

Investigating the Biological Role and Binding Modes of Histone-Like Proteins of  
Halophilic Archaea.

by

Saaz Sakrikar

University Program in Genetics and Genomics  
Duke University

Date: \_\_\_\_\_

Approved:

---

Amy Schmid, Supervisor

---

David Macalpine, Chair

---

Amy Grunden

---

Raluca Gordan

---

Richard Brennan

Dissertation submitted in partial fulfillment of  
the requirements for the degree  
of Doctor of Philosophy in the  
University Program in Genetics and Genomics in the Graduate School  
of Duke University

2022

ABSTRACT

Investigating the Biological role and Binding Modes of Histone-Like Proteins of  
Halophilic Archaea.

by

Saaz Sakrikar

University Program in Genetics and Genomics  
Duke University

Date: \_\_\_\_\_

Approved:

---

Amy Schmid, Supervisor

---

David Macalpine, Chair

---

Amy Grunden

---

Raluca Gordan

---

Richard Brennan

An abstract of a dissertation submitted in partial fulfillment of  
the requirements for the degree  
of Doctor of Philosophy in the  
University Program in Genetics and Genomics in the Graduate School  
of Duke University

2022

Copyright by  
Saaz Sakrikar  
2022

## Abstract

Protein-based compaction of the genome is a feature found in species across the tree of life. In Archaea, the majority of species contain a histone fold domain-containing protein, and these have been shown to compact DNA through the formation of nucleosomes and extended structures called hypernucleosomes. However, the role of the histone-like proteins of halophilic archaea is unclear. Previous work in the model species *Halobacterium salinarum* indicated that its sole histone gene, *hpyA*, is dispensable for growth and is expressed at very low levels. I hypothesize that the unique high-salt environment of halophilic archaea has selected for an alternative histone function, and that they function instead as transcription factors.

This hypothesis was addressed with genetic approaches including the creation of knockout and complementation strains, traditional microbiology techniques including growth assays and microscopy, and high-throughput genomics approaches: ChIP-Seq to study genome-wide binding, and RNA-Seq to study differential expression in  $\Delta hpyA$  strain. It was found that *hpyA* is required for optimal growth in hypo-osmotic conditions, and exhibits strong salt-dependent binding patterns and gene regulation. It directly regulates genes involved in iron uptake, and indirectly regulates genes in ion transport and nucleotide metabolism. These results validate the link between histone function and the high-salt environment of halophilic archaea.

Similar to *hpyA*, I found that the sole histone gene of another model halophile: *hstA* of *Haloferax volcanii*, could be deleted, and that knockout cells remained viable. The

genome-wide binding of both halophilic histones was studied, and compared with publicly available data regarding the binding patterns from transcription factors (TFs), nucleoid-associated proteins (NAPs), and eukaryotic histones. Halophilic histones bind in narrow, discrete, and relatively rare peaks, just like TFs; however, this binding is not enriched at the promoter, and they instead bind evenly in both intergenic and coding regions (like some NAPs). Their occupancy profile across gene start sites do not resemble those of histones or TFs. In terms of sequence specificity, HpyA exhibits a histone-like preference for 10bp periodicity, while HstA exhibits a TF-like trait in preferentially binding a palindromic sequence motif. When considering all the data, I conclude that halophilic histones blur the line between TFs, NAPs, and histones.

A major technical challenge in generating this data was the removal of rRNA prior to carrying out RNA-Seq. Several approaches were tested across four model species of halophiles, and the reasons for differences in performance for these approaches were analyzed. Methods that deliver efficient rRNA removal targeted to a particular species, or to halophilic archaea in general, are highlighted.

Together these results shed light on the unusual function and binding modes of the histone-like proteins of halophilic archaea. In combination with other recent work, they suggest that histone function is linked with the physical environment of archaeal species.

# Table of Contents

Abstract .....	iv
List of tables .....	ix
List of figures .....	x
<b>Introduction:</b> .....	<b>1</b>
1.1 Archaea have bacterial, eukaryotic, and unique molecular features.....	1
1.2 Halophilic archaea, members of the Euryarchaea, are excellent model systems for genetic and genomic studies .....	4
1.3 Specialized DNA binding proteins compact or organize genomes throughout the tree of life.....	7
1.4 Characterised archaeal histones compact DNA and form nucleosomes, with notable exceptions.....	11
1.5 Open questions on archaeal chromatin motivate this thesis work .....	17
<b>2. An archaeal histone-like protein regulates gene expression in response to salt stress</b> .....	<b>20</b>
2.1. Introduction .....	20
2.2. Materials and Methods.....	23
2.2.1 Strains, media and general culturing: .....	23
2.2.2 Growth and microscopy:.....	25
2.2.3 ChIP-seq experiments:.....	26
2.2.4. Analysis of ChIP-seq data:.....	27
2.2.5. RNA-seq experiments:.....	29
2.2.6. RNA-seq data analysis: .....	30
2.3. Results.....	31
2.3.1. HpyA is important for wild type growth and morphology in low salinity stress conditions. ....	31
2.3.3. HpyA functions primarily as an activator of genes encoding ion transport and metabolic proteins.....	40
2.4. DISCUSSION .....	45
2.5. Acknowledgements .....	51
<b>3. Haloarchaeal histone proteins blur the line between transcription factors and nucleoid-associated proteins</b> .....	<b>52</b>
3.1. Introduction .....	52

3.2. Results and Discussion.....	58
3.2.1. <i>Haloferax volcanii</i> histone HstA is not essential but is important for maintaining wild type growth rate.....	58
3.2.2. HstA genome-wide location analysis (ChIP-seq) reveals binding patterns similar to those of HpyA.....	63
3.2.3. Comparison of halophilic histone-like protein binding patterns with those for TFs, NAPs, and eukaryotic histones.....	65
3.2.4. Haloarchaeal histone-like protein occupancy curves surrounding start sites are unique relative to canonical histone and TF signals.....	70
3.2.5. Halophilic genomes lack the dinucleotide periodicity that indicates genome-wide optimization for histone binding.....	74
3.2.6. Halophilic histones differ in predicted DNA binding sequence specificity.....	80
3.3. Conclusions.....	82
3.4. Materials and methods.....	86
3.4.1. Strain construction:.....	86
3.4.2. Media, culturing, and phenotyping:.....	87
3.4.3. ChIP-seq experiment:.....	88
3.4.4. ChIP-seq analysis:.....	89
3.4.5. Generating <i>Hfx volcanii</i> HstA peak list:.....	90
3.4.6. Start site occupancy analysis:.....	91
3.4.7. Dinucleotide periodicity analysis:.....	92
3.4.8. Motif search:.....	93
3.5. Acknowledgements.....	94
<b>4. Comparative Analysis of rRNA removal methods for RNA-seq Differential Expression in Halophilic Archaea.....</b>	<b>95</b>
4.1. Introduction.....	95
4.2. Material and methods.....	98
4.2.1. Media, Strains, and Growth Conditions.....	98
4.2.2. RNA-seq experimental protocol.....	100
4.2.3. Data Analysis.....	102
4.3. Results.....	106
4.3.1. Discontinuation of the Illumina RiboZero kit is associated with a decline in published archaeal RNA-Seq studies:.....	106

4.3.2. Testing new rRNA depletion strategies on total RNA samples from <i>Halobacterium salinarum</i> (HBT).	108
4.3.3. Species-specific probe methods efficiently remove <i>Haloferax volcanii</i> (HVO) rRNA:	114
4.3.4. siTools Panarchaea kit efficiently removes rRNA from diverse halophilic archaeal species:	115
4.3.5. Choice of removal method does not affect per-gene read counts:	116
4.3.6. Utility of rRNA removal is seen in counts of non-rRNA genes:	118
4.4. Conclusions and discussion	120
<b>Conclusions and Future Directions</b>	<b>123</b>
5.1 Conclusions.	123
5.2 Future directions	126
<b>Appendix A</b>	<b>129</b>
<b>Appendix B</b>	<b>134</b>
<b>Appendix C</b>	<b>136</b>
<b>Appendix D</b>	<b>139</b>
<b>References:</b>	<b>145</b>

## List of tables

Table 1: Summarized version of the diversity of chromatin proteins .....	11
Table 2: Strains used in this study. ....	99
Table 3: All media recipes used for test organisms in this study .....	99
Table 4: Doubling time and Incubation time for different species in different media....	100
Table 5: Primers used to check for genomic contamination. ....	101
Table 6: rRNA-coding gene identifiers for each species of interest .....	104
Table 7: Median rRNA removal using different methods.....	107

## List of figures

Figure 1: Cartoon representation of the phylogeny of archaeal lineages.....	2
Figure 2: The $\Delta$ hpyA strain is impaired for growth under reduced salt conditions.....	32
Figure 3: No significant difference in growth rate was observed for WT (blue) and KO (red) strains grown in optimal salt. ....	33
Figure 4: OD600 is correlated with CFU/mL across strains and conditions.....	34
Figure 5: Circularity of Hbt. salinarum increases when hpyA is deleted under reduced salt. ....	36
Figure 6: ChIP-seq of HpyA shows salt and growth phase dependent binding patterns.. ....	37
Figure 7: HpyA binds without preference for coding vs non-coding regions. ....	39
Figure 8: HpyA regulates gene expression in a salt-dependent manner. ....	41
Figure 9: HpyA-dependent regulon shows diverse expression patterns in the conditions tested. ....	44
Figure 10: $\Delta$ hstA strain is impaired for growth in optimal conditions. ....	59
Figure 11: $\Delta$ hstA phenotype in response to diverse stress conditions. ....	61
Figure 12: hstA deletion can be complemented in-trans; HA tag does not interfere with HstA function.. ....	62
Figure 13: ChIP-seq binding signal for HpyA and HstA compared with TFs, NAPs, and eukaryotic histone. ....	65
Figure 14: HpyA and HstA bind in few, discrete peaks, like TFs, and contrasting with histones and NAPs.....	67
Figure 15: Genomic features of HpyA and HstA binding sites according to ChIP-seq data.....	69
Figure 16: Binding occupancy at start sites of selected DNA-binding proteins. ....	71
Figure 17: Binding occupancy at start sites of bacterial NAPs.. ....	72
Figure 18: Heatmap and average line of binding occupancy for yeast histone and HpyA .....	73
Figure 19: AA/TT/TA dinucleotide periodicity shows histone-linked pattern.. ....	76
Figure 20: Additional genome-wide AA/TT/TA and GC periodicities.....	77
Figure 21: Phylogenetic tree of selected archaeal species, with their genome-wide periodicities and histone presence/function shown.....	79
Figure 22: Sequence specificity of HpyA and HstA binding.. ....	81
Figure 23: Qualitative 3-D visual representation of binding characteristics of selected DNA-binding proteins investigated in this study.....	85
Figure 24: Slowdown in Archaeal RNA-Seq publications in recent years.....	106
Figure 25: Percentage of rRNA remaining in halophile RNA by using the discontinued Ribozero kit (RZ). ....	108
Figure 26: rRNA removal using alternative methods in Hbt salinarum.....	109
Figure 27: Analysis of sequencing depth, number of biological replicates, and detection of differentially expressed genes (2-fold differential expression) using the online tool Scotty.....	110

Figure 28: Increasing RNase digestion time is less important than probe sequence identity for efficient rRNA removal. ....	111
Figure 29: Species-specific probes efficiently remove rRNA from target species.....	114
Figure 30: Panarchaea kit offers efficient rRNA removal across halophilic species.....	115
Figure 31: Choice of removal method does not affect relative abundance of non-rRNA genes. ....	117
Figure 32: More complete rRNA removal leads to increased detection of lowly expressed genes. ....	119

## Acknowledgements

I want to thank Prof. Schmid for letting me change my research topic early on, thus freeing me from ever studying metabolism, and instead letting me dig into this weird “histone”. And for her guidance, mentorship, and hours (years?) of editing and making and re-making figures.

Thank you also to committee members (David Macalpine, Raluca Gordan, Richard Brennan, Amy Grunden) for their feedback and questions that helped focus and refine the project. I was lucky to have excellent professors during my undergraduate days: Profs. Pradeepkumar, Bhat, and Patankar, who are the reason for my fascination with molecular biology and genetics.

I’d also like to thank Anne Lacey and Liz Labriola, who guided me through a lot of complicated requirements and paperwork. My UPGG cohort, especially Martine Tremblay and Dan Grigsby, helped me settle in my first year in a (very) faraway country.

Every Schmid lab member has been great. Angie Vreugdenhill, Mar Martinez-Pastor, and Cynthia Darnell helped me out immensely. Everything, even PCRs and gels, was new to me; I wouldn’t have been able to start, let alone finish, this project without their patient guidance. Thanks especially to Cindy for your never-ending questions in lab meetings, forcing me to sharpen my thinking about my project and my presentations. Thank you Rylee Hackley for your ideas, and for listening to my nonsense, science-related or otherwise.

Away from Duke, it's been reassuring to have Bodhi, Nisheet, Rushil, and Srimukh, always available to speak to. Finally, my parents have answered my calls at all hours of the day and night, and I'm sure I wouldn't have completed this work without having them to fall back on.

## **Introduction:**

### **1.1 Archaea have bacterial, eukaryotic, and unique molecular features**

Archaea, earlier classified as a third domain of life<sup>1</sup>, but currently classified along with eukaryotes in a two-domain tree of life<sup>2,3</sup>, share molecular features with both bacteria and eukaryotes, as well as their own unique features. For example, the basal transcriptional machinery of archaea resembles that of eukaryotes, with RNA polymerase recruited to the promoter with the TATA-binding protein and transcription factor B that bind to specific sequences upstream of the promoter<sup>4,5</sup>. However, proteins involved in regulation of transcription (including many helix-turn-helix transcription factors) resemble those found in bacteria at the level of amino acid sequence<sup>6</sup>. Archaeal mRNA is not subject to the characteristic modifications seen in eukaryotes – the 5'-cap and 3' poly A-tail, while the translational process has features resembling both eukaryotes and bacteria<sup>7,8</sup>. Archaeal genome compaction, particularly the diversity of archaeal histones, which is discussed in more detail in subsections 1.3 and 1.4, contains elements of bacterial, eukaryotic, and uniquely archaeal strategies.

Archaeal species, particularly those that were initially characterized in the history of the field of archaeal molecular biology, dominate in extreme and energy-limited environments<sup>9</sup>. Archaeal extremophiles have been found at extremes of temperature, pH, and salinity<sup>10</sup>. However, subsequent work has discovered that archaea are present across diverse ecological niches<sup>11</sup>. Non-extremophilic archaea have been found associated with human, plant and animal microbiomes<sup>12</sup>, in soil, and in marine and

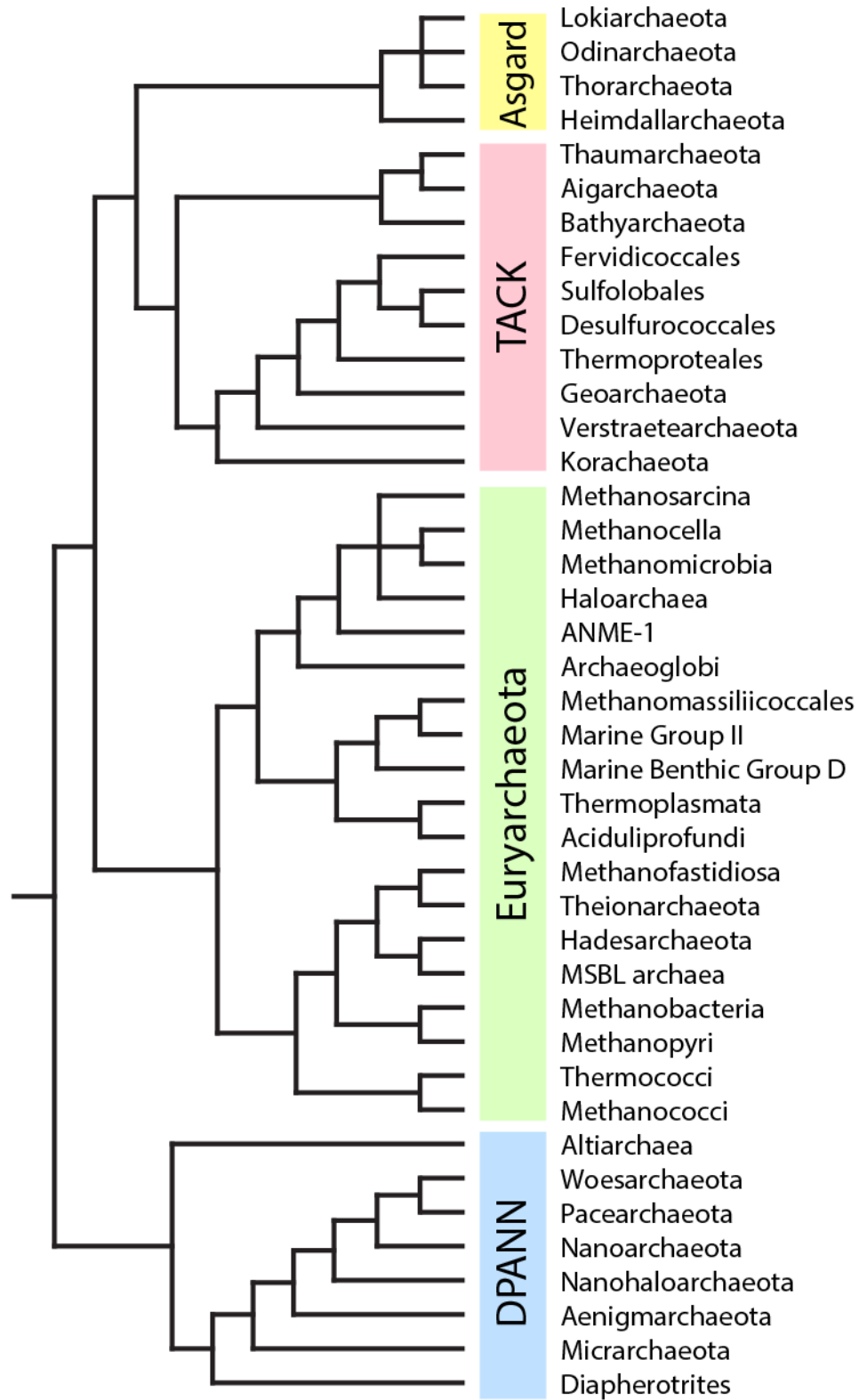


Figure 1: Cartoon representation of the phylogeny of archaeal lineages, adapted from Darnell et al<sup>15</sup>.

freshwater sediment<sup>13</sup>. These archaeal lifestyles are spread across currently known lineages of archaea, which are divided into 4 major superphyla<sup>14</sup> (Asgard, TACK, DPANN, Euryarchaeota; **Figure 1**). The Asgard archaea are considered a sister lineage to the eukaryotes<sup>14</sup>. They have been characterised on the basis of metagenomic sequencing, and only one species has been successfully cultured<sup>15</sup>. A number of genes previously thought to be eukaryote-specific have been discovered in many lineages within the Asgard archaea, including some genes that code for ribosomal proteins, certain RNA and DNA polymerases, ESCRT genes, actin and other cytoskeletal genes, and ubiquitination genes<sup>14</sup>. The TACK (Thaumarchaeota, Aigarchaeota, Crenarchaeota, Korarchaeota) superphylum<sup>16</sup> contains a number of well-studied species, including model species of the *Sulfolobus* genus, which belongs to the Crenarchaeota<sup>17</sup>. Many TACK species do not encode histones, though at least one species encoding histone genes has been found within each TACK lineage<sup>18,19</sup>. TACK archaea are generally monoploid<sup>20</sup>. The DPANN is a superphylum distinguished by small cell and genome sizes, which often are found in obligate symbiotic relationships with other archaea<sup>21</sup>. The Euryarchaeota are the best-studied phylum, and the majority of experimentally tractable species are members of this phylum. They contain methanogens, thermophiles, and halophiles. Most members of this phylum encode histone genes<sup>18</sup>. Many members, especially the halophiles, are polyploid<sup>22,23</sup>.

## **1.2 Halophilic archaea, members of the Euryarchaea, are excellent model systems for genetic and genomic studies**

Halophilic archaea, or halophiles, belong to the order Halobacteriales within the Euryarchaeota<sup>24</sup>. As their name suggests, they are adapted to hypersaline environments. A major adaptation to the high (2-5M Na<sup>+</sup>) extracellular salt concentrations is the “salt-in” strategy, where halophiles balance the external osmotic pressure with up to 3.6M K<sup>+</sup> ions in the cytoplasm<sup>25,26</sup>. The resultant highly ionic cytoplasm necessitated an evolution of an acidic proteome in these species to enable solubility – halophilic proteins boast a highly negative surface charge, with an over-representation of acidic residues<sup>25</sup>. Hence, halophilic enzymes and DNA-binding proteins require a suitably high salt concentration to carry out their functions, since these proteins are inactivated and can be denatured in reduced salt concentrations<sup>27,28</sup>. Within the hypersaline ecosystems of solar salterns and high-salt lakes, halophiles are exposed to other stresses like low and fluctuating availability of nutrients, temperature and pH fluctuations, UV radiation, and cycles of desiccation and rehydration<sup>29-31</sup>. Halophilic archaea have adapted to these stressors with a number of strategies, as detailed below.

The dynamic coordination of a response to multiple stressors in halophiles has led to an extensive network of transcriptional regulation, coordinated by over 70 known transcription factors (TFs)<sup>32</sup>. Specific stresses (like varying salt levels, heavy metal concentrations, or oxidative stress) are responded to by global changes in the activity of relevant transcription factors, which activate and repress a number of target genes, including other transcriptional regulators<sup>33-35</sup>. Together, these induce an overall

transcriptional response in a manner similar to the well-characterised eukaryotic stress response, with repression of genes involved in growth-related processes<sup>36</sup>. However, it remains unclear how chromatin proteins such as histones are involved in this dynamic network.

Other adaptations include those at the genomic structural level: polyploidy and high G+C content. A characteristic feature of the halophiles is that they carry a very high number of genome copies, for example, *Halobacterium salinarum* has a ploidy of ~25 during exponential growth and ~15 during stationary phase<sup>37</sup>. A number of explanations, linked with the stressful environment, have been proposed for this. Having multiple copies can allow easy repair of damaged DNA via homologous recombination; multiple copies can facilitate heterozygosity and hence potential for adaptation to stress; and, the genome can act as a reservoir of excess phosphate to be utilized during nutrient stress conditions<sup>23</sup>. Halophilic genomes have a high GC content, almost always above 60%<sup>38</sup>. It has been hypothesized that this is a protection against photodamage, due to the reduced frequency of radiation-susceptible T-T dimers in the genome<sup>38</sup>.

Halophiles are excellent extremophilic model organisms for understanding the molecular biology, genetics, and genomics of stress response, especially in comparison to other extremophiles. As facultative aerobes, they can readily be cultured aerobically in rich media (with the appropriate addition of 2-5M NaCl and other salts) at familiar temperatures for microbiology (37<sup>o</sup> -42<sup>o</sup>C). Their generation time of 3-6 hours is

relatively fast compared with other archaeal model organisms<sup>39</sup>. They do not require specialized equipment for laboratory cultivation, unlike thermophiles or specialized nutrient media and strict anaerobic conditions like methanogens. The first fully sequenced archaeal genome was that of a halophile, *Halobacterium salinarum*<sup>40</sup>. Since then, *Haloferax volcanii*<sup>41</sup> and four other halophiles have been fully sequenced, and are also used as model species in the lab. Like bacteria, they lack membrane-bound organelles and have circular genomes, rather than linear chromosomes<sup>42</sup>. Halophilic genomes are typically a few (2-4) megabases in length<sup>40,41</sup>, like those of bacteria, shorter than eukaryotic genomes. However, these genomes often have multiple origins of replication, as opposed to the single origin of bacterial chromosomes<sup>43</sup>.

Several halophilic species are highly experimentally tractable, with complete genetic toolkits (sequenced genome, ability to make knockouts, availability of expression vectors, etc.). Model halophiles include *Hbt. salinarum*<sup>44</sup>, *Hfx. volcanii*<sup>45,46</sup>, *Haloarcula hispanica*<sup>47</sup>, and *Haloferax mediterranei*<sup>47</sup>, all of which are currently used for comparative genetic and genomic studies in the Schmid lab. Halophiles are the only group of archaea for which so many experimentally and genetically tractable model species are readily available. The global transcriptional response of *Hbt. salinarum* to a variety of stressors has been well studied<sup>32</sup>, as have the knockout phenotypes of many transcription factors<sup>48</sup>, and high-throughput genomics methods such as transcriptomics (RNA-seq), genome-wide protein-DNA binding assays (ChIP-seq), and whole genome resequencing are well developed<sup>49</sup>.

Hence, halophiles are excellent and convenient model systems, especially for studying stress response and transcriptional regulatory networks. Having more than one model species within the clade allows comparison of a particular gene or the overall gene network between two species, both adapted to similar conditions but with significant evolutionary distance between them. They can also be used as models for understanding cell biological processes like cell division, away from the well-studied bacterial and eukaryotic models. Finally, as discussed in subsection 1.4, the role of histone proteins and the nature of chromatin in halophilic archaea remains an open question.

### **1.3 Specialized DNA binding proteins compact or organize genomes throughout the tree of life.**

Species across the tree of life use proteins to compact and organize their genomes<sup>50,51</sup>. A major reason for compaction across all life is that the uncompact genome is far larger than the cell (or nucleus) within which it is contained<sup>50</sup>. Within eukaryotes, the four core histones H2A, H2B, H3, and H4 are almost universally present architectural proteins that wrap the DNA into nucleosomes, the “beads-on-a-string” visible with electron microscopy<sup>50,52</sup>. These proteins are basic, facilitating their electrostatic interaction with negatively charged DNA<sup>53</sup>. They contain a conserved histone fold motif, which facilitates histone-histone dimerization, as well as histone-DNA binding<sup>54</sup>. This fold domain, consisting of three alpha helices and two loop domains, allows the formation of (H3-H4) and (H2A-H2B) heterodimers, which form the octamer around which ~147bp of DNA wraps to form a nucleosome<sup>52</sup>. Eukaryotic histones also contain a N-terminal tail, a target for post-translational modifications. These core histones are responsible for

genome condensation and decompaction, and hence also regulate DNA accessibility, gene expression and DNA replication<sup>52,55</sup>. Post-translational modifications can regulate gene expression by modulating histone-DNA binding, which is extensively studied as an epigenetic mechanism of gene regulation<sup>56</sup>. In addition to the core histones that form nucleosomes, eukaryotic DNA is also bound by other histone variants. Linker histone H1 (which, despite its name, lacks a histone fold domain), binds and stabilizes the nucleosome and surrounding “linker” DNA, and contributes to the formation of higher-order chromatin structures<sup>57</sup>. Other common eukaryotic histone families include derivatives of the core histones like CenH3 which is specialized to bind to centromeres, and H2A.Z which is involved in RNA polymerase recruitment for transcription initiation<sup>52</sup>.

In bacteria, the architectural proteins involved in genome compaction are far more varied, with involvement of multiple protein families and with differences between species and clades<sup>58,59</sup>. Bacteria lack histones and also lack “beads-on-a-string” nucleosomes; instead, they typically use a number of small, basic proteins collectively known as nucleoid-associated proteins (NAPs) to organize their genomes<sup>58,59</sup>. The model species *E. coli* contains multiple NAPs, including H-NS, Fis, HU, and IHF, as well as proteins from the Lrp and Dps families, all of which contribute to genome compaction and organization in a growth-phase dependent manner<sup>50,58-60</sup>. While histones wrap DNA, different NAPs interact with DNA differently: H-NS bridges different DNA strands; HU, IHF, Fis, and Dps bend DNA; while Lrp proteins wrap DNA<sup>50,58</sup>. Bending of the

genome by HU is necessary for the negative supercoiling that compacts the bacterial genome<sup>59,61</sup>. Together the bacterial NAPs dynamically determine the local structure of DNA and its accessibility<sup>58</sup>, and also regulate gene expression and DNA replication<sup>59</sup>.

Like bacteria, archaea also use a variety of chromatin proteins to organize their genomes. Histone proteins, containing a histone fold domain similar to that of eukaryotes have been found in species from all known archaeal superphyla (**Figure 1**)<sup>62,63</sup>. Indeed, the evolutionary origin of the histone fold domain has been traced to the Archaea<sup>62,64</sup>. Like eukaryotic histones, these proteins form dimers and wrap DNA into nucleosomes<sup>65</sup>, which are visible as “beads-on-a-string”<sup>66,67</sup>. Moreover, recent work has shown that, rather than eukaryotic nucleosomes of ~147bp formed by a core histone octamer, archaeal histones can polymerize and form extended structures called hypernucleosomes<sup>18,63,68</sup>. While most archaeal histones lack the N-terminal tails seen in eukaryotes and contain only the fold domain, the Asgard archaea do have N-terminal extensions, and the sequence characteristics of some of these resemble eukaryotic N-terminal tails<sup>63</sup>. A more detailed summary of the current state of knowledge regarding archaeal histones is given in subsection 1.4.

Many proteins other than histones are also involved in organizing archaeal chromatin, often with more than one chromatin protein present in the same organism (**Table 1**).

Alba is a protein family present in all archaeal lineages (except the Halobacteria and the genus *Methanosarcina*<sup>18</sup>), which can bind DNA in a sequence-independent manner via its positively-charged residues. Alba plays a role in chromatin organization of

thermophilic archaea, and regulation of gene expression at both the transcriptional and post-transcriptional level (the latter due to its ability to also bind RNA)<sup>69</sup>. Crenarchaea, which largely lack histones<sup>63</sup>, utilize Alba as well as some crenarchaeal-specific proteins from the Cren7 and Sul7d families (**Table 1**). Both are small, basic proteins that induce sharp kinks in DNA and can compact it at high concentrations by bending and bridging interactions<sup>18,70,71</sup>. TrmBL2 is a helix-turn-helix protein homologous to the TrmB transcription factor family, which in some Euryarchaeal species forms thick fibrous chromatin structure and also acts as a transcriptional repressor<sup>67,72</sup>. Yet other archaea possess bacterial-like chromatin proteins, like HTa, the HU family protein in *Thermoplasma acidophilum*<sup>73,74</sup>. Mc1 is an archaeal-specific protein present in the phyla Methanosarcina and Halobacteria<sup>75</sup>. It has been well-characterised in Methanosarcina, where it was found to be an architectural DNA-binding protein<sup>76,77</sup>. It binds by inducing kinks in DNA and is predicted to be involved in DNA compaction<sup>78,79</sup>.

Across all the domains of life, higher-order chromatin structures have been observed. In eukaryotes, these are assemblages of nucleosomes, forming fibers ~30nm in diameter, and ultimately resulting in the compact chromosomes visible during mitosis<sup>80</sup>. In bacteria, segments of the genome that are bridged or bent by the NAPs discussed above are further organized into chromosome interaction domains (CIDs), at a scale of tens or hundreds of kilobases<sup>59,81,82</sup>. Higher-order interaction domains have also recently been observed in archaeal species from both Crenarchaea and Euryarchaea<sup>83,84</sup>. In all three domains, proteins from the SMC family have been found to play a critical role in forming and regulating these higher-order structures<sup>80,83-86</sup>. In summary, protein-based

genome compaction is a feature shared among all species on the tree of life, although the specific molecular mechanisms that bring about compaction are highly diverse. Eukaryotes rely on histone proteins to compact DNA into well-defined nucleosomes, while in bacteria a diverse array of NAPs bend and bridge DNA into a compact chromosome. Many archaea show eukaryote-like chromatin with the formation of histone-based nucleosomes, but also contain a diverse array of NAPs which compact DNA instead of or in competition with histones (**Table 1**). In addition, higher-order chromatin is seen in all three clades, and is linked with SMC family proteins.

**Table 1: Summarized version of the diversity of chromatin proteins**

Protein or protein family	Lineages	Mode of binding
Histones	Eukaryotes, Archaea	Wrapping/nucleosome formation <sup>68,87</sup>
HU*, IHF, Fis	Bacteria	Bending <sup>50</sup>
H-NS	Bacteria	Bridging <sup>50</sup>
Lrp	Bacteria, Archaea	Wrapping <sup>50</sup> , bending <sup>88</sup>
ALBA**, TrmBL2	Thermophilic archaea	Protein-DNA fibres <sup>67,69,89</sup>
Mc1, Cren7/Sul7d	Some archaea	Bending, kinks in DNA <sup>70,71,90</sup>
*HU family proteins also present in some Archaea		
**ALBA present across all domains of life, but its DNA architectural role only characterised in thermophilic archaea; elsewhere its RNA-binding is characterised.		

#### **1.4 Characterised archaeal histones compact DNA and form nucleosomes, with notable exceptions.**

The histone fold domain has its evolutionary origin in the domain Archaea<sup>62</sup>. As discussed above, archaeal histone proteins contain a fairly well-conserved histone fold, but typically (outside the Asgard archaea) do not contain the N-terminal tails seen in eukaryotic histones<sup>62,64</sup>. The archaeal histones of some model species, particularly those

of thermophilic archaea, have been well studied for over three decades, while in other clades research is newer and less comprehensive, including halophilic archaea, the subject of focus here.

The pioneering work on archaeal histones was carried out by Kathleen Sandman and John Reeve in 1990<sup>66</sup>. This initial study isolated the histone protein of *Methanothermus fervidus* and showed that it was capable of compacting DNA *in-vitro*: it formed visible beads-on-a-string structures with DNA, and the compacted DNA-histone complex was able to migrate faster through a gel than naked DNA. Following this work, the histones of *M. fervidus* and *Thermococcus kodakarensis*, both thermophilic Euryarchaea, became molecules of focus in archaeal histone research<sup>91,92</sup>. Both species contain two histone genes, and the resultant proteins were shown to form both homo- and hetero-dimers capable of binding DNA<sup>91-93</sup>. Digestion of DNA by micrococcal nuclease (MNase) revealed a ladder pattern characteristic of nucleosome-bound DNA, but with a 30- and 60-bp increment, rather than the 147bp fragments seen in eukaryotes<sup>94,95</sup>. The MNase fragments obtained *in-vitro* with a recombinant histone-DNA complex were shown to be identical to those obtained with an *in-vivo* digestion, strongly suggesting that nucleosomes within these species are indeed formed by histone proteins<sup>95</sup>.

The X-ray crystal structure of the homodimers of the *M. fervidus* histones was solved, and found to be similar to the H3-H4 dimers seen in eukaryotic nucleosomes<sup>65</sup>.

However, in place of the N-terminal tail, these histones contain a proline tetrad that facilitates DNA binding<sup>62,65</sup>. More recently, crystal structures of archaeal histones with

DNA have been obtained<sup>68</sup>. While these do show a basic nucleosome similar to the eukaryotic nucleosome, they also demonstrated the formation of extended hypernucleosome structures<sup>68</sup>, which were shown to be facilitated by stacking interactions between neighbouring histones<sup>96</sup>. These structural results corroborate and clarify previous results obtained with nuclease digestion, showing the presence of protected fragments whose length was equal to many multiples of the 60bp basic nucleosome<sup>97</sup>.

Other studies have focused on the differences between the two histone genes within each species of thermophilic archaea. In *M. ferrovidus* cells with both copies present, it was found that different paralogs are expressed at different phases of growth<sup>91</sup>. It was also observed that while deletion of a single gene was viable, a double deletion (thus removing all histone genes from the cell) was impossible<sup>92</sup>. Indeed, in laboratory strains of *Methanothermobacter thermoautotrophicus*, one of the two histone paralogs had spontaneously mutated and lost its DNA-binding ability as a homodimer, but retained the ability to form histone-histone heterodimers that could compact DNA<sup>98</sup>. These results suggest that, at least to some extent, one of the histone genes can substitute for the other. In summary, work done largely on *M. ferrovidus* and *T. kodakarensis*, two evolutionarily diverse archaea, both adapted to high temperatures, revealed that they usually encoded two histone genes whose protein products performed the work of DNA compaction.

In recent years, there has been growing research interest on archaeal histones besides those of *M. fervidus* and *T. kodakarensis*. *Methanopyrus kandleri*, another thermophilic Euryarchaeon, contains a lone doublet histone gene- two histone fold domains in the same peptide<sup>99</sup>. This fused heterodimer also demonstrated DNA compaction via increased mobility of the DNA-protein complex<sup>99,100</sup>. Its X-ray crystal structure revealed a dimer of the fused heterodimer protein, analogous to (H3-H4)<sub>2</sub> tetramer seen in eukaryotes<sup>101</sup>. Interestingly, this protein lacks the proline tetrad present in other archaeal histones but absent in eukarya<sup>101</sup>. It has been suggested that it is the single fused heterodimer histone gene of *M. kandleri* that allowed for less constrained mutation and hence evolution in one of the two histone fold domains<sup>62</sup>. In summary, even though *M. kandleri* differs from *M. fervidus* and *T. kodakarensis* by having only one histone gene, it was found to be similar to them in the sense that it is another high temperature-adapted archaea whose histone forms nucleosome structures with DNA.

In sharp contrast to *M. kandleri* and its single histone gene, some species contain several histone proteins. Computational analysis of the seven histone paralogs of *Methanosphaera stadtmanae* showed that they vary sharply in their DNA binding affinity and ability to form histone-histone dimers, and their expression was found to be proportional to their ability to form dimers. Hence, it was proposed that different paralogs play very different functional roles, including as capstones inhibiting hypernucleosome formation, in a manner more similar to the specialized role for some histone variants seen in eukaryotes<sup>102</sup>. *M. stadtmanae* is a mesophile, unlike the

previously described species, but as described above, computational analysis has revealed that its histone genes too are capable of performing DNA compaction.

So far, all the histones discussed here have shown an ability to wrap and compact DNA similar to eukaryotic histones. Deletion of all the histone genes from a species was found to be impossible, corroborating their essential role in the fundamental process of genome compaction. However, in *Methanosarcina mazei*, a mesophilic euryarchaea, deletion of the single histone gene was found to be viable; indeed, there was no growth defect observed in optimal conditions<sup>103</sup>. However, the histone deletion strain was sensitive to UV stress and resulted in decreased transcription of ~25% of genes<sup>103</sup>. As discussed in section 1.3, the NAP Mc1 is suggested to play an important role in the chromatin of *Methanosarcina*<sup>79</sup>, explaining possibly why the histone is not essential.

Halophilic archaea all contain a single fused heterodimer histone gene (with two histone fold domains), like in *M. kandleri* discussed above<sup>62,104</sup>. Unlike all other histones, but similar to all other haloarchaeal proteins, halophilic histones are negatively charged<sup>104</sup>. In an early study, the chromosomal structure of *Hbt. salinarum* was isolated, suggesting that protein-bound “chromatinized” DNA resembles beads-on-a-string under high salt conditions *in vitro*<sup>105</sup>. However, naked DNA also adopts a beads-on-a-string conformation under high salt conditions *in vitro*, and DNA is known to be strongly condensed in the presence of cations<sup>106,107</sup>. In a more recent study in another model halophile, *Hfx. volcanii*, researchers observed protection of 30- and 60-bp fragments from MNase digestion, and interpreted this as histone-based nucleosome formation

somewhat similar to observations in other archaea with characterised histones<sup>108</sup>.

However, the link between the histone protein and the observed nuclease digestion pattern was not established, and alternative explanations, including the presence of a different chromatin protein, are possible. Hence, these studies left an open question about the chromatin structure and histone function in a high salt cytoplasm.

To test the hypothesis that halophilic histone plays a role in DNA compaction, prior work from the Schmid lab showed that *hpyA*, the histone gene of the model halophile *Hbt. salinarum* can be deleted with no growth defect<sup>104</sup>. This surprising result was also observed in stress conditions, where the histone deletion mutant grew identically to the parent strain in UV and oxidative stress, growth with bacitracin (an antibiotic that targets the cell wall), growth with novobiocin (antibiotic that targets DNA replication), and with mechanical stress. Moreover, mass spectrometry-based protein identification from chromatin enrichments as well as quantitative proteomics<sup>109</sup> revealed that HpyA as well as other putative chromatin proteins (Mc1, DpsA) were expressed at levels too low under standard laboratory growth conditions to perform the role of DNA compaction. The phenotypes that were observed for the deletion mutant included growth phase-dependent changes in cell morphology (from cylindrical to rounder cells) and in gene expression (subtle, bidirectional changes in <10% of the transcriptome). Therefore, we suggest that the prior studies were not detecting histone-based nucleosomes in halophilic archaea.

Summarizing, the majority of characterised archaeal histones, many from thermophilic archaea, perform the familiar function of genome compaction by wrapping DNA.

However, histone proteins encoded in some archaea, particularly halophiles, do not follow this trend. The precise function of the histone in the high salt biochemical environment of the halophile cytoplasm remains unclear and was therefore the focus of the current study.

## **1.5 Open questions on archaeal chromatin motivate this thesis work**

The current state of knowledge regarding histones and chromatin in halophilic archaea, summarized in subsection 1.4, leads to several unanswered questions. If histones are not essential and possibly do not perform the role of genome compaction, what is their role in these species, given that they are highly conserved within the halophiles?

I address this question by testing the hypothesis that it is the hypersaline environment of halophiles that has selected for an alternative histone function, and that halophilic histone instead functions as a transcription factor. I arrived at this hypothesis by combining several observations noted above: (1) salt plays a vital role in DNA compaction<sup>106</sup>; (2) *hpyA* of *Hbt. salinarum* is non-essential<sup>104</sup> (unlike most histone proteins); (3) HpyA is expressed at very low levels<sup>104</sup>. I tested this hypothesis with growth assays for the  $\Delta hpyA$  strain in reduced salt, and found a slight but significant defect in growth rate. This was corroborated by microscopy, showing increased circularity for  $\Delta hpyA$  in reduced salt. With a phenotype established, I carried out ChIP-Seq to characterise the genome-wide DNA binding locations for HpyA. HpyA binds in

discrete peaks, with a strong effect of external salt concentration and growth phase on the number of peaks. The functional role of HpyA was further explored with RNA-Seq. This transcriptomic data revealed that HpyA is involved in the direct regulation of iron transport genes, as well as indirect regulation of ion transport, nucleotide metabolism, and DNA repair and replication genes, in a salt-dependent manner. These results have been published<sup>110</sup> and constitute Chapter 2 of this thesis.

In Chapter 3, I explore how the genome-wide binding patterns of halophilic histones compare to those of other DNA-binding proteins, and attempt to classify halophilic histones based on their binding patterns. The scope of this study was expanded by including data from HstA, the sole histone of *Hfx volcanii*, which allows the results to be more generalizable across the haloarchaea. Just like *hpyA*, *hstA* was found to be a non-essential gene. However, it does show a growth defect in optimal conditions, and does not appear to have a salt-linked phenotype. Its genome-wide binding pattern, observed using ChIP-Seq, was similar to that of HpyA – it formed discrete peaks. These ChIP-Seq results were compared with ChIP-Seq data from bacterial NAPs and TFs, haloarchaeal TFs, as well as eukaryotic histones. In terms of peak width and number of peaks, halophilic histones most closely resemble TFs. However, their occupancy across gene start sites lacks both the characteristic patterns of eukaryotic histones and TFs. Indeed, they bind equally in promoter and coding regions, a trait seen in some NAPs. In terms of DNA sequence motifs, there appear to be interspecies differences, with HstA preferentially binding a TF-like palindromic motif, while HpyA preferentially binds sequences with 10bp periodicity of certain dinucleotides, a property of eukaryotic and

archaeal histones. Hence, the collective evidence puts halophilic histones in the blurred region between transcription factors and chromatin proteins.

In Chapter 4, I discuss the technical challenges in obtaining the data discussed above, in particular the removal of ribosomal RNA before carrying out RNA-Seq. As discussed in subsection 1.1, archaeal mRNA lacks the poly-A tail of eukaryotic mRNA, hence, selectively enriching mRNA from a pool of total RNA (~95% rRNA) is difficult.

Retention of high rRNA in the sample leads to poor sequencing depth of the target mRNA. In collaboration with Mar Martinez-Pastor, I tested various rRNA removal kits in model halophilic species, and analysed their removal efficiency. We also tested competing hypotheses to explain the poor efficiency of certain kits, and found that the choice of kit does not bias the observed gene expression. Finally, we quantified the advantage of efficient rRNA removal in terms of better coverage of low-expression genes.

These results together suggest a unique role for halophilic histones, regulating transcription with TF-like direct regulation as well as indirect regulation, and with binding characteristics that are a mixture TF-, NAP-, and eukaryotic histone-like features. I conclude from data presented here that the hypersaline environment has selected for an alternative function for the halophilic histone, as evidenced by the major changes to HpyA binding and gene regulation in reduced salt conditions.

## 2. An archaeal histone-like protein regulates gene expression in response to salt stress

This work is adapted from a manuscript published in *Nucleic Acids Research*<sup>110</sup>. The authors are Saaz Sakrikar and Amy Schmid. S.S. and A.K.S designed the project. S.S. carried out experimental work under supervision of A.K.S.; S.S. and A.K.S analyzed the data and wrote the manuscript.

### 2.1. Introduction

Phylogenetic analysis has shown that the histone fold domain originated in the Archaea<sup>62,64,111</sup>. Histone proteins play a vital role in genome compaction and regulation of gene expression in eukaryotes<sup>52</sup>. The four core eukaryotic histones (H3, H4, H2A, H2B) share a histone fold domain, which is involved in histone dimerization and DNA-binding<sup>54,112,113</sup>. Proteins containing the histone fold are present in all known major archaeal lineages<sup>18</sup>. Archaeal histone-like proteins have been most extensively characterized in species representing the euryarchaeal superphylum, with much work focusing on the thermophilic archaeal species *Methanothermus fervidus*<sup>66,91</sup> and *Thermococcus kodakarensis*<sup>92,97</sup>. *In vitro* structural studies from these species demonstrate strong conservation between archaeal and eukaryotic histones in terms of histone fold, multimeric protein structure, and DNA wrapping<sup>65,68,96</sup>. However, key differences from eukaryotes have also been noted<sup>18,63</sup>: archaeal histones form extended polymeric structures called hypernucleosomes<sup>68,96,97</sup>. These structural data help explain results from *in vivo* data. DNA digests with MNase yield fragments in multiples of 30-60 bp<sup>94,97,114</sup> and gene expression is significantly altered by histone binding<sup>115</sup>. Like eukaryotic

histones, archaeal histones can also hinder elongation<sup>115,116</sup> or inhibit the binding of site-specific transcription factors (TFs) through competition<sup>117</sup> to influence global transcription levels. These histones may act as the major chromatin protein. These studies led to the oft-noted hypothesis that many features of archaeal histone structure and function resemble those of eukaryotes in terms of genome compaction and gene expression, with some key differences<sup>18,97</sup>.

Recent evidence in other model systems call for further testing of this hypothesis. For example, a deletion mutant of the sole histone of *Methanosarcina mazei* was viable, but exhibited reduced growth when exposed to radiation<sup>103</sup>. Phylogenetics and molecular dynamics simulations in other model methanogens that encode multiple histone variants suggest various functions in the chromatin environment<sup>102</sup>. In other model species, a dynamic variable set of bacterial-like chromatin and archaeal-specific proteins rather than histones organize the genome<sup>72,74</sup>. Phylogenetic and proteomics evidence in hyperthermophiles suggests that chromatin compaction allows DNA stability to prevent unwanted transcription by promoter melting at high temperature<sup>118</sup>. This hypothesis has been substantiated *in vitro*: plasmid DNA dissociation is prevented by histone binding at 90°C<sup>119</sup>. Histone point mutants that cannot compact DNA exhibit differential expression of specific genomic regions<sup>114</sup>. Previous work from our group demonstrated an alternative regulatory function for HpyA, the sole histone of the hypersaline-adapted species *Halobacterium salinarum*<sup>104</sup>. HpyA is dispensable for cell viability but important for maintaining wild type gene expression and cell shape under optimum growth

conditions. HpyA protein levels were too low to facilitate genome-wide DNA compaction<sup>104</sup>. Together these findings reveal an expanding landscape of diverse histone functions across archaeal lineages selected for by the diverse and sometimes extreme environments of archaea. However, the function of histone-like proteins in hypersaline-adapted archaea remains understudied relative to other archaeal lineages. Halophilic archaea have adapted to survive extreme osmotic pressure (up to 5M NaCl) in their natural salt lake environments by counterbalancing with up to 4M potassium ions in the cytoplasm<sup>120</sup>. Due to the resultant highly ionic cytoplasm, haloarchaeal proteins, including the histone protein, have evolved a negatively charged surface<sup>40</sup>. This is in contrast to all other known species, where the positively charged surface of histones facilitates DNA-histone interactions<sup>64,104</sup>. It has previously been observed *in vitro* that naked DNA under moderate salinity tends to spontaneously form structures similar to the beads-on-a-string observed with histone-bound DNA, with increasing compaction and even aggregation occurring at higher concentrations<sup>106</sup>. These data call into question the need for protein-based genome compaction. In addition, proteomics data from our prior work demonstrated that the protein levels of HpyA are very low in *Hbt. salinarum*, and HpyA expression levels change little throughout growth<sup>104</sup>. Based on these results and given the unusual chemistry of the haloarchaeal saturated salt cytoplasm, here we hypothesize that the non-canonical function of HpyA in gene regulation is linked to the unique hypersaline cytoplasmic environment of *Hbt. salinarum*.

We tested this hypothesis using a battery of *in vivo* quantitative phenotyping and functional genomics assays. Growth rate and cell morphology in low sodium was

affected in the  $\Delta hpyA$  deletion strain, confirming an association between the presence of the gene (and its product) and the effects of low sodium concentration. Protein-DNA binding assays (ChIP-seq) revealed reproducible, salt-dependent, genome-wide binding of HpyA at nearly 60 discrete sites -- a binding pattern too sparse to coat or compact the genome. However, the high prevalence of binding within gene bodies suggests that the mechanism of regulation differs substantially from that of canonical TFs. Integration of DNA binding with transcriptomics data revealed direct regulation of iron uptake by HpyA. Global, indirect regulation of transport of other ions, biosynthesis of purines, and DNA replication and repair was also observed. Together, these results suggest that HpyA functions as a specific, direct transcriptional regulator of metal ion balance. HpyA thereby maintains growth rate and rod-shaped cell morphology during hypo-osmotic stress.

## **2.2. Materials and Methods**

### **2.2.1 Strains, media and general culturing:**

Strains used in this study have been described in Dulmage et al 2015<sup>104</sup>, summarized in **Appendix A**. All strains were constructed from a *Halobacterium salinarum* NRC-1 background with *ura3* (encodes uracil biosynthesis functions) gene deleted to enable uracil counterselection<sup>44</sup>. Growth assays were carried out using strain MDK407 ( $\Delta ura3$ ) as the parent strain (control, referred to here as wild type, or WT) and KAD100 ( $\Delta ura3\Delta hpyA$ ) as the  $\Delta hpyA$  deletion strain.

For ChIP-seq, strains carrying the *hpyA* gene tagged at its C-terminus with the hemagglutinin (HA) epitope were used<sup>104</sup>. The control strain was AKS134 ( $\Delta hpyA$  deletion carrying the empty vector pMTFCHA). The experimental strain was KAD128, which contained the pKAD17 plasmid expressing HpyA-HA driven by its native promoter (primers and plasmids given in **Appendix A**)<sup>121</sup>. pKAD17 was generated by: (a) insertion of *hpyA* into the pMTF-cHA plasmid upstream of the HA tag sequence (between the NdeI and HindIII restriction sites); and (b) replacement by isothermal ligation of the  $P_{idx}$  promoter of the plasmid with the  $P_{rpa200}$  native promoter sequence of HpyA at the KpnI site.

The media used for all experiments was *Hbt. salinarum* complete media (CM) containing 250g/L NaCl, 20 g/L MgSO<sub>4</sub>•7H<sub>2</sub>O, 3g/L trisodium citrate dihydrate, 2g/L KCl, 10g/L Bacteriological peptone (Oxoid). pH was adjusted to 6.8. Media were supplemented with 50 µg/mL uracil to compensate for the uracil auxotrophy of  $\Delta ura3$  parent and derivative strains. Reduced salt media was made identically except for NaCl, which was reduced to 199 g/L (3.4 M). For plasmid strains, 1 µg/mL mevinolin (AG Scientific) was added to liquid medium and 2.5 µg/mL to solid media to maintain selective pressure on the plasmid.

Cells were routinely streaked fresh from frozen stock onto solid medium. Individual colonies were picked from plates and inoculated into 5 mL CM (with additives when necessary) and allowed to grow for approximately 4 days at 42°C in a shaking incubator until stationary phase was reached. These starter cultures were diluted by sub-culturing

to OD<sub>600</sub> ~ 0.02 into 50 mL of media indicated in the figures and grown until harvesting as described below.

### **2.2.2 Growth and microscopy:**

For growth curve phenotyping, 9 biological replicates of  $\Delta ura3$  (MDK407) and  $\Delta hpyA$  (KAD100) strains were cultured in 125 mL flasks at 42°C in a shaking incubator. Optical density (OD) measurements were taken at time zero, then at 3-4 hour intervals following the initial lag phase of ~12 hours. Raw growth data are provided in **Appendix B**.

Resultant growth curves were fit by logistic regression to calculate the maximum instantaneous growth rate ( $\mu_{\max}$ ) using the R package *grofit*<sup>122</sup>. The code for analysis and visualization of these growth data are contained in

[https://github.com/amyschmid/HpyA\\_codes](https://github.com/amyschmid/HpyA_codes).

For microscopy, cultures of  $\Delta ura3$ ,  $\Delta hpyA$ , and  $\Delta hpyA$  / pKAD17 (strain KAD128) were each grown to mid-exponential phase. 8  $\mu$ l aliquots were placed on a thin, flat, agarose pad impregnated with 4.3M NaCl as described<sup>123</sup>. Cells were imaged at 100X using a Zeiss Axio Scope A1 microscope with a Pixelink PL-E421M camera. Images were analyzed for circularity using the MicrobeJ package within the ImageJ software<sup>124</sup>. In this context, circularity is defined as the measure of deviation from perfect circle, where 1 is a perfect circle and 0 is a polygon with 1 side infinitely longer than the other. Given that circularity distributions were skewed, adjusted bootstrap percentile corrected 95% confidence intervals were calculated by 1,000-fold ordinary non-parametric bootstrap

resampling of the median with replacement. The `boot()` package in the R coding environment was used for these calculations.

### **2.2.3 ChIP-seq experiments:**

One biological replicate colony of AKS134 (Empty vector control) and four replicates of KAD128 (expressing HpyA-HA) were cultured as described above. The 50mL cultures were grown in 125mL flasks and their growth was monitored by OD600 until the time for harvesting (exponential phase: 36-50 hours, OD~0.2-0.35, growth rate ~ 0.032 hr<sup>-1</sup>; stationary phase: ~70-140 hours, OD~1.4-1.7, growth rate ~0.017 hr<sup>-1</sup>). Strains were PCR-checked for the presence of the plasmid expressing *hpyA*-HA prior to each experiment (see **Appendix A** for primers).

Harvested cells (45mL) were immediately cross-linked using 1.4mL 37% formaldehyde (final concentration = 1% v/v) and immunoprecipitated using Abcam HA-specific antibody (catalog #ab9110) as described in Wilbanks et al, 2012<sup>49</sup>, with certain modifications to the protocol: the cross-linking reaction was allowed to proceed for 20 minutes, and cell pellets were resuspended in 800 µL lysis buffer. Resulting DNA was extracted with Phenol:Chloroform:Isoamyl alcohol (25:24:1) and then ethanol precipitation. Library preparation and single-end sequencing was carried out by the Duke Center for Genomic and Computational Biology Sequencing and Genomic Technologies core facility using the Illumina HiSeq4000 instrument.

#### 2.2.4. Analysis of ChIP-seq data:

Gzipped FastQ files (Accession: PRJNA703048, GEO: GSE182514) were analyzed using FastQC software. Information provided as input included read sequence quality, length distribution, and presence of adapters. Adapters were trimmed from the reads using Trim Galore!, and these trimmed sequences were aligned to the *Hbt. salinarum* NRC-1 genome (RefSeq ID GCF\_000006805.1, assembly ID ASM680v1) to generate a SAM file using Bowtie2 with default parameters. End-to-end alignment was suitable for trimmed reads<sup>125</sup>. FastQC and Trim Galore! are available online at <http://www.bioinformatics.babraham.ac.uk/projects/> (2015 version). The SAM files were converted to binary (BAM), sorted and indexed using SAMtools<sup>126</sup>. Sorted BAM files were used for peak calling. WIG files for easy visualization were also generated using SAMtools, with coverage recorded every 10 bp. All code used to analyze ChIP-seq data are available in File S1 at [https://github.com/amyschmid/HpyA\\_codes](https://github.com/amyschmid/HpyA_codes).

The sorted BAM files were used for peak-calling with MACS2<sup>127</sup> version 2.1.1 callpeak function. Parameters were: nomodel, qval=0.05 cutoff. Called peaks were combined across replicates using the *multiBedIntersect* function of the bedtools package<sup>128</sup>. Only peaks detected in at least two biological replicate experiments were kept in downstream analyses. Genes within 500 bp of these reproducible peaks were annotated using the IRanges package in R<sup>129</sup>. Resultant peaks were then manually curated to remove the following: (a) false positives caused by local variability in input control sequencing read depth; (b) local duplications and deletions associated with transposases and integrases;

(c) one peak that was also detected in the HA tag-alone input control; (d) peaks located nearby redundant genes. Details of the code and dependencies for the entire workflow for peak calling and visualization are noted in the github repository

[https://github.com/amyschmid/HpyA\\_codes](https://github.com/amyschmid/HpyA_codes).

Resultant peak regions (start to end of peak footprint) were then classified based on their genomic context (details in **Appendix B**). For this purpose, promoters were defined as the region from 500bp upstream of the translation start site [many halophile transcripts are leaderless<sup>130</sup>].

To classify binding peak center locations as “genic” or “intergenic” for the purpose of the intergenic test, the center of each peak (the mid-point between the peak start and stop determined above) was taken and classified as being within a gene (“coding”) or not (“intergenic”). The code used to make this classification is in

[https://github.com/amyschmid/HpyA\\_codes](https://github.com/amyschmid/HpyA_codes) and the results are shown in **Figure 7**.

To classify binding locations “genic” or “promoter”, the number of bp in the overlap between the ChIP-seq peak chromosomal coordinates and the genomic feature was calculated. If the peak overlapped both a genic and a promoter feature, the peak was classified as located within the feature with the largest overlap. Both features were counted in the case of ties. The code used to make this classification is in

[https://github.com/amyschmid/HpyA\\_codes](https://github.com/amyschmid/HpyA_codes). Operons were computationally predicted using the Operon-Mapper tool<sup>131</sup> and integrated with empirical predictions from Koide et al<sup>130</sup>. Classification of TrmB binding locations are given directly in reference <sup>132</sup> and

significance of enrichment was computed using the hypergeometric test in R.

Classification and computation of enrichment *p*-values for RosR binding locations [from Tonner et. al.<sup>133</sup>] were computed using BEDtools “fisher” function<sup>128</sup>.

### **2.2.5. RNA-seq experiments:**

Six biological replicate cultures of strains MDK407 (parent) and KAD100 ( $\Delta hpyA$ ) were cultivated as described above in either optimal salt (4.2M NaCl) or low salt (3.4M NaCl) media. Growth was monitored using OD600 until harvesting (exponential phase was defined as: ~31-34 hours of growth, OD~0.1-0.4 depending on the strain and medium).

A 4.2mL aliquot of each culture was removed and centrifuged for 30s at 21,000 x g in an Eppendorf tabletop centrifuge. The supernatant was discarded and the cell pellet was immediately plunged into liquid nitrogen and stored 1-7 days at -80°C. Extraction of RNA from these pellets was carried out using the Agilent Absolutely RNA Miniprep kit following the manufacturer’s protocol, with an extended on-column DNase incubation of 45-60 min. Resultant RNA samples were checked for: (a) genomic DNA contamination using PCR with 200ng input RNA and 35 amplification cycles using primers listed in **Appendix A**; (b) concentration using 260/280 nm ratio in a Nanodrop spectrophotometer; (c) quality using the Agilent Bioanalyzer RNA Nano 6000 chip (RNA Integrity Number (RIN) > 9.0). For each strain and condition, rRNA was removed from 3 replicates with NEBNext Bacteria rRNA Depletion Kit (New England Biolabs), while the other 3 were treated with NEBNext Depletion Core Reagent Set using custom probes targeted to *Haloferax volcanii* rRNA (Martinez-Pastor and Sakrikar, details in

Chapter 4). These custom probes were designed using the NEBNext Custom RNA Depletion Design Tool (<https://depletion-design.neb.com/>). rRNA depletion was verified using the Bioanalyzer RNA chip. The NEBNext Ultra II Directional RNA Library Prep Kit for Illumina was used for preparing sequencing libraries, and cDNA libraries were quality-checked using the High-Sensitivity DNA Bioanalyzer chip. Paired-end sequencing was carried out at the Duke Center for Genomic and Computational Biology Sequencing core facility using the Novaseq6000 instrument (Illumina).

### **2.2.6. RNA-seq data analysis:**

For analysis of sequencing data, paired FastQ files were trimmed and checked for quality using Trim Galore! (<http://www.bioinformatics.babraham.ac.uk/projects/>) and aligned to the genome using Bowtie2<sup>125</sup>. SAMtools was used to generate, sort, and index BAM files<sup>126</sup>. The count function of HTSeq<sup>134</sup> was used to create a file assigning the number of reads to each gene (see **File S1** within the Github repository for details). Outlier samples were removed from further analysis using Strong PCA<sup>135</sup> ([https://github.com/amyschmid/HpyA\\_codes](https://github.com/amyschmid/HpyA_codes)). The R package DESeq2<sup>136</sup> was used to normalize counts and batch correct across replicates for each strain and genotype (using DESeq2 default parameters). Significant differential gene expression analysis using DESeq2 applied three pairwise contrasts:  $\Delta hpyA$  vs WT in optimal salt,  $\Delta hpyA$  vs WT in reduced salt, and reduced vs optimal salt in a WT background. For each contrast, reproducibility and quality was checked across replicates using dispersion, MA, and volcano plots. For each contrast, Benjamini-Hochberg (BH) adjusted<sup>137</sup> Wald test  $p < 0.05$

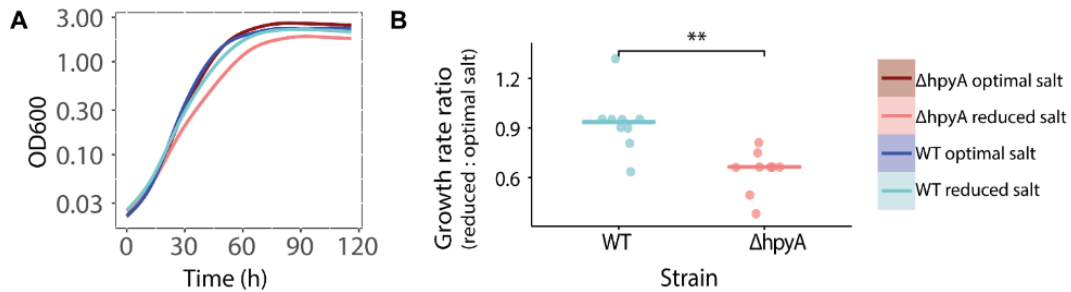
(default within DESeq2) was used as the criterion for significant differential expression (results in **Appendix B**).

Averaged normalized counts across biological replicates for each strain and stress treatment were then mean and variance standardized and subjected to Kmeans clustering using the factoextra package in R, which also determines the best value for  $K$ <sup>138</sup>. Resultant gene clusters were then subclustered using Kmeans and visualized using ggplot2<sup>139</sup> and pheatmap<sup>140</sup> functions in R ([https://github.com/amyschmid/HpyA\\_codes](https://github.com/amyschmid/HpyA_codes)). This clustering procedure was carried out twice, once with genes differentially expressed in both reduced and optimal salt, and then excluding genes differentially expressed in optimal salt. Results of the clustering are given in **Appendix B**. For analysis of gene functional enrichments, the hypergeometric test  $p$ -value of enrichment for differentially expressed genes was calculated. Resultant  $p$ -values were BH-corrected for multiple hypothesis testing. The archaeal Clusters of Orthologous Genes (arCOG) functional ontology was used for functional assignments<sup>141</sup>, results are listed in **Appendix B**.

## 2.3. Results

### 2.3.1. HpyA is important for wild type growth and morphology in low salinity stress conditions.

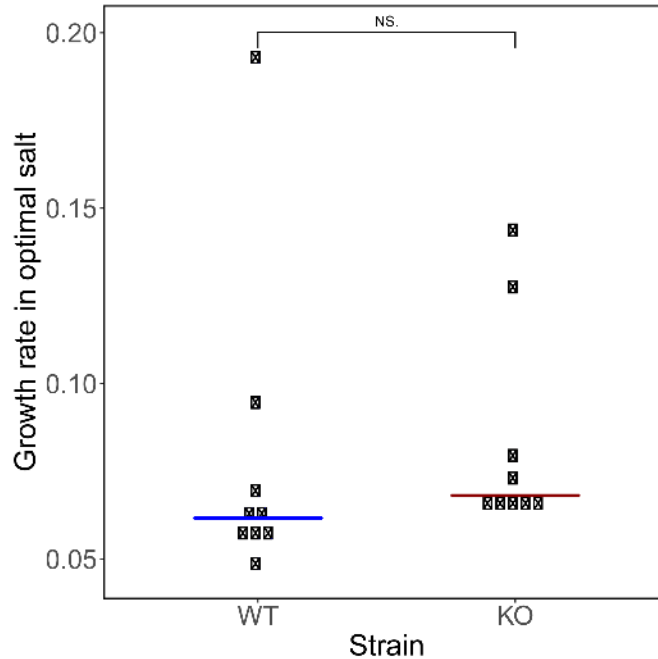
To test the hypothesis that HpyA plays a role in salt stress, we compared the growth rate of *Hbt. salinarum*  $\Delta$ *ura3* (parent strain, hereafter referred to as wild type or “WT”) to  $\Delta$ *hpyA* cells in rich complete medium with salt concentrations supporting optimal



**Figure 2: The  $\Delta hpyA$  strain is impaired for growth under reduced salt conditions. (A) Spline-smoothed growth curves for the  $\Delta ura3$  parent ('WT', blue curves) compared to the  $\Delta hpyA$  mutant (red curves) under optimal salt (dark colors) and reduced salt (light colors). (B) Dot plots of relative maximum instantaneous growth rate ( $\mu_{max}$ ) for each of the WT and  $\Delta hpyA$  strains. Each dot represents one of nine biological replicate trials measuring the  $\mu_{max}$  for each strain under reduced salt compared with its own growth in optimal conditions. Horizontal bars represent the median of each distribution.**

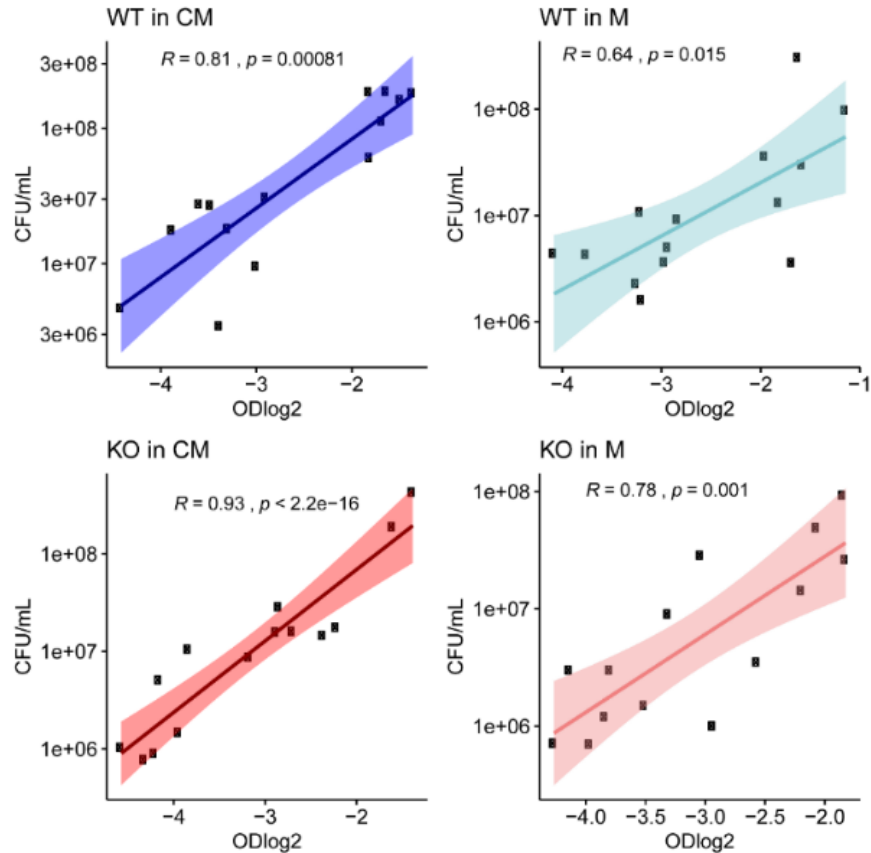
growth (CM, 4.2M NaCl) and CM with reduced salt (3.4M NaCl). As expected from previous observations<sup>104</sup>, instantaneous growth rate ( $\mu_{max}$ ) under optimal salt of the WT strain was statistically indistinguishable from that of  $\Delta hpyA$  (Fig. 2A, Appendix B, Fig. 3). Reduced salt slows the instantaneous growth rate ( $\mu_{max}$ ) of WT cultures to 89% of that in standard conditions. In contrast,  $\Delta hpyA$  cultures show significant growth impairment in reduced salt relative to WT, growing at 67% of their standard rate. (Fig. 2B; unpaired two-sample  $t$ -test  $p < 0.008$ ).

Cell morphology of *Hbt. salinarum* changes from rod-shaped to circular in the presence of low salt due to disruption of charges in the glycoprotein surface layer (S-layer)<sup>142-146</sup>. Our previous work demonstrated that the  $\Delta hpyA$  strain exhibits similar circularity in standard conditions<sup>104</sup>. To further test the hypothesis that HpyA plays a role in the salt stress response, we used phase contrast microscopy to visualize the combined effects of reduced salt and  $hpyA$  deletion on cell shape. From the images, we quantified circularity



**Figure 3: No significant difference in growth rate was observed for WT (blue) and KO (red) strains grown in optimal salt. Y-axis indicates maximum instantaneous growth rate ( $\mu_{max}$ ). Crossbars indicate the median of 9 biological replicate trials.**

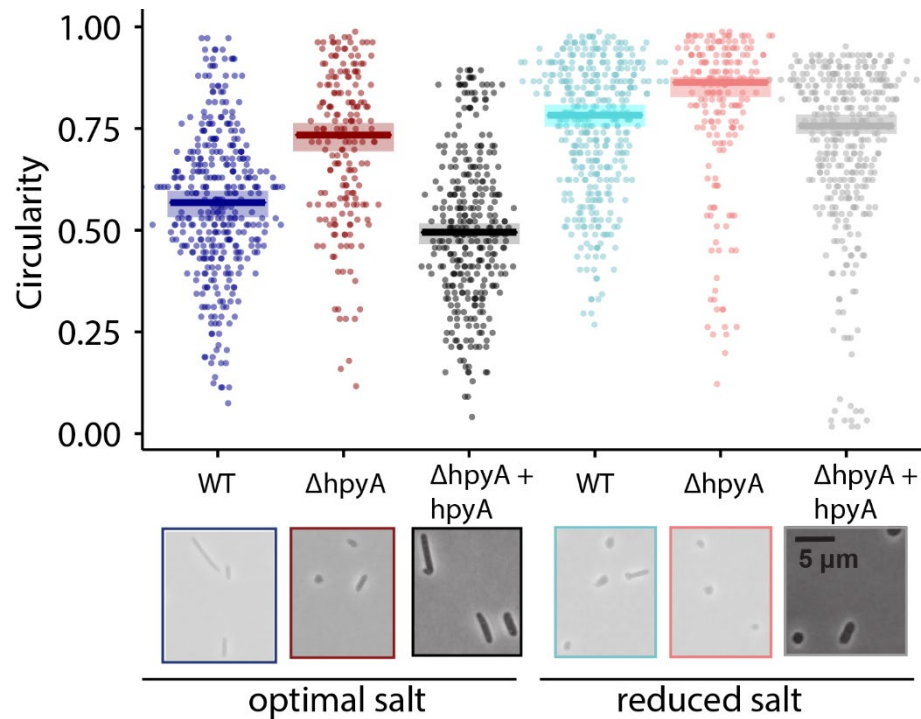
of individual cells (where 1 indicates a perfectly circular cell). In media containing optimal salt concentrations, WT cells are primarily rod-shaped, whereas the  $\Delta hpyA$  cells are significantly rounder (**Fig. 5**, non-parametric bootstrapped 95% confidence intervals of the medians of these distributions do not overlap, see Methods). In reduced salt, WT cell morphology was more circular: the median of the distribution was not significantly different from that of  $\Delta hpyA$  in optimal salt.  $\Delta hpyA$  morphology in reduced salt was the most circular of all strain-by-genotype combinations, indicating that this strain's morphology is strongly impacted by reduced salt.



**Figure 4: OD<sub>600</sub> is correlated with CFU/mL across strains and conditions.** To calculate colony forming units (CFU), six biological replicate cultures in selected growth phases (exponential and late exponential) were diluted  $1 \times 10^{-4}$  to  $1 \times 10^{-6}$  depending on the OD OD<sub>600</sub>. 100  $\mu$ L aliquots of each dilution were spread on CM plates. The number of colonies was counted after 7-10 days of growth in a 42°C incubator, and was related to the measured OD<sub>600</sub> at the time of plating. In each subpanel, log<sub>2</sub> OD<sub>600</sub> (x-axis) is plotted against the log<sub>10</sub> colony forming units (CFU) / mL (y-axis) for each of the  $\Delta$ hpyA knockout (KO) and  $\Delta$ ura3 parent strain (a.k.a. wild type, WT) in optimal salt complete medium (CM) or reduced salt medium (M) as indicated in each plot title. Spearman's correlation coefficients for OD vs CFU/mL and p-values of significance of each correlation are indicated at the top of each plot. Significant correlation between OD and CFU/mL is observed for each strain and condition, indicating strong correspondence between these two measures of growth, and therefore validating spectrophotometric measurements of growth.

Growth and morphology defects are significantly complemented by expression of *hpyA* from its native promoter *in trans* on a plasmid ( $\Delta$ hpyA + *hpyA*-HA, Fig. 5). Whole

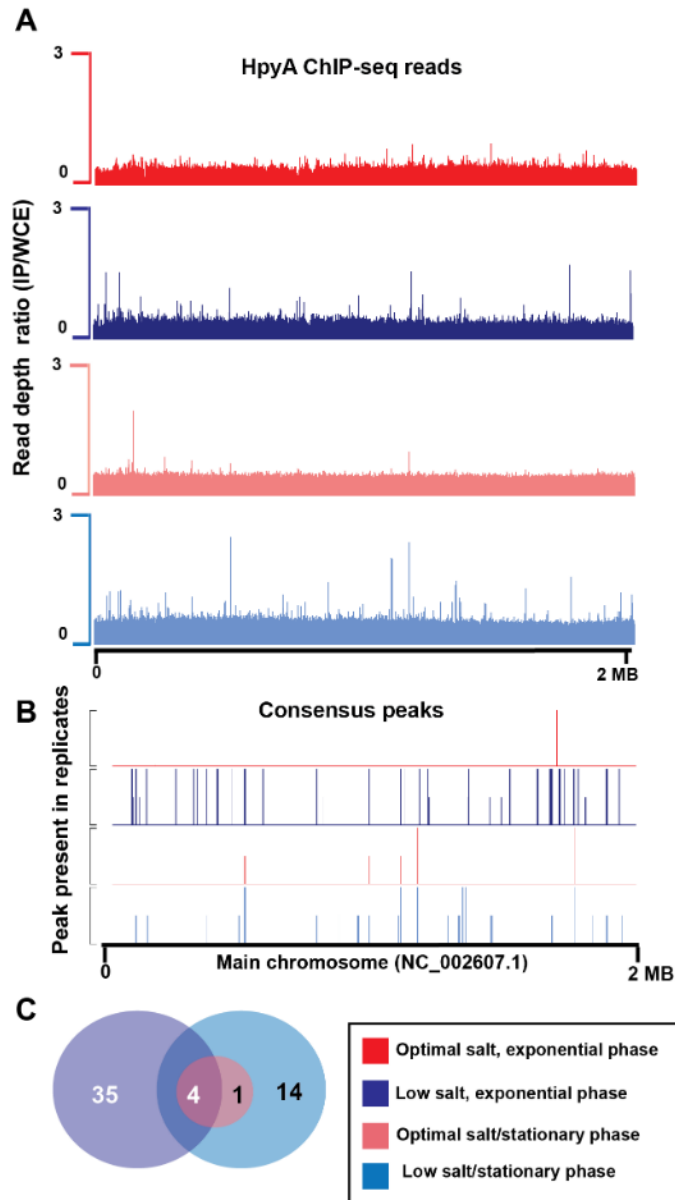
genome resequencing of the  $\Delta hpyA$  strain also demonstrated that: (a) second site suppressor mutations were absent; and (b) deletion of *hpyA* was complete through all chromosomal copies (**Appendix B**). *Hbt. salinarum* is highly polyploid<sup>23</sup>, which necessitates validation that all gene copies have been deleted. These results indicate that  $\Delta hpyA$  phenotypes are solely attributable to the deletion of *hpyA*. Because cell shape differences can lead to alterations in light scattering in a spectrophotometer<sup>147</sup>, as a control, we calculated CFU/mL by dilution plate counting. We found that OD600 measurements were well correlated with CFU counts for both strains and media preparations. In optimal salt, WT cultures CFU to OD Spearman correlation was  $\rho = 0.81$  ( $p = 0.00081$ ),  $\Delta hpyA$   $\rho = 0.64$  ( $p = 0.015$ ). In reduced salt, WT correlation was  $\rho = 0.93$  ( $p < 2.2 \times 10^{-16}$ ),  $\Delta hpyA$   $\rho = 0.78$  ( $p = 0.001$ ; **Fig. 4**). This indicates that the  $\Delta hpyA$  growth defect observed in reduced salt (as measured by optical density) is due to differences in growth and not an artefact of the shape change. Taken together, these batch culture (**Fig. 2**) and single cell microscopy (**Fig. 5**) quantitative phenotype data suggest that HpyA is important for maintaining wild type morphology and growth in response to hypo-osmotic salt stress.



**Figure 5: Circularity of *Hbt. salinarum* increases when *hpyA* is deleted under reduced salt. In dot plot, dots represent circularity measurements of individual cells. Horizontal bars are the median of the distribution in each strain under each condition. Shaded regions represent the 95% bias-corrected confidence interval from bootstrap resampling (see Methods). Below, representative micrograph images are shown for cells of WT,  $\Delta hpyA$ , and complemented strain ( $\Delta hpyA + hpyA$ , i.e. pKAD17) cells in optimal and reduced salt media. Scale bar is 5uM and consistent across images. Colours are as in Figure 2. Number of cells counted: WT in optimal salt, n=363; WT in reduced salt, n=383; *hpyA* in optimal cells, n=188; *hpyA* in reduced salt, n=187; complemented strain in optimal salt, n=313, complemented strain in reduced salt, n=360.**

### 2.3.2. *HpyA* binds genome-wide in a salt-specific manner

To determine which genes are potential targets of transcriptional regulation by *HpyA*, we performed genome-wide DNA binding location analysis using chromatin immunoprecipitation coupled to sequencing (ChIP-seq). For this purpose, we generated an  $\Delta hpyA$  strain expressing *in trans* *HpyA* translationally fused at its C-terminus to the hemagglutinin (HA) epitope tag. This fusion construct was driven by its native



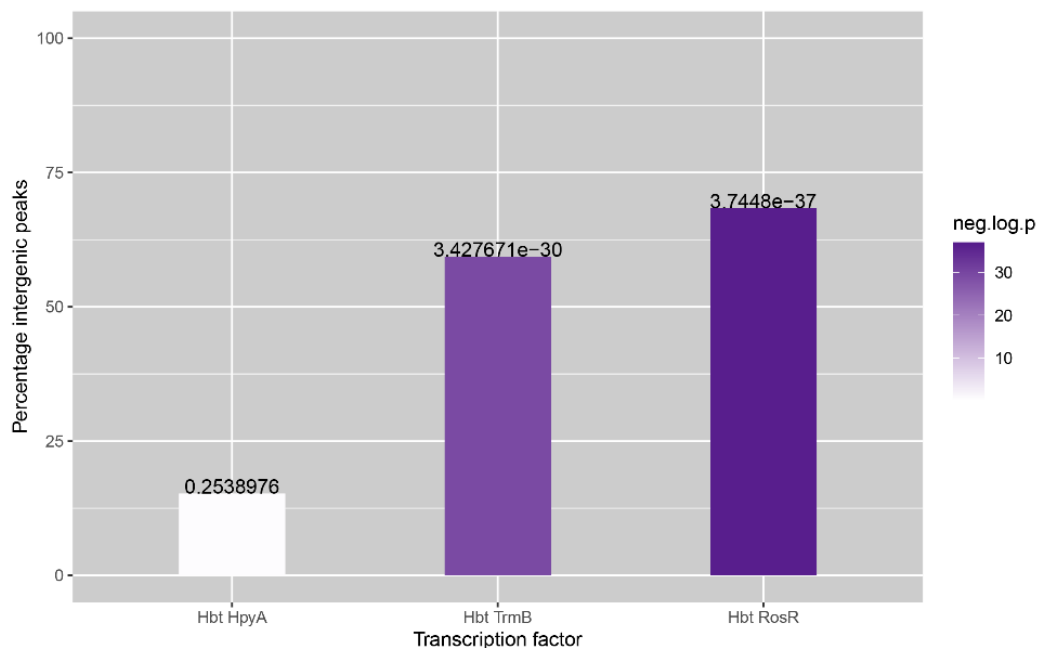
**Figure 6: ChIP-seq of HpyA shows salt and growth phase dependent binding patterns. (A) Chromosome-wide binding pattern (measured as read-depth of IP/Input) of HpyA-HA in optimal salt and exponential growth phase (red), optimal salt and stationary phase (pink), reduced salt and exponential phase (dark blue), reduced salt and stationary phase (light blue). (B) Reproducible peaks detected across at least 2 of 4 biological replicates for each condition – shorter peaks represent those found in 2 replicates only, while peak at full heights were detected in at least 3 replicates for that particular condition. Note that peaks shown in tag-alone control have been removed from the other conditions and from further analysis. (C) Venn diagram indicating the number of peaks detected in the different conditions. Circles are not scaled by number of peaks.**

promoter (see Methods and **Appendix A** for strain details). As described above, expression of HpyA-HA *in trans* complemented the circularity defect of  $\Delta hpyA$ , demonstrating that HA tag and plasmid-based expression does not interfere with wild type function of HpyA (**Fig. 5**). Based on the  $\Delta hpyA$  phenotypes observed (**Figs. 2 & 5**), the ChIP-seq experiments were performed at both physiological and reduced salt concentrations in both mid-exponential and stationary phase. HpyA binding was enriched relative to the background input control at a total of 59 discrete genomic locations (ChIP-seq peaks) across all conditions tested (**Appendix B**). These 59 peaks were consistently detected in reduced salt across growth phases and biological replicate experiments (**Figure 6**), but only 5 of these peaks remained bound in optimal salt conditions. Of the low salt peaks, 35 were detected exclusively during exponential growth phase, 14 exclusively during stationary phase, and 9 across both growth phases (**Figure 6C**). HpyA protein levels do not change significantly across the growth curve<sup>104,109</sup>. Because HpyA binds DNA primarily under low salt conditions, these results corroborate the growth and morphological impairments of  $\Delta hpyA$  cells observed in early log phase under reduced salt conditions (**Figure 2 and 5**).

HpyA binding peaks were located nearby 86 genes (within the gene coding region or 500 bp upstream of the gene start in the promoter region) (**Appendix B**). Few of the HpyA binding sites are located within non-coding regions of the genome (15.2%,  $p = 0.253$ ). In contrast, other previously characterized *Hbt. salinarum* TFs bind in a sequence-specific manner with significant preference non-coding regions (**Fig. 7**)<sup>132,133</sup>. Binding of

HpyA is also not statistically enriched within gene coding regions (84.8%,  $p = 0.615$ ). This high number was as expected because, like many archaeal genomes, the *Hbt. salinarum* genome is dense with coding sequences (86%).

Taken together these DNA binding results suggest that, unlike canonical histone proteins of eukaryotes and other archaeal species, HpyA binds in a salt-specific manner to a restricted set of sites genome-wide. However, unlike canonical TFs, HpyA binds apparently without preference for coding vs non-coding regions.

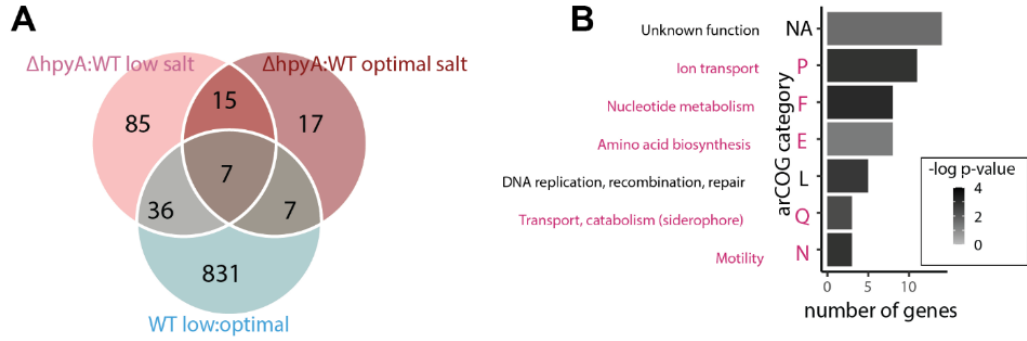


**Figure 7: HpyA binds without preference for coding vs non-coding regions. Height of the bar graph corresponds to percentage of ChIP-seq peaks in non-coding regions of the genome. Color of the bars are shaded by negative log<sub>10</sub> p-value of enrichment of peak locations in promoter regions (see scale at right). Actual p-values of enrichment calculated by hypergeometric test for each TF are written above each bar. HpyA binding locations (left) are compared with those for characterized TFs TrmB and RosR in *Hbt. salinarum* (Hbt) <sup>132,133</sup>.**

### 2.3.3. HpyA functions primarily as an activator of genes encoding ion transport and metabolic proteins.

Based on the quantitative phenotyping and ChIP-seq data, we reasoned that HpyA may regulate gene expression in response to salt stress. To test this hypothesis, we performed transcriptome profiling experiments in WT vs  $\Delta hpyA$  strains in both optimal and reduced salt using RNA-seq (see **Methods**). In the WT strain, over one-third of the genes in the transcriptome were significantly differentially expressed during exponential growth phase in reduced salt compared to optimal salt conditions ( $p < 0.05$ ; 882 genes; 37% of genome; **Appendix B**). Of the 37 genes previously identified by microarray analysis<sup>33</sup>, 22 genes were also identified as significantly differentially expressed in the current dataset. For these 22 genes, the fold-change in expression was strongly and significantly correlated across the two datasets ( $r = 0.86$ ,  $p < 2.2 \times 10^{-16}$ ). Our results therefore recapitulate but also extend previous observations that *Hbt. salinarum* mounts a strong, reproducible, and global regulatory response to hypo-osmotic stress.

To determine the extent of HpyA's regulatory reach, gene expression ratios ( $\Delta hpyA$ :WT) were calculated during mid-exponential growth in optimal salt and reduced salt conditions (in two separate DEseq2 analyses, see **Methods**). A total of 168 differentially expressed genes (DEG) were detected, 143 of which were significantly altered in reduced salt and 46 in optimal salt in  $\Delta hpyA$  vs WT (**Figure 8A, Appendix B**). Of these, 121 genes were uniquely differentially expressed in response to low salt. These genes are significantly enriched for a wide variety of functions critical to maintaining cell growth



**Figure 8: HpyA regulates gene expression in a salt-dependent manner. (A) Venn diagram illustrates the number of genes differentially expressed due to knockout of *hpyA* in different conditions. Genes with significant  $\Delta hpyA$ :WT ratios in optimal salt are shown in red, genes with significant  $\Delta hpyA$ :WT ratios low salt in pink, genes with significant low : optimal salt ratios in WT in blue. (B) arCOG enrichment of differentially expressed genes. X-axis shows the number of differentially expressed genes in each category that are annotated in the arCOG ontology, y-axis lists the arCOG category functions and short-hand single letter designations. Categories enriched in low salt are listed in pink text, categories enriched across conditions in black. Bars are shaded by Benjamini-Hochberg corrected<sup>137</sup> p-values of significance of enrichment according to the scale shown in the legend.**

and physiology in adverse conditions, especially ion transport and nucleotide metabolism (hypergeometric test  $p < 0.05$  enrichment in arCOG categories<sup>141</sup>, **Figure 8B**, **Appendix B**). Across both optimum and reduced salt conditions, the expression of 21 genes was significantly affected by *hpyA* deletion. These genes encode predicted functions in DNA recombination, replication, and repair pathways including RadA, DNA topoisomerase VI, and RPA family proteins (**Appendix B**).

To determine the role of HpyA in the activation or repression of these genes, we performed K-means clustering analysis of normalized read count data for gene expression across the four conditions tested ( $\Delta hpyA$  in low salt,  $\Delta hpyA$  in optimal salt, WT in low salt, WT in optimal salt, details in Materials and Methods). We first analyzed

the expression of the 21 genes that are differentially regulated the  $\Delta hpyA$  strain in both optimal and reduced salt conditions. These 21 genes fall into 2 clear categories – 10 genes downregulated in the  $\Delta hpyA$  strain and 11 genes upregulated (**Fig. 9A, Appendix B**). As noted above, genes across these two clusters are significantly enriched for DNA recombination, replication, and repair functions (8 genes). HpyA binding was detected in ChIP-seq by only one of these genes (*ssb*, encoding single-stranded DNA binding protein, **Appendix B**). HpyA binding was not detected for the 20 other genes in this cluster, indicating indirect regulation by HpyA.

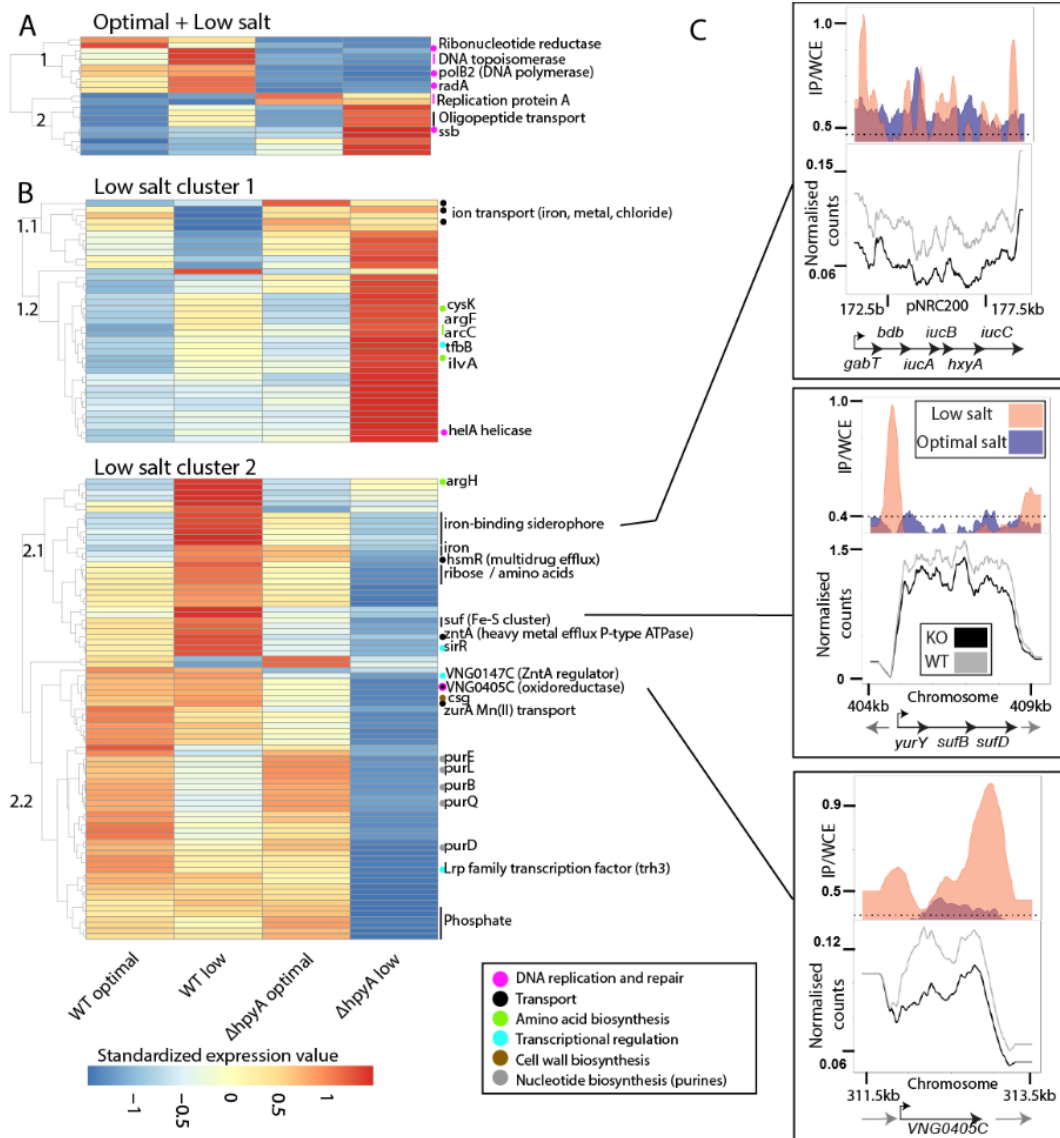
A separate clustering analysis of the 122 genes differentially expressed only in reduced salt in  $\Delta hpyA$  yielded two main patterns (**Figure 9B, Appendix B**). In cluster 1, genes are elevated in expression in the  $\Delta hpyA$  background under reduced salt relative to WT, whereas cluster 2 genes are downregulated. Cluster 2 includes 64% of genes differentially expressed in low salt, suggesting HpyA functions as an activator in the majority of cases in reduced-salt conditions.

To more clearly observe the gene expression patterns and the function of differentially expressed genes, we further divided these two main clusters, resulting in a total of 4 sub-clusters (**Figure 9B, Appendix B**). Subcluster 1.1 contains 11 genes whose expression pattern is downregulated in WT in reduced salt but upregulated in  $\Delta hpyA$ . This cluster includes 3 genes predicted to encode ion transport proteins (chloride, iron, and other metals). Subcluster 1.2 contains 28 genes that are upregulated in reduced salt in WT but more heavily upregulated in reduced salt in the knockout strain. The function of genes

in subcluster 1.2 are varied and not statistically enriched for a particular function.

However, notable among genes in cluster 1.2 include transcription factor B (TFB), four amino acid biosynthesis genes, and HelA ATP-dependent DNA helicase (**Figure 9A, Appendix B**). ChIP-seq enrichment for HpyA binding was not detected nearby any of the genes in subclusters 1.1 and 1.2, suggesting indirect regulation (**Figure 9B, Appendix B**). HpyA is therefore necessary but not sufficient for repression of cluster 1 in low salt conditions.

Cluster 2 contains many genes encoding transporters (21 genes across both subclusters 2.1 and 2.2). Notably, genes encoding known metal cation transporters exhibit tight clustering with their cognate transcriptional regulators SirR and VNG0147C<sup>34,148</sup>(**Appendix B**). Subcluster 2.1 consists of 37 genes modestly upregulated in reduced salt in the WT but strongly downregulated in reduced salt in the  $\Delta hpyA$  mutant. Interestingly, ChIP-seq enrichment for HpyA binding was detected at 4 sites nearby genes in this subcluster (**Appendix B**). Three of these 4 sites are nearby genes involved in maintenance of iron levels (**Fig. 9C**). These encode the siderophore (iron chelator) biosynthesis and transport operon, the Suf iron-sulfur cluster biosynthesis and transport system, and a putative oxidoreductase (VNG0405C). Surprisingly, the siderophore biosynthesis operon is bound at both the 5' and 3' ends by HpyA, which is associated with significant activation of this operon in low salt conditions (**Fig. 9C, top panel**). These results indicate that HpyA is required for direct activation of iron uptake under low sodium.



**Figure 9: HpyA-dependent regulon shows diverse expression patterns in the conditions tested. (A) Clustering heatmap of genes differentially expressed in response to hpyA deletion across both optimal and low salt conditions. Each column corresponds to the genotype in each condition and rows represent the averaged normalized counts for each gene. Each row is self-standardized for normalization. Genes labeled with certain colors represent gene functional categories (see legend for colors). Dots next to genes represent monocistronic genes, vertical bars indicate differentially expressed operons. (B) Clustering heatmap of genes differentially expressed in response to hpyA deletion in low salt conditions alone. (C) Normalized reads for 3 selected direct targets of HpyA. Each box corresponds to a particular gene target indicated in the heatmap. In each panel, ChIP-seq data are shown in the top box, RNA-seq data in the middle, genomic context at bottom. ChIP-seq y-axes represent the ratio of IP to input control (whole cell extract, or WCE). RNA-seq y-axes**

represent read depth for WT in reduced salt (grey traces) and KO in reduced salt (black). Genomic context images include the differentially expressed gene(s) (black arrows) and neighboring genes (grey arrows).

In subcluster 2.2, 46 genes are downregulated or constitutive across optimal and reduced salt in the WT, but more heavily downregulated in  $\Delta hpyA$  in reduced salt. All 8 differentially expressed nucleotide metabolism genes are found within this subcluster, and all encode *de novo* purine biosynthesis enzymes. However, only one of the 46 genes of subcluster 2.2 is a direct target of HpyA (*VNG0161G*, encoding glutamate dehydrogenase). This suggests that HpyA regulates purine biosynthesis and other functions in this subcluster in an indirect manner.

Together, these transcriptome profiling data integrated with ChIP-seq binding locations suggest an important role for HpyA as specific, direct activator of iron uptake, and an indirect global regulator of ion transport and nucleotide biosynthesis during hypo-osmotic stress.

## 2.4. DISCUSSION

Here we integrate quantitative phenotyping and functional genomics data to demonstrate that the sole histone-like protein encoded in the hypersaline adapted archaeal species *Hbt. salinarum* directly activates iron uptake transporters under hypo-osmotic stress. At other sites in the genome, HpyA also functions as an indirect, global activator of genes encoding functions central to cellular physiology in low ionic strength medium. These transcriptional effects enable cells to maintain rod-shaped cellular morphology and growth in hypo-osmotic conditions.

Two of the five operons under the direct transcriptional control of HpyA encode transmembrane ABC transporters that are predicted to import iron. One operon (VNG0524G-VNG0527C) encodes a putative iron-sulfur (Fe-S) cluster assembly system of the Suf family. The predicted encoded proteins exhibit moderate identity to the well-characterized *E. coli* Fe-S assembly proteins SufC, SufB, and SufD (45%, 56%, and 28%, respectively<sup>149</sup>). The other operon (*gabT/bdb/iucABC*, VNG6210-VNG6216) encodes siderophore biosynthesis and uptake. Siderophores are high-affinity iron binding chelators that are secreted from the cell and then imported via a dedicated ABC transporter<sup>150</sup>. In *Hbt. salinarum* and many bacteria, in addition to the ABC transporter, this operon includes a novel L-2,4-diaminobutyrate decarboxylase (DABA DC; encoded by *gabT*) and a DABA aminotransferase (encoded by *bdb*) for siderophore biosynthesis in lieu of synthesis via polyamines<sup>150</sup>. Because amino acids are precursors for DABA biosynthesis, down-regulation of *iucABC* in the  $\Delta$ *hpyA* mutant strain may also explain the indirect differential expression of amino acid biosynthesis genes during hypo-osmotic stress.

*Hbt. salinarum* is a facultative anaerobe capable of aerobic and anaerobic respiratory metabolism<sup>151</sup>. Across the tree of life, including *Hbt. salinarum*, iron is an essential cofactor for the function of respiratory complexes in the oxygen-accepting electron transport chain<sup>152</sup>. Because reduced salinity increases oxygen saturation in the medium, these conditions would favor aerobic respiratory metabolism over anaerobic metabolism, increasing the cellular demand for iron<sup>109</sup>. Indeed, we observe that these iron transport systems are induced in an HpyA-dependent manner under low salt

conditions (**Figure 9B**). Low levels of iron transport expression in the  $\Delta hpyA$  strain would therefore be expected to lead to low intracellular iron levels. Low intracellular iron has also been observed previously for strains deleted for *idr2*, which encodes a DtxR family iron-dependent TF in *Hbt. salinarum*. This TF also functions as a direct activator of the *iucABC* siderophore biosynthesis and transport operon, and intracellular iron levels are low in the  $\Delta idr2$  strain due to dysregulation of *iucABC*<sup>34,153</sup>. Idr2 is a member of a complex network of TFs that regulate the response to iron imbalance<sup>34,153</sup> and the current study suggests that HpyA is also involved in regulation of iron uptake. This mode of transcriptional regulation by HpyA explains the  $\Delta hpyA$  growth impairment observed in low sodium conditions tested here (**Fig. 2**).

The remaining three operons under direct HpyA regulation encode central metabolic functions (**Appendix B**). HpyA activates glutamate dehydrogenase and acyl-coA ligase enzymes, encoded by the *gdhB / alkK* operon. These enzymes control the entry of glutamate into the TCA cycle via the conversion of glutamate to 2-oxoglutarate.

Glutamate is also a key precursor for biosynthesis of many metabolites, including purines and other amino acids<sup>154,155</sup>. Direct control of this operon may explain the indirect transcriptional dysregulation of these pathways in the  $\Delta hpyA$  mutant strain.

HpyA activates an oxidoreductase gene (NAD-dependent epimerase predicted to act on nucleotide-sugar substrates; *VNG0405C*) and glycerol dehydrogenase gene and its associated operon (*VNG0161G / VNG0162G*), also encoding key components of core metabolism. The gene encoding a single-stranded DNA binding protein (*ssb*) is the only direct target predicted to be regulated by HpyA under both optimal and low salt

conditions, and repressed rather than activated. Although the precise relationship between these HpyA regulatory targets and the  $\Delta hpyA$  growth defect remains unclear, current knowledge of metabolism in *Hbt. salinarum* suggests that, in the  $\Delta hpyA$  mutant, disruption in the levels of key metabolic intermediates (glycerol, glutamate) may contribute to the growth impairment of this strain under low salt conditions.

Dysregulation of import and/or efflux of other ions (divalent metal cations, chloride, and other transporters) in the  $\Delta hpyA$  mutant may also explain the cell shape change in this strain (**Fig 5**). The proteinaceous surface layer (S-layer) is a key cell shape determinant of *Hbt. salinarum*<sup>146,156</sup>. The S-layer is pliable and allows for changes in cell shape under physical pressure and low salinity<sup>104,142</sup>. This shape change is exacerbated in  $\Delta hpyA$  (**Fig 5**), which we hypothesize is due to dysregulation of ion transport expression. Iron has also been shown to impact cell morphology in the related haloarchaeal species *Haloferax volcanii*, although the underlying mechanism remains unknown<sup>157</sup>. Expression of other pathways, for example, the S-layer (encoded by *csg*) and glycosylation enzymes (*VNG0140G*; **Fig 9 and Appendix B**) is reduced in the  $\Delta hpyA$  mutant under low salt conditions. However, these appear to play a more minor role in the  $\Delta hpyA$  morphology defect given that: (a) these genes are indirect targets of HpyA regulation; and (b) overall S-layer glycosylation is unaffected in strains deleted of *hpyA*<sup>104</sup>. Taken together, these data suggest that HpyA salt-dependent regulation of ionic balance is a major contributor to maintenance of wild type cell morphology and growth in reduced sodium environments.

Apart from these cases of direct regulation by HpyA, the majority of differentially expressed genes are located >500 bp away from HpyA binding sites. This can be explained in a number of ways. Several TFs are differentially expressed in the  $\Delta hpyA$  strain relative to WT in low salt (**Appendix B, Figure 9**). Therefore, the proximate cause of indirect differential gene expression can be inferred based on prior knowledge of the global gene regulatory network (GRN) in this organism<sup>32,48,133</sup>. For example, the general TF, TfbB, is differentially expressed in  $\Delta hpyA$  in low salt (cluster 1.1, **Fig 9**), and most of the genes in this cluster are indirectly regulated. TfbB is a direct regulator of several of the genes in this cluster, including *cysK*<sup>48</sup>. *Hbt. salinarum* encodes 7 paralogs of transcription factor B (TFB)<sup>158</sup>. Together, TFB and TATA binding protein recruit RNA polymerase to core promoters to initiate transcription<sup>6</sup>. The TFB network in *Hbt. salinarum* is highly interconnected: for example, TfbB directly activates TfbG, which in turn regulates other genes indirectly regulated by HpyA (e.g. metal transporter *VNG1744H*, **Appendix B**). Other indirect regulation by HpyA can be attributed to metal-responsive TFs. For example, SirR and VNG0147C, members of cluster 2.1, have previously been experimentally characterized as regulators of operons encoding metal transporters, specifically manganese uptake (*ZurA*) and the heavy metal efflux (*ZntA*), respectively<sup>34,148</sup>. Aside from indirect regulation as part of a transcriptional network, we note that our data do not exclude the possibility that HpyA may function as a co-regulator, perhaps by binding DNA through interaction with another TF. Hence, we propose that HpyA may, in part, achieve its global, indirect regulatory effect via its

regulation of genes encoding other TFs and/or through protein-protein interaction with other sequence-specific TFs.

In addition to transcriptional regulation of ion balance, HpyA may play other functional roles during hypo-osmotic stress. More than 40 HpyA binding sites were detected with no corresponding significant change in gene expression in the  $\Delta hpyA$  knockout (**Appendix B**). HpyA prefers to bind neither coding nor non-coding genomic regions, setting it apart from characterized haloarchaeal TFs that function by canonical, sequence-specific DNA binding to promoter regions [TrmB<sup>132</sup> or RosR<sup>133</sup>, **Fig. 7**]. We provide evidence of direct regulation both among targets bound in promoter and genic regions (**Appendix B**). Direct regulation of expression via binding in gene bodies has also been reported for *E. coli* regulator RutR<sup>159</sup>. The mechanism of RutR binding is complex and includes DNA bending, specific cis-regulatory sequence binding, and interaction with DNA-wrapping proteins<sup>160</sup>. Future biochemical studies on HpyA are therefore needed to elucidate its specific DNA binding mechanism. Our data do not rule out that non-canonical binding modes of HpyA could also influence other aspects of the transcription cycle, including elongation or termination. Bacterial nucleoid associated proteins (NAPs) bind DNA to regulate gene expression, remodel chromatin by bending or wrapping, and/or protect the nucleoid during stress<sup>58,161</sup>. For example, the *E. coli* transcription regulator CRP can function both as a canonical TF (site-specific gene regulation) for some genes, and as a DNA-bending chromatin remodeler at other genomic sites<sup>58,161</sup>. These newly-discovered and expanding roles for DNA binding

proteins calls for a broader perspective on the function of transcriptional regulators.

Likewise, further research is needed to explore such functional roles for HpyA.

Taken together, the results presented here strongly suggest that HpyA functions as a direct activator of iron regulatory genes and a global indirect regulator of diverse pathways. This function is markedly different than other characterized H3/H4-like histones in archaea and eukaryotes.

## **2.5. Acknowledgements**

I would like to thank current and former Schmid lab members (Cynthia Darnell, Rylee Hackley, Mar Martinez-Pastor, Angie Vreugdenhill, Peter Tonner, Sungmin Hwang, and Preeti Bhanap), for technical assistance with experimental methods and analysis, and for comments on the manuscript. I am grateful to my graduate thesis committee members (David Macalpine, Amy Grunden, Richard Brennan, and Raluca Gordan) for mentorship and comments on the manuscript. We thank Deyra Rodriguez and New England Biolabs for providing reagents and advice for rRNA depletion and RNA library preparation. We thank Antoine Hocher and Tobias Warnecke for scientific discussions during the preparation of the manuscript. We thank the Duke Sequencing and Genomic Technologies Core Facility for their technical expertise in generating the sequencing data reported here. Funding was provided by NSF MCB [1651117, 1936024, 1615685 to A.K.S.] and NIH T32 training grant [5T32GM007754 to the Duke University Program in Genetics and Genomics for S.S.].

### **3. Haloarchaeal histone proteins blur the line between transcription factors and nucleoid-associated proteins**

Chapter 3 is modified from a manuscript in preparation for publication. The authors are Saaz Sakrikar, Cynthia Darnell, Rylee Hackley, Angie Vreugdenhil Hayslette, Mar Martinez-Pastor, and Prof. Amy Schmid. S.S. and A.K.S. devised the project and wrote the manuscript. A.K.S. secured the funding for the project and contributed to data analysis. S.S. carried out strain creation, HstA ChIP-Seq, and data analysis under supervision of Prof. Schmid. C.D., A.H., R.H., and M.M.P. carried out ChIP-Seq of halophilic transcription factors, and provided technical assistance and expertise for experiments and analysis.

#### **3.1. Introduction**

Formation of chromatin and regulation of transcription are fundamental features of all domains of life, and are mediated by a variety of DNA binding proteins. These proteins are commonly characterised as chromatin proteins<sup>50</sup> and transcription factors (TFs)<sup>6,162,163</sup>, with some known to play both architectural and gene regulatory roles<sup>161</sup>. The amino acid sequences of many of these proteins are conserved throughout the domains of life.

Histones are ubiquitous in eukaryotes, where their architectural role is well-characterized<sup>52</sup>, and are also encoded in genomes across most archaeal lineages<sup>18</sup>.

Transcription factors with helix-turn-helix DNA binding domains are widespread in all domains of life<sup>6,164</sup>. In contrast, some chromatin proteins are specific to certain clades within the domain Archaea, for example Mc1<sup>90</sup> and Cren<sup>770</sup>.

However, conserved primary sequence and even protein structure may not necessarily imply conserved function. For instance, Alba, which can bind both DNA and RNA, is involved in chromatin architecture and in RNA structure stability in Archaea, while in eukaryotes it has a greater variety of functions, particularly in translation<sup>69</sup>. While Alba in both domains is capable of binding DNA and RNA, its specific biological roles have diverged. Similarly, the histone fold domain is present in all archaeal histones as well as eukaryotic core histones<sup>64</sup>, but is also detected in some eukaryotic transcription factors<sup>165</sup>. Again, while the DNA-binding role of the histone fold is conserved, its precise cellular function has diverged.

Hence, in understanding the function of these DNA-binding proteins, the question of which features of molecular function are conserved remains open: how do these proteins play architectural roles, how do they regulate transcription, or perform both functions? Certain hallmarks give clues as to the cellular functions of such proteins, including sequence determinants of binding, genomic positions relative to genomic features (coding or noncoding DNA), size of binding footprints, and shape of binding peaks. Many of these features are detectable in genome-wide DNA binding location analysis data, or ChIP-seq (chromatin immunoprecipitation coupled to sequencing)<sup>166-170</sup>. Here we address this question for the histone-fold containing proteins found in halophilic archaea using ChIP-seq and phenotypic analysis.

Across bacterial species, nucleoid-associated proteins (NAPs) share architectural roles in DNA, but differ in their mechanism of binding<sup>58,171,172</sup>. Some NAPs such as HU<sup>172</sup> or

Lrp<sup>173</sup> can bind without the need for a defined sequence, while others (like IHF<sup>172</sup>) bind a defined cis-regulatory sequence motif. NAPs are also diverse in terms of genomic binding locations: HU binds throughout the genome without discrimination between coding and non-coding regions<sup>172</sup>, whereas H-NS and Fis preferentially bind promoter regions<sup>171</sup>. Regardless of precise genomic location, NAPs bind frequently, covering 10-20% of the genome<sup>171,172</sup>. This is consistent with high NAP protein expression levels<sup>60</sup>, and their functions in genomic structural organization<sup>58</sup>. NAPs typically affect transcription globally, including indirect effects far from their binding loci<sup>171,172,174</sup>, but can also directly affect transcription by competitively inhibiting the binding of TFs at their target loci<sup>175</sup>.

In archaeal transcription, proteins that guide RNA polymerase to core promoters resemble those of eukaryotes (TATA binding protein, transcription factor B), whereas TFs that regulate gene expression in response to environmental perturbation more closely resemble those of bacteria<sup>6</sup>. Haloarchaeal and bacterial TFs typically regulate transcription of target genes by binding in a sequence-specific manner<sup>34,132</sup> proximal to gene promoters<sup>6</sup>. These proteins regulate expression of target genes by recruiting or hindering the basal transcriptional machinery<sup>6</sup>. However, in bacterial and archaeal species, some TFs bind hundreds of sites in the genome and can bind or loop DNA, such as Lrp family homologs<sup>166,176</sup>. DNA binding specificity is often weak when such TFs play architectural roles<sup>4,166,173,176</sup>.

In eukaryotes, the histone fold domain is required in the formation of the histone octamer (nucleosome core) and for histone-DNA binding<sup>54</sup>. Eukaryotic nucleosomes package DNA and regulate gene expression<sup>52</sup>. Like bacterial NAPs, eukaryotic core histones play an important role in genome architecture, binding frequently throughout the genome. Histone binding to DNA leads to nucleosome formation, but promoter regions form a “nucleosome-depleted region” with reduced histone binding, while regularly-spaced nucleosomes are present downstream of the gene start<sup>177-179</sup>. While these histones do not have a defined sequence motif, they disfavour poly-A sequences and preferentially bind sequences with ~10bp periodicity of A/T dinucleotides<sup>178-180</sup>. This corresponds with the length of the helical pitch of DNA<sup>181</sup>, and the positioning of the A/T dinucleotides is hypothesized to aid wrapping of the DNA helix around the nucleosome<sup>180</sup>. Like NAPs, histones are expressed at very high levels in eukaryotes (and in archaeal species in which histones play architectural roles) in order to bind and compact the genome<sup>118,182,183</sup>.

Phylogenetic evidence has traced the origin of the histone fold domain to the archaeal domain of life<sup>62,64,111</sup>. Studies primarily in thermophilic archaea representing the euryarchaeal superphylum suggested a eukaryotic-like DNA packaging function of these archaeal histones<sup>65,66,97</sup>, which can form nucleosomes (like in eukaryotes). Extended polymeric structures known as hypernucleosomes have also been observed that wrap and compact the genome in multiples of 30-60 bp<sup>63,68,97</sup>. The genomes of many of these well-studied species encode two histone paralogs; deletion of a single gene was found to be viable, while a deletion of both was not possible<sup>92</sup>. These archaeal histones favour

binding sites with 10-bp periodicity of AT-dinucleotides<sup>95</sup>, similar to eukaryotic histones<sup>180</sup>. Indeed, a genome-wide 10bp periodicity was detected in several archaeal species with histones<sup>184</sup>. A number of non-histone proteins of bacterial origin as well as archaeal-specific proteins also contribute to DNA architecture in the Archaea<sup>18</sup>. For example, the lineages of Crenarchaeota do not contain histone proteins but instead utilize the NAPs Alba, Cren7, and Sul7d for genome compaction<sup>69,71</sup>. *Thermoplasma acidophilium* does not encode histones. Instead the role of genome compaction is performed by a homolog of the bacterial-like NAP, HU<sup>74</sup>. In species of *Methanoarcina*, the histone protein was found to be dispensable for growth, and an archaeal protein, Mc1, was shown to be capable of performing an architectural role in the genome<sup>90,103</sup>. Hence, while archaeal histones perform their conserved role in chromatin architecture in many thermophilic species of archaea, their role in other lineages remains unclear.

In the hypersaline adapted (halophilic) lineage of archaea, the cellular function and DNA binding properties of histone-like proteins are understudied relative to those of eukaryotic histones or bacterial NAPs. So far evidence points to an alternative structure and function for halophilic histones. Halophilic archaeal genomes encode a sole histone protein with two histone fold domains<sup>100,104</sup>. This fused heterodimer forms a nucleosome with a structure resembling the (H3-H4) dimers present at the core of eukaryotic nucleosomes, but with a long flexible linker between monomers<sup>100</sup> (Schmid, unpublished data from RosettaFold predictions<sup>185</sup>). This fused dimer is strongly conserved across halophile genomes<sup>104</sup>. As an adaptation to their hypersaline environment, halophilic archaea have evolved a “salt-in” strategy by which the external sodium concentration is

balanced by a very high (~3-4 M KCl) potassium concentration within the cytoplasm<sup>25</sup>. The resultant highly ionic cytoplasm has led to further adaptations: the halophilic proteome is highly acidic to aid stability and solubility in this charged environment<sup>25</sup>. Hence, halophilic histone surface is highly acidic, unlike known eukaryotic histones in which the basic surface facilitates electrostatic attraction to DNA<sup>53,104</sup>.

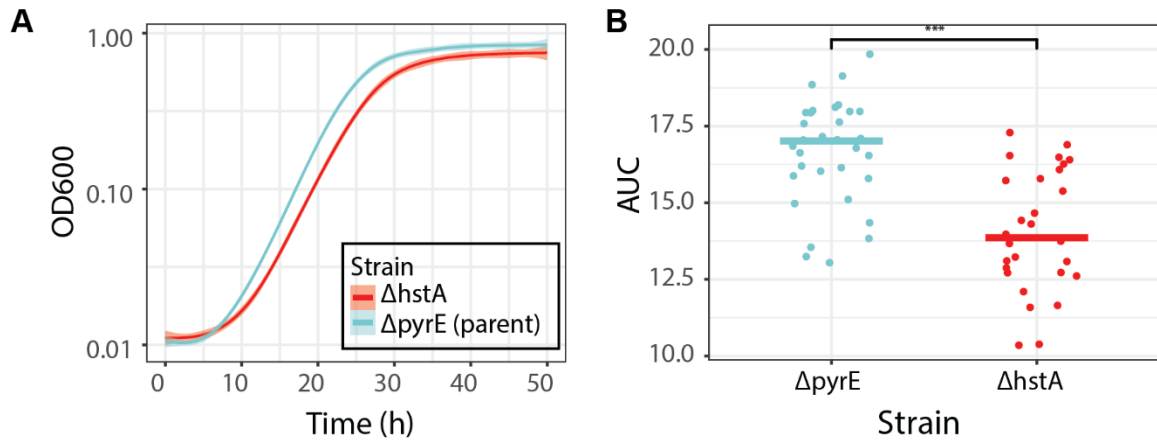
Our previous work demonstrated that putative DNA binding proteins encoded in *Hbt. salinarum*, including the histone-like protein HpyA and putative NAPs are expressed at low levels too and are therefore unlikely to compact the genome<sup>104</sup>. HpyA binding is sparse throughout the genome (~60 sites<sup>110</sup>) and can be deleted with no growth defect under standard conditions<sup>104</sup>. Instead, HpyA functions as a transcriptional regulator of inorganic ion transport and nucleotide metabolism by binding DNA in low sodium conditions<sup>110</sup>. HpyA is necessary for growth and cell shape maintenance in low salt, therefore linking its regulatory function to hypo-osmotic stress resilience<sup>110</sup>. Although HpyA clearly serves unique cellular roles, its DNA binding properties and conservation of functional features with other DNA binding proteins across the tree of life remain unclear.

To reveal the true evolutionary trajectory of DNA binding proteins that integrate genomic architecture with transcription, a more thorough understanding of archaeal histone function across the domain is required. Here we ask how and whether halophile histone binding properties are conserved across halophiles and the domains of life. The binding characteristics of these halophilic histone proteins are compared to those of

characterized bacterial NAPs, archaeal and bacterial TFs, and archaeal and eukaryotic histones. Based on these features of TFs, NAPs, and histones, we used quantitative phenotyping data, CHIP-seq, and genomic sequence data to evaluate halophilic histone function and binding attributes based on six criteria: (i) knockout phenotype; (ii) sequence specificity; (iii) binding location (promoter vs coding); (iv) binding frequency; and (v) binding peak size and shape. We examine the halophilic histone proteins from two model halophilic species, *Hbt. salinarum* and *Haloferax volcanii*. These species represent two different clades of the halophilic phylogeny and are therefore representative models for halophiles in general<sup>186</sup>. Our results suggest that haloarchaeal histone proteins are conserved across halophilic species, and possess functional attributes that differ from other known DNA binding proteins in some respects but are similar in others. Halophilic histones therefore blur the distinctions between TFs, nucleoid proteins, and histones, calling for more flexible definitions of DNA binding proteins.

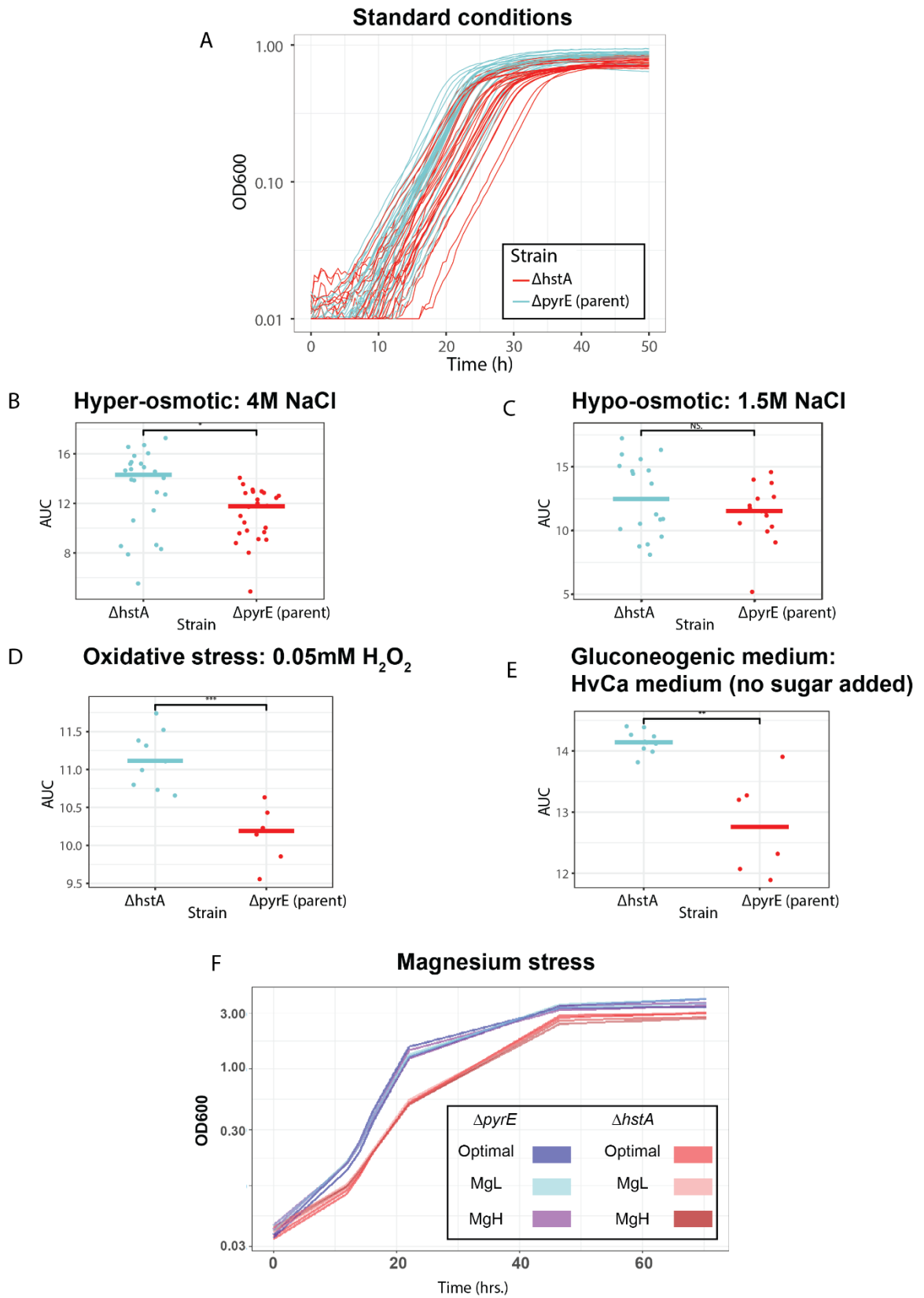
## **3.2. Results and Discussion**

### **3.2.1. *Haloferax volcanii* histone HstA is not essential but is important for maintaining wild type growth rate.**



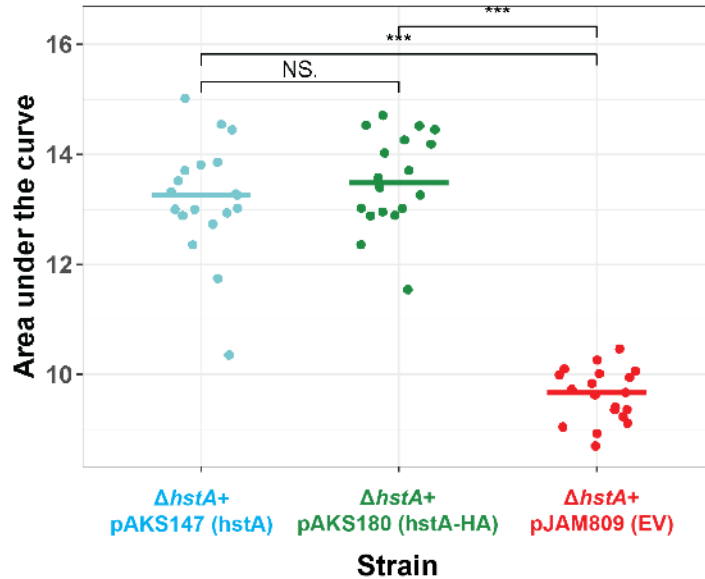
**Figure 10:  $\Delta hstA$  strain is impaired for growth in optimal conditions. (A) Growth of strains ( $\Delta pyrE$  blue,  $\Delta hstA$  red; 9 biological replicates with 2-3 technical replicates each) measured as optical density (OD600), curves represent 99% confidence interval derived from all replicates. (B) Area under the curve (AUC) of the growth curve for each strain under standard conditions as calculated by R package grofit; each dot represents one technical replicate. Horizontal lines represent the median of the distribution of points for each strain.**

HstA (HVO\_0520) is the sole histone protein encoded in the genome of the model halophile *Hfx. volcanii*. As we observed in our previous study<sup>104</sup>, HstA shares 65% sequence identity with HpyA histone-like protein of *Hbt. salinarum* and retains residues conserved across histones of nearly 80 sequenced halophile genomes. Here we used a genetic approach to determine if the observations previously made regarding non-essentiality of the *Hbt. salinarum* histone HpyA are generalizable to other species of halophiles<sup>104,110</sup>. *hstA* was readily deleted from *Hfx. volcanii* (details in Materials and Methods). However, unlike the  $\Delta hpyA$  deletion strain of *Hbt. salinarum*, this  $\Delta hstA$  strain exhibited a slight but significant growth defect compared to the parent strain in optimal conditions (rich media at 42°C, **Fig. 10A and B** and **Appendix B**, 84% of parent strain growth as measured by area under the curve; Welch two-sample t-test  $p < 1.5 \times 10^{-6}$ ).



**Figure 11:  $\Delta hstA$  phenotype in response to diverse stress conditions. (A) Raw growth curves of multiple biological replicate cultures under standard growth conditions (YPC medium, 2.5M NaCl, 0.3 M Mg<sup>2+</sup>, 42°C). (B-E) Each graph depicts the area under the log-transformed growth curve (AUC) under the conditions indicated at the top of each panel. Each point represents one growth curve, and the horizontal lines depict the median of the AUC distribution for each strain under each growth condition. (B) Growth in increased sodium (4M NaCl); (C) growth in reduced sodium (1.5M NaCl). (D) gluconeogenic conditions (no sugar added to HvCa medium). HvCa medium contains identical components to YPC18 described above, except without the addition of peptone or yeast extract<sup>46</sup>. (E) Oxidative stress (0.05M H<sub>2</sub>O<sub>2</sub>).  $\Delta hstA$ :parent growth ratio in all cases was 0.84 or higher, indicating *hstA* is not required for stress response. (F) Growth curves in high MgCl<sub>2</sub> (overall Mg<sup>2+</sup> 0.48M), low MgCl<sub>2</sub> (overall Mg<sup>2+</sup> 0.19M); visual analysis of the data confirmed that growth under Mg stress conditions was identical to non-stress conditions.**

Growth under a variety of stress conditions (sodium and magnesium stress, oxidative stress with peroxide, alternate nutrient conditions) was tested. The ratio of the area under the curve for  $\Delta hstA$  to parent under these conditions was found to be in the range 84%-91%, at or above the ratio for optimal conditions (**Figure 11**). These data indicate that the  $\Delta hstA$  growth defect under standard conditions is not further compounded under these stress conditions, suggesting that *hstA* is dispensable for growth under the stress conditions tested. Growth at high temperature resulted in an ambiguous phenotype, detailed in **Appendix D**. This growth defect is significantly complemented by the *in trans* expression of *hstA* alone or translationally fused in frame to the hemagglutinin epitope tag (**Figure 12**), indicating that the growth defect is attributable to the deletion of *hstA* and not due to polar effects on surrounding genes.



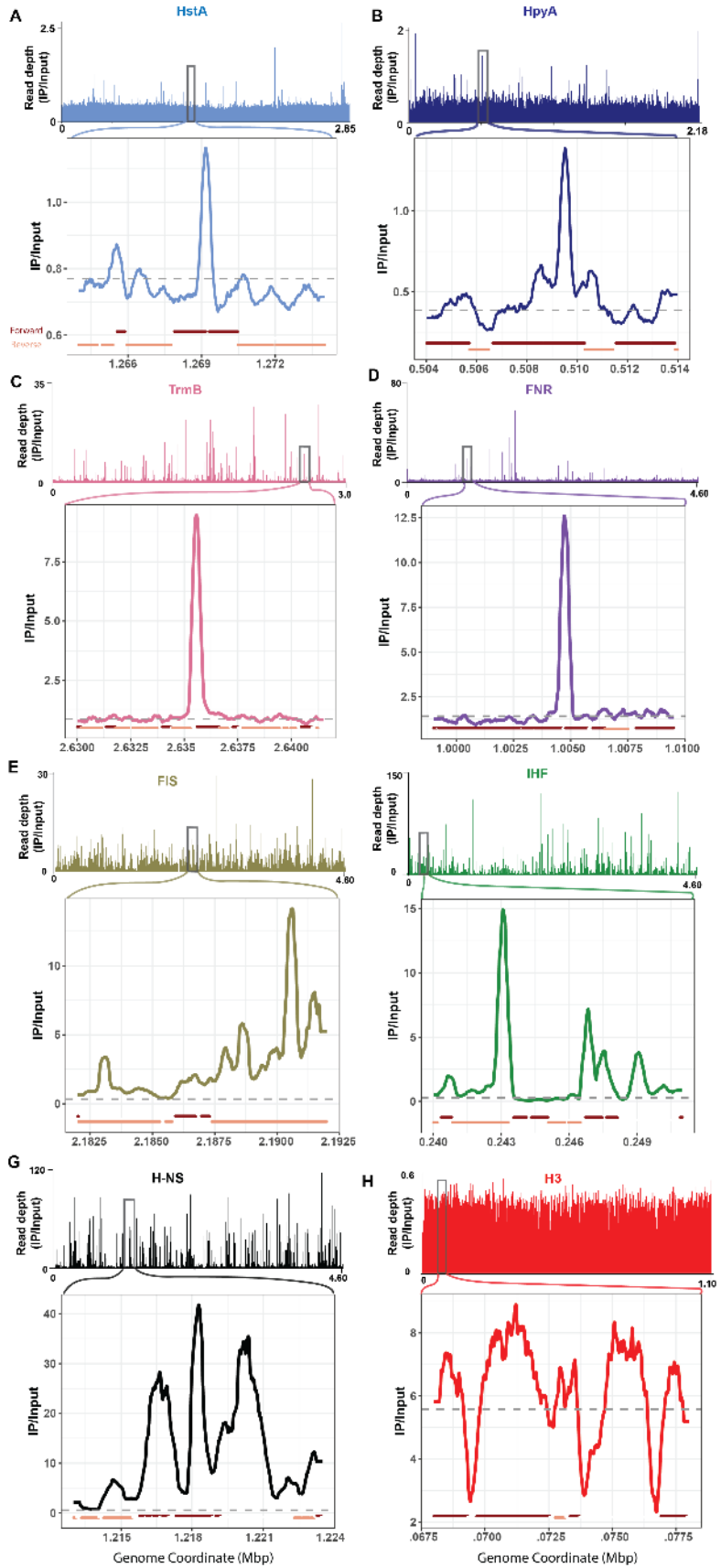
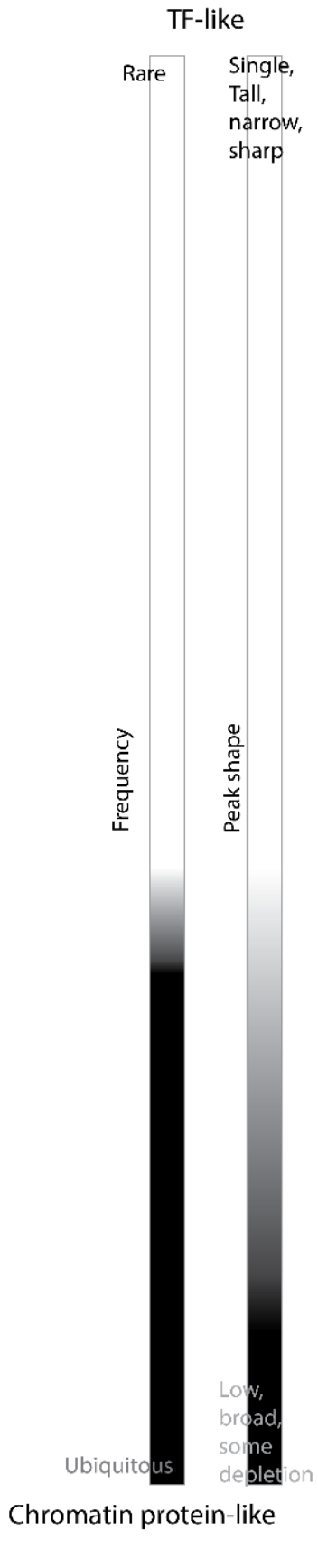
**Figure 12:** *hstA* deletion can be complemented *in-trans*; HA tag does not interfere with HstA function. Boxplot representing growth measured by area under the curve of plasmid-containing strains as calculated by the R package *grofit*; each dot represents one technical replicate; central line represents median value.

In addition, these data indicate that the C-terminal HA tag does not interfere with HstA function, allowing ChIP-seq with the tagged strain to be carried out. Whole-genome re-sequencing verified the absence of any secondary site mutations in this strain (**Appendix B**) and the complete absence of any wild type *hstA* copies from the genome (halophiles are highly polyploid<sup>187</sup>, necessitating such validation). Taken together, these results establish that *hstA* is a non-essential gene whose deletion causes a slight but significant growth defect under standard growth conditions. We conclude that HstA resembles *Hbt. salinarum* HpyA in that both are dispensable for growth. In contrast, known chromatin proteins are usually essential<sup>188,189</sup>, if not individually, then combinatorially. For example, in *Thermococcus kodakarensis*, with two histone genes, either single knockout is

viable but a double knockout is not possible<sup>92</sup>. HpyA and HstA therefore differ from other known histones with respect to their essentiality for viability.

### **3.2.2. HstA genome-wide location analysis (ChIP-seq) reveals binding patterns similar to those of HpyA.**

We carried out chromatin immunoprecipitation coupled to sequencing (ChIP-seq) to locate and compare genome-wide binding sites of HstA to HpyA<sup>110</sup> (details in Materials and Methods). Given that HstA is important for growth under standard conditions, we conducted ChIP-seq experiments in optimal conditions in exponential phase. Across the *Hfx. volcanii* genome, we observed infrequent HstA binding in discrete, sharp peaks of enrichment (**Fig 13A**). This binding frequency and peak shape were similar to those observed for *Hbt. salinarum* HpyA<sup>110</sup> (**Fig 13B**). Binding peak frequency and shape in ChIP-seq data correlates with function: such rare, 'sharp' narrow peak shapes are often observed for site-specific TFs, in contrast to the wide, broad peaks typical of histones<sup>167</sup>. The similarities between the binding profiles of the two histones extend further: the number of reproducible peaks for HstA was 32 (**Appendix B**), on the same order of magnitude seen for HpyA (59 peaks<sup>110</sup>). HstA peaks averaged 374 bp wide (see zoom-in of representative peak, **Fig 13A**), also of the same order of magnitude as the mean peak width observed for HpyA (299 bp, **Fig 13B, Appendix B**). For each of HpyA and HstA, these binding loci cover <1% of the genome of *Hbt. salinarum* and *Hfx. volcanii*, respectively, suggesting that the overwhelming majority of the genome is not bound by halophilic histones. Of these peaks, only 15.6% were in non-coding regions, corresponding with a genome that is



**Figure 13: ChIP-seq binding signal for HpyA and HstA compared with TFs, NAPs, and eukaryotic histone. In each panel, chromosome-wide binding patterns (measured as read-depth of IP/Input) are shown above and zoomed-in regions of representative peaks are shown below. All archaeal and bacterial genome views depict the main chromosome of each species. (A and B) show halophilic histone-like protein binding patterns: (A) *Hfx. volcanii* HstA (light blue, NCBI accession NC\_013967.1, peak centre 1.27Mb); (B) *Hbt. salinarum* HpyA (dark blue, NC\_002607.1, peak centre 0.51Mb). (C) Depicts halophilic TF *Haloarcula hispanica* TrmB (pink, NC\_015948.1, peak centre 2.64Mb). (D) Shows bacterial TF *E. coli* FNR (purple, NC\_000913.3, peak centre 1.01Mb). (E-G) *E. coli* NAPs: (E) H-NS (black, peak centre 1.22Mb); (F) IHF (green); (G) Fis (olive). (H) Shows yeast histone H3 (red), chromosome VII (NC\_001139.9). For the TFs and H-NS, known to directly regulate target genes<sup>132,171</sup>, peaks with a known functional role were chosen. For each genome-wide view and zoom-in, X-axis represents chromosomal coordinates in Mbp, Y-axis represents the read depth ratio of IP to input control (i.e. binding enrichment). Grey dotted line in the zoom-ins represent a baseline calculated from the average genome-wide IP/Input signal; dark red and tan lines below each zoom-in plot represent genomic context (forward and reverse strand genes, respectively). Scale at left indicates the classification of each DNA binding protein pattern based on features of frequency and peak shape.**

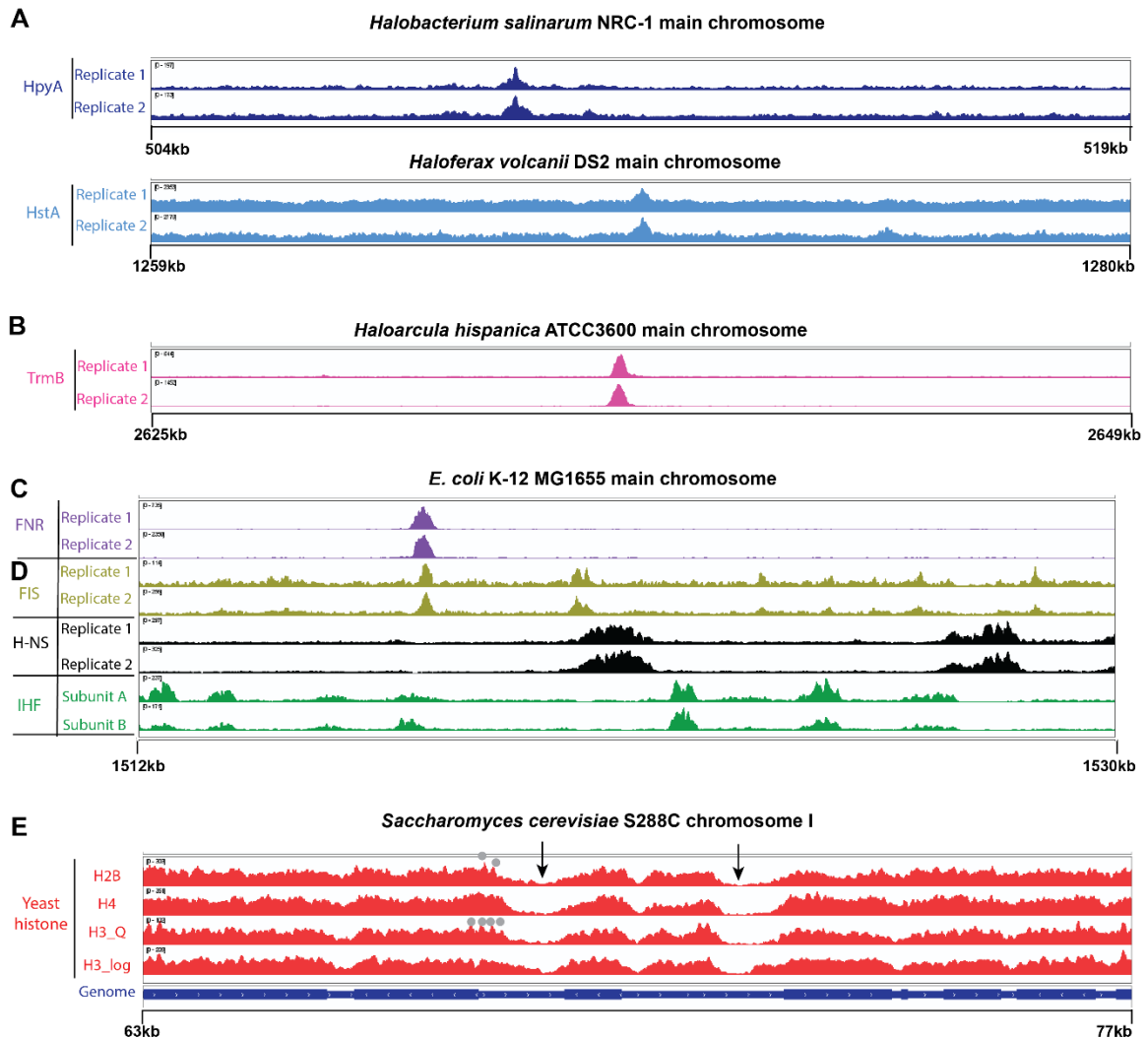
15.8% intergenic; hence, like HpyA<sup>110</sup>, HstA binding is also not enriched in intergenic regions ( $p$ -value >0.4; **Appendix C**). However, unlike HpyA, which regulates ion uptake<sup>110</sup>, the genes nearby HstA binding peaks were not enriched for a particular function according to archaeal clusters of orthologous genes (arCOG) categories. Taken together, the data suggests that the genome-wide DNA binding patterns of *Hbt. salinarum* and *Hfx. volcanii* histone-like proteins are conserved across halophiles, with their binding peaks resembling those of TFs.

### **3.2.3. Comparison of halophilic histone-like protein binding patterns with those for TFs, NAPs, and eukaryotic histones.**

To understand the DNA binding functions of halophilic histones relative to known DNA-binding proteins, we compared our ChIP-seq results to those of eukaryotic histones, bacterial NAPs and TFs, and halophilic TFs (details of datasets used in

Methods section and **Appendix B**). Together, these proteins encompass a wide spectrum of cellular functions and DNA binding modes, and hence provide a comprehensive set of comparisons for our data from halophilic histones.

We first analysed the similarities and differences between these proteins by visual inspection of the genome-wide binding patterns and representative zoomed-in regions (**Fig 13, Figure 14**). We focused on peak shape and genome-wide binding frequency: it has been observed that “canonical” TFs tend to exhibit tall sharp peaks and bind rarely<sup>167</sup>, consistent with their role in site-specific regulation of transcription initiation<sup>6</sup>. Consistent with this hypothesis, as described above, we observed that halophilic histone-like proteins binding enrichment forms discrete, tall, narrow peaks at relatively few locations in the genome (**Fig 13A, B, 14A**). We observe similarly sharp, discrete peaks binding rarely in the case of halophilic TF TrmB (**Fig 13C, 14B**) and bacterial TF FNR (**Fig 13D, 14C**). Such sharp peak shapes have been observed for sequence-specific TFs in eukaryotes as well<sup>167</sup>. In contrast, previous research has shown chromatin-like proteins (histones, NAPs) bind ubiquitously with broader, flatter peaks, and/or with areas of depletion, in keeping with their roles in DNA architecture and compaction<sup>4,167,169</sup>. Consistent with this, we observed that bacterial NAPs IHF, H-NS, and FIS bound frequently in the genome (**Figs 13E-G, Fig. 14D**), particularly for H-NS and IHF, consistent with previous reports that these NAPs cover ~17% and ~11% of the genome, respectively<sup>171</sup>. However, we observed a mixture of peak shapes for the NAPs. H-NS exhibited very broad peaks, often more than a kb in width (**Fig. 13G**), while there was mix of narrow and broad peaks seen for IHF (**Fig. 13F**) and FIS (**Fig. 13E**), with a

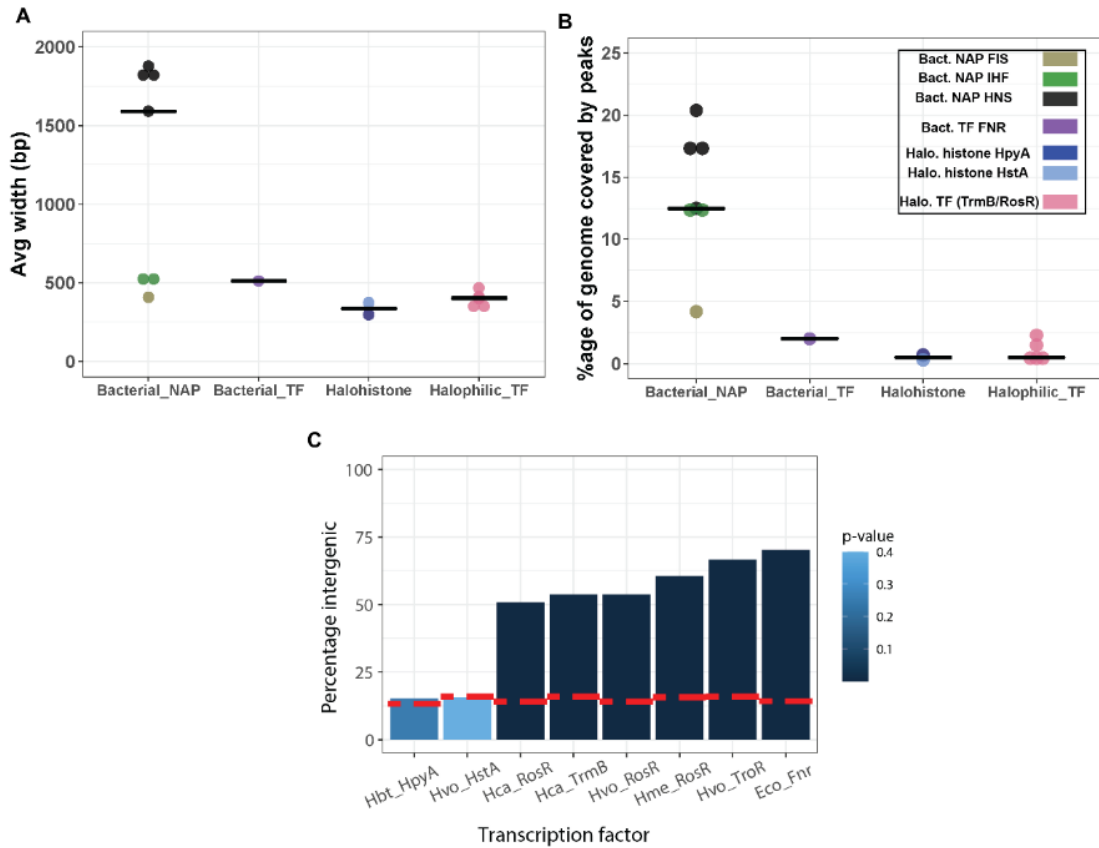


**Figure 14: HpyA and HstA bind in few, discrete peaks, like TFs, and contrasting with histones and NAPs. (A) HpyA (dark blue) and HstA (light blue) binding is observed as discrete reproducible peaks; shown here are 2 representative replicates each. (B) Halophilic TF TrmB (pink) binds in discrete reproducible peaks; 2 replicates shown (C) Bacterial TF FNR (purple) binds in discrete reproducible peaks (2 replicates shown). (D) Bacterial NAPs cover a large part of the genome, either with a large number of peaks: IHF (green), FIS (brown), broad peaks: HNS (black). (E) Yeast histone ChIP-Seq (red) shows depletions in promoter regions of some genes (marked with arrow) and ordered nucleosomes in the gene body (grey dots); shown here is data from histone H2B, H4, H3 in quiescent phase (H3\_Q), and H3 in logarithmic growth phase (H3\_log).**

very high frequency of peaks for IHF (also seen in Fig. 14D). Indeed, for yeast histones, we do not observe binding peaks at all in the genome-wide view (Fig 13H, 14E). Upon

closer inspection of local genomic regions, we observe broad, flat areas of enrichment punctuated by depletion at promoter regions, as expected from ubiquitous binding and nucleosome formation (except at promoters of transcribed genes)<sup>178,179</sup>.

To quantify these characteristics in order to gain further insight into halophilic histone function, we tallied and compared the number of binding events, average peak width, and percentage of the genome covered for each DNA binding protein (Methods, **Appendix B, Figure 15**). As discussed above, we detected 59 reproducible binding sites for HpyA<sup>110</sup> and 32 for HstA. This is at the lower end but within the range observed for haloarchaeal and bacterial TFs (36-253 peaks, **Appendix B**). The average width of HpyA and HstA peaks (299bp and 374bp, respectively) is comparable to the range of peak widths observed for haloarchaeal TFs (318-466bp; **Fig 15A**) and the bacterial TF FNR (511 bp average width). Halophilic histone peaks also most closely for HpyA and HstA, 0.4-2.2% for haloarchaeal TFs, 2% for bacterial TF; **Appendix B, Fig 15B**) resemble halophilic and bacterial TFs with respect to the percentage of the genome bound (<1%). By contrast, the average peak width and genome coverage is more variable across the various bacterial NAPs included in the comparison here (**Fig 15A,B**). On average, NAP binding sites are more numerous (mean number of peaks is 631), wider (average 1.2kb), and cover more of the genome (average 14%) than HpyA and HstA binding peaks, particularly in the case of H-NS (**Appendix B, Fig 15**). These numbers generated from our analysis correspond with published estimates that, as a consequence of their genome-wide architectural role, NAPs cover 10-20% of the genome<sup>171,172</sup>.



**Figure 15: Genomic features of HpyA and HstA binding sites according to ChIP-seq data. (A) Average width of all ChIP-seq peaks for a given DNA-binding protein, arranged into columns by type (bacterial NAPs, bacterial TFs, halophilic histones, halophilic TFs). (B) Percentage of genome covered by all ChIP-seq peaks of a given DNA-binding protein arranged into columns by type (bacterial NAPs, bacterial TFs, halophilic histones, halophilic TFs). (C) Bar graph of ChIP-seq peaks for HpyA, HstA, and transcription factors TrmB, RosR, and TroR. Species names are abbreviated: Hbt, *Hbt. salinarum*; Hvo, *Hfx. volcanii*; Hca, *Haloarcula hispanica*; Eco, *E. coli*. The height of each bar represents the percentage of peaks located in intergenic regions. Dotted red line indicates the percentage of each genome that is non-coding. The intensity of colour of the bars represents hypergeometric test p-values of significance for enrichment within promoter regions (see legend for colour scale).**

Because binding location often relates to molecular function, we next analysed the

location of the HpyA and HstA peaks relative to genomic features (Fig 15C). As

discussed above, neither HpyA nor HstA show a preference for certain genomic

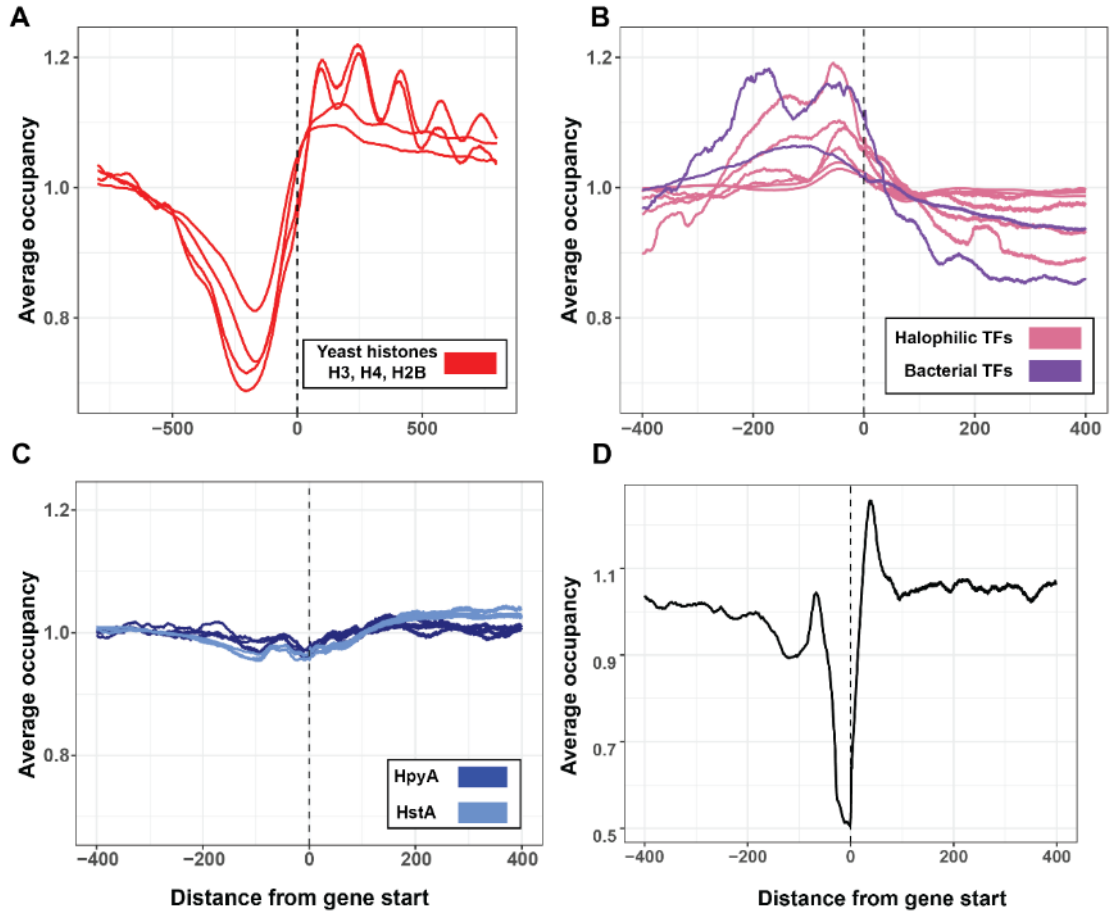
features, neither coding nor promoter sequences. In contrast, as expected from previous

studies<sup>132</sup>, haloarchaeal and bacterial TF binding sites were significantly overrepresented in intergenic regions relative to the genomic backgrounds of these species, which are 84-87% coding (hypergeometric test;  $p < 10^{-3}$ ) (**Appendix C**). For these TFs, the proportion of peaks binding to intergenic regions varied between 51% for *Hfx. mediterranei* TrmB to 70% for *E. coli* FNR (**Fig. 15C**). Hence, while halophilic histones appear to bind without preference for genic or intergenic regions, TF binding favours intergenic regions.

Taken together, these data demonstrate that halophilic histones meet many quantitative criteria for binding patterns like those of TFs (binding at discrete, narrow peaks at relatively few genomic sites). However, in their lack of preference for binding particular genomic features, HpyA and HstA resemble NAPs such as IHF and HU<sup>172</sup>. Thus, HpyA and HstA binding patterns resemble those of TFs in some respects but non-specific DNA binding proteins in others.

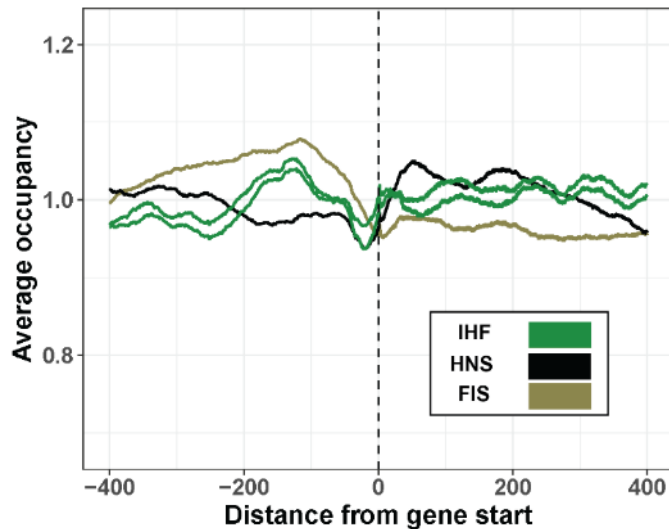
#### **3.2.4. Haloarchaeal histone-like protein occupancy curves surrounding start sites are unique relative to canonical histone and TF signals**

To further challenge the hypothesis that HpyA and HstA bind DNA like typical histones, the average occupancy (normalized read depth) at open reading frame (ORF) start sites were compared across DNA binding proteins (see Methods). As expected from previous studies of global nucleosome occupancy,<sup>170,177,178,190</sup> we detected a depletion in ChIP-seq binding signal at promoter regions but enrichment at regularly spaced nucleosomes in the gene body for *Saccharomyces cerevisiae* histones H3, H4, and H2B (**Fig. 16A**, data sources listed in **Appendix B**). Specifically, the upstream depletion



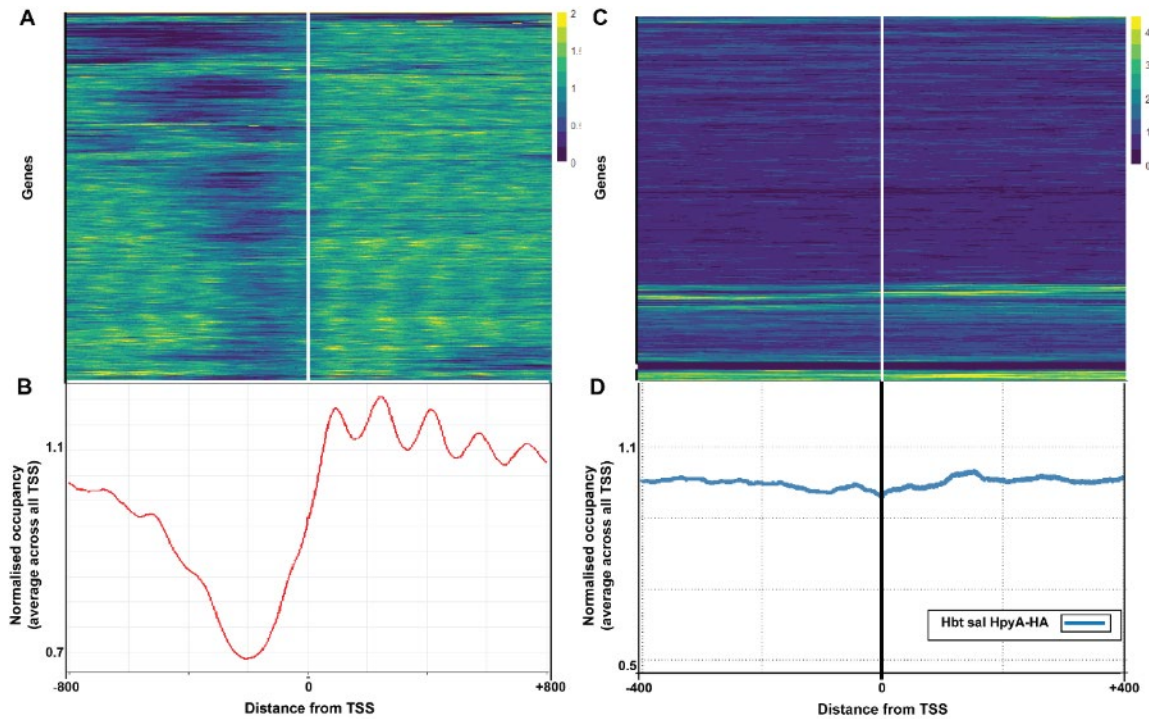
**Figure 16: Binding occupancy at start sites of selected DNA-binding proteins.** Average binding occupancy across all genes for (A) yeast histones (red), (B) HpyA (dark blue, 1 representative replicate from each condition tested) and HstA (light blue, 3 replicates, where each line is one replicate), and (C) Bacterial TFs *E. coli* FNR and *Brucella abortus* VjbR (purple) and Archaeal TFs *Hca. hispanica* TrmB, *Hfx. volcanii* TroR, *Hfx. med* RosR (pink, 2 replicates each). (D) Re-analysis of genome-wide average of micrococcal nuclease digestion (MNase-seq) pattern for *Hfx. volcanii*. In each panel, x-axis represents distance from start site (bp), y-axis represents average occupancy, as measured by read depth in genomic positions around the start site, normalized to average depth across the genome.

showed an occupancy minima at ~160-200 bp upstream of the start site for the histone proteins analyzed. In contrast, binding peaks representing the +1 to +3 nucleosomes bound downstream of the start site were detected in the expected positions for histone H3 (+1 at ~100 bp, with the subsequent peaks at ~150 bp intervals, Fig. 16A). The



**Figure 17: Binding occupancy at start sites of bacterial NAPs. Average binding occupancy across all genes for *E. coli* IHF (green, 2 subunits), H-NS (black), FIS (olive).**

nucleosome positions detected in our analysis therefore correspond with known positions of the nucleosome-free region and bound nucleosomes across gene start sites and into the 5'-end of gene coding regions. In the case of haloarchaeal TFs, we detected, as expected, increased occupancy ~50 bp upstream of gene start sites (**Fig. 16B**), which corresponds with prior data that TFs generally bind in gene promoters<sup>6</sup>. Depletion is observed within gene bodies, corresponding to the known relatively infrequent binding within genes<sup>132</sup>. The bacterial TF FNR also exhibited an upstream enrichment similar to archaeal TFs, again corresponding to the known preference for bacterial TF binding in promoter regions<sup>191</sup>. Bacterial NAPs showed a variety of profiles (**Fig. 17**), including upstream enrichment and/or depletion. In contrast to these other DNA binding proteins, haloarchaeal histone occupancy profiles are relatively flat across the ORF start site, (**Fig 16C**), consistent with our observations that neither HpyA nor HstA binding is enriched nearby and genomic features (Fig. 15C). Importantly, these profiles for HpyA and HstA



**Figure 18:** A) Heatmap of yeast histone H3, with each row representing the start site of one gene. Color scale at right represents average normalized occupancy. (B) Line graph of histone H3 showing the average across all genes. (C) HpyA heatmap with color scale as in A. (D) HpyA average line graph.

stand in sharp contrast to those of the other DNA binding proteins compared. Heatmap representations of occupancy data for each individual gene corroborate these findings (Fig. 18).

We note that no ChIP-seq data are available in the literature for archaeal histones other than HpyA or HstA. However, genome-wide mapping of nucleosomes [genomic regions resistant to micrococcal nuclease (MNase) digestion] has been carried out using MNase-Seq in *T. kodakarensis*<sup>95,97</sup>, *Methanothermobacter thermoautotrophicus*<sup>95</sup>, and *Hfx. volcanii*<sup>108</sup>. In all these data, a depletion in occupancy just upstream of the start site, similar to the nucleosome-free region observed for yeast promoters (and shown using ChIP-seq data

in **Fig. 16A**) was reported. We were able to reproduce the same TSS occupancy depletion using the publicly available *Hfx. volcanii* MNase data<sup>108</sup> (**Fig. 16D**); however, this MNase profile contrasts strongly with the HstA ChIP-seq occupancy profile (light blue lines, **Fig. 16C**). By contrast, in the case of yeast histones, the histone ChIP-seq occupancy showing upstream depletion and downstream ordered nucleosomes (**Fig. 16A**) correlates strongly with the known MNase digestion patterns<sup>170</sup>.

Taken together, these data suggest that the start site occupancy patterns of halophilic histones are unique relative those of canonical eukaryotic or archaeal histones as well as TFs, suggesting a divergent DNA binding function. The dissimilarity between *Hfx. volcanii* MNase-Seq and HstA ChIP-Seq occupancy suggests that HstA is not the chromatin protein for *Hfx. volcanii*. This corroborates mass spectroscopic evidence<sup>104,118</sup> that halophilic histone expression is too low to act as the main chromatin protein.

### **3.2.5. Halophilic genomes lack the dinucleotide periodicity that indicates genome-wide optimization for histone binding**

To determine whether halophilic genomes carry a genome-wide 10 bp dinucleotide periodicity signal (GPS) indicative of histone packaging<sup>179,180,184</sup>, we used power spectrum analysis to detect the GPS of AA/TT/TA dinucleotides in the genome sequences of various archaeal model species (**Fig. 19A**, Methods). The genomes of thermophilic species *M. fervidus* and *T. kodakarensis* exhibited a sharp peak in their respective spectral density curves at 10-10.3bp, indicative of periodicity that guides histone binding. This was as expected, as the well-characterised histones of these species are known to wrap DNA and function as chromatin packaging proteins<sup>68,97</sup>. In contrast, periodicity in the

same range was not detected for the four model halophilic species (*Hbt. salinarum*, *Hfx. volcanii*, *Hfx. mediterranei*, *Haloarcula hispanica*) even though their genomes encode histones. For further comparison, we considered organisms in different branches of the tree of life known to use other proteins besides histones to organize their genomes. *Methanosarcina mazei*<sup>103</sup> and *E. coli* show a periodicity of 10.7-11bp instead of 10bp (**Fig 19A**). This slightly longer periodic frequency may be indicative of negative supercoiling<sup>192</sup>. In contrast, *Sulfolobus solfataricus*, which encodes no histones but uses archaeal-specific proteins such as Alba and Cren7 to package its genome<sup>18</sup>, lacks both periodicities.

Having validated our method, we extended first by looking at more archaeal species with published information regarding chromatin proteins. We tested the genomes of *Methanosphaera stadtmanae*, *Methanocaldococcus janaschii*, *Methanothermobacter thermoautotrophicus*, all of which have published work regarding their histones<sup>95,102,193</sup>, as well as *Thermoplasma acidophilum* and *Pyrobaculum calidifontis*, which use non-histone chromatin proteins<sup>72,74</sup>. We again found visible 10bp periodicity in the histone-containing species, missing in those with non-histone chromatin (**Fig. 20A**). Within halophiles, we also examined the periodicity of the GC dinucleotide, because, unlike the genome of archaea known to use histones to package their genomes, halophilic genomes are > 60% G+C<sup>38</sup>. Therefore, the histone binding signal might be revealed by GC dinucleotides, which are also known to play a role in histone binding<sup>194</sup>. *Methanopyrus kandleri*, a GC-rich species<sup>195</sup> whose histone protein has been shown to be capable of nucleosome

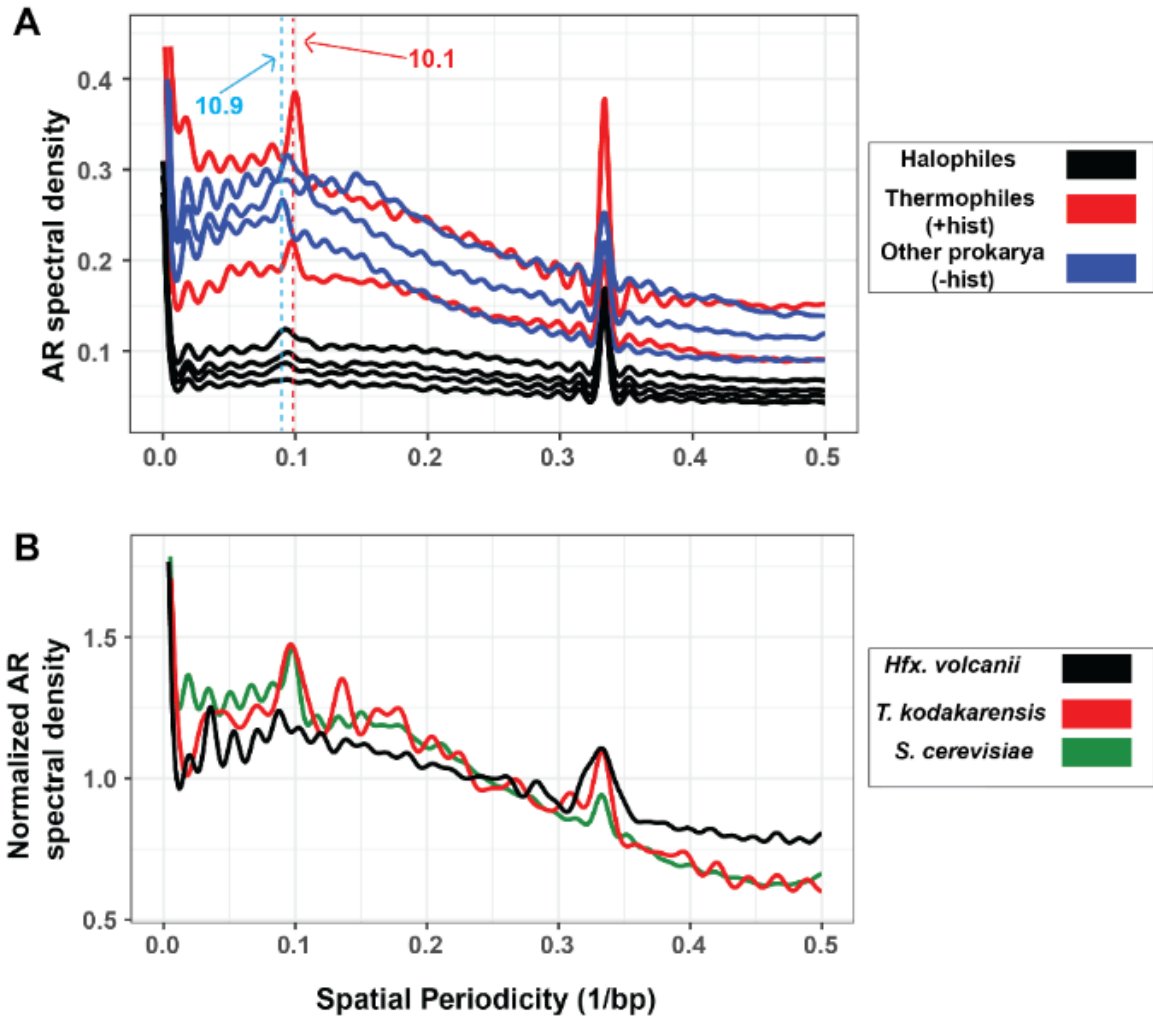
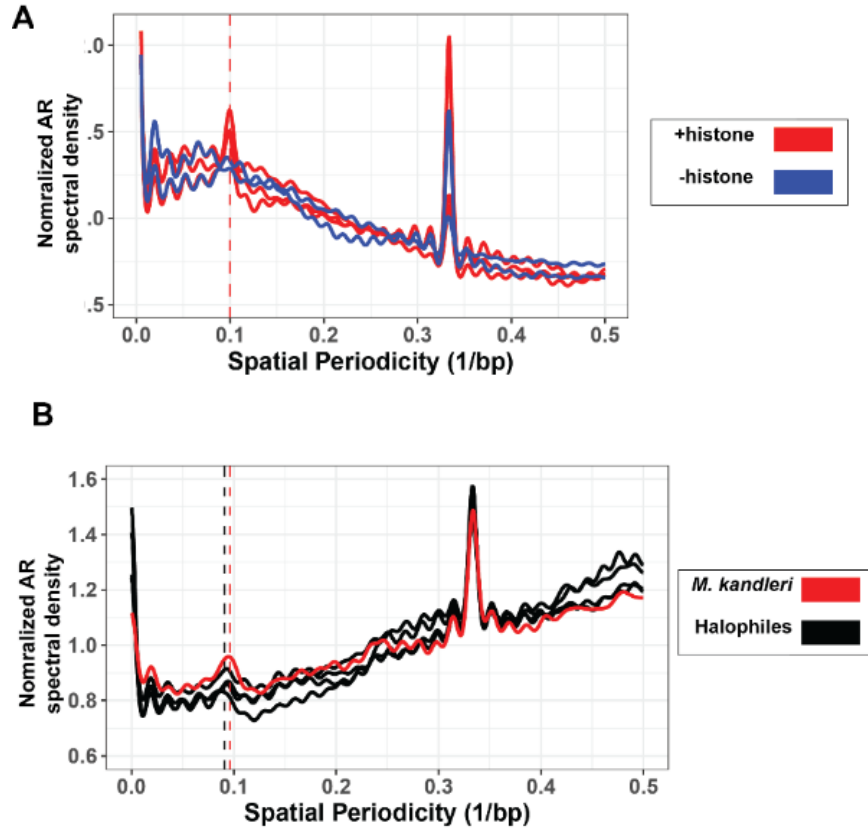


Figure 19: AA/TT/TA dinucleotide periodicity shows histone-linked pattern. (A) Autoregression spectra indicating genome-wide dinucleotide periodicity of thermophilic archaeal species with well-characterized histones (*M. fervidus*, *T. kodakarensis*; red lines), halophilic archaea that encode histones (*Hbt. salinarum*, *Hfx. volcanii*, *Hfx. mediterranei*, *Hca. hispanica*; black traces) and other prokaryotic species (blue traces) that lack histones (*E. coli*, *S. solfataricus*) or with non-histone chromatin (*M. mazei*). Dotted red line indicates ~10.1bp periodicity present in histone-utilizing species (red traces), dotted blue line represents ~10.9bp periodicity (i.e. from supercoiling) detected in some non-histone utilizing species (blue traces). (B) MNase protected regions (nucleosomes) of *S. cerevisiae* (green), *T. kodakarensis* (red), *Hfx. volcanii* (black). Dotted black line indicates the ~10bp periodicity peak (not detected for *Hfx. volcanii*). Normalized spectra are plotted in this panel to facilitate clarity in the visualization.

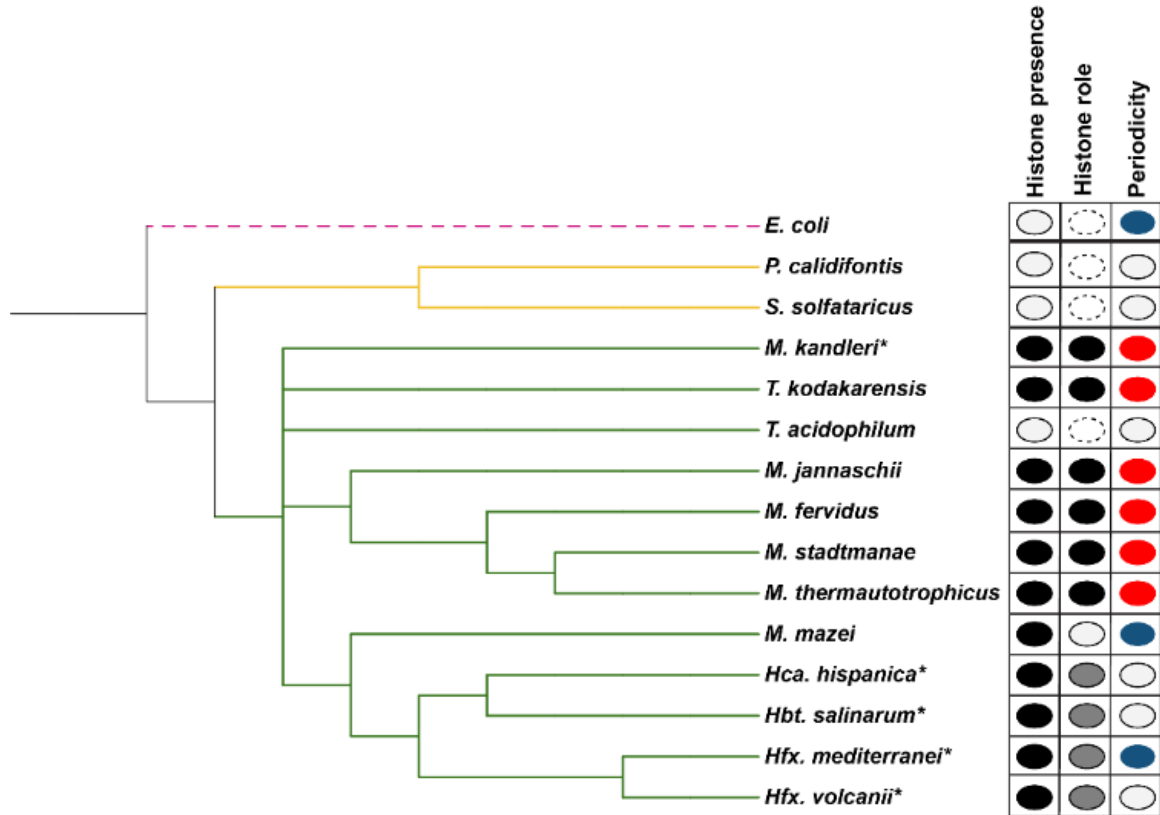


**Figure 20: Additional genome-wide AA/TT/TA and GC periodicities. (A) AA/TT/TA dinucleotide periodicity of archaeal species with characterised histones (red; *M. stadtmanae*, *M. thermoautotrophicus*, *M. janaschii*) and with non-histone chromatin (blue; *T. acidophilum*, *P calidifontis*) (B) Periodicity of GC dinucleotides in species with >60%G+C genomic content: histone-containing *M. kandleri* (red), model halophile species *Hbt. salinarum*, *Hfx. volcanii*, *Hfx. mediterranei*, *Hca. hispanica* (black).**

formation<sup>100</sup>, was also included in this comparison. We did not detect a GC dinucleotide 10bp periodicity in halophilic genomes. However, we did observe a periodicity peak very close to 11bp (Fig. 20B), which, as noted above, is more likely linked to supercoiling than histone binding. On the other hand, ~10bp GC periodicity was indeed observed for *M. kandleri*. In summary, across genomes of representative species from Euryarchaea and Crenarchaea, we observe that 10bp periodicity is strongly associated with encoded

histones that function as chromatin organizers (**Fig 21**). Histone-like proteins of halophiles are clear outliers in this regard.

We next used MNase-seq data from *T. kodakarensis*<sup>97</sup>, *Hfx. volcanii*<sup>108</sup> and the eukaryote *S. cerevisiae*<sup>178</sup> to test if the presence of this dinucleotide periodicity correlates with MNase-protected regions. These nuclease protected regions have been verified to show nucleosome- based binding periodicity *in vitro* for *S. cerevisiae*<sup>178</sup> and *T. kodakarensis*<sup>95</sup>, but not for *Hfx. volcanii*<sup>108</sup>. We detected a strong 10.2-10.4 bp GPS signal in the protected regions of the two species known to have histone-based chromatin (*T. kodakarensis* and *S. cerevisiae*, **Fig. 19C**). This signal is not detected in the nuclease-protected regions of *Hfx. volcanii*. This evidence combined with the discrepancy between *Hfx. volcanii* occupancy profiles from MNase vs HstA ChIP-seq data (**Fig. 16 C,D**) suggests that the histone protein in *Hfx. volcanii* is unlikely to have been the source of previously observed MNase-protected regions<sup>108</sup>. We note that a recent analysis of mass spectrometry<sup>118</sup> has uncovered other potential chromatin proteins in *Hfx. volcanii* that are expressed at much higher levels than HstA, and are therefore stronger candidates than HstA for generating the MNase-seq TSS depletion pattern (reported previously<sup>108</sup> and reproduced in **Fig. 16D**).



**Figure 21: Phylogenetic tree of selected archaeal species (with *E. coli* as outgroup). First column shows presence (black) or absence (white) of histone-encoding genes. Second column documents experimental characterization of the function of encoded histone based on previous publications: compaction (black), non-compaction (white), non-canonical histone function (grey), N/A (dotted line). The third column shows genome-wide AA/TT/TA dinucleotide periodicity: ~10bp (red), ~11bp (blue), no detectable periodicity (white). For halophiles (*Hbt. salinarum*, *Hfx. volcanii*, *Hfx. mediterranei*, *Hca. hispanica*) and *M. kandleri*, genome-wide GC periodicity (instead of AA/TT/TA) was also considered due to these genomes having >60% GC content.**

Taken together, these data suggest that a genome-wide enrichment for AA/TT/TA periodicity (GPS) of ~10bp is strongly and directly correlated with the presence of histones that function as the main chromatin packaging proteins in archaea. The absence of such GPS (and GC periodicity) in halophilic genomes and some archaeal species suggests that their genomic sequence is not optimized for genome-wide histone binding.

### 3.2.6. Halophilic histones differ in predicted DNA binding sequence specificity

Because the periodic signal associated with canonical histone binding was absent in halophile genomes, we next asked how HpyA and HstA may bind DNA using de novo searches for specific cis-regulatory sequence motifs. Site-specific TFs are usually guided to their target genes by preferential binding to a particular sequence motif<sup>161</sup>. This is also true of halophilic TFs, where palindromic motifs have been reported for RosR<sup>35</sup> and TrmB<sup>132</sup> of *Hbt. salinarum*. We used a variety of *de novo* motif searching methods to detect over-represented cis-regulatory sequences in HpyA- and HstA-bound regions in CHIP-seq data. This included *de novo* motif detection programs such as MEME<sup>196</sup>, DNA secondary structures, and over-represented k-mers (Methods; **Appendix C**).

In the case of HstA, MEME detected (E-value  $1.4 \times 10^{-14}$ ) a palindromic sequence in 31 of the 32 CHIP-seq peaks (**Fig 22A; Appendix C**). This motif, of the form TCGNSSNCGA (where S is G or C), was robust to correction for background di- and tri-nucleotide frequencies. Genome pattern scanning analysis using FIMO (part of the MEME suite) detected this motif at 11,630 locations genome-wide, suggesting that HstA may bind additional sites under alternate conditions. In contrast, exhaustive de novo computational searches using multiple methods were unable to detect a sequence-specific binding motif for *Hbt. salinarum* HpyA (See details in **Appendix C**). Instead we asked whether HpyA binding regions specifically exhibit the dinucleotide periodicity known to facilitate histone binding in other species.<sup>95,180,184</sup> Although this periodicity is not present at a genome-wide level, a periodicity of 10.4 bp is detected in HpyA-bound

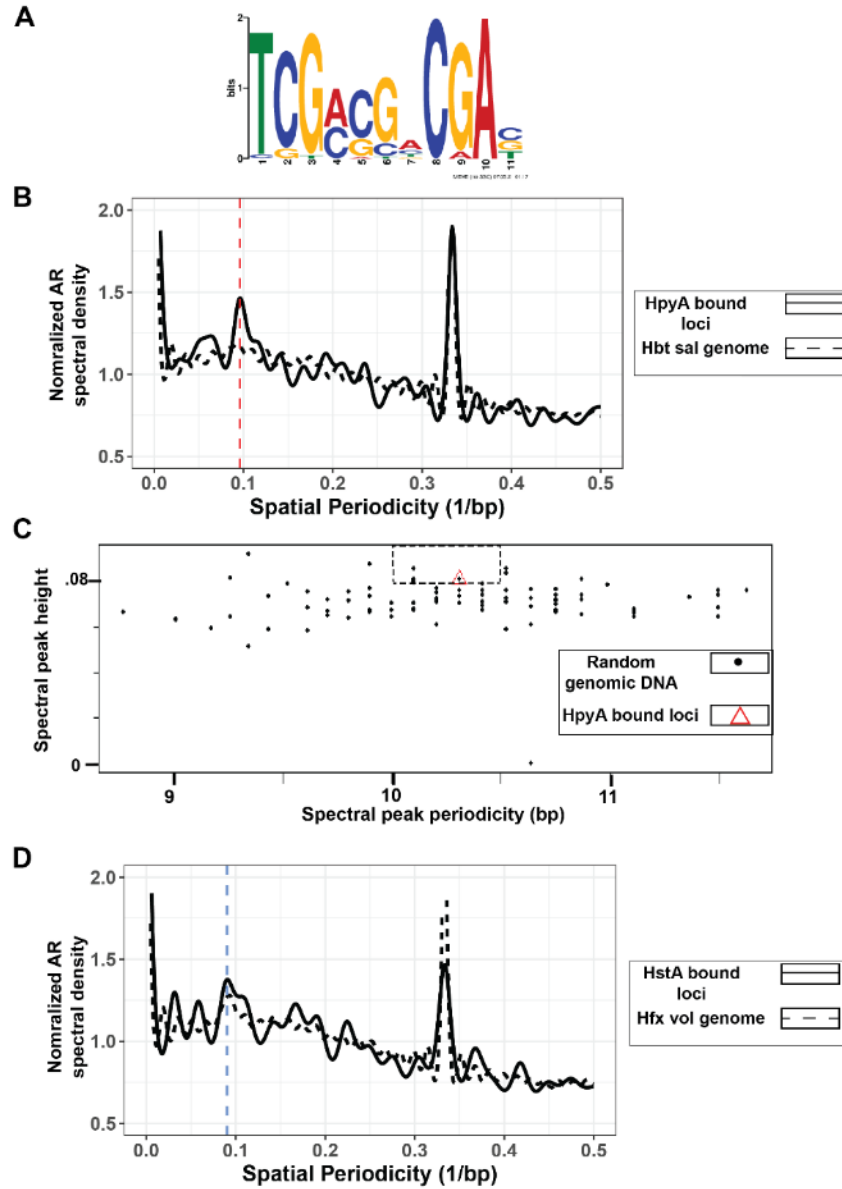


Figure 22: Sequence specificity of HpyA and HstA binding. (A) Motif logo of cis-regulatory sequence detected in HstA-bound sites. Bit scores are shown in the y-axis and bp positions on the x-axis. Motif logo generated by the MEME suite196 output. (B) 10.4bp periodicity present in HpyA-bound loci (solid line) but absent in the *Hbt. salinarum* genome as a whole (dotted line). (C) Comparing randomly-chosen regions of the genome (black dots) with the periodicity of the HpyA-bound loci (red triangle). Dotted rectangle includes those randomly-chosen sequences that show stronger periodicity than HpyA at relevant levels (10-10.5bp). (D) ~11bp frequency of HstA-

**bound loci (solid line) matches periodicity of the entire genome of *Hfx. volcanii* (dotted line).**

regions (**Fig. 22B**). Three of 100 randomly chosen sequences equal to the length of the HpyA-bound regions exhibited greater spectral peak height (indicating stronger periodicity) in the 10-10.5bp range (**Fig. 22C**), suggesting that HpyA may bind additional sites in the genome and/or under alternative growth conditions not yet investigated. This suggests that, like other histones, HpyA favours binding in DNA regions with a ~10bp dinucleotide periodicity. In contrast, HstA target loci exhibited 11-bp periodicity but not ~10bp periodicity (**Fig. 22D**). Indeed, the autoregression curve resembles that of the entire *Hfx. volcanii* genome, leaving the cis-regulatory sequence noted above (**Fig. 22A**) as the only determinant of HstA binding detected in our analysis. Taken together, these data suggest that HpyA favours binding to sequences with a ~10bp periodic presence of A/T dinucleotides, implicating a histone-like binding mode. In contrast, HstA binds a more sequence-specific manner to a palindromic motif, like TFs.

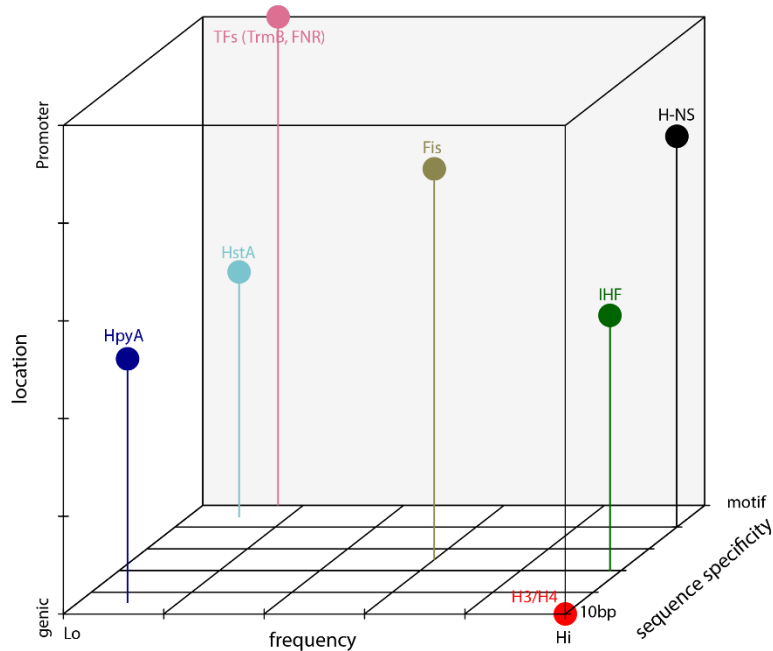
### **3.3. Conclusions**

Here we demonstrate strong functional conservation of histone-like proteins across related species of halophiles, with some subtle differences. The sole histone coding gene of both halophilic species is non-essential for growth (**Fig 10**). However, unlike HpyA of *Hbt. salinarum*, HstA of *Hfx. volcanii* is necessary for growth under optimum conditions. The genome-wide binding patterns of the HstA and HpyA are similar to one another with respect to their mode of binding, number and width of peaks, percentage of

genome covered and lack of preference for genomic features. Interestingly, HpyA and HstA differ in terms of their binding site sequence preferences. In the case of *Hbt. salinarum*, HpyA-bound regions were observed to contain 10bp dinucleotide periodicity. HstA, on the other hand, may bind a TF-like palindromic sequence motif. In comparing halophilic histones to DNA binding proteins across domains of life, we observe a pastiche of conserved features. Considering this evidence, we conclude that HpyA and HstA functions diverged from those of other archaeal and eukaryotic histones. This is consistent with and extends knowledge from our previous characterization of HpyA in *Hbt. salinarum*, where we demonstrated its function as a transcriptional regulator of ion uptake<sup>110</sup>. Given the strong sequence and structural conservation of histones across sequenced halophile genomes<sup>104</sup>, we posit that alternative DNA binding protein functions are likely for all halophilic histones. Our results are consistent with the hypothesis that archaeal histone function varies according to habitat by selection under extreme conditions<sup>118</sup>.

Chromatin proteins are highly expressed<sup>60,118</sup>, and ChIP-seq data demonstrates that their genome-wide binding results in a large number of peaks covering over 10% of the genome,<sup>171,172</sup> or in characteristic occupancy signals in the case of eukaryotic histones<sup>170,190</sup>. Previous proteomics mass spectrometry results from our lab<sup>104</sup> and others<sup>118</sup> demonstrated the low expression level of HpyA in *Hbt. salinarum* and HstA in *Hfx. volcanii* is comparable to that of a TF, and therefore too low to provide architectural organization of the genome. The sparse binding patterns of HpyA and HstA corroborate

this hypothesis (**Fig 13**). More broadly across the archaeal phylogenetic spectrum, histone expression level is strongly associated with chromatinization of the genome<sup>118</sup>. Hence, HpyA and HstA differ from chromatin proteins with respect to the binding frequency necessary for architectural functions. Taken together, these data suggest that the start site occupancy patterns of halophilic histones are unique relative those of canonical eukaryotic or archaeal histones as well as TFs, suggesting a divergent DNA binding function. These results therefore situate halophilic histone-like proteins in a growing group of DNA binding proteins that defy categorization according to the criteria commonly used to differentiate between them (**Fig 23**). Dorman and colleagues posited that traditional definitions of bacterial DNA binding proteins as “TFs” or “NAPs” are insufficient to capture the true continuum of functional characteristics observed for certain proteins<sup>161</sup>. For example, some proteins were defined as NAPs because of their ability to bind genome-wide and alter DNA structure; however, some NAPs can also bind in a highly sequence-specific manner (e.g. IHF). Some TFs like CRP exert sequence-specific control of certain loci, but bind to hundreds of other sites in the genome<sup>197,198</sup>.



**Figure 23: Qualitative 3-D visual representation of binding characteristics of selected DNA-binding proteins investigated in this study. Binding is classified on the basis of sequence specificity, ranging from preference for 10bp periodicity to strict cis sequence motif; frequency as measured by genome-wide coverage and number of peaks, ranging from low to high; location preference, ranging from promoter depletion to promoter preference.**

Such examples are not restricted to bacteria: newly discovered site-specific archaeal TFs are also likely to bend or loop DNA. Examples include the TetR family TF FadR<sup>199,200</sup> and archaeal Lrp family proteins<sup>176,201</sup>. Depending on the locus, Lrp family proteins can bind with or without sequence specificity<sup>173</sup>, exhibit direct or indirect effects of transcription<sup>174</sup>, combining features observed for TFs as well as halophilic histone-like proteins. The DNA binding proteins under investigation in the current study clearly require more flexible functional categorization. We conclude that halophilic histones, with their primary sequence homology to archaeal and eukaryotic histones<sup>62,104</sup>, their

role as transcription regulators<sup>104,110</sup>, and hybrid modes of DNA binding, lie within the unclear divide between TFs, histones, and nucleoid associated proteins.

### 3.4. Materials and methods

#### 3.4.1. Strain construction:

*Halobacterium salinarum* strains used in this study have been described previously<sup>104,110</sup>.

*Haloferax volcanii* wild type strain was DS2<sup>41</sup>. The strains created here used DS2

derivative  $\Delta pyrE$  (strain H26) as the parent strain. The  $\Delta hstA$  (*HVO\_0520*) knockout

strain AKS198 was created from parent H26 using vectors described by Allers *et al*<sup>46</sup> and

the pop-in pop-out double crossover counterselection strategy commonly used for *Hfx.*

*volcanii*<sup>45</sup>. Briefly, the pAKS145 knockout vector was generated by isothermal ligation of

sequences flanking the *hstA* gene into backbone vector pTA131 at the EcoRV site.

Strains, primers, and plasmids used for all strain constructions are noted in **Appendix**

**A.**

AKS214 was the strain used to test in *trans* complementation of  $\Delta hstA$  deletion growth

defect. It contains the pAKS147 plasmid, which was created by inserting *hstA* and 500bp

of its upstream sequence into the pJAM809 backbone at the XbaI and KpnI sites. Two

strains were generated for ChIP-seq experiments. AKS217, the negative control strain, is

the  $\Delta hstA$  background carrying the pJAM809 empty vector. AKS233 is the *hstA* strain

carrying plasmid pAKS180, which was derived from pAKS147 by addition of the

hemagglutinin (HA) tag using the NEB Q5 site-directed mutagenesis kit.

*hstA* deletion from the genome and *hstA* or *hstA-HA* presence *in-trans* was confirmed with PCR and Sanger sequencing of the flanking regions. Deletion was additionally confirmed with full-genome resequencing. Full-genome resequencing for the parent strain and  $\Delta hstA$  strain was analysed using the Breseq<sup>202</sup> analysis tool, results are given in **Appendix B**. The whole genome sequencing data for the  $\Delta hstA$  deletion strain have been deposited in the NCBI Sequence Read Archive at accession PRJNA773760.

### **3.4.2. Media, culturing, and phenotyping:**

*Hfx. volcanii* rich medium was used for routine growth across experiments: Yeast Peptone Casamino Acids (Hv-YPD), as described previously<sup>46</sup>. For plasmid maintenance, media were supplemented with Novobiocin (0.1  $\mu\text{g}/\text{mL}$ ). For construction of deletion mutants, media were supplemented with 5-FOA (300 $\mu\text{g}/\text{mL}$ ) in selection of the second crossover.

To measure growth rates of  $\Delta pyrE$  parent strain and  $\Delta hstA$  strains, at least 3 biological replicate individual colonies of H26 and AKS198 were picked from plates freshly streaked from frozen stock and precultured for 70-80 hrs in 5mL Hv-YPD at 42°C with 225 rpm shaking (referred to as “standard” or “optimum” conditions” throughout). To test growth phenotypes in standard conditions, precultures were diluted to  $\text{OD}_{600} \sim 0.025$  and then cultured in a BioScreen C (Growth Curves USA) at 42°C with fast shaking at maximum amplitude. Each biological replicate culture was inoculated into at least duplicate and up to quadruplicate wells of the microtiter plate to ensure technical reproducibility in the measurements.  $\text{OD}_{600}$  was measured by the BioScreen every 30

minutes over the growth curve. Further details of stress conditions tested and results, are given in **Figure 11**. Resultant growth curves were quantified by measuring the area under the log-transformed growth curve (AUC) Visualization and area under the curve (AUC) analysis of the growth curve was carried out as in

[https://github.com/saaz1291/Halophilic\\_histone\\_binding/tree/main/Growth\\_analysis](https://github.com/saaz1291/Halophilic_histone_binding/tree/main/Growth_analysis).

Growth data for  $\Delta hstA$  and parent strain in optimal conditions is provided in **Appendix B**.

### **3.4.3. ChIP-seq experiment:**

*Haloferax volcanii* HstA-HA ChIP-seq was carried out using the same method as described previously<sup>110</sup>. Briefly, three biological replicate cultures of AKS233 (*hstA*-HA) and 1 replicate of AKS217 as a negative control were grown in 50 mL of YPC18% and harvested at 15-17 hours post-inoculation at an optical density of 0.21-0.33 (mid-exponential phase). Cultures were cross-linked, immunoprecipitated by virtue of the HA tag, and DNA prepared as described previously. Strain details are provided in **Appendix A**. As before, the Duke Center for Genomic and Computational Biology carried out library preparation including adapter ligation. The only difference from previous protocol was use of the Illumina NovaSeq6000 to carry out paired-end sequencing. The ChIP-seq data have been deposited in the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) accession number GSE186415.

#### 3.4.4. ChIP-seq analysis:

Publicly available ChIP-seq data from the relevant TF, NAP, or histone was downloaded from the NCBI sequence Read Archive using the fastq-dump feature from SRAToolkit 2.9.0 (<https://hpc.nih.gov/apps/sratoolkit.html>). Details of published datasets used for bacterial and archaeal TFs, histones, and NAPs, including sequence read archive trace numbers, are provided in **Appendix B**. The bacterial NAP HU was excluded from this analysis because it does not show IP/Input enrichment, indicative of binding, like other TFs/NAPs do<sup>172</sup>. For ChIP-seq experiments done in the Schmid lab, the fastq files were available directly. Fastq files were converted to sorted BAM files, wig files, and per-base read-depth text files were generated as described previously<sup>110</sup> ([https://github.com/saaz1291/Halophilic\\_histone\\_binding/blob/main/README.md](https://github.com/saaz1291/Halophilic_histone_binding/blob/main/README.md)).

Sorted BAM files were used to generate peak lists using the R package MOSAiCS, and the MOSAiCS workflow is noted in [https://github.com/saaz1291/Halophilic\\_histone\\_binding/tree/main/Peak\\_calling](https://github.com/saaz1291/Halophilic_histone_binding/tree/main/Peak_calling). Peak lists were created for haloarchaeal TFs (*Hca hispanica* RosR and TrmB, *Hfx volcanii* RosR and TroR, and *Hfx mediterranei* RosR), selected bacterial NAPs (IHF, H-NS, FIS), and bacterial TF (FNR). The bacterial TF VjbR was excluded from this analysis because the vast majority of the peaks called by Mosaics could not be visually confirmed. For experiments where more than one replicate from the same conditions was present, multiIntersectBed from Bedtools<sup>128</sup> was used to combine peaks across replicates, and only peaks present in the majority of replicates were considered.

The peak list for HpyA was taken from previously published work<sup>110</sup>. The peak list for HstA was created as described above, but with some manual curation (details below).

The average width and total area covered by the peaks within these lists was calculated within Excel, and total area covered was expressed as a percentage of genome length

(**Appendix B**). This was then graphed with simple code

([https://github.com/saaz1291/Halophilic\\_histone\\_binding/tree/main/Bindingfeatures](https://github.com/saaz1291/Halophilic_histone_binding/tree/main/Bindingfeatures)).

We caveat this analysis by acknowledging that this measure of peak width is a gross simplification, but nonetheless was enough to distinguish the differences seen within this dataset.

Peaks were classified as “intergenic” or “coding” on the basis of where in the genome they were located. The centre of each peak was found, and was determined to be within or outside a coding region (as described by the list of genes in the NCBI gene table for that species). The code used to make this classification, and to graph the results, is in

[https://github.com/saaz1291/Halophilic\\_histone\\_binding/tree/main/Bindingfeatures](https://github.com/saaz1291/Halophilic_histone_binding/tree/main/Bindingfeatures).

The results of this classification were used as the basis of a hypergeometric test in R using the phyper function: `phyper(#peaks in non-coding regions, length of genome that is non-coding, length of coding genome, #total peaks)` to determine if peaks were over-represented in intergenic regions (**Appendix C**).

#### **3.4.5. Generating *Hfx volcanii* HstA peak list:**

As mentioned above, Mosaics was used to generate peak lists from HstA ChIP-Seq data, and peaks common in at least 2 of 3 replicates were retained to make a joint peak list.

This list was then curated manually in order to remove false positives caused by changes in input control sequencing, transposase and integrase-caused local duplications, and peaks common with the HA tag-alone input control. The final manually curated peak list for HstA is noted in **Appendix B**.

### **3.4.6. Start site occupancy analysis:**

Per-base read-depth text files were generated as described above. These text files were used as inputs for occupancy analysis, alongside genome annotations downloaded from NCBI (details in **Appendix B**). Code was written ([https://github.com/saaz1291/Halophilic\\_histone\\_binding/tree/main/TSSgraphs](https://github.com/saaz1291/Halophilic_histone_binding/tree/main/TSSgraphs)) that returns a matrix where each row corresponds to a single gene, and the columns represent sequence depth at positions from  $\pm 400$  bp of that start site, normalized to the average depth over the whole chromosome. (For yeast analysis, considering the larger size of intergenic regions, and the known length of DNA bound to a single nucleosome as 147bp, this analysis was repeated by changing the boundaries to  $\pm 800$ bp). The occupancy graph was generated by taking the average of occupancy across all start sites (rows in the matrix).

Note that the start site used here refers to the ORF translation start site, instead of the transcription start site (TSS) that is often used for these analyses. Three criteria motivated this choice: (a) ORF start sites are better annotated in most species including halophilic archaea; (b) ORF and transcription start sites are very close in halophiles, with

>60% of ORFs being leaderless in *Hfx. volcanii*<sup>203</sup>; (c) using ORF start sites, we were able to reproduce previously seen patterns for yeast TSS<sup>170</sup> (**Fig. 16A**).

### **3.4.7. Dinucleotide periodicity analysis:**

FASTA files containing the genome sequence of the relevant species were downloaded from the NCBI website (species and download details in **Appendix B**) and were analysed using custom R scripts. In brief, dinucleotides (AA/TT/TA) are detected in each genome, and binarized: locations with these dinucleotides are marked as 1 and the rest of the genome as 0. Then, the autoregression spectrum `spec.ar()` function in R (with default parameters) was used to estimate the spectral density of this binary signal, which indicates the periodicity of the selected dinucleotides using an autoregression fit. For facilitating clarity in visualization of autoregression curves, periodicity was normalized by the average signal in **Figures 19B, 20A, B, and 22 B,D**. The entire workflow is noted in

[https://github.com/saaz1291/Halophilic\\_histone\\_binding/tree/main/Periodicity\\_genome\\_wide](https://github.com/saaz1291/Halophilic_histone_binding/tree/main/Periodicity_genome_wide). The same analysis was carried out for GC dinucleotides to generate **Figure 20B**.

For nucleosome enrichment analysis, data regarding the centre of the nucleosomes was downloaded from supplementary information of Brogaard et al 2012<sup>178</sup> (for *Saccharomyces cerevisiae*), Maruyama et al 2013<sup>97</sup> (for *Thermococcus kodakarensis*), and Ammar et al 2012<sup>108</sup> (for *Haloflex volcanii*). Sequences of the length of a typical nucleosome (150bp for eukaryotes, 30-60bp for archaea) were isolated around each centre and the same analysis as above was carried out. This procedure is noted in

[https://github.com/saaz1291/Halophilic\\_histone\\_binding/tree/main/Periodicity\\_nucleosomes](https://github.com/saaz1291/Halophilic_histone_binding/tree/main/Periodicity_nucleosomes).

Note that the strong peak at  $0.33\text{bp}^{-1}$  (3bp) seen in all these spectra is linked to codon usage; it is present in all species and is not linked to histone binding<sup>204</sup>.

Depending on the AT content of the sequence being examined, some of the spectra have an increasing or decreasing slope resulting from slight deviations in A+T content locally; this too is not linked to histone binding<sup>204</sup>.

The results of the genome-wide periodicity obtained and shown in **Figures 19A** and **20** are summarized in **Fig 21**. The phylogenetic tree for this figure was made using the Integrated Tree of Life (iTOL, <https://itol.embl.de>).

#### **3.4.8. Motif search:**

Bed files containing peak locations from HstA and HpyA ChIP-seq data were converted to FASTA format using the Bedtools<sup>128</sup> getfasta command. These FASTA files were used as input for various motif and over-represented sequence determining programs. We used motif-detection with MEME<sup>196</sup> and Homer<sup>205</sup>, k-mer detection tool KMAC<sup>206</sup>, and a DNA secondary structure detection R-package called gquad (<https://cran.r-project.org/web/packages/gquad/index.html>). Finally, the fasta-get-markov tool of MEME was used to determine background mono-, di- and tri-nucleotide frequencies. A more detailed description of the parameters used for each program, and the results of the searches, is provided in **Appendix C**.

For obtaining periodicity of sequences bound by HpyA or HpyA, the FASTA file containing all the peaks (generated as above) was merged into a single line. Finally,

periodicity of the sequences in this FASTA file were analysed. A similar to the procedure used to analyze genome-wide periodicity was used here. The obtained periodicity was compared with randomly chosen sequences from the genome roughly equal in length to the width of ChIP-seq peaks (peak widths given in **Appendix B** and in reference<sup>110</sup>). These simulated peak lists were then analysed using the autoregression spectrum scripts, and results compared between the 100 simulated sequences and the empirically detected peaks. The full workflow is noted in

[https://github.com/saaz1291/Halophilic\\_histone\\_binding/tree/main/Periodicity\\_ChIPSeq](https://github.com/saaz1291/Halophilic_histone_binding/tree/main/Periodicity_ChIPSeq)

.

### **3.5. Acknowledgements**

The authors thank all Schmid lab members for their feedback on the study and comments on the manuscript. S.S. acknowledges the support of his graduate thesis committee for their comments and advice on the study (David McAlpine, Amy Grunden, Richard Brennan), and discussions with Antoine Hocher and Prof. Tobias Warnecke. Funding for the study was provided by grants MCB-1651117 and 1936024 from National Science Foundation to AKS.

## **4. Comparative Analysis of rRNA removal methods for RNA-seq Differential Expression in Halophilic Archaea**

Chapter 4 is modified from a manuscript in preparation for publication. The authors are Saaz Sakrikar, Mar Martinez-Pastor (co-first authors), and Prof. Amy Schmid. All three devised the project and wrote the manuscript. A.K.S. secured the funding for the project and contributed to data analysis. M.M.P. and S.S. carried out experimental work, and S.S. did data analysis, under supervision of A.K.S.

### **4.1. Introduction**

The expression of the genomic information of an organism depends on the cell status and environmental factors that determine the phenotype. The genes that are being transcribed (or read) define the transcriptome, and the compendium of methods that enable the study of the expression of large number of genes simultaneously is known as transcriptomics.

RNA sequencing (RNA-seq) has emerged as a widely used approach to transcriptome profiling with high throughput, sensitivity, dynamic range, and relatively low cost compared to former methods such as microarrays. The first successful RNA-seq experiments were performed using eukaryotic model organisms<sup>207-210</sup>; however, using this tool for understudied models such as archaea has been challenging despite their biological and evolutionary importance.

Archaea are prokaryotic microorganisms that were defined as the third branch of life in the late 70's, when Carl Woese and colleagues found enough genetic differences to

cluster them as a distinct group separate from bacteria and eukaryotes<sup>211</sup>. Even though the first archaea were identified in extreme environments, archaeal species are now known to be diverse and abundant, colonizing a vast array of habitats (from oceans to human skin to extreme environments<sup>13,212</sup>). Therefore, differential expression analysis using transcriptomics in archaea is an important step for a better understanding of responses to diverse environments<sup>36,213</sup>. Such studies would further advance knowledge of the unique molecular biology of archaea, since they combine the molecular characteristics of both bacteria and eukaryotes, such as transcriptional regulation<sup>6</sup>.

Despite previous progress on differential expression by RNA-seq in archaea<sup>214</sup>, this method has recently become unavailable because ribosomal RNA (rRNA) in archaeal transcriptomes can reach more than 90% of the total cellular RNA (**Fig. 26**). Previously, archaeal transcriptomics studies successfully depleted rRNA using commercially available reagent kits for rRNA removal in bacteria<sup>215-219</sup>. However, these kits were discontinued in 2018. Some studies report RNA-seq without rRNA removal, but these studies aimed at different purposes that are possible without rRNA removal (e.g. transcription start site mapping<sup>220</sup>, small RNA detection<sup>214</sup>, etc). However, as we report in the current work, ribodepletion is a key step for reliable RNA-seq results because high rRNA reads can preclude the detection of messenger RNA (mRNA) reads. rRNA removal enables higher sequencing depth of mRNA, leading to better detection of transcripts. This is critical for analyzing differential expression, particularly when detecting non-coding or low-expression RNAs<sup>221</sup>. Previous studies have suggested a minimum sequencing depth of two<sup>222</sup> to ten<sup>223</sup> million reads per sample for obtaining

reproducible results for differential expression, while the ENCODE consortium<sup>224</sup> mandates 30 million reads (albeit for much larger human genomes). Such sequencing depth enables sound statistical comparisons of differential expression on a per-gene basis: at least 5 reads per gene are typically needed to detect the significance of change in expression for a given gene<sup>221</sup>. Removing rRNA also substantially reduces the cost of RNA-seq, enabling extensive sample multiplexing in a single sequencing run, especially for relatively small archaeal genomes.

In this work, we have used four species of halophilic archaea that have been widely used as model organisms in the archaeal research community: *Halobacterium salinarum* (HBT) and *Haloarcula hispanica* (HAH) of the family Halobacteriaceae require salt concentrations close to saturation, whereas *Haloferrax volcanii* (HVO) and *Haloferrax mediterranei* (HFX) of the family Haloferacales colonize lower salinity environments. These four species are highly tractable models for extremophilic microorganisms given their relatively fast generation time (2-6 hours in rich medium), facile genetic tools<sup>44,47,225</sup>, and highly curated genomic annotations and databases<sup>109,226,227</sup>. Establishing a set of tools and best practices for transcriptomics methods would therefore greatly facilitate advances in this field.

Archaeal RNA, like that of bacteria, lacks a 3' polyA tail and so rRNA cannot be removed by polyT tagging. Here we test two methodologies for rRNA depletion in archaea using (a) biotinylated probes and (b) enzymatic digestion, with default as well as sequence-specific probes customized for particular species of interest. The first

approach (biotinylated probes/streptavidin beads) consists of a physical removal of rRNA by hybridizing with a pool of biotinylated oligo probes. These probes are then be captured and removed from the RNA sample using streptavidin-coated magnetic beads. The enzymatic removal of rRNA consists of generating DNA-rRNA hybrids by incubating specifically designed DNA probes complementary to rRNA. Hybrids are then treated with RNaseH that catalyzes the cleavage of RNA when it is bound to a DNA substrate.

Here we report that the two methods are equally successful for removing rRNA across the four species of halophilic archaea growing in diverse media. Both methods can be used successfully with probe sequences custom-designed for one species or with a broad probe pool designed to target multiple species simultaneously. We show that bacterial rRNA probes are sufficiently divergent in sequence to preclude the use of recently developed custom and commercial bacterial rRNA probe sets in archaea<sup>228</sup>. These methods are robust to varying culturing conditions (rich and defined media). This analysis has achieved the goal of identifying an efficient and broadly useful strategy for depleting undesirable archaeal rRNA prior to sequencing for successful transcriptomics.

## **4.2. Material and methods**

### **4.2.1. Media, Strains, and Growth Conditions**

All strains, media and growth conditions used in this study are summarized in **Tables 2** and **3**.

**Table 2: Strains used in this study.**

Name	Species abbreviation	Genotype	Reference genome
MDK407 <sup>44</sup>	HBT	$\Delta$ <i>ura3</i>	GCF_000006805.1_ASM680v1
DS2 <sup>45</sup>	HVO	$\Delta$ <i>pyrE</i>	GCF_000025685.1_ASM2568v1
ATCC33500 <sup>47</sup>	HFX	$\Delta$ <i>pyrE</i>	GCF_000306765.2_ASM30676v2
DF60 <sup>47</sup>	HAH	$\Delta$ <i>pyrF</i>	GCF_000223905.1_ASM22390v1

**Table 3: All media recipes used for test organisms in this study**

Name	Species	Ingredients (per L)	Supplement	pH
CM (rich media)	HBT	250g NaCl (Fisher Chemicals); 20g MgSO <sub>4</sub> .7H <sub>2</sub> O (Fisher Chemicals); C <sub>6</sub> H <sub>5</sub> Na <sub>3</sub> O <sub>7</sub> .2H <sub>2</sub> O (Fisher Chemicals); 2g KCl (Fisher Chemicals); 10g bacteriological peptone (Oxoid)	50 ml uracil (1mg/ml) (Acros Organics)	6.8
YPC 18% (rich media <sup>46</sup> )	HVO and HFX	144g NaCl (Fisher Chemicals); 4.2g KCl (Fisher Chemicals); 18g MgCl <sub>2</sub> .6 H <sub>2</sub> O (Fisher Chemicals); 20g MgSO <sub>4</sub> .7H <sub>2</sub> O (Fisher Chemicals); 12ml 1M Tris.HCl (Fisher Chemicals) pH7.5; 5g yeast extract (Fisher Chemicals); 1g 10g bacteriological peptone (Oxoid); 1g Casamino acids (VWR)	50 ml uracil (1mg/ml) (Acros Organics)	7.5
PR 18% (minimal media)	HVO	170g NaCl (Fisher Chemicals); 70g MgCl <sub>2</sub> .6 H <sub>2</sub> O (Fisher Chemicals); 7g KCl (Fisher Chemicals); 5ml 1M TrisHCl (Fisher Chemicals) pH7.5; 5 ml 1M NH <sub>4</sub> Cl; 2ml 0.25M K <sub>2</sub> HPO <sub>4</sub> ; 5ml 1M NaHCO <sub>3</sub> ; 0.8 ml thiamine (1mg/ml); 0.1 ml biotin (1mg/ml); 0.5% glucose.	50 ml uracil (1mg/ml) (Acros Organics)	7.2
YPC 23% (rich media <sup>229</sup> )	HAH	180g NaCl (Fisher Chemicals); 4.2g KCl (Fisher Chemicals); 18g MgCl <sub>2</sub> .6 H <sub>2</sub> O (Fisher Chemicals); 20g MgSO <sub>4</sub> .7H <sub>2</sub> O (Fisher Chemicals); 12ml 1M TrisHCl (Fisher Chemicals) pH7.5; 5g yeast extract (Fisher Chemicals); 1g 10g bacteriological peptone (Oxoid); 1g Casamino acids (VWR)	50 ml uracil (1mg/ml) (Acros Organics)	7.5

For routine culturing in these media, each species was freshly streaked from frozen stock. Single colonies from no more than two weeks old plates were inoculated in triplicate in 3 ml of rich or minimal liquid media (**Table 3**) and grown aerobically until saturation (stationary phase) at 42°C with continuous shaking at 225 rpm. From each saturated pre-culture, 50ml cultures were initiated from OD<sub>660</sub>=0.1 in 150 ml Pyrex flasks, and 3ml of each were harvested in mid exponential phase OD<sub>660</sub>=0.4-0.8 (doubling times and incubation times included in **Table 4**), by centrifugation in a tabletop centrifuge (5424, Eppendorf) at 21,130 x g for 3min. Supernatant was discarded, and pellets were immediately snap-frozen in liquid N<sub>2</sub> and stored no longer than 3 weeks at -80°C.

**Table 4: Doubling time and Incubation time for different species in different media.**

Species	media	doubling time (h)	days until lag phase
HBT	CM	6	3
HVO	YPC18%	3	2.5 (36h)
HVO	PR18%	12	3
HFX	YPC18%	2.5	2
HAH	YPC23%	6	3

#### **4.2.2. RNA-seq experimental protocol**

Total RNA was extracted from pellets using Absolutely RNA Miniprep kit (Agilent Technologies, Santa Clara, CA) according to manufacturer's instructions. The obtained RNA concentration and integrity was quantified by Nanodrop One (Thermo Scientific, Grand Island, NY) and RNA electropherograms, Bioanalyzer 2100 Instrument with the RNA 6000 Nano kit (Agilent Technologies, Santa Clara, CA), respectively. RNA was checked for DNA contamination by PCR using 200-300ng of input RNA and primers

given in **Table 5** for 30-35 cycles. Extracted RNA was in all the cases high quality, with Bioanalyzer RNA integrity number (RIN) greater than 8.

**Table 5: Primers used to check for genomic contamination.**

Species	Forward primer sequence 5'-3'	Reverse primer sequence 5'-3'	fragment size
HBT	CGACATTCGGGTTGCGTTGT G	GGCGTTGTTACGAAGCA	1372
HFX	CACATCAGCGAGGAGTTTG A	GACAGACGACGAGTTGGTC A	162
HVO	AGAAGTACAAGGGCGTCGA A	TTTTCGAACTCCTCGCTGAT	171
HAH	GCCGATTGCTCCGTCTACTA	ACTGCTCGGTGAGAAACGT C	161

Ribosomal RNA was removed using the following reagent kits and methodologies, abbreviated throughout the text and figures as indicated below:

1. Biotinylated probes with streptavidin bead pull-down:
  - a. Discontinued Ribo-Zero rRNA Removal Kit (Bacteria). Abbr: RZ
  - b. siTools HVO RiboPOOL™ with probes specific for HVO. Abbr: rP-HVO
  - c. siTools Pan-Archaea riboPOOL™ (probes included). Abbr: rP-PA
2. RNase H and enzymatic depletion-based protocols with magnetic bead pull-down:
  - a. Ribo-Zero Plus Kit (probes included). Abbr: RZ+
  - b. NEBNext Bacteria rRNA depletion Kit (New England Biolabs) with default probes from NEB. Abbr: NEB-B
  - c. NEBNext Depletion Core Reagent Set with customized sequence-specific probes for HVO (Table S3). These probes were designed using the NEB

web tool (<https://depletion-design.neb.com/>) and ordered from IDT technologies (idtdna.com). Abbr: NEB-HVO

RNA input to each depletion kit was 300-500 ng. Ribodepletion was performed according to the manufacturer's manuals using default or custom-designed probes as well as modifying time of enzymatic incubation with RNaseH. These details and ordering information are specified in **Appendix B**.

Library preparation from 1-10ng rRNA-depleted RNA was performed using NEBNext UltraII Directional RNA Library Preparation Kit (Illumina, #E7760) following the vendor protocols and complementing cleaning steps with NEBNext Sample Purification Beads (#E7767). An extra-cleaning step using the same type of beads was carried out when samples showed contamination with adaptor dimers. The obtained library quality and concentration was assessed by monitoring the distribution of the fragment sizes with a Bioanalyzer 2100 instrument using DNA High Sensitivity reagent kit (Agilent Technologies, Santa Clara, CA). This size and quantity information was used for pooling the libraries in equimolar concentrations to normalize each library. Libraries were subjected to HiSeq2500, HiSeq4000, or NovaSeq6000 by the Sequence and Genomics Technologies Facility at Duke University. Additional experimental metadata, results, and details are given in **Appendix B**.

### **4.2.3. Data Analysis**

#### **4.2.3.1. Publications on archaeal RNA-seq per year:**

Data regarding the number of publications yearly available from National Center for Biotechnology Information (NCBI) PubMed database

(<https://pubmed.ncbi.nlm.nih.gov/>) was searched with the phrases “archaea”, “RNA-Seq”, and “archaea RNA-Seq” Database hits were downloaded from the NCBI PubMed database on November 1, 2021. The publication of Carl Woese’s seminal paper regarding the classification of Archaea in 1977<sup>211</sup> was used as the starting date. The downloaded data is in **Appendix B**. The code used to generate **Fig. 24** is in [https://github.com/saaz1291/rRNA\\_analysis](https://github.com/saaz1291/rRNA_analysis).

#### **4.2.3.2. RNA-seq data processing:**

FASTQ files generated by sequencing were downloaded and processed as described previously<sup>110</sup>. Files were quality-checked using FastQC, adapter sequences were trimmed using TrimGalore! with cutadapt (FastQC and TrimGalore! downloaded from <http://www.bioinformatics.babraham.ac.uk/projects/>). Trimmed files were aligned to the reference genomes of the four species of interest (Table 2) using Bowtie2<sup>125</sup>. Resultant SAM files were converted into a compact BAM file using SAMtools<sup>126</sup> to generate, sort, and index reads. BAM files were used as the input for HTSeq-count<sup>134</sup> to generate a count file, assigning a numeric raw count of reads to each gene. Details regarding the full workflow are included in reference. To determine rRNA percentage remaining following depletion, the counts corresponding to each of 16S, 23S, and 5S rRNA genes was divided by the total number of raw counts mapping to all genes. The ratio was multiplied by 100 to yield a percentage. These genes are listed in **Table 6**.

**Table 6: rRNA-coding gene identifiers for each species of interest**

Species	rRNA type	Gene identifier(s)	Alternate gene identifier(s)
HBT	16S	VNG_RS09790	VNG_r02
	23S	VNG_RS09800	VNG_r03
	5S	<i>rrf</i>	VNG_r04
HVO	16S	HVO_RS13015, HVO_RS18920	HVO_3038, HVO_3064
	23S	HVO_RS13025, HVO_RS18910	HVO_3040, HVO_3062
	5S	<i>rrf</i>	HVO_3041, HVO_3061
HFX	16S	HFX_RS14380, HFX_RS08900	HFX_1820, HFX_2933
	23S	HFX_RS14370, HFX_RS08910	HFX_1822, HFX_2931
	5S	<i>rrf</i>	HFX_2930, HFX_1823
HAH	16S	HAH_RS08910, HAH_RS01110	HAH_1834, HAH_0232
	23S	HAH_RS08905, HAH_RS01120	HAH_1833, HAH_0234
	5S	<i>rrf</i>	HAH_1832, HAH_0235

The results, expressing all rRNA, 16S rRNA, 23S rRNA, and 5S rRNA as a percentage of total reads, are listed in **Appendix B**, with a shorter summary of results in **Table 7**. The code used to generate **Figs. 25,26, 28-31** is in

[https://github.com/saaz1291/rRNA\\_analysis](https://github.com/saaz1291/rRNA_analysis), and the input to the code is also given in **Appendix B** under the appropriate tabs.

#### **4.2.3.3. Probe specificity analysis:**

Sequences of probes custom-designed for HVO rRNA removal using the NEB website (<https://depletion-design.neb.com/>) were compared to HBT strain NRC-1 genome sequence using NCBI BlastN search with default parameters (NCBI taxonomy ID: 64091; NCBI access date May 4, 2021). The resultant sequence identity (expressed as a percentage) was noted for each of the 117 sequences. These data were classified into 4 categories: 100% identity, 90-99% identity, <80% identity, and no significant similarity.

The probe sequences, BLAST results, and identity percentages are listed in **Appendix B**

and results are shown in **Fig. 28**. The code used to generate **Fig. 28** is in [https://github.com/saaz1291/rRNA\\_analysis](https://github.com/saaz1291/rRNA_analysis) and the specific inputs to generate this figure are in the appropriate tabs within **Table S3**.

#### **4.2.3.4. Count correlations:**

RNA-seq read counts corresponding to all genes outside of rRNA genes for different rRNA removal methods and replicates in HBT and HVO were calculated as described above. Each gene's count was expressed as a percentage of total counts, and the arithmetic average of all replicates using a particular method was calculated. These average values for each gene for a given removal method were then noted in **Appendix**

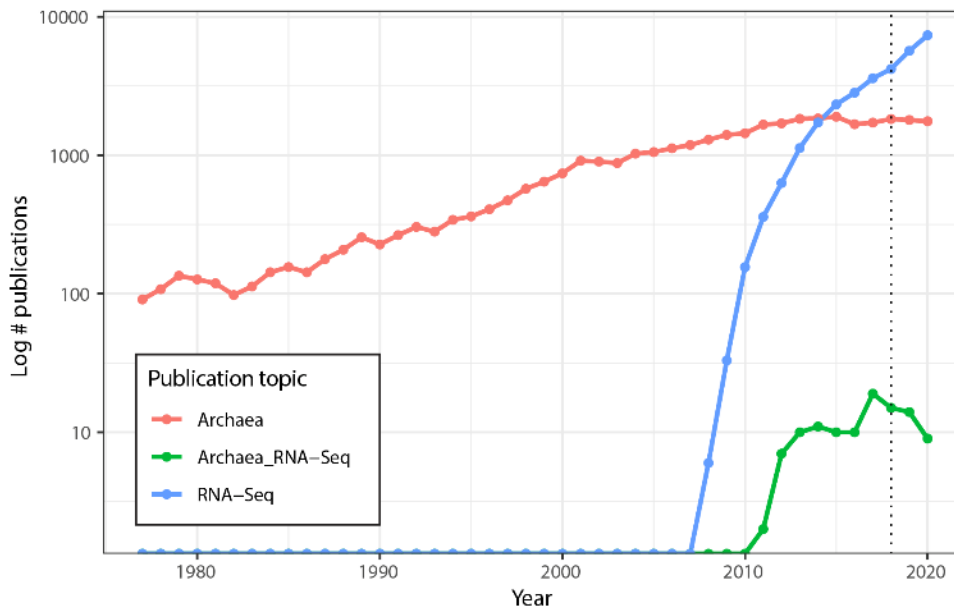
**B**. The code used to generate **Fig. 31** is in [https://github.com/saaz1291/rRNA\\_analysis](https://github.com/saaz1291/rRNA_analysis) and the specific inputs to generate this figure are in the appropriate tabs within **Appendix B**.

#### **4.2.3.5. Power analysis:**

RNA-Seq data generated from a pilot run for a published project<sup>110</sup> from the Schmid lab was inputted into the power optimization tool Scotty ([scotty.genetics.utah.edu](http://scotty.genetics.utah.edu))<sup>230</sup>. This was used to assess power for differential expression experiments involving upto 6 biological replicates with between 1 and 15 million reads mapping to genes for each replicate, so that at least 75% of 2-fold differentially expressed genes could be detected at  $p < 0.01$ .

## 4.3. Results

### 4.3.1. Discontinuation of the Illumina RiboZero kit is associated with a decline in published archaeal RNA-Seq studies:



**Figure 24: Slowdown in Archaeal RNA-Seq publications in recent years. Lines depicting number of publications per year detected in the NCBI PubMed databased searched with the terms “Archaea” (red), “RNA-Seq” (blue), and “Archaea RNA-Seq” (green), plotted on log-scale y-axis. Dotted line at 2018 marks discontinuation of Ribozero kit.**

RNA-Seq of archaeal species belonging to diverse clades has previously been facilitated by rRNA depletion using the bacterial Ribo-Zero kit from Illumina<sup>215-219</sup> (Methods). However, the kit was discontinued in 2018. To determine the impact of this discontinuation, we conducted a comprehensive literature search on the PubMed database for articles reporting on archaea (1977-present) and on RNA-seq in archaea (2010-present). The discontinuation of the Ribo-Zero kit appears to correlate with a plateau and decline of papers published on the topic of RNA-seq in archaea, even as the

**Table 7: Summary of percentage of rRNA removal left using different methods on the 4 model species; medians of 3-4 biological replicates (abbreviations from Methods)**

Species	rRNA depletion method	%age rRNA remaining
HBT	None	95.02
HBT	RZ	35.3
HBT	RZ+	91.8
HBT	NEB-B	85.75
HBT	NEB-HVO	80.5
HBT	NEB-HVO-120	75
HBT	rP-PA	3.55
HVO	RZ	0
HVO	rP-PA	4E-4
HVO	rP-HVO	8E-6
HVO	NEB-HVO	8E-3
HME	rP-PA	0.02
HCA	rP-PA	0.4

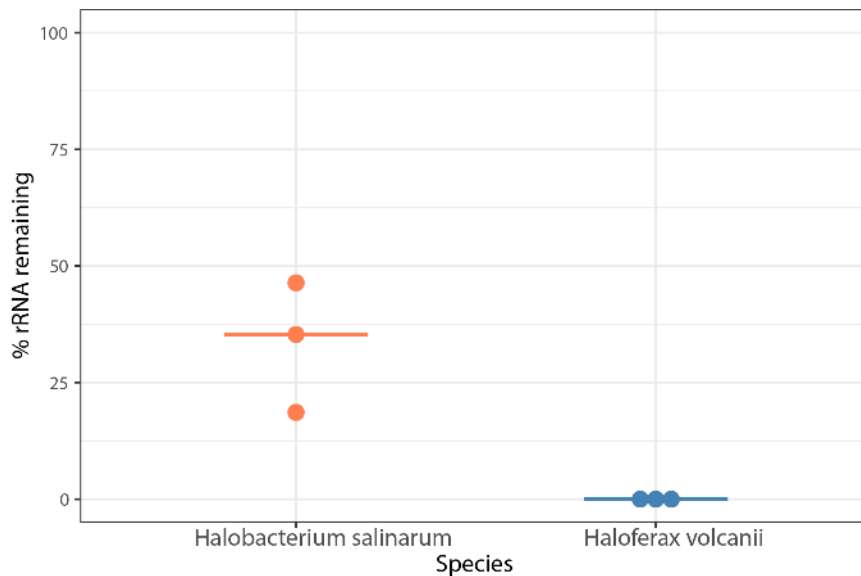
number of publications on archaea in general and on RNA-Seq in other domains of life has grown (**Fig. 24**).

Within our lab, we had successfully used this kit on two model halophile species, HBT<sup>216</sup> and HVO (Mar Martinez-Pastor, unpublished data). The RiboZero kit used biotinylated RNA probes designed to deplete abundant rRNA transcripts from bacterial total RNA with streptavidin beads. We observed 100% removal of rRNA from HVO total RNA samples (**Fig. 25**). In contrast, removal from HBT was variable, with a median rRNA value of 35% (range 18.7% - 46.4%; **Fig. 25; Table 7**), at a level which allowed analysis of differential expression<sup>216</sup>. Because RNA-seq transcriptomic profiling studies across halophilic archaea are valuable to understand responses to environmental perturbation, we were hence motivated to find a suitable replacement capable of matching or bettering this performance across four model species of

halophiles routinely used in our lab (HBT, HVO, HFX and HAH, abbreviations listed in Table 2).

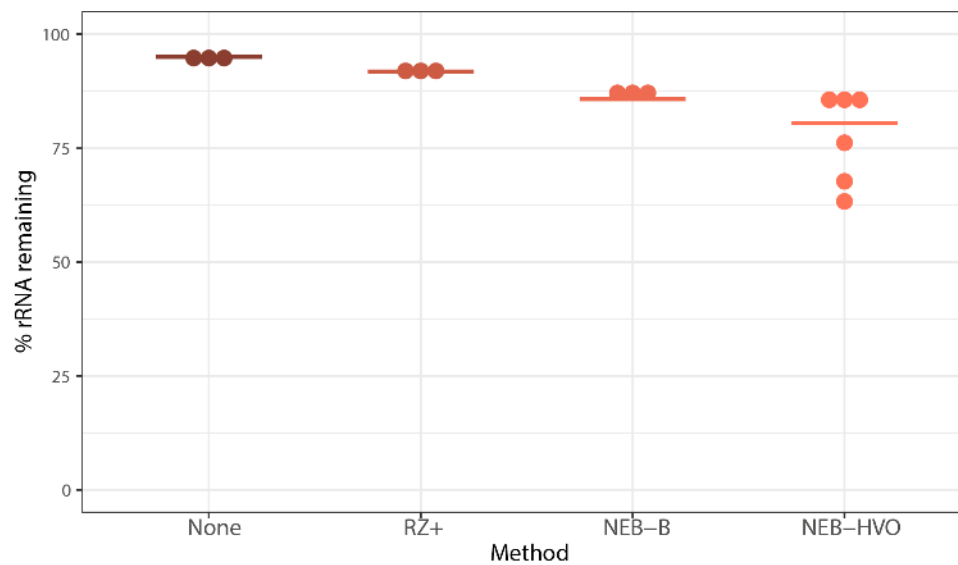
#### 4.3.2. Testing new rRNA depletion strategies on total RNA samples from *Halobacterium salinarum* (HBT).

We first began with a quantification of rRNA removal in HBT to allow continuation of ongoing differential gene expression experiments<sup>110</sup>. We used three enzymatic digestion-based rRNA depletion approaches from the following commercial kits (details in Appendix B and Methods): (a) NEBNext Bacterial rRNA Removal Kit (probes included, abbreviated throughout as “NEB-B”); (b) NEBNext rRNA Core Depletion Reagent Set (with user-designed probes specific for



**Figure 25: Percentage of rRNA remaining in halophile RNA by using the discontinued Ribozero kit (RZ). Each dot denotes one sample, with light orange dots representing *Hbt. salinarum* and blue dots representing *Hfx. volcanii* samples. Horizontal bars represent the median value.**

HVO, another model organism used in our lab, method abbreviated throughout as “NEB-HVO”); and (c) the newly released Ribo-Zero Plus kit from Illumina (includes probes allowing universal depletion across bacteria and eukaryotes, “RZ+”). Following rRNA removal, resultant RNA samples were subjected to Next Generation sequencing, and the number of rRNA reads removed was quantified as compared to an untreated RNA control (Methods).



**Figure 26: rRNA removal using alternative methods in *Hbt salinarum*. Each dot represents percentage of counts mapping to rRNA genes after using no removal (brown), New Ribozero kit (RZ+, dark orange), NEBNext kit with bacterial probes (NEB-B, orange), and NEBNext kit with HVO probes (NEB-HVO, pink). Horizontal bars represent the median value.**

We observed that ~95% of reads from sequenced untreated RNA correspond to rRNA (Fig. 26, Table 7). RZ+ treatment achieved a negligible reduction of rRNA to ~92%. A slightly more substantial reduction was seen with the NEB-B method, with a median remaining rRNA percentage of 86%. Of these methods, the best results were obtained using NEBNext with customized probes designed to bind HVO rRNA sequences (NEB-

HVO), although high levels of rRNA still remained (median remaining rRNA 80.5%, range 63% to 86%. We note that using no removal, RZ+, and NEB-HVO methods result in a range of ~1.5-3.6M reads mapping to non-rRNA genes per sample (with 12 total samples run on one lane, **Appendix B**). Based on our power analysis using online tools<sup>230</sup>, this level of sequencing depth would require 5-6 biological replicates for reliable detection of 75% of differentially expressed genes (FDR < 0.05, log fold change  $\geq$  2.0) (**Fig. 27**). Since this depth was achieved with 12 samples multiplexed per lane, a requirement of 4-5 samples of each type would restrict RNA-Seq experimental design to a single

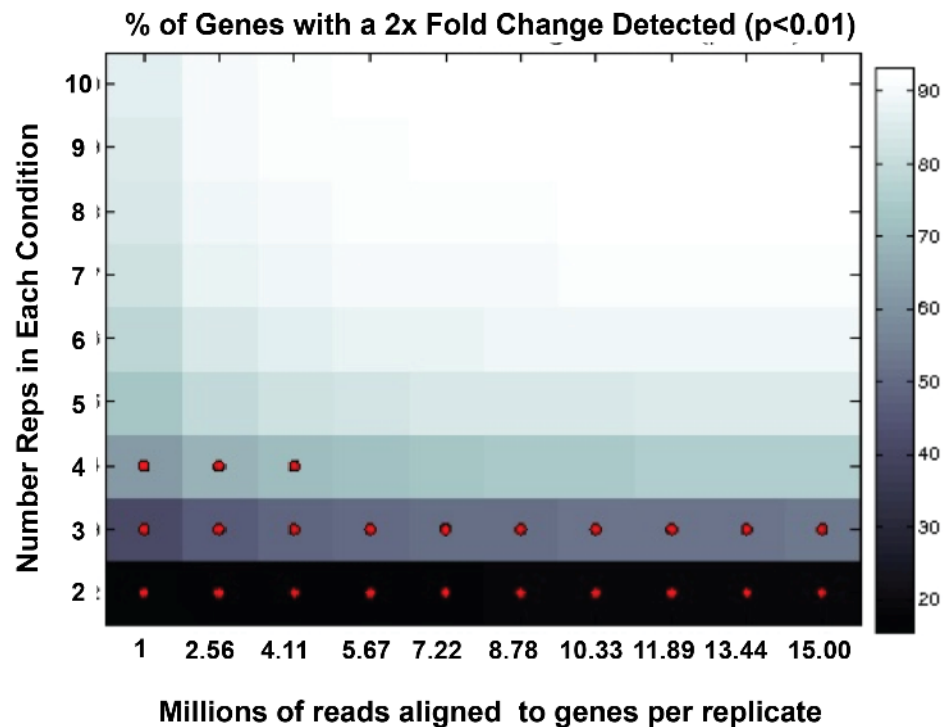
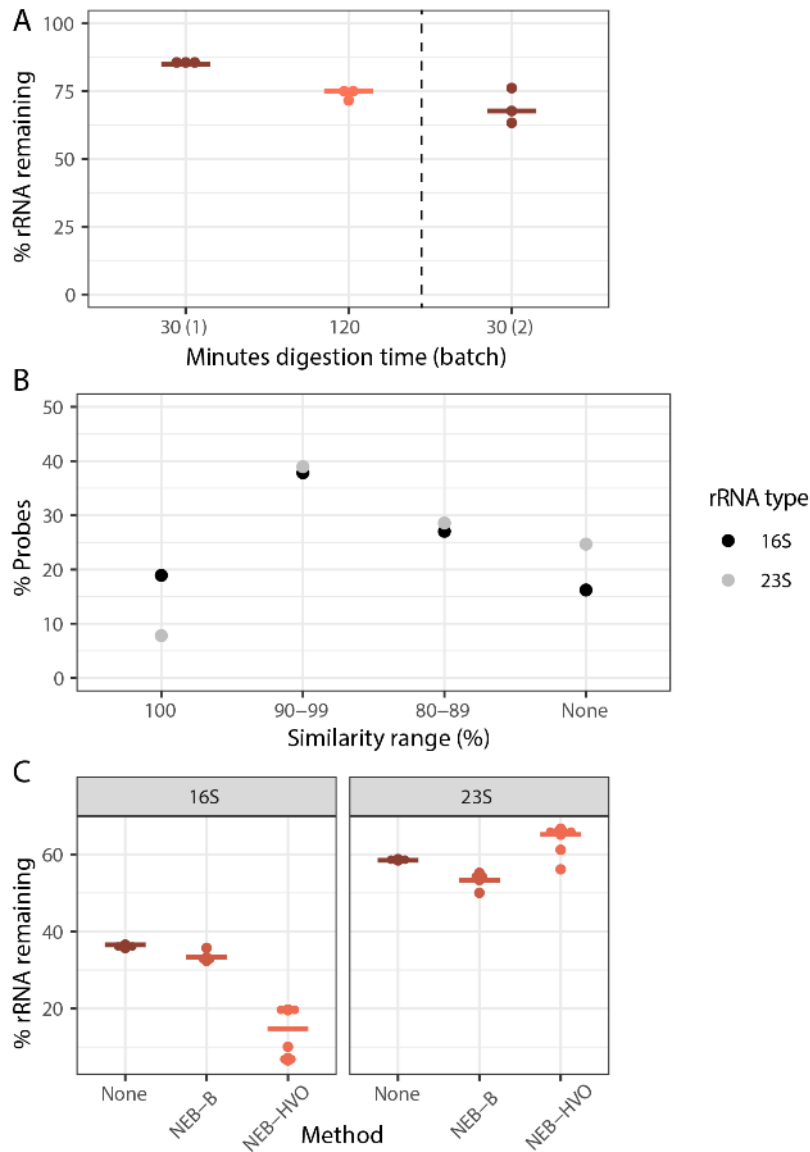


Figure 27: Analysis of sequencing depth, number of biological replicates, and detection of differentially expressed genes (2-fold differential expression) using the online tool Scotty. Squares with red dots are predicted to have <75% detection of diff expr genes.



**Figure 28: Increasing RNase digestion time is less important than probe sequence identity for efficient rRNA removal. (A) Dotplot showing percentage of counts mapping to rRNA genes after using NEB-HVO method on HBT total RNA samples after 30 minutes (brown) or minutes 120 (light orange) of RNaseH digestion. NEB30\_2 category to the right of the dotted line denotes samples processed and sequenced in a different batch, showing that batch effect dominates over RNase digestion time effect. Horizontal bars represent the median value. (B) Percentage of custom-designed HVO probes classified into 16S (black) and 23S (grey). Levels of sequence identity of HVO probes with *Hbt. salinarum* (HBT)16S and 23S rRNA genes are shown on the X-axis and percentage of total probes at each sequence identity level is shown on the Y-axis. (C) Percentage of total reads mapping to either 16S (left panel) or 23S rRNA (right panel) genes of HBT using 3 different rRNA removal methods - none (brown), NEB-B (dark orange), NEB-HVO (light orange).**

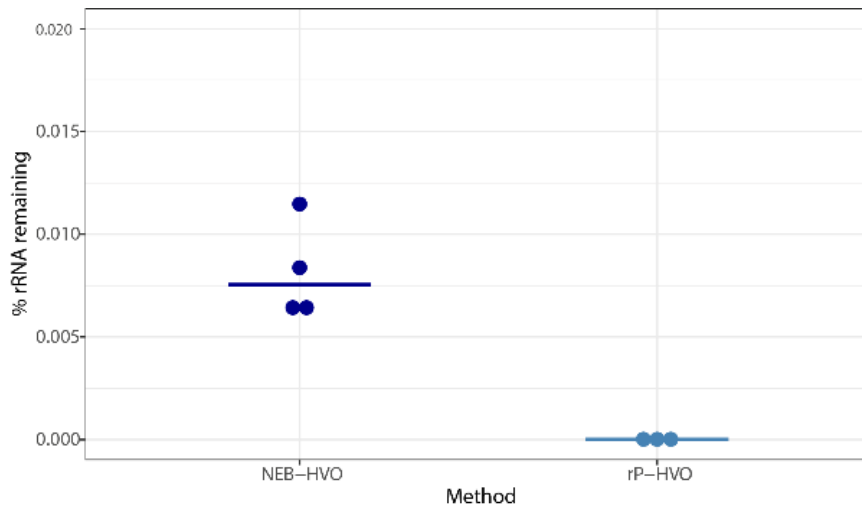
comparison (for example, two genotypes in one condition or two conditions for the same genotype) per lane. Hence, the inefficient rRNA removal severely limits the extent to which samples can be multiplexed, increasing costs even in modern high-throughput sequencing instruments used here (**Appendix B**). We hypothesized that poor rRNA removal may stem from either the incomplete RNase H digestion or the imperfect sequence match between the HVO rRNA probes used in the NEB-HVO method and the rRNA genes of HBT. To test the efficacy of RNase H digestion, we carried out this digestion over 30 minutes (manufacturer protocol) and 120 mins (extended digestion) using the NEB-HVO method. Each digestion time used the same extracted RNA sample (split into two different aliquots for digestion), and was performed in biological triplicate within the same sequencing batch. A marked improvement rRNA removal is seen in the 120-minute digestion (**Fig. 28A**), with 75% median rRNA remaining, as compared to 85% for the 30 minute samples. However, when comparing the results between different batches of sequencing, we found that the batch effect was stronger than the RNase effect: 30 minute RNase H digestion from a different batch produced a rRNA range of 63-76% (median 68%), better than even the 120 minute digestion from the first batch. Hence, while longer RNase H digestion could potentially improve rRNA removal, this effect is inconsistent.

Based on these results, we then tested the hypothesis that this relatively poor rRNA removal (compared to the discontinued RZ method) was associated with sequence mismatches between probes and rRNA. Using the NEB-HVO method, we observed that the probe sequences custom-designed for HVO rRNA matched HBT 16S rRNA

sequences better than to 23S probe sequences (**Fig. 28B**, Table S3). 19% of 16S HVO probe sequences had 100% identity with HBT 16S rRNA, compared to only 8% for 23S rRNA. Conversely, 25% of 23S probe sequences shared no sequence similarity with HBT 23S rRNA, while this was only 16% for 16S. Corresponding with these different levels of sequence identity, we observed that 16S rRNA removal was more effective than 23S rRNA removal (**Fig. 28C**, **Appendix B**). Hence, there is a strong relation between probe sequence and rRNA removal, with even slight increases in probe specificity (**Fig. 28B**) resulting in profound differences in rRNA removal (**Fig. 28C**).

We conclude from these experiments that the NEBNext Core Reagent Set kit with probes custom-designed for the related species HVO (NEB-HVO) is the best of reagent kits that we tested for HBT rRNA removal. RiboZero Plus (RZ+) and NEBNext Bacterial kit using the bacterial probes (NEB-B) led to less efficient rRNA removal for HBT. Targeting of custom probes specifically for HBT would likely result in better rRNA removal.

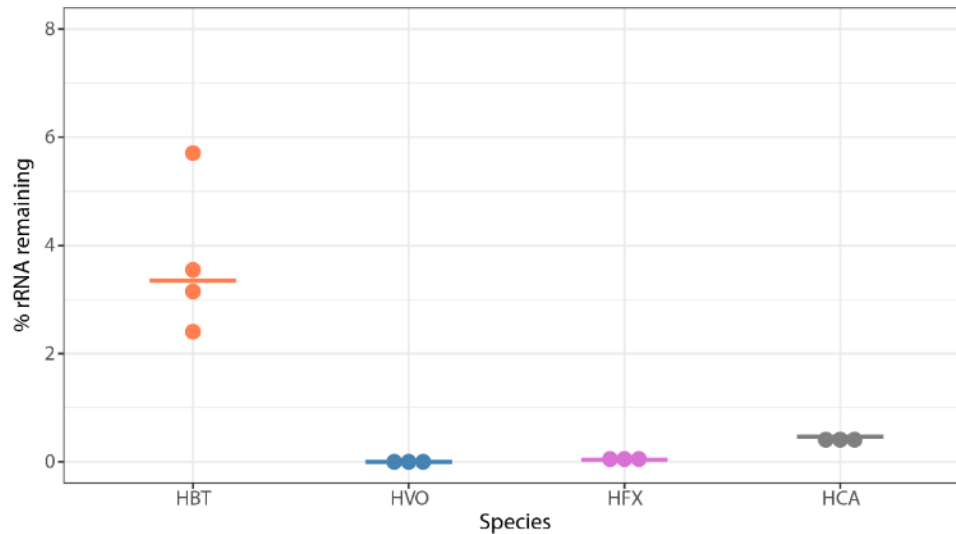
### 4.3.3. Species-specific probe methods efficiently remove *Haloferax volcanii* (HVO) rRNA:



**Figure 29: Species-specific probes efficiently remove rRNA from target species. Dotplots showing percentage of rRNA remaining after using probes with sequences specific for *Hfx. volcanii* (HVO) rRNA. Dark blue dots represent %rRNA remaining in individual replicate samples depleted with NEBNext Core Reagent Set (“NEB-HVO” method). Light blue dots represent %rRNA remaining in individual replicate samples depleted with the siTools RiboPool kit (“rP-HVO”). Horizontal bars represent the median value.**

Having shown the importance of probe sequence specificity, we next tested two different methods with rRNA probes targeted to HVO against HVO total RNA samples: (a) NEBNext Core Reagent Set (“NEB-HVO” method); and (b) the siTools RiboPool kit (“rP-HVO”). Unlike the enzymatic NEB-HVO method, rP-HVO uses streptavidin-based removal of rRNA hybridized to biotinylated probes. For both methods, we used probes custom-designed to be specific to HVO rRNA sequences (see Methods). We observed that both methods achieved nearly complete rRNA digestion: median values of 0.008% and 0.000008% rRNA remaining were observed using NEB-HVO and rP-HVO methods, respectively (Fig. 29; Table 7). These results with near-complete rRNA depletion in

HVO with species-specific probes is line with the observations above that the limiting factor with these probe-based methods is the identity of probe sequences with target rRNA sequences. Overall, we found that using probes targeted to HVO with either method resulted in efficient and near-complete removal of rRNA from HVO samples.



**Figure 30: Panarchaea kit offers efficient rRNA removal across halophilic species.** Dotplots showing percentage of remaining counts mapping to rRNA genes in *Hbt. salinarum* (HBT, orange), *Hfx. volcanii* (HVO, blue), *Hfx. mediterranei* (HFX, purple), *Hca. hispanica* (H CA, grey). Horizontal bars represent the median value of three biological replicate samples.

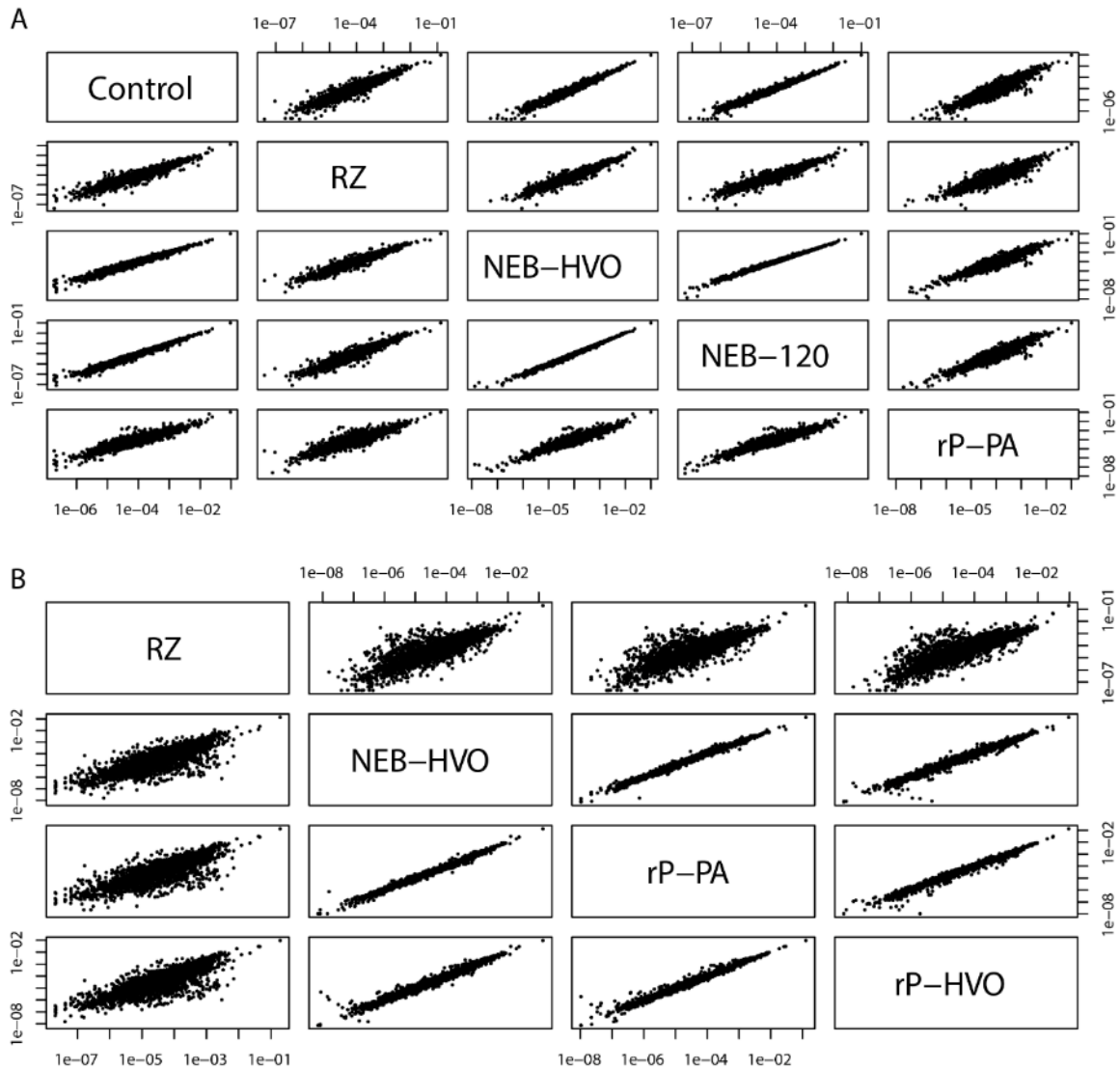
#### 4.3.4. siTools Panarchaea kit efficiently removes rRNA from diverse halophilic archaeal species:

To expand our analysis to other model species, we then tested the siTools riboPOOL Panarchaea kit (rP-PA, **Appendix B**, methods). The probe set associated with this kit is composed of high complexity pools of biotinylated DNA probes with sequences designed to deplete rRNA from a broad spectrum of archaea, including several classes of Euryarchaeota and Proteoarchaeota (<https://sitoolsbiotech.com/ribopools.php>). The Panarchaea riboPOOL probes have been shown to remove 99% of rRNA from *Sulfolobus*

*solfatarius* and *Sulfolobus acidocaldarius* (<https://sitoolsbiotech.com/pdf/microbes-ribopools-072021.pdf>), but to our knowledge have not been published for euryarchaeal species like the four model halophiles of interest here. After using this kit for ribodepletion, we observed that all tested RNA samples across the four species contained <10% rRNA, with median values of 3.3%, 0.0002%, 0.04%, and 0.5% for HBT, HVO, HFX, and HAH, respectively (**Fig. 30, Table 7**). This extensive rRNA removal is more effective for HBT than for any previously tested methods (**Fig. 25, 26**), and equally as effective as NEB-HVO and rP-HVO methods for HVO (**Fig. 29**). The other two species had not been previously tested, and no other RNA-seq results (other than for HFX small RNAs, which does not require rRNA removal<sup>231</sup>) are available for comparison in the literature. Taken together, these results demonstrate that the Panarchaea method (rP-PA) efficiently removed rRNA for four different model species of halophilic archaea.

#### **4.3.5. Choice of removal method does not affect per-gene read counts:**

It was observed previously that using different rRNA removal methods can affect relative read counts of some non-rRNA genes<sup>232,233</sup>. We therefore tested whether rRNA removal and the choice of removal method changes the relative levels of mRNA. We calculated gene counts from each sample as a percentage of the total (non-rRNA) counts from that sample, and correlated these relative counts obtained from different rRNA removal techniques (see Methods, **Fig. 31**). We observed strong correlations of normalized relative counts of non-rRNA genes among different rRNA removal methods, as well as with untreated total RNA which was available for HBT (**Fig. 31A**). The



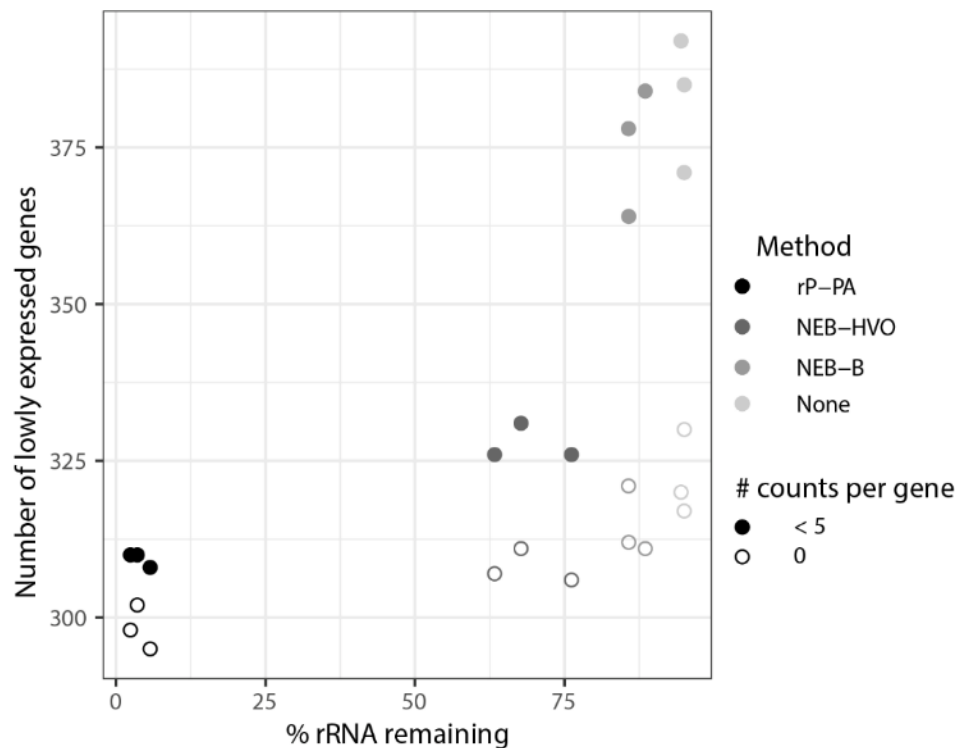
**Figure 31: Choice of removal method does not affect relative abundance of non-rRNA genes.** Correlations between relative abundance of each gene after different rRNA removal methods in (A) *Hbt. salinarum* (HBT) and (B) *Hfx. volcanii* (HVO). Each dot represents the percent of total normalized reads for each gene (see Methods section). Methods shown here are “Control” (no removal), “RZ” (using discontinued RiboZero kit), “NEB-HVO” (using NEBNext kit with custom HVO probes), “NEB-120” (NEBNext kit with custom HVO probes and 120 mins of RNase digestion), “rP-PA” (siTools riboPOOL method using Panarchaeal probes), and “rP-HVO” (siTools riboPOOL method using HVO-specific probes).

Pearson's correlation coefficients between per-gene normalized read counts across different methods used on HBT were in the range 0.91-0.99, with an average value of 0.95. Correlation with control (untreated) samples was >0.92. Similar results were seen with HVO: 0.94-0.99, average 0.97 (Fig. 31B, Appendix B). Based on this analysis, we conclude that rRNA removal and the choice of removal method does not change the number of reads on a per-gene basis in halophilic archaea. These rRNA removal methods can therefore be used for downstream applications such as differential gene expression analysis.

#### **4.3.6. Utility of rRNA removal is seen in counts of non-rRNA genes:**

Previous studies have suggested a minimum sequencing depth of two<sup>222</sup> to ten<sup>223</sup> million reads per sample for obtaining reproducible results for differential expression. On a per-gene basis, 5 reads is considered a threshold below which differential expression analysis is unreliable<sup>221</sup>. We sought to understand how rRNA removal affects transcript detection using data from HBT, from which we have data for a wide array of rRNA removal methods (including no removal), and a large range of rRNA remaining in sequenced samples (2%-95%). Across these samples, we calculated the number of annotated genes with no mapped reads as well as <5 mapped reads. For consistency, only samples that had been sequenced on the same machine (NovaSeq6000) were considered, so that the comparison isn't biased by massive changes in total (rRNA+non-rRNA) reads.. We observed that more complete rRNA removal generally leads to increased detection of genes (**Fig. 32**). All numbers that follow are median values, obtained from **Appendix B**. For untreated RNA (~95% rRNA), ~320 genes showed no

reads, and this reduced to ~312 for RNA treated with NEB- (~86% rRNA) and further to ~307 with NEB+ (~68% rRNA). The Panarchaea kit (~3% rRNA) reduced the number of undetected genes to ~298. When using <5 count genes, there was a more dramatic change, from ~385 genes for untreated RNA, but only ~310 for Panarchaea kit. This held even though the total number of reads for all genes (including rRNA) had relatively similar median values of ~30M and ~27M reads, respectively (Table S1), showing that the improved detection of lowly expressed genes is associated with



**Figure 32: More complete rRNA removal leads to increased detection of lowly expressed genes. Number of genes with zero (open circle) and <5 (filled circle) reads mapping to them in different rRNA removal methods in *Hbt. salinarum*. The darker the circle color, the more complete the rRNA removal for each method: riboPOOL Panarchaea (rP-PA, black); NEBNext with HVO-specific probes (NEB-HVO, dark grey); NEBNext with bacterial probes (NEB-B, grey); no removal (none, light grey).**

more complete rRNA removal rather than by deeper overall sequencing of the samples. These above results indicate that better rRNA removal improves per-gene counts and detection of lowly expressed genes, which is important when detecting differential expression or annotating non-coding transcripts, and thus can help increase multiplexing and therefore reduce cost when sequencing.

#### **4.4. Conclusions and discussion**

The main technical challenge for prokaryotic transcriptomics is the low ratio of mRNA:rRNA. Historically, different methods have been used to eliminate rRNA without biasing mRNA reads: from digestion with exonucleases that preferentially degrade rRNA relying on 5' monophosphate; to subtractive hybridization that captures rRNA binding to antisense oligonucleotides<sup>234,235</sup>; to poly(A) tail addition to discriminate rRNA or reverse transcription with rRNA primers followed by RNaseH digestion<sup>228,236</sup>. However, none of these methods has been successfully utilized for haloarchaea. Until the end of 2019, the Ribo-Zero kit from Illumina, based on sequence-specific biotinylated probes that hybridize with a pool of microbe rRNA sequences and then pull the hybrids out using streptavidin-coated magnetic beads, enabled removal of ~70% of rRNA for several archaeal species<sup>215-219</sup>. After this commercially available kit was discontinued, archaeal transcriptomics had undergone to a period of difficulty. Here we invested time to troubleshoot this problem, test, and directly compare newly available tools to help the archaeal community to move on with transcriptomic studies. Our investigation provides a guide for choosing a suitable application depending on the model organism or the

combination of archaeal species of interest (e.g. communities, labs using multiple cultured species, metatranscriptomics).

We found that both RNaseH-based and biotin-based methods are efficient for rRNA removal. Certain commercially available kits from NEB and siTOOLS are most effective when probes are designed that target archaeal species of interest. For HVO, the RiboPool kit as well as the NEB-Next kit with custom-designed probes that target HVO (**Fig. 29**) resulted in almost complete rRNA removal. A similar number of total reads was observed after sequencing with no detectable bias in lowly expressed transcripts (**Fig 31**). In general, when using targeted kits, we found that the most important factor in determining rRNA removal efficiency was percentage identity of the target rRNA with the probe sequence (**Fig. 28, 29**). We found that the Panarchaea kit from siTOOLS provides very good rRNA depletion across all four species tested here (**Fig. 30**) and we anticipate that these can be effectively used for metatranscriptomics of archaeal communities. Targeted methods from both NEB and SiTOOLS as well as the Panarchaea probe set provide comparable performance to the discontinued RiboZero kit for HVO, with rRNA percentages close to 0. We further note that the Panarchaea kit exceeds the performance of the RZ method for HBT (**Table 7, Fig. 25 vs Fig. 30**). In the future, continuing to deposit raw RNAseq data from the archaeal community into online data repositories such as NCBI Gene Expression Omnibus is critical for progress in the area of transcriptomics, which would facilitate future efforts to predict rRNA removal success depending on probe sequence identity.

One of the most important advantages of choosing an efficient rRNA removal method is when analyzing differential expression of low-count genes. In the current study, we show that in undepleted RNA, around 385 genes had <5 counts but efficiently depleted RNA resulted in a decreased number of low-count genes to ~310 (**Fig. 32**). This improvement in coverage of low-count genes enables correct statistical analysis of differential expression<sup>221-223</sup>. Accurate detection of lowly expressed transcripts is also important when using RNA-Seq to map the transcriptome<sup>214,237</sup> including in some metatranscriptomic protocols<sup>236</sup>.

The rapid pace of discovery of new archaeal species<sup>238,239</sup> as well as the use of novel archaeal model organisms in lab will bring further challenges for transcriptomics experiments. However, the methods tested here provide sufficient flexibility to solve such challenges. For example, it is possible that new challenging archaeal species may have rRNA sequences divergent from commercially available primer sets such as siTOOLS Panarchaea. The NEBNext Core Reagent Set using the custom probe design tool (<https://depletion-design.neb.com/>) would therefore be an appropriate choice in this case. Removal of rRNA enables increased detection of rare transcripts and extensive multiplexing. The methods tested here will therefore facilitate rapid progress in understanding the transcriptional response of a wide diversity of archaea to their environment.

## Conclusions and Future Directions

### 5.1 Conclusions

I have started this study with the hypothesis that the hypersaline environment of halophilic archaea has selected for an alternative histone function, viz. that of a transcription factor (TF). This hypothesis has been first tested with traditional microbiology methods including growth assays and microscopy. These results were used as the basis for high-throughput genome-wide methods like RNA-Seq quantification of differential expression, and the first published instance of ChIP-Seq to study the genome-wide binding of archaeal histone-like proteins. The results in Chapter 2 and Chapter 3 paint a complex picture, providing partial, but not complete, support for this hypothesis.

There is a growth defect and morphology change for the  $\Delta hpyA$  strain of *Halobacterium salinarum* in reduced salt, which corresponded with its increased genome-wide binding and differential expression of target genes within the reduced salt condition. These results link *hpyA* function and external salt concentration, supporting the hypothesis that halophilic histone function has been selected for by hypersaline conditions. On the other hand, varying external salt concentration did not affect *hstA* phenotype, as its growth defect relative to the parent strain was not worsened in either optimal or reduced salt conditions. This could be interpreted in two ways. One explanation is that *hstA* function is distinct from *hpyA* function and is unrelated to salt, implying lineage-specific histone function within halophilic archaea. Alternatively, we speculate that the

noted growth defect of *hstA* in *Hfx. volcanii* optimal conditions (2.5M external Na<sup>+</sup>), which is not seen in optimal conditions of *Hbt. salinarum* (4.2M external Na<sup>+</sup>) could itself be a result of histone function being linked to salt concentration. This could be because the internal K<sup>+</sup> concentrations for *Hfx. volcanii* are lower and vary over a smaller range<sup>240</sup> as compared to those of *Hbt. salinarum*<sup>26</sup>. Hence, overall, the link between halophilic histone function and external salt concentration has been validated in *Hbt. salinarum*, and may or may not hold in *Hfx. volcanii*.

These results constitute an interesting addition to the growing literature on archaeal histones. A link between chromatin protein and species environment has also been observed by a recent study of expression of chromatin proteins across the Archaea<sup>118</sup>. In particular, the expression of chromatin proteins including histones was generally found to be correlated with optimum growth temperature. This hypothesis connects previous observations from several archaeal species. *Methanothermus fervidus* and *Thermococcus kodakarensis*, which contain the best-characterised archaeal histones<sup>65,66,68,97</sup>, are thermophilic archaea. *Methanopyrus kandleri*<sup>99,101</sup>, *Methanocaldococcus jannaschii*<sup>193</sup>, and *Methanothermobacter thermoautotrophicus*<sup>98</sup>, whose histones have been studied *in-vitro*, are also thermophiles. All these histones were found to be capable of genome compaction. By contrast, in *Methanosarcina mazei*, a mesophile, the histone gene was found to be non-essential<sup>103</sup>. *Methanosphaera stadtmanae* is the only mesophilic archaeon whose histone was demonstrated<sup>102</sup> to be capable of genome compaction. In this context, this thesis, building on previous work<sup>104</sup> from Keely Dulmage also from the Schmid lab, characterises halophilic histones as largely sequentially similar but functionally

divergent from the well-studied archaeal histones of thermophiles, with a role in gene regulation in hypersaline conditions, rather than in genome compaction. All of this suggests that histones performing genome compaction are a necessary adaptation to thermophilic conditions, but their presence across the Archaea including in non-thermophilic species could be a result of a histone “addiction”. This speculative hypothesis is similar to a hypothesis about DNA gyrase in archaea<sup>241</sup>: removal of the conserved histone gene does not negatively impact the cell’s fitness, instead, it has developed niche-specific functions.

In terms of the role of halophilic histones as transcription factors, we find support for this hypothesis from several results that indicate TF-like properties– their low expression<sup>104</sup>, their binding in the form of discrete, narrow, and relatively sparse peaks, their non-essentiality, (in the case of HpyA) condition-dependent regulation of various pathways. However, this does not paint a complete picture. Neither HpyA nor HstA showed a preference for binding in promoter regions, which was observed for both bacterial and haloarchaeal TFs. A majority of the differentially-expressed genes for *ΔhpyA* were found to be far from any HpyA binding loci, and this indirect regulation conflicts with the classical model of TFs regulating transcription by binding at the promoter of the target genes, and instead resembles results seen for some nucleoid-associated proteins (NAPs) including IHF<sup>172</sup>, FIS<sup>171</sup>, and Lrp<sup>174</sup>. Additionally, no cis-regulatory sequence motif could be determined for HpyA, which was found to preferentially bind to sequences containing ~10bp dinucleotide periodicity, resembling the preferences seen for archaeal<sup>95,184</sup> and eukaryotic<sup>178,180</sup> histones. HstA, on the other

hand, did show a preference for a palindromic sequence and not for dinucleotide periodicity, suggesting some divergence within the halophilic histones. Put together, these results did not place HpyA and HstA in the classification of “transcription factor”, instead, they suggested that halophilic histones lie within a blurred line dividing TFs and chromatin proteins. This agrees with recent work<sup>161</sup> observing that the TF-NAP distinction in Bacteria is complex, with many TFs capable of DNA architectural roles<sup>242</sup> and of binding without a strict sequence motif<sup>243</sup>, while many NAPs are capable of TF-like functions (promoter binding<sup>171</sup> and RNA polymerase recruitment<sup>244</sup>, sequence specificity<sup>172</sup>, etc). A strong example of the inadequacy of the binary classification is provided by proteins of the Lrp family which are present in both Bacteria and Archaea. They share many features of TFs (relatively low protein expression compared to other NAPs<sup>60</sup>, ligand-based functional changes<sup>245</sup>, cis-regulatory sequence motif<sup>246</sup>, direct control of target gene regulation<sup>174</sup>) and also with NAPs (non-sequence specific binding<sup>173</sup>, indirect control of gene regulation<sup>174</sup>). After this study, we can place halophilic histones as another family of proteins that defy binary classification.

## **5.2 Future directions**

A number of paths forward are presented by this work, including experiments proposed at the start of the thesis work which could not be carried out, as well as logical follow-up work from the results obtained.

*In-vitro* analysis of the binding of halophilic histones and DNA could answer several questions and corroborate some of the results here. Is this histone, with its acidic surface,

still capable of compacting DNA (even if it is expressed at too low a level to be able to perform this function)? Can the dinucleotide periodicity preference for HpyA and the cis-regulatory motif for HstA be validated *in-vitro* too? Such an experiment would obviously require high expression of HpyA or HstA from a suitable host, and with a cleavable tag to enable purification. Unfortunately, while I was able to create *E. coli* and *Hfx. volcanii* strains containing HstA fused to a (His)<sub>6</sub> tag and under the control of an inducible promoter, no expression of the protein was observed in either a Coomassie stained gel or a western blot, hence that aspect of the thesis had to be abandoned.

The results from Chapter 3 (see also **Appendix D**) regarding a potential temperature phenotype for HstA are a promising lead for future work. This phenotype, observed best in a plate-reader at 56°C, should be checked more thoroughly at higher temperatures, as well as in flask growth. A temperature-dependence of histone function would align well with recent literature<sup>118</sup>.

This work did not, and was not intended to, address the question of which protein does perform the role of DNA compaction in halophilic archaea (assuming it is performed in the first place). Mc1, an archaeal chromatin protein, could be cleanly deleted from *Hbt. salinarum*, but the  $\Delta mc1$  strain of *Hfx. volcanii* was found to have a secondary site mutation. This might suggest a possible essentiality of this gene (a trait common to many chromatin proteins), however, this can only be corroborated if further deletions are made and sequenced. The  $\Delta mc1$  strain that was created did have a strong growth

defect in high temperature, and this could also be explored further (details regarding the mutation and growth defect in **Appendix D**).

Finally, in terms of the broader field of archaeal histones, an interesting avenue for future research is in the Asgard archaea, the closest relative of the Eukarya.

Metagenomic sequences of their histone genes revealed possible N-terminal tails, in some cases similar to that of eukaryotes, and in some cases much shorter<sup>63</sup>. The function of this tail, its potential for post-translational modifications, and the potential of Asgard histones to form hypernucleosomes are questions that have implications in the divergence of Eukaryotes from Archaea.

## Appendix A

**Table A1: List of primers used in this study and their purpose; primers created by Keely Dulmage (KD), Cynthia Darnell (CD), or Saaz Sakrikar (SS)**

Primer Name	Direction	Purpose	Sequence (5'-3')	Source	Created by
K2	F	HpyA KO check, sequencing; gDNA detection in RNA	GGCGTTGTTACGAAGCA	This study	KD
K3	R	HpyA KO check, sequencing; gDNA detection in RNA	CGACATTCGGGTTGCGTTG TG	This study	KD
1044_GFP_CHA_Kpn1_fwd91	F	pMTFcHA Plasmid screening, sequencing	CGCGGAAACGATGAAATG CG	Dulmage et al 2015 <sup>104</sup>	KD
seq_reporter	R	pMTFcHA Plasmid screening, sequencing	AGAACGAGTAGCACACCA AAG	Dulmage et al 2015	KD
HArev	R	pMTFcHA Plasmid screening, sequencing	GCGTAGTCCGGGACGTCGT AC	This study	CD
K71	F	HpyA insert for overexpression plasmid	ACGTACGT <i>CAT</i> <i>ATGAGCGTCGAACTCCCGT</i> TC	Dulmage et al 2015	KD
HpyA OE R2	R	HpyA insert for overexpression plasmid	ATCGTGAC <i>AAGCTT</i> TTCCACGAGCGTGACGTAC GTCTG	This study	KD
pRPA200_iso	F	200bp <i>rpa</i> promoter fragment	CCTCCCATGCCACTCTTCA CACGCGGTAC TGGTCCGCAAGCCA	Dulmage 2015 <sup>121</sup>	KD
pRPA_hpyA_iso	R	200bp <i>rpa</i> promoter fragment	GAGTTCGACGCTCAT TGTCGGTTCAGGCCA	Dulmage 2015	KD
HpyA_ATG_F	F	Fusing <i>hpyA</i> insert with <i>rpa</i> promoter	TGGCCTGAACCGACA <i>ATGAGCGTCGAACTC</i>	Dulmage 2015	KD
HpyA_tail_pMTF_iso	R	Fusing <i>hpyA</i> insert with <i>rpa</i> promoter	CGTACGGGTACGCCGAAA GCTTGGATCCG TCATTCCACGAGCGTG	Dulmage 2015	KD

hsta-KO-up1-new	F	<i>hstA</i> deletion (upstream amplification)	CTCGAGGTCGACGGTATCG ATAAGCTTGATGAAGGCCT CGGCCGCTTC	This study	SS
hstA-KO-up2	R	<i>hstA</i> deletion (upstream amplification)	GGA GAA GTA GGT CTC GAT GTC CTC ACT CAT ACC CAC ATA TCG GGT CCG	This study	SS
hstA-KO-d2	F	<i>hstA</i> deletion (downstream amplification)	CGG ACC CGA TAT GTG GGT ATG AGT GAG GAC ATC GAG ACC TAC TTC TCC	This study	SS
hstA-KO-d1	R	<i>hstA</i> deletion (downstream amplification)	AGTGGATCCCCCGGGCTGC AGGAATTCGATCTCGTGGA TGGACGTGTAGAACAC	This study	SS
hstA-native-rescue-up	F	Amplification of <i>hstA</i> and upstream promoter for cloning	CGGTCCGACAACAACCCC CGATCCAAGCTTGATAGA CCGCCGGTCCG	This study	SS
hstA-rescue-down	R	Amplification of <i>hstA</i> for cloning	CTTCACTTCTCGAACTGCG GGTGCGACCAGTTTCGCTG TAGCCGAATTGCATTATTC	This study	SS
hstArescue exchange_F	F	Add HA tag to <i>hstA</i> on pAKS147 plasmid	GCCGGATTATGCGTAATGC AATTCGGCTACAG	This study	SS
hstArescue exchange_R	R	Add HA tag to <i>hstA</i> on pAKS147 plasmid	ACATCATACGGATATTCGA AGAGGGAGAAGTAG	This study	SS
HvhstAseq up	F	<i>hstA</i> KO check, sequencing	GGA CGG ATA GAC CGC CGG	This study	SS
HvhstAseq down	R	<i>hstA</i> KO check, sequencing	CTCGTGGATGGACGTGTAG AACAC	This study	SS
pJAM809f_SS	F	pJAM809 plasmid screening, sequencing	GTCGGACAACAACCCCCG	This study	SS
C186	R	pJAM809 plasmid screening, sequencing	CGGTACGCTGCGCGTAA C	This study	CD
C33	F	pTA131 plasmid screening, sequencing	GCGCGTAATACGACTCACT A	This study	CD
pTA131r_SS	R	pTA131 plasmid screening, sequencing	CAAGCGCGCAATTAACCC	This study	SS

**Table A2: List of strains created/used in this study; strains created by Keely Dulmage (KD) or Saaz Sakrikar (SS)**

Strain	Schmid lab ID	Species	Genotype	Description	Reference	Created by	Notes
MDK 407	HS149	<i>Hbt sal</i>	$\Delta$ <i>ura3</i>	Parent strain	Peck et al 2000	-	Additional mutations; see B1
KAD 100	HS90	<i>Hbt sal</i>	$\Delta$ <i>ura3</i> $\Delta$ <i>hpyA</i>	<i>hpyA</i> knockout	Dulmage et al 2015	KD	
KAD 128	HS117	<i>Hbt sal</i>	$\Delta$ <i>ura3</i> $\Delta$ <i>hpyA</i> /pK AD17	HpyA-HA under its native promoter on a plasmid, $\Delta$ <i>hpyA</i> background; Ura- Mev+	This study	KD	
KAD 103	HS191	<i>Hbt sal</i>	$\Delta$ <i>ura3</i> $\Delta$ <i>hpyA</i> /pK AD03	HpyA under its native promoter on a plasmid, $\Delta$ <i>hpyA</i> background; Ura- Mev+	Dulmage et al 2015	KD	
KAD 101	HS189	<i>Hbt sal</i>	$\Delta$ <i>ura3</i> $\Delta$ <i>hpyA</i> /pM TFcHA	Empty vector, $\Delta$ <i>hpyA</i> background; Ura- Mev+	Dulmage et al 2015	KD	
H26	Hv28	<i>Hfx vol</i>	$\Delta$ <i>pyrE</i>	Parent strain	Hartman et al 2010	-	Additional mutations; see B1
AKS1 98	Hv253	<i>Hfx vol</i>	$\Delta$ <i>pyrE</i> $\Delta$ <i>hstA</i>	Deletion of HstA (HVO_0520)	This study	SS	
AKS2 14	Hv269	<i>Hfx vol</i>	$\Delta$ <i>pyrE</i> $\Delta$ <i>hstA</i> /p AKS147	HstA under its native promoter; $\Delta$ <i>hstA</i>	This study	SS	

				background; Ura-Nov+			
AKS2 17	Hv272	<i>Hfx vol</i>	$\Delta pyrE$ $\Delta hstA/pJ$ AM809	pJAM809 vector with - HA tag, $\Delta hstA$ background; Ura-Nov+	This study	SS	
AKS2 33	Hv288	<i>Hfx vol</i>	$\Delta pyrE$ $\Delta hstA/p$ AKS180	HstA-HA under its native promoter; $\Delta hstA$ background; Ura-Nov+	This study	SS	

**Table A3: List of plasmids created/used in this study; plasmids created by Keely Dulmage (KD) or Saaz Sakrikar (SS)**

Plasmid	Description	Purpose	Reference	Created by
pMTFchA	$P_{idx} Mev^r$	Expression plasmid for <i>Hbt. sal</i>	Wilbanks et al 2012 <sup>49</sup>	-
pKAD03	pMTFchA:: $P_{rpa}$ 200::hpyA	Complementation of HpyA with native HpyA promoter	Dulmage et al 2015 <sup>104</sup>	KD
pKAD17	pMTFchA:: $P_{rpa}$ 200::hpyA-HA	HpyA-HA expressed from plasmid driven by native promoter (200 bp upstream of ATG)	This study	KD
pTA131	pBluescript II with BamHI-XbaI fragment from pGB70 containing <i>pyrE</i> 2 under ferredoxin promoter	Shuttle vector for <i>Hfx. vol</i> gene deletion	Allers et al 2004 <sup>46</sup>	-
pJAM809		Expression plasmid for <i>Hfx. vol</i>	Humbard et al 2009 <sup>247</sup>	-
pAKS145	pTA131:HVO_0520_flanking	Deletion of HVO_0520 ( <i>hstA</i> ) from <i>Hfx vol</i>	This study	SS

pAKS147	pJAM809:P <sub>hstA</sub> :: hstA	Expression of hstA from native promoter	This study	SS
pAKS180	pJAM809:P <sub>hstA</sub> :: hstA-HA	Expression of HA-tagged hstA from native promoter	This study	SS

## Appendix B

Tables that have been deposited in Duke Data Repository- doi:10.7924/r4jd5093j

**Table B1:** List of mutations in parent and knockout strains, detected by Breseq

**Table B2:** Raw growth data for parental and  $\Delta hpyA$  cells in optimal and reduced salt (9 biological replicates), measured as optical density (OD600).

**Table B3:** List of peaks obtained by HpyA ChIP-seq; arranged by peak (**B3\_simplified**) and by overlap of peak and genomic feature (**B3\_full**)

**Table B4:** List of genes differentially expressed in  $\Delta hpyA$  in reduced and optimal salt conditions.

**Table B5:** List of genes in each subcluster obtained by clustering of expression patterns of differentially expressed genes.

**Table B6:** arCOG enrichments for genes nearest the HpyA ChIP-seq peaks, and for genes differentially expressed in  $\Delta hpyA$

**Table B7:** List of ChIP-seq peaks within 500 bp of differentially expressed genes.

**Table B8:** Growth data (OD600) for  $\Delta hstA$  and  $\Delta pyrE$  parent strain in optimal conditions.

**Table B9:** Manually curated list of ChIP-seq peaks for HstA.

**Table B10:** Details of characteristics (number of peaks, average width, total area covered by peaks) of ChIP-seq peaks for HpyA, HstA, and other DNA-binding proteins (shown graphically in main text figures).

**Table B11:** List of ChIP-seq datasets used for start site occupancy and peak width/coverage analysis, (with SRA trace where available), and genomes analyzed for periodicity (with link to the relevant NCBI assembly).

**Table B12:** Results of all rRNA removal experiments, including species, removal method, %rRNA (5S, 16S, 23S) remaining, total and non-rRNA aligned reads. Tabs within the table use the same data and are made for convenience of figure generation.

**Table B13:** Data downloaded from NCBI regarding number of publications per year mentioning “Archaea”, “RNA-Seq”, and “Archaea RNA-Seq”.

**Table B14:** Effect of probe specificity on rRNA removal: %age alignment of each custom-designed probe to *Hbt sal* 16S, 23S, and 5S rRNA sequences, and results of removal using these custom probes and non-targeted probes.

**Table B15:** Correlations between gene counts for different rRNA removal methods.

## Appendix C

### C1. Hypergeometric test results for peak location enrichments for haloarchaeal histones and transcription factors

Name	Total peaks	Peaks in non-coding regions	Coding genome length	Non-coding genome length	Hypergeometric test p-value
Hbtsal_HpyA	59	10	2229110	341900	0.15
Hfxvol_HstA	32	5	3376577	636323	0.4
Hfxvol_RosR	91	49	3376577	636323	<b>1.4e-17</b>
Hfxvol_TroR	36	24	3376577	636323	<b>9.9e-13</b>
Hfxmed_RosR	38	23	3293494	611213	<b>4.7e-11</b>
Hcahis_TrmB	221	119	3340391	549614	<b>2e-44</b>
Hcahis_RosR	59	30	3340391	549614	<b>4.1e-12</b>
Ecoli_FNR	188	132	3981787	659865	<b>6.9e-69</b>

### C2. Methods and results of search for motifs in bound regions (ChIP-Seq peaks) of HpyA and HstA

Software	Parameters used	Notes	HpyA results	HstA results
MEME	revcomp	Reverse complement of the sequence also considered for motif finding	Trinucleotide repeat [note 3]	-
	markov_order 0, 1 and 2	Correcting for nucleotide, dinucleotide, and trinucleotide frequencies in peaks (using the full genome as the background)	-	Palindrome-like sequence [note 4]
	mod anr	More than one motif instance per peak allowed	-	-
	pal	Look only for palindromic motifs	-	-

Memechip		Combining different parts of the MEME suite (MEME, Centrimo, FIMO)	-	-
Homer		Used the findmotifs.pl command, with full-genome fasta file as background	Motif in ~5% of peaks, flagged as possible false positive	-
KMAC		Used the --neg_seq option with the full genome fasta file, to provide a background k-mer frequency.	In ~10% of peaks, CG dinucleotide repeat <sup>[note 5]</sup>	-
gquad		[R package] Look for prevalence of G-Quadruplex and Z-DNA forming sequences, in ChIP-Seq peaks and the full genome.	Same frequency in peaks and genome	Same frequency in peaks and genome
MEME suite		Used fasta-get-markov to determine mono-, di- and trinucleotide frequencies in bound regions and the full genome.	Slight differences in some dinucleotides <sup>(note 6)</sup>	Same frequency in peaks and genome

Notes:

1. In the case of HpyA, MEME searches were made for all peaks, as well as condition-specific peaks (reduced salt only, stationary phase only, exponential phase only) peaks, and also by using the promoter sequences of proximal genes instead of directly-bound regions.
2. For MEME, all the above parameters were used in all possible combinations, in addition to using nmotifs=3 to force an output of at least 3 motifs.
3. Sequence was a trinucleotide repeat of the form SSW, where S=C or G and W= A or T, found in 35/59 peaks. It was not present when adding the genome-wide di- or tri-nucleotide frequencies as background (Markov\_order 1 or 2 respectively), and FIMO analysis showed similar sequences were present in 20,000 other locations genome-wide. Hence, it was not considered to be a true motif.
4. Motif is of the form TCGnssnCGA, where n = any nucleotide and s = C or G. Unlike the trinucleotide motif of note 3, it was found only after subtracting the dinucleotide background, and could be obtained after correcting for trinucleotide frequencies too. As a palindrome, it also resembles known

haloarchaeal TFs. It was present in 31 of 32 peaks. All of this suggested it could be a plausible motif. It is shown in Figure 22A.

However, FIMO analysis showed it was present in 15,000 other locations genome-wide. Hence, it is uncertain if this is a strict binding motif.

5. A CG-dinucleotide repeat sequence was found after correcting for genome-wide k-mer frequencies. However, it was only present in 7 of 59 peaks.
6. AA, AT, and TT are slightly under-represented in the peaks (2.16%, 2.19%, 2.16%) compared to genome-wide (2.74%, 2.89%, 2.74%), while GC is slightly over-represented (1.17% vs 1.02%). Due to the small magnitude of the differences, this was not investigated further.

## Appendix D

### D.1. Mc1 knockout in *Haloferax volcanii*:

*mc1* is a chromatin protein present in all halophilic archaea<sup>75</sup>. It is hypothesized to be the chromatin protein of *Methanosarcina mazei*, whose histone gene is also dispensable<sup>90,103</sup>.

Given this background, and the non-essentiality of *hpyA* in *Halobacterium salinarum*, I decided to also study *mc1* when I investigated *hstA* in *Haloferax volcanii*. As with *hstA*, a  $\Delta mc1$  strain was made and Sanger sequencing around the *mc1* locus revealed no abnormalities. The strain was phenotyped and was found to be similar to  $\Delta hstA$ : slight growth defect in optimal conditions, no additional defect in a variety of stress conditions including low and high salt, peroxide stress, and gluconeogenic conditions. However, when whole-genome sequencing of the gDNA of the knockout strain was analyzed using Breseq (similar to analysis done for  $\Delta hstA$ ), this revealed a secondary mutation (G>A) at position 970,322 on the main chromosome, within the promoter region of the gene HVO\_1062. This lack of a clean deletion leads to 3 possibilities:

1. *mc1* is an essential gene; this secondary mutation is necessary for the *mc1* to be deleted
2. *mc1* is not essential, but observed phenotype is influenced by the secondary mutation.
3. *mc1* isn't essential, and observed phenotype is unaffected by the secondary mutation.

These possibilities would ideally be resolved with the creation of a fresh  $\Delta mc1$  strain.

Work on that is ongoing but will be incomplete before I finish my thesis work.

Appearance of another secondary site mutation, especially at the same locus, would strongly suggest essentiality of *mc1*, and link its function with the HVO\_1062 gene. In this case, complementing  $\Delta mc1$  with *mc1* expressed in-trans would clarify if observed

phenotypes are due to the deletion itself, and not the secondary mutation. A clean deletion, on the other hand, would allow re-investigation of  $\Delta mc1$  phenotype.

**Table D1: List of primers used to create and sequence  $\Delta mc1$ . All primers created for this study by Saaz Sakrikar (SS)**

Primer Name	Direction	Purpose	Sequence (5'-3')
Hvmc1seq up	F	Mc1 KO check, sequencing,	CGTCGAACCGATGGCGC
Hvmc1seq dn	R	Mc1 KO check, sequencing.	AGTCGGTTCGAGACGATACGG
Hvmc1SN P_f	F	Sequencing secondary site mutation in $\Delta mc1$	ACCGGCTACTTCACCGAC
Hvmc1SN P_r	R	Sequencing secondary site mutation in $\Delta mc1$	CGAGCGTCTGGTCGGT
mc1-KO-up1	F	<i>mc1</i> deletion (upstream amplification)	CTCGAGGTCGACGGTATCGATAA GCTTGATGAGGCAGGACGC GTTGAA
mc1-KO-up2	R	<i>mc1</i> deletion (upstream amplification)	TGCCATGTGGTTCATTGCCATCGA CCGACGTTACATTCTTATCG
mc1-KO-d1	R	<i>mc1</i> deletion (downstream amplification)	AGTGGATCCCCCGGGCTGCAGGA ATTCGATAGTCGGTTCGAGACGA TACGG
mc1-KO-d2	F	<i>mc1</i> deletion (downstream amplification)	CGATAAGAATGTGAACGTCGGTC GATGGCAATGAACCACATGGCA
mc1-native-rescue-up	F	Amplification of <i>mc1</i> and upstream region for cloning	CGGTCCGACAACAACCCCGATC CAAGCTTCCGTTGGTTCGCCGC
mc1-rescue-down	R	Amplification of <i>mc1</i> for cloning	CTTCACTTCTCGAACTGCGGGTGC GACCAGCGATAAGAATGTGAACG TCGGTCG
Hvmc1SN P_f	F	Sequencing secondary mutation of AKS197	ACCGGCTACTTCACCGAC
Hvmc1SN P_r	R	Sequencing secondary mutation of AKS197	CGAGCGTCTGGTCGGT

**Table D2: List of *mc1*-related plasmids made. All plasmids made for this study by Saaz Sakrikar (SS).**

Plasmid	Description	Purpose
pAKS146	pTA131:HVO_2941_ flanking	Suicide vector for deletion of <i>mc1</i> (HVO_2941) from <i>Hfx. volcanii</i>
pAKS148	pJAM809:P <sub>mc1</sub> :: <i>mc1</i>	Expression of <i>mc1</i> from its native promoter
pAKS149	pJAM809:P <sub>mc1</sub> :: <i>mc1</i> -HA	Expression of <i>mc1</i> -HA from its native promoter

**Table D3: List of *mc1*-related strains made. All strains created in *Hfx. volcanii* for this study by Saaz Sakrikar (SS).**

Strain	Schmid lab ID	Genotype	Description
AKS197	Hv252	$\Delta pyrE\Delta mc1$	Deletion of <i>mc1</i> (HVO_2941). Has secondary mutation.
AKS215	Hv270	$\Delta pyrE\Delta mc1$ /pAKS148	<i>Mc1</i> under its native promoter; $\Delta mc1$ background; Ura-Nov+
AKS216	Hv271	$\Delta pyrE\Delta mc1$ /pJAM809	pJAM809 vector with -HA tag, $\Delta mc1$ background; Ura-Nov+
AKS218	Hv273	$\Delta pyrE\Delta mc1$ /pAKS149	<i>Mc1</i> -HA under its native promoter; $\Delta mc1$ background; Ura-Nov+
AKS236	Hv291	$\Delta pyrE\Delta mc1\Delta hstA$	Double deletion of <i>hstA</i> (HVO_0520) and <i>mc1</i> (HVO_2941)
AKS304	Hv393	$\Delta pyrE\Delta mc1$	Deletion of <i>mc1</i> (HVO_2941). Not verified with whole genome resequencing.
AKS305	Hv394	$\Delta pyrE\Delta mc1$	Deletion of <i>mc1</i> (HVO_2941). Not verified with whole genome resequencing.

Strain AKS197 was found to have the secondary site mutation mentioned above, and it was also detected in strain AKS236. Since AKS236 was derived from AKS197, this was to be expected. Similarly, I expect the mutation to also be present for AKS215, AKS216, and AKS218. Strains AKS304 and AKS305 are independent, fresh, deletions which have not yet been sequenced at the time of this study, and may not have the same secondary mutations.

## D.2. Heat shock phenotype in *Haloferax volcanii* chromatin knockout strains:

As discussed in the Introduction, the best-characterized archaeal histones are from species that thrive at high temperature. By contrast, the three species with dispensable histones (*Methanosarcina mazei*<sup>103</sup>, *Halobacterium salinarum*<sup>104</sup>, *Haloferax volcanii* (this study)) are all mesophilic species. These observations, combined with the fact that I was unable to detect a salt-related phenotype for  $\Delta hstA$  in *Hfx. volcanii*, motivated an investigation of the response of  $\Delta hstA$  to high temperature. In addition, since I was also investigating the function of *mc1* (see section D.1 above), I also investigated its response to heat shock.

This was tested as follows: strains were plated and pre-cultured as usual (see section 3.4.1.). They were then diluted to OD~0.05 and grown (as usual) at 42°C in a BioScreen, for ~14 hours (till end of lag phase). At this point, the temperature was increased to a shock temperature, and growth was observed. This shock was done at three temperatures: 48°C, 56°C, and 58°C. Phenotype was measured by calculating area under the curve, taking OD600 data only after the shock, after normalizing it by subtracting OD600 of that replicate at the time of the shock.

No phenotype was apparent at 48°C for any strain. For 56°C, the ratio of  $\Delta hstA$  to parent was 0.52 (**Figure D1**), a much stronger phenotype than was seen in standard conditions (corresponding ratio = 0,84, **Figure 10**). This was observed from 7 biological replicates, each with 2-3 technical replicates, over two batches. However, the same ratio was 0.70

for 58°C, and hence no clear trend was apparent. Regardless, this phenotype warrants further investigation.

$\Delta mc1$  had an even greater growth defect at 56°C (**Figure D1**), with  $\Delta mc1$ :parent ratio of 0.43. However, as discussed above, due to the presence of a secondary site mutation, this cannot be credibly attributed to the effect of *mc1* deletion alone.

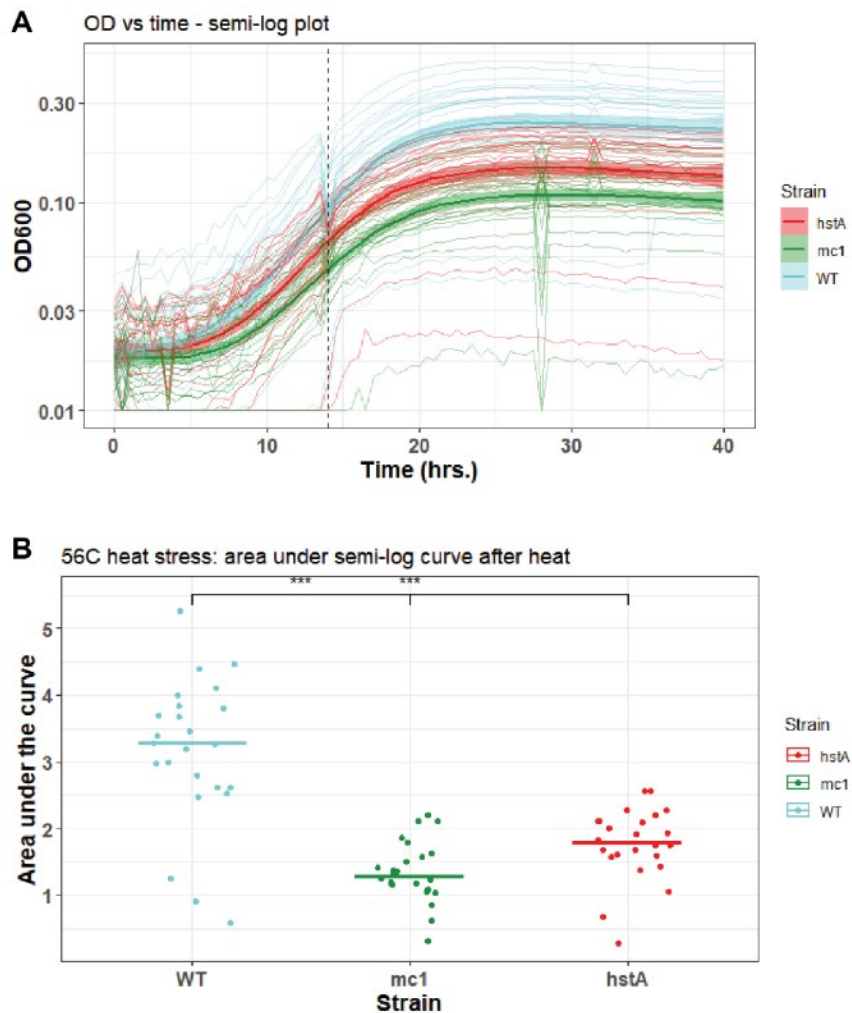


Figure D1: (A) Raw growth data for  $\Delta hstA$  (red),  $\Delta mc1$  (green) and parent (blue), with temperature raised to 56°C after 14 hours (dotted line). Y-axis is on a log-scale. Each line represents one technical replicate, and thick line represents 99% confidence interval of all replicates. (B) Area under the curve for the 3 strains (measured after start of shock); each dot represents one technical replicate.

A recent analysis of the chromatin proteins across the Archaea uncovered a link between growth temperature and expression of chromatin proteins, with the resulting hypothesis that growth temperature is the driver of chromatinization<sup>118</sup>. This study also revealed that HstA protein expression levels in *Hfx. volcanii* are very low, and while Mc1 is slightly higher, two as-yet-uninvestigated potential chromatin proteins have much higher expression levels. The question of the role of temperature in the requirement for chromatin-bound DNA and the identity of the principal chromatin protein in *Hfx. volcanii* hence requires further investigation, with the data shown here providing a basis for future work.

## References:

- 1 Woese, C. R., Kandler, O. & Wheelis, M. L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences* **87**, 4576, doi:10.1073/pnas.87.12.4576 (1990).
- 2 Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173-179, doi:10.1038/nature14447 (2015).
- 3 Raymann, K., Brochier-Armanet, C. & Gribaldo, S. The two-domain tree of life is linked to a new root for the Archaea. *Proceedings of the National Academy of Sciences* **112**, 6670, doi:10.1073/pnas.1420858112 (2015).
- 4 Peeters, E., Driessen, R. P. C., Werner, F. & Dame, R. T. The interplay between nucleoid organization and transcription in archaeal genomes. *Nature Reviews Microbiology* **13**, 333-341, doi:10.1038/nrmicro3467 (2015).
- 5 Gehring, A. M., Walker, J. E. & Santangelo, T. J. Transcription Regulation in Archaea. *Journal of Bacteriology* **198**, 1906, doi:10.1128/JB.00255-16 (2016).
- 6 Martinez-Pastor, M., Tonner, P. D., Darnell, C. L. & Schmid, A. K. Transcriptional Regulation in Archaea: From Individual Genes to Global Regulatory Networks. *Annual Review of Genetics* **51**, 143-170, doi:10.1146/annurev-genet-120116-023413 (2017).
- 7 Schmitt, E. *et al.* Recent Advances in Archaeal Translation Initiation. *Frontiers in Microbiology* **11**, 2259 (2020).
- 8 Benelli, D., La Teana, A. & Londei, P. in *RNA Metabolism and Gene Expression in Archaea* (ed Béatrice Clouet-d'Orval) 71-88 (Springer International Publishing, 2017).
- 9 Valentine, D. L. Adaptations to energy stress dictate the ecology and evolution of the Archaea. *Nature Reviews Microbiology* **5**, 316-323, doi:10.1038/nrmicro1619 (2007).
- 10 Schmid, A. K., Allers, T. & DiRuggiero, J. SnapShot: Microbial Extremophiles. *Cell* **180**, 818-818.e811, doi:<https://doi.org/10.1016/j.cell.2020.01.018> (2020).

- 11 DeLong, E. F. Everything in moderation: Archaea as 'non-extremophiles'. *Current Opinion in Genetics & Development* **8**, 649-654, doi:[https://doi.org/10.1016/S0959-437X\(98\)80032-4](https://doi.org/10.1016/S0959-437X(98)80032-4) (1998).
- 12 Bang, C. & Schmitz, R. A. Archaea: forgotten players in the microbiome. *Emerging Topics in Life Sciences* **2**, 459-468, doi:10.1042/ETLS20180035 (2018).
- 13 Cai, M. *et al.* Asgard archaea are diverse, ubiquitous, and transcriptionally active microbes. *bioRxiv*, 374165, doi:10.1101/374165 (2018).
- 14 Eme, L., Spang, A., Lombard, J., Stairs, C. W. & Ettema, T. J. G. Archaea and the origin of eukaryotes. *Nature Reviews Microbiology* **15**, 711-723, doi:10.1038/nrmicro.2017.133 (2017).
- 15 Imachi, H. *et al.* Isolation of an archaeon at the prokaryote–eukaryote interface. *Nature* **577**, 519-525, doi:10.1038/s41586-019-1916-6 (2020).
- 16 Guy, L. & Ettema, T. J. G. The archaeal TACK superphylum and the origin of eukaryotes. *Trends in Microbiology* **19**, 580-587, doi:10.1016/j.tim.2011.09.002 (2011).
- 17 Quehenberger, J., Shen, L., Albers, S.-V., Siebers, B. & Spadiut, O. Sulfolobus – A Potential Key Organism in Future Biotechnology. *Frontiers in Microbiology* **8**, 2474 (2017).
- 18 Laursen, S. P., Bowerman, S. & Luger, K. Archaea: The Final Frontier of Chromatin. (2021).
- 19 Cubonová, L. u., Sandman, K., Hallam, S. J., DeLong, E. F. & Reeve, J. N. Histones in crenarchaea. *Journal of bacteriology* **187**, 5482-5485, doi:10.1128/JB.187.15.5482-5485.2005 (2005).
- 20 Soppa, J. Ploidy and gene conversion in Archaea. *Biochemical Society Transactions* **39**, 150-154, doi:10.1042/BST0390150 (2011).
- 21 Dombrowski, N., Lee, J.-H., Williams, T. A., Offre, P. & Spang, A. Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. *FEMS Microbiology Letters* **366**, fnz008, doi:10.1093/femsle/fnz008 (2019).
- 22 Spaans, S. K., van der Oost, J. & Kengen, S. W. M. The chromosome copy number of the hyperthermophilic archaeon *Thermococcus kodakarensis* KOD1. *Extremophiles* **19**, 741-750, doi:10.1007/s00792-015-0750-5 (2015).

- 23 Zerulla, K. & Soppa, J. Polyploidy in haloarchaea: advantages for growth and survival. *Frontiers in Microbiology* **5**, 274 (2014).
- 24 Cavalier-Smith, T. & Chao, E. E. Y. Multidomain ribosomal protein trees and the planctobacterial origin of neomura (eukaryotes, archaeobacteria). *Protoplasma* **257**, 621-753, doi:10.1007/s00709-019-01442-7 (2020).
- 25 Oren, A. Microbial life at high salt concentrations: phylogenetic and metabolic diversity. *Saline Systems* **4**, 2, doi:10.1186/1746-1448-4-2 (2008).
- 26 Vauclare, P. *et al.* Molecular adaptation and salt stress response of Halobacterium salinarum cells revealed by neutron spectroscopy. *Extremophiles* **19**, 1099-1107, doi:10.1007/s00792-015-0782-x (2015).
- 27 Mevarech, M., Frolow, F. & Gloss, L. M. Halophilic enzymes: proteins with a grain of salt. *Biophysical Chemistry* **86**, 155-164, doi:[https://doi.org/10.1016/S0301-4622\(00\)00126-5](https://doi.org/10.1016/S0301-4622(00)00126-5) (2000).
- 28 Bergqvist, S., Williams, M. A., Brien, R. & Ladbury, J. E. Halophilic adaptation of protein-DNA interactions. *Biochemical Society Transactions* **31**, 677 (2003).
- 29 Stan-Lotter, H. & Fendrihan, S. Halophilic Archaea: Life with Desiccation, Radiation and Oligotrophy over Geological Times. *Life* **5**, doi:10.3390/life5031487 (2015).
- 30 Jones, D. L. & Baxter, B. K. DNA Repair and Photoprotection: Mechanisms of Overcoming Environmental Ultraviolet Radiation Exposure in Halophilic Archaea. *Frontiers in Microbiology* **8**, 1882 (2017).
- 31 Matarredona, L., Camacho, M., Zafrilla, B., Bonete, M.-J. & Esclapez, J. The Role of Stress Proteins in Haloarchaea and Their Adaptive Response to Environmental Shifts. *Biomolecules* **10**, 1390, doi:10.3390/biom10101390 (2020).
- 32 Bonneau, R. *et al.* A Predictive Model for Transcriptional Control of Physiology in a Free Living Cell. *Cell* **131**, 1354-1365, doi:10.1016/j.cell.2007.10.053 (2007).
- 33 Coker, J. A., DasSarma, P., Kumar, J., Müller, J. A. & DasSarma, S. Transcriptional profiling of the model Archaeon Halobacterium sp. NRC-1: responses to changes in salinity and temperature. *Saline systems* **3**, 6-6, doi:10.1186/1746-1448-3-6 (2007).
- 34 Martinez-Pastor, M., Lancaster, W. A., Tonner, P. D., Adams, Michael W. W. & Schmid, A. K. A transcription network of interlocking positive feedback loops

- maintains intracellular iron balance in archaea. *Nucleic Acids Research* **45**, 9990-10001, doi:10.1093/nar/gkx662 (2017).
- 35 Sharma, K., Gillum, N., Boyd, J. L. & Schmid, A. The RosR transcription factor is required for gene expression dynamics in response to extreme oxidative stress in a hypersaline-adapted archaeon. *BMC Genomics* **13**, 351, doi:10.1186/1471-2164-13-351 (2012).
- 36 Hackley, R. K. & Schmid, A. K. Global Transcriptional Programs in Archaea Share Features with the Eukaryotic Environmental Stress Response. *Journal of Molecular Biology* **431**, 4147-4166, doi:<https://doi.org/10.1016/j.jmb.2019.07.029> (2019).
- 37 Breuert, S., Allers, T., Spohn, G. & Soppa, J. Regulated Polyploidy in Halophilic Archaea. *PLOS ONE* **1**, e92, doi:10.1371/journal.pone.0000092 (2006).
- 38 Jones, D. L. & Baxter, B. K. Bipyrimidine Signatures as a Photoprotective Genome Strategy in G + C-rich Halophilic Archaea. *Life* **6**, doi:10.3390/life6030037 (2016).
- 39 Robinson, J. L. *et al.* Growth kinetics of extremely halophilic archaea (family halobacteriaceae) as revealed by arrhenius plots. *Journal of bacteriology* **187**, 923-929, doi:10.1128/JB.187.3.923-929.2005 (2005).
- 40 Ng, W. V. *et al.* Genome sequence of Halobacterium species NRC-1. *Proceedings of the National Academy of Sciences* **97**, 12176, doi:10.1073/pnas.190337797 (2000).
- 41 Hartman, A. L. *et al.* The Complete Genome Sequence of Haloferax volcanii DS2, a Model Archaeon. *PLOS ONE* **5**, e9605, doi:10.1371/journal.pone.0009605 (2010).
- 42 Grogan, D. W. in *Brenner's Encyclopedia of Genetics (Second Edition)* (eds Stanley Maloy & Kelly Hughes) 180-182 (Academic Press, 2013).
- 43 Wu, Z., Liu, J., Yang, H. & Xiang, H. DNA replication origins in archaea. *Frontiers in Microbiology* **5** (2014).
- 44 Peck, R. F., DasSarma, S. & Krebs, M. P. Homologous gene knockout in the archaeon Halobacterium salinarum with ura3 as a counterselectable marker. *Molecular Microbiology* **35**, 667-676, doi:<https://doi.org/10.1046/j.1365-2958.2000.01739.x> (2000).
- 45 Bitan-Banin, G., Ortenberg, R. & Mevarech, M. Development of a Gene Knockout System for the Halophilic Archaeon Haloferax volcanii by Use of the pyrE Gene. *J Bacteriol* **185**, 772-778, doi:10.1128/JB.185.3.772-778.2003 (2003).

- 46 Allers, T., Ngo, H.-P., Mevarech, M. & Lloyd, R. G. Development of additional selectable markers for the halophilic archaeon *Haloferax volcanii* based on the *leuB* and *trpA* genes. *Appl Environ Microbiol* **70**, 943-953, doi:10.1128/AEM.70.2.943-953.2004 (2004).
- 47 Liu, H., Han, J., Liu, X., Zhou, J. & Xiang, H. Development of *pyrF*-based gene knockout systems for genome-wide manipulation of the archaea *Haloferax mediterranei* and *Haloarcula hispanica*. *Journal of Genetics and Genomics* **38**, 261-269, doi:<https://doi.org/10.1016/j.jgg.2011.05.003> (2011).
- 48 Facciotti, M. T. *et al.* General transcription factor specified global gene regulation in archaea. *Proceedings of the National Academy of Sciences* **104**, 4630, doi:10.1073/pnas.0611663104 (2007).
- 49 Wilbanks, E. G. *et al.* A workflow for genome-wide mapping of archaeal transcription factors with ChIP-seq. *Nucleic Acids Research* **40**, e74-e74, doi:10.1093/nar/gks063 (2012).
- 50 Luijsterburg, M. S., White, M. F., van Driel, R. & Dame, R. T. The Major Architects of Chromatin: Architectural Proteins in Bacteria, Archaea and Eukaryotes. *Critical Reviews in Biochemistry and Molecular Biology* **43**, 393-418, doi:10.1080/10409230802528488 (2008).
- 51 Kundu, T. K. & Sikder, S. Evolution of genome organization and epigenetic machineries. *Journal of Biosciences* **43**, 239-242, doi:10.1007/s12038-018-9742-9 (2018).
- 52 Talbert, P. B. & Henikoff, S. Histone variants ancient wrap artists of the epigenome. *Nat Rev Mol Cell Biol* **11**, 264-275, doi:[http://www.nature.com/nrm/journal/v11/n4/suppinfo/nrm2861\\_S1.html](http://www.nature.com/nrm/journal/v11/n4/suppinfo/nrm2861_S1.html) (2010).
- 53 Dueva, R. *et al.* Neutralization of the Positive Charges on Histone Tails by RNA Promotes an Open Chromatin Structure. *Cell Chemical Biology* **26**, 1436-1449.e1435, doi:10.1016/j.chembiol.2019.08.002 (2019).
- 54 Arents, G. & Moudrianakis, E. N. The histone fold: a ubiquitous architectural motif utilized in DNA compaction and protein dimerization. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 11170-11174 (1995).
- 55 Kornberg, R. D. & Lorch, Y. Twenty-Five Years of the Nucleosome, Fundamental Particle of the Eukaryote Chromosome. *Cell* **98**, 285-294, doi:10.1016/S0092-8674(00)81958-3 (1999).

- 56 Ramazi, S., Allahverdi, A. & Zahiri, J. Evaluation of post-translational modifications in histone proteins: A review on histone modification defects in developmental and neurological disorders. *Journal of Biosciences* **45**, 135, doi:10.1007/s12038-020-00099-2 (2020).
- 57 Zhou, B.-R. & Bai, Y. Chromatin structures condensed by linker histones. *Essays in Biochemistry* **63**, 75-87, doi:10.1042/EBC20180056 (2019).
- 58 Hołowka, J. & Zakrzewska-Czerwińska, J. Nucleoid Associated Proteins: The Small Organizers That Help to Cope With Stress. *Frontiers in Microbiology* **11**, 590 (2020).
- 59 Dame, R. T., Rashid, F.-Z. M. & Grainger, D. C. Chromosome organization in bacteria: mechanistic insights into genome structure and function. *Nature Reviews Genetics* **21**, 227-242, doi:10.1038/s41576-019-0185-4 (2020).
- 60 Ali Azam, T., Iwata, A., Nishimura, A., Ueda, S. & Ishihama, A. Growth Phase-Dependent Variation in Protein Composition of the Escherichia coli Nucleoid. *Journal of Bacteriology* **181**, 6361-6370, doi:10.1128/JB.181.20.6361-6370.1999 (1999).
- 61 Tanaka, H. *et al.* Role of HU proteins in forming and constraining supercoils of chromosomal DNA in Escherichia coli. *Molecular and General Genetics MGG* **248**, 518-526, doi:10.1007/BF02423446 (1995).
- 62 Malik, H. S. & Henikoff, S. Phylogenomics of the nucleosome. *Nat Struct Mol Biol* **10**, 882-891 (2003).
- 63 Henneman, B., van Emmerik, C., van Ingen, H. & Dame, R. T. Structure and function of archaeal histones. *PLOS Genetics* **14**, e1007582, doi:10.1371/journal.pgen.1007582 (2018).
- 64 Sandman, K. & Reeve, J. N. Archaeal histones and the origin of the histone fold. *Current Opinion in Microbiology* **9**, 520-525, doi:<http://dx.doi.org/10.1016/j.mib.2006.08.003> (2006).
- 65 Decanniere, K., Babu, A. M., Sandman, K., Reeve, J. N. & Heinemann, U. Crystal structures of recombinant histones HMfA and HMfB from the hyperthermophilic archaeon Methanothermus fervidus. *Journal of Molecular Biology* **303**, 35-47, doi:<http://dx.doi.org/10.1006/jmbi.2000.4104> (2000).
- 66 Sandman, K., Krzycki, J. A., Dobrinski, B., Lurz, R. & Reeve, J. N. HMf, a DNA-binding protein isolated from the hyperthermophilic archaeon Methanothermus

- fervidus, is most closely related to histones. *Proceedings of the National Academy of Sciences of the United States of America* **87**, 5788-5791 (1990).
- 67 Maruyama, H. *et al.* Histone and TK0471/TrmBL2 form a novel heterogeneous genome architecture in the hyperthermophilic archaeon *Thermococcus kodakarensis*. *Molecular Biology of the Cell* **22**, 386-398, doi:10.1091/mbc.e10-08-0668 (2010).
- 68 Mattioli, F. *et al.* Structure of histone-based chromatin in Archaea. *Science* **357**, 609, doi:10.1126/science.aaj1849 (2017).
- 69 Goyal, M., Banerjee, C., Nag, S. & Bandyopadhyay, U. The Alba protein family: Structure and function. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1864**, 570-583, doi:<https://doi.org/10.1016/j.bbapap.2016.02.015> (2016).
- 70 Guo, L. *et al.* Biochemical and structural characterization of Cren7, a novel chromatin protein conserved among Crenarchaea. *Nucleic Acids Research* **36**, 1129-1137, doi:10.1093/nar/gkm1128 (2008).
- 71 Zhang, Z. *et al.* Archaeal Chromatin Proteins Cren7 and Sul7d Compact DNA by Bending and Bridging. *mBio* **11**, e00804-00820, doi:10.1128/mBio.00804-20.
- 72 Maruyama, H. *et al.* Different Proteins Mediate Step-Wise Chromosome Architectures in *Thermoplasma acidophilum* and *Pyrobaculum calidifontis*. *Frontiers in Microbiology* **11**, 1247 (2020).
- 73 DeLange, R. J., Williams, L. C. & Searcy, D. G. A histone-like protein (HTa) from *Thermoplasma acidophilum*. II. Complete amino acid sequence. *Journal of Biological Chemistry* **256**, 905-911, doi:[https://doi.org/10.1016/S0021-9258\(19\)70065-9](https://doi.org/10.1016/S0021-9258(19)70065-9) (1981).
- 74 Hocher, A., Rojec, M., Swadling, J. B., Esin, A. & Warnecke, T. The DNA-binding protein HTa from *Thermoplasma acidophilum* is an archaeal histone analog. *eLife* **8**, e52542, doi:10.7554/eLife.52542 (2019).
- 75 Sandman, K. & Reeve, J. N. Archaeal chromatin proteins: different structures but common function? *Current Opinion in Microbiology* **8**, 656-661, doi:<https://doi.org/10.1016/j.mib.2005.10.007> (2005).
- 76 Chartier, F., Laine, B., Belaïche, D., Touzel, J.-P. & Sautière, P. Primary structure of the chromosomal protein MC1 from the archaeobacterium *Methanosarcina* sp. CHTI 55. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression* **1008**, 309-314, doi:[https://doi.org/10.1016/0167-4781\(89\)90021-3](https://doi.org/10.1016/0167-4781(89)90021-3) (1989).

- 77 Laine, B., Chartier, F., Imbert, M., Lewis, R. & Sautiere, P. Primary structure of the chromosomal protein HMB from the archaeobacteria *Methanosarcina barkeri*. *European Journal of Biochemistry* **161**, 681-687, doi:<https://doi.org/10.1111/j.1432-1033.1986.tb10493.x> (1986).
- 78 De Vuyst, G., Aci, S., Genest, D. & Culard, F. Atypical Recognition of Particular DNA Sequences by the Archaeal Chromosomal MC1 Protein. *Biochemistry* **44**, 10369-10377, doi:10.1021/bi0474416 (2005).
- 79 Cam, E. L., Culard, F., Larquet, E., Delain, E. & Cognet, J. A. H. DNA Bending Induced by the Archaeobacterial Histone-like Protein MC1. *Journal of Molecular Biology* **285**, 1011-1021, doi:<https://doi.org/10.1006/jmbi.1998.2321> (1999).
- 80 Woodcock, C. L. & Ghosh, R. P. Chromatin higher-order structure and dynamics. *Cold Spring Harb Perspect Biol* **2**, a000596-a000596, doi:10.1101/cshperspect.a000596 (2010).
- 81 Le Tung, B. K., Imakaev Maxim, V., Mirny Leonid, A. & Laub Michael, T. High-Resolution Mapping of the Spatial Organization of a Bacterial Chromosome. *Science* **342**, 731-734, doi:10.1126/science.1242059 (2013).
- 82 Marbouty, M. *et al.* Condensin- and Replication-Mediated Bacterial Chromosome Folding and Origin Condensation Revealed by Hi-C and Super-resolution Imaging. *Mol Cell* **59**, 588-602, doi:10.1016/j.molcel.2015.07.020 (2015).
- 83 Takemata, N., Samson, R. Y. & Bell, S. D. Physical and Functional Compartmentalization of Archaeal Chromosomes. *Cell* **179**, 165-179.e118, doi:10.1016/j.cell.2019.08.036 (2019).
- 84 Cockram, C., Thierry, A., Gorlas, A., Lestini, R. & Koszul, R. Euryarchaeal genomes are folded into SMC-dependent loops and domains, but lack transcription-mediated compartmentalization. *Mol Cell* **81**, 459-472.e410, doi:10.1016/j.molcel.2020.12.013 (2021).
- 85 Badrinarayanan, A., Lesterlin, C., Reyes-Lamothe, R. & Sherratt, D. The *Escherichia coli* SMC Complex, MukBEF, Shapes Nucleoid Organization Independently of DNA Replication. *Journal of Bacteriology* **194**, 4669-4676, doi:10.1128/JB.00957-12 (2012).
- 86 Hirano, T. SMC protein complexes and higher-order chromosome dynamics. *Current Opinion in Cell Biology* **10**, 317-322, doi:[https://doi.org/10.1016/S0955-0674\(98\)80006-9](https://doi.org/10.1016/S0955-0674(98)80006-9) (1998).

- 87 Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251-260, doi:10.1038/38444 (1997).
- 88 Wang, Q. & Calvo, J. M. Lrp, a major regulatory protein in Escherichia coli, bends DNA and can organize the assembly of a higher-order nucleoprotein structure. *The EMBO journal* **12**, 2495-2501 (1993).
- 89 Lurz, R., Grote, M., Dijk, J., Reinhardt, R. & Dobrinski, B. Electron microscopic study of DNA complexes with proteins from the Archaeobacterium Sulfolobus acidocaldarius. *The EMBO Journal* **5**, 3715-3721, doi:<https://doi.org/10.1002/j.1460-2075.1986.tb04705.x> (1986).
- 90 Paquet, F. *et al.* Model of a DNA-Protein Complex of the Architectural Monomeric Protein MC1 from Euryarchaea. *PLOS ONE* **9**, e88809, doi:10.1371/journal.pone.0088809 (2014).
- 91 Sandman, K., Grayling, R. A., Dobrinski, B., Lurz, R. & Reeve, J. N. Growth-phase-dependent synthesis of histones in the archaeon Methanothermus fervidus. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 12624-12628 (1994).
- 92 Čuboňová, L. *et al.* An Archaeal Histone Is Required for Transformation of Thermococcus kodakarensis. *Journal of Bacteriology* **194**, 6864-6874, doi:10.1128/JB.01523-12 (2012).
- 93 Marc, F., Sandman, K., Lurz, R. & Reeve, J. N. Archaeal Histone Tetramerization Determines DNA Affinity and the Direction of DNA Supercoiling. *Journal of Biological Chemistry* **277**, 30879-30886, doi:10.1074/jbc.M203674200 (2002).
- 94 Tomschik, M., Karymov, M. A., Zlatanova, J. & Leuba, S. H. The Archaeal Histone-Fold Protein HMf Organizes DNA into Bona Fide Chromatin Fibers. *Structure* **9**, 1201-1211, doi:[http://dx.doi.org/10.1016/S0969-2126\(01\)00682-7](http://dx.doi.org/10.1016/S0969-2126(01)00682-7) (2001).
- 95 Nalabothula, N. *et al.* Archaeal nucleosome positioning in vivo and in vitro is directed by primary sequence motifs. *BMC genomics* **14**, 391-391, doi:10.1186/1471-2164-14-391 (2013).
- 96 Henneman, B. *et al.* Mechanical and structural properties of archaeal hypernucleosomes. *Nucleic Acids Research* **49**, 4338-4349, doi:10.1093/nar/gkaa1196 (2021).

- 97 Maruyama, H. *et al.* An alternative beads-on-a-string chromatin architecture in *Thermococcus kodakarensis*. *EMBO Reports* **14**, 711-717, doi:10.1038/embor.2013.94 (2013).
- 98 Sandman, K., Louvel, H., Samson, R. Y., Pereira, S. L. & Reeve, J. N. Archaeal chromatin proteins histone HMtB and Alba have lost DNA-binding ability in laboratory strains of *Methanothermobacter thermautotrophicus*. *Extremophiles* **12**, 811, doi:10.1007/s00792-008-0185-3 (2008).
- 99 Slesarev, A. I., Belova, G. I., Kozyavkin, S. A. & Lake, J. A. Evidence for an early prokaryotic origin of histones H2A and H4 prior to the emergence of eukaryotes. *Nucleic Acids Research* **26**, 427-430 (1998).
- 100 Pavlov, N. A., Cherny, D. I., Jovin, T. M. & Slesarev, A. I. Nucleosome-like Complex of the Histone from the Hyperthermophile *Methanopyrus Kandleri* (MkaH) with Linear DNA. *Journal of Biomolecular Structure and Dynamics* **20**, 207-214, doi:10.1080/07391102.2002.10506836 (2002).
- 101 Fahrner, R. L., Cascio, D., Lake, J. A. & Slesarev, A. An ancestral nuclear protein assembly: Crystal structure of the *Methanopyrus kandleri* histone. *Protein Science* **10**, 2002-2007, doi:10.1110/ps.10901 (2001).
- 102 Stevens, K. M. *et al.* Histone variants in archaea and the evolution of combinatorial chromatin complexity. *Proceedings of the National Academy of Sciences* **117**, 33384, doi:10.1073/pnas.2007056117 (2020).
- 103 Weidenbach, K. *et al.* Deletion of the archaeal histone in *Methanosarcina mazei* Gö1 results in reduced growth and genomic transcription. *Molecular Microbiology* **67**, 662-671, doi:10.1111/j.1365-2958.2007.06076.x (2008).
- 104 Dulmage, K. A., Todor, H. & Schmid, A. K. Growth-Phase-Specific Modulation of Cell Morphology and Gene Expression by an Archaeal Histone Protein. *mBio* **6**, doi:10.1128/mBio.00649-15 (2015).
- 105 Takayanagi, S. *et al.* Chromosomal structure of the halophilic archaeobacterium *Halobacterium salinarium*. *Journal of Bacteriology* **174**, 7207-7216 (1992).
- 106 Eickbush, T. H. & Moudrianakis, E. N. The compaction of DNA helices into either continuous supercoils or folded-fiber rods and toroids. *Cell* **13**, 295-306, doi:[https://doi.org/10.1016/0092-8674\(78\)90198-8](https://doi.org/10.1016/0092-8674(78)90198-8) (1978).

- 107 Widom, J. Physicochemical studies of the folding of the 100 Å nucleosome filament into the 300 Å filament: Cation dependence. *Journal of Molecular Biology* **190**, 411-424, doi:[https://doi.org/10.1016/0022-2836\(86\)90012-4](https://doi.org/10.1016/0022-2836(86)90012-4) (1986).
- 108 Ammar, R. *et al.* Chromatin is an ancient innovation conserved between Archaea and Eukarya. *eLife* **1**, e00078, doi:10.7554/eLife.00078 (2012).
- 109 Schmid, A. K. *et al.* The anatomy of microbial cell state transitions in response to oxygen. *Genome research* **17**, 1399-1413, doi:10.1101/gr.6728007 (2007).
- 110 Sakrikar, S. & Schmid, Amy K. An archaeal histone-like protein regulates gene expression in response to salt stress. *Nucleic Acids Research* **49**, 12732-12743, doi:10.1093/nar/gkab1175 (2021).
- 111 Brunk, C. F. & Martin, W. F. Archaeal Histone Contributions to the Origin of Eukaryotes. *Trends in Microbiology* **27**, 703-714, doi:10.1016/j.tim.2019.04.002 (2019).
- 112 Mariño-Ramírez, L., Kann, M. G., Shoemaker, B. A. & Landsman, D. Histone structure and nucleosome stability. *Expert Review of Proteomics* **2**, 719-729, doi:10.1586/14789450.2.5.719 (2005).
- 113 Arents, G. & Moudrianakis, E. N. Topography of the histone octamer surface: repeating structural motifs utilized in the docking of nucleosomal DNA. *Proc Natl Acad Sci U S A* **90**, 10489-10493, doi:10.1073/pnas.90.22.10489 (1993).
- 114 Sanders, T. J. *et al.* Extended Archaeal Histone-Based Chromatin Structure Regulates Global Gene Expression in *Thermococcus kodakarensis*. *Frontiers in Microbiology* **12**, 1071 (2021).
- 115 Sanders, T. J., Marshall, C. J. & Santangelo, T. J. The Role of Archaeal Chromatin in Transcription. *Journal of Molecular Biology* **431**, 4103-4115, doi:<https://doi.org/10.1016/j.jmb.2019.05.006> (2019).
- 116 Xie, Y. & Reeve, J. N. Transcription by an Archaeal RNA Polymerase Is Slowed but Not Blocked by an Archaeal Nucleosome. *Journal of Bacteriology* **186**, 3492, doi:10.1128/JB.186.11.3492-3498.2004 (2004).
- 117 Wilkinson, S. P., Ouhammouch, M. & Geiduschek, E. P. Transcriptional activation in the context of repression mediated by archaeal histones. *Proceedings of the National Academy of Sciences* **107**, 6777, doi:10.1073/pnas.1002360107 (2010).

- 118 Hocher, A. *et al.* Growth temperature is the principal driver of chromatinization in archaea. *bioRxiv*, doi:10.1101/2021.07.08.451601 (2021).
- 119 Soares, D. *et al.* Archaeal histone stability, DNA binding, and transcription inhibition above 90°C. *Extremophiles* **2**, 75-81, doi:10.1007/s007920050045 (1998).
- 120 Wagner, G., Hartmann, R. & Oesterhelt, D. Potassium uniport and ATP synthesis in *Halobacterium halobium*. *Eur J Biochem* **89**, 169-179, doi:10.1111/j.1432-1033.1978.tb20909.x (1978).
- 121 Dulmage, K. Large-scale Effectors of Gene Expression and New Models of Cell Division in the Haloarchaea. (2015).
- 122 Kahm, M., Hasenbrink, G., Lichtenberg-Fraté, H., Ludwig, J. & Kschischo, M. grofit: Fitting Biological Growth Curves with R. *2010* **33**, 21, doi:10.18637/jss.v033.i07 (2010).
- 123 Darnell, C. L. *et al.* The Ribbon-Helix-Helix Domain Protein CdrS Regulates the Tubulin Homolog ftsZ2 To Control Cell Division in Archaea. *mBio* **11**, e01007-01020, doi:10.1128/mBio.01007-20 (2020).
- 124 Ducret, A., Quardokus, E. M. & Brun, Y. V. MicrobeJ, a tool for high throughput bacterial cell detection and quantitative analysis. *Nat Microbiol* **1**, 16077, doi:10.1038/nmicrobiol.2016.77 (2016).
- 125 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).
- 126 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
- 127 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137, doi:10.1186/gb-2008-9-9-r137 (2008).
- 128 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).
- 129 Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9**, e1003118, doi:10.1371/journal.pcbi.1003118 (2013).
- 130 Koide, T. *et al.* Prevalence of transcription promoters within archaeal operons and coding sequences. *Mol Syst Biol* **5**, 285, doi:10.1038/msb.2009.42 (2009).

- 131 Taboada, B., Estrada, K., Ciria, R. & Merino, E. Operon-mapper: a web server for precise operon identification in bacterial and archaeal genomes. *Bioinformatics* **34**, 4118-4120, doi:10.1093/bioinformatics/bty496 (2018).
- 132 Schmid, A. K., Reiss, D. J., Pan, M., Koide, T. & Baliga, N. S. A single transcription factor regulates evolutionarily diverse but functionally linked metabolic pathways in response to nutrient availability. *Mol Syst Biol* **5**, 282-282, doi:10.1038/msb.2009.40 (2009).
- 133 Tonner, P. D., Pittman, A. M. C., Gulli, J. G., Sharma, K. & Schmid, A. K. A Regulatory Hierarchy Controls the Dynamic Transcriptional Response to Extreme Oxidative Stress in Archaea. *PLOS Genetics* **11**, e1004912, doi:10.1371/journal.pgen.1004912 (2015).
- 134 Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169, doi:10.1093/bioinformatics/btu638 (2015).
- 135 Chen, X., Zhang, B., Wang, T., Bonni, A. & Zhao, G. Robust principal component analysis for accurate outlier sample detection in RNA-Seq data. *BMC Bioinformatics* **21**, 269, doi:10.1186/s12859-020-03608-0 (2020).
- 136 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).
- 137 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289-300 (1995).
- 138 factoextra: Extract and Visualize the Results of Multivariate Data Analyses v. R package version 1.0.7 (2020).
- 139 Wickham, H. *ggplot2: elegant graphics for data analysis*. (Springer Verlag, 2016).
- 140 pheatmap: Pretty Heatmaps v. R package version 1.0.12 (2019).
- 141 Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Archaeal Clusters of Orthologous Genes (arCOGs): An Update and Application for Analysis of Shared Features between Thermococcales, Methanococcales, and Methanobacteriales. *Life* **5**, doi:10.3390/life5010818 (2015).

- 142 Eun, Y. J. *et al.* Archaeal cells share common size control with bacteria despite noisier growth and division. *Nat Microbiol* **3**, 148-154, doi:10.1038/s41564-017-0082-6 (2018).
- 143 Vauclare, P., Natali, F., Kleman, J. P., Zaccai, G. & Franzetti, B. Surviving salt fluctuations: stress and recovery in *Halobacterium salinarum*, an extreme halophilic Archaeon. *Sci Rep* **10**, 3298, doi:10.1038/s41598-020-59681-1 (2020).
- 144 Dyall-Smith, M. (ed Mike Dyall-Smith) (2009).
- 145 Guan, Z., Naparstek, S., Calo, D. & Eichler, J. Protein glycosylation as an adaptive response in Archaea: growth at different salt concentrations leads to alterations in *Haloferax volcanii* S-layer glycoprotein N-glycosylation. *Environ Microbiol* **14**, 743-753, doi:10.1111/j.1462-2920.2011.02625.x (2012).
- 146 Todor, H. *et al.* A transcription factor links growth rate and metabolism in the hypersaline adapted archaeon *Halobacterium salinarum*. *Mol Microbiol* **93**, 1172-1182, doi:10.1111/mmi.12726 (2014).
- 147 Stevenson, K., McVey, A. F., Clark, I. B. N., Swain, P. S. & Pilizota, T. General calibration of microbial growth in microplate readers. *Scientific Reports* **6**, 38828, doi:10.1038/srep38828 (2016).
- 148 Kaur, A. *et al.* A systems view of haloarchaeal strategies to withstand stress from transition metals. *Genome Res* **16**, 841-854, doi:10.1101/gr.5189606 (2006).
- 149 Outten, F. W. Recent advances in the Suf Fe-S cluster biogenesis pathway: Beyond the Proteobacteria. *Biochim Biophys Acta* **1853**, 1464-1469, doi:10.1016/j.bbamcr.2014.11.001 (2015).
- 150 Burrell, M., Hanfrey, C. C., Kinch, L. N., Elliott, K. A. & Michael, A. J. Evolution of a novel lysine decarboxylase in siderophore biosynthesis. *Mol Microbiol* **86**, 485-499, doi:10.1111/j.1365-2958.2012.08208.x (2012).
- 151 Andrei, A. S., Banciu, H. L. & Oren, A. Living with salt: metabolic and phylogenetic diversity of archaea inhabiting saline ecosystems. *FEMS Microbiol Lett* **330**, 1-9, doi:10.1111/j.1574-6968.2012.02526.x (2012).
- 152 Gonzalez, O. *et al.* Systems analysis of bioenergetics and growth of the extreme halophile *Halobacterium salinarum*. *PLoS Comput Biol* **5**, e1000332, doi:10.1371/journal.pcbi.1000332 (2009).

- 153 Schmid, A. K., Pan, M., Sharma, K. & Baliga, N. S. Two transcription factors are necessary for iron homeostasis in a salt-dwelling archaeon. *Nucleic Acids Res* **39**, 2519-2533, doi:10.1093/nar/gkq1211 (2011).
- 154 Falb, M. *et al.* Metabolism of halophilic archaea. *Extremophiles* **12**, 177-196, doi:10.1007/s00792-008-0138-x (2008).
- 155 Todor, H., Gooding, J., Ilkayeva, O. R. & Schmid, A. K. Dynamic Metabolite Profiling in an Archaeon Connects Transcriptional Regulation to Metabolic Consequences. *PLoS One* **10**, e0135693, doi:10.1371/journal.pone.0135693 (2015).
- 156 Mescher, M. F. & Strominger, J. L. Structural (shape-maintaining) role of the cell surface glycoprotein of *Halobacterium salinarium*. *Proc Natl Acad Sci U S A* **73**, 2687-2691, doi:10.1073/pnas.73.8.2687 (1976).
- 157 de Silva, R. T. *et al.* Improved growth and morphological plasticity of *Haloferax volcanii*. *Microbiology (Reading)* **167**, doi:10.1099/mic.0.001012 (2021).
- 158 Baliga, N. S. *et al.* Is gene expression in *Halobacterium* NRC-1 regulated by multiple TBP and TFB transcription factors? *Mol Microbiol* **36**, 1184-1185, doi:10.1046/j.1365-2958.2000.01916.x (2000).
- 159 Shimada, T., Ishihama, A., Busby, S. J. W. & Grainger, D. C. The Escherichia coli RutR transcription factor binds at targets within genes as well as intergenic regions. *Nucleic Acids Research* **36**, 3950-3955, doi:10.1093/nar/gkn339 (2008).
- 160 Nguyen Ple, M., Bervoets, I., Maes, D. & Charlier, D. The protein-DNA contacts in RutR\*carAB operator complexes. *Nucleic Acids Res* **38**, 6286-6300, doi:10.1093/nar/gkq385 (2010).
- 161 Dorman, C. J., Schumacher, M. A., Bush, M. J., Brennan, R. G. & Buttner, M. J. When is a transcription factor a NAP? *Current Opinion in Microbiology* **55**, 26-33 (2021).
- 162 de Mendoza, A. & Sebé-Pedrós, A. Origin and evolution of eukaryotic transcription factors. *Current Opinion in Genetics & Development* **58-59**, 25-32, doi:<https://doi.org/10.1016/j.gde.2019.07.010> (2019).
- 163 Seshasayee, A. S. N., Sivaraman, K. & Luscombe, N. M. in *A Handbook of Transcription Factors* (ed Timothy R. Hughes) 7-23 (Springer Netherlands, 2011).

- 164 Aravind, L., Anantharaman, V., Balaji, S., Babu, M. M. & Iyer, L. M. The many faces of the helix-turn-helix domain: Transcription regulation and beyond\*. *FEMS Microbiology Reviews* **29**, 231-262, doi:10.1016/j.fmrr.2004.12.008 (2005).
- 165 Romier, C., Cocchiarella, F., Mantovani, R. & Moras, D. The NF-YB/NF-YC Structure Gives Insight into DNA Binding and Transcription Regulation by CCAAT Factor NF-Y. *Journal of Biological Chemistry* **278**, 1336-1345, doi:10.1074/jbc.M209635200 (2003).
- 166 Ashworth, J., Plaisier, C. L., Lo, F. Y., Reiss, D. J. & Baliga, N. S. Inference of expanded Lrp-like feast/famine transcription factor targets in a non-model organism using protein structure-based prediction. *PLoS One* **9**, e107863, doi:10.1371/journal.pone.0107863 (2014).
- 167 Cremona, M. A. *et al.* Peak shape clustering reveals biological insights. *BMC Bioinformatics* **16**, 349, doi:10.1186/s12859-015-0787-6 (2015).
- 168 Kleinman, C. L. *et al.* ChIP-seq analysis of the LuxR-type regulator VjbR reveals novel insights into the Brucella virulence gene expression network. *Nucleic acids research* **45**, 5757-5769, doi:10.1093/nar/gkx165 (2017).
- 169 Mendoza-Parra, M. A., Nowicka, M., Van Gool, W. & Gronemeyer, H. Characterising ChIP-seq binding patterns by model-based peak shape deconvolution. *BMC Genomics* **14**, 834, doi:10.1186/1471-2164-14-834 (2013).
- 170 McKnight, J. N., Boerma, J. W., Breeden, L. L. & Tsukiyama, T. Global Promoter Targeting of a Conserved Lysine Deacetylase for Transcriptional Shutoff during Quiescence Entry. *Mol Cell* **59**, 732-743, doi:10.1016/j.molcel.2015.07.014 (2015).
- 171 Kahramanoglou, C. *et al.* Direct and indirect effects of H-NS and Fis on global gene expression control in Escherichia coli. *Nucleic Acids Research* **39**, 2073-2091, doi:10.1093/nar/gkq934 (2011).
- 172 Prieto, A. I. *et al.* Genomic analysis of DNA binding and gene regulation by homologous nucleoid-associated proteins IHF and HU in Escherichia coli K12. *Nucleic Acids Research* **40**, 3524-3537, doi:10.1093/nar/gkr1236 (2012).
- 173 Peterson, S. N., Dahlquist, F. W. & Reich, N. O. The Role of High Affinity Non-specific DNA Binding by Lrp in Transcriptional Regulation and DNA Organization. *Journal of Molecular Biology* **369**, 1307-1317, doi:<https://doi.org/10.1016/j.jmb.2007.04.023> (2007).

- 174 Kroner, G. M., Wolfe, M. B. & Freddolino, P. L. Escherichia coli Lrp Regulates One-Third of the Genome via Direct, Cooperative, and Indirect Routes. *Journal of bacteriology* **201**, e00411-00418, doi:10.1128/JB.00411-18 (2019).
- 175 Myers, K. S. *et al.* Genome-scale Analysis of Escherichia coli FNR Reveals Complex Features of Transcription Factor Binding. *PLOS Genetics* **9**, e1003565, doi:10.1371/journal.pgen.1003565 (2013).
- 176 Vassart, A. *et al.* Sa-Lrp from Sulfolobus acidocaldarius is a versatile, glutamine-responsive, and architectural transcriptional regulator. *Microbiologyopen* **2**, 75-93, doi:10.1002/mbo3.58 (2013).
- 177 Chereji, R. V., Ocampo, J. & Clark, D. J. MNase-Sensitive Complexes in Yeast: Nucleosomes and Non-histone Barriers. *Mol Cell* **65**, 565-577.e563, doi:10.1016/j.molcel.2016.12.009 (2017).
- 178 Brogaard, K., Xi, L., Wang, J.-P. & Widom, J. A map of nucleosome positions in yeast at base-pair resolution. *Nature* **486**, 496-501, doi:10.1038/nature11142 (2012).
- 179 Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442**, 772-778, doi:10.1038/nature04979 (2006).
- 180 Struhl, K. & Segal, E. Determinants of nucleosome positioning. *Nat Struct Mol Biol* (2013).
- 181 Trifonov, E. N. & Sussman, J. L. The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proceedings of the National Academy of Sciences of the United States of America* **77**, 3816-3820, doi:10.1073/pnas.77.7.3816 (1980).
- 182 Bonnet, J. *et al.* Quantification of Proteins and Histone Marks in Drosophila Embryos Reveals Stoichiometric Relationships Impacting Chromatin Regulation. *Developmental Cell* **51**, 632-644.e636, doi:<https://doi.org/10.1016/j.devcel.2019.09.011> (2019).
- 183 Marguerat, S. *et al.* Quantitative Analysis of Fission Yeast Transcriptomes and Proteomes in Proliferating and Quiescent Cells. *Cell* **151**, 671-683, doi:<https://doi.org/10.1016/j.cell.2012.09.019> (2012).
- 184 Bailey, K. A., Pereira, S. L., Widom, J. & Reeve, J. N. Archaeal histone selection of nucleosome positioning sequences and the procaryotic origin of histone-dependent genome evolution. *Journal of molecular biology* **303**, 25-34, doi:10.1006/jmbi.2000.4128 (2000).

- 185 Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871-876, doi:10.1126/science.abj8754 (2021).
- 186 Becker, E. A. *et al.* Phylogenetically driven sequencing of extremely halophilic archaea reveals strategies for static and dynamic osmo-response. *PLoS Genet* **10**, e1004784, doi:10.1371/journal.pgen.1004784 (2014).
- 187 Ludt, K. & Soppa, J. Polyploidy in halophilic archaea: regulation, evolutionary advantages, and gene conversion. *Biochemical Society Transactions* **47**, 933-944, doi:10.1042/BST20190256 (2019).
- 188 Dai, J. *et al.* Probing Nucleosome Function: A Highly Versatile Library of Synthetic Histone H3 and H4 Mutants. *Cell* **134**, 1066-1078, doi:10.1016/j.cell.2008.07.019 (2008).
- 189 Zhang, C., Phillips, A. P. R., Wipfler, R. L., Olsen, G. J. & Whitaker, R. J. The essential genome of the crenarchaeal model *Sulfolobus islandicus*. *Nature Communications* **9**, 4908, doi:10.1038/s41467-018-07379-4 (2018).
- 190 Cole, H. A., Ocampo, J., Iben, J. R., Chereji, R. V. & Clark, D. J. Heavy transcription of yeast genes correlates with differential loss of histone H2B relative to H4 and queued RNA polymerases. *Nucleic acids research* **42**, 12512-12522, doi:10.1093/nar/gku1013 (2014).
- 191 Balleza, E. *et al.* Regulation by transcription factors in bacteria: beyond description. *FEMS Microbiology Reviews* **33**, 133-151, doi:10.1111/j.1574-6976.2008.00145.x (2009).
- 192 Schieg, P. & Herzel, H. Periodicities of 10-11bp as indicators of the supercoiled state of genomic DNA. (2004).
- 193 Li, W.-T., Sandman, K., Pereira, S. L. & Reeve, J. N. MJ1647, an open reading frame in the genome of the hyperthermophile *Methanococcus jannaschii*, encodes a very thermostable archaeal histone with a C-terminal extension. *Extremophiles* **4**, 43-51, doi:10.1007/s007920050006 (2000).
- 194 Wang, J.-P. Z. & Widom, J. Improved alignment of nucleosome DNA sequences using a mixture model. *Nucleic Acids Research* **33**, 6743-6755, doi:10.1093/nar/gki977 (2005).
- 195 Slesarev Alexei, I. *et al.* The complete genome of hyperthermophile *Methanopyrus kandleri* AV19 and monophyly of archaeal methanogens.

- Proceedings of the National Academy of Sciences* **99**, 4644-4649, doi:10.1073/pnas.032671499 (2002).
- 196 Bailey, T. L. *et al.* MEME Suite: tools for motif discovery and searching. *Nucleic Acids Research* **37**, W202-W208, doi:10.1093/nar/gkp335 (2009).
- 197 Latif, H. *et al.* ChIP-exo interrogation of Crp, DNA, and RNAP holoenzyme interactions. *PLoS One* **13**, e0197272, doi:10.1371/journal.pone.0197272 (2018).
- 198 Grainger, D. C., Hurd, D., Harrison, M., Holdstock, J. & Busby, S. J. Studies of the distribution of Escherichia coli cAMP-receptor protein and RNA polymerase along the E. coli chromosome. *Proc Natl Acad Sci U S A* **102**, 17693-17698, doi:10.1073/pnas.0506687102 (2005).
- 199 Maklad, H. R., Gutierrez, G. J., Esser, D., Siebers, B. & Peeters, E. Phosphorylation of the acyl-CoA binding pocket of the FadR transcription regulator in Sulfolobus acidocaldarius. *Biochimie* **175**, 120-124, doi:10.1016/j.biochi.2020.05.007 (2020).
- 200 Wang, K. *et al.* A TetR-family transcription factor regulates fatty acid metabolism in the archaeal model organism Sulfolobus acidocaldarius. *Nat Commun* **10**, 1542, doi:10.1038/s41467-019-09479-1 (2019).
- 201 Liu, H. *et al.* BarR, an Lrp-type transcription factor in Sulfolobus acidocaldarius, regulates an aminotransferase gene in a beta-alanine responsive manner. *Mol Microbiol* **92**, 625-639, doi:10.1111/mmi.12583 (2014).
- 202 Deatherage, D. E., Traverse, C. C., Wolf, L. N. & Barrick, J. E. Detecting rare structural variation in evolving microbial populations from new sequence junctions using breseq. *Front Genet* **5**, 468, doi:10.3389/fgene.2014.00468 (2014).
- 203 Gelsinger, D. R. *et al.* Ribosome profiling in archaea reveals leaderless translation, novel translational initiation sites, and ribosome pausing at single codon resolution. *Nucleic Acids Research* **48**, 5201-5216, doi:10.1093/nar/gkaa304 (2020).
- 204 Mrázek, J. Comparative Analysis of Sequence Periodicity among Prokaryotic Genomes Points to Differences in Nucleoid Structure and a Relationship to Gene Expression. *Journal of Bacteriology* **192**, 3763, doi:10.1128/JB.00149-10 (2010).
- 205 Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010).

- 206 Guo, Y., Tian, K., Zeng, H., Guo, X. & Gifford, D. K. A novel k-mer set memory (KSM) motif representation improves regulatory variant prediction. *Genome Research* (2018).
- 207 Weber, A. P. M., Weber, K. L., Carr, K., Wilkerson, C. & Ohlrogge, J. B. Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. *Plant Physiol* **144**, 32-42, doi:10.1104/pp.107.096677 (2007).
- 208 Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (New York, N.Y.)* **320**, 1344-1349, doi:10.1126/science.1158441 (2008).
- 209 Bainbridge, M. N. *et al.* Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* **7**, 246, doi:10.1186/1471-2164-7-246 (2006).
- 210 Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**, 621-628, doi:10.1038/nmeth.1226 (2008).
- 211 Woese, C. R. & Fox, G. E. 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 (1977).
- 212 Angel, R., Claus, P. & Conrad, R. Methanogenic archaea are globally ubiquitous in aerated soils and become active under wet anoxic conditions. *The ISME Journal* **6**, 847-862, doi:10.1038/ismej.2011.141 (2012).
- 213 Browne, P. D. & Cadillo-Quiroz, H. Contribution of transcriptomics to systems-level understanding of methanogenic Archaea. *Archaea* **2013**, 586369-586369, doi:10.1155/2013/586369 (2013).
- 214 Gelsinger, D. R. & DiRuggiero, J. Transcriptional Landscape and Regulatory Roles of Small Noncoding RNAs in the Oxidative Stress Response of the Haloarchaeon *Haloferax volcanii*. *Journal of bacteriology* **200**, e00779-00717, doi:10.1128/JB.00779-17 (2018).
- 215 Cai, M. *et al.* Diverse Asgard archaea including the novel phylum Gerdarchaeota participate in organic matter degradation. *Science China Life Sciences* **63**, 886-897, doi:10.1007/s11427-020-1679-1 (2020).

- 216 Dulmage, K. A., Darnell, C. L., Vreugdenhil, A. & Schmid, A. K. Copy number variation is associated with gene expression change in archaea. *Microbial Genomics* **4**, doi:<https://doi.org/10.1099/mgen.0.000210> (2018).
- 217 Qi, L. *et al.* Genome-wide mRNA processing in methanogenic archaea reveals post-transcriptional regulation of ribosomal protein synthesis. *Nucleic acids research* **45**, 7285-7298, doi:10.1093/nar/gkx454 (2017).
- 218 Zhou, X. *et al.* The transcriptome response of the ruminal methanogen *Methanobrevibacter ruminantium* strain M1 to the inhibitor lauric acid. *BMC Research Notes* **11**, 135, doi:10.1186/s13104-018-3242-8 (2018).
- 219 Zhou, Z. *et al.* Genomic and transcriptomic insights into the ecology and metabolism of benthic archaeal cosmopolitan, Thermoprofundales (MBG-D archaea). *The ISME Journal* **13**, 885-901, doi:10.1038/s41396-018-0321-8 (2019).
- 220 Babski, J. *et al.* Genome-wide identification of transcriptional start sites in the haloarchaeon *Haloferax volcanii* based on differential RNA-Seq (dRNA-Seq). *BMC Genomics* **17**, 629, doi:10.1186/s12864-016-2920-y (2016).
- 221 Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. & Conesa, A. Differential expression in RNA-seq: A matter of depth. *Genome Research* **21**, 2213-2223 (2011).
- 222 Baccarella, A., Williams, C. R., Parrish, J. Z. & Kim, C. C. Empirical assessment of the impact of sample number and read depth on RNA-Seq analysis workflow performance. *BMC Bioinformatics* **19**, 423, doi:10.1186/s12859-018-2445-2 (2018).
- 223 Liu, Y., Zhou, J. & White, K. P. RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics* **30**, 301-304, doi:10.1093/bioinformatics/btt688 (2014).
- 224 Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Research* **46**, D794-D801, doi:10.1093/nar/gkx1081 (2018).
- 225 Allers, T., Barak, S., Liddell, S., Wardell, K. & Mevarech, M. Improved strains and plasmid vectors for conditional overexpression of His-tagged proteins in *Haloferax volcanii*. *Appl Environ Microbiol* **76**, 1759-1769, doi:10.1128/AEM.02670-09 (2010).
- 226 Pfeiffer, F. *et al.* Genome information management and integrated data analysis with HaloLex. *Arch Microbiol* **190**, 281-299, doi:10.1007/s00203-008-0389-z (2008).

- 227 Pfeiffer, F. & Dyall-Smith, M. Open Issues for Protein Function Assignment in *Haloferax volcanii* and Other Halophilic Archaea. *Genes (Basel)* **12**, doi:10.3390/genes12070963 (2021).
- 228 Culviner, P. H., Guegler, C. K. & Laub, M. T. A Simple, Cost-Effective, and Robust Method for rRNA Depletion in RNA-Sequencing Studies. *mBio* **11**, e00010-00020, doi:10.1128/mBio.00010-20 (2020).
- 229 Schwarzer, S., Rodriguez-Franco, M., Oksanen, H. M. & Quax, T. E. F. Growth Phase Dependent Cell Shape of Haloarcula. *Microorganisms* **9**, doi:10.3390/microorganisms9020231 (2021).
- 230 Busby, M. A., Stewart, C., Miller, C. A., Grzeda, K. R. & Marth, G. T. Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. *Bioinformatics* **29**, 656-657, doi:10.1093/bioinformatics/btt015 (2013).
- 231 Paya, G. *et al.* Small RNAs of *Haloferax mediterranei*: Identification and Potential Involvement in Nitrogen Metabolism. *Genes (Basel)* **9**, doi:10.3390/genes9020083 (2018).
- 232 Harrington, C. A. *et al.* RNA-Seq of human whole blood: Evaluation of globin RNA depletion on Ribo-Zero library method. *Scientific Reports* **10**, 6271, doi:10.1038/s41598-020-62801-6 (2020).
- 233 Adiconis, X. *et al.* Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nature methods* **10**, 623-629, doi:10.1038/nmeth.2483 (2013).
- 234 Pang, X. *et al.* Bacterial mRNA Purification by Magnetic Capture-Hybridization Method. *Microbiology and Immunology* **48**, 91-96, doi:<https://doi.org/10.1111/j.1348-0421.2004.tb03493.x> (2004).
- 235 Su, C. & Sordillo, L. M. A simple method to enrich mRNA from total prokaryotic RNA. *Molecular Biotechnology* **10**, 83-85, doi:10.1007/BF02745865 (1998).
- 236 He, S. *et al.* Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nature Methods* **7**, 807-812, doi:10.1038/nmeth.1507 (2010).
- 237 Giannoukos, G. *et al.* Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biology* **13**, r23, doi:10.1186/gb-2012-13-3-r23 (2012).

- 238 Farag, I. F., Zhao, R. & Biddle, J. F. "Sifarchaeota," a Novel Asgard Phylum from Costa Rican Sediment Capable of Polysaccharide Degradation and Anaerobic Methyloctrophy. *Appl Environ Microbiol* **87**, doi:10.1128/AEM.02584-20 (2021).
- 239 Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353-358, doi:10.1038/nature21031 (2017).
- 240 Meury, J. & Kohiyama, M. ATP is required for K<sup>+</sup> active transport in the archaebacterium *Haloferax volcanii*. *Archives of Microbiology* **151**, 530-536, doi:10.1007/BF00454870 (1989).
- 241 Villain, P. *et al.* The hyperthermophilic archaeon *Thermococcus kodakarensis* is resistant to pervasive negative supercoiling activity of DNA gyrase. *Nucleic Acids Research* **49**, 12332-12347, doi:10.1093/nar/gkab869 (2021).
- 242 Schumacher Maria, A., Choi Kang, Y., Zalkin, H. & Brennan Richard, G. Crystal Structure of LacI Member, PurR, Bound to DNA: Minor Groove Binding by  $\alpha$  Helices. *Science* **266**, 763-770, doi:10.1126/science.7973627 (1994).
- 243 Dorman, C. J. & Dorman, M. J. Control of virulence gene transcription by indirect readout in *Vibrio cholerae* and *Salmonella enterica* serovar Typhimurium. *Environmental Microbiology* **19**, 3834-3845, doi:<https://doi.org/10.1111/1462-2920.13838> (2017).
- 244 Bokal, A. J. t., Ross W Fau - Gourse, R. L. & Gourse, R. L. The transcriptional activator protein FIS: DNA interactions and cooperative interactions with RNA polymerase at the *Escherichia coli* *rrnB* P1 promoter. (1995).
- 245 Brinkman, A. B., Ettema, T. J. G., De Vos, W. M. & Van Der Oost, J. The Lrp family of transcriptional regulators. *Molecular Microbiology* **48**, 287-294, doi:<https://doi.org/10.1046/j.1365-2958.2003.03442.x> (2003).
- 246 Wang, Q. & Calvo, J. M. Lrp, a Global Regulatory Protein of *Escherichia coli*, Binds Co-operatively to Multiple Sites and Activates Transcription of *ilvIH*. *Journal of Molecular Biology* **229**, 306-318, doi:<https://doi.org/10.1006/jmbi.1993.1036> (1993).
- 247 Humbard, M. A., Zhou, G. & Maupin-Furlow, J. A. The N-terminal penultimate residue of 20S proteasome  $\alpha$ 1 influences its N( $\alpha$ ) acetylation and protein levels as well as growth rate and stress responses of *Haloferax volcanii*. *Journal of bacteriology* **191**, 3794-3803, doi:10.1128/JB.00090-09 (2009).

## Biography

I was born in 1991 in India. I received my integrated M.Sc. in Chemistry from the Indian Institute of Technology, Bombay (IIT-B) in 2015, after which I joined Duke University's PhD program in Genetics and Genomics. This work has been supported by a Biology Grant-in-aid (2018), a Duke Graduate School Travel Fellowship (2019), and Biology One-semester fellowship (2021).

### Publications:

1. Sakrikar S., Schmid AK. An Archaeal Histone-Like Protein Regulates Gene Expression in Response to Salt Stress. *Nucl Acids Res.* **2021**; 49(22).
2. Diveshkumar KV, Sakrikar S, Rosu F, Harikrishna S, Gabelica V, Pradeepkumar PI. Specific Stabilization of c-MYC and c-KIT G-Quadruplex DNA Structures by Indolylmethyleneindanone Scaffolds. *Biochemistry* **2016**; 55(25).
3. Diveshkumar KV, Sakrikar S, Harikrishna S, Damodharan V, Pradeepkumar PI. Targeting Promoter G-Quadruplex DNAs by Indenopyrimidine-Based Ligands. *ChemMedChem* **2014**; 9(12)