeschön an das Mentoring-Team meiner Doktorarbeit, Prof. Dr. Christine Ziegler und Prof. Dr. Bettina Siebers. Ein Dank im Voraus gilt außerdem Prof. Dr. Klaus Grasser für die Anfertigung des Zweitgutachtens. Bei allen Labor- und Bürokollegen und allen anderen Mitarbeitern des Lehrstuhls möchte ich mich für die angenehmen Atmosphäre bedanken. Danke an alle ehemaligen und aktuellen Postdocs, Doktoranden und Studenten, den TAs Wolfgang Forster und Renate Richau und allen anderen, die etwas zu den Manuskripten beigetragen haben. Insbesondere gilt der Dank auch PD Dr. Sébastien Ferreira-Cerca und seiner Gruppe um Robert Knüppel und Michael Jüttner für die Einführung und alle Diskussionen rund um die archaeelle Ribosomenbiogenese. Nicht vergessen möchte ich das RIGeL Management Office, vor allem Kinga Ay, für die unkomplizierte Unterstützung bei jeglichen administrativen Fragen. Danke außerdem an die Graduate Research Academy RNA Biology für die Bewilligung von Reisezuschüssen. Zuletzt möchte ich meiner Familie und Ari für die Unterstützung während der Doktorarbeit danken und für das fortwährende Interesse daran, was ich denn da eigentlich genau mache.