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
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

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Modes of H-NS Transcription Mediation</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>XATAC-seq Method</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>Tagmentation</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td>Replicate Correlations</td>
<td>14</td>
</tr>
<tr>
<td>5</td>
<td>Quantitative Footprint Evaluation</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>XATAC-seq Footprinting Captures Transcription Factor Binding Sites</td>
<td>16</td>
</tr>
<tr>
<td>7</td>
<td>Promoter Enrichment</td>
<td>16</td>
</tr>
<tr>
<td>8</td>
<td>Recapitulation of EPODS</td>
<td>18</td>
</tr>
<tr>
<td>9</td>
<td>Timeline of XATAC-seq and IPOD</td>
<td>19</td>
</tr>
<tr>
<td>10</td>
<td>DNA-only ATAC Control</td>
<td>20</td>
</tr>
<tr>
<td>11</td>
<td>Effects of Tagmentation Time</td>
<td>21</td>
</tr>
<tr>
<td>12</td>
<td>Effects of Reverse Crosslinking</td>
<td>22</td>
</tr>
<tr>
<td>13</td>
<td>H-NS Correlation</td>
<td>23</td>
</tr>
<tr>
<td>14</td>
<td>Rok of B. subtilis</td>
<td>24</td>
</tr>
<tr>
<td>15</td>
<td>Effects of AT Content on Binding in Untreated and</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rifampicin-Treated Cells</td>
<td>28</td>
</tr>
<tr>
<td>16</td>
<td>Virulence Gene Binding</td>
<td>29</td>
</tr>
<tr>
<td>17</td>
<td>RNA-seq Differential Expression Venn Diagram</td>
<td>30</td>
</tr>
<tr>
<td>18</td>
<td>GAS POL Comparison</td>
<td>30</td>
</tr>
<tr>
<td>19</td>
<td>Differential Expression vs. Differential Binding</td>
<td>32</td>
</tr>
<tr>
<td>20</td>
<td>Gene binding vs. Expression</td>
<td>33</td>
</tr>
<tr>
<td>21</td>
<td>Mass Spectroscopy Probe Design</td>
<td>34</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 1: Correlation of XATACh-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.
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 \textit{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 \textit{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 \(\mu\)m-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
implanted. 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. *E. coli* has around 450 nucleic structural domains, estimated to be between 10kb and 100kb, with variable boundaries, distributed sporadically along the chromosome. Highly transcribed genes appear to be involved in defining domain boundaries by spatially isolating DNA regions and restricting the diffusion of supercoiling.

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* and *Bacillus subtilis*, 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. 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\(^\text{13}\). ATAC-seq takes advantage of a hyperactive mutant of the bacterial Tn5 transposase (Tnp)\(^\text{14}\). 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 \textit{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\textsuperscript{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\textsuperscript{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\textsuperscript{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\textsuperscript{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\textsuperscript{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 \textit{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\(^\text{25}\). 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\(^\text{26}\). 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\(^\text{27}\). At low concentrations, HU increases DNA flexibility over short regions; at high concentrations, it increases DNA stiffness and rigidity\(^\text{28}\). 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\(^\text{29}\). Finally, HU appears to play a role in initiating DNA replication\(^\text{30}\).

1.2.2 IHF

Integration host factor (IHF) is one of the most abundant sequence-specific binding proteins in *E. coli\(^\text{15}\). 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
\(\sigma^{54}\) 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 \textit{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 \textit{E. coli}) protein common to enteric bacteria,
particularly *E. coli* and its close relatives\textsuperscript{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\textsuperscript{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\textsuperscript{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\textsuperscript{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\textsuperscript{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\textsuperscript{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\textsuperscript{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\textsuperscript{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)\textsuperscript{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\textsuperscript{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\textsuperscript{41,66–70}. Additionally, it has been implicated with widespread repression of intragenic transcription, thereby preventing spurious RNA synthesis\textsuperscript{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. 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. 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 environmental signals by adjusting the pool of H-NS-like dimers.

Other H-NS-like protein families include the Lsr2 family in *Mycobacteria* and the mvaT family in *Pseudomonas*. 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.](image)

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\textsuperscript{78}.

The transposition step involves transposase binding at the target site, a transposase-mediated nucleophilic attack on the phosphodiester bonds along the backbone of both DNA strands, and transposase release, followed by nick repair (Fig. 3)\textsuperscript{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)\textsuperscript{14}. The Tn5 transposase has an insertion preference (A-G-N-T/T/A/T-A/G-A-N-T/C) that is mirrored to a small degree in the bias of transposition events\textsuperscript{70–81}. 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\textsuperscript{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.
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 XATA-C-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 XATA-C-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](image)

**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* XATA-C-seq, 13% of footprints align with transcription factor binding sites compiled in the model organism database EcoCyc (Fig 6)\(^3\). 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 CAPTURES 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. 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

<table>
<thead>
<tr>
<th>Nucleoid-Associated Protein</th>
<th>Pearson Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>HU</td>
<td>0.21</td>
</tr>
<tr>
<td>IHF</td>
<td>-0.10</td>
</tr>
<tr>
<td>FIS</td>
<td>-0.07</td>
</tr>
<tr>
<td>H-NS</td>
<td>0.60</td>
</tr>
</tbody>
</table>

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\textsuperscript{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 than 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\textsuperscript{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 Tagmentation Bias Assessment

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](image)

**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 ФМ1Т1Z phage, which encodes the virulence factor streptodornase and differentiates the М1Т1 serotype from other closely related М1 strains (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 XATACT-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:
\[ p/t = 0.00379e^{0.06317T} \]

Where T is the temperature in degrees Celsius, t the time in minutes, and p the percentage of crosslinks reversed. This suggests that at 95°C, 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.

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).

![E. coli ATAC-seq vs. H-NS ChIP-seq](image)

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

![B. subtilis](image)

**Figure 14**: Rok of *B. subtilis*. Relationship between XATA-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. 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. 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. 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.

<table>
<thead>
<tr>
<th>Total BLAST Scores</th>
<th>H-NS</th>
<th>Rok</th>
<th>MvaT</th>
<th>Lsr2</th>
</tr>
</thead>
<tbody>
<tr>
<td>H-NS</td>
<td>276</td>
<td>0</td>
<td>52.3</td>
<td>0</td>
</tr>
<tr>
<td>Rok</td>
<td>0</td>
<td>381</td>
<td>23.9</td>
<td>0</td>
</tr>
<tr>
<td>MvaT</td>
<td>52.3</td>
<td>23.9</td>
<td>249</td>
<td>26.2</td>
</tr>
<tr>
<td>Lsr2</td>
<td>0</td>
<td>0</td>
<td>26.2</td>
<td>221</td>
</tr>
</tbody>
</table>
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\(^9^0\). M1T1 is the GAS serotype most frequently associated with severe infections, which are often difficult to treat with antibiotics and may require surgical intervention\(^9^1,9^2\). 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\(^9^3\). In total, CovRS is responsible for regulation of approximately 15% of the genes in GAS\(^9^4\) – CovS is a membrane-associated histidine kinase that controls the phosphorylation state of CovR, the response regulator of the system\(^9^4\). 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\(^8^7,9^5,9^6\). SIC repression appears to be covS-dependent\(^9^7\).
We have applied XATAc-seq to study the changes in protein binding associated with the transition from non-pathogenic M1T1 GAS to the hypervirulent \( \text{covR} \) deletion mutant (\( \Delta \text{covR} \)) as well as another hypervirulent mutant containing a \( \text{covS} \) 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. 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 rifampicin 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.
**Virulence Gene Binding in GAS Mutants**

![Gene XATAC-seq Binding](image)

**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. 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 considerably 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.

<table>
<thead>
<tr>
<th>Virulence Factor</th>
<th>Gene</th>
<th>Fold-Change in Expression vs. AP</th>
<th>Adjusted p-value</th>
<th>Fold-Change in Expression vs. covR</th>
<th>Adjusted p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>M Protein</td>
<td>emm</td>
<td>0.27</td>
<td>3.00E-05</td>
<td>0.26</td>
<td>2.00E-06</td>
</tr>
<tr>
<td>Hyaluronase</td>
<td>hasA</td>
<td>0.01</td>
<td>4.90E-93</td>
<td>0.01</td>
<td>1.07E-70</td>
</tr>
<tr>
<td>Hyaluronase Capsule</td>
<td>hasB</td>
<td>0.06</td>
<td>4.08E-08</td>
<td>0.08</td>
<td>4.64E-07</td>
</tr>
<tr>
<td>Hyaluronase Capsule</td>
<td>hasC</td>
<td>0.01</td>
<td>3.11E-111</td>
<td>0.02</td>
<td>6.17E-95</td>
</tr>
<tr>
<td>Exotoxin J</td>
<td>speJ</td>
<td>0.19</td>
<td>1.44E-10</td>
<td>0.26</td>
<td>1.84E-07</td>
</tr>
<tr>
<td>Streptolysin O</td>
<td>slo</td>
<td>0.02</td>
<td>1.32E-112</td>
<td>0.03</td>
<td>9.52E-47</td>
</tr>
<tr>
<td>ADP-Ribosyltransferase</td>
<td>spyA</td>
<td>0.03</td>
<td>5.84E-62</td>
<td>0.04</td>
<td>7.23E-56</td>
</tr>
<tr>
<td>Streptokinase A</td>
<td>ska</td>
<td>0.01</td>
<td>1.56E-231</td>
<td>0.03</td>
<td>7.23E-34</td>
</tr>
<tr>
<td>NAD Glycohydrolase</td>
<td>nga</td>
<td>0.04</td>
<td>3.82E-67</td>
<td>0.03</td>
<td>5.25E-90</td>
</tr>
<tr>
<td>IgG endopeptidase</td>
<td>ideS</td>
<td>0.01</td>
<td>2.06E-58</td>
<td>0.04</td>
<td>1.69E-16</td>
</tr>
<tr>
<td>interleukin 8 protease</td>
<td>spyCEP</td>
<td>0.27</td>
<td>2.60E-02</td>
<td>1.00</td>
<td>-</td>
</tr>
<tr>
<td>Alpha2-Macroglobulin-Binding Protein</td>
<td>grab</td>
<td>0.52</td>
<td>6.10E-06</td>
<td>1.00</td>
<td>-</td>
</tr>
<tr>
<td>Collagen-Like Surface Protein</td>
<td>sclA</td>
<td>0.07</td>
<td>8.94E-08</td>
<td>0.13</td>
<td>1.70E-04</td>
</tr>
<tr>
<td>Fibronectin-Binding Protein</td>
<td>fbaA</td>
<td>0.28</td>
<td>3.40E-02</td>
<td>1.00</td>
<td>-</td>
</tr>
<tr>
<td>Streptococcal inhibitor of Complement</td>
<td>sic</td>
<td>0.14</td>
<td>2.47E-08</td>
<td>0.20</td>
<td>1.29E-05</td>
</tr>
<tr>
<td>CSa Peptidase</td>
<td>scpA</td>
<td>0.15</td>
<td>2.30E-04</td>
<td>0.17</td>
<td>4.50E-04</td>
</tr>
</tbody>
</table>
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\textsuperscript{47}. This implies that global binding of protein in GAS, on the whole, has the same repressive effect as that of H-NS in \textit{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.

\textbf{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.
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.

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

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. 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. For this reason, it is currently classified as a specific regulator. Despite this, Mga lacks a consensus binding sequence. 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, \textit{in vitro} studies have shown that Mga binds regions of \textit{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 XATA-seq findings, lends itself well to the hypothesis that Mga acts in a role more similar to that of H-NS-family proteins\(^9\). 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 37°C with shaking.

*Bacillus subtilis* cells were grown in rich CH media at 37°C 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 37°C with shaking.

*Clostridium ljungdahlii*, were cultured in carbon monoxide and high-fructose media at 27°C 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 37°C, 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 4°C. 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 25μg/mL Rifampicin and incubated at 37°C 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:

- 98°C 2 min.
- 30x: 98°C 25 sec.
- 43°C 15 sec.
- 72°C 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 H2O.

**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 4°C. 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 4°C. 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 reverse-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 with 1.8X volume AMPure bead solution.
2) Sample was incubated with 1 ul proteinase K at 65°C 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. 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. 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. 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


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


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


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,


