Title
XATAC-seq: Genome-wide Protein Occupancy Assay

Author
Chapin, Nate

Publication Date
2017

UNIVERSITY OF CALIFORNIA, SAN DIEGO

XATAC-seq: Genome-wide Protein Occupancy Assay

A Thesis submitted in partial satisfaction of the requirements for the degree Master of Science

in

Bioengineering

by

Nathaniel Stephen Chapin

Committee in charge:

Professor Karsten Zengler, Chair
Professor Christian Metallo, Co-Chair
Professor Xiaohua Huang

2017

The Thesis of Nathaniel Stephen Chapin is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California, San Diego

2017

DEDICATION

I would like to dedicate this thesis to my parents for all their love and support throughout my education.

TABLE OF CONTENTS

Signature Page iii

Dedication iv

Table of Contents v

List of Abbreviations vii

List of Figures viii

List of Tables ix

Acknowledgments x

Abstract of the Thesis xi

Introduction 1

Chapter 1: DNA-binding Proteins 4

1.1 Transcription Factors 4

1.2 Nucleoid-Associated Proteins 5

1.2.1 HU 6

1.2.2 IHF 6

1.2.3 Fis 7

1.2.4 H-NS 7

Chapter 2: XATAC-seq 11

2.1 Validation 14

2.1.1 XATAC-seq Captures Transcription Factor Binding 15

2.1.2 XATAC-seq Captures Nucleoid-Associated Protein Binding 16

2.1.3 XATAC-seq Recapitlates IPOD Results 17

2.1.4 Tagmentation Bias Assessment 19

2.1.5 Reverse-Crosslinking Evaluation 21

Chapter 3: XATAC-seq and H-NS-like Nucleoid-Associated Proteins 23

3.1: Rok of *Bacillus subtilis* 24

3.2: Mga of Group A *Streptococcus pyogenes* 26

3.2.1: GAS XATAC-seq and AT Content 27

3.2.2: GAS XATAC-seq and Gene Expression 28

3.2.3: Mass Spectroscopy Analysis 33

Appendix 36

A1: Bacterial Strains and Growth Conditions 36

A2: Experimental Methods 36

A3: Data Analysis 39

References 40

LIST OF ABBREVIATIONS

ATAC: Assay for Transposase-Accessible Chromatin

ChIP: Chromatin Immunoprecipitation

NAP: Nucleoid-associated protein

HU: Heat Unstable Protein

IHF: Integration Host Factor

Fis: Factor for Inversion Stimulation

H-NS: Histone-like Nucleoid-Structuring Protein

POL: Protein Occupancy Landscape

LIST OF FIGURES

Figure 1: Modes of H-NS Transcription Mediation 9

Figure 2: XATAC-seq Method 11

Figure 3: Tagmentation 13

Figure 4: Replicate Correlations 14

Figure 5: Quantitative Footprint Evaluation 15

Figure 6: XATAC-seq Footprinting Captures Transcription Factor Binding Sites 16

Figure 7: Promoter Enrichment 16

Figure 8: Recapitulation of EPODS 18

Figure 9: Timeline of XATAC-seq and IPOD 19

Figure 10: DNA-only ATAC Control 20

Figure 11: Effects of Tagmentation Time 21

Figure 12: Effects of Reverse Crosslinking 22

Figure 13: H-NS Correlation 23

Figure 14: Rok of *B. subtilis* 24

Figure 15: Effects of AT Content on Binding in Untreated and

Rifampicin-Treated Cells 28

Figure 16: Virulence Gene Binding 29

Figure 17: RNA-seq Differential Expression Venn Diagram 30

Figure 18: GAS POL Comparison 30

Figure 19: Differential Expression vs. Differential Binding 32

Figure 20: Gene binding vs. Expression 33

Figure 21. Mass Spectroscopy Probe Design 34

LIST OF TABLES

Table 1: Correlation of XATAC-seq to ChIP-seq of Various NAPs. 17

Table 2: Homology of H-NS-family Proteins 25

Table 3: GAS Virulence Genes 26

Table 4: GAS RNA-seq 31

Table 5: Mass Spectroscopy Results 34

ACKNOWLEDGMENTS

I would like to thank Dr. Karsten Zengler for his support as an advisor and as the chair of my committee. His encouragement and guidance have proved tremendously valuable to me.

I would like to acknowledge the work and wisdom of Dr. Mahmoud Al-Bassam, with whom I have worked on material that appears in this manuscript and that is at the present time being prepared for publication.

I would also like to thank Nina Gao for her contribution on aspects of this work related to *Streptococcus pyogenes*, both experimental work and discussion.

Chapter 2, in part, is currently being prepared for submission for publication of the material. Al-Bassam, Mahmoud; Chapin, Nate; Gao, Nina; Zengler, Karsten; Nizet, Victor. The thesis author is co-first author of this paper.

Chapter 3, in part, is currently being prepared for submission for publication of the material. Al-Bassam, Mahmoud; Chapin, Nate; Gao, Nina; Zengler, Karsten; Nizet, Victor. The thesis author is co-first author of this paper.

ABSTRACT OF THE THESIS

XATAC-seq: Genome-wide Protein Occupancy Assay

by

Nathaniel Stephen Chapin

Master of Science in Bioengineering

University of California, San Diego, 2017

Professor Karsten Zengler, Chair

The binding of protein to DNA is central to the regulation of gene expression and the organization of chromosomal DNA. To date, there exist few techniques for the determination of genome-wide protein binding in prokaryotes, and none that are simultaneously simple, high-resolution, and rapid. I describe XATAC-seq, an adaptation of the eukaryotic assay for transposase-accessible chromatin with sequencing (ATAC-seq), combining formaldehyde crosslinking of DNA-protein complexes, adapter-loaded transposase treatment for next-generation sequencing library generation, and high-throughput sequencing to interrogate these genome-wide binding patterns in bacteria. The technique captures the binding of both major classes of prokaryotic DNA-binding proteins–transcription factors and nucleoid-associated proteins–genome-wide at the resolution of individual binding sites. XATAC-seq was applied to determine the protein occupancy landscapes of several bacterial species. Remarkably, the landscapes show a high degree of fidelity to specific nucleoid-associated proteins and demonstrate several conserved characteristics, including extended domains of high enrichment and preferential enrichment of AT-rich regions. This has led to the speculation that these nucleoid-associated proteins are members of a common high-level functional group, and that this class of nucleoid-associated protein is prevalent among a significantly wider range of prokaryotes than previously realized. In particular, the Mga protein of *Streptococcus pyogenes* is proposed to serve the high-level function of suppression of ectopic expression in an analogous fashion to the H-NS protein in *E. coli*. This work represents the first assessment of protein occupancy landscapes in gram-positive bacteria and a significant technical improvement over existing techniques for assaying genome-wide protein binding in prokaryotes.

INTRODUCTION

The binding of proteins to DNA is central to the cellular processes of DNA replication and repair, gene expression and its regulation, and DNA compaction and structural organization. The majority of research on protein-DNA interactions has focused on the binding behavior and binding loci of individual proteins. Understanding systems-level behaviors, such as genome replication, chromosomal organization, and regulatory network dynamics, requires observations encompassing the entire system.

DNA must be compacted many fold to fit inside the volume of a cell. For example, the DNA of *E. coli*, if fully extended, would reach a length of 1 mm[1]. Its collapse to fit within a 2 μm-long cell requires compaction of 3 orders of magnitude[2]. Negative supercoiling provides part of the answer, causing DNA to take on an interwound, plectonemic conformation, with branches extending outward from a central hub. Supercoiling provides only a partial solution, however. In eukaryotes, the majority of compaction is accomplished through the action of histones, which wrap DNA and organize it into nucleosomes[3]. Higher-order packaging of nucleosomes provides further reduction in size. Prokaryotes, by contrast, lack histones and accomplish DNA compaction through the action of a class of DNA-binding protein called nucleoid-associated proteins (NAPs), many of which bend or wrap DNA[4].

Nucleoid-associated proteins have a strong impact on the overall determination of chromosome architecture in prokaryotes. Specifically, some NAPs are capable of forming boundary elements between chromosomal domains[5]. The organization of chromosomal DNA into domains affects the way genetic information is accessed, interpreted, and implamented. For example, it prevents the spreading or propagation of looping or relaxation of one genome segment into the entire genome, allowing regional differences. In addition, domain organization can co-localize or spatially segregate transcription factors and their target genes, potentially making these domains adjustable functional units of gene regulation[6]. *E. coli* has around 450 nucleic structural domains, estimated to be between 10kb and 100kb, with variable boundaries, distributed sporadically along the chromosome[7,8]. Highly transcribed genes appear to be involved in defining domain boundaries by spatially isolating DNA regions and restricting the diffusion of supercoiling[9].

To date, nucleoid-associated protein binding and its effects on global gene expression and chromatin conformation remain understudied, largely due to technical limitations. There currently exist few techniques for comprehensive identification and assessment of NAP binding.

Hi-C and similar techniques are powerful tools for determination of three-dimensional chromosome structure, but despite breakthroughs in the elucidation of the chromatin structures of *Caulobacter crescentus*[10] and *Bacillus subtilis*[11], such investigations remain effortful. In particular, Hi-C is limited by its technical and bioinformatic intricacy as well as the requirement for highly synchronous cell cultures. As such, the study of bacterial chromatin remains challenging due to the difficulty of synchronizing most bacterial species, and Hi-C has not been widely adopted for the study of prokaryotic nucleoids.

In vivo protein occupancy display has been shown to be capable of detecting individual protein binding sites, as well as large-scale regions of enrichment[12]. However, the technique relies on several loss-prone reaction steps and a low-resolution method of sequence information extraction. This, in addition to the requirement for microarray design and post-processing of data, appears to have precluded its adoption, given that no studies have been reported beyond the original.

Herein, I describe XATAC-seq, a modification of the assay for transposase-accessible chromatin using sequencing (ATAC-seq) originally designed to interrogate nucleosome-free regions of eukaryotic chromosomes[13]. ATAC-seq takes advantage of a hyperactive mutant of the bacterial Tn5 transposase (Tnp)[14]. Rather than a single transposable element (transposon), the transposase dimer is loaded *in vitro* with a pair of double-stranded sequencing adaptors. As such, the transposition event results in simultaneous fragmentation and tagging of DNA segments for later amplification and sequencing. Because protein binding to DNA sterically hinders the transposase, the probability of transposition events is heavily weighted towards open, unbound regions of chromatin. The result is a distribution of fragment sizes depending on the region from which the DNA originates. Specifically, short fragments are associated with protein-free regions, whereas longer fragments are associated with regions bound by individual proteins or higher-order nucleoprotein complexes.

CHAPTER 1: DNA-BINDING PROTEINS

In all domains of life, classes of proteins called transcription factors bind to regulatory regions of DNA and modulate the expression of target genes. In general, these factors target a specific DNA-sequence (consensus sequence) and regulate a small number of genes, although there are global regulators for which this is not the case. In bacteria, there is an additional class of DNA-binding proteins called nucleoid-associated proteins, which typically bind with little or no sequence specificity, targeting features of the DNA structure rather than a particular base sequence[15].

1.1 TRANSCRIPTION FACTORS

Transcription factors are proteins that regulate gene expression by mediating transcription initiation through binding at specific, high-affinity *cis* regulatory elements in the vicinity of their target genes. This binding activity may be either activatory or inhibitory to gene expression, depending on the binding protein and the target site. Generally speaking, repressors bind directly to the promoter sequence, interfering with RNA polymerase binding. In contrast, activators generally bind upstream of the promoter and facilitate the recruitment of sigma factors[16].

There are several DNA-binding motifs that are well-conserved among transcription factors; in bacteria, the most common is the helix-turn-helix[17]. In addition, many transcription factors have domains responsible for signal-sensing, through ligand binding or protein-protein interactions[18]. The majority of prokaryotic transcription factors contain all required functional domains, but a major exception is two-component systems, in which the signal-sensing and DNA-binding/transcriptional regulation roles are accomplished by separate protein partners[19]. Typically, transcription factors bind DNA as homo- or heterodimers, which is reflected in the fact that many consensus sequences contain palindromic or direct repeats[20]. In addition, they typically interact with the major groove of DNA, along which, in contrast to the minor groove, the pattern of hydrogen bond donors and acceptors and hydrophobic regions differs significantly depending on the base-pair[21].

1.2 NUCLEOID-ASSOCIATED PROTEINS

Nucleoid-associated proteins (NAPs) are important regulators of gene expression and chromatin structure in bacterial cells. Even the most reduced of bacterial genomes encode at least one NAP, and many contain a variety[15,22,23]. In general, NAPs bind with low sequence-specificity throughout the genome, making their binding more widespread and less focused than that of transcription factors. Some NAPs have been shown to be contained to particular chromatin macrodomains, though these appear to be specialized cases[24]. Many have been shown to exhibit a preference for DNA with particular structural features rather than base compositions, and nearly all NAP binding impacts DNA structure significantly. In addition to their effect on chromatin structure, NAPs have been shown to act as global regulators, mediating changes between growth phase or responses to particular environmental stressors by regulating the expression of large numbers of genes. The major nucleoid-associated proteins of *E. coli* will now be discussed, as they are the best studied.

1.2.1 HU

Heat unstable protein (HU) is the most highly conserved of bacterial NAPs, and is the bacterial protein with the most sequence homology to eukaryotic histones. It exists as both a homodimer and heterodimer in *E. coli*, depending on the growth phase, and the cell is able to tune the properties of its HU dimers by altering the relative concentration of the monomers produced[25]. HU lacks any strong sequence specificity, targeting bent DNA segments, and is able to wrap DNA upon binding. HU binds single- and double-stranded DNA, as well as RNA[26]. In the case of double-stranded DNA, HU proteins engage the double helix at a convex surface, with multiple exposed cationic side-chains. This surface provides electrostatic and steric complementarity for B DNA and has been confirmed as the nucleic acid binding site[27]. At low concentrations, HU increases DNA flexibility over short regions; at high concentrations, it increases DNA stiffness and rigidity[28]. HU binding has been shown to increase the thermal stability of double-stranded DNA[4]. In addition, HU interacts with topoisomerase I to regulate DNA superhelicity[29]. Finally, HU appears to play a role in initiating DNA replication[30].

1.2.2 IHF

Integration host factor (IHF) is one of the most abundant sequence-specific binding proteins in *E. coli*[15]. The structurally important amino acids are conserved between HU and IHF, and they share the same basic tertiary structure[4]. Both proteins bend DNA, but IHF does so to a greater degree than HU, inducing a ~160º U-turn conformation. IHF, like HU, is predominantly a heterodimer. Another point of similarity is that IHF also impacts DNA replication from the chromosomal origin. In gram-negative bacteria, bending of DNA by IHF is associated with transcriptional activation of many σ^{54} promoters by bringing enhancer-binding proteins into proximity with RNA polymerase[31]. IHF can also induce open complex formation by restricting superhelical twist at its binding site, transmitting this torsional energy to neighboring regions where it facilitates transcription initiation[32]. The primary role of IHF appears to involve remodeling of local DNA structure.

1.2.3 FIS

The factor for inversion stimulation (Fis) is the most abundant NAP during exponential growth in *E. coli*[15]. Fis binds to an AT-rich consensus sequence as a homodimer, and its binding induces branched plectonemes. One of its major functions appears to be inactivating inessential genes during rapid growth[22]. In addition, it appears to be necessary for the transcription of rRNA and tRNA genes[15]. Fis functions as an activator of transcription initiation by either direct interaction with RNA polymerase or alteration of local DNA topology in the promoter region in a DNA structural transmission mechanism similar to IHF[33]. Fis also interacts with both major topoisomerases and therefore indirectly affects FDNA superhelicity[34]. In addition, like HU and IHF, Fis plays a role in initiation of chromosomal replication[35].

1.2.4 H-NS

Histone-like nucleoid-structuring protein (H-NS), so-named because of its effect on bacterial chromatic rather than homology to eukaryotic histones, is as mall (~15kD), highly abundant (~20,000 copies/cell in *E. coli*) protein common to enteric bacteria, particularly *E. coli* and its close relatives[36]. Several families of proteins sharing functional homology with H-NS have been identified in gram-negative bacteria, including the *Mycobacteriaceae* and *Pseudomonadaceae* families, though their similarity at the sequence level is minimal[37–39]. H-NS-like proteins share nonspecific DNA-binding behavior, targeting to the minor groove of DNA, along which the differences between bases are less pronounced, and exhibiting a preference for AT-rich regions of DNA[40–43]. H-NS is a pleiotropic repressor, regulating approximately 5% of *E. coli* genes, with 80% of that regulation being repressive, including autorepression of the *hns* gene[44–47]. This autorepression has been shown to act as a mechanism to ensure that the ratio of H-NS to DNA remains relatively constant throughout growth phases, although there is some contradiction as to that point[36,48,49]. H-NS comprises a C-terminal DNA-binding domain, and N-terminal dimerization domain, and a central linker domain involved in higher-order oligomerization[43,50–53]. It has been shown to act as a silencer of horizontally-acquired DNA, which for enteric bacteria generally has higher AT-content than that of the host genome. H-NS binding to DNA occurs in two steps: binding initiates at high-affinity sites followed by oligomerization and expansion of the nucleoprotein filament to cover less well-suited binding sites and form a nucleoprotein structure conducive to silencing[54–56]. The fundamental units of such nucleoprotein structures are believed to be dimers, which combine in head-to-head and tail-to-tail fashion. It has been demonstrated that mutations to the oligomerization domain of H-NS disrupt its ability to silence expression, and the several models of H-NS silencing support this finding[57,58]. Briefly, H-NS oligomers can bind to and occlude promoter sequences from RNA polymerase (Fig 1A), H-NS bridge formation can loop DNA and trap RNA polymerase at the promoter site (Fig 1B), binding within genes can stall RNA polymerase and lead to Rho-dependent transcriptional termination (Fig 1C), seed binding may occur at distal regions to the promoter, with oligomerization ultimately bringing H-NS protein into direct contact with RNA polymerase (Fig 1D), and channeling of RNA polymerase toward promoter sites in AT-rich regions of ambiguity (Fig. 1E)[54,59–63]. Of these direct mechanisms of transcriptional regulation, only the last is activatory. Interaction with accessory proteins of the Hha/YdgT family, which lack DNA-binding activity of their own, has been shown to facilitate H-NS oligomerization and H-NS-mediated gene silencing[64,65].

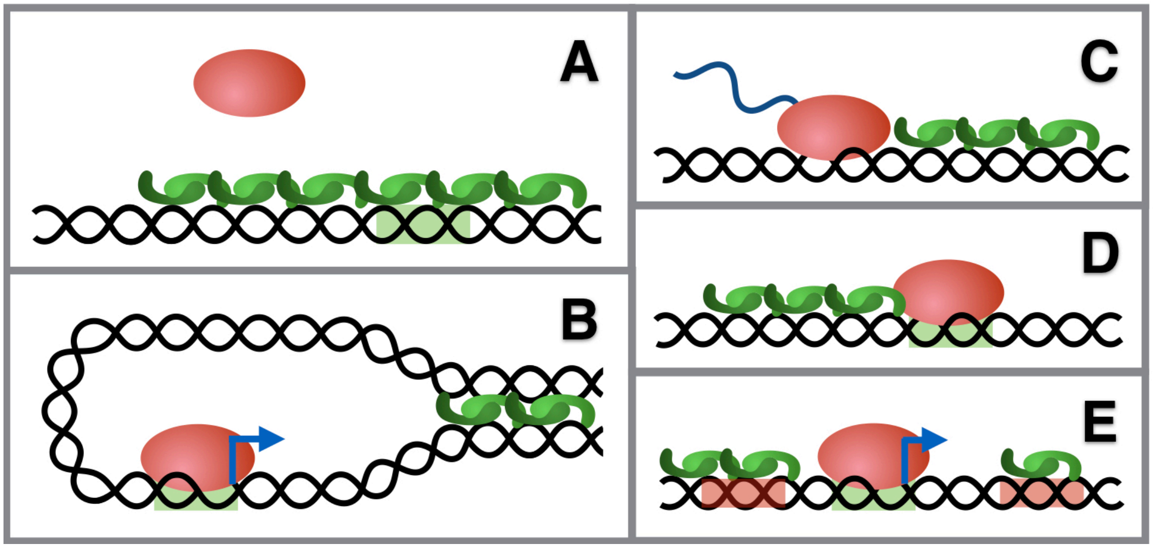


Figure 1: Modes of H-NS Transcription Mediation. H-NS oligomers shown in green; RNA polymerase (RNAp) in light red. Green and red regions of DNA are correct and incorrect promoter sites, respectively. A) Promoter exclusion. B) RNAp trapping. C) Transcription termination. D) Direct interaction with RNAp. E) Channeling of RNAp to canonical promoter.

H-NS acts as a xenogeneic silencer in *E. coli*, repressing horizontally-acquired genes until they can become properly integrated into the regulatory network of the cell[41,66–70]. Additionally, it has been implicated with widespread repression of intragenic transcription, thereby preventing spurious RNA synthesis[71]. In fact, nearly half of all transcripts (46%) repressed by H-NS in *E. coli* originate in intragenic regions, and a significant portion of those emanating from intergenic regions are non-coding RNAs[71]. A large part of the fitness cost associated with the loss of H-NS is due to this ability; when widespread intragenic transcription is allowed, the cell's supply of RNA polymerase is sequestered at these promoters, making it unavailable for the transcription of required genes[66]. These two functions in combination posit H-NS as an important regulator of transcription genome-wide and as integral to cellular fitness. In agreement with this, many bacterial species encode multiple H-NS molecules, allowing them to modulate their response to environment al signals by adjusting the pool of H-NS-like dimers[72].

Other H-NS-like protein families include the Lsr2 family in *Mycobacteria* and the mvaT family in *Pseudomonas.*[73,74]. All share similar binding preference and the ability for oligomerization, although as stated previously their homology at the amino–acid level is low.

CHAPTER 2: XATAC-SEQ

XATAC-seq is an adaptation of the assay for transposase-accessible chromatin (ATAC-seq) originally designed to interrogate nucleosome-free regions of DNA in eukaryotes[13]. It relies on formaldehyde treatment to crosslink DNA to protein and subsequent treatment of cell lysate with a hyperactive Tn5 transposase to simultaneously fragment DNA and ligate adapters in a process termed tagmentation[14]. The resulting fragments are PCR-enriched without explicit reverse-crosslinking and sequenced.

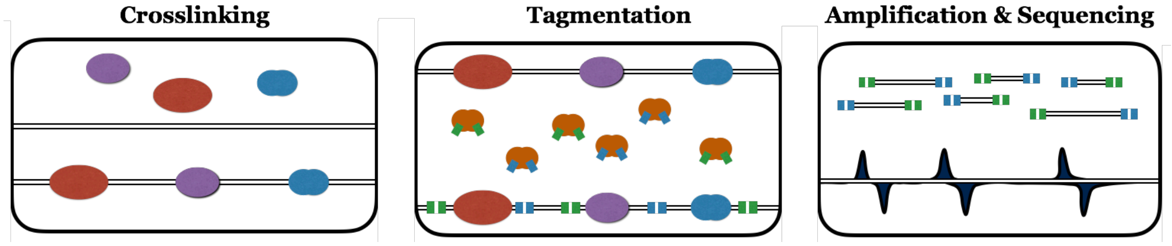


Figure 2: XATAC-seq Method.

The major methodological difference between XATAC-seq and ATAC-seq is treatment with formaldehyde, which forms a methylene bridge between DNA and protein and is used to ensure that protein-DNA complexes are not disrupted by the chemical steps they undergo through the course of the procedure. The amino acids that undergo cross linking are cysteine, tryptophan, lysine, and histidine with dA, dC , or dG, with the most prominent reaction being between lysine and dG[75]. Lysine is extremely common in DNA-binding proteins because it facilitates interactions with the phosphate backbone, but formaldehyde crosslinking efficiency can still vary significantly between proteins[76,77]. The advantages of formaldehyde as a crosslinking reagent include cell permeability, fast crosslinking kinetics, short crosslink length, and controlled reversibility[76]. In addition, because crosslinking occurs very rapidly, crosslinked complexes are faithful to the protein-DNA interactions occurring in live cells[78].

The transposition step involves transposase binding at the target site, a transposase-mediated nucleophilic attack on the phophodiester bonds along the backbone of both DNA strands, and transposase release, followed by nick repair (Fig. 3)[79]. As a result of the final step, 9 base pairs are duplicated on either side of the inserted adapters, which becomes important in downstream data analysis (Fig. 3 and Appendix A3)[14]. The Tn5 transposase has an insertion preference (A-G-N-T-T/C-A/T-A/G-A-N-T/C) that is mirrored to a small degree in the bias of transposition events[79–81]. However, the average information content within 10 bases of the tagmentation site, on a two-bit scale, is 0.049, compared to 0.0056 and 0.018 for sonication and endonuclease treatment, respectively[82]. Therefore, the bias associated with transposase-mediated library construction is higher than that generated by other procedures, but only to a small degree.

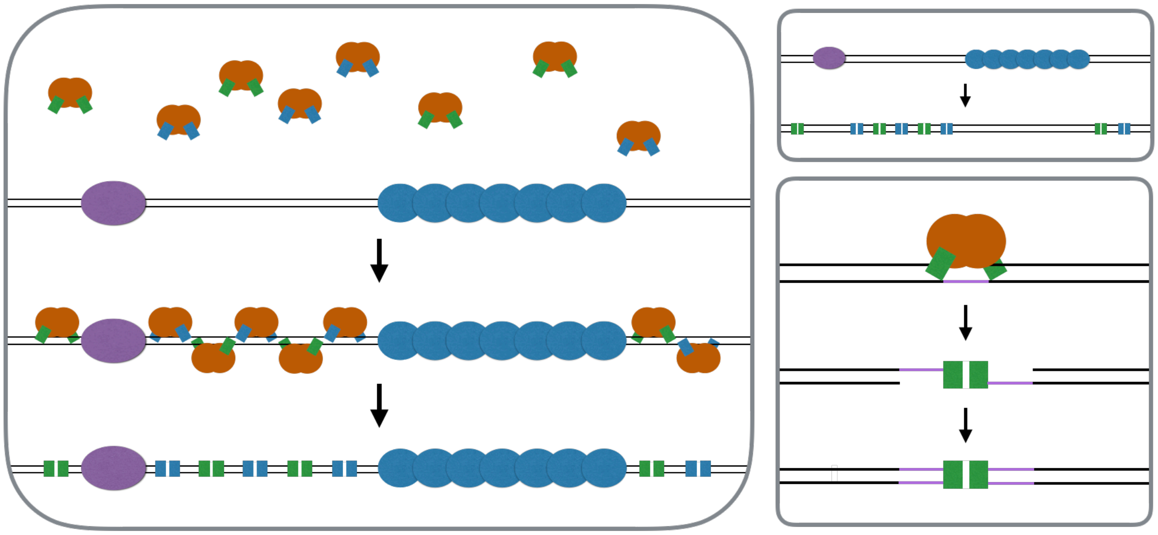


Figure 3: Tagmentation. Left: model of tagmentation reaction in which transposases saturate available DNA, ultimately limiting the minimum fragment size to ~38 bp due to steric hindrance between attacking transposases. Top right: fragment length is indicative of the state of binding in the region of origin of that fragment. Bottom right: the transposase's active site interacts with 9 bases of DNA, ultimately causing their duplication. Orange – transposase; purple – individual bound protein; blue – oligomerized bound protein.

The technique is highly reproducible, with replicates demonstrating Pearson correlation coefficients of 0.91 on average (Fig. 4). This consistency strongly suggests that the interaction of total protein with DNA, not just those with high sequence-specificity, is very precise and well-regulated.

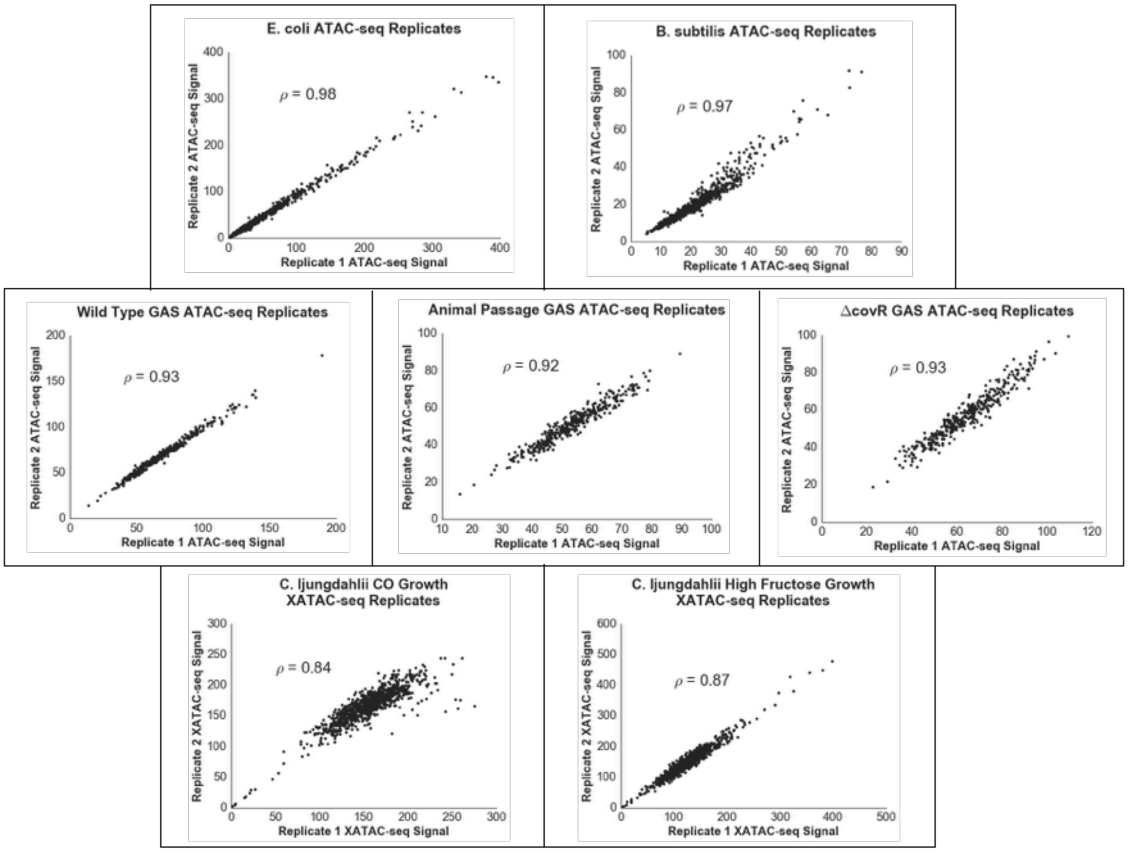


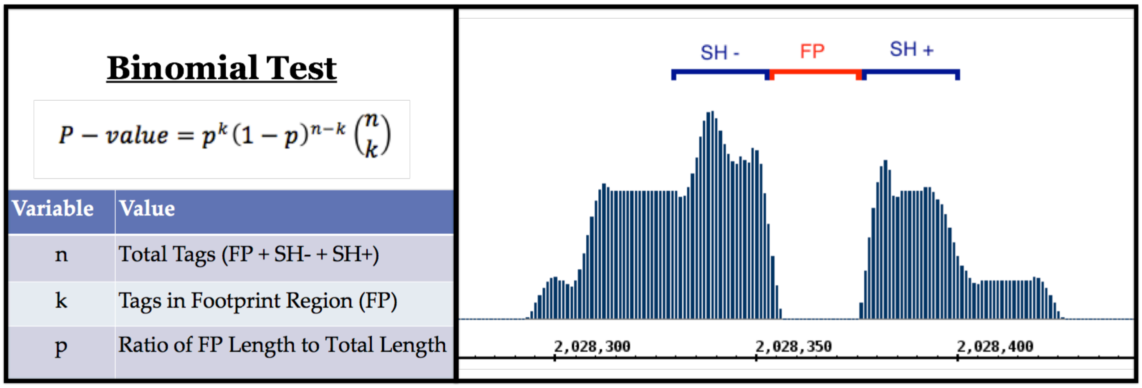
Figure 4: Replicate Correlations. Average XATAC-seq signal for sets of replicates is plotted over 5kb bins; correlations shown are for un-binned data. Note: this is not tagmentaion sites, but full read signal.

2.1 VALIDATION

In order to evaluate the ability of XATAC-seq to capture protein binding events, we have evaluated its ability to capture both transcription factor binding sites and nucleoid-associated protein binding. As further validation, we have compared our technique to IPOD, the only existing technique for genome-wide protein occupancy determination in bacteria. Finally, we have performed tests to ensure that XATAC-seq signal is not significantly impacted by tagmentation bias or inefficacy of reverse-crosslinking.

2.1.1 XATAC-SEQ CAPTURES TRANSCRIPTION FACTOR BINDING

Similar to the results obtained using the original ATAC-seq protocol, gaps in tagmentation are expected wherever protein is bound along the genome (greater than that between adjacent transposases, see Fig. 3). Therefore, the exact sites of binding can be accurately determined by evaluating the site of transposition events. Specifically, a binomial test is used to determine the significance of a potential footprint motif. The test compares the XATAC-seq signal immediately upstream and downstream of the putative binding site with the signal within to determine the degree of non-uniformity (Fig. 5). The test iterates through all possible footprint start positions and distances between a peak pair in order to determine which is most likely to represent the exact binding site.



Binomial Test

$$P - value = p^k (1-p)^{n-k} \binom{n}{k}$$

Variable	Value
n	Total Tags (FP + SH- + SH+)
k	Tags in Footprint Region (FP)
p	Ratio of FP Length to Total Length

Figure 5: Quantitative Footprint Evaluation. Footprints, indicative of protein binding, are evaluated using a binomial test in order to assign to each a degree of confidence. The test compares total signal within the putative footprint region (FP) with that in shoulder regions immediately upstream and downstream (SH).

When a p-value threshold of 1e-10 is imposed on the footprints from *E. coli* XATAC-seq, 13% of footprints align with transcription factor binding sites compiled in the model organism database EcoCyc (Fig 6)[83]. An additional 20% align with sigma factor binding sites. In addition, as shown in Figure 7, there is significant enrichment of tagmentation in non-coding regions relative to coding regions, indicating that these alignments are not coincidental.

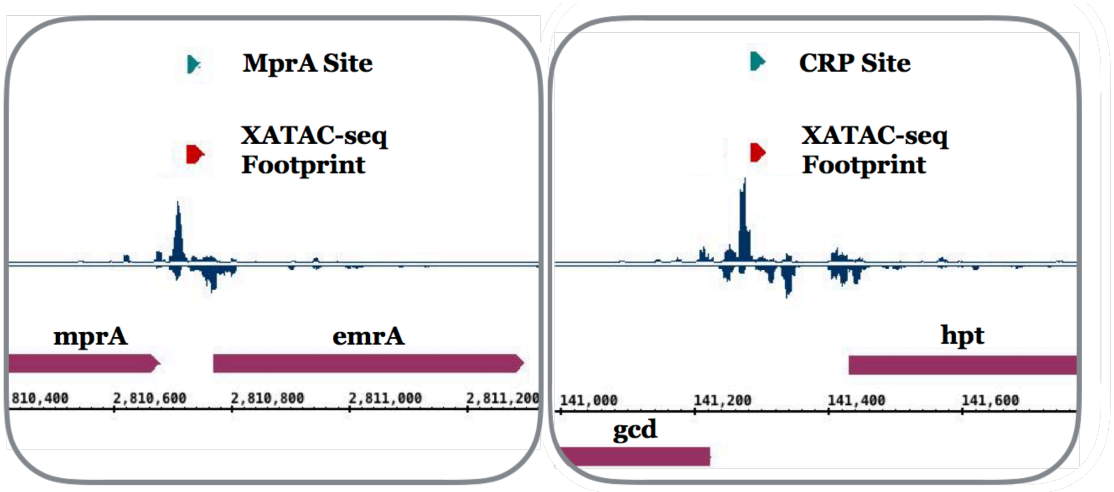


Figure 6: XATAC-seq Footprinting Captures Transcription Factor Binding Sites. XATAC-seq signal and the footprints resulting from assessment of this signal are shown in comparison to transcription factor binding sites compiled from the literature (EcoCyc).

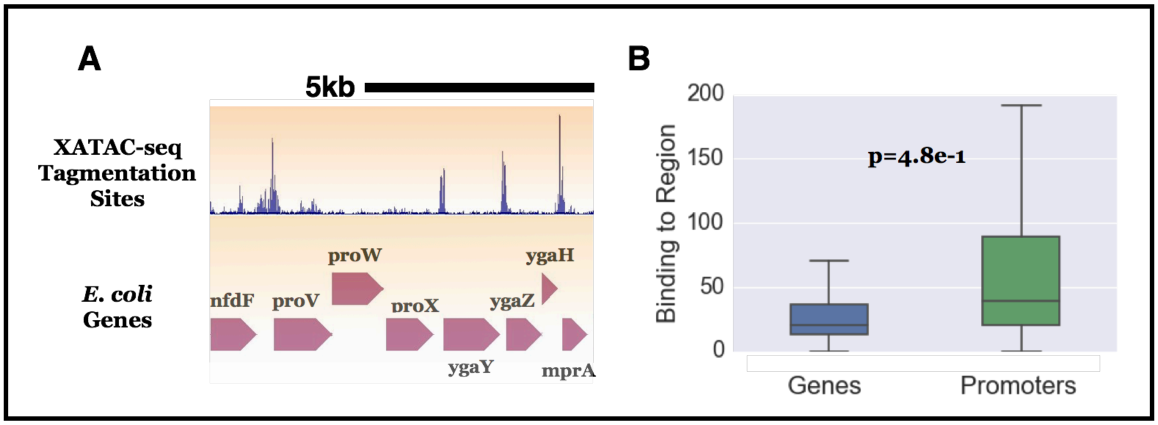


Figure 7: Promoter Enrichment. A. View of XATAC-seq tagmentation sites. B. Average XATAC-seq signal per unit length of genes and intergenic regions in *E. coli*. The p-value was calculated using the Mann-Whitney U test.

2.1.2 XATAC-SEQ CAPTUES NUCLEOID-ASSOCIATED PROTEIN BINDING

In order to assess the degree to which nucleoid-associated proteins impact XATAC-seq signal, we compared our data to ChIP-seq datasets from the literature for the predominant NAPs in *E. coli* – HU, IHF, Fis and H-NS[84,85]. ChIP-seq data was used because it was the highest resolution available. As can be seen in Table 1, XATAC-seq is correlated with the binding of both HU and H-NS, with a significantly stronger correlation to H-NS.

Table 1: Correlation of XATAC-seq to ChIP-seq of Various NAPs. HU and IHF ChIP-seq experiments were performed by Prieto *et al.* and H-NS and Fis ChIP-seq experiments were performed by Kahramanoglou *et al.*[84,85]

Nucleoid-Associated Protein	Pearson Correlation Coefficient
HU	0.21
IHF	-0.10
FIS	-0.07
H-NS	0.60

2.1.3 XATAC-SEQ RECAPITULATES IPOD RESULTS

Chromatin immunoprecipitation with high-throughput sequencing (ChIP-seq) provides comprehensive binding information for a single factor under a given set of conditions, but fails to provide specificity as to exact binding loci. ChIP-exo expands on the original ChIP methodology by adding double and single-strand-specific exonuclease digestions that digest DNA up to the binding site and thereby allow the technique to provide binding information to near single-base resolution [86]. The drawback to this method remains that it is capable of surveying only a single binding protein at a time, and therefore that a comprehensive understanding of the protein occupancy landscape (POL) is considerably challenging to assemble. For example, in *E. coli* there are 271 identified transcription factors, of which any number may be active under any given set of conditions, making such analysis by ChIP-based techniques unfeasible [17].

One technique that addresses this limitation is *in vivo* protein occupancy display (IPOD) – a genome-wide assay for protein occupancy that relies on formaldehyde crosslinking, DNase I treatment, phenol:chloroform isolation of protein-DNA complexes, reverse-crosslinking, and detection by array hybridization in order to identify regions of protein binding[12]. The detection of sequences by array hybridization results in lower resolution than next generation sequencing.

XATAC-seq is able to re-capture the same regions of enrichment identified in IPOD (Fig. 8). It is apparent from the figure that XATAC-seq is significantly higher resolution that IPOD. In addition, IPOD is considerably more involved that XATAC-seq, requiring 7 major chemical steps and an estimated 9 hours to perform, from cell pellets to array-ready DNA (Fig. 9), compared to 3 hours and 2 reaction steps for XATAC library preparation from cell pellets, plus additional time for array hybridization and scanning and next-generation sequencing, respectively[12].

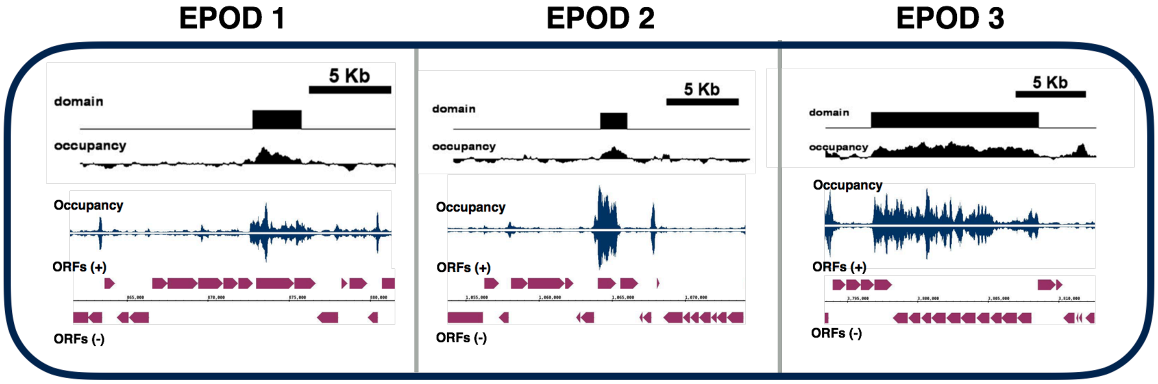


Figure 8: Recapitulation of EPODS. Enriched protein occupancy domains identified by IPOD are shown with their associated IPOD signal and the XATAC-seq signal in the same region.

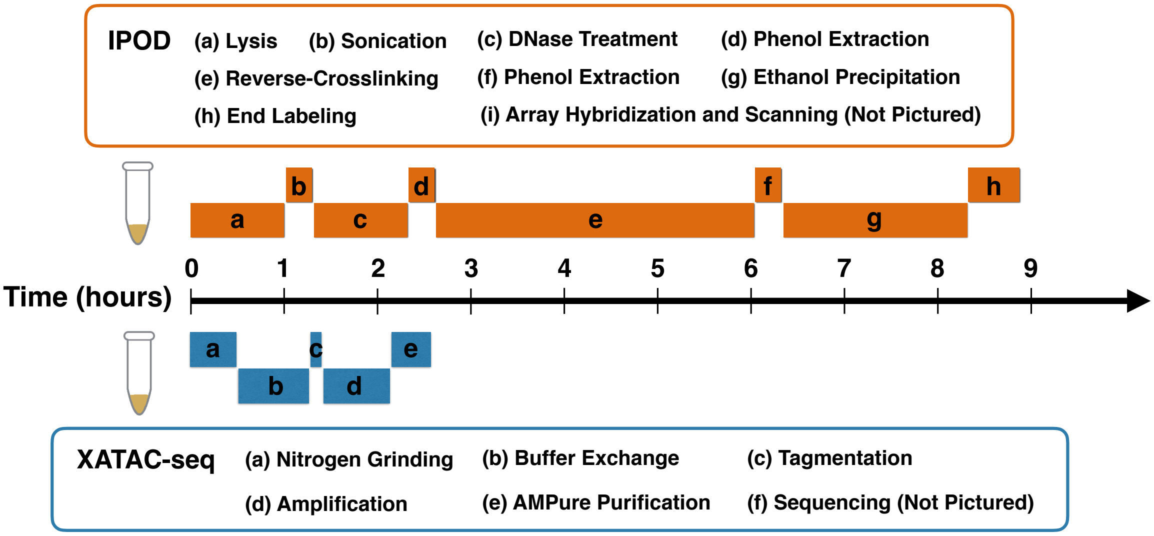


Figure 9: Timeline of XATAC-seq and IPOD, beginning with formaldehyde-crosslinked cell pellets, the preparation of which is common to both techniques. Not pictured are sequence detection methods for each technique, which take approximately 16 hours for microarrays and a variable amount of time for next-generation sequencing, which is typically outsourced.

Multiple experiments from different conditions or organisms may be multiplexed in a single sequencing run, whereas a separate microarray is required for each. Furthermore, an entirely new microarray must be designed for each new organism under study. Overall, we believe that XATAC-seq represents a significant methodological improvement over IPOD, the only existing technique for POL determination.

2.1.4 TAGMENTAION BIAS ASSMENT

Tagmentation bias was assessed using a pure-DNA control. Briefly, DNA was extracted as in the standard XATAC-seq protocol using nitrogen grinding, but without formaldehyde crosslinking to secure protein-DNA complexes. A phenol:chloroform extraction was then used to isolated pure DNA in the aqueous phase, and this DNA was used as input for the Illumina Nextera kit, as in the standard procedure. The results show a small bias in tagmentation toward GC-rich regions, so it can be concluded that the enrichment of XATAC-seq for AT (discussed in Chapter 3) is due to the preference of binding proteins and not that of the transposase (Fig. 10). For example, the correlation between the DNA-only control and XATAC-seq signal for wild-type GAS is only 0.17. It is worth noting that this control captures both tagmentation bias and differences in copy number of different regions of the genome, for example enrichment of origin-proximal regions caused by concurrent rounds of replication.

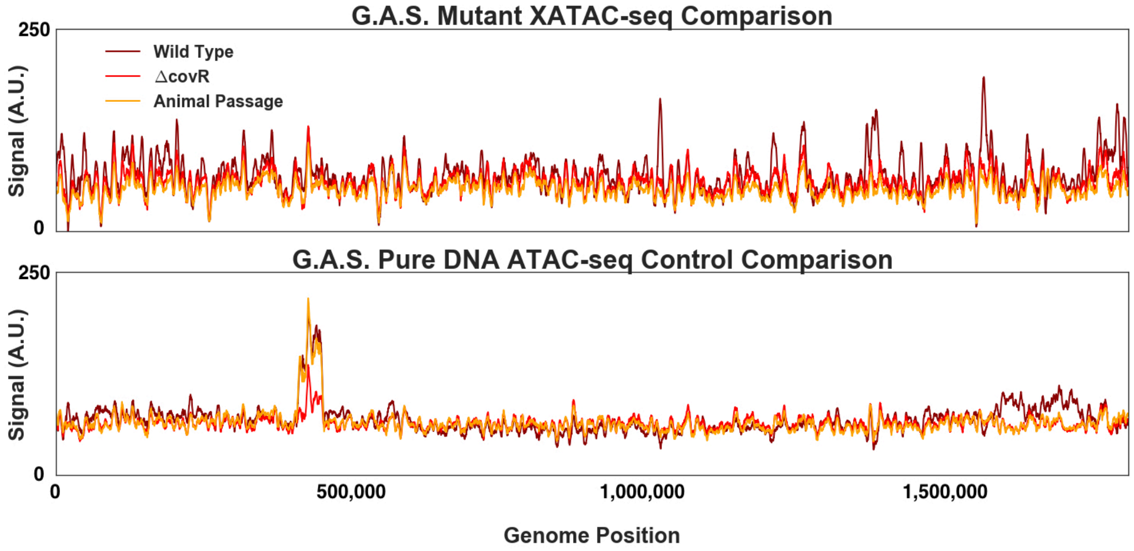


Figure 10: DNA-only ATAC Control. It is evident that signal intensity is greatly reduced without protein bound, and the signal is to a great degree more uniform across the genome. The peak near position 450,000 is the well-characterized ΦM1T1Z phage, which encodes the virulence factor streptodornase and differentiates the M1T1 serotype from other closely related M1 strains[87] (discussed in Chapter 3).

The effects of the length of incubation of the transposase reaction step on library size distribution was assessed by performing the incubation at lengths of 3, 5, 7, and 9 minutes. The results show that the resulting size distribution is not greatly impacted by the tagmentation time (Fig. 11), but it appears that there is significantly more noise with very short tagmentation times. However, this is confounded slightly by the fact that the library generated from 3-minute tagmentation was not sequenced as deeply as the others.

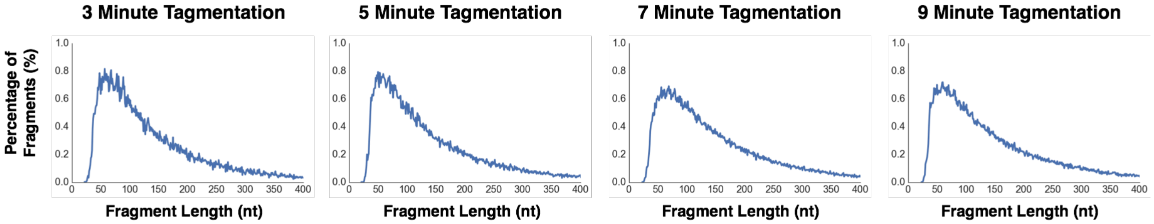


Figure 11: Effects of Tagmentation Time. The length of fragments constituting each library is plotted against the percentage of library fragments at that length.

2.1.5 REVERSE-CROSSLINKING EVALUATION

To determine whether fragments were being lost due to the inability of PCR to effectively reverse protein-DNA complexes and amplify DNA, a variation of XATAC-seq was performed with an explicit reverse-crosslinking step. The results showed that more long fragments were retained than in the standard procedure, but there is significant loss of short fragments (shift right in Fig. 12). These large fragments must have originated from extended regions of occupancy, preventing the transposase from digesting them into smaller pieces. Therefore, the number of crosslinks, and thus the difficulty of reverse-crosslinking, increases with the length of the fragment, and the results of this experiment imply that a significant number of nucleoprotein filaments are lost in the standard method because they cannot be effectively reverse-crosslinked. As such, an extension of the pre-PCR heating step may be advisable as a simple means of increasing crosslink reversal. According to a study of formaldehyde crosslink reversal rate, the percentage of crosslinks reversed over time fits an exponential function of temperature[88]. They also demonstrate that the reversal rate is constant over time at a given temperature, and so overall:

$$p/t = 0.00379e^{0.0631T}$$

Where T is the temperature in degrees Celsius, t the time in minutes, and p the percentage of crosslinks reversed. This suggests that at 95ºC, 1.5% of crosslinks are reversed per minute.

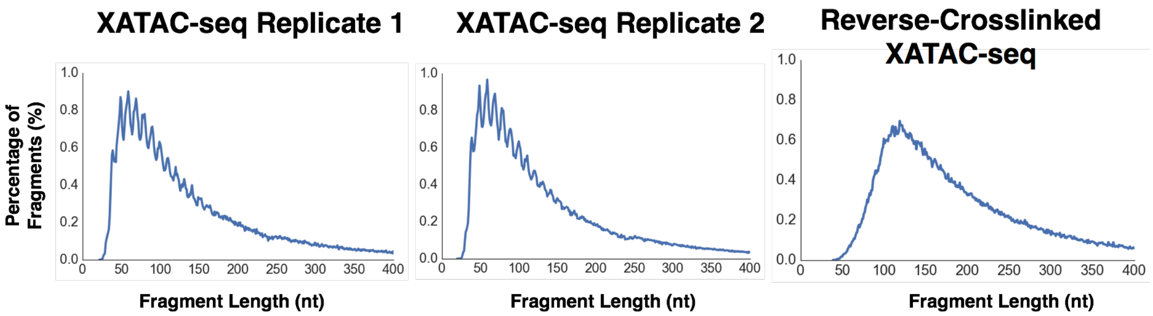


Figure 12: Effects of Reverse Crosslinking. The fragment size distributions of duplicate XATAC-seq libraries are compared with that of the reverse-crosslinked XATAC-seq library. It can be seen that the fragment lengths in the reverse-crosslinked library are more evenly distributed and that the average fragment is longer in this library than the standard.

It is interesting to note the oscillatory behavior of the fragment length distribution (Fig. 12). The same was observed in ATAC-seq results at both the scale of the DNA helical pitch and the length of DNA wrapped by a nucleosome. In the case of XATAC-seq, it appears that the period is roughly 10 nt, corresponding to the helical pitch of DNA as it extends away from the transposase[82].

Acknowledgments: Chapter 2, in part, is currently being prepared for submission for publication of the material. Al-Bassam, Mahmoud; Chapin, Nate; Gao, Nina; Zengler, Karsten; Nizet, Victor. The thesis author is co-first author of this paper.

CHAPTER 3: XATAC-SEQ AND H-NS-LIKE NUCLEOID-ASSOCIATED PROTEINS

In general, XATAC-seq signal is dominated by NAP binding, with H-NS-like proteins in particular being associated with the majority of signal. Correlations to the major NAPs in *E. coli* are shown in Table 1. For example, the genome-wide correlation between an H-NS ChIP-seq dataset obtained from work by Kahramanoglou and colleagues and *E. coli* XATAC-seq is 0.60 (Fig. 13).

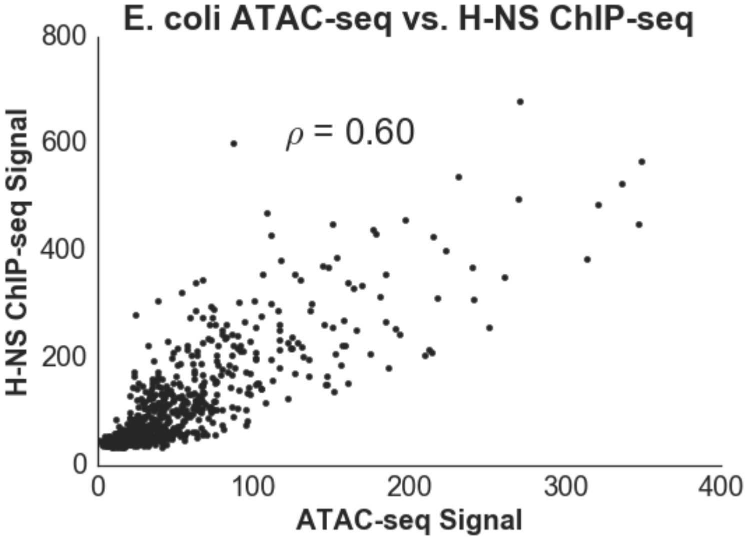


Figure 13: H-NS Correlation. Correlation between XATAC-seq in E. coli and H-NS ChIP-seq; data points are 5kb average signal, but correlation is genome-wide at the single nucleotide level.

Because H-NS is known to perform functions that appear to address fundamental difficulties faced by cells, namely the silencing of horizontally-acquired DNA and prevention of spurious expression, we wondered if similar proteins perform the same functions in a wide range of bacteria. Specifically, it appears that an AT-binding repressor would provide advantages regardless of the species' native AT-content. For those with a high genomic GC, foreign DNA is more likely to be AT-rich by comparison. In contrast, a high-AT genome would be expected to include more intragenic promoter-like sequences. As such, we decided to apply XATAC-seq to determine if proteins exhibiting widespread binding to AT-rich DNA and broad repression of transcription are more universal than currently realized.

3.1 ROK OF *BACILLUS SUBTILIS*

Bacillus subtilis is a model gram-positive, spore-forming bacterium. Among gram-positives, it is one of the most well studied and is a widely-used species in industry.

When XATAC-seq was applied to *B. subtilis* cells, a correlation was observed between XATAC-seq signal and regional AT content. A large portion of the singal appears to be contributed by the Rok protein – the correlation between Rok ChIP-seq and XATAC-seq was found to be 0.27.

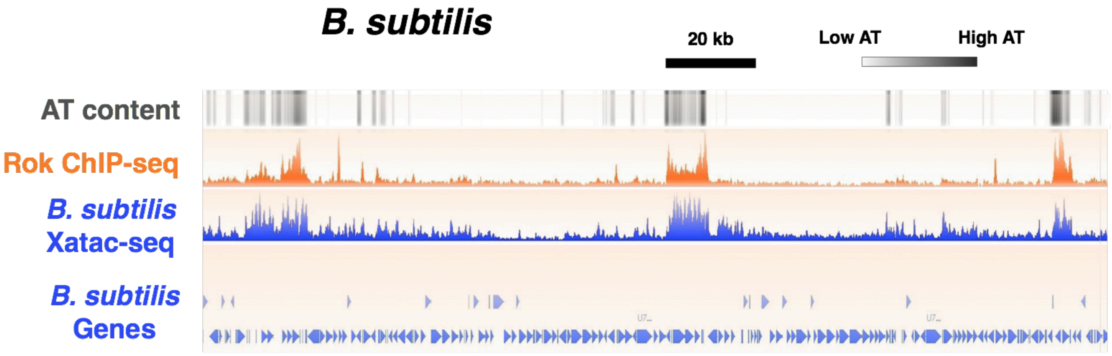


Figure 14: Rok of *B. subtilis*. Relationship between XATAC-seq, rok binding, and AT content of the *B. subtilis* genome.

Rok is a repressor of many genes in *B. subtilis*, including those associated with competence development[89]. Like H-NS, it binds to extended, AT-rich genome regions[73].

In addition, Rok, like H-NS, is an autorepressor[89]. The C-terminal domain of Rok has been identified as responsible and sufficient for DNA binding and Rok has been shown to possess the same AT hook DNA-binding motif as H-NS[43,73]. The degree to which Rok binding contributes to genome-wide binding, as measured by XATAC-seq, indicates that it may have a larger role in determining cellular phenotype than previously recognized.

Overall, there is considerable evidence that Rok is a functional homolog of H-NS in *Bacillus subtilis*. Because of the fact that it is restricted to a small number of closely related *Bacillus* species, it has been speculated that Rok was acquired by lateral gene transfer sometime in *B. subtilis's* recent evolutionary history[73]. However, when the sequence homology of H-NS and Rok is compared using BLAST, there is no significant agreement. In fact, when the homology of all major H-NS-like protein families is compared, none share the degree of sequence similarity that would be expected were they to have a common evolutionary origin (Table 2). As such, these proteins, and specifically their DNA-binding properties, appear to be a result of convergent evolution, reinforcing the idea that the functions of AT-binding proteins meet fundamental cellular needs.

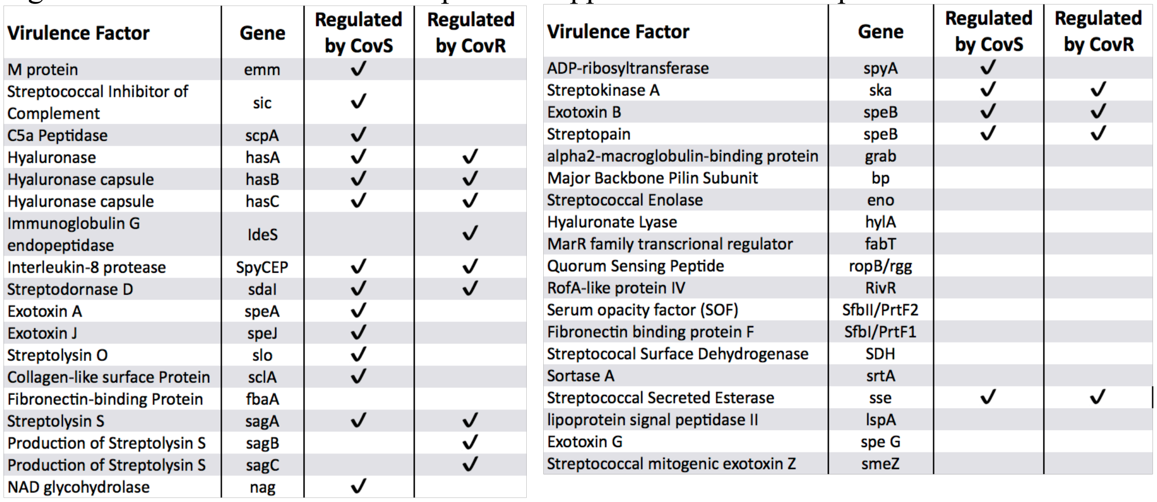
Table 2: Homology of H-NS-family Proteins. Homology was assessed using BLAST two-sequence comparison. The total score is the sum of the quality of alignments of individual segments of the protein. In many cases where no significant similarity was observed.

Total BLAST Scores	**H-NS**	**Rok**	**MvaT**	**Lsr2**
H-NS	**276**	0	52.3	0
Rok	0	**381**	23.9	0
MvaT	52.3	23.9	**249**	26.2
Lsr2	0	0	26.2	**221**

3.2 MGA OF GROUP A *STREPTOCOCCUS PYOGENES*

Group A *Streptococcus pyogenes* (GAS) is a human pathogen responsible for numerous invasive diseases that cause an estimated 163,000 deaths each year[90]. M1T1 is the GAS serotype most frequently associated with severe infections, which are often difficult to treat with antibiotics and may require surgical intervention[91,92]. Invasive bacterial disease is dependent upon the action of virulence factors that moderate interactions of bacteria with host tissues and facilitate subversion of the innate immune system. Pathogenesis in GAS serotype M1T1 is potentiated by mutations in the genes encoding the two-component CovRS system, causing upregulation of the several virulence-associated genes it regulates[93]. In total, CovRS is responsible for regulation of approximately 15% of the genes in GAS[94] – CovS is a membrane-associated histidine kinase that controls the phosphorylation state of CovR, the response regulator of the system[94]. A summary of virulence genes in M1 GAS is provided in Table 3, along with known effects of CovR or CovS on each, if applicable.

Table 3: GAS Virulence Genes. CovR negatively regulates hyaluronic acid capsule synthesis, streptolysin S, streptodornase D, streptokinase, spyCEP, IdeS, and positively regulates exotoxin B[87,95,96]. SIC repression appears to be covS-dependent[97].



Virulence Factor	Gene	Regulated by CovS	Regulated by CovR
M protein	emm	✓	
Streptococcal Inhibitor of Complement	sic	✓	
C5a Peptidase	scpA	✓	
Hyaluronase	hasA	✓	✓
Hyaluronase capsule	hasB	✓	✓
Hyaluronase capsule	hasC	✓	✓
Immunoglobulin G endopeptidase	IdeS		✓
Interleukin-8 protease	SpyCEP	✓	✓
Streptodornase D	sdaI	✓	✓
Exotoxin A	speA	✓	
Exotoxin J	speJ	✓	
Streptolysin O	slo	✓	
Collagen-like surface Protein	sclA	✓	
Fibronectin-binding Protein	fbaA		
Streptolysin S	sagA	✓	✓
Production of Streptolysin S	sagB		✓
Production of Streptolysin S	sagC		✓
NAD glycohydrolase	nag	✓	

Virulence Factor	Gene	Regulated by CovS	Regulated by CovR
ADP-ribosyltransferase	spyA	✓	
Streptokinase A	ska	✓	✓
Exotoxin B	speB	✓	✓
Streptopain	speB	✓	✓
alpha2-macroglobulin-binding protein	grab		
Major Backbone Pilin Subunit	bp		
Streptococcal Enolase	eno		
Hyaluronate Lyase	hylA		
MarR family transcrional regulator	fabT		
Quorum Sensing Peptide	ropB/rgg		
RofA-like protein IV	RivR		
Serum opacity factor (SOF)	SfbII/PrtF2		
Fibronectin binding protein F	SfbI/PrtF1		
Streptococal Surface Dehydrogenase	SDH		
Sortase A	srtA		
Streptococcal Secreted Esterase	sse	✓	✓
lipoprotein signal peptidase II	lspA		
Exotoxin G	spe G		
Streptococcal mitogenic exotoxin Z	smeZ		

We have applied XATAC-seq to study the changes in protein binding associated with the transition from non-pathogenic M1T1 GAS to the hypervirulent *covR* deletion mutant (Δ*covR*) as well as another hypervirulent mutant containing a *covS* point-mutation at position 877, procured from subcutaneous animal passage (AP).

3.2.1 GAS XATAC-SEQ AND AT CONTENT

XATAC-seq in each of the mutants shows a similar AT-preference, each with a correlation of approximately 0.2. To ensure that this signal preference is due to affinity of the binding proteins, we conducted an experiment in which we treated each of the GAS strains with rifampicin, an antibiotic that prevents transcriptional elongation beyond 2-3 nt[98]. In theory, this should allow DNA-binding proteins access to essentially all genomic DNA, and resulting XATAC-seq signal should be representative of their binding preferences. With rifampicin, the association between XATAC-seq signal and AT increases significantly (Fig. 15). This leads to the conclusion that expression prevents binding of proteins in GAS from accessing some high-affinity targets, and implies an inverse relationship between binding and transcription at these genes. Overall, this finding lends credence to the hypothesis that an H-NS-like AT-binding protein, or multiple, exist in GAS. In addition, the similarity of rifampicin-treated XATAC-seq signal between mutants implies that the same DNA-binding proteins are active in both phenotypes.

`

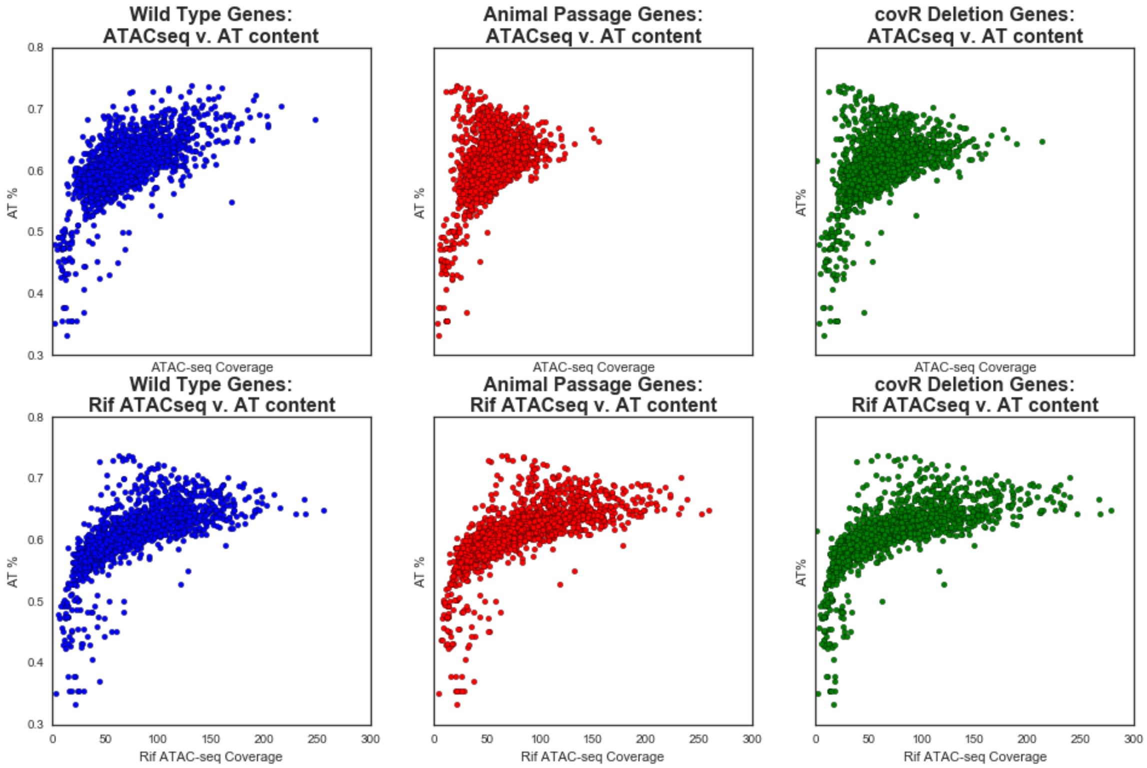


Figure 15: Effects of AT Content on Binding in Untreated and Rifampicin-Treated Cells. The correlation between XATAC-seq and AT improves significantly upon rifampoicin treatment in both of the mutants. With rifampicin treatment, XATAC-seq signal converges to become nearly identical.

3.2.2: GAS XATAC-SEQ AND GENE EXPRESSION

In order to determine if the putative H-NS-like protein also shares its repressive capacity, we compared the binding of virulence genes between mutants. This analysis revealed that on the whole, virulence-associated genes are bound significantly more in wild-type GAS than in the mutants (Fig. 16). This, in agreement with the changes observed upon rifampicin treatment, suggests that the activity of AT-binding proteins in GAS is generally repressive.

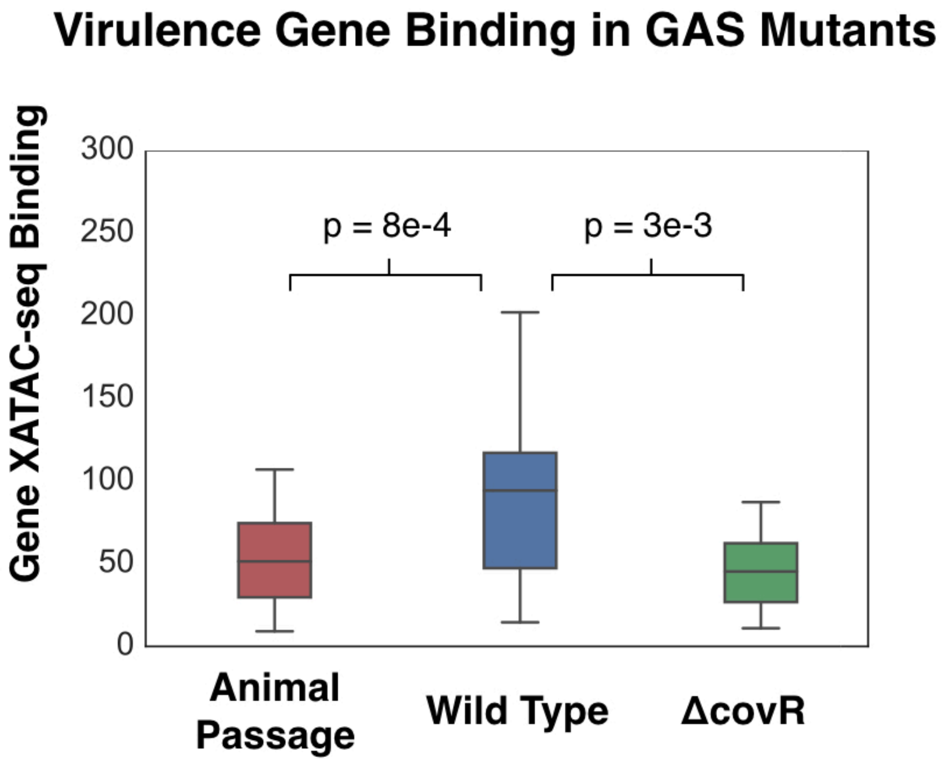


Figure 16: Virulence Gene Binding. Binding (average XATAC-seq signal normalized by gene length). The p-value was calculated using the Mann-Whitney U test; the difference in binding between the hypervirulent mutants is not statistically significant.

Clearly, since binding within genes antagonizes their transcription (see Fig. 2C), this represents an increase in expression of these genes, as has been shown previously for similar serotypes and as corroborated by RNA-seq experiments performed on these strains (Table 4).

Differential expression was determined from the RNA-seq data using DESEq. Briefly, DESeq models the number of reads assigned to each gene as a binomial distribution, estimating the mean and variance from the data so that it is able to accurately infer whether differences in expression are the product of noise or a true, relevant difference between the samples[99]. In total, 158 genes are differentially expressed between the wild type and both of the mutants (Fig 17). Among these, 16 are virulence genes.

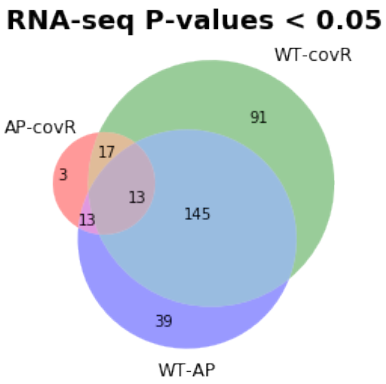


Figure 17: RNA-seq Differential Expression Venn Diagram. As expected, few genes are differentially expressed between the two hypervirulent mutants (pink region), and many genes were commonly differentially expressed between the wild type and each mutant (light blue region).

Interestingly, these differences in expression are mirrored by overall differences in the protein occupancy landscapes of the strains (Fig. 18). The correlation between the hypervirulent mutants is on the same order as that of replicates (0.94), whereas that between each of the mutants and the wild type is considerable lower (~0.6).

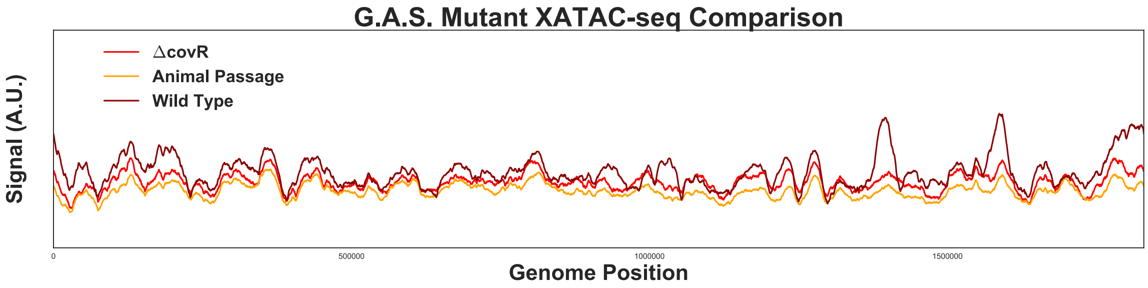
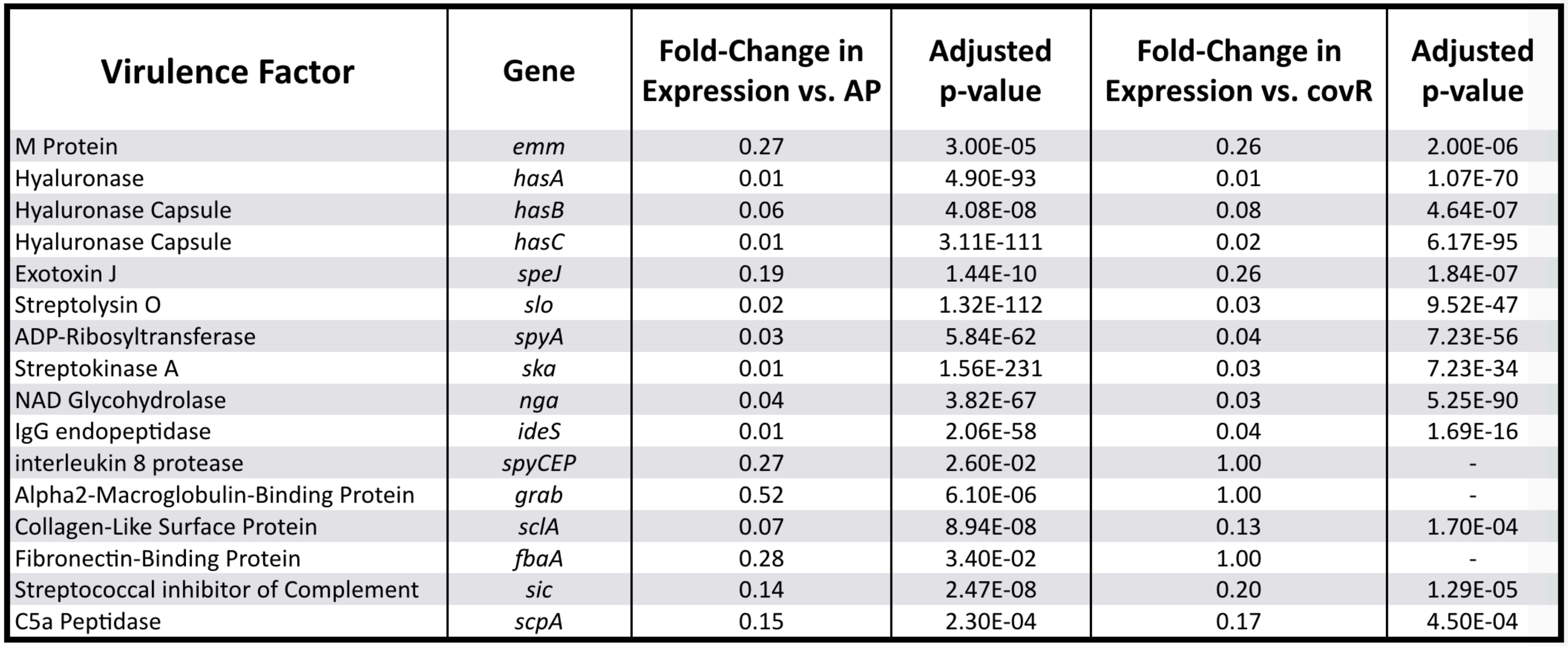


Figure 18: GAS POL Comparison. GAS XATAC-seq signal is plotted across the entire genome. The data shown has been smoothed in order to make it easier to visualize.

Table 4: GAS RNA-seq. RNA-seq reveals up-regulation of a number of virulence genes in the hypervirulent mutant strains with respect to the wild type.



Virulence Factor	Gene	Fold-Change in Expression vs. AP	Adjusted p-value	Fold-Change in Expression vs. covR	Adjusted p-value
M Protein	*emm*	0.27	3.00E-05	0.26	2.00E-06
Hyaluronase	*hasA*	0.01	4.90E-93	0.01	1.07E-70
Hyaluronase Capsule	*hasB*	0.06	4.08E-08	0.08	4.64E-07
Hyaluronase Capsule	*hasC*	0.01	3.11E-111	0.02	6.17E-95
Exotoxin J	*speJ*	0.19	1.44E-10	0.26	1.84E-07
Streptolysin O	*slo*	0.02	1.32E-112	0.03	9.52E-47
ADP-Ribosyltransferase	*spyA*	0.03	5.84E-62	0.04	7.23E-56
Streptokinase A	*ska*	0.01	1.56E-231	0.03	7.23E-34
NAD Glycohydrolase	*nga*	0.04	3.82E-67	0.03	5.25E-90
IgG endopeptidase	*ideS*	0.01	2.06E-58	0.04	1.69E-16
interleukin 8 protease	*spyCEP*	0.27	2.60E-02	1.00	-
Alpha2-Macroglobulin-Binding Protein	*grab*	0.52	6.10E-06	1.00	-
Collagen-Like Surface Protein	*sclA*	0.07	8.94E-08	0.13	1.70E-04
Fibronectin-Binding Protein	*fbaA*	0.28	3.40E-02	1.00	-
Streptococcal inhibitor of Complement	*sic*	0.14	2.47E-08	0.20	1.29E-05
C5a Peptidase	*scpA*	0.15	2.30E-04	0.17	4.50E-04

Differential gene binding, as assessed by DESeq of XATAC-seq data, shows a negative correlation with differential expression (Fig. 19). This correlation (-0.4) is on the same order of magnitude as that observed from ChIP-chip experiments performed with RNA polymerase and H-NS[47]. This implies that global binding of protein in GAS, on the whole, has the same repressive effect as that of H-NS in *E. coli.* In particular, many of the virulence genes identified as active in the mutants by RNA-seq are differentially bound in the mutants compared to the wild type. Despite this result, we do not have a clear mechanistic understanding of the relationship of binding to expression, nor is it clear whether changes in gene binding are a cause or an effect of changes in gene expression.

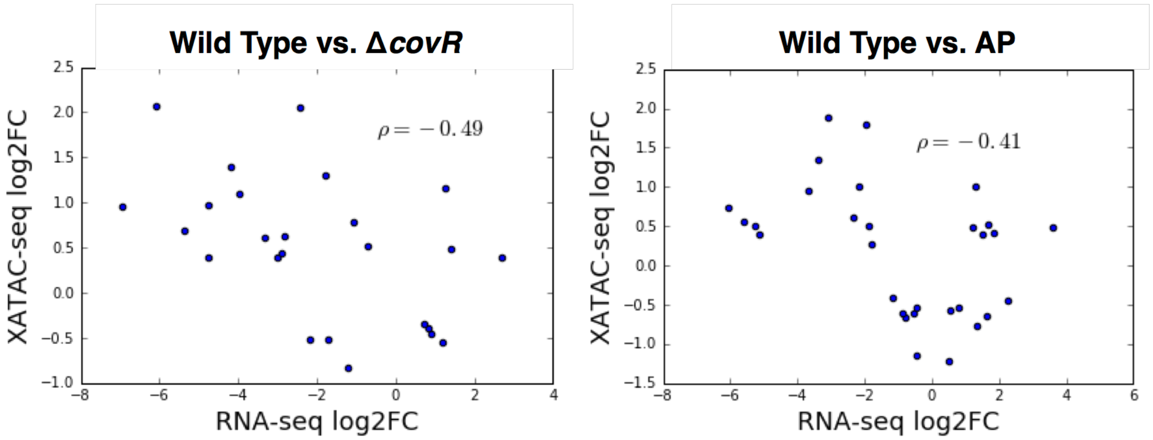


Figure 19: Differential Expression vs. Differential Binding. A negative correlation between change in binding and change in expression is observed at a significance threshold of 1e-3.

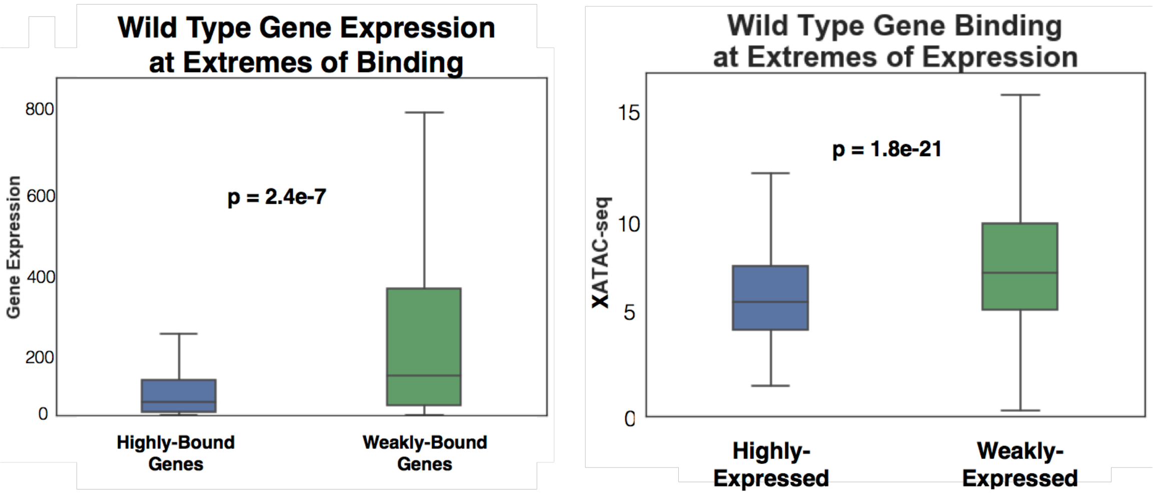


Figure 20: Gene binding vs. Expression. Left: comparison of the average RNA-seq signal (expression) for genes whose average XATAC-seq signal (binding) is above the 75th percentile and below the 25th percentile. Right: comparison of the average XATAC-seq signal (binding) for genes whose average RNA-seq signal (expression) is above the 75th percentile and below the 25th percentile. P-values were calculated using the Mann-Whitney U-test.

3.2.3 MASS SPECTROSCOPY ANALYSIS

In order to determine whether there is a protein of a similar functional nature to H-NS in GAS, we performed mass spectroscopy (MS) analysis on protein isolated from wild type GAS. DNA probes were amplified *in vitro* from regions of the GAS genome corresponding to bound, closed chromatin, and unbound, tagmented chromatin (Fig. 21). Briefly, probes were biotinylated, then mixed with cell lysate in order to fish out proteins with binding affinity for this region. Upon analysis, it was found that the Mga protein was the most enriched in the test sample compared to the control (Table 5). In addition, it was significantly more enriched than the next highest protein, and the only member of the ten most enriched that is known to be a DNA-binding protein. This indicates that Mga is the sole AT-binding protein responsible for a large portion of XATAC-seq signal, and therefore genome-wide binding in GAS.

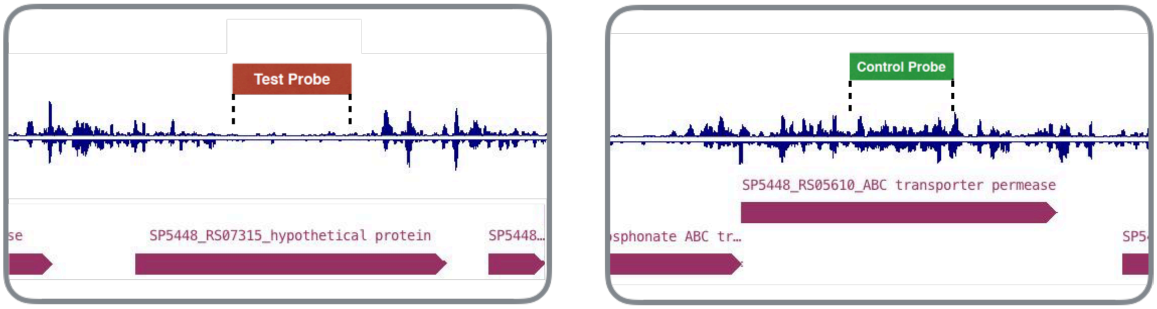

Test Probe
se
SP5448_RS07315_hypothetical protein
SP5448…
Control Probe
SP5448_RS05610_ABC transporter permease
sphonate ABC tr…
SP5

Figure 21. Mass Spectroscopy Probe Design. Test probes were amplified from regions of minimal tagmentation, representing occupied DNA. Control probes were amplified from regions of frequent tagmentation, representing accessible DNA. Two of each were used.

Table 5: Mass Spectroscopy Results. The number of peptide spectral matches (PSMs) associated with each protein in the test and control samples is compared to determine which is most responsible for binding as assessed by XATAC-seq.

Gene	Protein	Control PSMs	Sample PSMs	Difference
SP5448_RS08410	**Mga**	**13**	**41**	**28**
SP5448_RS02735	Ribonuclease Y	9	15	6
SP5448_RS01725	Excinuclease ABC subunit A	9	15	6
SP5448_RS06210	50S ribosomal protein L35	6	11	5
SP5448_RS00320	50S ribosomal protein L2	4	8	4
SP5448_RS00340	50S ribosomal protein L16	7	10	3
SP5448_RS05870	GTP-binding protein	3	6	3
SP5448_RS04825	Signal recognition particle protein	3	6	3
SP5448_RS01250	30S ribosomal protein S12	16	18	2
SP5448_RS00060	Cell division protein FtsH	12	14	2

Mga is a large (62kD) DNA-binding protein in GAS. It binds DNA at its N-terminal domain, which contains a pair of helix-turn-helices[100]. It is known to repress genes associated with sugar metabolism, as well as directly bind the promoter regions of several virulence genes associated with cell-surface proteins and interactions with host tissues in order to activate their expression[101–103]. For this reason, it is currently classified as a specific regulator. Despite this, Mga lacks a consensus binding sequence[104]. In addition, the regulatory behavior is modulated by the phosphorylation state of the protein[105]. In particular, the phosphorylation state of Mga is known to determine whether or not it is capable of oligomerizing, which for H-NS is known to determine its ability to repress expression. Another point of similarity is that Mga autorepresses *mga*, just as H-NS does *hns*[106]. This indicates that it is important for the cell to carefully control the ratio of Mga to DNA. Furthermore, *in vitro* studies have shown that Mga binds regions of *E. coli* DNA that are strongly bound by H-NS; and conversely that H-NS strongly binds a promoter regulated by Mga[107]. Overall, Mga has been shown to regulates approximately 10% of the GAS genome, which, in combination with our XATAC-seq findings, lends itself well to the hypothesis that Mga acts in a role more similar to that of H-NS-family proteins[93]. As such, we conclude that Mga is likely to be a functional homolog of H-NS in GAS, and that the current understanding of its function and role is incomplete.

Acknowledgments: Chapter 3, in part, is currently being prepared for submission for publication of the material. Al-Bassam, Mahmoud; Chapin, Nate; Gao, Nina; Zengler, Karsten; Nizet, Victor. The thesis author is co-first author of this paper.

APPENDIX

A1. BACTERIAL STRAINS AND GROWTH CONDITIONS

E. coli K12 cells were grown to mid-log phase (OD_{600} = 0.5) in LB media at 37ºC with shaking.

Bacillus subtilis cells were grown in rich CH media at 37ºC with stirring[108].

Group A. Streptococcus pyogenes strain 5448 Wild type (serotype M1T1), *ΔcovR* mutant, and *covS* point-mutation at position 877 - "Animal Passage" mutant were used. Cells were grown in Todd Hewitt broth (Hardy Diagnostics) at 37ºC with shaking.

Clostridium ljungdalhii, were cultured in carbon monoxide and high-fructose media at 27ºC with stirring.

A2. EXPERIMENTAL METHODS

A2.1 XATAC-seq

Crosslinking and Protein-DNA Complex Isolation

Bacterial cultures were grown to mid exponential phase (OD_{600} = 0.3-0.5) in appropriate media (Appendix A1) at 37ºC, with shaking. Crosslinking was achieved by treatment with 1% formaldehyde for 20-30 minutes. Cells were pelleted by centrifugation and cell pellets were lysed by grinding in liquid nitrogen. 500 uL SET buffer (75mM NaCl, 25mM EDTA pH 8, 20nM Tris-Hcl pH 7.5) were used for grinding. Lysate was resuspended in 2X protease inhibitor solution (cOmplete mini, Roche) and centrifuged for 10 min. at 14,000 rpm and 4ºC. 25 uL of supernatant was used for buffer exchange with Tris-EDTA (10M Tris, 1mM EDTA, pH 8) with a 45 minute incubation period at room temperature.

Xatac-seq Library Preparation

700 pg DNA were used as input for the Illumina Nextera kit. After library preparation, AMPure beads were used to purify the library as recommended by the manufacturer.

Sequencing

Libraries were sequenced on Illumina MiSeq or HiSeq for 100 or 150 cycles in paired-end mode.

A2.2 DNA Control ATAC-seq

The procedure for performing the DNA-only control experiment is the same as that for XATAC-seq, with the following exceptions and additional steps:

1) The formaldehyde crosslinking step is skipped.
2) After cell lysis by nitrogen grinding, 2 rounds of phenol:chloroform extraction are performed.
3) Ethanol precipitation is performed with glycerol.

700 pg of this genomic DNA are used as input to the Illumina Nextera kit, as in XATAC-seq.

A2.3 Rifampicin Treatment Experiments

Rifampicin treatment experiments are performed using the standard XATAC-seq protocol, but before treatment with formaldehyde, cells are treated with a final concentration of 25ug/mL Rifampicin and incubated at 37ºC for 30 minutes, with shaking.

A2.4 RNA-seq

RNA Extraction

S. pyogenes cultures were grown to mid exponential phase in Todd Hewitt media. Cells were pelleted by centrifugation and cell pellets were lysed by grinding in liquid nitrogen with 300 µl RLT buffer (Qiagen). Lysates were resuspended in 1 ml Trizol and 200 uL chloroform. Solution was vortex mixed and centrifuged to separate phases, after which the aqueous phase was extracted. Finally, the sample were purified with Qiagen RNEasy columns.

rRNA Removal and RNA-seq Library Preparation

2 µg of total RNA were used as input to the RiboZero kit (Illumina). 50 ng of purified, rRNA-depleted RNA was used as input to the KAPA Stranded RNA-seq Library Preparation Kit.

Sequencing

Libraries were sequenced on Illumina MiSeq or HiSeq for 100 or 150 cycles in paired-end mode.

A2.5 Protein Extraction for MS

Probe Amplification

Biotinylated primers for select regions of the GAS genome (2 control, 2 test) were requisitioned from IDT. Cell lysate was prepared as per protocol outlined in "*Crosslinking and Protein-DNA Complex Isolation*" of A2.1. Each set of primers was mixed with 1 μl lysate, dNTPs, and Q5 DNA polymerase in Q5 buffer. The following PCR thermocycler program was run for 30 cycles in order to amplify target regions:

98ºC 2 min.

30x: 98ºC 25 sec.

43ºC 15 sec.

72ºC 15 sec.

Bait Purification

PCR products were washed with Quiagen columns. 5X PBS added to each sample containing amplified primer to bind columns. Columns were washed twice with PE, and DNA was eluted in 25 μl H_2O.

Protein Extraction

10 μl of each test bait mixed and added to 500 μl cell lysate. The same was repeated for the controls. Solutions were incubated on a rotating stand mixed for 1 hr. at 4ºC. Dynabeads were washed according to manufacturer's instructions. 100 μl bead solution was added to each sample, and washed 6 times with wash buffer (50mM Tris, 250mM NaCl, 0.1% Triton 100X) at 4ºC. Proteins were eluted by incubation in 2.5M NaCl solution for 1 hr. at room temperature.

A2.6: XATAC-seq with Reverse-Crosslinking

The procedure for performing the revere-crosslinking experiment is the same as that for XATAC-seq, with the following additional steps:

1) After quenching the transposase reaction with NT buffer, sample was purified wirg 1.8X volume AMPure bead solution.
2) Sample was incubated with 1 ul proteinase K at 65ºC overnight to reverse crosslinks.
3) Sample was purified again to remove protein debris and protease, using 1.8X volume AMPure bead solution.
4) Nextera PCR amplification and subsequent steps were performed as in the standard XATAC-seq protocol.

A3. DATA ANALYSIS

In the general procedure, primers and adapter sequences are removed using trim_galore in paired-end mode (--paired) with the quality cutoff (-q) set to 22 and -fastqc enabled. Next, reads are aligned to the reference genome using bowtie2, with the maximum length limit (-X) set to 1000[109]. Wig files containing the number of mappings at each genome position are then generated using the samtools mpileup command and normalized by reads per million (RPM). The resulting wig files are then processed using in-house Python scripts.

When evaluating differential expression (or differential binding) of genes, trimming is performed as in the general procedure. Next, featureCounts is used to determine the number of fragments corresponding to each region of interest (features), which could be a gene or promoter. A minimum of 2/3 of each read must be within the gene in order for it to be assigned (--fracoverlap 0.66). DESeq is then implemented in R to determine the level and significance of differential signal for each feature using a negative binomial distribution[99]. Further analysis, including imposition of significance thresholds and sorting by magnitude of differential signal, is performed using custom Python scripts.

In order to accurately determine the location of specific transposition events so as to precisely pinpoint individual binding sites, mapped reads must be trimmed to a single base. Therefore, trimming and alignment are performed as in the general case using trim_galore and bowtie2. Afterward, the position field, sequence field, and CIGAR field of the sam file are adjusted appropriately. An additional offset of +4 bases for reads on the forward strand and -5 for those on the reverse strand is applied because the Tn5 transposase introduces a 9bp gap on either side of its transposition site which is subsequently duplicated and must be corrected for in order to obtain the true site of transposition[14]. Once this is done, wig files can be generated from the modified sams with samtools commands as usual, and footprints are detected and evaluated in Python.

REFERENCES

1. Deng, S., Stein, R. A. & Higgins, N. P. Organization of supercoil domains and their reorganization by transcription. *Mol. Microbiol.* **57,** 1511–21 (2005).

2. Holmes, V. F. & Cozzarelli, N. R. Closing the ring: links between SMC proteins and chromosome partitioning, condensation, and supercoiling. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 1322–4 (2000).

3. Kornberg, R. D. Chromatin Structure: A Repeating Unit of Histones and DNA. *Science (80-.).* **184,** (1974).

4. Drlica, K. & Rouviere-Yaniv, J. Histonelike Proteins of Bacteria. *Microbiol. Rev.* **51,** 301–319 (1987).

5. Noom, M. C., Navarre, W. W., Oshima, T., Wuite, G. J. L. L. & Dame, R. T. H-NS promotes looped domain formation in the bacterial chromosome. *Curr. Biol.* **17,** R913–R914 (2007).

6. Dekker, J. & Heard, E. Structural and functional diversity of Topologically Associating Domains. *FEBS Lett.* **589,** 2877–84 (2015).

7. Postow, L., Hardy, C. D., Arsuaga, J. & Cozzarelli, N. R. Topological domain structure of the Escherichia coli chromosome. *Genes Dev.* **18,** 1766–79 (2004).
8. Worcel, A., Burgi, E. & Burgti, E. On the Structure of the Folded Chromosome of Escherichia coli. *J. Mol. Biol.* **71,** 127–147 (1972).

9. Le, T. B. & Laub, M. T. Transcription rate and transcript length drive formation of chromosomal interaction domain boundaries. *EMBO J.* **35,** 1582–1595 (2016).

10. Le, T. B. K., Imakaev, M. V, Mirny, L. A. & Laub, M. T. High-resolution mapping of the spatial organization of a bacterial chromosome. *Science* **342,** 731–4 (2013).

11. Wang, X., Le, T. B. K., Lajoie, B. R., Dekker, J., Laub, M. T. & Rudner, D. Z. Condensin promotes the juxtaposition of DNA flanking its loading site in Bacillus subtilis. *Genes Dev.* **29,** 1661–75 (2015).

12. Vora, T., Hottes, A. K., Tavazoie, S., Cornet, F., Boccard, F., McLeod, S. M., Marko, J. F., Johnson, R. C., Hannett, N., Kanin, E. & al., et. Protein occupancy landscape of a bacterial genome. *Mol. Cell* **35,** 247–53 (2009).

13. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*

10, 1213–1218 (2013).

14. Vaezeslami, S., Sterling, R. & Reznikoff, W. S. Site-directed mutagenesis studies of tn5 transposase residues involved in synaptic complex formation. *J. Bacteriol.* **189,** 7436–41 (2007).

15. Ali Azam, T., Iwata, A., Nishimura, A., Ueda, S. & Ishihama, A. Growth phase-dependent variation in protein composition of the Escherichia coli nucleoid. *J. Bacteriol.* **181,** 6361–70 (1999).

16. Collado-Vides, J., Magasanik, B. & Gralla2, J. D. Control Site Location and Transcriptional Regulation in Escherichia coli. *Microbiol. Rev.* **55,** 371–394 (1991).

17. Madan Babu, M. & Teichmann, S. A. Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucleic Acids Res.* **31,** 1234–44 (2003).

18. Martínez-Antonio, A., Chandra Janga, S., Salgado, H. & Collado-Vides, J. Internal-sensing machinery directs the activity of the regulatory network in Escherichia coli. doi:10.1016/j.tim.2005.11.002

19. Balleza, E., Opez-Bojorquez, L. N., Martínez-Antonio, A., Resendis-Antonio, O., Lozada-C Avez, I., Balderas-Martínez, Y. I., Encarnací On, S. & Collado-Vides, J. Regulation by transcription factors in bacteria: beyond description. (2008). doi:10.1111/j.1574-6976.2008.00145.x

20. Browning, D. F. & W Busby, S. J. Local and global regulation of transcription initiation in bacteria. *Nat. Publ. Gr.* **14,** (2016).

21. Alberts, B., Johnson, A. & J, L. *Molecular Biology of the Cell.* (Garland Science, 2002).

22. Browning, D. F., Grainger, D. C. & Busby, S. J. Effects of nucleoid-associated proteins on bacterial chromosome structure and gene expression. *Curr. Opin. Microbiol.* **13,** 773–780 (2010).

23. Dorman, C. J. Regulation of transcription by DNA supercoiling in Mycoplasma genitalium: global control in the smallest known self-replicating genome. *Mol. Microbiol.* **81,** 302–304 (2011).

24. Dame, R. T., Kalmykowa, O. J. & Grainger, D. C. Chromosomal Macrodomains and Associated Proteins: Implications for DNA Organization and Replication in Gram Negative Bacteria. *PLoS Genet.* **7,** e1002123 (2011).

25. Dillon, S. C. & Dorman, C. J. Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. (2010). doi:10.1038/nrmicro2261

26. Balandina, A., Kamashev, D. & Rouviere-Yaniv, J. The bacterial histone-like protein HU specifically recognizes similar structures in all nucleic acids. DNA, RNA, and their hybrids. *J. Biol. Chem.* **277,** 27622–8 (2002).

27. Serban, D., Arcineigas, S. F., Vorgias, C. E., Thomas, G. J. & Jr. Structure and dynamics of the DNA-binding protein HU of B. stearothermophilus investigated by Raman and ultraviolet-resonance Raman spectroscopy. *Protein Sci.* **12,** 861–70 (2003).

28. Luijsterburg, M. S., White, M. F., van Driel, R. & Dame, R. T. The Major Architects of Chromatin: Architectural Proteins in Bacteria, Archaea and Eukaryotes. *Crit. Rev. Biochem. Mol. Biol.* **43,** 393–418 (2008).

29. Bensaid, A., Almeida, A., Drlica, K. & Rouviere-Yaniv, J. Cross-talk Between Topoisomerase I and HU in Escherichia coli. *J. Mol. Biol.* **256,** 292–300 (1996).

30. Dixon, N. E. & Kornberg, A. Protein HU in the enzymatic replication of the chromosomal origin of Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **81,** 424–8 (1984).

31. Huo, Y.-X., Zhang, Y.-T., Xiao, Y., Zhang, X., Buck, M., Kolb, A. & Wang, Y.-P. IHF-binding sites inhibit DNA loop formation and transcription initiation. *Nucleic Acids Res.* **37,** 3878–3886 (2009).

32. Sheridan, S. D., Benham, C. J. & Hatfield, G. W. Activation of Gene Expression by a Novel DNA Structural Transmission Mechanism That Requires Supercoiling-induced DNA Duplex Destabilization in an Upstream Activating Sequence*. *J. Biol. Chem.* **273,** 21298–21308 (1998).

33. Opel, M. L., Aeling, K. A., Holmes, W. M., Johnson, R. C., Benham, C. J. & Hatfield, G. W. Activation of transcription initiation from a stable RNA promoter by a Fis protein-mediated DNA structural transmission mechanism. *Mol. Microbiol.* **53,** 665–674 (2004).

34. Weinstein-Fischer, D. & Altuvia, S. Differential regulation of *Escherichia coli* topoisomerase I by Fis. *Mol. Microbiol.* **63,** 1131–1144 (2007).

35. Leonard, A. C. & Grimwade, J. E. Building a bacterial orisome: emergence of new regulatory features for replication origin unwinding. *Mol. Microbiol.* **55,** 978–85 (2005).

36. Spassky, A., Rimsky, S., Garreau, H. & Buc, H. H1a, an E. coli DNA-binding

protein which accumulates in stationary phase, strongly compacts DNA in vitro. *Nucleic Acids Res.* **12,** 5321–40 (1984).

37. Bertin, P., Benhabiles, N., Krin, E., Laurent-Winter, C., Tendeng, C., Turlin, E., Thomas, A., Danchin, A. & Brasseur, R. The structural and functional organization of H-NS-like proteins is evolutionarily conserved in Gram-negative bacteria. *Mol. Microbiol.* **31,** 319–329 (1999).

38. Navarre, W. W. Selective Silencing of Foreign DNA with Low GC Content by the H-NS Protein in Salmonella. *Science (80-.).* **313,** 236–238 (2006).

39. Singh, K., Milstein, J. N. & Navarre, W. W. Xenogeneic Silencing and Its Impact on Bacterial Genomes. *Annu. Rev. Microbiol* **70,** 199–213 (2016).

40. Owen-Hughes, T. A., Pavitt, G. D., Santos, D. S., Sidebotham, J. M., Hulton, C. S. J., Hinton, J. C. D. & Higgins, C. F. The Chromatin-Associated Protein H-NS Interacts with Curved DNA to Influence DNA Topology and Gene Expression. *Cell* **71,** 255–265 (1992).

41. Oshima, T., Ishikawa, S., Kurokawa, K., Aiba, H. & Ogasawara, N. Escherichia coli Histone-Like Protein H-NS Preferentially Binds to Horizontally Acquired DNA in Association with RNA Polymerase. *DNA Res.* **13,** 141–153 (2006).

42. Grainger, D. C., Hurd, D., Goldberg, M. D. & Busby, S. J. W. Association of nucleoid proteins with coding and non-coding segments of the Escherichia coli genome. *Nucleic Acids Res.* **34,** 4642–4652 (2006).

43. Gordon, B. R. G., Li, Y., Cote, A., Weirauch, M. T., Ding, P., Hughes, T. R., Navarre, W. W., Xia, B. & Liu, J. Structural basis for recognition of AT-rich DNA by unrelated xenogeneic silencing proteins. *Proc. Natl. Acad. Sci.* **108,** 10690–10695 (2011).

44. Ueguchi, C., Mizuno, T. & Buc, H. The Escherichia coli nucleoid protein H-NS functions directly as a transcriptional repressor. *EMBO J.* **1,** 39–1046 (1993).

45. Falconi, M., Higgins, N. P., Spurio, R., Pon, C. L. & Gualerzi, C. O. Expression of the gene encoding the major bacterial nucleoid protein H-NS is subject to transcriptional auto-repression. *Mol. Microbiol.* **10,** 273–282 (1993).

46. Hommais, F., Krin, E., Laurent-Winter, C., Soutourina, O., Malpertuy, A., Le Caer, J. P., Danchin, A. & Bertin, P. Large-scale monitoring of pleiotropic regulation of gene expression by the prokaryotic nucleoid-associated protein, H-NS. *Mol. Microbiol.* **40,** 20–36 (2001).

47. Lucchini, S., Rowley, G., Goldberg, M. D., Hurd, D., Harrison, M. & Hinton, J. C.

D. H-NS Mediates the Silencing of Laterally Acquired Genes in Bacteria. *PLoS Pathog.* **2,** e81 (2006).

48. Free, A. & Dorman, C. J. Coupling of Escherichia coli hns mRNA levels to DNA synthesis by autoregulation: implications for growth phase control. *Mol. Microbiol.* **18,** 101–113 (1995).

49. Afflerbach, H., Schroder, O. & Wagner, R. Effects of the Escherichia coli DNA-binding protein H-NS on rRNA synthesis in vivo. *Mol. Microbiol.* **28,** 641–653 (1998).

50. Ueguchi, C., Suzuki, T., Yoshida, T., Tanaka, K.-I. & Mizuno, T. Systematic Mutational Analysis Revealing the Functional Domain Organization of Escherichia coli Nucleoid Protein H-NS. *J. Mol. Biol* **263,** 149–162 (1996).

51. Bloch, V., Yang, Y., Margeat, E., Chavanieu, A., Augé, M. T., Robert, B., Arold, S., Rimsky, S. & Kochoyan, M. The H-NS dimerization domain defines a new fold contributing to DNA recognition. *Nat. Struct. Biol.* **10,** 212–218 (2003).

52. Esposito, D., Petrovic, A., Harris, R., Ono, S., Eccleston, J. F., Mbabaali, A., Haq, I., Higgins, C. F., Hinton, J. C. D., Driscoll, P. C. & Ladbury, J. E. H-NS Oligomerization Domain Structure Reveals the Mechanism for High Order Self-association of the Intact Protein. *J. Mol. Biol.* **324,** 841–850 (2002).

53. Dorman, C., Hinton, J. & Free, A. Domain Organization and Oligomerization Among H-NS-like Nucleoid-Associated Proteins in Bacteria. *TRENDS Microbiol.* **124,** (1999).

54. Rimsky, S., Zuber, F., Buckle, M. & Buc, H. A molecular mechanism for the repression of transcription by the H-NS protein. *Mol. Microbiol.* **42,** 1311–1323 (2001).

55. Ulissi, U., Fabbretti, A., Sette, M., Giuliodori, A. M. & Spurio, R. Time-resolved assembly of a nucleoprotein complex between Shigella flexneri virF promoter and its transcriptional repressor H-NS. *Nucleic Acids Res.* **42,** 13039–50 (2014).

56. Lang, B., Blot, N., Bouffartigues, E., Buckle, M., Geertz, M., Gualerzi, C. O., Mavathur, R., Muskhelishvili, G., Pon, C. L., Rimsky, S., Stella, S., Babu, M. M. & Travers, A. High-affinity DNA binding sites for H-NS provide a molecular basis for selective silencing within proteobacterial genomes. *Nucleic Acids Res.* **35,** 6330–7 (2007).

57. Spurio, R., Falconi, M., Brandi, A., Pon, C. L. & Gualerzi, C. O. The oligomeric structure of nucleoid protein H-NS is necessary for recognition of intrinsically curved DNA and for DNA bending. *EMBO J.* **16,** 1795–1805 (1997).

58. Winardhi, R. S., Fu, W., Castang, S., Li, Y., Dove, S. L. & Yan, J. Higher order oligomerization is required for H-NS family member MvaT to form gene-silencing nucleoprotein filament. *Nucleic Acids Res.* **40,** 8942–8952 (2012).

59. Dame, R. T., Wyman, C., Wurm, R., Wagner, R. & Goosen, N. Structural Basis for H-NS-mediated Trapping of RNA Polymerase in the Open Initiation Complex at the rrnB P1. *J. Biol. Chem.* **277,** 2146–2150 (2001).

60. Shin, M., Song, M., Rhee, J. H., Hong, Y., Kim, Y.-J., Seok, Y.-J., Ha, K.-S., Jung, S.-H. & Choy, H. E. DNA looping-mediated repression by histone-like protein H-NS: specific requirement of Esigma70 as a cofactor for looping. *Genes Dev.* **19,** 2388–98 (2005).

61. Kotlajich, M. V, Hron, D. R., Boudreau, B. A., Sun, Z., Lyubchenko, Y. L., Landick, R., Yamazaki, T., Marchadier, E., Hoebeke, M., Aymerich, S., Becher, D., Bisicchia, P., Botella, E., Delumeau, O., Doherty, G., Denham, E., Fogg, M., Fromion, V., Goelzer, A. Hansen, A., Härtig, E., Harwood, CR., Homuth, G., Jarmer, H., Jules, M., Klipp, E., Chat, L. Le, Lecointe, F., Lewis, P., Liebermeister, W., March, A., Mars, RA., Nannapaneni, P., Noone, D., Pohl, S., Rinn, B., Rügheimer, F., Sappa, PK., Samson, F., Schaffer, M., Schwikowski, B., Steil, L., Stülke, J., Wiegert, T., Devine, KM., Wilkinson, AJ., Dijl, JM. van, Hecker, M., Völker, U., Bessières, P., Noirot, P. Bridged filaments of histone-like nucleoid structuring protein pause RNA polymerase and aid termination in bacteria. *Elife* **4,** 1199–1208 (2015).

62. Shin, M., Lagda, A. C., Lee, J. W., Bhat, A., Rhee, J. H., Kim, J.-S., Takeyasu, K. & Choy, H. E. Gene silencing by H-NS from distal DNA site. *Mol. Microbiol.* **86,** 707–719 (2012).

63. Singh, S. S. & Grainger, D. C. H-NS Can Facilitate Specific DNA-binding by RNA Polymerase in AT-rich Gene Regulatory Regions. *PLoS Genet.* **9,** e1003589 (2013).

64. Ueda, T., Takahashi, H., Uyar, E., Ishikawa, S., Ogasawara, N. & Oshima, T. Functions of the Hha and YdgT Proteins in Transcriptional Silencing by the Nucleoid Proteins, H-NS and StpA, in Escherichia coli. *DNA Res.* **20,** 263–271 (2013).

65. Nieto, J. M., Madrid, C., Miquelay, E., Parra, J. L., Rodríguez, S. & Juárez, A. Evidence for Direct Protein-Protein Interaction between Members of the Enterobacterial Hha/YmoA and H-NS Families of Proteins. *J. Bacteriol.* **184,** 629–635 (2002).

66. Lamberte, L. E., Baniulyte, G., Singh, S. S., Stringer, A. M., Bonocora, R. P., Stracy, M., Kapanidis, A. N., Wade, J. T. & Grainger, D. C. Horizontally acquired

AT-rich genes in Escherichia coli cause toxicity by sequestering RNA polymerase. *Nat. Microbiol.* **2,** 16249 (2017).

67. Dorman, C. J. H-NS-like nucleoid-associated proteins, mobile genetic elements and horizontal gene transfer in bacteria. *Plasmid* **75,** 1–11 (2014).

68. Navarre, W. W., McClelland, M., Libby, S. J. & Fang, F. C. Silencing of xenogeneic DNA by H-NS-facilitation of lateral gene transfer in bacteria by a defense system that recognizes foreign DNA. *Genes Dev.* **21,** 1456–71 (2007).

69. Ali, S. S., Xia, B., Liu, J. & Navarre, W. W. Silencing of foreign DNA in bacteria. *Curr. Opin. Microbiol.* **15,** 175–181 (2012).

70. Higashi, K., Tobe, T., Kanai, A., Uyar, E., Ishikawa, S., Suzuki, Y., Ogasawara, N., Kurokawa, K. & Oshima, T. H-NS Facilitates Sequence Diversification of Horizontally Transferred DNAs during Their Integration in Host Chromosomes. *PLoS Genet.* (2016).

71. Singh, S. S., Singh, N., Bonocora, R. P., Fitzgerald, D. M., Wade, J. T. & Grainger, D. C. Widespread suppression of intragenic transcription initiation by H-NS. *Genes Dev.* **28,** (2014).

72. Williams, R. M., Rimsky, S. & Buc, H. Probing the Structure, Function, and Interactions of the Escherichia coli H-NS and StpA Proteins by Using Dominant Negative Derivatives. *J. Bacteriol.* **178,** 4335–4343 (1996).

73. Smits, W. K. & Grossman, A. D. The Transcriptional Regulator Rok Binds A+T-Rich DNA and Is Involved in Repression of a Mobile Genetic Element in Bacillus subtilis. *PLoS Genet.* **6,** e1001207 (2010).

74. Gordon, B. R. G., Imperial, R., Wang, L., Navarre, W. W. & Liu, J. Lsr2 of Mycobacterium represents a novel class of H-NS-like proteins. *J. Bacteriol.* **190,** 7052–9 (2008).

75. Lu, K., Ye, W., Zhou, L., Collins, L. B., Chen, X., Gold, A., Ball, L. M. & Swenberg, J. A. Structural Characterization of Formaldehyde-Induced Cross-Links Between Amino Acids and Deoxynucleosides and Their Oligomers. *J. Am. Chem. Soc.* **132,** 3388–3399 (2010).

76. Hoffman, E. A., Frey, B. L., Smith, L. M. & Auble, D. T. Formaldehyde Crosslinking: A Tool for the Study of Chromatin Complexes *. (2015). doi:10.1074/jbc.R115.651679

77. Gavrilov, A., Razin, S. V & Cavalli, G. In vivo formaldehyde cross-linking: it is time for black box analysis. (2014). doi:10.1093/bfgp/elu037

78. Toth, J. & Biggin, M. D. The specificity of protein–DNA crosslinking by formaldehyde: in vitro and in Drosophila embryos. *Nucleic Acids Res.* **28,** (2000).

79. Goryshin, I. Y. & Reznikoff, W. S. Tn5 in vitro transposition. *J. Biol. Chem.* **273,** 7367–74 (1998).

80. Goryshin, I. Y., Miller, J. A., Kil, Y. V, Lanzov, V. A. & Reznikoff, W. S. Tn5/IS50 target recognition. *Genetics* **95,** 10716–10721 (1998).

81. York, D. & Reznikoff, W. S. DNA binding and phasing analyses of Tn5 transposase and a monomeric variant. *Nucleic Acids Res.* **25,** 2153–60 (1997).

82. Adley, A., Morrison, H. G., Asan, Xun, X., Kitzman, J. O., Turner, E. H., Stackhouse, B., MacKenzie, A. P., Caruccio, N. C., Zhang, X. & Schendure, J. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol.* **11,** (2010).

83. Keseler, I. M., Mackie, A., Peralta-Gil, M., Santos-Zavaleta, A., Gama-Castro, S., Bonavides-Martinez, C., Fulcher, C., Huerta, A. M., Kothari, A., Krummenacker, M., Latendresse, M., Muniz-Rascado, L., Ong, Q., Paley, S., Schroder, I., Shearer, A. G., Subhraveti, P., Travers, M., Weerasinghe, D.*,* Weiss, V., Collado-Vides, J., Gunsalus, R. P., Paulsen, I., Karp, P. D.EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res.* **41,** D605–D612 (2013).

84. Kahramanoglou, C., Seshasayee, A. S. N., Prieto, A. I., Ibberson, D., Schmidt, S., Zimmermann, J., Benes, V., Fraser, G. M. & Luscombe, N. M. Direct and indirect effects of H-NS and Fis on global gene expression control in Escherichia coli. *Nucleic Acids Res.* **39,** 2073–91 (2011).

85. Prieto, A. I., Kahramanoglou, C., Ali, R. M., Fraser, G. M., Seshasayee, A. S. N. & Luscombe, N. M. Genomic analysis of DNA binding and gene regulation by homologous nucleoid-associated proteins IHF and HU in Escherichia coli K12. *Nucleic Acids Res.* **40,** 3524–37 (2012).

86. Rhee, H. S. & Pugh, B. F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147,** 1408–19 (2011).

87. Cole, J. N., Barnett, T. C., Nizet, V. & Walker, M. J. Molecular insight into invasive group A streptococcal disease. *Nat. Rev. Microbiol.* **9,** 724–736 (2011).

88. Kennedy-Darling, J. & Smith, L. M. Measuring the formaldehyde Protein-DNA cross-link reversal rate. *Anal. Chem.* **86,** 5678–81 (2014).

89. Hoa, T. T., Tortosa, P., Albano, M. & Dubnau, D. Rok (YkuW) regulates genetic competence in Bacillus subtilis by directly repressing comK. *Mol. Microbiol.* **43,**

15–26 (2002).

90. Carapetis, J. R., Steer, A. C., Mulholland, E. K. & Weber, M. The global burden of group A streptococcal diseases. *Lancet Infect. Dis.* **5,** 685–694 (2005).

91. Young, M. H., Aronoff, D. M. & Engleberg, N. C. Necrotizing fasciitis: pathogenesis and treatment. *Expert Rev. Anti. Infect. Ther.* **3,** 279–294 (2005).

92. Aziz, R. K. & Kotb, M. Rise and persistence of global M1T1 clone of Streptococcus pyogenes. *Emerg. Infect. Dis.* **14,** 1511–7 (2008).

93. Sumby, P., Whitney, A. R., Graviss, E. A., DeLeo, F. R. & Musser, J. M. Genome-wide analysis of group a streptococci reveals a mutation that modulates global phenotype and disease specificity. *PLoS Pathog.* **2,** e5 (2006).

94. Dalton, T. L. & Scott, J. R. CovS inactivates CovR and is required for growth under conditions of general stress in Streptococcus pyogenes. *J. Bacteriol.* **186,** 3928–37 (2004).

95. Levin, J. C. & Wessels, M. R. Identification of csrR/csrS , a genetic locus that regulates hyaluronic acid capsule synthesis in group A Streptococcus. *Mol. Microbiol.* **30,** 209–219 (1998).

96. Heath, A., DiRita, V. J., Barg, N. L. & Engleberg, N. C. A two-component regulatory system, CsrR-CsrS, represses expression of three Streptococcus pyogenes virulence factors, hyaluronic acid capsule, streptolysin S, and pyrogenic exotoxin B. *Infect. Immun.* **67,** 5298–305 (1999).

97. Kansal, R. G., Datta, V., Aziz, R. K., Abdeltawab, N. F., Rowe, S. & Kotb, M. Dissection of the Molecular Basis for Hypervirulence of an In Vivo–Selected Phenotype of the Widely Disseminated M1T1 Strain of Group A Streptococcus Bacteria. *J. Infect. Dis.* **201,** 855–865 (2010).

98. Campbell, E. A., Korzheva, N., Mustaev, A., Murakami, K., Nair, S., Goldfarb, A. & Darst, S. A. Structural Mechanism for Rifampicin Inhibition of Bacterial RNA Polymerase. *Cell* **104,** 901–912 (2001).

99. Anders, S. & Huber, W. Differential expression analysis for sequence count data. doi:10.1186/gb-2010-11-10-r106

100. McIver, K. S. & Myles, R. L. Two DNA-binding domains of Mga are required for virulence gene activation in the group A streptococcus. *Mol. Microbiol.* **43,** 1591–1601 (2002).

101. Ribardo, D. A. & McIver, K. S. Defining the Mga regulon: comparative

transcriptome analysis reveals both direct and indirect regulation by Mga in the group A streptococcus. *Mol. Microbiol.* **62,** 491–508 (2006).

102. Terao, Y., Kawabata, S., Kunitomo, E., Murakami, J., Nakagawa, I. & Hamada, S. Fba, a novel fibronectin-binding protein from Streptococcus pyogenes, promotes bacterial entry into epithelial cells, and the fba gene is positively transcribed under the Mga regulator. *Mol. Microbiol.* **42,** 75–86 (2008).

103. Hondorp, E. R. & McIver, K. S. The Mga virulence regulon: infection where the grass is greener. *Mol. Microbiol.* **66,** 1056–1065 (2007).

104. Hondorp, E. R., Hou, S. C., Hempstead, A. D., Hause, L. L., Beckett, D. M. & Mciver, K. S. Characterization of the Group A Streptococcus Mga Virulence Regulator Reveals a Role for the C-terminal Region in Oligomerization and Transcriptional Activation. *Mol. Microbiol.* **83,** 953–967 (2012).

105. Hondorp, E. R., Hou, S. C., Hause, L. L., Gera, K., Lee, C.-E. & Mciver, K. S. PTS Phosphorylation of Mga Modulates Regulon Expression and Virulence in the Group A Streptococcus. *Mol. Microbiol.* **88,** 1176–1193 (2013).

106. Mciver, K. S., Thurman, A. S. & Scott, J. R. Regulation of mga Transcription in the Group A Streptococcus: Specific Binding of Mga within Its Own Promoter and Evidence for a Negative Regulator. *J. Bacteriol.* **181,** 5373–5383 (1999).

107. Baker, B. M., Creamer, T. P., Piepenbrink, K. H., Lucius, A. L., Juárez, A., Bravo, A., Solano-Collado, V., Hüttener, M. & Espinosa, M. MgaSpn and H-NS: Two Unrelated Global Regulators with Similar DNA-Binding Properties. **3,** (2016).

108. Harwood, C. & Cutting, S. *Molecular biological methods for Bacillus*. (Wiley, 1990).

109. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).