Functional and Mechanistic Characterization of Heterochromatin-like Domains in Bacteria

by

Haley Minami Amemiya

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Cellular and Molecular Biology)
in the University of Michigan
2021

Doctoral Committee:

Assistant Professor Peter L. Freddolino, Chair
Professor Gyorgyi Csankovszki
Professor Ursula Jakob
Associate Professor Xiaoxia Lin
Professor Lyle A. Simmons

Haley Minami Amemiya

hamemiya@umich.edu

ORCID iD: 0000-0002-2610-4436

Dedication

To Mr. Eric Thiel: There is no greater gift to give to another person than empowering them to believe in themselves. Thank you for giving that to me and so many others. Rest in peace.

To the Young Scientist: you got this.

Acknowledgements

I would not be the scientist I am today without the support and mentorship of many incredible individuals. I addition, the work presented here would not have come to fruition without the unconditional friendship and love that I am so fortunate to receive.

My love of research and science stemmed from my experience at the University of Washington working with two of the best scientists, people, and mentors I know, Dr. Bonita Brewer and Dr. M.K. Raghuraman (Raghu). Thank you for everything, especially for pairing me with Dr. Joseph Sanchez, who always pushed me to think for myself but gave me the tools to do so. To Gina Alvino, queen of NY, thank you for encouraging me to set high standards for myself and never letting me feel like I couldn't reach them. Thank you, Summer, Kelsey and Liz, who supported me in every meeting and celebration. To Mackenzie Croy and Madison Miller, it is so inspiring to see all that you do today.

To the Heist, my unwavering group of friends, thank you for always putting a smile on my face and accepting me for who I am, never passing up a chance for guac, tequila, or "The Cookie", and making life so awesome. Each of you has put your brain and heart into caring for others, it is amazing to see. I love you! Thank you to Pure Barre for exposing me to great people and maintaining my sanity.

To the group of the most amazing scientists I've had the honor of being peers with, Taylor Nye, Emily Yaroz, Marlena Bannick, Bonnie Cheng, Zena Lapp, Lindsay Moritz, Brooke Wolford, Kelly Sovacool, Anne Hakim, Bouchet Buds, and many more, thank you for being someone to look up to every day. Thank you to the many senior scientists and staff who carved out time and space to support me. Thank you Yashar and Theo for cheering me on.

Thank you to Peter Freddolino, the only person I know who can summarize 10 years of work in one breath, I am so grateful for the opportunities you let me grasp to chase my dreams. To the rascals of the Freddolino lab, Christine and Yulduz, thank you for making COVID-19 lab work bearable and fun. Thank you Lyle, Ursula, Gyorgyi, and Nina for your support and guidance.

To my anchors in life, my family, there are not enough words I can say to truly capture my gratitude. Mom and Dad, you are the most hardworking and devoted people I know to any craft you take on. My sister and friend, Jamie, my mentor since day one, thank you for encouraging me to advocate for myself and use my voice. To Elena, the only person who can take my father's jokes, thank you for always being open to adventures. Gus and Ada, you are the best dogs on the planet.

To my amazing partner in crime, Michael Schmidt, thank you for not contaminating my LB... and being the most supportive partner, I could ever ask for. I love you so much!

Lastly, to my Grandmother Florence Lewis: I will continue to live as best I can to the words you lived by, "To thine own self be true."

Table of Contents

Dedication……ii

Acknowledgements……iii

List of Tables…… vii

List of Figures…… viii

Abstract…… x

Chapter 1…… 1

A New Perspective on Bacterial Architecture..... 1

Abstract…… 1

Introduction…… 1

The organized bacterial nucleoid…… 4

Bacterial transcription depends on a single RNA polymerase..... 6

Bacterial chromatin influences transcription..... 6

Nucleoid associated proteins mediate the formation of bacterial chromatin..... 7

Understanding the state of the cell: tools to profile genome architecture and regulation. ……13

Conclusions and future perspectives……14

References……17

Chapter 2…… 29

Dynamic Landscape of Protein Occupancy Across the *Escherichia coli* Chromosome…… 29

Abstract…… 29

Introduction…… 30

Results…… 33

Discussion…… 57

Materials and Methods…… 60

Acknowledgments……75

Supplementary Text……..…… 75
Supplementary Figures……..……...79
Supplementary Tables……..…… 84
References……..……... 87
Chapter 3……..…… 93
Distinct Heterochromatin-like Domains Promote Transcriptional Memory and Silence Parasitic Genetic Elements in Bacteria……..…..... 93
Abstract……..…… 93
Introduction……..…… 94
Results……..…… 96
Discussion……..……...120
Materials and Methods……..…... 122
Acknowledgments……..……..129
Supplementary Figures and Tables……..…… 130
Supplementary Text……..…….. 139
References……..……..140
Chapter 4……..…… 146
Interplay of Hfq and Polyphosphate in Bacterial Heterochromatin Formation…… 146
Abstract and Introduction……..…… 146
Results……..……147
Discussion……..…….. 158
Materials and Methods……..……..158
Supplemental Figures and Tables……..…… 168
References……..……..174
Chapter 5……..…… 177
Concluding Remarks and Future Research……..... 177
Introduction……..…… 177
Investigating the impact of methylases on EPOD maintenance.....…… 178
Elucidating the mechanism underlying transcriptional memory response in exotic carbon source exposure. ……..…… 179
Identify the proteins and post-translational modifications that define the structure of EPODs……..……...… 180

Investigate the mechanism of EPODs in the silencing of harmful genetic elements. ……..182
Conclusion……..183
References……..185
Appendix……..187
The ENCODE Blacklist: Identification of Problematic Regions of the Genome.....187
Abstract……187
Introduction……..187
Results……..188
Discussion……..196
Materials and Methods……..197
References……..199

List of Tables

TABLE

S2.1: Mass spectrometry identified peptide counts showing abundances of proteins pulled down by biotinylated bait DNA from the *sdaC* promoter region, after pruning of likely contaminants (see Methods for details)…….....……..85
S2.2: Summary of EPOD characteristics across experimental conditions…….....…....86
S3.1: HMM class enrichments…….....……136
S3.2: MDS42 regions containing prophages…….....……136
S3.3: Strains used in this study…….....……..137
S3.4: Plasmids used in this study…….....……..137
S3.5: Primers used in this study…….....138
S4.1: Strains used in this study…….....……171
S4.2: Plasmids used in this study…….....……172
S4.3: Primers used in this study…….....……173

List of Figures

FIGURE

1.1: Key proteins mediate structure across domains of life……2
1.2: Nucleoid associated proteins mediate a variety of DNA conformations and are abundant in different stages of growth……..……4
2.1: Schematic of IPOD-HR technology and detection of context-dependent binding by transcription factor PurR……34
2.2: IPOD-HR profiles reveal rich high-resolution occupancy dynamics and large-scale structural features across the chromosome……..…….37
2.3: IPOD-HR profiles reveal global binding activity of known transcription factors and sigma factors...……40
2.4: Experimental identification of the protein bound to a novel occupancy peak upstream of the *sdaC* promoter……44
2.5: Genome-wide de novo discovery of sequence specificity motifs for actively bound transcription factors……48
2.6: EPODs define stable genomic structures and are associated with many distinct features……...53
2.7: EPODs are statistically enriched for genes in specific functional categories……56
S2.1: Effect of peak calling threshold on coverage and enrichment of known transcription factor binding sites (TFBSs)……..79
S2.2: Interplay of H-NS occupancy, EPOD locations, and transcription…….....……...80
S2.3: Identification of RNA polymerase vs. non-RNA polymerase protein occupancy……81
S2.4: Overlaps of EPOD sets resulting from different calling methods…….82
S2.5: Effects of rifampin treatment on protein occupancy of a highly transcribed region……83
S2.6: Effects of rifampin treatment on protein occupancy of a transcriptionally silent region…..……..……...84
3.1: EPODs are highly robust across growth conditions……....99
3.2: Loss of nucleoid associated proteins (NAPs) leads to changes in EPODs……...102
3.3: Changes in EPODs are induced in specific conditions……..107
3.4: EPODs mediate transcriptional memory……..……109
3.5: Nucleoid associated proteins contribute to protein occupancy at EPODs that contain prophages……111
3.6: Loss of Fis and Hfq is lethal in a prophage-dependent manner……...114
3.7: IPODHR in Bacillus subtilis reveals Rok-bound and SMC-bound domains……118
S3.1: Pathway analysis of EPODs across WT conditions……130

S3.2: Loss of NAPs results in increases in RNA polymerase occupancy and decreases in overlapping EPODs……..……..131
S3.3: Changes in protein occupancy across the genome…….132
S3.4: Deletion of hns and stpA impact EPODs across the genome……133
S3.5: H-NS and StpA mediate silencing of the idn operon……...134
S3.6: Growth deficiency of Δfis Δhfq cells……...135
4.1: Loss of ppk leads to an induction of prophages and mobile elements and sensitivity to DNA damaging agents……..…….148
4.2: polyP and Hfq show evidence of epistasis…….……..152
4.3: Loss of polyphosphate kinase impacts Hfq binding across prophage regions……154
4.4: Polyphosphate facilitates distinct oligomeric species of Hfq hexamers……..157
S4.1: Hfq RNA chaperone targets are not impacted by loss of ppk…….……168
S4.2: DNA damage epistatic response is specific to ppk and hfq……169
S4.3: Loss of ppk does not change the amount of Hfq protein in the cell……..170
5.1. Workflow for assessing presence of PTMs and accessory factors……..181
5.2. Examining the impact of genome location in EPOD formation…….....183
A.1: Blacklist regions are tightly distributed across the chromosome and sequester high read mapping signals……..190
A.2: Blacklist regions account for a significant portion of ChIP-seq reads, are driven by artifacts in genome assemblies, and removal of these regions is essential to removing noise in genomics assays…….……..191
A.3: Justification of thresholds for automated blacklist generation……..192
A.4: Comparison across different “blacklists”……..194

Abstract

Every organism faces the challenge of organizing immense amounts of genetic information into a small physical space that is the cell. In eukaryotes, this process is facilitated by histones that wrap DNA into small units. While it has been historically assumed that bacteria do not have an organized genome, increasing evidence implicates a robust structure that enables bacteria to quickly cope with a variety of environmental pressures. The organization and regulation of DNA is incredibly important to engineer bacteria for biotechnological purposes and to understand bacteria that cause disease. However, while bacteria impact almost every aspect of human life, we do not fully understand their genomes. In this thesis I investigated genome organization with a specific focus in bacteria. To improve our understanding of bacterial genomes, I helped design a high-throughput tool, *in-vivo* protein occupancy display at high-resolution (IPOD-HR), that allows resolution of how proteins bind across the whole length of a bacterial chromosome. By applying this tool to a number of different bacteria, we discovered conserved areas of the genome that are densely bound by proteins but are transcriptionally silent – similarly to heterochromatin in eukaryotes. I show that these regions, termed extended protein occupancy domains (EPODs), have functional roles in bacteria that enable them to use new carbon sources for energy and provide an immune defense against viruses in *Escherichia coli* (*E. coli*). I show that EPODs are occupied by nucleoid associated proteins (NAPs). By performing deletions of single NAPs, I identified the key NAPs that bind to specific regulons. In *E. coli*, I find that EPODs silence a number of metabolic pathways and toxic prophages. I induced changes of particular EPODs by exposing cells to exotic carbon sources and find that EPODs mediate a transcriptional memory affect, where upon a second exposure to an exotic carbon source mounts a faster growth rate and de-repression of genes required for metabolism. In addition, I show one essential role of the formation of EPODs by NAPs in *E. coli* is to silence harmful genetic elements that have integrated into the

genome, such as mobile elements and prophages, that can be potentially toxic to the cell. I define novel prophage silencers, Hfq and Fis that are required for silencing specific prophages. In collaboration with the Jakob Lab, I employed biochemistry, genetics, and bioinformatics and discovered that Hfq binds with a poly-anion, polyphosphate (polyP), to DNA to silence prophages. Biochemical results suggest a model in which polyphosphate acts as an Hfq chaperone in order to permit appropriate silencing at EPODs. These results provided the first evidence that polyP might act in DNA damage control by either directly or indirectly suppressing the expression of genetic mobile elements and prophages, and a mechanism by which bacterial heterochromatin enables regulation during times of stress. Ultimately, my work defines the importance of genome organization in bacteria and provides a scaffold for further investigation into mechanisms underlying the establishment heterochromatin-like domains in bacteria.

Chapter 1

A New Perspective on Bacterial Architecture

Abstract

Genome architecture has proven to be incredibly important in determining gene regulation across almost all domains of life. While many of the key components and mechanisms of eukaryotic genome organization have been described, bacterial DNA organization has only begun to be recognized. Only until the early 2000's has it been appreciated that bacteria may have a structured genome, and the tools to better understand structure and function are limited. Much of what we know about bacterial organization relies on *in vitro* characterization of nucleoid associated proteins (NAPs), however, it remains a mystery as to how these findings translate *in vivo*. Additionally, tools to study individual NAPs have also been incredibly cumbersome, as they promiscuously bind nucleic acids and have overlapping binding sites. Here, I summarize the importance of genome organization in gene regulation and the state of the field of bacterial genome architecture.

Introduction

Every cell must solve an incredible challenge: to organize negatively charged genetic material, DNA, that is orders of magnitude longer than the width of the cell. The DNA must be compacted and neutralized, but accessible to maintain proper physiological states and easily changed if environmental stimuli require it. In eukaryotes, the compaction of the DNA is mediated by histones- basic proteins that in an octamer form tight structural units of DNA called nucleosomes[1], these nucleosomes then wrap to

form chromatin fibers[2,3]. There are two broad categories of chromatin that largely relate to the accessibility of DNA to RNA-polymerase binding and transcription: euchromatin which is loosely packed and accessible for transcription vs. heterochromatin which is inaccessible and silent[4] (**Fig. 1.1A,B**). Histones, chromatin-associated proteins, and amino-terminal modifications form the 'histone code' that in turn impacts chromatin structure and gene regulation[2,3] (**Fig. 1.1A,B**). The flexibility of this code can be appreciated when considering post-translational modifications to histones, which can both repress transcription or promote initiation in accessible promoter regions[5] (**Fig. 1.1A,B**). In addition to maintaining organization, this genome architecture facilitates cell-type memory as tissue-specific cells pass information to daughter cells. Cell fate decisions are largely determined by gene expression and transcription, that may change in response to environmental stimuli[6]. This passage of information, often termed 'transcriptional memory' is critical for the viability of the organism[6,7]. In humans, changes in chromatin lead to global genome misregulation and instability have been linked to cancer[8–11], implicating chromosome structure as an essential unit of genome maintenance and health.

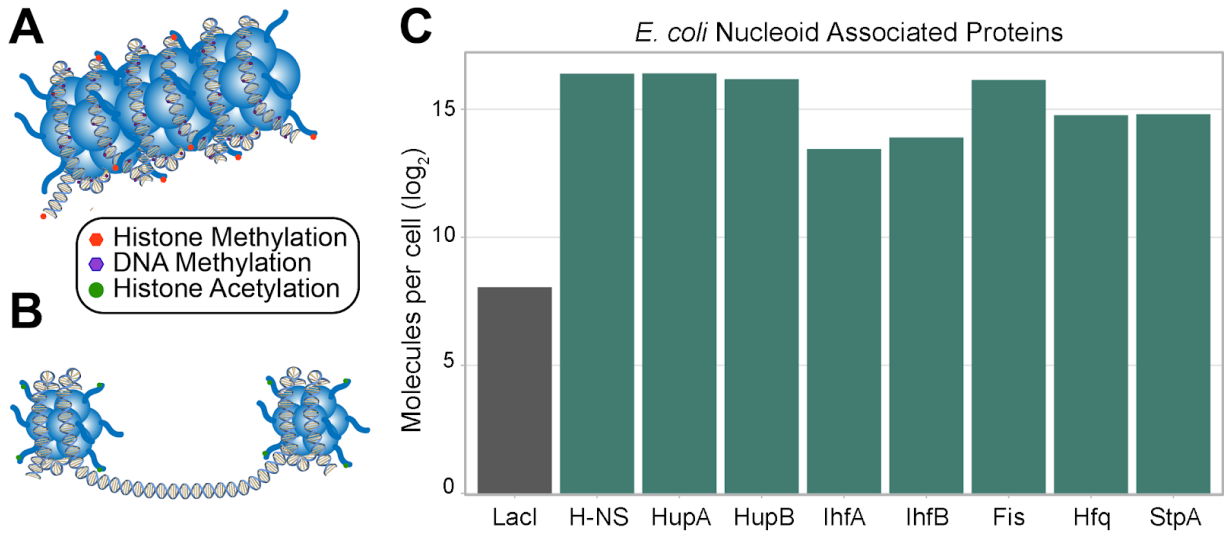

Figure 1.1: Key proteins mediate structure across domains of life. (A) A model of heterochromatin in eukaryotes exhibiting both histone methylation and DNA methylation, which facilitates the assembly of heterochromatin structures[194,195]. **(B)** In contrast, a model of euchromatin exhibits histone acetylation which diminishes electrostatic affinity between histones and can lead to accessibility of DNA For

transcription[196–198]. **(C)** Molecules per cell (log_2) of nucleoid associated proteins (NAPs; green) and the transcription factor LacI (grey) in *E. coli* grown in rich media[199].

In the case of bacteria, DNA organization has only begun to be described, as it was previously believed that bacteria have a relatively accessible genome that exists in a membrane-free nucleoid[12–14] (**Fig. 1.2**). With advances in technology, it was discovered that bacteria maintain a discrete nucleoid shape that changes through different growth phases or conditions[15–17] (**Fig. 1.2**). The tools that exist to study genome architecture in eukaryotes cannot simply be applied to bacteria, so much of our understanding of the genome organization in bacteria, and how it may impact gene regulation, is incredibly limited. Overall compaction of the DNA has been known to be facilitated by highly abundant nucleoid associated proteins (NAPs)[14,18] (**Fig. 1.1C, Fig. 1.2**), however their broad binding specificity and overlapping binding regions make it difficult to profile a single NAP's contribution to gene regulation. Even comprehensively studied NAPs such as H-NS remain a mystery in terms with how they interact with DNA in the cell. Here, we will explore the current state of the field of bacterial genome organization: the many modes by which genome organization impacts gene regulation in bacteria, the key proteins that mediate chromosomal organization, and the tools that exist to better understand genome regulation.

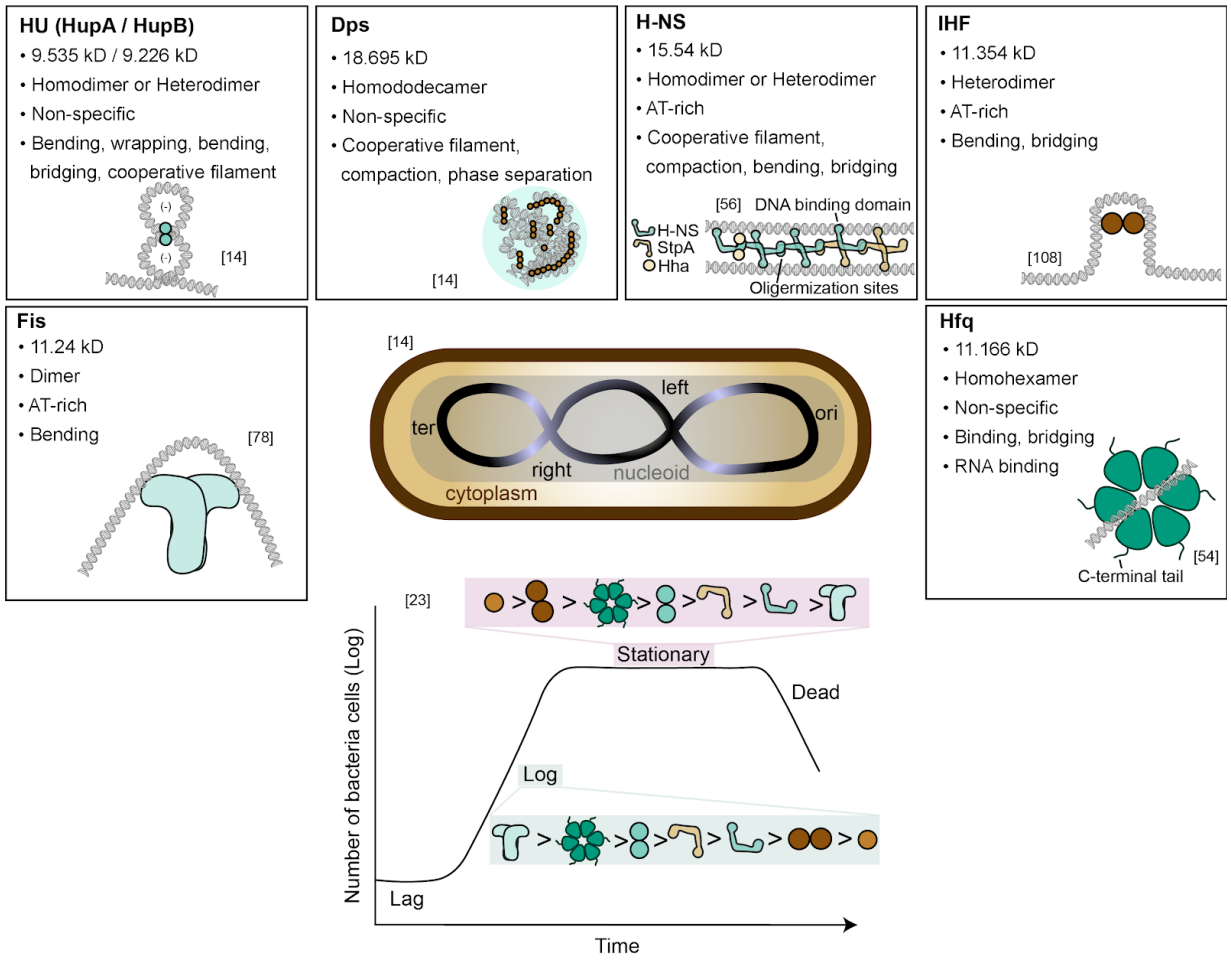

Figure 1.2: Nucleoid associated proteins mediate a variety of DNA conformations and are abundant in different stages of growth. Overview of highly abundant NAPs and their relative abundance in the different stages of growth in *E. coli*. Single subunit kDs, binding associations, and binding preferences were referenced from [199], binding capabilities (filaments, bendings, etc.) were referenced from [14,199]. The H-NS binding depicted is an example of a bridged DNA filament formed between Hha, StpA, and H-NS.

The organized bacterial nucleoid.

The hallmark of what defines a prokaryote vs a eukaryote is the absence of membrane bound organelles. Instead, the (in most cases) single circular chromosome is contained within the nucleoid: an irregularly shaped, organized mixture of negatively charged DNA[14,17,19]. The *E. coli* chromosome contains one origin of replication and is organized into four macrodomains: Ori, Ter, Right and Left that mediate the separation

of sister chromatids during replication[20,21] (**Fig. 1.2**). The overall shape of the nucleoid is robust, and rearranges during stress, replication, and growth phase[15,16]. The determining factors of the overall shape of the nucleoid are nucleoid associated proteins (NAPs) and transcription[18,22], intertwining the importance of structure, proteins, and transcription for proper cell viability. Many parallels have been drawn between histones and NAPs, since NAPs have DNA binding domains, broad specificity, and the ability to compact DNA. However, the *in vivo* interactions NAPs have with DNA have not been completely resolved. In *E.coli*, there are around a dozen NAPs that interact with the DNA and other nucleic acids in various ways[13,14,18,23–25] (**Fig. 1.2**). While more deeply summarized in the following paragraphs, NAPs facilitate the formation of DNA bridges, filaments, bending, supercoiling, RNA mediated interactions, puncta, phase separated droplets, and impede RNA polymerase directly [14,26] . These interactions can promote and suppress transcription, however, in general NAPs are thought to be the major silencers of bacterial DNA[17,18]. NAPs are highly abundant in the cell, however their abundance may differ as the nucleoid responds to different conditions. For instance, the NAP Dps is highly abundant in stationary phase and becomes the main component of the nucleoid[23], sequestering iron and protecting DNA from damage[27,28] (**Fig. 1.2**). In contrast, transcription itself can impact the shape of the nucleoid, as transcription generates positive supercoils and can influence the binding of NAPs[14]. An experiment to beautifully exhibit the influence transcription has on nucleoid shape, is to treat cells with rifampin, an antibiotic that blocks and inhibits RNA polymerase. The inhibition of transcription initially results in a compaction of the nucleoid due to the reduction in expansion from transcription-translation interactions, and then an expansion occurs from ribosomes and chromosome mixing[29,30]. Together, NAPs, transcription, and DNA form the organized nucleoid, all influencing each other in the process of growth and stress. For the remainder of this thesis, the focus will be *E. coli*, as it is one of the hallmark model microbial organisms, harnessed for biotechnology purposes, and pathogenic.

Bacterial transcription depends on a single RNA polymerase.

While the central dogma is conserved across life, bacteria have a specific order of events to initiate transcription. Unlike eukaryotes which contain membrane bound organelles, prokaryotic transcription, translation, etc. occur in the cytoplasm of the cell and transcription relies on a single RNA polymerase (compared to the three present in eukaryotic organisms) [31,32]. This single RNA polymerase is composed of four subunits: two alpha[33], one beta and one beta' that bind with a number of sigma factors (σ factor) to form a complete RNA polymerase holoenzyme that can bind to a promoter sequence and initiate transcription[34]. During exponential growth, sigma 70 is the primary σ factor consisting of 60-95% of the total of factors, however in other stages of growth and different stressors (heat, osmotic stress, etc.), sigma 70 decreases[35–37]. The change in the σ factor composition of the cell impacts binding of the RNA polymerase holoenzyme, thus leading to large changes in the expression profile of the cell[38–40].

Bacterial chromatin influences transcription.

Historically, chromatin is defined as a complex of tightly packed DNA made up of histones, and found exclusively in eukaryotic cells[3]. However, new technologies have uncovered evidence that heterochromatin-like regions- areas that are densely bound by protein but transcriptionally silent- exist in bacteria[14,41,42]. If the definition of chromatin is expanded to capture regions of the chromosome that are densely packed and transcriptionally inert, it is important to begin to define mechanisms that maintain and regulate chromatin in bacteria. The structure of bacterial chromatin is largely defined by supercoiling and compaction of the DNA mediated by NAPs[19,29,43–46], however, much of what occurs *in vivo* still remains incomplete and will require adaptation of tools to capture 3D structure. The five modes by which bacterial chromatin impacts transcription (specifically thinking in the lens of how it may impact RNA polymerase binding) are summarized in [14], but will briefly be discussed here for its relevance: (1) occlusion of RNA polymerase binding: proteins bound to a promoter or

transcription start site that prohibit RNA polymerase touch down, (2) blocking RNA polymerase progression: RNA polymerase is able to bind and initiate transcription, but cannot proceed due to a protein roadblock, (3) DNA topology: (+) supercoiling is generated in front of an elongation complex and (-) supercoiling is generated behind; (-) supercoiling supports DNA unwinding and thus facilitates transcription initiation and inhibits termination, and the opposite is true for (+) supercoiling[47,48], (4) RNA-mediated silencing: Transcription factors that bind to nascent RNA transcripts can interfere with RNA polymerase termination, translocation, and pausing[49,50],(5) phase-separation: the formation of DNA condensates has been primarily shown in eukaryotes to control transcription[51], where DNA is compacted in droplets, however if phase separation mediates transcription in bacteria is unclear. Each one of these modes has been linked to NAPs, as almost all NAPs influencing DNA topology via supercoiling, however specific NAP roles are described in detail below.

Nucleoid associated proteins mediate the formation of bacterial chromatin.

NAPs mediate chromosomal structure and DNA compaction across bacterial species[18,52,53]. What they all have in common broad DNA binding specificity (usually preferring curved DNA and / or AT rich DNA) and high abundance in the cell[18] (**Fig. 1.2**). The majority form homodimers, with some, H-NS and StpA, binding together to form DNA filaments[14,18](**Fig. 1.2**)[14,18]. Crystal structures and *in vitro* experiments of NAPs have led to insights into some of the interactions between nucleic acids and proteins[54–56], however the method by which particular NAPs bind DNA *in vivo* has largely been uncharacterized. Increasing evidence has implicated NAPs serving important functions as regulators of horizontal gene expression and pathogenesis[57–62]. H-NS (Histone-like nucleoid structuring protein), one of the most well studied NAPs, and has been shown to silence horizontally acquired DNA[63–65]. H-NS and it's paralog StpA will be more deeply explored in the next sections. However, even as a comprehensively studied NAP, it still remains a mystery of how H-NS binds to DNA inside the cell and mediates silencing of foreign DNA. Even less is known about the

remaining NAPs of *E. coli*: Fis, HU, IHF, Dps, Hfq, Lrp, IciA, DnaA, Cbp (**Fig. 1.2**). Proteomic analysis suggests that H-NS, HU, IHF, and Fis activities may be influenced by posttranslational modifications, however, the mechanisms underlying that response remain unclear[66]. Here, I will briefly summarize what is known for the NAPs that make up the main components of the nucleoid (Fis, HU, IHF, Dps, H-NS) and the highly conserved RNA chaperone, Hfq.

Factor for inversion stimulation (Fis) was named and first identified for its role in inversion of the bacteriophage Mu[67,68], but has a broader role in organization and maintenance of nucleoid structure[69,70], primarily acting through binding as a homodimer to DNA or modulating gyrase and topoisomerase I[71]. Fis is one of the most abundant NAPs in the cell (>60,0000 copies per cell) during exponential growth and optimal conditions, however falls to drastically lower amounts during later stages of growth (<100 copies per cell) [23,72] (**Fig. 1.2**). Fis is largely autoregulated, and it's role in modulating gyrase and topoisomerase I is linked with its induction by high supercoiling levels in the cell[73]. Fis is made up of an α-helical core with four helices and an N-terminal domain that has a β-hairpin arm that facilitates DNA inversion[74–77]. As a homodimer, Fis has been shown to bend DNA to as large as a 90° bend, which stabilizes DNA looping, thus leading to compaction of DNA and effects on transcription[61,78,79]. Fis can have both inducing [80–82]and suppressing effects [83–85] on gene expression, which can largely depend on Fis intracellular abundance[71]. It is one of the major gene regulators of the cell, with 894 Fis associated regions across the *E. coli* genome[69]. The global binding of Fis negatively correlates with transcriptional propensity, measured by a randomly inserted reported construct across the genome[86]. Authors in [86] show that rather than impact transcriptional propensity by interactions at the promoter region, Fis acts on the integration site. In total, it is clear that Fis is a major gene regulator of the *E. coli* cell, impacts transcription through its DNA binding capacity, and thus can change the overall shape and gene expression of the cell. It remains unclear if Fis binds or interacts with other nucleoid associated proteins, however, Fis serves an overlapping role of silencing toxic regions of the genome compared to other NAPs[85]. As a general theme with all NAPs, it is becoming

increasingly clear that NAPs serve a variety of functions, many overlapping, to maintain overall cell viability and health.

Heat-unstable nucleoid protein (**HU**), the most abundant and highly conserved NAP[87], forms a heterodimer with subunits HupA and HupB encoded by *hupA* and *hupB*, respectively. Homodimers between each subunit can form as well[88]. Like many other NAPs, HU is known as a global regulator and organizer of the *E. coli* nucleoid. HU can bind linear dsDNA with low affinity and RNA, but prefers DNA forks, sharp bends, bulges, and kinks[14]. *In vitro*, HU determines polyamine DNA condensates, facilitating the formation to rod structures[89], however it is unclear if the same pattern would be observed *in vivo*. HU plays largely a repressive role as an accessory factor that regulates key pathways involved in replication initiation[90], stress response[91], the Gal respressome[92], and outer membrane maintenance[93]. HU's ability to bend the DNA and form higher order nucleoprotein complexes at promoters stabilizes dense structures that prohibit the ability of transcription initiation at that site[92,94,95]. There is no known inducer for HU, but similarly to Fis, HU is one of the major components of the nucleoid during exponential growth and lowers in content during later stages of growth[23]. HU remodels the nucleoid during growth phases and fosters the formation of a dense condensed core, hypothesized to facilitate coordinate gene regulation during times of growth and environmental changes[96]. One of HU's main modes of impacting gene expression comes from the introduction of negative supercoiling in the presence of topoisomerase I[97–99], or in some cases, stimulating topoisomerase I to remove negative supercoiling[100] (**Fig. 1.2**). The wide variety of regulatory roles, binding capacity and modes leaves an incomplete picture of how HU and the variety of dimers directly mediate nucleoid structure and regulation.

Integration host factor (IHF) similarly to HU forms a heterodimer with subunits IhfA and IhfB, contributes to DNA supercoiling, is more abundant in exponential phase of growth compared to stationary phase[23], has a preference for curved DNA[101], plays a role in polyamine DNA condensation[102], and is largely an accessory factor to stabilize nucleoprotein complexes[103]. It has been found to play a role in major

processes such as DNA replication, recombination, and gene expression[103–105]. IHF, as the name suggests, was initially discovered to be an essential factor for site-specific recombination of phage λ[106]. The binding and bending (which can be up to a 160° bend) capacity of IHF positions and stabilizes the DNA, directing bivalent integrase molecules to bind to the DNA[107] (**Fig. 1.2**). This interaction is surprisingly not mediated by any direct protein-protein interactions[107]. The crystal structure has been resolved, and the binding specificity to DNA seems to be determined by the inherent structure of DNA imposed by A/T-rich regions[108–112]. In terms of motif specificity, Fis and IHF can bind the same sites across the genome, and can lead to both repressive and activated effects[113]. The variation in the regulatory effect, whether it be from the mode by which NAPs bind to the proteins or the amount of protein bound, remains unclear.

DNA protection during starvation (Dps) forms a ferritin-like dodecamer with a hollow core [114], has low sequence specificity for DNA, required for starvation response[28], and makes up half of the nucleoid in stationary phase of growth[115]. DNA and Dps form a DNA-protein crystal [116] that aids in protection of the DNA[28,117,118]. Dps is a main facilitator of DNA compaction during later stages of growth, largely combating the actions of Fis, which controls DNA gyrase and topoisomerase I and prevents large-scale condensation of the nucleoid in exponential phase[70,119–121]. Similarly to ferritin, Dps serves an important role in iron acquisition and contributes to oxidative damage protection[122–125]. Dps is induced post-transcriptionally in times of carbon or nitrogen starvation and oxidative stress[119,126,127], and in pathogenic *E. coli*, Dps promotes acid tolerance[128].The regulation of Dps has been extensively studied and summarized[129–135]. During exponential growth, Dps is degraded by proteases ClpXP, however in times of carbon starvation proteolysis is halted to maintain proper levels of Dps for DNA protection and compaction[126]. The compaction mediated by Dps does not repress transcription, supporting the idea that compaction may form phase separated droplets that still enables RNA polymerase accessibility[136] (**Fig. 1.2**). More exploration into the impact Dps has on genome regulation is required to fully appreciate its impact on the cell.

Histone-like nucleoid structuring protein (H-NS) is a small basic protein known to be a major regulator of the cell[137] as a transcriptional repressor[138,139]. There is no known direct inducer of H-NS, however H-NS is thought to be autoregulator and is known to regulate transcription of major regulators in response to a number of different processes such as acid resistance[140], flagella biogenesis[141], rRNA components[142], proteases[143], and metabolism[144]. The wide range of systems H-NS regulates and sequences it binds suggests that the regulator role of H-NS relies on its role in impacting chromosomal structure. H-NS has a strong preference for A/T rich regions, horizontally acquired DNA, and regulates newly acquired DNA[140,145]. Conversely when compared to Fis, negatively correlated with transcriptional propensity and positively correlates with integration of a reporter and A/T content, further showing that H-NS silences newly acquired DNA[86]. H-NS contributes to the compaction[146,147] and organization[148] of the nucleoid, is capable of supercoiling DNA[46,149–151] and forms different types of DNA filaments[14,56]. H-NS can form homodimers or heterodimers with its paralog StpA[152]. While sharing similar sequence and structural features[56,153], H-NS is more highly abundant in the cell and has a lower DNA binding affinity[154]. The loss of *hns* leads to an increase in expression of *stpA*[154,155], however the loss of *stpA* shows minimal phenotypic effects largely hypothesized to be due to its lower amount in the cell and overlapping regulatory roles with H-NS. StpA has been shown to partially compensate[156] for the loss of *hns*, and repress similar genes. *In vitro*, H-NS forms both linear (binding one DNA fragment) and bridge (“bridging” two fragments of DNA) filaments across dsDNA (**Fig. 1.2**), that have the capacity to impede transcription by interfering with RNA polymerase[14,56]. Both linear and bridged filaments can impede transcription initiation by binding throughout the promoter and transcription start site[56]. Only bridged filaments promote RNA polymerase backtrack pausing and subsequent ρ-dependent termination[56]. StpA can form these types of filaments with H-NS, and in some species Hha, which belongs to the family of H-NS proteins but does not have a DNA binding domain, supports the formation of bridged H-NS filaments[56]. StpA has also been linked to RNA chaperone activity, further deepening the mechanistic options these proteins have on impacting

bacterial chromatin[157]. Along with the other NAPs, the recruitment of H-NS to particular regions, the binding mode of H-NS in the cell, and the regulation of H-NS as whole is undetermined.

Host factor for phage Q beta (Hfq): RNA chaperone or DNA binding protein? Hfq is a well-documented, conserved RNA-binding protein, whose homohexamer ring has the propensity to bind RNA in a number of different conformations[158–161]. Sequence analysis of Hfq revealed that it was related to Sm proteins found in eukaryotes and archaea, which similarly form ring structures that are the main unit of spliceosomal small nuclear ribonucleoproteins (snRNPs)[162,163]. These snRNPs are key building blocks of the spliceosome, which splice and process RNA[164]. Hfq facilitates and stabilizes small RNA interactions that repress mRNA translation and promote degradation for a number of RNA transcripts[158,165]. Through these interactions it has been shown that Hfq is mediates the translation of RpoS - a stress induced sigma factor in both *Salmonella typhimurium* and *E. coli [166–168].*The similarities between Sm proteins and Hfq, while incredibly fruitful and interesting when considering the evolutionary link across domains of life, has led to a bias in studies focusing solely on Hfq's RNA interactions, leaving its DNA binding capabilities largely uncharacterized. Hfq was originally identified as a gene required for phage Q beta propagation and RNA-directed synthesis of infected *E. coli* [169,170] but has been connected to a number of different processes. For instance, Hfq has been shown to associate with nascent transcripts and RNA polymerase, but the connection has not been made *in vivo* [171,172].The loss of *hfq* exhibits pleiotropic phenotypes, such as impacting cell division and decreasing negative supercoiling, osmosensitivity, largely thought to be linked heavily to its RNA chaperone activity[173–177]. Hfq has been shown to form foci in response to starvation[26] and plays a role in stress-induced mutagenesis[173]. However, Hfq, especially with increasing amounts of hexamers, has been shown to compact dsDNA, and recently a structure was resolved showing its interaction with DNA molecules[54] (**Fig. 1.2**). Like H-NS, Hfq is able to bridge dsDNA[178]. Understanding Hfq's role in regulating genes at the level of occupancy across the genome will give insight into the mechanism behind the wide variety of effects Hfq has on the cell.

Understanding the state of the cell: tools to profile genome architecture and regulation.

High-throughput assays have greatly improved our ability to understand gene regulation in a variety of species, conditions, and scales. When considering the central dogma of life and the tools to profile each level, DNA-sequencing, RNA-sequencing, and mass spectrometry begin to tease apart the content of a cell at a given time[179,180]. ChIP-seq enables the assessment of binding of specific proteins, to gain knowledge on binding motifs and regulators such as transcription factors across the entire genome. While all of these methods and others have enhanced the understanding of the genome, only until recently was there a method to capture the structure of the genome. Conceptually, it has only recently been appreciated that the genome structure mattered, if it existed at all. Three dimensional analysis of the genome using chromosome conformation capture techniques has exploded the ability to understand genome domains, folding, and looping [181–183]. Despite having an expansive toolkit, it is still poorly understood how structure and function are linked[184].

Employing the proper bioinformatic tools and pipelines is essential to gaining biological insight to any high-throughput tool. In the case of both eukaryotes and prokaryotes, this involves an intimate understanding of the genome. For instance, when applying high-throughput assays such as ChIP-seq to eukaryotic organisms, after processing the DNA, aligning it to the genome, and taking a glimpse of where peaks of DNA are found in a dataset, one would find distinct areas across the genome that always retain an obnoxiously high singal[185,186]. These signals are present no matter what condition or protein being assessed[186], and are largely attributed to high repetitive sequences across the genome. To resolve this issue, in the Appendix I will share my work defining the ENCODE Blacklist, areas of the genome that sequester signal due to repetitive content, across a variety of eukaryotic species.

In the case of prokaryotic organisms, the tools that exist are limited in comparison to their eukaryotic counterparts. While seemingly backwards when considering microbes have smaller genomes and are relatively easier to manipulate than, let’s say a human,

the focus for genome annotation and downstream tools has a spotlight on eukaryotic genomes. As an example, despite being one of the most well studied organisms on the planet, to date, ~50% of the genes in *Escherichia coli* (*E. coli*) are of unknown function[187–189]. Even less information is known about the genome of common microbial pathogens and commensals- leaving a pool of genetic information that remains, largely, a mystery. Of the tools listed above, DNA-seq, RNA-seq, ChIP-seq have all successfully been applied to bacterial contexts. However, implementation of genome architecture tools has been challenging. While in some cases, 3D profiling has been successfully implemented[190–193], the dynamics of the bacterial genome require tools to profile changes in proteins that mediate DNA architecture.

Conclusions and future perspectives

Chromatin, here defined as packaging of the DNA, is essential for proper genome maintenance and gene regulation across all species[12]. In eukaryotes, this is largely mediated by histones and posttranslational modifications[2,3], similarly, bacteria contain histone-like proteins that impact DNA structure and compaction[13,14]. The organization of the DNA within the nucleoid imposes transcriptional effects that may change given environmental cues[26], replication status[22], or growth[15,23], reminiscent of heterochromatin in eukaryotes[12]. The histone-like proteins in bacteria are nucleoid associated proteins (NAPs) that mediate changes in DNA structure by forming linear DNA filaments[56], bridged DNA filaments[56], promote supercoiling[12,97], bend and wrap the DNA[12], and may form phase-separated condensates[14]. The variety in binding leads to a range of effects on gene expression, notably by interfering with RNA polymerase binding, translocation, or progression, and / or binding to nascent transcripts[14]. Furthermore, NAPs such as H-NS and StpA specifically silence horizontally acquired genes and mediate stress responses[14,25,59,63,64], leading to interesting hypotheses about whether a function of heterochromatin-like regions in bacteria is to promote diversification of the species, while also initially protecting the organism from toxic elements. Hfq, only recently being

deeply explored for its DNA binding capacity, also mediates the stress induced mutagenesis response[173], and in pathogenic organisms has been shown to play a role in virulence[58]. Together, these results show the impact nucleoid associated proteins have on DNA organization and gene regulation, but the link between *in vitro* characteristics and *in vivo* mechanisms has not been fully realized.

The low binding specificity and overlapping binding characteristics of NAPs make them especially difficult to study. Assessment of some NAP deletions have also resulted in no large changes in transcription and / or physiology, leading to increased confusion as to the role of NAPs in binding or compensation[136,156]. To decipher the changes in genome-wide binding of NAPs, implementation of a series of ChIP-seq experiments seems like the likely next step, however, since there is not an antibody for each NAP, we cannot profile each NAP in this manner. Additionally, ChIP-seq may not retain information about overall binding profiles of the genome that may provide further information on how NAPs impact binding of other proteins. Therefore, I helped create a method termed *in vivo* protein occupancy display at high-resolution (IPOD-HR) presented in Chapter 2 that enables genome-wide assessment of protein occupancy[42], and reveals heterochromatin-like regions, termed extended protein occupancy domains (EPODs) in *E. coli*. IPOD-HR does not rely on an antibody, and can be applied to a number of different organisms to assess the presence of heterochromatin-like regions across the genome[42]. In Chapter 3, I will implement IPOD-HR to investigate the changes in protein occupancy due to the deletions of key NAPs and pair IPOD-HR with RNA-seq to better understand the role of NAPs in gene regulation. I identify the key NAPs regulating metabolic pathways and xenogeneic silencing in *E. coli* and the distantly related species *Bacillus subtilis*. Furthermore, I find that induction of metabolic pathways via carbon source exposure may mediate a transcriptional memory response, again linking the heterochromatin-like regions in bacteria to serve similar functions as seen in eukaryotes. Lastly, I show that Fis and Hfq serve overlapping roles in silencing prophages and mobile elements and attempting to delete both *fis* and *hfq* leads to inviability. In Chapter 4, I investigate the mechanism behind Hfq's role as a prophage silencer in connection to a polyanion, polyphosphate,

which seems to mediate Hfq binding to prophages and mobile elements. The work presented in the remainder of this thesis provides new functional and mechanistic insights into bacterial chromatin *in vivo* across the genome. The conservation of many of these NAPs across species, and as shown here the similar functions of EPODs across species, will enable this work to impact biotechnology and health.

References

1. Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature. 1997;389: 251–260.
2. Jenuwein T, Allis CD. Translating the histone code. Science. 2001;293: 1074–1080.
3. Grewal SIS, Jia S. Heterochromatin revisited. Nat Rev Genet. 2007;8: 35–46.
4. Elgin SC. Heterochromatin and gene regulation in Drosophila. Curr Opin Genet Dev. 1996;6: 193–202.
5. Bowman GD, Poirier MG. Post-translational modifications of histones that influence nucleosome dynamics. Chem Rev. 2015;115: 2274–2295.
6. Francis NJ, Kingston RE. Mechanisms of transcriptional memory. Nat Rev Mol Cell Biol. 2001;2: 409–421.
7. Ferraro T, Esposito E, Mancini L, Ng S, Lucas T, Coppey M, et al. Transcriptional Memory in the Drosophila Embryo. Curr Biol. 2016;26: 212–218.
8. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100: 57–70.
9. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144: 646–674.
10. Carone DM, Lawrence JB. Heterochromatin instability in cancer: from the Barr body to satellites and the nuclear periphery. Semin Cancer Biol. 2013;23: 99–108.
11. Zhao Z, Shilatifard A. Epigenetic modifications of histones in cancer. Genome Biol. 2019;20: 1–16.
12. Luijsterburg MS, White MF, van Driel R, Dame RT. The major architects of chromatin: architectural proteins in bacteria, archaea and eukaryotes. Crit Rev Biochem Mol Biol. 2008;43: 393–418.
13. Dame RT, Dorman CJ. Bacterial Chromatin. Springer Science & Business Media; 2009.
14. Shen BA, Landick R. Transcription of Bacterial Chromatin. J Mol Biol. 2019;431: 4040–4066.
15. Hadizadeh Yazdi N, Guet CC, Johnson RC, Marko JF. Variation of the folding and dynamics of the Escherichia coli chromosome with growth conditions. Mol Microbiol. 2012;86: 1318–1333.
16. Berlatzky IA, Rouvinski A, Ben-Yehuda S. Spatial organization of a replicating bacterial chromosome. Proc Natl Acad Sci U S A. 2008;105: 14136–14140.
17. Thanbichler M, Wang SC, Shapiro L. The bacterial nucleoid: a highly organized and dynamic structure. J Cell Biochem. 2005;96: 506–521.
18. Dillon SC, Dorman CJ. Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. Nat Rev Microbiol. 2010;8: 185–195.
19. Dame RT, Tark-Dame M. Bacterial chromatin: converging views at different scales. Curr Opin Cell Biol. 2016;40: 60–65.
20. Duigou S, Boccard F. Long range chromosome organization in Escherichia coli: The position of the replication origin defines the non-structured regions and the Right and Left macrodomains. PLoS Genet. 2017;13: e1006758.
21. Postow L, Hardy CD, Arsuaga J, Cozzarelli NR. Topological domain structure of the Escherichia coli chromosome. Genes Dev. 2004;18: 1766–1779.

22. Cagliero C, Grand RS, Jones MB, Jin DJ, O'Sullivan JM. Genome conformation capture reveals that the Escherichia coli chromosome is organized by replication and transcription. Nucleic Acids Res. 2013;41: 6058–6071.
23. 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. 1999;181: 6361–6370.
24. Ueguchi C, Suzuki T, Yoshida T, Tanaka K, Mizuno T. Systematic mutational analysis revealing the functional domain organization of Escherichia coli nucleoid protein H-NS. J Mol Biol. 1996;263: 149–162.
25. Ueguchi C, Mizuno T. The Escherichia coli nucleoid protein H-NS functions directly as a transcriptional repressor. EMBO J. 1993;12: 1039–1046.
26. McQuail J, Switzer A, Burchell L, Wigneshweraraj S. The assembly of Hfq into foci-like structures in response to long-term nitrogen starvation in Escherichia coli. Cold Spring Harbor Laboratory. 2020. p. 2020.01.10.901611. doi:10.1101/2020.01.10.901611
27. Nair S, Finkel SE. Dps protects cells against multiple stresses during stationary phase. J Bacteriol. 2004;186: 4192–4198.
28. Almirón M, Link AJ, Furlong D, Kolter R. A novel DNA-binding protein with regulatory and protective roles in starved Escherichia coli. Genes Dev. 1992;6: 2646–2654.
29. Bakshi S, Choi H, Weisshaar JC. The spatial biology of transcription and translation in rapidly growing Escherichia coli. Front Microbiol. 2015;6: 636.
30. Bakshi S, Choi H, Mondal J, Weisshaar JC. Time-dependent effects of transcription- and translation-halting drugs on the spatial distributions of the Escherichia coli chromosome and ribosomes. Mol Microbiol. 2014;94: 871–887.
31. Baumberg S. Prokaryotic Gene Expression. OUP Oxford; 1999.
32. Wagner R. Transcription Regulation in Prokaryotes. Oxford University Press on Demand; 2000.
33. Zhang G, Darst SA. Structure of the Escherichia coli RNA polymerase alpha subunit amino-terminal domain. Science. 1998;281: 262–266.
34. Young BA, Gruber TM, Gross CA. Minimal machinery of RNA polymerase holoenzyme sufficient for promoter melting. Science. 2004;303: 1382–1384.
35. Jishage M, Iwata A, Ueda S, Ishihama A. Regulation of RNA polymerase sigma subunit synthesis in Escherichia coli: intracellular levels of four species of sigma subunit under various growth conditions. J Bacteriol. 1996;178: 5447–5451.
36. Ozaki M, Wada A, Fujita N, Ishihama A. Growth phase-dependent modification of RNA polymerase in Escherichia coli. Mol Gen Genet. 1991;230: 17–23.
37. Wade JT, Struhl K. Association of RNA polymerase with transcribed regions in Escherichia coli. Proc Natl Acad Sci U S A. 2004;101: 17777–17782.
38. Sun Z, Cagliero C, Izard J, Chen Y, Zhou YN, Heinz WF, et al. Density of σ70 promoter-like sites in the intergenic regions dictates the redistribution of RNA polymerase during osmotic stress in Escherichia coli. Nucleic Acids Res. 2019;47: 3970–3985.
39. Magnusson LU, Nystrom T, Farewell A. Underproduction of sigma 70 mimics a stringent response. A proteome approach. J Biol Chem. 2003;278: 968–973.

40. Fujita N, Nomura T, Ishihama A. Promoter selectivity of Escherichia coli RNA polymerase. Purification and properties of holoenzyme containing the heat-shock sigma subunit. J Biol Chem. 1987;262: 1855–1859.
41. Vora T, Hottes AK, Tavazoie S. Protein occupancy landscape of a bacterial genome. Mol Cell. 2009;35: 247–253.
42. Freddolino PL, Goss TJ, Amemiya HM, Tavazoie S. Dynamic landscape of protein occupancy across the Escherichia coli chromosome. Cold Spring Harbor Laboratory. 2020. p. 2020.01.29.924811. doi:10.1101/2020.01.29.924811
43. Jin DJ, Cagliero C, Martin CM, Izard J, Zhou YN. The dynamic nature and territory of transcriptional machinery in the bacterial chromosome. Front Microbiol. 2015;6: 497.
44. Joyeux M. Compaction of bacterial genomic DNA: clarifying the concepts. J Phys Condens Matter. 2015;27: 383001.
45. Lagomarsino MC, Espéli O, Junier I. From structure to function of bacterial chromosomes: Evolutionary perspectives and ideas for new experiments. FEBS Lett. 2015;589: 2996–3004.
46. Japaridze A, Renevey S, Sobetzko P, Stoliar L, Nasser W, Dietler G, et al. Spatial organization of DNA sequences directs the assembly of bacterial chromatin by a nucleoid-associated protein. J Biol Chem. 2017;292: 7607–7618.
47. Liu LF, Wang JC. Supercoiling of the DNA template during transcription. Proc Natl Acad Sci U S A. 1987;84: 7024–7027.
48. Ma J, Bai L, Wang MD. Transcription under torsion. Science. 2013;340: 1580–1583.
49. Zhang J, Landick R. A Two-Way Street: Regulatory Interplay between RNA Polymerase and Nascent RNA Structure. Trends Biochem Sci. 2016;41: 293–310.
50. Vitiello CL, Kireeva ML, Lubkowska L, Kashlev M, Gottesman M. Coliphage HK022 Nun protein inhibits RNA polymerase translocation. Proc Natl Acad Sci U S A. 2014;111: E2368–75.
51. Hnisz D, Shrinivas K, Young RA, Chakraborty AK, Sharp PA. A Phase Separation Model for Transcriptional Control. Cell. 2017;169: 13–23.
52. Ohniwa RL, Ushijima Y, Saito S, Morikawa K. Proteomic analyses of nucleoid-associated proteins in Escherichia coli, Pseudomonas aeruginosa, Bacillus subtilis, and Staphylococcus aureus. PLoS One. 2011;6: e19172.
53. Hołówka J, Zakrzewska-Czerwińska J. Nucleoid Associated Proteins: The Small Organizers That Help to Cope With Stress. Front Microbiol. 2020;11: 590.
54. Orans J, Kovach AR, Hoff KE, Horstmann NM, Brennan RG. Crystal structure of an Escherichia coli Hfq Core (residues 2-69)-DNA complex reveals multifunctional nucleic acid binding sites. Nucleic Acids Res. 2020;48: 3987–3997.
55. Arold ST, Leonard PG, Parkinson GN, Ladbury JE. H-NS forms a superhelical protein scaffold for DNA condensation. Proc Natl Acad Sci U S A. 2010;107: 15728–15732.
56. Boudreau BA, Hron DR, Qin L, van der Valk RA, Kotlajich MV, Dame RT, et al. StpA and Hha stimulate pausing by RNA polymerase by promoting DNA-DNA bridging of H-NS filaments. Nucleic Acids Res. 2018;46: 5525–5546.

57. Kakoschke TK, Kakoschke SC, Zeuzem C, Bouabe H, Adler K, Heesemann J, et al. The RNA Chaperone Hfq Is Essential for Virulence and Modulates the Expression of Four Adhesins in Yersinia enterocolitica. Sci Rep. 2016;6: 29275.
58. Liu Y, Wu N, Dong J, Gao Y, Zhang X, Mu C, et al. Hfq is a global regulator that controls the pathogenicity of Staphylococcus aureus. PLoS One. 2010;5. doi:10.1371/journal.pone.0013069
59. Higashi K, Tobe T, Kanai A, Uyar E, Ishikawa S, Suzuki Y, et al. H-NS Facilitates Sequence Diversification of Horizontally Transferred DNAs during Their Integration in Host Chromosomes. PLoS Genet. 2016;12: e1005796.
60. Perez JC, Latifi T, Groisman EA. Overcoming H-NS-mediated transcriptional silencing of horizontally acquired genes by the PhoP and SlyA proteins in Salmonella enterica. J Biol Chem. 2008;283: 10773–10783.
61. Finkel SE, Johnson RC. The Fis protein: it's not just for DNA inversion anymore. Mol Microbiol. 1992;6: 3257–3265.
62. Sheikh J, Hicks S, Dall'Agnol M, Phillips AD, Nataro JP. Roles for Fis and YafK in biofilm formation by enteroaggregative Escherichia coli. Mol Microbiol. 2001;41: 983–997.
63. Navarre WW, McClelland M, Libby SJ, Fang FC. Silencing of xenogeneic DNA by H-NS-facilitation of lateral gene transfer in bacteria by a defense system that recognizes foreign DNA. Genes Dev. 2007;21: 1456–1471.
64. Lucchini S, Rowley G, Goldberg MD, Hurd D, Harrison M, Hinton JCD. H-NS mediates the silencing of laterally acquired genes in bacteria. PLoS Pathog. 2006;2: e81.
65. Navarre WW, Porwollik S, Wang Y, McClelland M, Rosen H, Libby SJ, et al. Selective silencing of foreign DNA with low GC content by the H-NS protein in Salmonella. Science. 2006;313: 236–238.
66. Dilweg IW, Dame RT. Post-translational modification of nucleoid-associated proteins: an extra layer of functional modulation in bacteria? Biochem Soc Trans. 2018;46: 1381–1392.
67. Koch C, Kahmann R. Purification and properties of the Escherichia coli host factor required for inversion of the G segment in bacteriophage Mu. J Biol Chem. 1986;261: 15673–15678.
68. Johnson RC, Bruist MF, Simon MI. Host protein requirements for in vitro site-specific DNA inversion. Cell. 1986;46: 531–539.
69. Cho B-K, Knight EM, Barrett CL, Palsson BØ. Genome-wide analysis of Fis binding in Escherichia coli indicates a causative role for A-/AT-tracts. Genome Res. 2008;18: 900–910.
70. Schneider R, Lurz R, Lüder G, Tolksdorf C, Travers A, Muskhelishvili G. An architectural role of the Escherichia coli chromatin protein FIS in organising DNA. Nucleic Acids Res. 2001;29: 5107–5114.
71. Weinstein-Fischer D, Altuvia S. Differential regulation of Escherichia coli topoisomerase I by Fis. Mol Microbiol. 2007;63: 1131–1144.
72. Xu J, Johnson RC. Fis activates the RpoS-dependent stationary-phase expression of proP in Escherichia coli. J Bacteriol. 1995;177: 5222–5231.

73. Schneider R, Travers A, Muskhelishvili G. The expression of the Escherichia coli fis gene is strongly dependent on the superhelical density of DNA. Mol Microbiol. 2000;38: 167–175.
74. Cheng YS, Yang WZ, Johnson RC, Yuan HS. Structural analysis of the transcriptional activation region on Fis: crystal structures of six Fis mutants with different activation properties. J Mol Biol. 2000;302: 1139–1151.
75. Yuan HS, Finkel SE, Feng JA, Kaczor-Grzeskowiak M, Johnson RC, Dickerson RE. The molecular structure of wild-type and a mutant Fis protein: relationship between mutational changes and recombinational enhancer function or DNA binding. Proc Natl Acad Sci U S A. 1991;88: 9558–9562.
76. Kostrewa D, Granzin J, Koch C, Choe HW, Raghunathan S, Wolf W, et al. Three-dimensional structure of the E. coli DNA-binding protein FIS. Nature. 1991;349: 178–180.
77. Safo MK, Yang WZ, Corselli L, Cramton SE, Yuan HS, Johnson RC. The transactivation region of the fis protein that controls site-specific DNA inversion contains extended mobile beta-hairpin arms. EMBO J. 1997;16: 6860–6873.
78. Skoko D, Yoo D, Bai H, Schnurr B, Yan J, McLeod SM, et al. Mechanism of chromosome compaction and looping by the Escherichia coli nucleoid protein Fis. J Mol Biol. 2006;364: 777–798.
79. Travers A, Muskhelishvili G. DNA microloops and microdomains: a general mechanism for transcription activation by torsional transmission. J Mol Biol. 1998;279: 1027–1043.
80. Ross W, Thompson JF, Newlands JT, Gourse RL. E.coli Fis protein activates ribosomal RNA transcription in vitro and in vivo. EMBO J. 1990;9: 3733–3742.
81. Opel ML, Aeling KA, Holmes WM, Johnson RC, Benham CJ, Hatfield GW. Activation of transcription initiation from a stable RNA promoter by a Fis protein-mediated DNA structural transmission mechanism. Mol Microbiol. 2004;53: 665–674.
82. Xu J, Johnson RC. Activation of RpoS-dependent proP P2 transcription by the Fis protein in vitro. J Mol Biol. 1997;270: 346–359.
83. Browning DF, Grainger DC, Beatty CM, Wolfe AJ, Cole JA, Busby SJW. Integration of three signals at the Escherichia coli nrf promoter: a role for Fis protein in catabolite repression. Mol Microbiol. 2005;57: 496–510.
84. Zusman T, Speiser Y, Segal G. Two Fis regulators directly repress the expression of numerous effector-encoding genes in Legionella pneumophila. J Bacteriol. 2014;196: 4172–4183.
85. Karambelkar S, Swapna G, Nagaraja V. Silencing of toxic gene expression by Fis. Nucleic Acids Res. 2012;40: 4358–4367.
86. Scholz SA, Diao R, Wolfe MB, Fivenson EM, Lin XN, Freddolino PL. High-Resolution Mapping of the Escherichia coli Chromosome Reveals Positions of High and Low Transcription. Cell Syst. 2019;8: 212–225.e9.
87. Macvanin M, Adhya S. Architectural organization in E. coli nucleoid. Biochim Biophys Acta. 2012;1819: 830–835.
88. Ramstein J, Hervouet N, Coste F, Zelwer C, Oberto J, Castaing B. Evidence of a thermal unfolding dimeric intermediate for the Escherichia coli histone-like HU proteins: thermodynamics and structure. J Mol Biol. 2003;331: 101–121.

89. Sarkar T, Vitoc I, Mukerji I, Hud NV. Bacterial protein HU dictates the morphology of DNA condensates produced by crowding agents and polyamines. Nucleic Acids Res. 2007;35: 951–961.
90. Lee H, Kim HK, Kang S, Hong CB, Yim J, Hwang DS. Expression of the seqA gene is negatively modulated by the HU protein in Escherichia coli. Mol Gen Genet. 2001;264: 931–935.
91. Oberto J, Nabti S, Jooste V, Mignot H, Rouviere-Yaniv J. The HU regulon is composed of genes responding to anaerobiosis, acid stress, high osmolarity and SOS induction. PLoS One. 2009;4: e4367.
92. Semsey S, Tolstorukov MY, Virnik K, Zhurkin VB, Adhya S. DNA trajectory in the Gal repressosome. Genes Dev. 2004;18: 1898–1907.
93. Painbeni E, Caroff M, Rouviere-Yaniv J. Alterations of the outer membrane composition in Escherichia coli lacking the histone-like protein HU. Proc Natl Acad Sci U S A. 1997;94: 6712–6717.
94. Aki T, Adhya S. Repressor induced site-specific binding of HU for transcriptional regulation. EMBO J. 1997;16: 3666–3674.
95. Azam TA, Hiraga S, Ishihama A. Two types of localization of the DNA-binding proteins within the Escherichia coli nucleoid. Genes Cells. 2000;5: 613–626.
96. Remesh SG, Verma SC, Chen J-H, Ekman AA, Larabell CA, Adhya S, et al. Nucleoid remodeling during environmental adaptation is regulated by HU-dependent DNA bundling. Nat Commun. 2020;11: 2905.
97. Yan Y, Leng F, Finzi L, Dunlap D. Protein-mediated looping of DNA under tension requires supercoiling. Nucleic Acids Res. 2018;46: 2370–2379.
98. Shindo H, Furubayashi A, Shimizu M, Miyake M, Imamoto F. Preferential binding of E.coli histone-like protein HU alpha to negatively supercoiled DNA. Nucleic Acids Res. 1992;20: 1553–1558.
99. Rouvière-Yaniv J, Yaniv M, Germond JE. E. coli DNA binding protein HU forms nucleosomelike structure with circular double-stranded DNA. Cell. 1979;17: 265–274.
100. Ghosh S, Mallick B, Nagaraja V. Direct regulation of topoisomerase activity by a nucleoid-associated protein. Nucleic Acids Res. 2014;42: 11156–11165.
101. Liu G, Ma Q, Xu Y. Physical properties of DNA may direct the binding of nucleoid-associated proteins along the E. coli genome. Math Biosci. 2018;301: 50–58.
102. Sarkar T, Petrov AS, Vitko JR, Santai CT, Harvey SC, Mukerji I, et al. Integration host factor (IHF) dictates the structure of polyamine-DNA condensates: implications for the role of IHF in the compaction of bacterial chromatin. Biochemistry. 2009;48: 667–675.
103. Swinger KK, Rice PA. IHF and HU: flexible architects of bent DNA. Curr Opin Struct Biol. 2004;14: 28–35.
104. Dhavan GM, Crothers DM, Chance MR, Brenowitz M. Concerted binding and bending of DNA by Escherichia coli integration host factor. J Mol Biol. 2002;315: 1027–1037.
105. Freundlich M, Ramani N, Mathew E, Sirko A, Tsui P. The role of integration host factor in gene expression in Escherichia coli. Mol Microbiol. 1992;6: 2557–2563.

106. Miller HI, Friedman DI. An E. coli gene product required for lambda site-specific recombination. Cell. 1980;20: 711–719.
107. Moitoso de Vargas L, Kim S, Landy A. DNA looping generated by DNA bending protein IHF and the two domains of lambda integrase. Science. 1989;244: 1457–1461.
108. Rice PA, Yang S, Mizuuchi K, Nash HA. Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn. Cell. 1996;87: 1295–1306.
109. Hales LM, Gumport RI, Gardner JF. Examining the contribution of a dA+dT element to the conformation of Escherichia coli integration host factor-DNA complexes. Nucleic Acids Res. 1996;24: 1780–1786.
110. Friedman DI. Integration host factor: a protein for all reasons. Cell. 1988;55: 545–554.
111. Goodrich JA, Schwartz ML, McClure WR. Searching for and predicting the activity of sites for DNA binding proteins: compilation and analysis of the binding sites for Escherichia coli integration host factor (IHF). Nucleic Acids Res. 1990;18: 4993–5000.
112. Ellenberger T, Landy A. A good turn for DNA: the structure of integration host factor bound to DNA. Structure. 1997;5: 153–157.
113. Monteiro LMO, Sanches-Medeiros A, Westmann CA, Silva-Rocha R. Unraveling the Complex Interplay of Fis and IHF Through Synthetic Promoter Engineering. Front Bioeng Biotechnol. 2020;8: 510.
114. Grant RA, Filman DJ, Finkel SE, Kolter R, Hogle JM. The crystal structure of Dps, a ferritin homolog that binds and protects DNA. Nat Struct Biol. 1998;5: 294–303.
115. Talukder A, Ishihama A. Growth phase dependent changes in the structure and protein composition of nucleoid in Escherichia coli. Sci China Life Sci. 2015;58: 902–911.
116. Wolf SG, Frenkiel D, Arad T, Finkel SE, Kolter R, Minsky A. DNA protection by stress-induced biocrystallization. Nature. 1999;400: 83–85.
117. Frenkiel-Krispin D, Levin-Zaidman S, Shimoni E, Wolf SG, Wachtel EJ, Arad T, et al. Regulated phase transitions of bacterial chromatin: a non-enzymatic pathway for generic DNA protection. EMBO J. 2001;20: 1184–1191.
118. Martinez A, Kolter R. Protection of DNA during oxidative stress by the nonspecific DNA-binding protein Dps. J Bacteriol. 1997;179: 5188–5194.
119. Ohniwa RL, Morikawa K, Kim J, Ohta T, Ishihama A, Wada C, et al. Dynamic state of DNA topology is essential for genome condensation in bacteria. EMBO J. 2006;25: 5591–5602.
120. Weinstein-Fischer D, Elgrably-Weiss M, Altuvia S. Escherichia coli response to hydrogen peroxide: a role for DNA supercoiling, topoisomerase I and Fis. Mol Microbiol. 2000;35: 1413–1420.
121. Sato YT, Watanabe S, Kenmotsu T, Ichikawa M, Yoshikawa Y, Teramoto J, et al. Structural change of DNA induced by nucleoid proteins: growth phase-specific Fis and stationary phase-specific Dps. Biophys J. 2013;105: 1037–1044.
122. Rychlewski L, Zhang B, Godzik A. Functional insights from structural predictions: analysis of the Escherichia coli genome. Protein Sci. 1999;8: 614–624.

123. Zhao G, Ceci P, Ilari A, Giangiacomo L, Laue TM, Chiancone E, et al. Iron and hydrogen peroxide detoxification properties of DNA-binding protein from starved cells. A ferritin-like DNA-binding protein of Escherichia coli. J Biol Chem. 2002;277: 27689–27696.
124. Ilari A, Ceci P, Ferrari D, Rossi GL, Chiancone E. Iron incorporation into Escherichia coli Dps gives rise to a ferritin-like microcrystalline core. J Biol Chem. 2002;277: 37619–37623.
125. Ceci P, Di Cecca G, Falconi M, Oteri F, Zamparelli C, Chiancone E. Effect of the charge distribution along the “ferritin-like” pores of the proteins from the Dps family on the iron incorporation process. J Biol Inorg Chem. 2011;16: 869–880.
126. Stephani K, Weichart D, Hengge R. Dynamic control of Dps protein levels by ClpXP and ClpAP proteases in Escherichia coli. Mol Microbiol. 2003;49: 1605–1614.
127. Lomovskaya OL, Kidwell JP, Matin A. Characterization of the sigma 38-dependent expression of a core Escherichia coli starvation gene, pexB. J Bacteriol. 1994;176: 3928–3935.
128. Jeong KC, Hung KF, Baumler DJ, Byrd JJ, Kaspar CW. Acid stress damage of DNA is prevented by Dps binding in Escherichia coli O157:H7. BMC Microbiol. 2008;8: 181.
129. Ivanova AB, Glinsky GV, Eisenstark A. Role of rpoS regulon in resistance to oxidative stress and near-UV radiation in delta oxyR suppressor mutants of Escherichia coli. Free Radic Biol Med. 1997;23: 627–636.
130. Rockabrand D, Livers K, Austin T, Kaiser R, Jensen D, Burgess R, et al. Roles of DnaK and RpoS in starvation-induced thermotolerance of Escherichia coli. J Bacteriol. 1998;180: 846–854.
131. Altuvia S, Almirón M, Huisman G, Kolter R, Storz G. The dps promoter is activated by OxyR during growth and by IHF and sigma S in stationary phase. Mol Microbiol. 1994;13: 265–272.
132. Gérard F, Dri A-M, Moreau PL. Role of Escherichia coli RpoS, LexA and H-NS global regulators in metabolism and survival under aerobic, phosphate-starvation conditions. Microbiology. 1999;145 (Pt 7): 1547–1562.
133. Michán C, Manchado M, Dorado G, Pueyo C. In vivo transcription of the Escherichia coli oxyR regulon as a function of growth phase and in response to oxidative stress. J Bacteriol. 1999;181: 2759–2764.
134. Bechtloff D, Grünenfelder B, Akerlund T, Nordström K. Analysis of protein synthesis rates after initiation of chromosome replication in Escherichia coli. J Bacteriol. 1999;181: 6292–6299.
135. Yamamoto K, Ishihama A, Busby SJW, Grainger DC. The Escherichia coli K-12 MntR miniregulon includes dps, which encodes the major stationary-phase DNA-binding protein. J Bacteriol. 2011;193: 1477–1480.
136. Janissen R, Arens MMA, Vtyurina NN, Rivai Z, Sunday ND, Eslami-Mossallam B, et al. Global DNA Compaction in Stationary-Phase Bacteria Does Not Affect Transcription. Cell. 2018;174: 1188–1199.e14.
137. Dorman CJ. H-NS: a universal regulator for a dynamic genome. Nat Rev Microbiol. 2004;2: 391–400.

138. Fang FC, Rimsky S. New insights into transcriptional regulation by H-NS. Curr Opin Microbiol. 2008;11: 113–120.
139. Hommais F, Krin E, Laurent-Winter C, Soutourina O, Malpertuy A, Le Caer JP, et al. Large-scale monitoring of pleiotropic regulation of gene expression by the prokaryotic nucleoid-associated protein, H-NS. Mol Microbiol. 2001;40: 20–36.
140. 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. 2006;13: 141–153.
141. Bertin P, Terao E, Lee EH, Lejeune P, Colson C, Danchin A, et al. The H-NS protein is involved in the biogenesis of flagella in Escherichia coli. J Bacteriol. 1994;176: 5537–5540.
142. Afflerbach H, Schröder O, Wagner R. Effects of the Escherichia coli DNA-binding protein H-NS on rRNA synthesis in vivo. Mol Microbiol. 1998;28: 641–653.
143. Forns N, Juárez A, Madrid C. Osmoregulation of the HtrA (DegP) protease of Escherichia coli: an Hha-H-NS complex represses HtrA expression at low osmolarity. FEMS Microbiol Lett. 2005;251: 75–80.
144. Rimsky S, Spassky A. Sequence determinants for H1 binding on Escherichia coli lac and gal promoters. Biochemistry. 1990;29: 3765–3771.
145. Yousuf M, Iuliani I, Veetil RT, Seshasayee ASN, Sclavi B, Cosentino Lagomarsino M. Early fate of exogenous promoters in E. coli. Nucleic Acids Res. 2020;48: 2348–2356.
146. Dame RT, Wyman C, Goosen N. H-NS mediated compaction of DNA visualised by atomic force microscopy. Nucleic Acids Res. 2000;28: 3504–3510.
147. Dame RT. The role of nucleoid-associated proteins in the organization and compaction of bacterial chromatin. Mol Microbiol. 2005;56: 858–870.
148. Wang W, Li G-W, Chen C, Xie XS, Zhuang X. Chromosome organization by a nucleoid-associated protein in live bacteria. Science. 2011;333: 1445–1449.
149. Zimmerman SB. Cooperative transitions of isolated Escherichia coli nucleoids: implications for the nucleoid as a cellular phase. J Struct Biol. 2006;153: 160–175.
150. McLeod SM, Johnson RC. Control of transcription by nucleoid proteins. Curr Opin Microbiol. 2001;4: 152–159.
151. Tupper AE, Owen-Hughes TA, Ussery DW, Santos DS, Ferguson DJ, Sidebotham JM, et al. The chromatin-associated protein H-NS alters DNA topology in vitro. EMBO J. 1994;13: 258–268.
152. Leonard PG, Ono S, Gor J, Perkins SJ, Ladbury JE. Investigation of the self-association and hetero-association interactions of H-NS and StpA from Enterobacteria. Mol Microbiol. 2009;73: 165–179.
153. Zhang A, Belfort M. Nucleotide sequence of a newly-identified Escherichia coli gene, stpA, encoding an H-NS-like protein. Nucleic Acids Res. 1992;20: 6735.
154. Sonnenfield JM, Burns CM, Higgins CF, Hinton JC. The nucleoid-associated protein StpA binds curved DNA, has a greater DNA-binding affinity than H-NS and is present in significant levels in hns mutants. Biochimie. 2001;83: 243–249.
155. Sonden B, Uhlin BE. Coordinated and differential expression of histone-like proteins in Escherichia coli: regulation and function of the H-NS analog StpA. EMBO J. 1996;15: 4970–4980.

156. Shi X, Bennett GN. Plasmids bearing hfq and the hns-like gene stpA complement hns mutants in modulating arginine decarboxylase gene expression in Escherichia coli. J Bacteriol. 1994;176: 6769–6775.
157. Doetsch M, Gstrein T, Schroeder R, Fürtig B. Mechanisms of StpA-mediated RNA remodeling. RNA Biol. 2010;7: 735–743.
158. Zhang A, Wassarman KM, Ortega J, Steven AC, Storz G. The Sm-like Hfq protein increases OxyS RNA interaction with target mRNAs. Mol Cell. 2002;9: 11–22.
159. Møller T, Franch T, Højrup P, Keene DR, Bächinger HP, Brennan RG, et al. Hfq: a bacterial Sm-like protein that mediates RNA-RNA interaction. Mol Cell. 2002;9: 23–30.
160. Arluison V, Derreumaux P, Allemand F, Folichon M, Hajnsdorf E, Régnier P. Structural Modelling of the Sm-like Protein Hfq from Escherichia coli. J Mol Biol. 2002;320: 705–712.
161. Sauter C, Basquin J, Suck D. Sm-like proteins in Eubacteria: the crystal structure of the Hfq protein from Escherichia coli. Nucleic Acids Res. 2003;31: 4091–4098.
162. Hermann H, Fabrizio P, Raker VA, Foulaki K, Hornig H, Brahms H, et al. snRNP Sm proteins share two evolutionarily conserved sequence motifs which are involved in Sm protein-protein interactions. EMBO J. 1995;14: 2076–2088.
163. Salgado-Garrido J, Bragado-Nilsson E, Kandels-Lewis S, Séraphin B. Sm and Sm-like proteins assemble in two related complexes of deep evolutionary origin. EMBO J. 1999;18: 3451–3462.
164. Will CL, Lührmann R. Spliceosomal UsnRNP biogenesis, structure and function. Curr Opin Cell Biol. 2001;13: 290–301.
165. Melamed S, Peer A, Faigenbaum-Romm R, Gatt YE, Reiss N, Bar A, et al. Global Mapping of Small RNA-Target Interactions in Bacteria. Mol Cell. 2016;63: 884–897.
166. Brown L, Elliott T. Efficient translation of the RpoS sigma factor in Salmonella typhimurium requires host factor I, an RNA-binding protein encoded by the hfq gene. J Bacteriol. 1996;178: 3763–3770.
167. Muffler A, Fischer D, Hengge-Aronis R. The RNA-binding protein HF-I, known as a host factor for phage Qbeta RNA replication, is essential for rpoS translation in Escherichia coli. Genes Dev. 1996;10: 1143–1151.
168. Soper T, Mandin P, Majdalani N, Gottesman S, Woodson SA. Positive regulation by small RNAs and the role of Hfq. Proc Natl Acad Sci U S A. 2010;107: 9602–9607.
169. Kajitani M, Ishihama A. Identification and sequence determination of the host factor gene for bacteriophage Q beta. Nucleic Acids Res. 1991;19: 1063–1066.
170. Su Q, Schuppli D, Tsui HcT, Winkler ME, Weber H. Strongly reduced phage Qbeta replication, but normal phage MS2 replication in an Escherichia coli K12 mutant with inactivated Qbeta host factor (hfq) gene. Virology. 1997;227: 211–214.
171. Sukhodolets MV, Garges S. Interaction of Escherichia coli RNA polymerase with the ribosomal protein S1 and the Sm-like ATPase Hfq. Biochemistry. 2003;42: 8022–8034.

172. Kambara TK, Ramsey KM, Dove SL. Pervasive Targeting of Nascent Transcripts by Hfq. Cell Rep. 2018;23: 1543–1552.
173. Al Mamun AAM, Lombardo M-J, Shee C, Lisewski AM, Gonzalez C, Lin D, et al. Identity and function of a large gene network underlying mutagenic repair of DNA breaks. Science. 2012;338: 1344–1348.
174. Muffler A, Traulsen DD, Fischer D, Lange R, Hengge-Aronis R. The RNA-binding protein HF-I plays a global regulatory role which is largely, but not exclusively, due to its role in expression of the sigmaS subunit of RNA polymerase in Escherichia coli. J Bacteriol. 1997;179: 297–300.
175. Wachi M, Takada A, Nagai K. Overproduction of the outer-membrane proteins FepA and FhuE responsible for iron transport in Escherichia coli hfq::cat mutant. Biochem Biophys Res Commun. 1999;264: 525–529.
176. Takada A, Wachi M, Nagai K. Negative regulatory role of the Escherichia coli hfq gene in cell division. Biochem Biophys Res Commun. 1999;266: 579–583.
177. Tsui HC, Leung HC, Winkler ME. Characterization of broadly pleiotropic phenotypes caused by an hfq insertion mutation in Escherichia coli K-12. Mol Microbiol. 1994;13: 35–49.
178. Malabirade A, Partouche D, El Hamoui O, Turbant F, Geinguenaud F, Recouvreux P, et al. Revised role for Hfq bacterial regulator on DNA topology. Sci Rep. 2018;8: 16792.
179. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016;107: 1–8.
180. García-Sancho M. Biology, Computing, and the History of Molecular Sequencing: From Proteins to DNA, 1945-2000. Palgrave Macmillan; 2012.
181. Kempfer R, Pombo A. Methods for mapping 3D chromosome architecture. Nat Rev Genet. 2020;21: 207–226.
182. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326: 289–293.
183. Tiana G, Giorgetti L. Modeling the 3D Conformation of Genomes. CRC Press; 2019.
184. Oudelaar AM, Higgs DR. The relationship between genome structure and function. Nat Rev Genet. 2020. doi:10.1038/s41576-020-00303-x
185. Carroll TS, Liang Z, Salama R, Stark R, de Santiago I. Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. Front Genet. 2014;5: 75.
186. Pickrell JK, Gaffney DJ, Gilad Y, Pritchard JK. False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. Bioinformatics. 2011;27: 2144–2146.
187. Ghatak S, King ZA, Sastry A, Palsson BO. The y-ome defines the 35% of Escherichia coli genes that lack experimental evidence of function. Nucleic Acids Res. 2019;47: 2446–2454.
188. Merlin C, McAteer S, Masters M. Tools for characterization of Escherichia coli genes of unknown function. J Bacteriol. 2002;184: 4573–4581.
189. Serres MH, Gopal S, Nahum LA, Liang P, Gaasterland T, Riley M. A functional update of the Escherichia coli K-12 genome. Genome Biol. 2001;2: RESEARCH0035.

190. Crémazy FG, Rashid F-ZM, Haycocks JR, Lamberte LE, Grainger DC, Dame RT. Determination of the 3D Genome Organization of Bacteria Using Hi-C. Methods Mol Biol. 2018;1837: 3–18.
191. Hofmann A, Heermann DW. Processing and Analysis of Hi-C Data on Bacteria. Methods Mol Biol. 2018;1837: 19–31.
192. Walker DM, Freddolino PL, Harshey RM. A Well-Mixed E. coli Genome: Widespread Contacts Revealed by Tracking Mu Transposition. Cell. 2020;180: 703–716.e18.
193. Umbarger MA, Toro E, Wright MA, Porreca GJ, Baù D, Hong S-H, et al. The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation. Mol Cell. 2011;44: 252–264.
194. Regulation of heterochromatin by histone methylation and small RNAs. Curr Opin Cell Biol. 2004;16: 230–238.
195. Gonzalo S, García-Cao M, Fraga MF, Schotta G, Peters AHFM, Cotter SE, et al. Role of the RB1 family in stabilizing histone methylation at constitutive heterochromatin. Nat Cell Biol. 2005;7: 420–428.
196. Grunstein M. Histone acetylation in chromatin structure and transcription. Nature. 1997;389: 349–352.
197. Gräff J, Tsai L-H. Histone acetylation: molecular mnemonics on the chromatin. Nat Rev Neurosci. 2013;14: 97–111.
198. Görisch SM, Wachsmuth M, Tóth KF, Lichter P, Rippe K. Histone acetylation increases chromatin accessibility. J Cell Sci. 2005;118: 5825–5834.
199. Keseler IM, Mackie A, Santos-Zavaleta A, Billington R, Bonavides-Martínez C, Caspi R, et al. The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. Nucleic Acids Res. 2017;45: D543–D550.

Chapter 2

Dynamic Landscape of Protein Occupancy Across the *Escherichia coli* Chromosome

Abstract

Free living bacteria adapt to environmental change by reprogramming gene expression through precise interactions of hundreds of DNA-binding proteins. A predictive understanding of bacterial physiology requires us to globally monitor all such protein-DNA interactions across a range of environmental and genetic perturbations. Here, we show that such global observations are possible using a modification of *in vivo* protein occupancy display technology (IPOD-HR) applied to *E. coli.* We observe that the *E. coli* protein-DNA interactome organizes into two distinct prototypic features: (**1**) highly dynamic condition-dependent transcription factor occupancy by dedicated transcriptional regulators, and (**2**) condition-invariant kilobase scale occupancy by nucleoid factors, forming silencing domains analogous to eukaryotic heterochromatin. We show that occupancy dynamics across a range of conditions can rapidly reveal the global transcriptional regulatory organization of a bacterium. Beyond discovery of previously hidden regulatory logic, we show that these observations can be utilized to computationally determine sequence-specificity models for the majority of active transcription factors. Our study demonstrates that global observations of protein occupancy combined with statistical inference can rapidly and systematically reveal the

The contents of this chapter are in revision at *Plos Biology* by Peter L. Freddolino, Haley M. Amemiya, Thomas J. Goss, and Saeed Tavazoie. Conceptualization, S.T.; Methodology, P.L.F. and S.T.; Investigation, P.L.F., T.J.G., and H.A.; Data Analysis and Curation, P.L.F. and S.T.; Writing - Original Draft, P.L.F. and S.T.; Writing -- Review & Editing, P.L.F., H.A., and S.T.; Funding Acquisition, P.L.F. and S.T. All authors reviewed the manuscript. I specifically performed the no rifampin experiments (IPOD-HR assay) as an important control to examine the originally termed silent and active EPODs. I contributed to the data analysis, manuscript writing, and manuscript preparation. I also performed extensive validation experiments that will be included in the reviewed document of this manuscript, where we utilize a reporter to measure activity of the *sdaC* promoter with scrambled binding sites for *yieP*.

transcriptional regulatory and structural features of a bacterial genome. This capacity is particularly crucial for non-model bacteria which are not amenable to routine genetic manipulation.

Introduction

Transcriptional regulation plays a central role in establishing adaptive gene expression states. In bacteria, the dominant regulators are transcription factors (TFs) [1,2] and sigma factors, which direct the activity of RNA polymerase holoenzyme to a specific subset of promoters [3,4]. The phenotypic state of the bacterial cell is determined in large part by its transcriptional regulatory state which, in turn, is dictated by the binding pattern of TFs and sigma factors across the chromosome, likely in interplay with structural factors such as the local supercoiling state [5].

At present, however, our knowledge of the complete wiring of bacterial transcriptional regulatory networks remains insufficient to fully predict or design regulatory responses to arbitrary environmental conditions. The case of *Escherichia coli* serves as an illustrative case study: due to its status as a pre-eminent model organism and human pathogen, the *E. coli* transcriptional regulatory network has been an intense subject of investigation for several decades. As a result, researchers have obtained an increasingly comprehensive and detailed map of the binding specificities and physiological roles of transcriptional regulators in this organism [6]. However, roughly one quarter of the ~250 transcription factors in *E. coli* have no available binding or regulatory data [7], and many more are virtual unknowns in terms of the signals that might alter their regulatory activity. Likely as a result of this knowledge gap, Larsen and colleagues recently found that despite our broad knowledge of the potential regulatory targets of *E. coli* transcription factors, our ability to predict regulatory behavior on the basis of expression levels of transcription factors is no better than it would be for random networks. The authors attribute this partly to the fact that even when a TF is expressed, in many cases it will not bind its targets in the absence of additional signals

[8]. Furthermore, *E. coli* represents a best-case scenario in terms of our knowledge state for a bacterial transcriptional regulatory network, and for most species current databases lag far behind.

Expanding our capability to predict, and ultimately design, bacterial regulatory responses will be critical for controlling bacterial pathogenesis and engineering synthetic microbes in biotechnology applications. Achieving such a complete predictive understanding, however, requires substantial additional information both on the binding sites of as-yet uncharacterized TFs, and the actual physical occupancy of sites for both known and uncharacterized factors across conditions. Widely used methods such as ChIP-seq pose difficulties on both fronts: they demand a combinatorial explosion of experiments to study many transcription factors across a variety of conditions, and require either an antibody against each TF of interest or genetic manipulation sufficient to add an epitope tag to each target TF.

In order to significantly advance our understanding of transcriptional network dynamics and chromosomal structure, we sought to monitor, in parallel, the occupancy states of all DNA-binding proteins across a set of genetic and environmental perturbations. We argue that such comprehensive observations are critical for defining the global modes of transcriptional regulation and determining the regulatory logic that underlies adaptive reprogramming of gene expression, particularly given the importance of combinatorial logic by many factors and sites in dictating transcriptional output [9]. In order to achieve our goal, we decided to employ the concept of *in vivo* protein occupancy display (IPOD) which we, in a previous proof-of-concept study, demonstrated to reveal global occupancy of protein binding sites across the *E. coli* chromosome [10]. However, we had to introduce critical modifications and enhancements in order to deconvolve distinct contributions from sequence-specific TFs and RNA-polymerase and define binding sites at high-resolution. We will refer to this second generation IPOD technology as IPOD-HR. IPOD-HR enables efficient coverage of a large range of physiological conditions in relatively few experiments (one experiment per condition, rather than the one experiment per TF per condition that would be required for ChIP-seq). As we

demonstrate below, a single IPOD-HR experiment can reveal the occupancy dynamics of dozens of known and novel active TFs genome-wide, permitting rapid profiling of global transcriptional regulatory logic across different conditions. Furthermore, the comprehensive nature of IPOD-HR profiles enables efficient statistical inference of sequence specificity models (TFBS motifs) for active transcription factors, both recapitulating well-known regulatory logic and revealing the presence and condition-dependent activities of novel regulatory elements.

Here, we characterized the dynamics of the global protein-DNA interactome of *E. coli* across a range of three physiological conditions and three genetic perturbations. Our observations allowed us to infer, in parallel, the activities of most annotated transcription factors across conditions, and provided a catalogue of many additional likely regulatory sites and DNA sequence motifs for uncharacterized TFs. With the compact set of experiments, we reveal the dramatic regulatory dynamics of dozens of transcription factors that collectively shape the response of *E. coli* to changing environments. In sharp contrast, we find that at the kilobase scale, the genome is characterized by a set of relatively static structural domains, which consist of transcriptionally silent loci with dense protein occupancy that appear mostly constitutive across a range of physiological conditions. These regions, which we refer to as EPODs (extended protein occupancy domains) following the nomenclature of Vora *et al. [10]*, appear to act, at least partially, to suppress prophages and mobile genetic elements.

Because our approach does not rely on prior knowledge of TFs of interest or genetic manipulation of the target organism, but rather only on essential physico-chemical properties of protein-DNA complexes, we expect that it will be broadly applicable across bacterial species, even those which cannot be cultured or genetically manipulated. Our approach lays the technical and analytic foundation to rapidly characterize the regulatory and structural features of any bacterial chromosomes.

Results

Global high-resolution profiling of condition-dependent transcription factor occupancy across the *E. coli* chromosome

The IPOD-HR procedure is shown in schematic form in **Figure 2.1A**: cells are grown under a physiological condition of interest, fixed using formaldehyde, and then lysed. Heavy digestion of the chromosomal DNA provides minimized DNA fragments that may be in either a protein bound or unbound state. The protein bound DNA fragments are subsequently isolated using a phenol-chloroform extraction. Under appropriate buffer conditions, the amphipathic protein-DNA complexes are depleted from the aqueous phase and partition to a robust disc at the aqueous-organic interface [10].

As we will demonstrate below, the measurements enabled by IPOD-HR can subsequently be used for a broad range of downstream analyses, such as simultaneous monitoring of the activities of characterized TFs, large-scale inference of binding motifs for previously uncharacterized DNA-binding proteins, and identification of key occupancy sites driving previously unrecognized gene regulatory logic. To accomplish these objectives, it is essential to separate out the occupancy signal of RNA polymerase from that of specific regulatory factors of interest. Otherwise, the strong occupancy signal caused by RNA polymerase could mask changes in protein occupancy that in fact provide regulatory information (e.g., if a repressor becomes unbound but RNA polymerase subsequently binds to the same location, the occupancy signal would be nearly unchanged). To deconvolve occupancy caused by sequence-specific TFs and that of RNA polymerase, we subtract the normalized RNA polymerase ChIP-seq signal from that of the normalized raw IPOD-HR signal (see Methods for details), generating a corrected IPOD-HR profile that is a more precise representation of the cell's dynamic regulatory state (**Fig. 2.1B**). IPOD-HR is conceptually similar to the original IPOD method [10] in terms of overall workflow, but contains critical optimizations and extensions designed to permit genome-wide identification of binding by TFs and organizing factors such as nucleoid-associated proteins in a condition-specific manner; these optimizations arise both in the sample preparation procedure itself (through

enhanced washing and additional measures to track and separately factor out RNA polymerase occupancy), and an expanded analytical framework for use in inferring both existing and new regulatory logic based on whole-genome protein occupancy profiles.

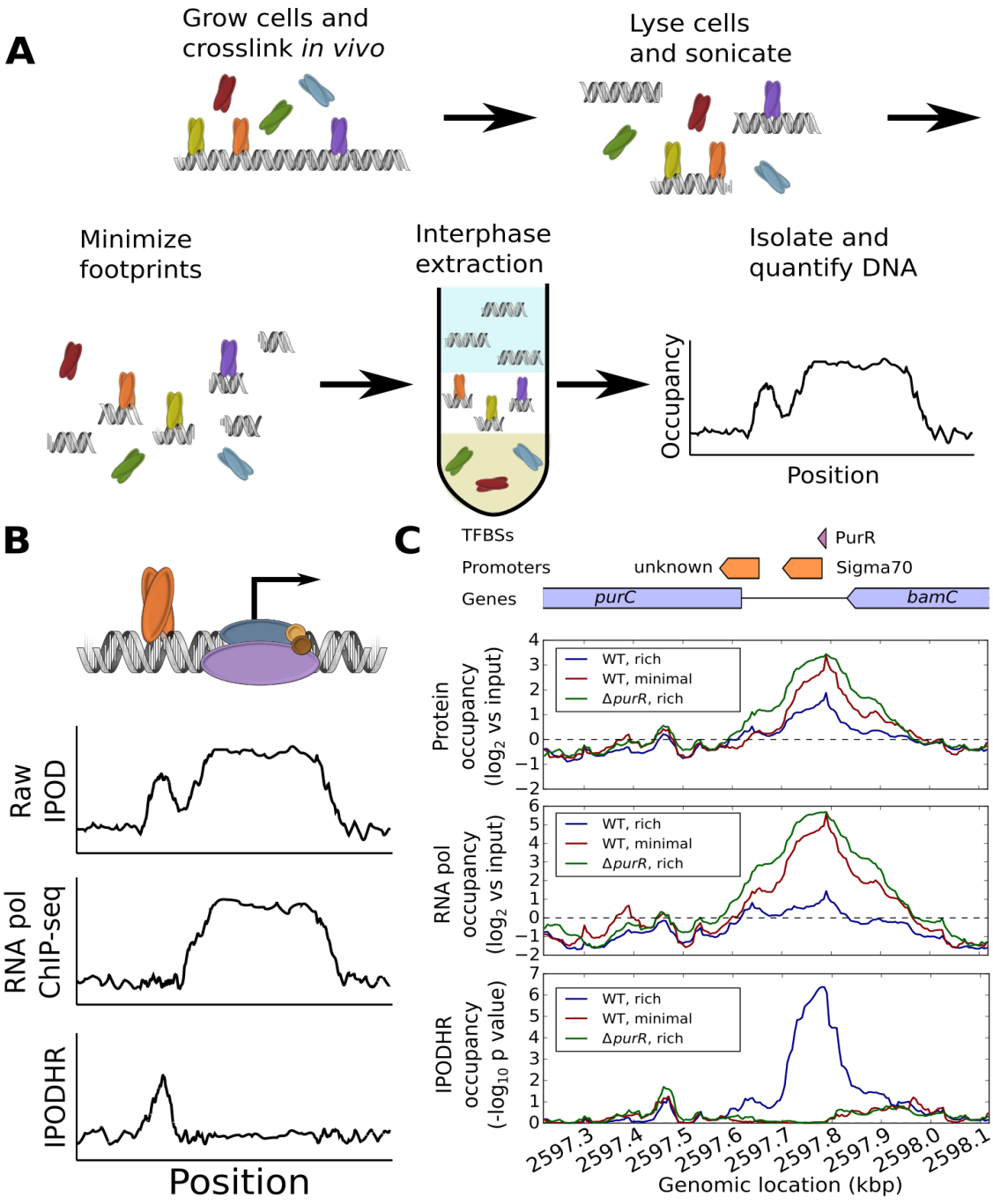

Figure 2.1: Schematic of IPOD-HR technology and detection of context-dependent binding by transcription factor PurR. (A) Overall workflow for isolation of the IPOD-HR fraction and quantification of total protein occupancy. **(B)** The final IPOD-HR signal

is obtained by subtracting a normalized RNA polymerase occupancy signal from the raw IPOD-HR protein occupancy, resulting in a polymerase-corrected signal. **(C)** Example of RNA-polymerase corrected IPOD-HR profile upstream of the *purC* gene, where subtraction of RNA polymerase occupancy from the raw IPOD-HR signal properly reveals a PurR binding site in rich media that is lost upon deletion of *purR* or transition to minimal media. In the schematic above the plots, blue regions show genes, orange regions show promoters, and purple regions show annotated transcription factor binding sites.

We note in passing that, at first glance, IPOD may seem to share superficial similarities with FAIRE (formaldehyde-assisted isolation of regulatory elements, originally described in [11]). However, FAIRE experiments were designed to detect regions of nucleosome-depleted DNA in eukaryotic chromosomes. IPOD was independently developed to detect occupancy of individual factors in prokaryotic chromosomes [10], and IPOD-HR contains further optimizations and additional experimental and computational steps to improve performance in detecting both localized and large-scale protein occupancy in bacteria.

An illustrative example of the ability of IPOD-HR to identify regulatory protein occupancy, its dynamics across conditions, and the importance of factoring out the RNA polymerase signal, is shown in **Figure 2.1C**. We consider the IPOD-HR occupancy profiles for the promoter region upstream of the *purC* gene in wild type (WT) and Δ*purR* cells during growth in rich defined medium. Based on the characterized behavior of PurR (which binds DNA in response to exogenous purine supplementation [12,13]), under this growth condition, transcription of *purC* should be repressed by binding of PurR to its promoter. However, if one considers only the raw IPOD-HR occupancy profiles (top panel), binding to the PurR site in this region is apparent in both WT and Δ*purR* cells. The resolution to this seeming paradox becomes apparent through inclusion of the correction for RNA polymerase occupancy (middle panel), which is substantially higher in Δ*purR* cells. As expected, the resulting corrected IPOD-HR occupancy profiles (bottom panel) reveal a protein occupancy peak directly on top of the annotated PurR binding site in this region in the WT cells, and no detectable occupancy in the Δ*purR* cells. This demonstrates the ability of IPOD-HR to reveal condition-dependent TF occupancy dynamics even in regions that may overlap with RNA

polymerase binding. In the following sections, IPOD-HR refers to the RNA polymerase-corrected occupancy signal, unless otherwise noted.

Local and large-scale protein occupancy patterns across the *E. coli* chromosome

To benchmark our ability to quantitatively profile protein occupancy at high spatial resolution, we performed IPOD-HR on *E. coli* cells from mid-exponential growth in rich defined medium (**Fig. 2.2A**). Over the length of the chromosome we observed a large number of small peaks, presumably corresponding to protein binding events at individual regulatory sites. In addition, we observed many large-scale (> 1 Kb) regions of high occupancy which we refer to as extended protein occupancy domains (EPODs), following the nomenclature of (Vora et al., 2009). An example of condition-dependent changes in binding of local TFs is shown in **Fig. 2.2B-C**. Examination of a ~50 kb slice of the genome reveals dozens of small occupancy peaks, with a visually apparent enrichment in intergenic regions (**Fig. 2.2B**). Many such peaks, which presumably correspond to individual protein binding events, coincide with known TF binding sites (TFBSs). For example, the region upstream of *argA* (**Fig. 2.2C**) shows strong occupancy at known ArgR binding sites, and condition-appropriate occupancy dynamics including weakening of binding in arginine-poor conditions [14,15] and loss of occupancy upon deletion of the *argR* gene. At the same time, similar occupancy patterns can be observed at many sites lacking an annotated TFBS, as seen in **Fig. 2.2D**, where conditionally dynamic binding sites are apparent upstream of *lgt* and *rppH*. These peaks likely indicate the presence of previously unrecognized TFBSs, as we will discuss in more detail below. As expected, at a genome-wide scale IPOD-HR signals show both higher occupancy in intergenic regions relative to coding regions, and higher occupancy at annotated TFBSs relative to other regions of the chromosome (**Fig. 2.2E**), demonstrating a strong overlap of the observed protein occupancy with transcriptional regulatory sites. Indeed, applying peak calling to the IPOD-HR signal demonstrates an increasingly strong overlap with known TFBSs as the threshold for peak calling is increased (**Fig. S2.1**).

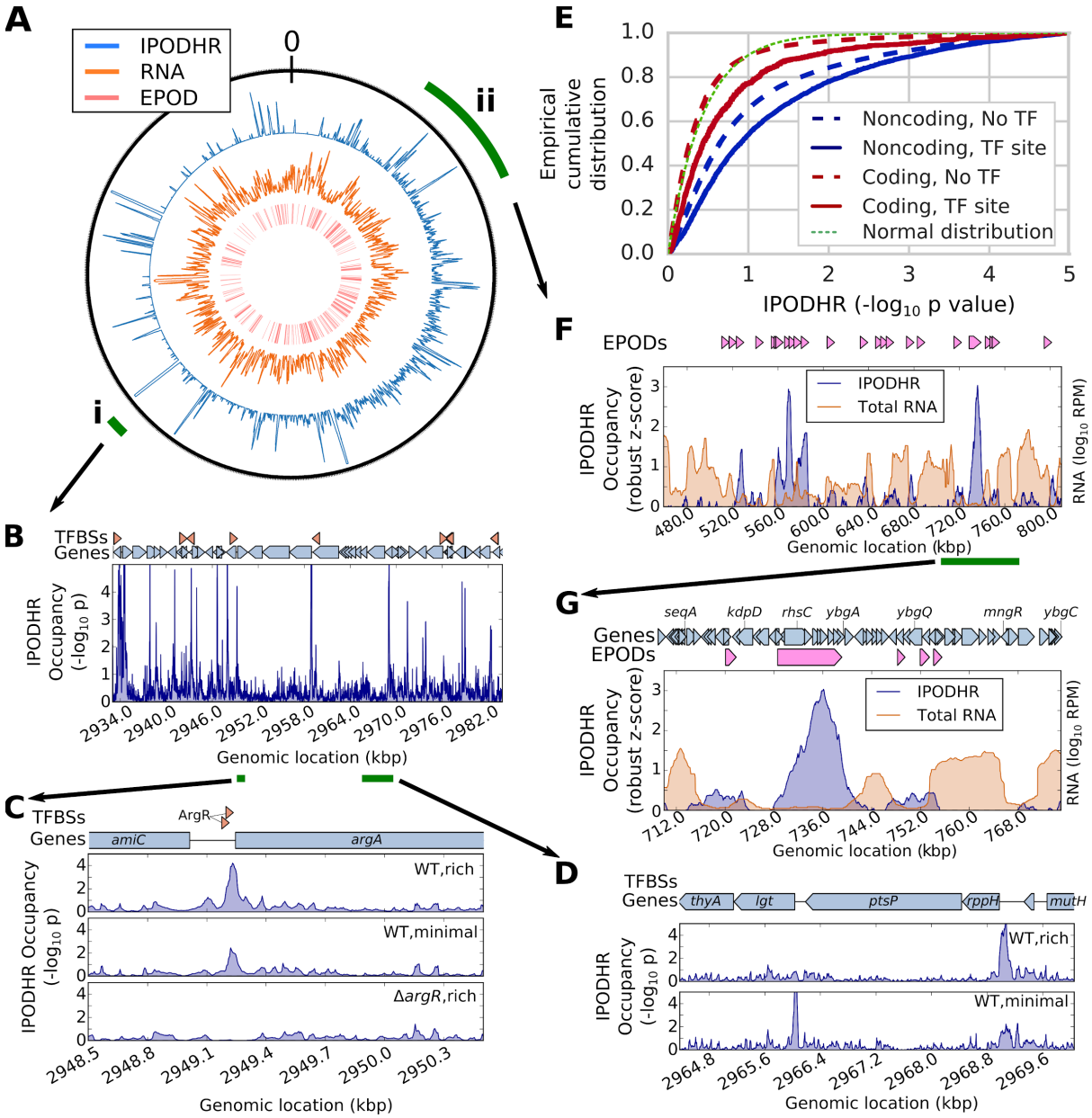

Figure 2.2: IPOD-HR profiles reveal rich high-resolution occupancy dynamics and large-scale structural features across the chromosome. (A) *Outer track:* IPOD-HR occupancy (robust Z-scores, 5 kb moving average); *middle track:* total RNA read density (5 kb moving average); *inner track*: locations of inferred EPODs. The outer green wedges mark the portion of the chromosome shown in subsequent panels. The origin of the coordinate system is oriented at the top of the plot. **(B)** IPOD-HR occupancy measured during growth in glucose rich defined medium, in the vicinity of wedge i from panel **A**. Green segments below the genomic coordinates indicate the regions highlighted in panels **C-D**. **(C)** Condition-dependent occupancy changes at the ArgR binding sites upstream of *argA*. **(D)** Identification of condition-specific occupancy of likely TFBSs upstream of *lgt* and *rppH*. **(E)** Cumulative histograms showing RNA polymerase ChIP-subtracted IPOD-HR occupancy in coding vs. noncoding regions, and at sites that match known transcription factor binding sites from RegulonDB [7],

compared with the curve that would be expected from a standard normal distribution of scores. **(F)** Occupancy (blue) and total RNA abundance (orange) for a selected sector of the genome (wedge ii from panel **A**), showing the presence of several EPODs in regions corresponding to low RNA abundance; rolling medians over a 5 kb window are plotted, with RNA read densities shown in units of reads per million (RPM). (**G)** Magnification of the region highlighted by the green bar in panel F, illustrating a silenced region in and around *rhsC*, alongside flanking areas of low IPOD-HR occupancy and high gene expression. A 5 kb rolling median is plotted.

It is also apparent by inspection of the genome-wide occupancy shown in **Fig. 2.2A** that many extended regions of high protein occupancy coincide with regions of relatively low transcription. For example, in **Figure 2.2F**, we show a typical ~300 kb region with alternating segments of high protein occupancy that have relatively low transcription, with those of low protein occupancy and relatively high transcription (also apparent in the higher-resolution plot in **Fig. 2.2G**. Thus, in addition to revealing occupancy at the level of individual regulatory sites, IPOD-HR enables tracking of the behavior of large, densely protein occupied regions of the chromosome that appear to coincide with transcriptionally silent loci. We will explore both of these prototypic classes of occupancy, in more detail, below.

Transcription factor and sigma factor occupancy dynamics across genetic and environmental perturbations

Since IPOD-HR occupancy profiles show highly enriched overlaps with known TFBSs (**Fig. 2.2E**), we asked whether IPOD-HR profiles can be used to reveal the occupancy dynamics for known *E. coli* TFs across a set of conditions. Indeed, we find that IPOD-HR reveals consistent and condition-appropriate regulatory logic at the level of individual regulons, and patterns of regulatory behavior across regulons. As expected, strains with each of three single TF deletions (*argR*, *lexA*, *purR*) show global loss of occupancy at the ensemble of annotated sites for the corresponding TFs (**Fig. 2.3A**). Analysis of condition-dependent changes in the occupancy of binding sites for single transcription factors likewise recapitulates expected behavior; for example, ArgR [16], PurR [12,17], and TyrR [18] all show enhanced binding to DNA in the presence of amino acid and/or nucleobase ligands which are supplied directly in our rich media

conditions, and the IPOD-HR occupancy signal shows global loss of occupancy for binding sites of all three of these TFs in nutrient depleted conditions (minimal media and stationary phase) when compared with exponential growth in rich media (**Fig. 2.3A**). In contrast, RutR shows increased overall occupancy in minimal media relative to rich media conditions, consistent with the known inhibition of RutR binding by thymine and uracil [19]. Sites for MetR, which is dependent upon homocysteine as a co-regulator [20], likewise show large increases in occupancy in minimal media and in stationary phase (**Fig. 2.3A**). As homocysteine is the final intermediate in methionine biosynthesis [21], its levels would naturally be expected to rise upon methionine starvation, and consistently, transcript levels of the canonical homocysteine-dependent MetR-activated target *metE [22]* rise six-fold in our stationary phase RNA-seq data and more than 100-fold in our minimal media RNA-seq data (data not shown).

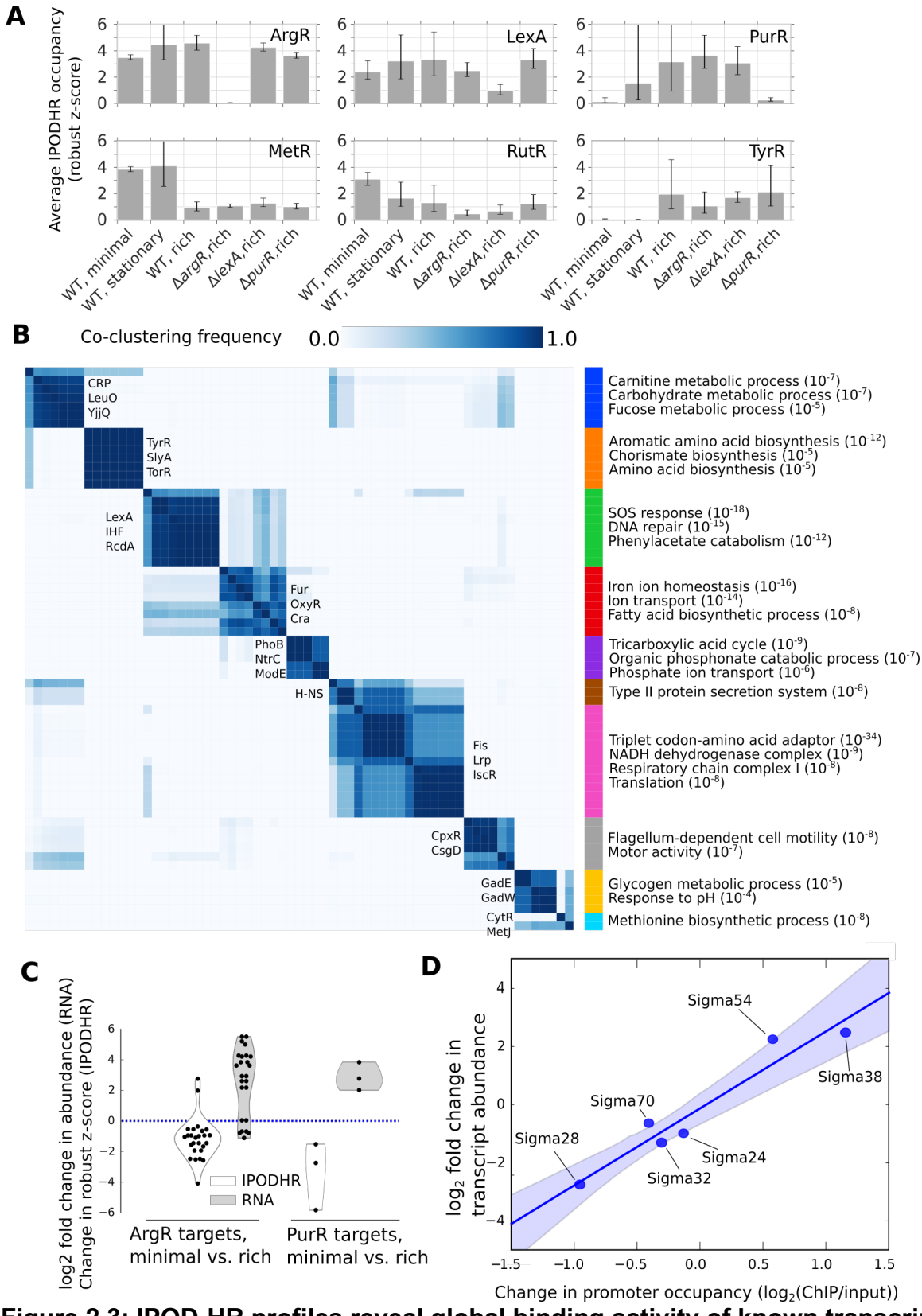

Figure 2.3: IPOD-HR profiles reveal global binding activity of known transcription factors and sigma factors. (**A**) Average (geometric mean) occupancies for all

annotated binding sites of the six indicated transcription factors under each indicated condition. Error bars indicate a 95% confidence interval based on parametric bootstrapping with pessimistic assumptions; see Methods for details. (**B**) Heat map showing the consensus clustering (co-occurrence frequencies) of the pattern of occupancy dynamics for the regulons of all considered TFs across the varied nutrient conditions in this study (see Methods for details). Consensus division into ten clusters via agglomerative clustering is shown at right; for each cluster, representative TFs (on matrix) and regulated gene ontology terms (right) are shown, with numbers in parentheses indicating the $\log_{10}$ p-value for enrichment of that GO term. **(C)** Changes in occupancy and target gene transcript level for all annotated repressive binding sites of ArgR and PurR (for minimal media vs. rich media), in each case demonstrating the strong anti-correlation of binding and regulatory effects across the regulons. (**D**) Correlation of promoter-level occupancy changes (measured by RNA polymerase ChIP-seq) and changes in transcript abundance, shown for the WT stationary phase condition compared with exponential phase. Shaded area shows a bootstrap-based 95% confidence interval.

By applying an unsupervised clustering approach (see Methods for details), we identified transcriptional regulatory modules that show consistent co-regulation across the conditions in our study. We found clustering of TFs with highly similar behavior (**Fig. 2.3B**) that coordinate, for example, amino acid metabolism (orange), the core translational apparatus (pink), and iron homeostasis (red). We also observe several cases where regulatory cascades are clustered together; for example, YjjQ and its transcriptional activator LeuO (blue), or the tightly intertwined acid response regulators GadE, GadW, and GadX (yellow). We thus find that IPOD-HR occupancy profiles can provide detailed, site-level, condition-specific information on regulatory protein occupancy across the entire chromosome. By comparing changes in protein occupancy with changes in transcript levels across conditions, we can relate changes in protein occupancy to their positive or negative regulatory consequences. This can be seen for two nutrient sensing transcriptional repressors with sites annotated in RegulonDB, ArgR and PurR, across changes in nutrient conditions (**Fig. 2.3C**). As expected, given the annotated repressive role of these sites, in both cases we observe a strong anti-correlation between changes in protein occupancy and target transcription.

Since each IPOD-HR global protein occupancy data set is performed alongside an RNA polymerase ChIP-seq experiment, we can easily track promoter occupancy alongside TFBS occupancy. The use of rifampin permits transcriptional initiation, but prevents

promoter clearance. Thus, these data sets are ideal for identifying regulation at the level of RNA polymerase (e.g. via different sigma factors). The differential patterns of RNA polymerase occupancy show strong correlations with transcript levels for each Sigma factor's regulon across a range of conditions. As shown in **Figure 2.3D**, when comparing logarithmic vs. stationary phase conditions, the changes in transcript abundance and RNA polymerase promoter occupancy show a Spearman correlation of 0.83 ($p=0.042$); a similar comparison for changes in occupancy vs. expression for cells grown in minimal media yields equivalent results (data not shown).

Global occupancy dynamics reveals the action of new DNA binding proteins

Despite extensive annotation efforts, at present fewer than 1,100 of the 3,560 annotated transcriptional units present in the RegulonDB database have any annotated regulation by transcription factors assigned to them [7]. While several recent notable efforts have sought to expand the completeness of these regulatory annotations by studying the DNA binding preferences of purified TFs [6,23,24], or via computational inference of likely additional regulation [25] and regulatory modules [26], none of these methods provides either direct evidence for binding *in vivo*, or information on condition-dependent changes in occupancy. IPOD-HR, in contrast, can provide both. Furthermore, the protein occupancy signals thus obtained provide information on occupancy of both well-characterized and uncharacterized proteins. In fact, a large fraction of dynamic IPOD-HR peaks occur in promoters with no previous annotation for TF binding sites, as we will discuss in detail in the following section.

A representative example of an orphan occupancy peak is seen upstream of the gene *sdaC* (**Fig. 2.4**). In our RNA-seq data, *sdaC* transcript levels are nearly twenty-fold higher during exponential growth in rich media (317.3 transcripts per million (TPM)) compared with either exponential growth in minimal media (17.9 TPM) or stationary phase in rich media (16.7 TPM). Despite a lack of annotated TFBSs upstream of *sdaC*, IPOD-HR occupancy profiles (**Fig. 2.4A**) show a likely transcriptional activator binding site upstream of the *sdaC* core promoter, which shows strong occupancy in the WT

M9/RDM/glu conditions but not the related conditions where *sdaC* expression is lower. To identify the transcription factor(s) responsible for that occupancy, we used a biotinylated bait DNA matching the sequence of the *sdaC* promoter region to isolate proteins bound to that region from *E. coli* cells grown in the WT M9/RDM/glu condition (**Fig. 2.4B**). Mass spectrometry on isolated bait-dependent bands revealed two poorly characterized transcription factors, UlaR and YieP, that showed highly enriched binding to the *sdaC* promoter (see **Table S2.1**). While UlaR proved difficult to purify due to poor solubility, and was thus excluded from further analysis, we found that purified YieP does indeed show specific shifting of the *sdaC* promoter in an electrophoretic mobility shift assay (**Fig. 2.4C**). Consistently, recent RNA-seq data on a Δ*yieP* strain shows a significant drop in *sdaC* transcript levels (2.7-fold change; $q=7.6 * 10^{-18}$) relative to isogenic cells with a plasmid-born reintroduction of YieP during growth in LB media (C. Bianco and C. Vanderpool, personal communication).

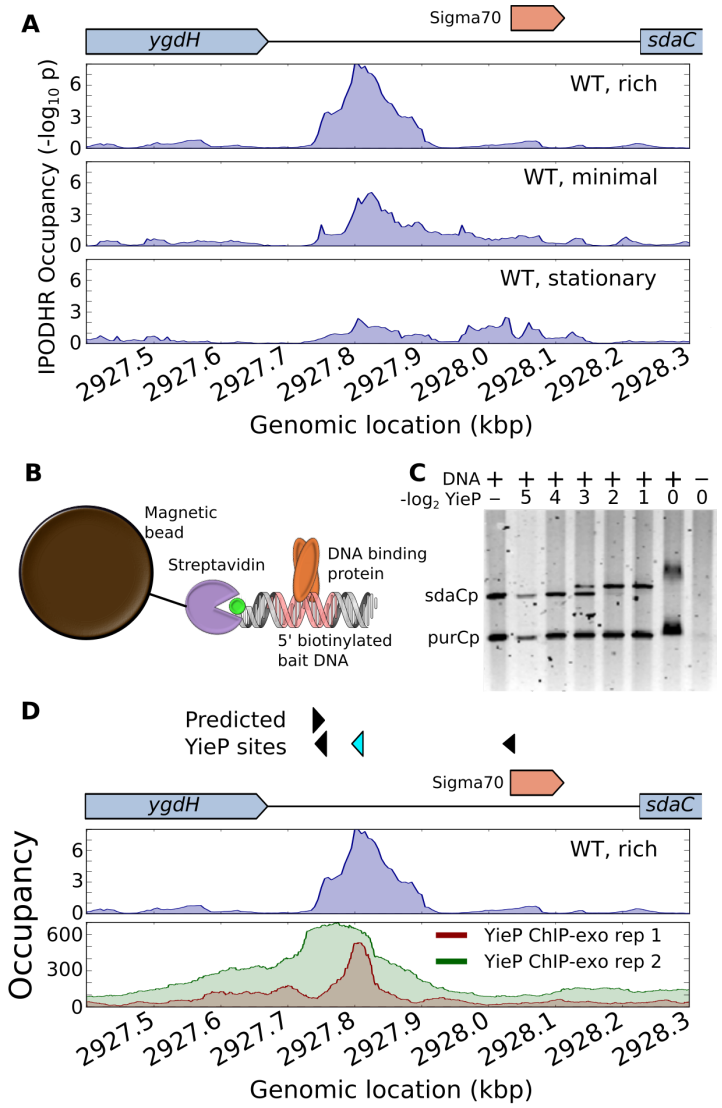

Figure 2.4: Experimental identification of the protein bound to a novel occupancy peak upstream of the *sdaC* promoter. (A) IPOD-HR profiles upstream of *sdaC* in rich (M9/RDM/glu) media, minimal (M9/glu) media, and in rich media in stationary phase. No annotated transcription factor binding sites are present in the displayed region. **(B)** Schematic of pulldown/mass spectrometry experiments used to identify factors binding the *sdaC* promoter. **(C)** Gel shift experiments showing specific interaction of YieP with the *sdaC* promoter. Increasing concentrations of purified His_6-YieP are incubated with a mixture of fluorescein-labeled promoter regions from *sdaC* and *purC* and then run on a gel, demonstrating specific shifting of the *sdaC* promoter region. YieP concentrations are given as the number of two-fold dilutions relative to full strength. **(D)** Comparison of IPOD-HR occupancy profile (as in panel **A**) with ChIP-exo data from [25], with the latter given as total read counts (parsed from GEO accessions GSM3022131 and GSM3022132). The top track of predicted YieP sites shows significant hits for the YieP motif identified based on that ChIP-exo data set. Out of 1,025 potential YieP sites in the genome, the location highlighted in cyan is tied for 10th highest score (identified using

FIMO; see Methods for details). Occupancy signal is given as $-\log_{10}(p)$ for the IPOD-HR track, or raw counts (averaged across strands) for the ChIP-exo tracks.

YieP was recently (and independently) selected by Palsson and co-workers as a validation case to be used in their consideration of computational methods for identifying the binding sites of orphan TFs, and subjected to ChIP-exo analysis on cells grown in glucose minimal media using epitope-tagged YieP [25]. Indeed, their data demonstrate both strong direct YieP occupancy, and a high confidence YieP motif match, at the precise position of the occupancy peak detected in our IPOD-HR data set (**Fig. 2.4D**). Based on the relative intensity at that position across conditions, combined with the expression data noted above, we infer that YieP binds to the *sdaC* promoter in nutrient-replete conditions and acts as a transcriptional activator (explaining the solitary strong peak in our "WT,rich" condition), whereas in other conditions, YieP binding is weakened (but not abolished) and additional factors likely bind downstream of the YieP site to repress *sdaC* transcription. We must emphasize that the discovery of YieP binding sites through IPOD-HR and subsequent mass spectrometry experiments (by us) occurred in parallel with the ChIP-exo experiments of Gao and colleagues, and indeed, represent highly complementary paths for identification of the binding sites for orphan transcription factors, with one centered on a candidate protein and the other on candidate sites.

The example presented here of regulation of *sdaC* by the uncharacterized transcription factor YieP highlights the broad potential for using IPOD-HR to rapidly identify and characterize previously cryptic regulatory connections. IPOD-HR thus complements the multitude of other approaches noted above (based on, e.g., promoter libraries or computational inference), and provides the unique benefit of directly assessing binding to DNA *in vivo*, at native loci, under physiological conditions of interest.

The utility of IPOD-HR in identifying the activity of previously un-characterized transcription factors motivates its extension to a genome-wide scale, providing an *in vivo* complement to high-throughput *in vitro* screening methods such as genomic SELEX [6]. By applying peak calling to our IPOD-HR data sets across the six conditions

considered in the present study, we were able to identify thousands of likely TFBSs, many of which are not identifiable based on existing databases. To compare the peak sets identified from IPOD-HR data with our existing state of knowledge, we divided the peak calls obtained from IPOD-HR into a set of annotated TFBSs from RegulonDB, and a set of binding sites predicted using all known PWMs available in the SwissRegulon database (see Methods for details). We find that across a range of thresholds, approximately half of the binding sites identified by IPOD-HR overlap with either known or predicted sites, whereas the other half represent novel binding sites which likely (as in the case of the YieP site described above) reflect the activity of poorly annotated or orphan TFs. Pooling the newly identified binding sites across conditions, our IPOD-HR data sets are able to provide a total of 19,068 putative TFBSs which are occupied *in vivo* under at least one condition (and track the dynamics of occupancy of those sites across conditions). This extensive map of chromosomal occupancy and its dynamics provides the community with a wealth of known and putative novel regulatory interactions that can be further explored and validated by follow-up experiments such as those shown in **Fig. 2.4**.

Global *de novo* discovery of sequence-specificity motifs for active transcription factors

While the peak calls obtained from IPOD-HR data show strong enrichments with known TFBSs (**Fig. S2.1**), roughly half of the called peaks do not match any known or predicted transcription factor binding sites (as detailed in **Figure 2.5A**), and likely correspond either to unknown sites for well-characterized TFs or binding sites for previously uncharacterized TFs. Given that the majority of the newly inferred binding sites appear not to correspond to known or predicted sites for annotated TFs, we hypothesized that the regulons corresponding to those motifs would likely show enrichments for poorly annotated genes, as we expect here to reveal the regulatory logic driving typically under-studied pathways. We thus identified likely regulatory targets of each newly called peak, divided them between poorly annotated genes (those with UniProt annotation scores of 1 or 2 out of 5 [27]) and well-annotated genes, and

then examined the proportion of poorly annotated targets for occupancy peaks matching RegulonDB binding sites compared with all other peaks. As shown in **Figure 2.5B**, peaks that do not correspond to RegulonDB-annotated binding sites are strongly enriched upstream of poorly annotated genes, whereas those matching annotated binding sites are enriched for well-annotated genes. Thus, examination of occupancy peaks derived from IPOD-HR enables identification of a large number of new putative regulatory sites, with a particular abundance of possible regulators of poorly-annotated genes.

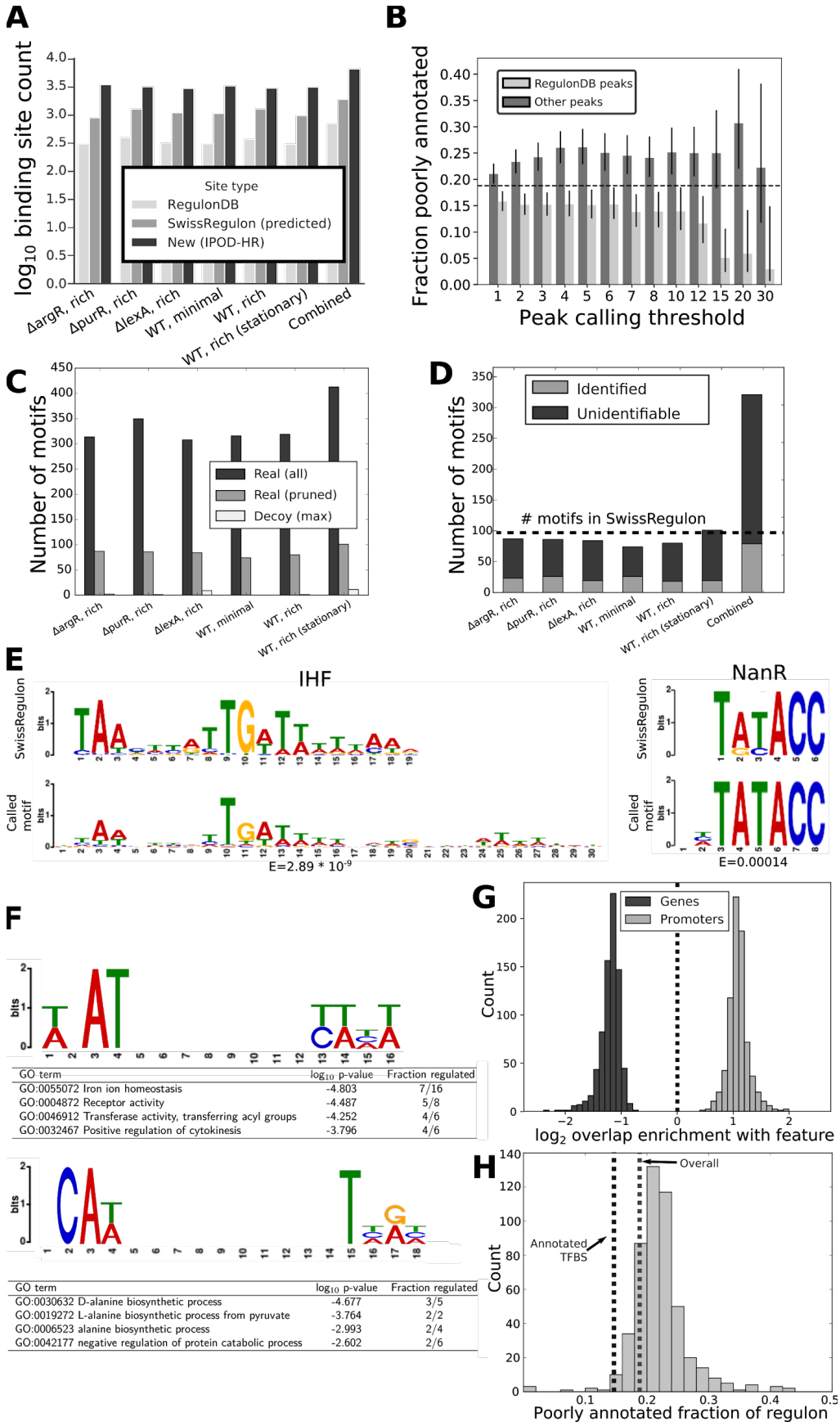
A
log10 binding site count
Site type
RegulonDB
SwissRegulon (predicted)
New (IPOD-HR)
ΔargR, rich
ΔpurR, rich
ΔlexA, rich
WT, minimal
WT, rich
WT, rich (stationary)
Combined
B
Fraction poorly annotated
RegulonDB peaks
Other peaks
Peak calling threshold
C
Number of motifs
Real (all)
Real (pruned)
Decoy (max)
D
Number of motifs
Identified
Unidentifiable
motifs in SwissRegulon
E
IHF
SwissRegulon
Called motif
E=2.89 * 10-9
NanR
E=0.00014
F
GO term | log10 p-value | Fraction regulated
GO:0055072 Iron ion homeostasis | -4.803 | 7/16
GO:0004872 Receptor activity | -4.487 | 5/8
GO:0046912 Transferase activity, transferring acyl groups | -4.252 | 4/6
GO:0032467 Positive regulation of cytokinesis | -3.796 | 4/6
GO term | log10 p-value | Fraction regulated
GO:0030632 D-alanine biosynthetic process | -4.677 | 3/5
GO:0019272 L-alanine biosynthetic process from pyruvate | -3.764 | 2/2
GO:0006523 alanine biosynthetic process | -2.993 | 2/4
GO:0042177 negative regulation of protein catabolic process | -2.602 | 2/6
G
Genes
Promoters
Count
log2 overlap enrichment with feature
H
Overall
Annotated TFBS
Count
Poorly annotated fraction of regulon

Figure 2.5: Genome-wide *de novo* discovery of sequence specificity motifs for actively bound transcription factors. (**A**) At a peak calling threshold of 4 (c.f. **Figure S2.1**), we show the number of identified binding sites that overlap with annotated sites in RegulonDB ("RegulonDB"), motif-based predicted binding sites ("SwissRegulon"), or novel ("New"). The "Combined" category represents peak sets where the peaks at a given threshold identified across all conditions are merged, prior to comparisons with the RegulonDB and predicted databases. Qualitatively similar results are observed at all tested peak calling thresholds (data not shown). (**B**) All called IPOD-HR occupancy peaks across the conditions shown in panel **A** were combined, and then partitioned based on whether they overlap with a known or inferred binding site in RegulonDB (RegulonDB peaks) or not (Other peaks). Peaks were then considered to have regulatory potential if they fell within 50 bp of an annotated transcription start site, and the fraction of the genes potentially regulated by each peak category plotted across different peak calling threshold. Error bars show 95% credible intervals calculated assuming that the incidence of poorly annotated genes in the inferred regulon is a binomial random variable, using Bayesian inference with a Beta(1,1) prior. The dashed line shows the overall fraction of poorly annotated genes included in the analysis (that is, those belonging to transcripts regulated by at least one annotated transcription start site in RegulonDB). (**C**) Number of motifs discovered *de novo* using IPOD-HR occupancies under each condition in our study. "All" and "pruned" refer to all discovered motifs and those surviving cluster-based filtering by RSAT (see Methods for details), respectively. "Real" shows the motif counts discovered in real data, and "Decoy" shows the maximum discovered motif count across 20 independent circular permutations of the data under each condition. (**D**) Classification of non-redundant motifs across conditions as "Identified" (match to an existing motif from the SwissRegulon database, via TOMTOM, with E-value < 0.5) or "Unidentified" (no matches found with E<0.5). "Combined" refers to the full set of motifs discovered after pooling all motifs across all conditions and redundancy filtering; a horizontal dashed line shows the total number of known motifs present in SwissRegulon. (**E**) Example cases of "Identified" matches of IPOD-HR-inferred motifs with motifs from the SwissRegulon database, showing good correspondence with annotated IHF (left) and NanR motifs. E values arising from the TOMTOM search pairing newly discovered motifs with similar known motifs are shown beneath each inferred motif. y axes for motifs in this and the following panel show information content in bits. (**F**) Examples of two newly inferred motifs that do not have identifiable hits in the SwissRegulon database (as assessed using TOMTOM). In each case, the gene ontology (GO) terms showing most significant enrichments amidst the predicted regulon associated with that motif are shown (see Methods for details). (**G**) Overlap of predicted binding sites for IPOD-HR inferred motifs with either coding regions (genes) or promoters (both as annotated in RegulonDB); shown are the $\log_2$ fold enrichment or depletion of the overlap as compared with that expected by chance. (**H**) For the predicted regulon of each newly inferred motif, we show the fraction of regulon members that are poorly annotated (Uniprot annotation score of 1 or 2 out of 5); for comparison, dashed lines are shown for the values obtained when the same statistic is calculated for all annotated TF-gene interactions in RegulonDB ("Annotated TFBS"), and for the genome as a whole ("Overall").

Our large-scale identification of new TFBSs also raises the important possibility that new TF binding motifs might likewise be identifiable through *de novo* computational motif discovery in the set of all sequences within IPOD-HR peaks. Indeed, the application of the FIRE [28] motif discovery algorithm to peak locations obtained from IPOD-HR data reveals dozens of *de novo* discovered sequence motifs that are informative of strong occupancy sites, even after pruning of redundant motifs (**Fig. 2.5C**). Upon cross-referencing with a database of known *E. coli* TFBS motifs using TOMTOM [29], we find that approximately 25% of the discovered motifs can be matched with known motifs (87/97 of the annotated motifs in the *E. coli* SwissRegulon database are matched by at least one inferred motif from the set present prior to redundancy pruning, and 63/97 match at least one motif present in our inferred set after pruning), while at the same time nearly 200 novel motifs are called with similar confidence (**Fig. 2.5D**). To provide estimates of the false discovery rate (FDR) arising from our motif inference, we performed an identical motif discovery procedure for each biological condition on 20 “decoy” data sets in which the underlying *E. coli* genomic sequence was rotated by a random distance relative to the peak calls, thus preserving the correlation structure of both the data and sequence with respect to themselves (light bars in **Figure 2.5C**). Our decoy data sets gave rise to no more than 11 motifs under any condition, and usually far fewer, giving rise to an average effective FDR (across shuffles and conditions) of 0.2% for the unpruned motifs or 1.0% for the pruned motifs. Using only the novel motifs (that is, motifs which did not have detectable similarity to any motifs in the SwissRegulon database) in a genome-wide search for potential binding sites using FIMO, we find that 32.1% of all IPOD occupancy peaks at a peak calling threshold of 4 can be explained by binding sites for the novel motifs, compared with 9.3% that can be explained by annotated binding sites from RegulonDB. Thus, the newly inferred motifs provide a substantially expanded ability to assign the observed profile of protein binding across the chromosome. Of the occupancy peaks not identifiable based on either our novel motifs or RegulonDB binding sites, 43.7% fall in EPODs, and thus are attributable to locally concentrated binding of the EPOD constituent proteins (likely H-NS and other nucleoid-associated proteins, as discussed

elsewhere). Taken together, the combination of newly called motifs, known binding sites from RegulonDB, and EPODs accounts for 64.3% of all occupancy peak locations at a calling threshold of 4.

In **Figure 2.5E** we show two representative examples of discovered motifs that show strong matches with annotated motifs, demonstrating that the motifs for well-characterized transcriptional regulators such as IHF and NanR can be inferred directly from IPOD-HR data. For comparison, in **Figure 2.5F**, we show two newly inferred motifs that do not match any known motifs in the *E. coli* SwissRegulon database. Intriguingly, the pattern of binding sites across the *E. coli* chromosome for both of these novel motifs illustrates a potential regulatory function, with the first motif associating with a substantial fraction of the genes involved in iron ion acquisition, and the second apparently involved in alanine metabolism (and in particular synthesis of the cell wall constituent D-alanine).

We further assessed the regulatory capacity of all newly called sequence motifs by comparing their genome-wide distribution of binding sites with annotated genes (coding regions) and promoters. We would expect that binding sites for functional transcriptional regulators would be enriched within promoters and depleted from coding regions, as was the case for overall IPOD-HR occupancy (**Fig. 2.2E**). Indeed, the overlap distributions of binding sites for our newly inferred motifs are uniformly enriched for annotated promoters and depleted for ORFs (**Fig. 2.5G**), demonstrating that motifs inferred directly from IPOD-HR occupancy data occur primarily in likely regulatory regions. Equivalent results were obtained even after excluding all of the newly inferred motifs with identifiable similarity to SwissRegulon motifs (as assessed using TOMTOM; data not shown).

Given that the majority of the newly inferred motifs appear not to correspond to annotated TFs, and our findings above regarding the enrichment of poorly annotated genes downstream of orphan binding peaks, we hypothesized that the regulons corresponding to our newly inferred motifs would likely show enrichments for poorly

annotated genes, as we expect here to reveal the regulatory logic driving typically under-studied pathways. We thus calculated the fractions of the hypothetical regulons of each newly inferred motif that consist of poorly annotated genes (defined as noted above). As shown in **Figure 2.5H**, we found that the regulons of the newly inferred motifs were significantly enriched for poorly annotated genes when compared with both the annotated *E. coli* transcriptional regulatory network in RegulonDB ($p<2.2*10^{-16}$, Wilcoxon signed rank test), and the overall average rate of poorly annotated genes throughout the chromosome ($p<2.2*10^{-16}$, Wilcoxon signed rank test). Taken together, we see that IPOD-HR enables inference of a large number of sequence motifs, many of which likely correspond to functional, but currently under-studied, transcriptional regulators in *E. coli*, providing a substantial resource for ongoing investigation of this transcriptional regulatory network.

Extended protein occupancy domains define distinct and largely stable transcriptionally silent regions with unique sequence features

One of the most striking findings enabled by the original application of IPOD was the discovery of extended protein occupancy domains (EPODs): large regions of the *E. coli* chromosome that show unusually dense levels of protein occupancy over kilobase or longer scales [10]. EPODs are also clearly apparent in all our IPOD-HR data sets, and appear to correspond functionally to the transcriptionally silent tsEPODs of Vora *et al. [10]*. The profile of protein occupancy and EPODs, along with the accompanying impacts on transcript levels, for a representative region of the genome is shown in **Figure 2.6A**. Indeed, we found that many highly protein occupied regions measured using the original IPOD method (in particular, the highly expressed extended protein occupancy domains, or heEPODs) represent RNA polymerase occupancy, whereas the EPODs now revealed by IPOD-HR consist solely of large domains of occupancy by proteins other than RNA polymerase (which typically also exclude RNA polymerase). We discuss these differences and the details of the approach used in IPOD-HR to remove contributions from RNA polymerase in **Text S2.1.** The specific resolution of tsEPODs afforded by the IPOD-HR method, and the coverage of multiple genetic and

nutrient perturbations in the present data sets, allow us to fully investigate the nature and condition-dependent occupancy of these chromosomal structures.

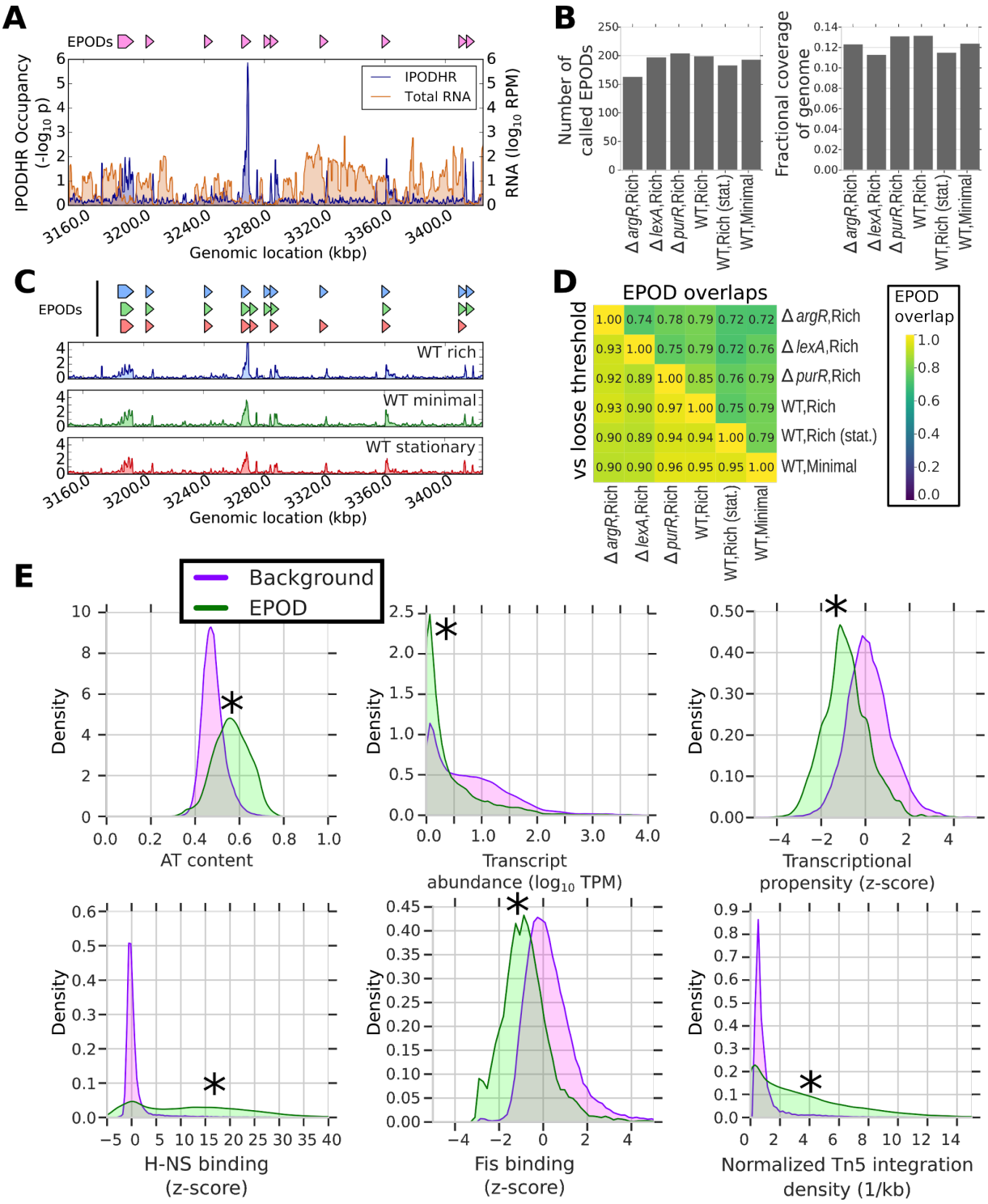

Figure 2.6: EPODs define stable genomic structures and are associated with many distinct features. (**A**) EPOD calls from a representative genomic region in the WT Rich media condition, along with protein occupancy and RNA levels smoothed with

a 1 kb rolling median. (**B**) Number of called EPODs by condition (left) and fraction of the genome covered by EPODs (right). (**C**) IPOD-HR occupancies (shown over a 1 kb rolling median) and associated EPOD calls under three different conditions, in the same genomic region shown in panel **A**. EPOD calls are shown above the occupancy, in the same order as the data tracks. (**D**) Upper triangle: Overlap of EPOD calls (see text for definition) between each pair of the studied conditions. In the lower triangle, each entry shows the fraction of the EPOD calls (at a 5 bp resolution) from the sample defining that row that is contained in a relaxed set of EPOD calls (see text) of the sample defining that column. (**E**) Density plots showing normalized histograms (smoothed by a kernel density estimator) of the specified quantities for regions of the genome that are in EPODs versus those that are not (Background), as assessed in the WT M9/RDM/glu (WT, rich) condition. '*' indicates FDR-corrected $p<0.005$ via a permutation test (against a null hypothesis of no difference in medians). Significance calling and additional comparisons are shown in **Table S2.2**

The identified EPODs show remarkable stability (**Fig. 2.6B**), with ~180 EPODs in each condition, and similar fractions of the genome contained in EPODs in each case. Furthermore, the locations of individual EPODs are likewise well maintained, even across very different physiological conditions. For example, in **Figure 2.6C** we show IPOD-HR occupancy across the same region as shown in **Figure 2.6A**, comparing exponential growth in rich vs. minimal media, and stationary phase cells. In contrast with the condition-dependent occupancy of individual TFs, at the ~kilobase scale the occupancy traces are nearly superimposable, and show that most EPODs called under the various conditions overlap. Furthermore, out of the subset of EPOD calls that are missing from the 'WT,Rich' condition but present in the others, all but one are also present among calls made in the 'WT,Rich' condition using a relaxed threshold, suggesting that the small differences in EPOD locations that do appear between EPOD calls under different conditions are in fact due to thresholding effects. We observe the same trends genome-wide: 72-85% of genomic locations (at the base pair level) that are called as EPODs under any one condition are likewise EPOD calls under any other condition (**Fig. 2.6D**); furthermore, at least 89% (and typically much more) of the EPODs called in one condition are contained within the relaxed threshold calls under any other condition (*n.b.* the 'relaxed' threshold used here corresponds with the original EPOD definition from [10]). It is also worth noting, in this context, that 90% of the tsEPOD-occupied locations from [10] are contained within the new "WT,Rich" relaxed

threshold EPOD set, in line with the observed concordance across experimental conditions in our new data sets.

Several defining characteristics of EPODs are readily apparent upon cross-referencing with other genome-wide datasets (**Fig. 2.6E**): they represent regions of high AT content, which are both associated with low levels of native transcripts and decreased transcriptional propensity (that is, expression of standardized integrated reporters [30]). Consistent with our original findings [10], EPODs also show high occupancy of H-NS, HU, and LRP; low occupancy of Fis; and are associated with high efficiency of Tn5 integration (**Fig. 2.6E**). While the latter might seem surprising given that highly protein occupied regions on eukaryotic chromatin tend to exclude Tn5 (as is used to great effect in ATAC-seq [31]), we note that bacterial H-NS occupancy has previously been shown to facilitate Tn5 insertion [32]. Additional characteristics of EPODs, such as reduced densities of possible Dam methylation sites (consistent with the expected blocking of Dam methylase by bound proteins, previously shown in *in vivo* methylase protection experiments [33]) and a characteristic pattern of DNA structural parameters including decreased minor groove width, are shown in **Table S2.2**.

The remarkable condition-invariance of the locations of EPODs outlined above, even across such dramatic changes as transition from exponential to stationary phase, suggests that EPODs predominantly represent fixed structural features of the *E. coli* chromosome, rather than highly dynamic regulatory structures. We thus examined the classes of genes (assessed using gene ontology, or GO, terms) most strongly enriched or depleted in EPODs. As illustrated in **Figure 2.7A**, EPODs show strong enrichments for mobile elements (GO:0006313) and prophage genes (specifically lytic pathways; GO:0019835), and are depleted for core metabolic pathways such as ribosome components (GO:0030529). Indeed, EPODs are associated with the silencing of many prophages (e.g., **Fig. 2.7B**) and even smaller operons of unknown function (e.g., **Fig. 2.7C**).

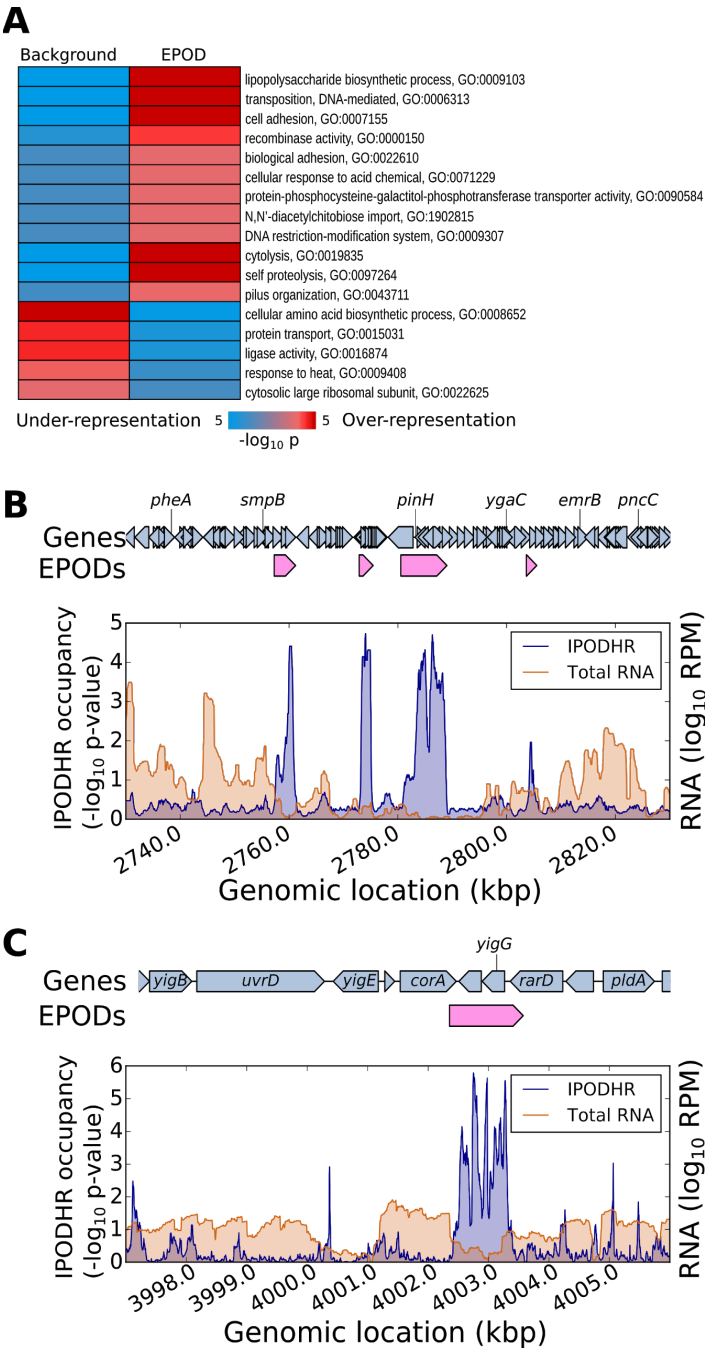

Figure 2.7: EPODs are statistically enriched for genes in specific functional categories. (**A**) The genome was split into EPOD and background regions as in **Figure 2.6**; we then applied iPAGE [45] to identify gene ontology terms showing significant mutual information with occupancy in EPODs. All shown GO terms were significant according to the built-in tests in iPAGE. **(B)** Multiple EPODs are associated with silencing of the CP4-57 prophage. Shown are the IPOD-HR occupancy and transcript levels in the vicinity of the prophage locus during growth in rich defined media with glucose, with EPOD locations indicated above the plots. **(C)** Association of a small EPOD with two genes of unknown function, *yigE* and *yigF*; data tracks defined as in panel **B**.

Our findings regarding EPODs, particularly the high levels of H-NS binding in EPODs, and the known role of H-NS as a xenogeneic silencer [34,35], are highly consistent with prior information regarding the silencing role of H-NS. In order to determine the extent to which H-NS silenced regions and EPODs as defined here overlap, we compared the distribution of EPODs across the genome with H-NS ChIP-seq data from [36]. Using an unsupervised clustering method to divide genomic intervals into high, medium, and low levels of H-NS occupancy. We found that 66.7% of EPODs fall into the high H-NS category, compared with 3.9% of non-EPOD regions (**Fig. S2.2A**). Nevertheless, when considering the average transcript levels observed as arising from the same genomic intervals, the EPODs from the low H-NS and medium H-NS categories still showed significantly lower expression than non-EPOD regions with similar H-NS levels (**Fig. S2.2B**), and the small number of highly H-NS bound regions which are not part of EPODs are in fact more silent than highly H-NS bound EPODs. Taken together, we thus observe that while many EPODs represent chromosomal regions silenced by H-NS, roughly one-third of EPODs do not show the characteristics of highly H-NS occupied regions, but are nevertheless transcriptionally silenced by an extended stretch of high protein occupancy. The possibility of course exists that H-NS is repositioned in the conditions of our study, which differ from those of [36], to cover the remainder of the EPODs identified here; however, we have shown in practice that this is not the case, as the non H-NS EPODs identified here persist even in a *hns*/*stpA* strain (Amemiya and Freddolino, manuscript in preparation). The mechanism of silencing at these non-H-NS dependent EPODs will likely be a fruitful area for future investigation.

Discussion

The study of bacterial transcriptional regulatory networks has long benefitted from bottom-up approaches such as DNase footprinting, ChIP-chip, and ChIP-seq to map the behavior of individual factors and regulons. At the same time, however, the insight provided by such approaches has been inherently limited by the need to specify *a priori* the target of investigation, either in terms of the regulator, regulated gene, or both.

However, as we hope to have demonstrated here, a global agnostic strategy (as exemplified by IPOD-HR) provides a unique top-down complement to existing methods by permitting rapid profiling of the protein occupancy landscape of a bacterial chromosome. We have demonstrated that IPOD-HR simultaneously enables resolution of individual changes in transcription factor binding at specific sites, inference of new regulatory motifs which likely correspond to functional but poorly characterized transcriptional regulators, and large-scale patterns of protein occupancy indicative of constitutively silenced genomic regions. IPOD-HR thus falls into the same family as methods such as DNase I hypersensitivity [37], MNase-seq [38], and ATAC-seq [31], but developed, tuned, and validated for the unique molecular and biophysical features of bacterial chromosomes.

We expect that all three key capabilities of IPOD-HR highlighted above will prove to be of substantial utility in investigating all cultivable bacterial transcriptional regulatory networks, and could potentially even be applied to environmental samples to study occupancy landscapes in uncultivable bacteria. The ability to directly track the occupancy of TFBSs for a large set of transcriptional regulators in parallel provides the missing link that has previously stymied efforts to predict the transcriptional output of *E. coli* across conditions, as consideration of only the expression levels of TFs to predict the behavior of their regulons has yielded mixed results [8,39]. Furthermore, the ability to identify likely regulatory sites even in the absence of prior knowledge, as shown both for isolated promoters (**Fig. 2.4**) and inference of entire regulons (**Fig. 2.5**), will substantially accelerate our ability to complete a wiring diagram for the *E. coli* transcriptional regulatory network and to rapidly approach the networks of other less well-characterized bacterial species. Here IPOD-HR provides a powerful high-throughput *in vivo* approach tracking occupancy at native sites, complementing methods based on screening with purified proteins [23], computational inference [25], or reporter assays [24].

Our study of diverse experimental conditions across different genetic and physiological states provides a comprehensive view of the protein-DNA interactome of *E. coli*. As we

have shown, the majority of discovered occupancy events do not correspond to previously known or annotated sites of protein-DNA interactions. We have further shown that these global occupancy profiles can be used for wholesale discovery of sequence specificity for the set of TFs active under these conditions. These occupancy maps and the corresponding DNA motifs provide the community with a rich catalogue of likely regulatory events to study, targeting either particular genes or larger pathways. Indeed, our finding that the novel occupancy sites and DNA motifs are highly enriched upstream of genes that are under-studied promises to discover and expand the physiological and regulatory modules of *E. coli*, beyond those that have been targeted by decades of previous research.

Our study also provides significant additional evidence for the presence of large, transcriptionally silent, high occupancy chromosomal domains in *E. coli.* Many such EPODs clearly correspond to regions of H-NS binding, which has previously been shown to form several types of filaments that silence horizontally acquired DNA [5,35,40–42]. On the other hand, we also observe a substantial fraction of EPODs that do not correspond to H-NS binding, and yet are still associated with transcriptionally silent regions of the chromosome. Numerous questions regarding the nature and role of those EPODs remain for future work, including: what is the protein composition of non H-NS EPODs? What rules dictate their formation on specific sites? We are also tempted to speculate that in some contexts non-H-NS EPODs may undergo condition-dependent changes in occupancy that drive transcriptional regulation, although no such cases could be definitively identified in the conditions studied here. Such behavior has already been observed for H-NS filaments in various enterobacteria [43,44]. Ongoing application of IPOD-HR to a broader range of physiological conditions in *E. coli* should provide further insight into the overall landscape of large-scale protein occupancy across conditions, allowing tracking both of occupancy associated with H-NS (and the related protein StpA) and other classes of EPODs in a single experiment.

Our IPOD-HR strategy for mapping the global dynamics of the *E. coli* protein-DNA interactome relies only on simple physico-chemical principles for isolating protein-DNA

complexes. As such, it is easily transferable to other bacterial species. The rich and comprehensive data sets generated by such studies, and application of statistical inference during data processing as exemplified here, will provide particularly important regulatory roadmaps in organisms with less well studied transcriptional regulatory networks. In the future, more applications to a broader range of physiological conditions (in *E. coli*) and to other bacterial strains and species will provide important information on the role of large-scale nucleoprotein assemblies on gene regulation, and pave the way for more comprehensive and predictive models of transcriptional regulatory logic, particularly for non-model bacterial species of clinical and industrial importance.

Materials and Methods

Strain construction

The base strain for all experiments used here is an MG1655 stock obtained from H. Goodarzi, which is isogenic with ATCC 700926 [46]. All specified gene knockouts were obtained by P1 transduction [47] of the FRT-flanked *kanR* marker from the corresponding knockout strain of the Keio collection [48], followed by Flp recombinase mediated excision of the marker using the pCP20 plasmid [49] to leave a small scar in place of the original open reading frame. Candidate isolates for each deletion were grown overnight at 42^{o} C to drop the pCP20 plasmid, and **then** replica plated onto appropriate selective plates to ensure loss of both the plasmid and kanamycin resistance marker. Knockouts were confirmed by PCR fragment sizing and/or sequencing across the marker scar. Note that the Δ*lexA* strain that we refer to is in fact Δ*lexA*/Δ*sulA*, as loss of *lexA* is lethal in the presence of a functional *sulA* gene [50,51].

Media/culture conditions

For routine cloning applications and for recovery of cryogenically preserved cells, we used LB (Lennox) media (10 g/L tryptone, 5 g/L yeast extract, 5 g/L NaCl), with bacteriological agar (15g/L) added as appropriate.

For physiological experiments, we made use of a variety of supplemented versions of M9 defined medium (6 g/L Na_2HPO_4, 3 g/L KH_2PO_4, 1 g/L NH_4Cl, 0.5 g/L NaCl, 1 mM $MgSO_4$) [47]. Our M9 minimal media condition (M9/min) additionally includes 0.2% (w/v) glucose, 0.4 mM $CaCl_2$, 40 μM ferric citrate, and the micronutrient mixture typically incorporated in MOPS minimal media [52]. Our M9 rich defined medium condition (M9/rdm) instead incorporates into the M9 base 0.4% (w/v) glucose, MOPS micronutrients (as above), 4 μM $CaCl_2$, 40 μM ferric citrate, and 1x supplements ACGU and EZ as used in MOPS rich defined medium [52].

Cell growth and harvest for IPOD-HR

The cells of interest were grown overnight in the media of interest after inoculation from an LB plate. In the morning, the culture was back-diluted into fresh, prewarmed media to an OD600 of 0.003. The culture was then grown to the target OD600 (0.2, except in the case of stationary phase samples, which are described below), at which point a 200 μL aliquot was removed and preserved in 1 mL of DNA/RNA Shield (Zymo Research) following the manufacturer's instructions.

The remainder of the culture was treated with rifampin to a final concentration of 150 μg/mL, and incubated for 10 minutes under the same culture conditions as the main growth to immobilize initiating RNA polymerase at active promoters and permit completion of transcripts in progress. The culture was then rapidly mixed with concentrated formaldehyde/sodium phosphate (pH 7.4) buffer sufficient to yield a final concentration of 10 mM NaPO4 and 1% v/v formaldehyde. Crosslinking was allowed to

proceed for 5 minutes at room temperature with vigorous shaking, followed by quenching with an excess of glycine (final concentration 0.333 M) for 5 minutes with shaking at room temperature. The crosslinked cells were subsequently chilled on ice, and washed twice with ice cold phosphate buffered saline, 10 mL per wash. The fully washed pellets were carefully dried, any remaining media pipetted away, and then the pellets were snap-frozen in a dry ice-ethanol bath and stored at -80 C.

In the case of our stationary phase samples, cells were grown as described above in terms of back-dilution and growth to an OD600 of 0.2, and then grown for an additional three hours prior to RNA harvest, rifampin treatment, and crosslinked as described above.

Cell lysis and DNA preparation

Frozen cell pellets were resuspended in 1x lPOD lysis buffer (10 mM Tris HCl, pH 8.0; 50 mM NaCl) containing 1x protease inhibitors (Roche Complete Mini, EDTA free) and 52.5 kU/mL of ready-lyse (Epicentre); 600 µL per pellet (stationary phase cells were diluted 10x prior to lysis, and only 1/10 of the resulting material used, due to the much higher biomass of those pellets). We incubated the resuspended pellet for 15 minutes at 30 C, and then placed it on ice. We then sonicated the cells using a Branson digital sonicator at 25% power, using three 10 second bursts with 10 second pauses between bursts. The cells were maintained in a wet ice bath throughout sonication.

We then performed a calibrated DNA digestion to sub-200 bp fragments, by adding to the sonicated lysates 60 µg RNase A (Thermo Fisher), 6 µL DNase I (Fisher product #89835), 5.4 µL 100 mM MnCl2, and 4.5 µL 100 mM CaCl2, and then incubating on ice. While the appropriate digestion time must be calibrated for each particular sample type and batch of DNase, 30 minutes of digestion proved appropriate for all samples here. Reactions were quenched after completion by the addition of 50 µL 500 mM EDTA (pH 8.0), typically yielding 50-200 bp fragments.

IPOD-HR Interface Extraction

Prior to interface extraction, samples were clarified by centrifugation for 10 minutes at 16,9000xg at 4° C. After clarification, a 50 microliter input sample was diluted 1:9 in elution buffer (50 mM Tris, pH 8.0; 10 mM EDTA; 1% SDS) and kept on ice until the reverse crosslinking step. The remainder of the lysate was mixed with 1 volume of 100 mM Tris base and 2 volumes of 25:24:1 phenol:chloroform:isoamyl alcohol, vortexed, and then incubated for 10 minutes at room temperature. After incubation, the sample was spun at 21,130xg for two minutes at room temperature, allowing formation of a white disc at the aqueous-organic interface enriched for protein-DNA complexes [10,11].

The complete aqueous phases were removed and discarded, and the remaining disc washed again with 350 microliters TE (10 mM Tris, pH 8.0; 1 mM EDTA), 350 microliters 100 mM Tris base, and 700 microliters 24:1 chloroform:isoamyl alcohol. The resulting mixture was vortexed vigorously, and again centrifuged for 2 minutes at 21,130xg. All liquid was again removed, and the wash was repeated using 700 microliters TE and 700 microliters 24:1 chloroform:isoamyl alcohol. After vortexing, centrifugation, and removal of the final wash (exactly as above), any residual liquid was removed by wicking with a laboratory wipe (if any substantial pools of liquid were present). Finally, the interface was resuspended in 500 microliters of elution buffer (described above), vortexed vigorously, and kept on ice until reverse crosslinking (no more than a few hours).

We caution the reader that the separation of the interface layer from the liquid on either side of it is crucial to success with this method. We have found it most effective to tilt the microcentrifuge tube toward while pipetting out the organic layer from beneath, at which point the interface will adhere to the tube wall and allow easy removal of the aqueous layer. We have also found that the handling characteristics of the interface vary greatly with the plasticware in use. For the work described here, we have used 2 mL microcentrifuge tubes from USA Scientific for all interface handling, as the interfaces

adhere nicely to the tube wall (other plasticware may yield variable results); at the same time, the use of low-retention pipette tips appears to reduce binding of the interface to the tip.

RNA polymerase chromatin immunoprecipitation

DNA for RNA polymerase ChIP-seq experiments was prepared as described above for IPOD-HR interface extraction up through the lysate clarification stage. Whenever possible, we used frozen pellets obtained from the same culture for matched IPOD-HR and ChIP-seq experiments, in which case the lysates were pooled and mixed immediately prior to removal of a single input sample. ChIP procedures here were modeled on those of [53].

The digested lysates were mixed 1:1 with 2x IP buffer (200 mM Tris, pH 8.0; 600 mM NaCl; 4% Triton X-100; 2x Roche Complete EDTA-free protease inhibitors), and then kept on ice for no more than a few hours prior to antibody addition. We added 10 microliters of purified anti-*E. Coli* RNA polymerase antibody (Neoclone WP023), and incubated overnight with rocking at 4 C. Near the end of the incubation period, we resuspended an aliquot of 50 microliters of protein G dynabeads (Invitrogen) and equilibrated the protein G beads with 1x IP buffer lacking protease inhibitors. The bead aliquot was added to the antibody-lysate mixture, and then incubated 2 hours with rocking at 4^{o} C. The bead-antibody-target complexes were subsequently subjected to the following series of washes, with 1 mL used per wash. All washes were at room temperature, and involved manual resuspension of the beads in the new wash buffer followed by immediate re-separation.

- 1x Wash buffer A (100 mM Tris, pH 8.0; 250 mM LiCl; 2% Triton X-100; 1 mM EDTA)
- 1x Wash buffer B (100 mM Tris, pH 8.0; 500 mM NaCl; 1% Triton X-100; 0.1% sodium deoxycholate; 1 mM EDTA)
- 1x Wash buffer C (10 mM Tris, pH 8.0; 500 mM NaCl; 1% Triton X-100; 1 mM EDTA)

- 1x TE (10 mM Tris, pH 8.0; 1 mM EDTA)

The antigens were subsequently eluted by adding 500 microliters of elution buffer (composition described above) and incubating 30 minutes at 65^{o} C, with vigorous vortexing every 5-10 minutes.

Crosslinking reversal and recovery of DNA

The DNA from the input, IPOD-HR, and ChIP fractions described above was recovered using identical procedures: samples diluted in elution buffer (see above) were incubated overnight (6--16 hours) at 65^{o} C to reverse formaldehyde crosslinks. After allowing the samples to cool to room temperature, we then added 100 μg of RNase A (Thermo-Fisher), incubated 2 hours at 37^{o} C, then added 200 μg of proteinase K (Fermentas) and incubated an additional 2 hours at 50^{o} C. DNA was then recovered via standard phenol-chloroform extraction and ethanol precipitation, following protocols from [47]. We used Glycoblue (Ambion) as a co-precipitant, NaCl as a precipitating salt (due to the presence of SDS in our solution), and washed with ice-cold 95% ethanol to avoid loss of low molecular weight DNA.

Recovered DNA was quantified via fluorescent quantitation (using either the Invitrogen PicoGreen or Promega QuantIT system), and samples of sufficiently high concentration were also run on a 2% agarose gel for fragment size assessment. Typical total yields from the procedure above were on the order of 1 µg of DNA for the input samples, 100-200 ng for the IPOD-HR samples, and 1-10 ng for the ChIP samples.

Preparation of next-generation sequencing (NGS) libraries

Except as otherwise noted, all DNA samples were prepared for Illumina sequencing using the NEBNext Ultra DNA Library Prep Kit (NEB product #E7370), with either single index or dual index primers also obtained from NEB. We followed the manufacturer's instructions except for the following variations:

- Cleanups prior to adapter ligation were performed using a Zymo Clean&Concentrator 5 spin column kit or Zymo Oligo Clean&Concentrator spin column kit instead of Ampure beads, in order to avoid the loss of low molecular weight DNA
- We used Ampure and Axygen PCR cleanup beads interchangeably, having established in side-to-side comparisons that they were functionally identical for the steps in the NEB sequencing prep (data not shown). The final cleanup step was in some cases repeated to remove obvious populations of adapter dimers.

All libraries were sequenced on either an Illumina HiSeq or NextSeq instrument. A small number of samples were prepared for sequencing using an Illumina Truseq Nano kit instead of the NEBnext kit noted above; we found that upon calculation of correlations between the coverages of a broad range of IPOD, input, and RNA polymerase ChIP-seq samples prepared using various sequencing preparation kits that the Truseq Nano samples were indistinguishable from NEBNext Ultra samples, whereas other sequencing preparation methods (notably including standard Illumina Truseq samples) lead to detectable non-biological differences in observed coverage.

RNA isolation and RNA-seq sample preparation

As noted above, samples for RNA isolation were preserved immediately prior to rifampin addition by dilution in a 5x excess of DNA/RNA Shield (Zymo); the RNA samples were then stored at -80^{o} C until purification. RNA was isolated using a Zymo QuickRNA microprep kit following the manufacturer's instructions, including the on-column DNase digestion. Purified RNA was quantified using RiboGreen (Invitrogen), and then ribosome-depleted using the Illumina RiboZero Gram-negative bacteria kit according to the manufacturer's instructions, with the input RNA amount and all reaction volumes cut in half. Final recovery of the ribo-depleted RNA was accomplished using the modified Zymo spin column protocol present in the RiboZero documentation. Ribo-depleted RNA was then prepared for sequencing using the NEBNext Ultra Directional

RNA kit (NEB product E7420),and sequenced as described above for the DNA samples.

Analysis of NGS data

All NGS data was preprocessed using a common pipeline, after which DNA and RNA data sets were processed separately. The reference genome in all cases was the most recent version of the *E. coli* MG1655 genome (GenBank U00096.3), with gene, transcription factor binding site, and transcription start site annotations from RegulonDB [7]. Data processing was automated using in-house python and bash scripts, and parallelized where possible using GNU parallel [54] or the python multiprocessing library.

Read Quality Control and Preprocessing

All reads were subjected to adapter removal using cutadapt 1.8.1 [55] to cut the common sequence of Illumina Truseq adapters, and then trimmed to remove low-quality read ends with Trimmomatic 0.33 [56], using the trimming steps '`TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:10`'. Samples were subjected to additional manual quality checks using FastQC [57] and MultiQC [58] to identify any irregularities in terms of sequence content, quality, or duplication.

DNA Sequencing and Protein Occupancy Calling

Surviving DNA reads were aligned to the U00096.3 genome using bowtie2 version 2.1.0 with "very sensitive" end-to-end alignment presets, and dovetail alignments allowed. Only concordant paired-end reads were retained for subsequent quantitation. Separate occupancy tracks were calculated both for the aligned reads of each sample, using the parsing method of Kroner *et al.* [59], and scaling the basepair-wise contribution of each read by the inverse of its length (thus, each read contributed the same total amount of occupancy signal to the traces). All sample-wise occupancy data were normalized by

quantile-normalizing the original datasets (acting separately for the input, IPOD interface, and RNA polymerase ChIP-seq tracks). In order to correct for copy number variations, for each biological condition, we fitted a periodic smoothing spline with four evenly spaced knots to the input samples for that condition; all occupancy values were divided by the spline-smoothed abundances prior to further processing. After abundance normalization, all data tracks were rescaled to have matching means, and then all replicates for each sample type/biological condition combination were averaged to generate a composite occupancy track (yielding, for example, one input data track for the WT M9/RDM/glu condition, one IPOD data track for the WT M9/RDM/glu condition, etc.).

The displayed IPOD and ChIP data tracks were then obtained as $\log_2$ ratios of the extracted (interphase or ChIP) to input samples for each condition; we refer to these tracks as the "IPOD" and "ChIP" signals below. Upon viewing the correlation between total protein occupancy and RNA polymerase occupancy, two protein-occupied subpopulations were apparent (**Figure S2.3**): a linear subpopulation where total protein and RNA polymerase are well-correlated, and a second subpopulation of positions where the total protein occupancy is much higher than expected based on the RNA polymerase occupancy. We interpret the former set of positions as protein occupancy due directly to RNA polymerase binding, and the latter as non-RNA polymerase occupancy (as schematized in **Fig. 2.1B**). To obtain the fully processed IPOD-HR signal for non-RNA polymerase occupancy, we applied a LOESS [60] fit to the scatter plot of IPOD vs ChIP signal for each sample type (fitting to a 500-fold downsampled data set for the sake of efficiency), and subtract from the IPOD signal at each position 1.1 times the value predicted from the LOESS model based on the corresponding ChIP occupancy signal (an arbitrary scaling factor that makes the calling of non-RNA polymerase occupancy more conservative); no subtraction was performed at positions where the observed RNA polymerase occupancy was negative. We refer to the resulting ChIP-subtracted IPOD signal as the IPOD-HR signal; for analysis and display we further standardize the signal by calculating robust z scores, where the robust z-score z_i at position i is defined as

$$z_i = (x_i - \text{median}(\mathbf{X})) / \text{mad}(\mathbf{X})$$

for a IPOD-HR data vector **X**, and mad() indicates the median absolute deviation. In many cases a more useful signal for visualization is a p-value for enrichment at each site; $\log_{10}$ p-values are calculated under the null hypothesis that the distribution of the robust z-scores is standard normal. To provide uncertainty estimates grounded in observed levels of biological variability across replicates, for each data point we also constructed an interval between the lowest and highest values that could have been obtained for our occupancy statistics using any combination of biological replicates (potentially different replicates for the IPOD, ChIP, and input samples to construct the largest possible range). To calculate the error bars shown in **Figure 2.3A**, we then used parametric bootstrapping to generate confidence intervals for the parameters of interest, assuming that the occupancy of each TFBS followed a normal distribution with a mean of the observed mean and standard deviation of one-quarter the range between the highest and lowest replicate-wise values (thus treating the range of the pessimistic replicate-wise possible values as an interval expected to contain ~95% of observed points); 95% confidence intervals for the average site-wise occupancies were then calculated from 1,000 bootstrap replicates.

Feature calling

To identify peaks in the robust z-scaled IPOD-HR data, we applied continuous wavelet-transform (CWT) peak calling [61] (as implemented in the scipy.signal package), with a range of widths from 25 bp to 125 bp (at 5 bp increments) used to generate the CWT matrix, and refer to peaks based on the minimum signal-to-noise ratio threshold at which they appear as peak calls. Each peak call was padded by 30 bp on each side to define the peak region used in subsequent analysis.

Extended protein occupancy domains (EPODs) were called using an approach similar to that in (Vora et al., 2009): we identified EPOD seed regions as any region at least 1,024 bp in length, over which a 512 bp rolling median exceeded the overall kth percentile of a 256 bp rolling median across the entire chromosome (in all cases acting

on the robust z-scored IPODHR data); we used k=90 for the main EPOD calls made in the text, and k=75 for the relaxed version used in threshold analysis. In the case of overlapping seeds, only the one with the highest median was retained. Seed regions were then expanded in both directions as far as possible while maintaining the median over the entire EPOD call above the threshold noted above, and without crossing any location with a robust z-score ≤ 0.

Transcription factor co-clustering analysis

For the transcription factor occupancy and co-clustering data, we performed consensus clustering (inspired by [62]). For each biological condition, we assigned each transcription factor a score given by the geometric mean of the site-level occupancies (IPOD-HR $-\log_{10}$ p values) for annotated binding sites of that TF in that condition (using a minimum value of 0.01 for each site-level value); the condition-wise average occupancies for each TF were then divided by the highest average occupancy for that TF across all conditions, yielding an occupancy score on the interval (0,1] for each TF-condition combination. The occupancy profiles of TFs across conditions were clustered 100 times using K-means clustering at each number of clusters between 8 and 12 (inclusive); the 'co-clustering frequency' κ is defined as the fraction of those 100 trials in which a given pair of TFs were assigned to the same cluster. We then used the quantity (1-κ) as a distance measure in a final hierarchical clustering, assigning the TFs to 10 clusters, to provide the cluster identities shown in **Figure 2.3B**.

RNA Sequencing and Differential Expression Calling

RNA-seq data sets were subjected to the same initial preprocessing and quality control steps as outlined above for the DNA samples, and then gene-level expression was quantified using kallisto v0.43 [63] on a version of the MG1655 (GenBank NC_000913) genome with all ribosomal RNAs removed. Gene-level transcript per million (TPM) values from kallisto were used for all downstream analysis unless otherwise noted. To generate high-resolution occupancy plots, reads were instead aligned with bowtie2 as

described above for DNA reads, and read occupancies quantified using the genomecov command of bedtools2 [64].

TFBS Comparison

Binding sites identified from IPOD-HR peaks (as described above in the “Feature calling” paragraph) were cross-referenced with known and predicted TFBSs using bedtools v2.17.0 [64]. “Known” sites comprise all binding sites contained in the RegulonDB release 9.4 BindingSetSet.txt file [7]; “predicted” sites are identified by scanning the MG1655 (GenBank NC_000913) genome with FIMO [65] using all *E. coli* position weight matrices from SwissRegulon [66] (as distributed by the MEME project); each PWM was applied separately, and all sites with a q-value less than 0.2 were retained. Default settings were used for FIMO, except that the background was a second-order Markov model based on the NC_000913 genome, and the number of maximum stored scores was set to 10,000,000.

Motif Identification

Novel sequence motifs implied by IPOD-HR data were identified using an inference pipeline built off of FIRE [67]. Occupancy peaks and associated discrete threshold scores in the IPOD-HR traces were called using the CWT-based approach described above, using a score threshold of 4; each peak was assigned a discrete score corresponding to the average IPOD-HR occupancy score within that peak, rounding down. We then generated a background distribution of unbound sequences drawn from the portion of the genome not included in peaks, matching the length distribution of the peaks but with three times as many locations; all such background regions were assigned a score of 0 to distinguish them from the various thresholded peak regions.

Motifs were called using FIRE with two separate variations: FIRE_gapped (with parameters `--kungapped=6 --gap=0-10 --jn_t_gapped=4 --minr=0.5`), which searches for gapped motifs typical of prokaryotic transcription factor binding sites;

and FIRE_maxdeg (with parameters `--jn_t=8 --minr=1.5 --maxdeg=1.8`), which searches for motifs while preserving information content above a specified threshold. We applied additional empirical filters to specifically enrich for peaks corresponding to binding sites: all peaks identified via FIRE were required to have the motif significantly depleted from the background population ($p<0.01$). To assess the false discovery rate of our methods, we also generated 20 decoy peak sets by shuffling the locations of the real peaks observed in each condition, along with corresponding randomized unbound sets for each, and then applied identical peak calling procedures to each decoy set.

To avoid repeated reporting of very similar motifs which might be identified by our pipelines, we applied the matrix-clustering module of RSAT [68] (using recommended thresholds `-lth cor 0.7 -lth w 5 -lth Ncor 0.4`) to obtain non-redundant motif sets for downstream analysis. We compared all called motifs with previously known motifs from the SwissRegulon database using TOMTOM [29] with default parameters, requiring an E-value of 0.5 or lower for 'Identified' hits, and labeling other identified motif matches as 'Ambiguous' (counted together with TFs lacking any hits). For the identification of predicted regulons associated with each motif, we applied the FIMO program [65] to identify potential binding sites on the *E. coli* K12 genome, with a q-value threshold of 0.2. For the purposes of our analysis of the potential regulatory networks of novel motifs), we marked each transcriptional unit in *E. coli* as being regulated by a particular motif if and only if a predicted binding site for that motif overlapped the gene's core promoter.

In vitro pulldown of unidentified transcription factors

In order to identify the protein(s) binding to the *sdaC* promoter (as in **Fig. 2.4**), we first prepared biotinylated bait DNA by cloning a fragment of the *sdaC* promoter (running from positions 2927790 to 2927975 in the U00096.3 genome) into a pAZ3-based cloning vector [69], and then amplifying that region of the plasmid using a primer pair where one primer contained a 5' biotinylation. The resulting 486 bp fragment was

treated with Exonuclease I (Affymetrix) according to the manufacturer's instructions to remove unreacted primer, and then purified using a Zymo Clean & Concentrate 25 kit.

The biotinylated bait DNA was then bound to equilibrated Dynabeads MyOne Streptavidin C1 beads (Invitrogen). Beads were equilibrated by washing three times with 1x B&W buffer (5 mM Tris Cl, pH 7.5; 0.5 mM EDTA; 1 M NaCl) and then resuspended in five volumes of 2x B&W buffer, using 42 µL of the original resuspended bead solution per reaction. The equilibrated beads were combined with 8 µg of biotinylated bait DNA plus an appropriate volume of water to yield a final 1x B&W solution, and incubated 15 minutes at room temperature with gentle rocking to allow for bait binding. The beads were then washed three times with 500 µL of 1x B&W buffer, twice in 500 µL of 1x BMg/THS buffer (5 mM HEPES, pH 7.5; 5 mM MgCl2; 50 mM KCl; 31 mM NaCl; 1x cOmplete EDTA-free protease inhibitors (Roche)), once with 500 µL of 1x BMg/THS/EP buffer (1x BMg/THS buffer supplemented with 20 mM EGTA (pH 8.0) and 10 µg/mL poly d(IC) (Sigma)). The beads were then resuspended in 200 µL of BMg/THS/EP buffer and gently mixed by hand for one minute to complete equilibration.

Cell extracts were prepared by growing to an OD_{600} of 0.2 in M9/RDM/glucose media (following the same procedures as those given for IPOD-HR experiments). Once reaching the target OD, the cells were chilled 10 minutes on ice, and then pelleted by spinning for 10 minutes at 5,500xg while at 4^{o} C. Supernatant was removed, and the cells were flash-frozen in a dry ice/ethanol bath. Cells were then lysed by resuspending the frozen pellet resulting from 82 mL of culture in 160 µL of B-PER II bacterial protein extraction reagent (Thermo Scientific). We then added 3.8 mL of 1x BMg/THS buffer and 0.8 µL of ReadyLyse lysozyme solution (Lucigen), 40 µL of 10 mg/mL RNase A, 20 µL of $CaCl_2$, and 200 µL of micrococcal nuclease (NEB; 2,000,000 gel units/mL). The lysis/digestion reaction was allowed to proceed for 30 minutes at room temperature, and then clarified by centrifugation at 30 minutes at 16,100x g held at 4^{o} C. The reaction was halted by the addition of 444 µL of 5 M NaCl, and then the entire volume applied to a 3 kDa MWCO spin filter (Amicon Ultra; Millipore) and centrifuged at (3,200 x g held at 4^{o} C) until ~400 µL of retentate remained. We then added 3.6 mL of BMg/THS (lacking

NaCl and KCl, but containing 5 mM CaCl2) and filtered to 400 µL of retentate. Retained liquid was then recovered, and diluted to a final volume of 4.0 mL with addition of salt-free BMg/THS + 5 mM $CaCl_2$. The retained lysate was then incubated 30 minutes at room temperature to permit further activity of micrococcal nuclease on remaining DNA in the sample, and then quenched with 168 µL of 500 mM EGTA. The volume of the sample was reduced to approximately 1.6 mL by ultrafiltration as above, and further supplemented with 10 µg/mL of poly d(IC) and 1 mM dithiothreitol.

Probing of the lysates was then accomplished by combining the equilibrated bait-bead complexes (described above) with the lysates, and incubating 30 minutes with rocking at room temperature. The supernatant was then removed, and the beads washed twice with 200 µL of BMg/THS/20 mM EGTA/10 µg/mL poly d(IC), and once with 200 µL of BMg/THS/20 mM EGTA. Proteins were then eluted from the beads through progressive washes of elution buffer (25 mM Tris HCl, pH 7.5) with 100 mM NaCl, 200 mM NaCl, 400 mM NaCl, and 1 M NaCl, with 50 µL used for each elution.

We successively probed the lysates described here with probes containing promoter sequences from *lexA*, *purR*, and finally, *sdaC* (each containing identical plasmid-derived flanking sequences). A ~25 kDa band of interest appeared in the 400 mM and 1 M NaCl *sdaC* eluates but not eluates from a parallel experiment performed under identical conditions with a segment of the *thiC* promoter; these bands were excised from a silver-stained gel. The 400 mM gel slice was then subjected to proteomic analysis at the University of Michigan Proteomics & Peptide Synthesis core facility. The gel slice was processed using a ProGest robot (DigiLab) to wash with 25 mM ammonium bicarbonate followed by acetonitrile, reduce with 10 mM dithiothreitol at 60^{o} C followed by alkylation with 50 mM iodoacetamide at room temperature, digested with trypsin (Promega) at 37^{o} C for 4 hours, and then quenched with formic acid. The digest was then analyzed by nano LC-MS/MS with a Waters NanoAcquity HPLC system interfaced to a ThermoFisher Q Exactive. Peptides were loaded on a trapping column and eluted over a 75 µm analytical column at 350 nL/min; both columns were packed with Jupiter Proteo resin (Phenomenex). The injection volume was 30 µL. The mass spectrometer was

operated in data-dependent mode, with the Orbitrap operating at 60,000 FWHM and 17,500 FWHM for MS and MS/MS respectively. The fifteen most abundant ions were selected for MS/MS. Data were searched using a local copy of Mascot, and Mascot DAT files were parsed into the Scaffold software for validation, filtering and to create a non-redundant list per sample. Data were filtered using 90% protein and 95% peptide probability thresholds (Prophet scores) and requiring at least two unique peptides per protein). The resulting mass spectrometry analysis is given in **Table S3.1** after manual pruning by core staff of common contaminants (e.g. human keratin).

Data Availability

The raw and processed sequencing data used in this study have been deposited in the Gene Expression Omnibus under accession GSE142291.

Acknowledgments

This work was supported by NIH grants R00-GM097033 and R35-GM128637 (to PLF) and R01-AI077562 (to ST). We are grateful to Dr. Alison Hottes for technical assistance and suggestions on the original IPOD method.

Supplementary Text

Supplementary Text S2.1: Effects of RNA polymerase on large-scale protein occupancy

In the IPOD-HR method, we address the contribution of RNA polymerase occupancy to the overall protein occupancy signal through two steps: first, immediately prior to crosslinking, cells are treated with rifampin for 10 minutes. We have calibrated the

rifampin treatment to allow sufficient time for in-progress transcripts to finish while minimizing perturbation of cellular physiology (building on data from [70], as well as direct calculation of the required time based on the elongation rate of RNA polymerase and lengths of *E. coli* transcripts). As rifampin inhibits promoter clearance but not the transcriptional initiation nor completion of already-elongating transcripts [71], the result will thus be to cause an accumulation of RNA polymerase at active promoters, while clearing the majority of polymerase occupancy from gene bodies, and thus substantially simplifying the identification of RNA polymerase occupancy. In addition, all IPOD-HR experiments described here were performed in parallel with RNA polymerase ChIP-seq experiments under the same conditions, permitting calibrated subtraction of RNA polymerase occupancy from active promoters to reveal changes in regulatory protein occupancy (as schematized in **Figure 2.1B-C**). However, the fact that transcriptional inhibition is known to affect nucleoid condensation [72–75] prompted us to directly inspect the effects of rifampin on genome-wide protein binding, with a particular emphasis on the effects on EPOD formation and transcription factor binding.

To illustrate the effects of rifampin treatment on regions of extremely high or low RNA polymerase occupancy, we performed IPOD experiments under our baseline growth condition (WT cells undergoing exponential growth in M9/RDM/glu medium) following the same procedure as used for all other samples in the present study, but omitting the rifampin treatment (*n.b.* the data sets of [10] were obtained under a slightly different condition, during exponential growth in LB medium). An overall analysis of the resulting EPOD calls (**Fig. S2.4**) demonstrates that the -RIF EPOD set arising from our methods show good correlations with both the heEPOD and tsEPOD calls from [10], whereas the +RIF EPOD set from our method (used throughout the rest of text) shows a much stronger correlation with the Vora tsEPOD set and a weaker correlation with heEPODs. It is also important to note that the fraction of locations from the Vora tsEPOD set that is contained within our relaxed +RIF EPOD calls (0.90) is in line with the equivalent overlaps between the EPOD sets observed across different conditions in our main data sets (lower triangle of **Fig. 2.6D**), further indicating that the EPODs identified via IPOD-HR closely resemble the original Vora tsEPODs.

To obtain a more detailed picture of the effects of rifampin on protein occupancy in our assays, we show an example of a highly expressed EPOD (heEPOD) from the original IPOD data sets in **Figure S2.5**. In the original IPOD data set [10], strong protein occupancy was noted throughout the cluster of ribosomal protein operons running from *rplQ* to *rpsJ*. Here, we see from the rifampin-omitted (-RIF) samples that the protein occupancy profile in this region is dominated by RNA polymerase occupancy, closely matching the bounds of the originally called heEPOD. On the other hand, in the rifampin-treated samples (+RIF), the vast majority of occupancy in this region is lost, with peaks only apparent at a few points within the region of interest (likely corresponding to highly active promoters). After subtraction of the scaled RNA polymerase occupancy to yield the IPOD-HR signal, the only prominent peak remaining in the region is in the *gspA-gspC* intergenic region, which has been demonstrated to be repressed by H-NS binding (although the precise binding location was previously unknown [76]).

The occupancy observed in the ribosomal protein operon cluster described above contrast strongly with that seen in the transcriptionally silent EPODs (tsEPODs) found in [10]. In the region of the tsEPOD shown in **Figure S2.6**, for example, we observe that in the -RIF samples, there is continuous protein occupancy but essentially no RNA polymerase occupancy throughout the large tsEPOD that was originally observed to span the *waaQGPSBOJYZU* operon, whereas both strong IPOD occupancy and strong RNA polymerase occupancy are apparent on a nearby heEPOD covering the *rpmBG* operon. Treatment with rifampin does not substantially alter the high level of overall protein occupancy throughout the *waa* region, whereas it restricts occupancy near *rpmBG* to active promoters only. As a result, in the ChIP-subtract IPOD-HR signal, strong occupancy remains throughout the *waaQGPSBOJYZU* operon, resulting in an EPOD call nearly identical to the original tsEPOD in that region from [10]. One feature of the +RIF samples that requires consideration is the fact that several RNA polymerase occupancy peaks appear in regions such as tsEPODs where no comparable occupancy is apparent in -RIF samples. We attribute these additional peaks to the fact that during

rifampin treatments, concentrations of free RNA polymerase will rise substantially due to the immobilization of polymerase at transcription start sites; thus, occupancy at normally weak promoters through the chromosome will increase. We find, however, that the ChIP-subtraction step of our IPOD-HR data processing pipeline accurately removes RNA polymerase occupancy at both normally-active promoters and those showing RNA polymerase binding only in the presence of rifampin treatment (as seen by the well-calibrated removal of RNA polymerase occupancy in the +RIF tracks of **Figures S2.6-S2.7**, for example). Overall, we find that even in the +RIF samples, RNA polymerase occupancy remains very well correlated with transcript levels (**Figure 2.3D**), can be cleanly subtracted to yield condition-appropriate changes in transcription factor occupancy (**Figure 2.1C**); furthermore, brief rifampin treatment does not appear to substantively alter large-scale protein occupancy other than that directly attributable to RNA polymerase binding (**Figures S2.5-S2.7**). Given that rifampin treatment permits cleaner subtraction of RNA polymerase occupancy, while not perturbing either the local or large-scale protein binding patterns that are the subject of our interest, we make use of it throughout the IPOD-HR data sets shown in the present work.

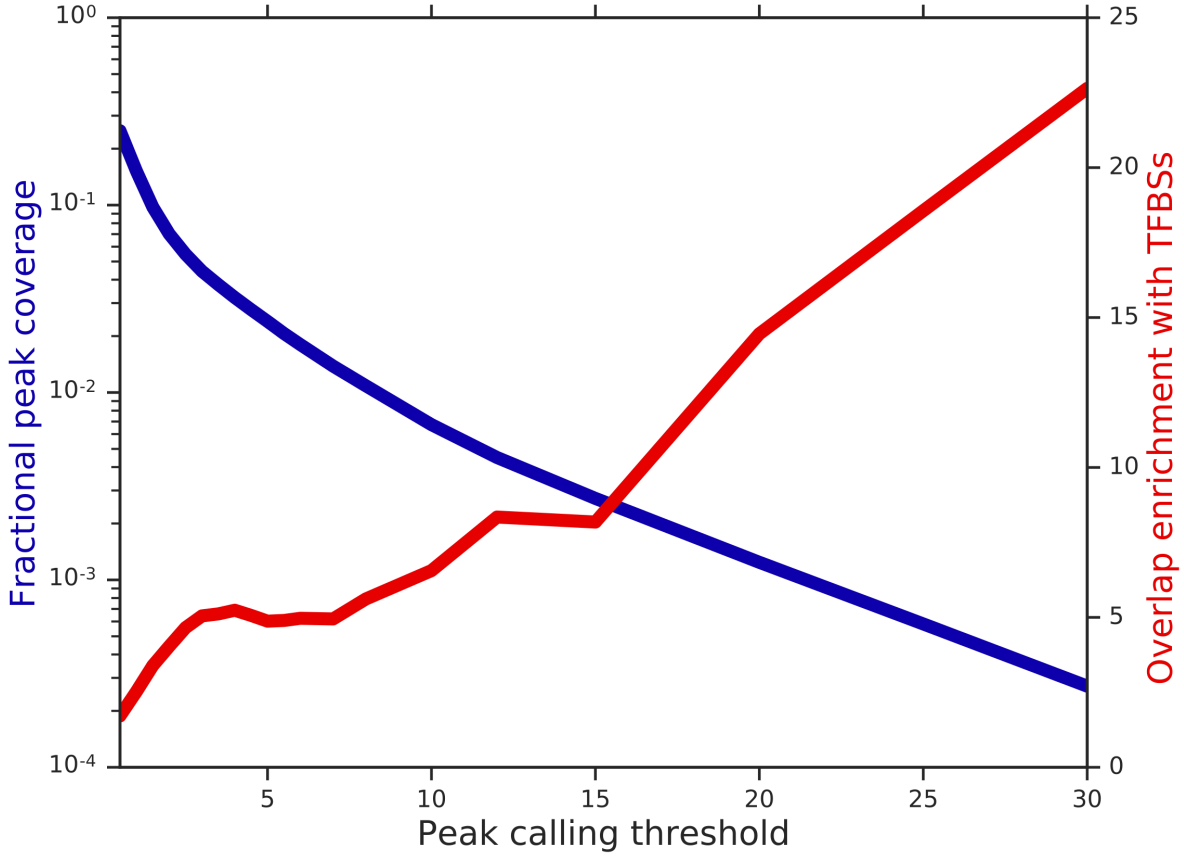

Figure S2.1: Effect of peak calling threshold on coverage and enrichment of known transcription factor binding sites (TFBSs). Data are shown for the wild type cells in the rich defined medium condition. Shown is the fraction of the entire genome contained in peak calls (left vertical axis, blue line) or the enrichment of TFBSs overlapping those peak calls relative to that expected by chance (right vertical axis, red line). Overlaps at all shown thresholds were statistically significant ($p<0.01$, permutation test in each case).

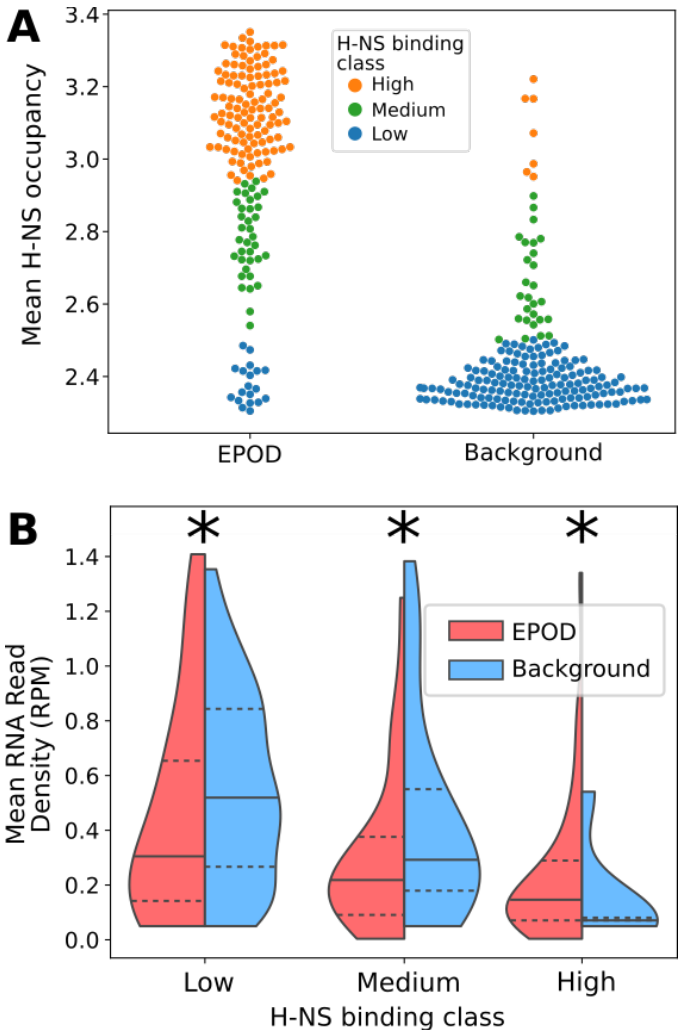

Figure S2.2: Interplay of H-NS occupancy, EPOD locations, and transcription. (A) Mean levels of H-NS binding (data from [36]) for all EPODs called in the WT rich media condition; each point shows either an EPOD or a single contiguous non-EPOD region. Each point is colored by its classification into high, medium, or low H-NS binding using a Gaussian mixture model with three groups, after removal of outliers using the local outlier factor [77] as implemented in the python scikit-learn module [78], using 25 neighbors and default settings for other parameters. **(B)** Distributions of mean RNA read density stratified by the H-NS binding categories shown in panel **A**, with each case divided by EPOD status. The median of each group is shown by a dashed line, and the 25th and 75th quartiles by dotted lines. '*' indicates a significant difference between the medians of the EPOD vs. background groups ($p<0.05$, Mann-Whitney U test).

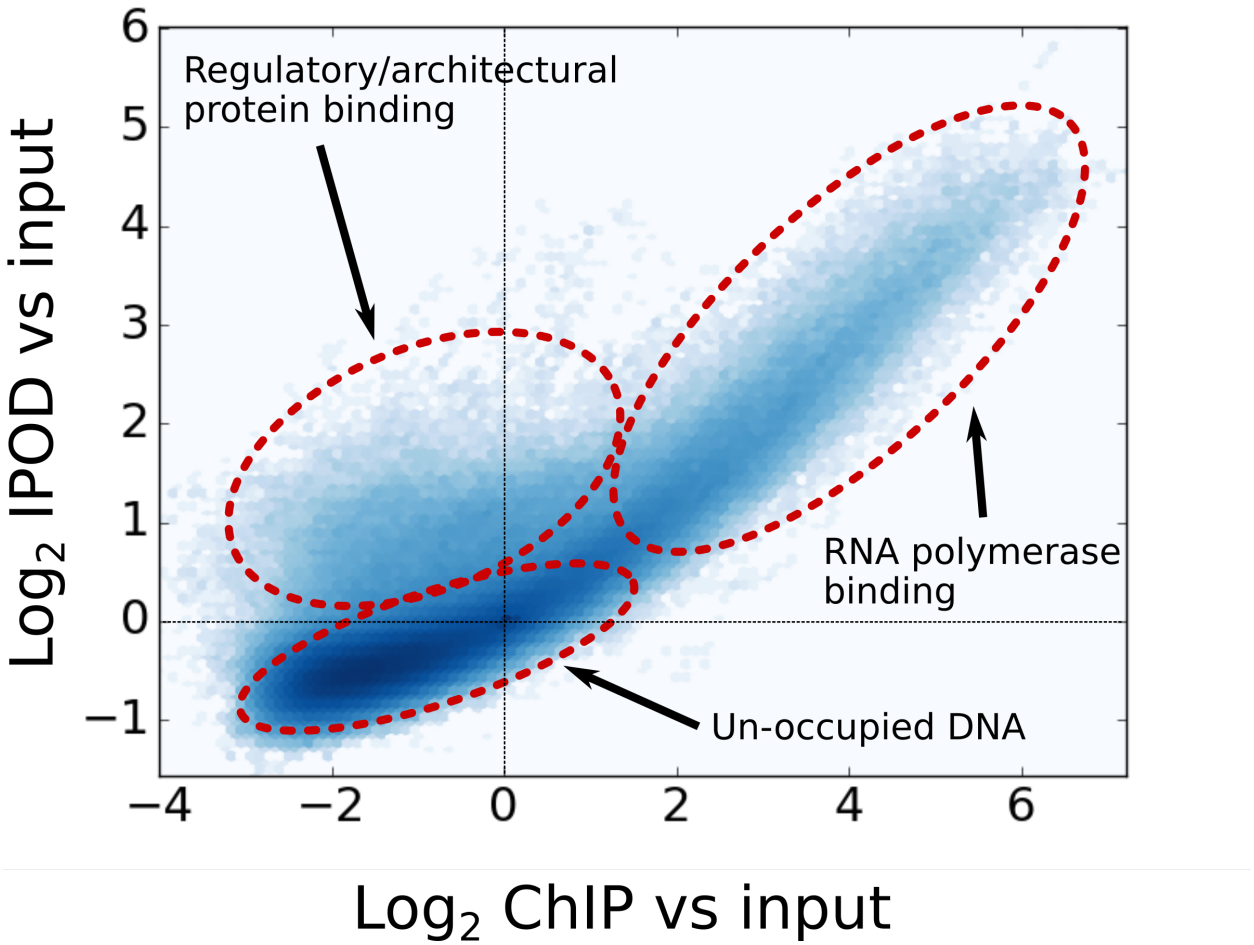

Figure S2.3: Identification of RNA polymerase vs. non-RNA polymerase protein occupancy. Shown is a density plot of the $\log_2$(IPOD/Input) signal vs. $\log_2$(RNA polymerase ChIP/Input) signal, demonstrating the presence of three subpopulations of genomic positions: unbound positions (without enrichment using either protein occupancy profiling method), RNA polymerase occupancy (part of a highly correlated region of high IPOD occupancy and high RNA polymerase occupancy), and occupancy with other proteins (which shows high IPOD occupancy but low RNA polymerase occupancy). Note that there is no corresponding population of high RNA polymerase occupancy but low IPOD occupancy; rather, the RNA polymerase-bound regions are a subset of the regions detected by IPOD.

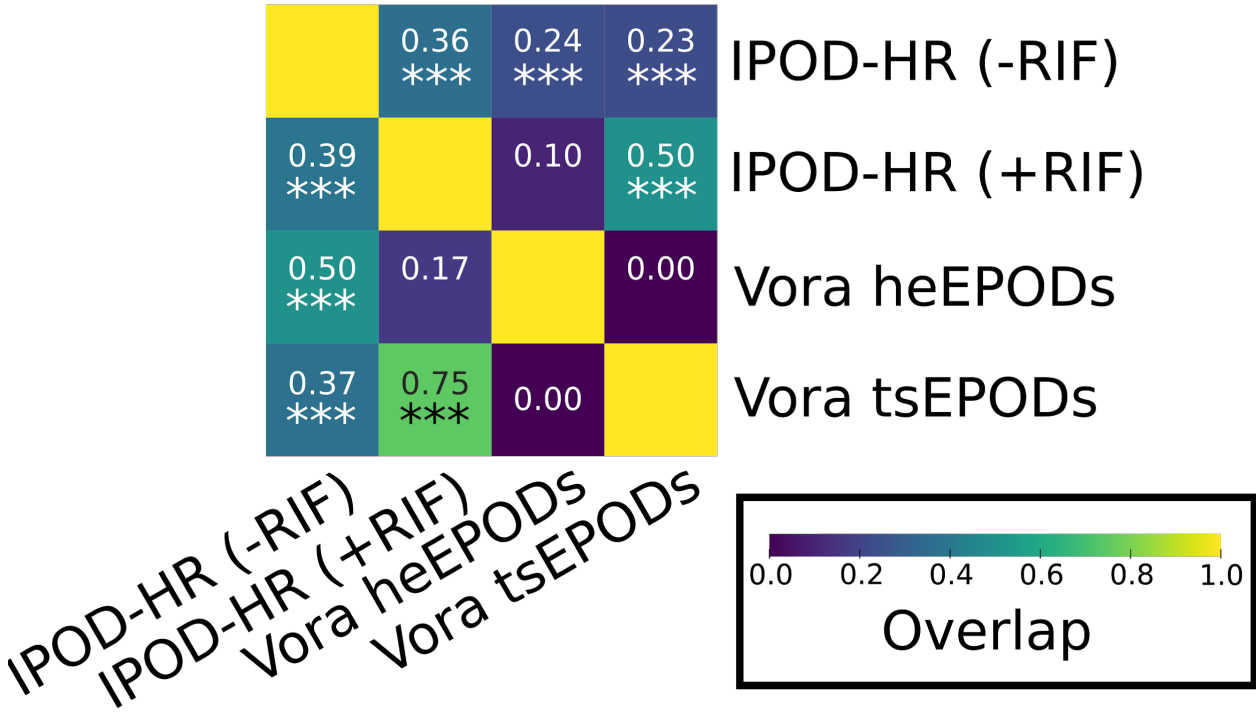

Figure S2.4: Overlaps of EPOD sets resulting from different calling methods. Shown in the heatmap are the fraction of EPODs from the EPOD set defined by the row label that overlap the EPOD set defined by the column label. Asterisks reflect p-values arising from a Monte Carlo permutation test (1000 random circular permutations; * p<0.05, ** p<0.01, *** p<0.001). p-values for the overlaps between the +RIF IPOD-HR EPOD set and the Vora heEPODs were >0.8 for both directions of comparisons.

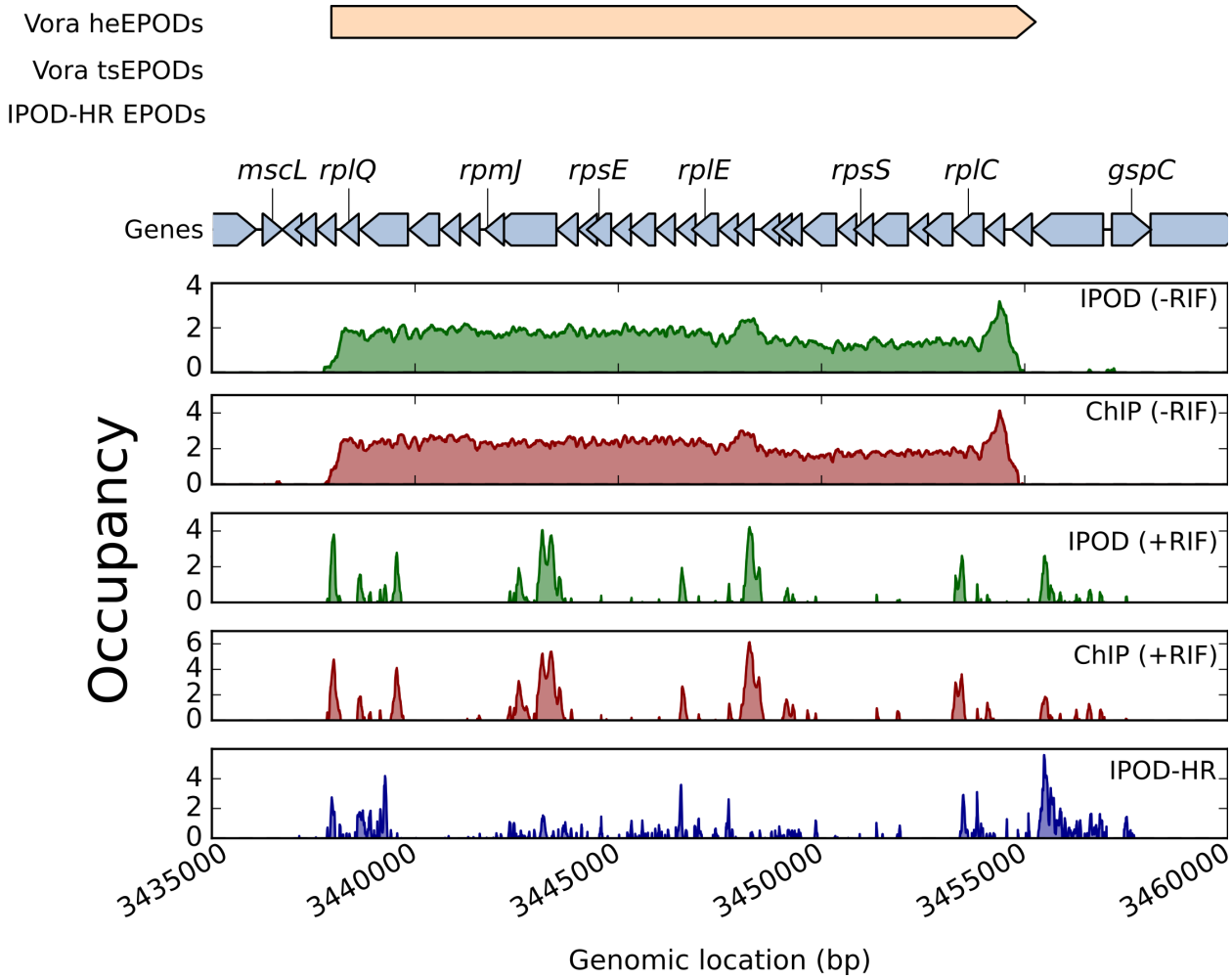

Figure S2.5: Effects of rifampin treatment on protein occupancy of a highly transcribed region. Shown are occupancy signals for interphase-extracted, RNA polymerase ChIP, and ChIP-subtracted IPOD occupancy (IPOD-HR) samples in the vicinity of a large cluster of ribosomal protein genes (running from *rplQ* to *rpsJ*). Signals are log2 extracted:input ratios (for IPOD and ChIP samples), or ChIP-subtracted robust z scores (IPOD-HR).

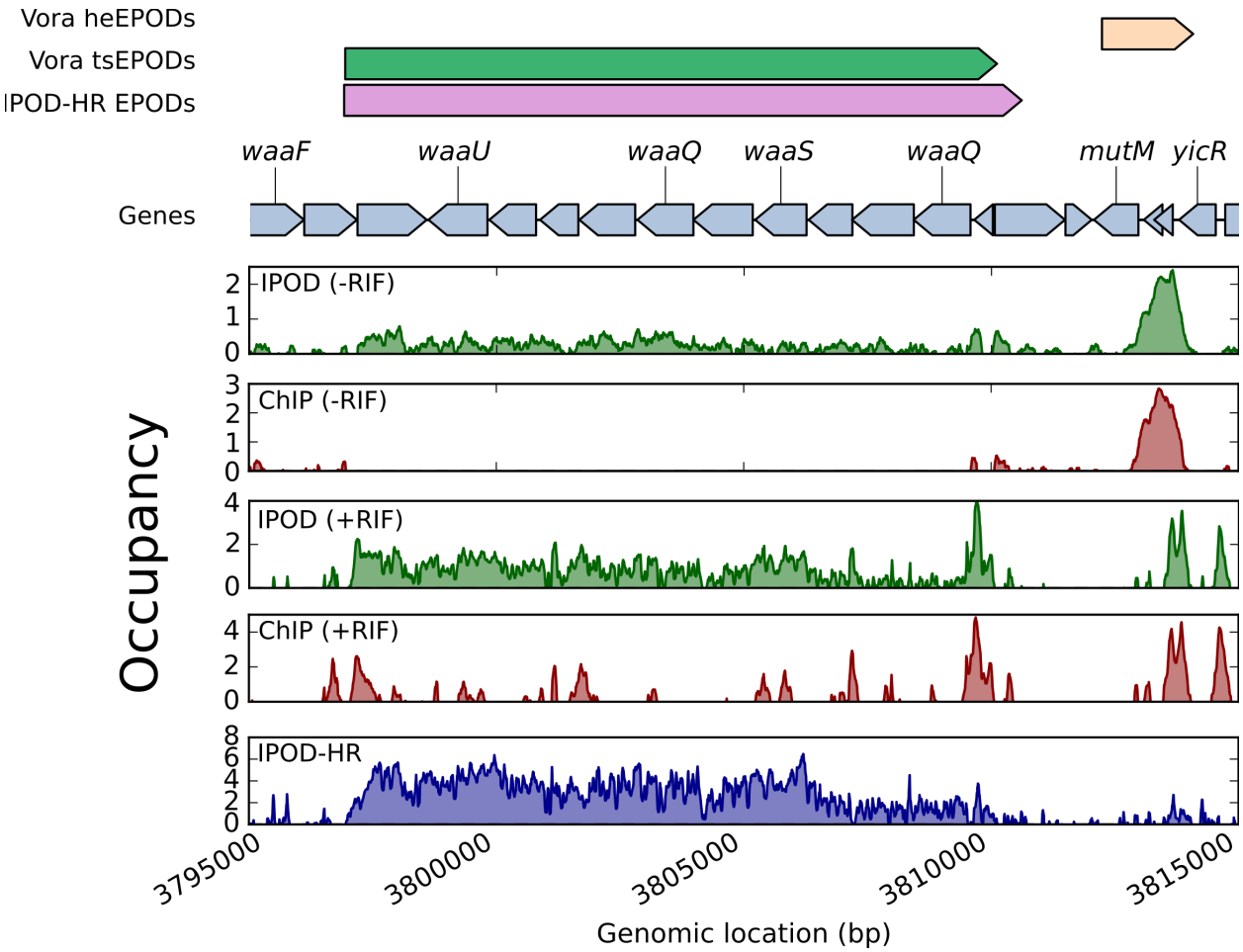

Figure S2.6: Effects of rifampin treatment on protein occupancy of a transcriptionally silent region. Shown are occupancy signals for interphase-extracted, RNA polymerase ChIP, and ChIP-subtracted IPOD occupancy (IPOD-HR) samples in the vicinity of the *waaQGPSBOJYZU* operon, which was identified as a strong tsEPOD in [10]. Signals are log2 extracted:input ratios (for IPOD and ChIP samples), or ChIP-subtracted robust z scores (IPOD-HR).

Supplementary Tables

Identified protein	Accession Number	Molecular Weight	Total Spectrum Count
UlaR	tr|C3SFV2|C3SFV2_ECOLX	28 kDa	16
YiеP	tr|E2QHU3|E2QHU3_ECOLX	26 kDa	9
RpsC	tr|C3SQX2|C3SQX2_ECOLX	26 kDa	7
RpoC	tr|C3SIA2|C3SIA2_ECOLX	155 kDa	11

RpoB	tr|E2QJ13|E2QJ13_ECOLX	151 kDa	7
RuvA	tr|C3T5R2|C3T5R2_ECOLX	22 kDa	10
FliA	tr|C3SDE6|C3SDE6_ECOLX	28 kDa	6
RpoA	tr|C3SR67|C3SR67_ECOLX	37 kDa	7
RpsD	tr|C3SR62|C3SR62_ECOLX	23 kDa	3
RdgC	tr|E2QGD8|E2QGD8_ECOLX	34 kDa	3
UvrA	tr|C3SHF7|C3SHF7_ECOLX	104 kDa	5
CysB	tr|C3TC57|C3TC57_ECOLX	36 kDa	5
FabR	tr|E2QIZ8|E2QIZ8_ECOLX	24 kDa	4
GroL	tr|Q548M1|Q548M1_ECOLX	57 kDa	3
IhfB	tr|Q14F22|Q14F22_ECOLX	11 kDa	4
RplC	tr|C3SQU2|C3SQU2_ECOLX	22 kDa	2
AccB	tr|C3SRL7|C3SRL7_ECOLX	17 kDa	4
AmiA	tr|E2QPR8|E2QPR8_ECOLX	31 kDa	3
IhfA	tr|Q14F23|Q14F23_ECOLX	11 kDa	2
Ppx	tr|C3T027|C3T027_ECOLX	58 kDa	2
RplB	tr|C3SQV7|C3SQV7_ECOLX	30 kDa	2
RpoE	tr|Q0P6M2|Q0P6M2_ECOLX	22 kDa	2
FabZ	tr|C3TPH7|C3TPH7_ECOLX	17 kDa	3
PolA	tr|E2QI51|E2QI51_ECOLX	103 kDa	2
TufB	tr|E2QFJ4|E2QFJ4_ECOLX (+1)	43 kDa	2

Table S2.1: Mass spectrometry identified peptide counts showing abundances of proteins pulled down by biotinylated bait DNA from the *sdaC* promoter region, after pruning of likely contaminants (see Methods for details).

Feature	Median difference	p value	q value	Reference
Transcriptional propensity	-1.0347	0.001	0.001	[30]
Supercoiling density	-0.1259	0.766	0.766	[80]
Normalized Tn5 integration density	4.5285	0.001	0.001	[30]
Fis binding	-1.0107	0.001	0.001	[36]
H-NS binding	11.8366	0.001	0.001	[36]
HU binding	-0.1193	0.042	0.048	[81]
LRP binding	0.3464	0.268	0.286	[59]
SeqA binding	-0.2629	0.003	0.004	[82]
Dam sites	-0.6745	0.001	0.001	Sequence analysis
AT content	1.8548	0.001	0.001	Sequence analysis
RNA	-0.6668	0.001	0.001	Present study
MGW	-0.0843	0.001	0.001	Calculated using DNAshapeR [83]
HelT	0.0707	0.001	0.001	
ProT	-0.1692	0.001	0.001	
Roll	-0.0910	0.001	0.001	

Table S2.2: Summary of EPOD characteristics across experimental conditions. The “Median difference” column refers to the difference in median robust Z-scores between EPODs and all other sites in the genome, with positive values indicating higher levels within EPODs. P-values for a significant difference are obtained using a resampling test, with 1000 random circular permutations of the EPOD locations on the genome (thus preserving the correlation structure of genomic features); q-values are obtained by correction of the p-values using the Benjamini-Hochberg method [79]. All sequence features were subjected to a 500 bp rolling mean prior to overlap calculation.

References

1. Browning DF, Butala M, Busby SJW. Bacterial Transcription Factors: Regulation by Pick “N” Mix. J Mol Biol. 2019. doi:10.1016/j.jmb.2019.04.011
2. Lee DJ, Minchin SD, Busby SJW. Activating transcription in bacteria. Annu Rev Microbiol. 2012;66: 125–152.
3. Gruber TM, Gross CA. Multiple sigma subunits and the partitioning of bacterial transcription space. Annu Rev Microbiol. 2003;57: 441–466.
4. Murakami KS, Darst SA. Bacterial RNA polymerases: the wholo story. Curr Opin Struct Biol. 2003;13: 31–39.
5. Shen BA, Landick R. Transcription of Bacterial Chromatin. J Mol Biol. 2019. doi:10.1016/j.jmb.2019.05.041
6. Ishihama A, Shimada T, Yamazaki Y. Transcription profile of Escherichia coli: genomic SELEX search for regulatory targets of transcription factors. Nucleic Acids Res. 2016;44: 2058–2074.
7. Gama-Castro S, Salgado H, Santos-Zavaleta A, Ledezma-Tejeida D, Muñiz-Rascado L, García-Sotelo JS, et al. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic Acids Res. 2016;44: D133–43.
8. Larsen SJ, Röttger R, Schmidt HHHW, Baumbach J. E. coli gene regulatory networks are inconsistent with gene expression data. Nucleic Acids Res. 2019;47: 85–92.
9. Mejía-Almonte C, Busby SJW, Wade JT, van Helden J, Arkin AP, Stormo GD, et al. Redefining fundamental concepts of transcription initiation in bacteria. Nat Rev Genet. 2020. doi:10.1038/s41576-020-0254-8
10. Vora T, Hottes AK, Tavazoie S. Protein occupancy landscape of a bacterial genome. Mol Cell. 2009;35: 247–253.
11. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res. 2007;17: 877–885.
12. Meng LM, Nygaard P. Identification of hypoxanthine and guanine as the co-repressors for the purine regulon genes of Escherichia coli. Mol Microbiol. 1990;4: 2187–2192.
13. Cho B-K, Federowicz SA, Embree M, Park Y-S, Kim D, Palsson BØ. The PurR regulon in Escherichia coli K-12 MG1655. Nucleic Acids Res. 2011;39: 6456–6464.
14. Tian G, Lim D, Carey J, Maas WK. Binding of the arginine repressor of Escherichia coli K12 to its operator sites. J Mol Biol. 1992;226: 387–397.
15. Charlier D, Roovers M, Van Vliet F, Boyen A, Cunin R, Nakamura Y, et al. Arginine regulon of Escherichia coli K-12. A study of repressor-operator interactions and of in vitro binding affinities versus in vivo repression. J Mol Biol. 1992;226: 367–386.
16. Van Duyne GD, Ghosh G, Maas WK, Sigler PB. Structure of the oligomerization and L-arginine binding domain of the arginine repressor of Escherichia coli. J Mol Biol. 1996;256: 377–391.
17. Rolfes RJ, Zalkin H. Purification of the Escherichia coli purine regulon repressor and identification of corepressors. J Bacteriol. 1990;172: 5637–5642.

18. Pittard AJ, Davidson BE. TyrR protein of Escherichia coli and its role as repressor and activator. Mol Microbiol. 1991;5: 1585–1592.
19. Shimada T, Hirao K, Kori A, Yamamoto K, Ishihama A. RutR is the uracil/thymine-sensing master regulator of a set of genes for synthesis and degradation of pyrimidines. Mol Microbiol. 2007;66: 744–757.
20. Urbanowski ML, Stauffer GV. Role of homocysteine in metR-mediated activation of the metE and metH genes in Salmonella typhimurium and Escherichia coli. J Bacteriol. 1989;171: 3277–3281.
21. Soda K. Microbial sulfur amino acids: an overview. Methods Enzymol. 1987;143: 453–459.
22. Cai XY, Redfield B, Maxon M, Weissbach H, Brot N. The effect of homocysteine on MetR regulation of metE, metR and metH expression in vitro. Biochem Biophys Res Commun. 1989;163: 79–83.
23. Shimada T, Ogasawara H, Ishihama A. Genomic SELEX Screening of Regulatory Targets of Escherichia coli Transcription Factors. Methods Mol Biol. 2018;1837: 49–69.
24. Belliveau NM, Barnes SL, Ireland WT, Jones DL, Sweredoski MJ, Moradian A, et al. Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. Proc Natl Acad Sci U S A. 2018;115: E4796–E4805.
25. Gao Y, Yurkovich JT, Seo SW, Kabimoldayev I, Dräger A, Chen K, et al. Systematic discovery of uncharacterized transcription factors in Escherichia coli K-12 MG1655. Nucleic Acids Res. 2018;46: 10682–10696.
26. Sastry AV, Gao Y, Szubin R, Hefner Y, Xu S, Kim D, et al. The Escherichia coli transcriptome mostly consists of independently regulated modules. Nat Commun. 2019;10: 5536.
27. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, et al. The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res. 2006;34: D187–91.
28. Elemento O, Slonim N, Tavazoie S. A Universal Framework for Regulatory Element Discovery across All Genomes and Data Types. Molecular Cell. 2007. pp. 337–350. doi:10.1016/j.molcel.2007.09.027
29. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome Biol. 2007;8: R24.
30. Scholz SA, Diao R, Wolfe MB, Fivenson EM, Lin XN, Freddolino PL. High-Resolution Mapping of the Escherichia coli Chromosome Reveals Positions of High and Low Transcription. Cell Syst. 2019;8: 212–225.e9.
31. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biol. 2015;109: 21.29.1–9.
32. Whitfield CR, Wardle SJ, Haniford DB. The global bacterial regulator H-NS promotes transpososome formation and transposition in the Tn5 system. Nucleic Acids Res. 2009;37: 309–321.
33. Tavazoie S, Church GM. Quantitative whole-genome analysis of DNA-protein interactions by in vivo methylase protection in E. coli. Nat Biotechnol. 1998;16: 566–571.

34. Navarre WW, McClelland M, Libby SJ, Fang FC. Silencing of xenogeneic DNA by H-NS-facilitation of lateral gene transfer in bacteria by a defense system that recognizes foreign DNA. Genes Dev. 2007;21: 1456–1471.
35. Navarre WW, Porwollik S, Wang Y, McClelland M, Rosen H, Libby SJ, et al. Selective silencing of foreign DNA with low GC content by the H-NS protein in Salmonella. Science. 2006;313: 236–238.
36. Kahramanoglou C, Seshasayee ASN, Prieto AI, Ibberson D, Schmidt S, Zimmermann J, et al. Direct and indirect effects of H-NS and Fis on global gene expression control in Escherichia coli. Nucleic Acids Research. 2011. pp. 2073–2091. doi:10.1093/nar/gkq934
37. Ling G, Waxman DJ. DNase I Digestion of Isolated Nulcei for Genome-Wide Mapping of DNase Hypersensitivity Sites in Chromatin. Methods in Molecular Biology. 2013. pp. 21–33. doi:10.1007/978-1-62703-284-1_3
38. Mieczkowski J, Cook A, Bowman SK, Mueller B, Alver BH, Kundu S, et al. MNase titration reveals differences between nucleosome occupancy and chromatin accessibility. Nat Commun. 2016;7: 11485.
39. Fang X, Sastry A, Mih N, Kim D, Tan J, Yurkovich JT, et al. Global transcriptional regulatory network for Escherichia coli robustly connects gene expression to transcription factor activities. Proc Natl Acad Sci U S A. 2017;114: 10286–10291.
40. Singh SS, Singh N, Bonocora RP, Fitzgerald DM, Wade JT, Grainger DC. Widespread suppression of intragenic transcription initiation by H-NS. Genes Dev. 2014;28: 214–219.
41. Kotlajich MV, Hron DR, Boudreau BA, Sun Z, Lyubchenko YL, Landick R. Bridged filaments of histone-like nucleoid structuring protein pause RNA polymerase and aid termination in bacteria. Elife. 2015;4. doi:10.7554/eLife.04970
42. Landick R, Wade JT, Grainger DC. H-NS and RNA polymerase: a love--hate relationship? Curr Opin Microbiol. 2015;24: 53–59.
43. Atlung T, Ingmer H. H-NS: a modulator of environmentally regulated gene expression. Mol Microbiol. 1997;24: 7–17.
44. Ono S, Goldberg MD, Olsson T, Esposito D, Hinton JCD, Ladbury JE. H-NS is a part of a thermally controlled mechanism for bacterial gene regulation. Biochem J. 2005;391: 203–213.
45. Goodarzi H, Elemento O, Tavazoie S. Revealing global regulatory perturbations across human cancers. Mol Cell. 2009;36: 900–911.
46. Freddolino PL, Amini S, Tavazoie S. Newly identified genetic variations in common Escherichia coli MG1655 stock cultures. J Bacteriol. 2012;194: 303–306.
47. Ausubel FM. Current Protocols in Molecular Biology. John Wiley & Sons; 1998.
48. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, et al. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol. 2006;2: 2006.0008.
49. Cherepanov PP, Wackernagel W. Gene disruption in Escherichia coli: TcR and KmR cassettes with the option of Flp-catalyzed excision of the antibiotic-resistance determinant. Gene. 1995;158: 9–14.

50. Huisman O, D'Ari R. An inducible DNA replication–cell division coupling mechanism in E. coli. Nature. 1981;290: 797–799.
51. Huisman O, D'Ari R, Gottesman S. Cell-division control in Escherichia coli: specific induction of the SOS function SfiA protein is sufficient to block septation. Proc Natl Acad Sci U S A. 1984;81: 4490–4494.
52. Neidhardt FC, Bloch PL, Smith DF. Culture medium for enterobacteria. J Bacteriol. 1974;119: 736–747.
53. Mooney RA, Davis SE, Peters JM, Rowland JL, Ansari AZ, Landick R. Regulator trafficking on bacterial transcription units in vivo. Mol Cell. 2009;33: 97–108.
54. Tange O, Others. Gnu parallel-the command-line power tool. The USENIX Magazine. 2011;36: 42–47.
55. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17: 10–12.
56. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30: 2114–2120.
57. Andrews S, Others. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom; 2010.
58. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32: 3047–3048.
59. Kroner GM, Wolfe MB, Freddolino PL. Escherichia coli Lrp regulates one-third of the genome via direct, cooperative, and indirect routes. J Bacteriol. 2018. doi:10.1128/JB.00411-18
60. Cleveland WS, Devlin SJ. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. J Am Stat Assoc. 1988;83: 596–610.
61. Du P, Kibbe WA, Lin SM. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. Bioinformatics. 2006;22: 2059–2065.
62. Slonim N, Atwal GS, Tkacik G, Bialek W. Information-based clustering. Proc Natl Acad Sci U S A. 2005;102: 18297–18302.
63. Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. Nat Methods. 2017;14: 687–690.
64. Quinlan AR. BEDTools: the Swiss-army tool for genome feature analysis. Curr Protoc Bioinformatics. 2014;47: 11–12.
65. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27: 1017–1018.
66. Pachkov M, Balwierz PJ, Arnold P, Ozonov E, van Nimwegen E. SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. Nucleic Acids Res. 2013;41: D214–20.
67. Elemento O, Tavazoie S. Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. Genome Biol. 2005;6: R18.

68. Castro-Mondragon JA, Jaeger S, Thieffry D, Thomas-Chollier M, van Helden J. RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. Nucleic Acids Res. 2017;45: e119.
69. Kawano M, Aravind áL, Storz G. An antisense RNA controls synthesis of an SOS-induced toxin evolved from an antitoxin. Mol Microbiol. 2007;64: 738–754.
70. Herring CD, Raffaelle M, Allen TE, Kanin EI, Landick R, Ansari AZ, et al. Immobilization of Escherichia coli RNA polymerase and location of binding sites by use of chromatin immunoprecipitation and microarrays. J Bacteriol. 2005;187: 6166–6174.
71. Campbell EA, Korzheva N, Mustaev A, Murakami K, Nair S, Goldfarb A, et al. Structural mechanism for rifampicin inhibition of bacterial rna polymerase. Cell. 2001;104: 901–912.
72. Cabrera JE, Cagliero C, Quan S, Squires CL, Jin DJ. Active transcription of rRNA operons condenses the nucleoid in Escherichia coli: examining the effect of transcription on nucleoid structure in the absence of transertion. J Bacteriol. 2009;191: 4180–4185.
73. Pettijohn DE, Hecht R. RNA molecules bound to the folded bacterial genome stabilize DNA folds and segregate domains of supercoiling. Cold Spring Harb Symp Quant Biol. 1974;38: 31–41.
74. Cabrera JE, Jin DJ. The distribution of RNA polymerase in Escherichia coli is dynamic and sensitive to environmental cues. Mol Microbiol. 2003;50: 1493–1505.
75. Dworsky P, Schaechter M. Effect of rifampin on the structure and membrane attachment of the nucleoid of Escherichia coli. J Bacteriol. 1973;116: 1364–1374.
76. Francetic O, Belin D, Badaut C, Pugsley AP. Expression of the endogenous type II secretion pathway in Escherichia coli leads to chitinase secretion. EMBO J. 2000;19: 6697–6703.
77. Breunig MM, Kriegel H-P, Ng RT, Sander J. LOF: identifying density-based local outliers. Proceedings of the 2000 ACM SIGMOD international conference on Management of data. New York, NY, USA: Association for Computing Machinery; 2000. pp. 93–104.
78. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12: 2825–2830.
79. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Series B Stat Methodol. 1995;57: 289–300.
80. Lal A, Dhar A, Trostel A, Kouzine F, Seshasayee ASN, Adhya S. Genome scale patterns of supercoiling in a bacterial chromosome. Nat Commun. 2016;7: 11055.
81. Prieto AI, Kahramanoglou C, Ali RM, Fraser GM, Seshasayee ASN, Luscombe NM. Genomic analysis of DNA binding and gene regulation by homologous nucleoid-associated proteins IHF and HU in Escherichia coli K12. Nucleic Acids Res. 2012;40: 3524–3537.

82. Joshi MC, Magnan D, Montminy TP, Lies M, Stepankiw N, Bates D. Regulation of sister chromosome cohesion by the replication fork tracking protein SeqA. PLoS Genet. 2013;9: e1003673.
83. Chiu T-P, Comoglio F, Zhou T, Yang L, Paro R, Rohs R. DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. Bioinformatics. 2016;32: 1211–1213.

Chapter 3

Distinct Heterochromatin-like Domains Promote Transcriptional Memory and Silence Parasitic Genetic Elements in Bacteria

Abstract

There is increasing evidence that microbes maintain a structured chromosome, which in turn impacts gene expression. However, tools to profile the genome landscape and binding of key architectural proteins have been limited. We recently discovered densely occupied, multi-kilobase regions in *E. coli* that are transcriptionally silent, similar to eukaryotic heterochromatin. These regions, termed EPODs, overlap metabolic pathways and parasitic elements such as prophages. Here, we investigate the contributions of nucleoid associated proteins (NAPs) to these domains by examining the impacts of deleting NAPs on EPODs genome-wide in *E. coli* and an evolutionarily distant species, *Bacillus subtilis*. We identify key NAPs contributing to the silencing of specific EPODs, where deletion of a particular NAPs opens a chromosomal region to RNA polymerase binding. In *E. coli,* we distinguish novel xenogeneic silencing NAPs, Fis and Hfq, that are essential for cell viability in the presence of domesticated prophages. Furthermore, we show that changes in EPODs facilitate an extra layer of transcriptional regulation to prepare cells for exposure to exotic carbon sources. Our findings demonstrate a new suite of mechanisms through which genomic architecture primes bacteria for changing metabolic environments and silences harmful genomic elements.

The contents of this chapter have been submitted to Review Commons by Haley M. Amemiya, Thomas J. Goss, Taylor M. Nye, Rebecca Hurto, Lyle A. Simmons, and Peter L. Freddolino. H.M.A .and P.L.F.; Methodology, H.M.A., T.G., and P.L.F; Investigation, H.M.A., T.G., T.N., L.A.S., R.H., and P.L.F.; Data Analysis and Curation, H.M.A. and P.L.F.; Writing -- Original Draft, H.M.A., T.N., L.A.S., and P.L.F.; Funding Acquisition, L.S. and P.L.F. All authors reviewed the manuscript. For this manuscript, I performed all experiments from strain construction to assay, to computational analysis with a few exceptions: T.G. performed initial replicates of IPOD-HR for WT, Δ*dps,* Δ*stpA,* Δ*hns,* Δ*hupAB,* and deep stationary samples. T.N. and R.H. performed IPOD-HR in *B. subtilis*. I created all figures and wrote the manuscript.

Introduction

All organisms must organize immense amounts of genetic information into a relatively small physical space within the cell. Paradoxically, the maintenance and accessibility of the DNA architecture is required for efficient DNA replication, repair, and transcription, and thus for proper cell division required to sustain life. In the bacterium *Escherichia coli* (*E. coli*), chromosome structure is mediated by roughly a dozen small, basic nucleoid-associated proteins (NAPs) [1–3] that work in concert to modulate supercoiling, DNA looping, and distant chromosomal contacts [4–8]. Despite substantial research effort, it is largely unclear how individual NAPs impact the overall structure of the chromosome, due in part to their promiscuous and overlapping binding across the genome [3,9]. Our previous findings, consistent with many studies in the literature, have implicated H-NS as a dominant gene silencer [1,5,10,11]. H-NS has the capacity to form filaments, tightly compacting dsDNA. *In vitro*, it has been shown that the two types of filaments, linear or bridged, block transcription initiation, while only bridged filaments block transcription via blocking elongation[12–14] . Other proteins, such as Hha and the H-NS paralog StpA, promote the formation of H-NS filaments and modulate their structural properties [6,15]. In addition to H-NS's role in global gene silencing, it is well documented to specifically silence horizontally acquired DNA [11,16–19]. While the dominant form of filaments *in vivo* remains unknown, it is clear that H-NS plays a major role in transcriptional silencing in bacteria.

In order to gain a comprehensive understanding of the contribution of NAPs to global protein occupancy and gene expression, we used *in vivo* protein occupancy display at high resolution (IPOD-HR). IPOD-HR is a method that provides a global snapshot of areas in the genome where proteins are bound to DNA, thus yielding insight into genome-wide regulation and structure [11,20]. IPOD-HR revealed approximately two hundred regions of the *E. coli* genome that are densely packed with protein but exclude RNA polymerase, inviting striking comparisons to heterochromatin found in eukaryotes [11]. We refer to these densely-packed regions as extended protein occupancy domains (EPODs) [20]. While EPODs have been shown to be partially occupied by NAPs and in some cases overlap with known binding sites for NAPs such as H-NS [11], the

contributions of individual NAPs on EPOD formation remains unexplored. Recent data suggests that EPODs contribute to both the regulation of metabolic pathways and the silencing of horizontally acquired DNA [11,21], however the NAPs mediating this response for individual EPODs are unknown. Additionally, evidence suggests that bacterial genomes have integration hotspots for horizontally acquired DNA [22]. We have found that EPODs have a higher integration frequency compared to the rest of the genome[11], perhaps giving insight that EPODs may be the functional unit that serve as hotspots for foreign DNA. Understanding the key protein components of EPODs and EPODs' roles in genome organization will shed light on the relationship between the regulation of transcription and genome architecture in bacterial genomes.

Here, we investigate the contributions of the major NAPs in *E. coli* to maintaining the pattern of global protein occupancy. By tracking global protein occupancy across the chromosome, our approach fundamentally differs from (and complements) what would be provided by a series of ChIP-seq experiments. Rather than simply seeing where one particular protein binds, we monitor the complete set of changes in protein occupancy caused by loss of any single NAP. Loss of a NAP may result in changes across the genome due to the combination of its direct binding, physical cooperation and competition with other factors, and regulatory effects. In fact, we show that loss of any single NAP results in changes to the global landscape of protein occupancy, albeit often less dramatic than might be expected. We find that NAPs are able to compensate for each other to maintain EPOD structure even after single NAP knockouts, often acting in pairings. Deletion of specific combinations of NAPs further supports this hypothesis and reveals key silencers for specific processes, such as the combined activities of StpA and H-NS in silencing many H-NS targets. We document one particular case where an H-NS dependent EPOD both maintains properly silencing of the operon for metabolizing a rare carbon source under baseline conditions, and facilitates a transcriptional memory response that enable them to induce the same operon faster upon a second exposure to the same carbon source, providing yet another analogy to eukaryotic chromatin.

In the process, we also uncover additional evidence consistent with the well documented contribution of H-NS[11,16] in the silencing of harmful genetic elements that have integrated into the genome, such as latent bacteriophages [6,16]. However, we show that H-NS is not unique in its role as a xenogeneic silencer, and that a variety of other NAPs contribute to the silencing of prophages in a locus-specific manner. Of particular importance, the loss of two specific NAPs, Hfq and Fis, leads to inviability in a prophage dependent manner, underscoring the importance of different NAPs in establishing EPODs to maintain cellular health.

Although originally identified in *E. coli,* we find that EPOD-like structures exist in a broad range of bacteria, including the evolutionarily distinct Gram-positive Firmicute *Bacillus subtilis*. Similarly to *E. coli*, we show that nucleoid associated proteins in *B. subtilis* facilitate silencing of metabolic pathways and horizontally acquired genes. Overall, we present a unifying conceptual framework in which heterochromatin-like domains across diverse bacterial species serve both to provide architectural regulation of metabolism and as a bacterial 'innate immune system' for potentially toxic DNA.

Results

Large-scale patterns of protein occupancy are highly maintained across conditions and laboratory evolution

The ~200 EPODs identified across the *E. coli* MG1655 (WT) genome show significantly enriched overlaps with loci encoding genes involved in metabolism and silencing of mobile elements [11]. As remodeling of nucleoid organization has been previously detected in response to environmental changes, such as media richness and growth phase [23–25], we tested whether EPODs were one of the functional units that mediated changes across growth conditions. To examine the robustness of these domains under various physiological conditions, we performed IPOD-HR in MG1655 strains, varying the media type, growth phase, and parental strain origin. IPOD-HR was

performed as previously described [11]; samples were taken at mid-exponential growth phase or deep stationary growth phase (D.S.; described in Methods below), in both rich defined media and minimal media with glucose (see Methods). For all experiments, we used defined media due to maintain consistency in physiological experiments with buffer conditions, salts, etc and to avoid the noted variation between Lysogeny Broth (LB medium) [26,27]. Even within *E. coli* MG1655 strains, there are genetic differences from variation in parental origin [28] that have led to differences in what is deemed "WT" across different laboratories. To examine the robustness of EPODs in a relatively similar strain with minor genetic differences, we utilized another MG1655 variant (labeled MG1655 (2)) (described in Methods below).

The number and overall genomic coverage EPODs across conditions and genetic backgrounds remained relatively stable (**Fig. 3.1A,B**), with the greatest increases in coverage and counts in cells in deep stationary phase and the MG1655 (2) variant. To examine the changes in the locations and boundaries of EPODs, we calculated a measure that we refer to as the symmetrized overlap distance, given by the difference between unity and the geometric mean of the A-B and B-A EPOD overlaps:

$$1 - \sqrt{A_B \times B_A}$$

Where A_B is the fraction of condition B's EPODs overlapped by the relaxed threshold EPODs contained in condition A, and B_A is the fraction of condition A EPODs overlapped by the relaxed threshold EPODs from condition B. A value of 0 indicates identical EPOD locations, since both overlaps would be 1. The use of strict vs. loose comparisons avoids overstating the differences between conditions based on minor thresholding differences and emphasizes large and systematic changes in occupancy. The symmetrized overlap distances between all pairs of conditions examined here are shown in **Figure 3.1C**. Values within the heatmap displayed are calculated using the fraction of EPODs contained in the relaxed threshold EPODs, defined as in [11], for each condition. Hierarchical clustering analysis of these distances reveal that the major differences stem decreasingly from growth phase, genotype, and then media changes (**Fig. 3.1C**). Notably, the EPOD location profiles for the two distinct MG1655 lineages

considered here were more similar to each other than the deep stationary phase condition, but in all cases a substantial majority of EPODs are conserved across conditions. These findings support ongoing research that shows major changes in nucleoid composition during later phases of growth[8,9]. Previous investigation of EPODs under baseline growth conditions (exponential growth in rich defined media) demonstrated that genes contained in EPODs were enriched for several gene functionalities, including DNA transposition, cytolysis, and LPS biosynthesis [11] .To identify the pathways maintained across all conditions considered in **Figure 3.1**, we ran iPAGE [29] to associate which gene ontology (GO) terms fell within EPODs relative to Background. We found that three GO terms remained highly enriched in EPODs under all four conditions shown here: cytolysis, LPS biosynthetic pathway, and cellular response to acid chemical, demonstrating that EPOD-mediated regulation of these functionalities persists across both physiological condition and even cell lineages (**Fig. S3.1**).

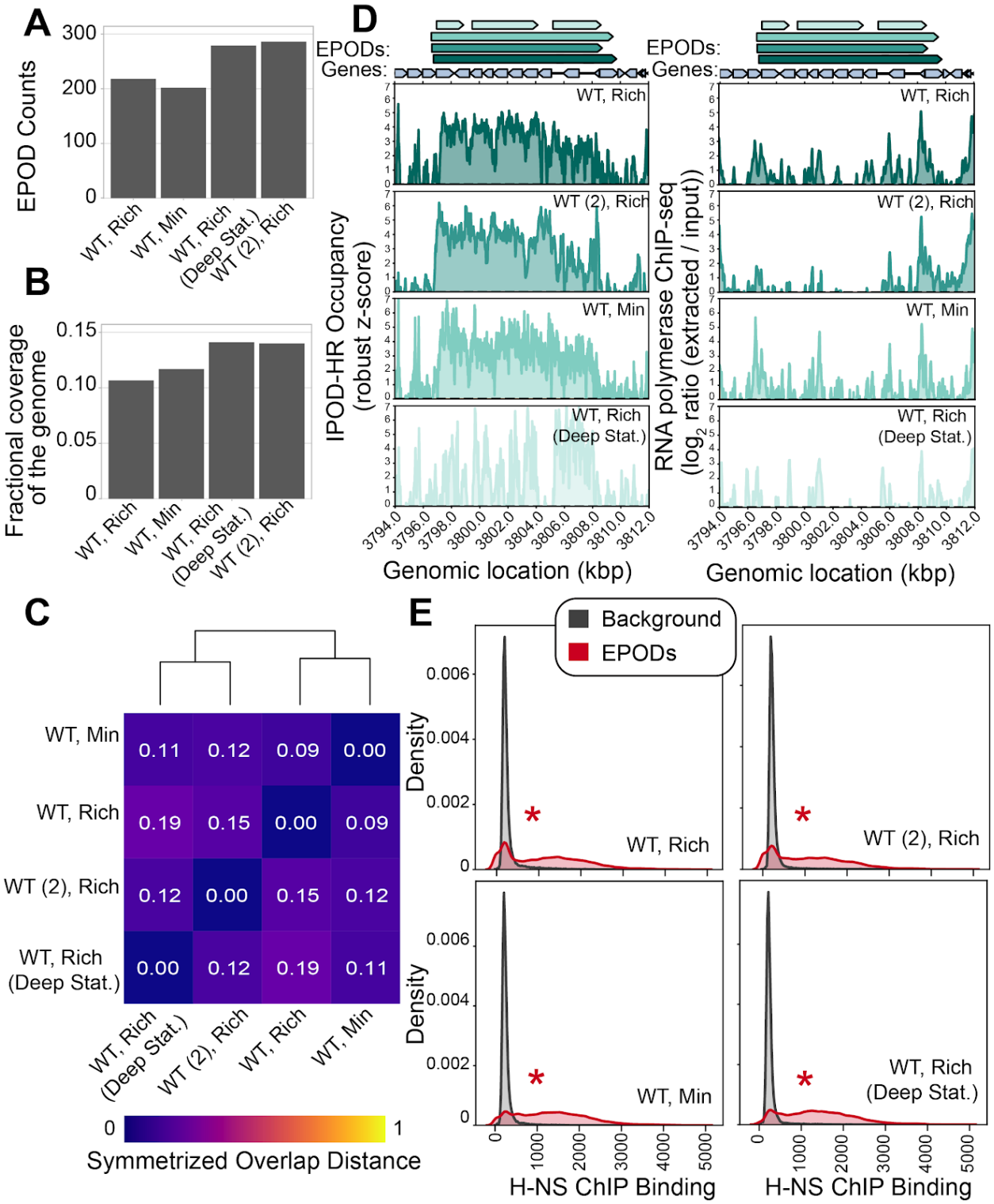

Figure 3.1: EPODs are highly robust across growth conditions. (A) EPOD counts across WT (MG1655) *E. coli* cells are similar across different conditions. Cells were grown in: Rich (Rich Defined Media (See Methods)), Min (Minimal Media) and collected at log phase growth or deep stationary phase (Deep Stat.). WT (2) is another MG1655 variant that comes from a different lab's parental strain and has a few genetic

differences, as described in Methods. **(B)** Fractional coverage of EPODs across the genome does not largely vary across conditions, however cells grown in the Min condition have a slightly larger coverage. **(C)** To assess the similarity between EPOD calls in each condition, we calculated the symmetrized overlap distance of EPODs. A value of 0 indicates that the set of EPODs are identical. Hierarchical clustering reveals that growth phase impacts EPOD location in this dataset. **(D)** Specific locations across all conditions remain occupied by protein where EPODs were called (left) and unoccupied by RNA polymerase (right). The quantile normalized robust z scores of the protein occupancy at each 5 bp are represented by the IPOD-HR occupancy. **(E)** Density plots displaying the normalized histograms (smoothed by a kernel density estimator) of H-NS ChIP [17] for regions of the genome within EPODs versus Background (in which each contiguous non-EPOD block is treated as a single data point). H-NS binding shows a strong overlap with EPOD locations measured in all conditions. (*) indicates FDR-corrected $p<0.005$ via permutation test (against a null hypothesis of no difference in medians).

Silencing of the non-functional LPS gene pathway is maintained across conditions

To better understand the dynamics of protein occupancy across different conditions at a locus that shows a heavily maintained EPOD, we examined the changes in occupancy at an LPS biosynthesis locus across conditions (**Fig. 3.1D**). The LPS pathway was noted to be silenced by EPODs in our initial findings [11]. We hypothesize that these components of the LPS pathway are silenced due to an insertion element in *wbbL* in MG1655, an upstream component of the LPS pathway. Typically, *E. coli* express these pathways as part of O antigen biosynthesis, a highly beneficial cellular component that increases resistance to phage infection and environmental pressures [30–35]. Without functional genes in this pathway, there are no biological benefits to expressing downstream genes under various physiological conditions, and silencing may have been selected for over the course of the subsequent evolution of MG1655. Consistent with prior observations, we observe robust protein occupancy and minimal RNA polymerase occupancy across all conditions (**Fig. 3.1D**), although a qualitative change in the locations of the high-occupancy regions is apparent in deep stationary phase, perhaps reflecting a turnover in the predominant NAPs present under that condition.

Nucleoid associated proteins are the main components of EPODs.

We hypothesized that the maintenance of EPODs was largely driven by the NAP occupancy. H-NS has been widely described as the major component of the *E. coli* nucleoid in exponential growth, specifically inhibiting transcription and silencing mobile elements and prophages [6,10,16]. We compared H-NS binding within EPODs compared to background and found significant enrichments of H-NS binding in EPODs across all considered conditions (**Fig. 3.1E**), consistent with previous observations [20]. Thus, H-NS is facilitating silencing of regions across the genome robustly in varying media, growth phase, and slight genotype differences. However, we also found previously that while the majority of EPODs overlap with known H-NS binding regions, a substantial fraction do not [11]. In addition, the fact that H-NS is present at a particular EPOD does not necessarily mean that it is the only factor (or even a necessary or sufficient factor) in forming that EPOD and silencing the genes there. We thus sought to assess whether other NAPs facilitate silencing of specific EPODs or if H-NS is the major silencing factor of EPODs across the *E. coli* genome.

The binding locations and biological roles of nucleoid associated proteins (NAPs) have been difficult to define largely because of their promiscuous binding across the genome. Due to their propensity to bind DNA and their high abundance in the cell, we hypothesized that multiple NAPs contribute to EPODs. To examine the contributions of a range of *E. coli* NAPs to EPODs, we performed single deletions of the most abundant *E. coli* NAPs (*hns*, *stpa*, *fis*, *hfq*, *ihfAB*, *dps*, and *hupAB*) and performed IPOD-HR in the deletion strains. Since StpA is a known paralog of H-NS and forms bridged filaments across DNA [6,7], we also created an *hns/stpA* double knock-out and performed IPOD-HR. As the NAP Dps is primarily expressed during stationary phase of growth [36,37], we performed IPOD-HR in WT and Δ*dps* cells that were collected during deep stationary phase stage of growth (defined in Methods) as well as in exponential phase. EPOD counts and coverage slightly varied across genotypes tested (**Fig. 3.2A,B**), with the largest loss of coverage observed in Δ*stpA*Δ*hns.* To examine shifts in EPOD locations, we again measured the symmetrized overlap distance (defined for **Fig. 3.1C**) and performed hierarchical clustering (**Fig. 3.2C**). The greatest changes in EPOD locations relative to the baseline condition (WT cells in exponential phase) occur within

Δ*stpA*Δ*hns* and Δ*ihf,* both of which cause profound changes in the profile of EPOD locations. The deep stationary phase samples cluster together, indicating similar shifts in EPOD locations. Interestingly, Δ*hns* and Δ*hfq* cluster together, perhaps suggesting a similar role in silencing at some EPODs. Several other NAP deletions show minimal effects on EPODs relative to WT cells, including deletions of *hupAB*, *stpA*, and *dps* (in exponential phase), indicating that at least under baseline conditions, these proteins do not contribute strongly to defining EPODs.

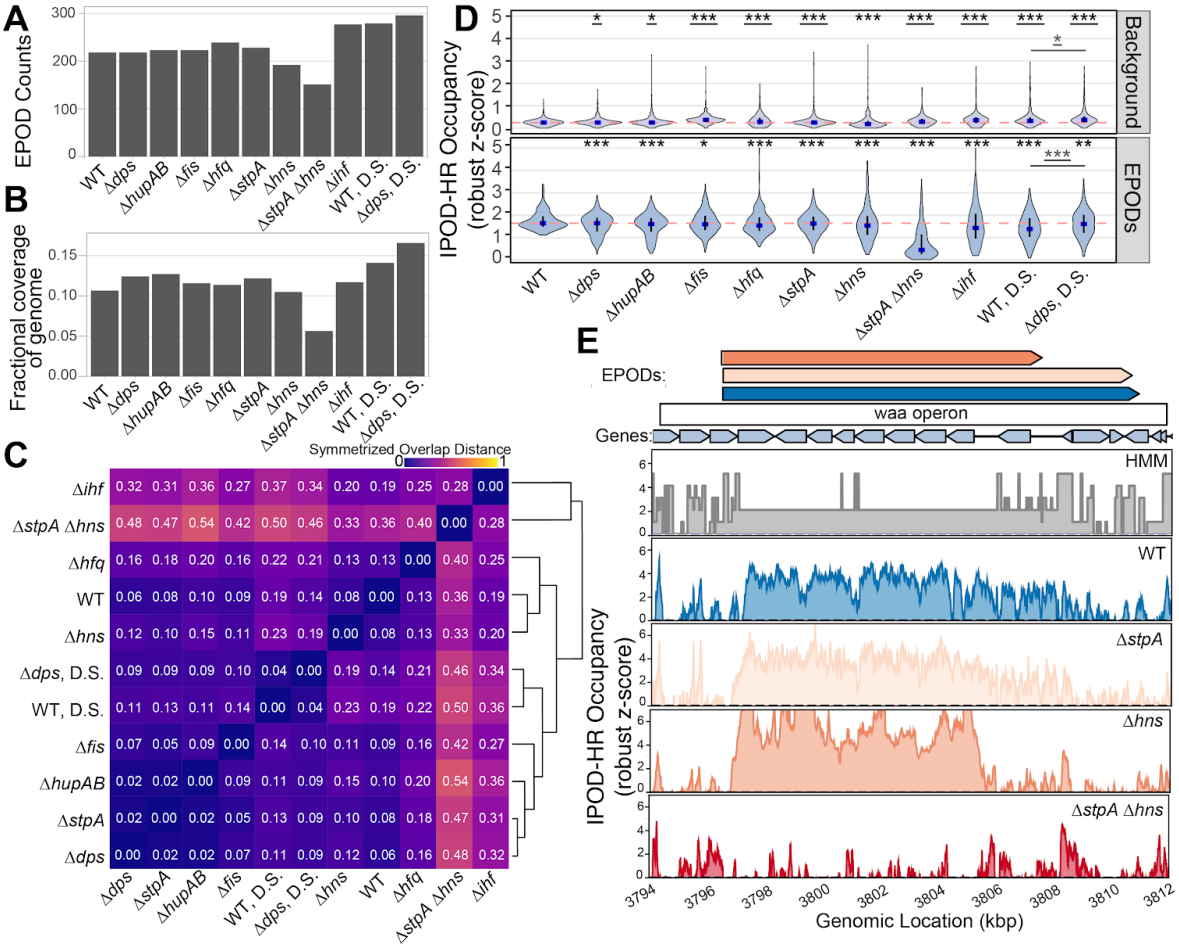

Figure 3.2: Loss of nucleoid associated proteins (NAPs) leads to changes in EPODs. (A) Number of EPODs called in different NAP deletions. D.S. denotes genotypes where cells were collected in the deep stationary phase of growth. **(B)** Fractional coverage of EPODs for each genotype across the genome. **(C)** Symmetrized overlap statistic comparing each pair of samples in the NAP deletion dataset. The symmetrized overlap denotes similarity between EPOD locations, where a value of 0 = identical. Hierarchical clustering was performed to group like-genotypes. **(D)** Distributions of mean protein occupancy at WT EPOD boundaries and background. The blue dots denote the median and the black line displays the interquartile ranges in each condition. The dashed pink line represents the WT median. (*) indicate the Wilcoxon

Rank Sum test p value comparing the change in median vs. WT for each condition that has been adjusted using the Benjamini and Hochberg method (against a null hypothesis of no difference in medians). The grey line denotes the same comparison between the D.S. conditions. Underlined (*)'s indicate a gain in the median compared to baseline conditions, while no underline indicates a loss in the median compared to baseline conditions. p value < 0.05 = *, <0.005=**, <0.0005=***. **(E)** Protein occupancy over the *waa* operon. The quantile normalized robust z scores of the protein occupancy at each 5 bp are represented by the IPODHR occupancy. The EPOD over the waa operon is lost in the Δ*stpA*Δ*hns* condition which results in increased accessibility and RNA polymerase occupancy (**Fig. S3.3A**).

Different EPODs are comprised of distinct combinations of NAPs

To identify the relative contribution of different NAPs in maintaining a 'standard' set of EPODs present in WT cells during exponential phase growth, we calculated the average of the total protein occupancy signal across every EPOD location and other genomic regions for each genotype (**Fig. 3.2D**). When comparing the WT median (pink dashed line) vs. the deletion mutant medians (blue), NAPs contributing to protein occupancy at EPODs normally present in exponential phase display a significant dip in occupancy. Consistent with our prior observations, the loss of occupancy at normal EPOD boundaries in Δ*stpa*Δ*hns* cells is particularly profound; notably, however, the *hns* single mutant shows only a minor loss of occupancy, indicating that StpA can largely compensate for its loss. Several other NAP deletions also showed significant drops in occupancy at the standard EPOD locations, often with a bimodal distribution of occupancy changes suggesting that only some subset of EPODs were affected in each case. At the same time, in many cases, RNA polymerase occupancy at a subset of affected EPODs is observed to rise (**Fig. S3.2A**), indicating a derepression of some EPODs upon deletion of *hupAB*, *hfq*, *hns*, *ihf*, or (especially) in the *stpA/hns* double knockout. Additionally, all NAPs exhibited changes in EPOD locations that impacted the overlap of called EPOD regions. We examined the fraction of EPODs contained in the relaxed threshold EPODs and found the overlap between EPOD locations among different genotypes dipped as low as 38% (**Fig. S3.2C**). Changes in protein occupancy compared to WT at all regions of the genome and downstream hierarchical clustering suggests a similar role of Hfq and H-NS in silencing specific regions of the genome (**Fig. S3.3**).

Inspection of the patterns of both total protein occupancy and RNA polymerase occupancy across the *waa* operon provides an instructive example. Whereas single deletions of *hns* or *stpA* show a minor effect on the integrity of the *waa* EPOD, the Δ*stpa*Δ*hns* cells show nearly complete loss of occupancy in this region (**Fig. 3.2E**). At the same time, there is a concomitant gain in RNA polymerase occupancy and induction of RNA expression of *waa* operon genes (**Fig. S3A,B**), demonstrating that H-NS and StpA act jointly to maintain silencing of the *waa* operon. While either can compensate for the other, the loss of both silencers leads to substantial de-repression of genes in this region.

Our data thus permit us to identify which NAPs regulate specific EPODs across the genome. To provide an automated high-level classification of the regions across the genome specific to particular NAPs, we used our IPOD-HR occupancy and RNA polymerase ChIP-seq datasets across the NAP deletions and trained a Hidden Markov Model (HMM) that split the genome up into six classes (see Methods for details). We were able to identify three classes (2, 3, and 5) that were associated with EPODs, which exhibited significant enrichments of IHF motifs[38], Hfq binding (See note in Methods), H-NS binding[17], and Fis binding[17], as well as a low abundance of motifs for the known DNA methylases - Dam and Dcm [38] (**Table S3.1**). In particular, HMM class 2 is especially strongly associated with H-NS binding and likely represents H-NS dependent EPODs (with IHF also apparently contributing); HMM class 5 is associated with high levels of Hfq and Fis binding and may represent EPODs comprised in part by these two factors. In addition, class 5 had a high abundance of transcription factor binding sites and promoters[38], further implicating this class serving a regulatory role; HMM class 3 represents yet another category of EPODs that at present cannot be assigned. The utility of the HMM classification is apparent. For example, the *waa* operon is almost entirely associated with HMM class 2 (**Fig. 3.2E**), which we assign as an H-NS filament. On the other hand, the borders of the EPOD instead fall into HMM class 3 or 5, and, consistently, show loss of occupancy in an *hns* deletion strain even if *stpA* is intact (**Fig. 3.2E**). Together, we find that StpA and H-NS together contribute to large protein regions

across the genome, and are the main components silencing the LPS biosynthesis pathway. At the same time, we observe the presence of two distinct classes of EPOD occupancy that appear largely H-NS independent, and likely represent different types of large-scale repressive protein occupancy.

Metabolic pressures induce changes in EPODs

Many EPODs overlap operons involved in metabolic pathways, and thus silence pathways that may not be actively used in the cell under the conditions that we studied. For example, during growth in glucose rich defined media we observed an EPOD, associated with HMM class 2 (H-NS filament), overlapping the *idnDOTR* operon, specifically in the promoter of the operon. The *idn* operon is essential for the metabolism of carbon sources such as idonate and 5-keto gluconate, which are not present under typical laboratory conditions. In particular the *idnD* gene codes for the enzyme L-idonate 5-dehydrogenase, which catalyzes the oxidation of L-idonate to 5-ketogluconate [39–41]. The *idnDOTR* operon and *idnK* are known to be transcriptionally regulated by CRP, IdnR, GlaR, and GntR, but to our knowledge no connection to regulation by NAP occupancy has previously been described [42]. Upstream of the *idnDOTR* operon, there is a 215-bp regulatory region that lies in between *idnK* and *idnD* [39]. In this region, there is a single putative IdnR/GntR binding site, CRP binding site, and an UP element [39]. Bausch *et al.* previously showed that induction of this pathway can occur due to exposure to L-idonate or 5-keto-gluconate (5KG) [39]. The local positive regulator, IdnR, is activated by 5KG [41], and promotes the induction of the rest of the operon. Following exposure to 5KG in the absence of glucose, IdnT enables uptake of the carbon source, IdnD (as previously mentioned) catalyzes the reversible reduction of L-idonate to 5-ketogluconate, IdnO catalyzes the oxidation of 5-ketogluconate to D-gluconate, and idnK catalyzes the phosphorylation of D-gluconate which then proceeds through the Entner-Doudoroff pathway to be metabolized [41]. Since the *idn* operon appears to be controlled by three local and one global regulators, the question arises of what additional regulatory role might be played by the apparently silencing EPOD covering the *idn* promoter under normal conditions.

To explore the mechanisms of silencing at the *idnDOTR* operon, we first referred to our NAP deletion dataset to see whether specific NAP(s) silenced the operon. Data from a previous H-NS ChIP-seq dataset [17] plus the classification of this region as a type 2 EPOD in our HMM suggested the presence of H-NS bound to the *idn* promoter region. We performed RNA-seq in the Δ*stpa*Δ*hns* background compared with the parental cells, and discovered that H-NS and StpA indeed repress expression of the *idnDOTR* operon (**Fig. S3.5A,B**). We leveraged this knowledge to address whether we could induce changes in protein occupancy at the EPOD covering the *idn* promoter by performing a carbon source shift experiment outlined in (**Fig. 3.3A**; further explanation in Methods). Briefly, cells were grown in minimal media with 0.2% glucose, shifted to minimal media with 0.2% 5-keto-D-gluconate (5KDG) as a sole carbon source, and shifted back to minimal media with 0.2% glucose as the carbon source. In all conditions, cells were collected at an OD600 of ~0.1 for both IPOD-HR to examine changes in EPODs and RNA-seq for changes in expression. Notably, there was a severe lag for growth in 5KDG. We found that growth in 5KDG led to a reduction in protein occupancy and loss of the EPOD within the *idnDOTR* operon promoter (**Fig. 3.3B**). Upon shifting the cells back to glucose, both the original pattern of protein occupancy and the EPOD were restored (**Fig. 3.3B**). The loss of EPOD occupancy in the 5KDG condition was accompanied by an induction of expression of the *idnDOTR* operon, and repression when EPOD occupancy was restored upon the return to glucose as a carbon source (**Fig. 3.3C**). Due to the long lag in growth during the transfer from glucose to 5KDG as a carbon source, we examined the correlation between the expression of all genes in the three conditions to see whether there were broad changes in expression when cells were forced to metabolize an exotic carbon source (**Fig. 3.3D**, top panel). The Spearman correlation was extremely high when comparing all conditions (>0.9), suggesting that changes are localized to the *idn* operon and the small set of other genes specifically regulated in response to the 5KDG carbon source. Similarly, the symmetrized overlap distances comparing the variation in EPOD locations across conditions were low, again supporting the notion that changes are specific to the

induced operon (**Fig. 3.3D**, bottom panel) and that no global rearrangement of protein occupancy occurs.

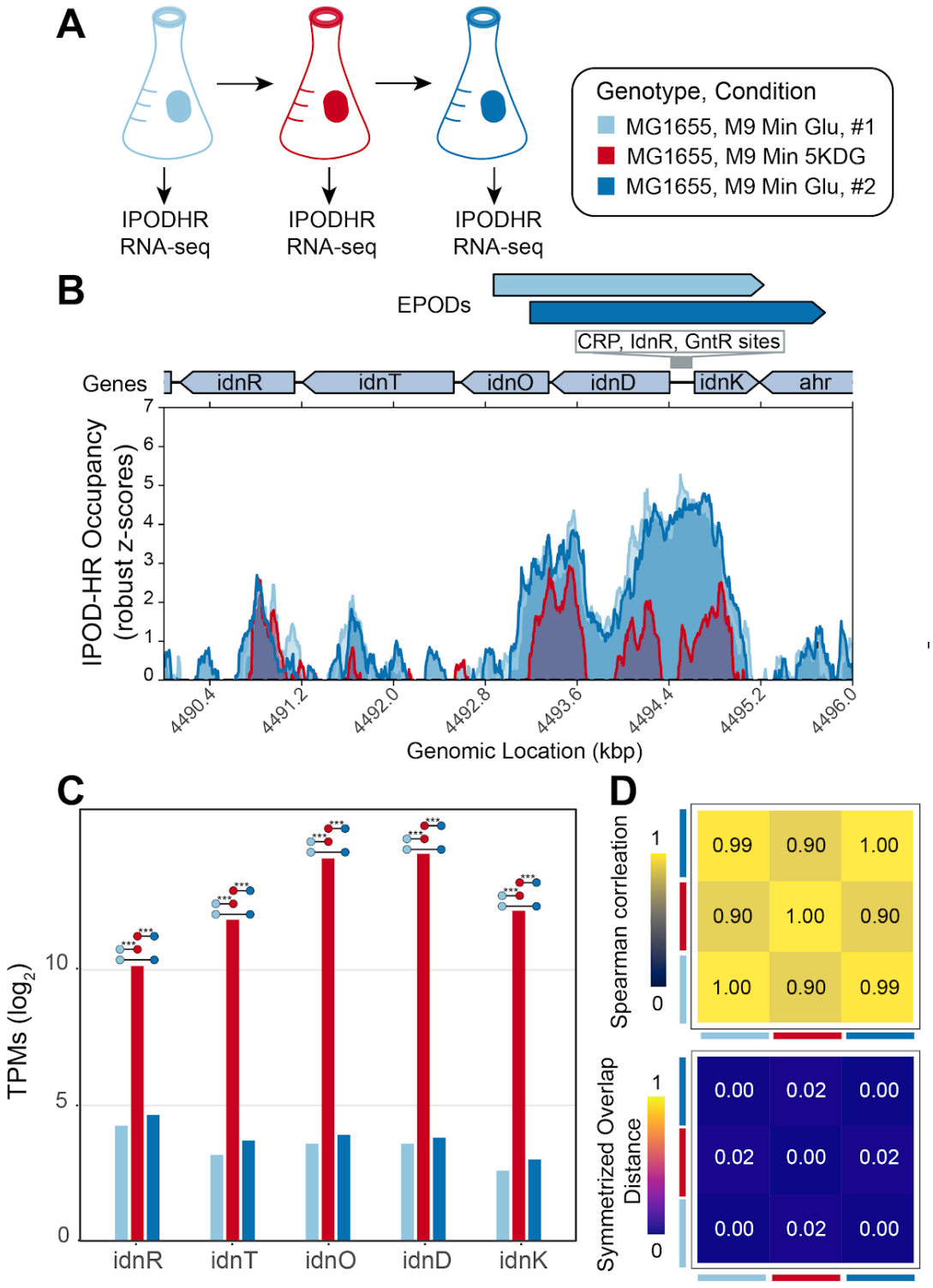

Figure 3.3: Changes in EPODs are induced in specific conditions. (A) Experimental overview. WT, MG1655 cells were grown in M9 Minimal Media with 0.2% glucose,

samples were collected at mid log phase of growth (OD600 ~0.2) for RNA-seq and IPODHR. The cells were back diluted to an OD600 of ~0.1 in M9 Minimal Media with 0.2% 5-Keto-D-gluconic-acid (5KDG), grown to an OD600 of ~0.2, collected for RNA-seq and IPODHR. In the final shift, the cells were back diluted to an OD600 of ~0.003 in M9 Minimal Media with 0.2% glucose, grown to OD600 of 0.2 and collected for RNA-seq and IPODHR. Two biological replicates were performed. **(B)** Protein occupancy over the *idn* operon for each condition (colors are denoted in **(A)**)**.** The quantile normalized robust z scores of the protein occupancy at each 5 bp are represented by the IPODHR occupancy. There is a large loss in protein occupancy when cells are shifted to 5KDG, leading to the loss of the called EPOD. Protein occupancy is restored once cells are grown in the second glucose condition. **(C)** To examine the expression of the idn operon at each shift, RNA-seq was performed. RNA-seq expression estimates (from Rockhopper) $\log_2$ scaled for *idn* operon genes for WT cells grown in each condition colored in **(A).** Comparisons are denoted with colored dots with significance stars representing the q-value calculated by Rockhopper [72] using a negative binomial distribution (against the null hypothesis that the expression of the transcript in two conditions is the same). where (***) signify q-values <0.0005. **(D)** Spearman correlations are represented with the heatmap comparing the expression profiles in each condition, where identical expression values for every gene show a spearman correlation of 1. The symmetrized overlap distance was calculated for all EPODs for each condition, where a value of 0 is identical. The colored squares on the sides of the heatmaps denote the condition with the colors represented in **(A).**

Given the sophisticated regulatory logic implemented at the *idn* promoter by a combination of local (GlaR, GntR, IdnR) and global (CRP) regulators, it is unclear what additional function is played by the apparently repressive EPOD covering this region. Drawing inspiration from the behavior of chromatin modifications in eukaryotes (e.g. [43–46]), we hypothesized that one function of EPODs could be to facilitate transcriptional memory. To test this, we measured growth in a new shift experiment, and added another shift back to 5KDG (**Fig. 3.4A**). The median lag time for the first shift into 5KDG across three biological replicates was 48.6 hrs, with a range of 30.6 - 52.1 hrs. The cells were then back diluted into minimal media with 0.2% glucose, grown to mid exponential phase (OD600 ~0.2; around 6 doublings), and back diluted into 5KDG again. Here, the median lag time for cells previously exposed to 5KDG were 10.2 hrs with a range of 6.4-11.5hrs (**Fig. 3.4B**). This dramatic change in lag time between the first and second exposures to KDG suggests that expression of the gene products needed to metabolize 5KDG occurs faster upon the second exposure to the carbon source, despite the transcriptional regulatory state having reset to be virtually

indistinguishable from the original round of growth in glucose minimal media prior to the second 5KDG challenge (**Fig. 3.3**). Given the long outgrowth (5.5-6 doublings) between the first and second 5KDG challenges, it is expected that essentially all of the Idn proteins would have been diluted to irrelevant levels. While additional direct evidence is needed, our findings are consistent with the possibility that the structure of the EPOD in this region is such that transcriptional initiation is faster upon second induction within some time window after an initial induction, providing a transcriptional memory that facilitates responses to repeated stresses. Such a memory could be implemented, for example, by formation of bridged (after long timescales) vs. unbridged (for some time period after formation) H-NS filaments in this region, or by post-translational modification of the H-NS comprising the EPOD (analogous to the histone code[47]). The far longer lag time of the cells upon their second growth period in glucose minimal media relative to the first exposure (see **Fig. 3.4B**) also suggests that some of the gene products induced to metabolize 5KDG may themselves be detrimental under normal conditions, and thus another role of the EPOD at the *idn* promoter may be to ensure sufficiently tight silencing of these genes except when they are needed.

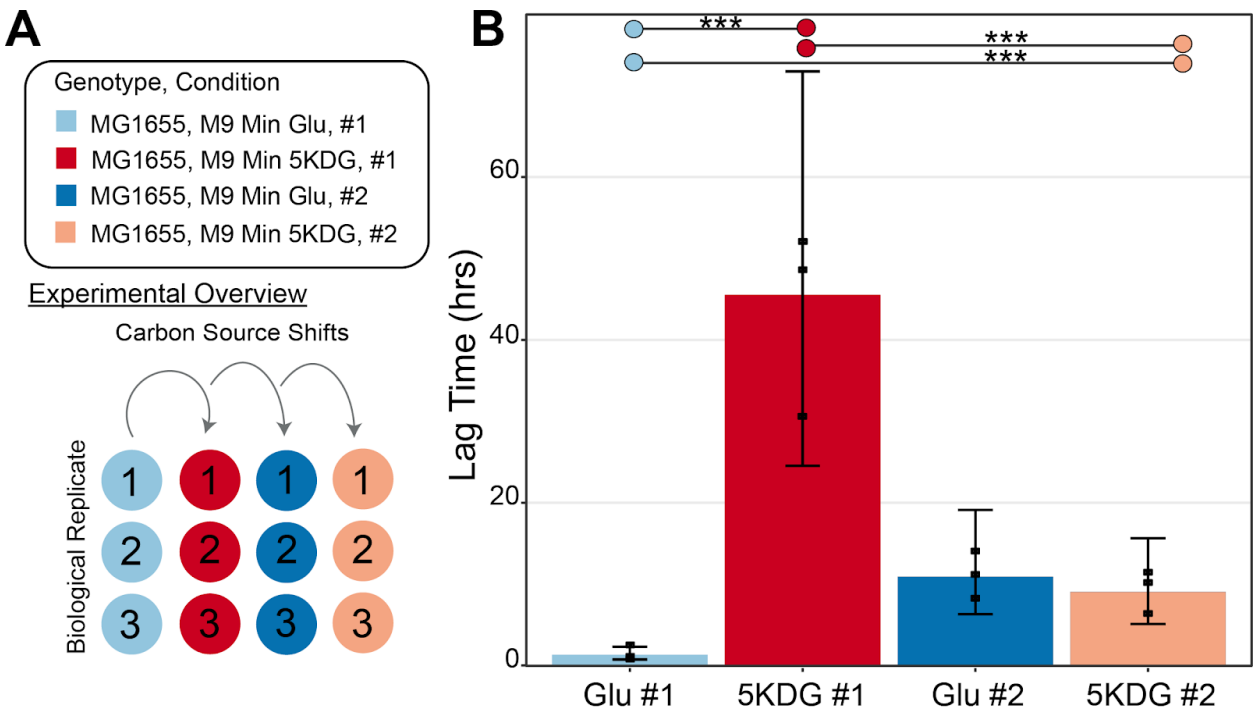

Figure 3.4: EPODs mediate transcriptional memory. (A) Experimental overview of growth curve shift experiment. At every arrow, there is a back dilution of the previous

carbon source to start the growth in the new media type. **(B)** Lag times in hours for each condition, individual values are plotted as dots. (***) is defined as where the mean posterior probability of difference >0.999, assessed using a Bayesian model (see Methods for details).

Multiple NAPs contribute to the silencing of prophages

In addition to metabolic processes, we also found that genes within EPODs were overrepresented in Gene Ontology (GO) terms associated with annotated prophages [11]. Across the *E. coli* genome, there are a number of xenogeneic elements that have been integrated and can be potentially toxic to the cell, although maintenance of these elements can also be beneficial, as they can promote resistance in the face of antibiotics [48]. H-NS is known to silence cryptic prophages, and is likely to contribute to silencing the majority of horizontally acquired DNA [48,49]. IPOD-HR successfully resolved known H-NS silenced prophages (e.g. **Fig. 3.5A**), where large reductions in protein occupancy and corresponding increases in accessibility to RNA polymerase are observed in an *hns* knockout strain. However, as noted above, not all EPODs correspond to H-NS binding, and likely not all silenced prophages correspond to H-NS repressed regions. To examine the role of other nucleoid associated proteins in silencing prophages, we calculated the mean protein occupancy across WT EPODs that overlap prophages, and how those occupancies changed upon deletion of different NAPs. Decreases in median occupancy across prophage-containing EPODs are observed upon deletion of *dps, hupAB, hfq, stpA, hns, stpa/hns, ihf,* and among both deep stationary phase samples compared to WT (**Fig. 3.5B**).

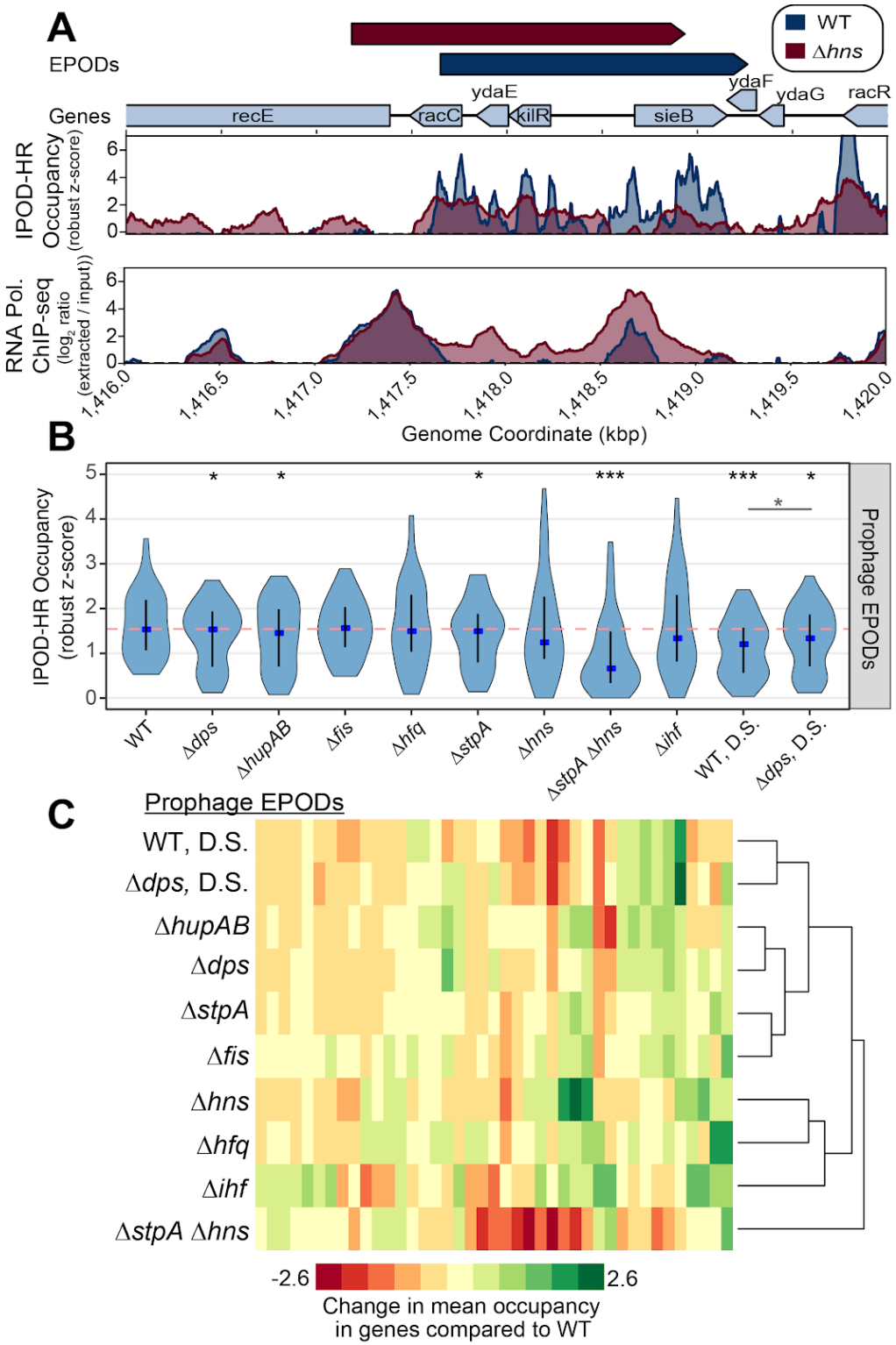

Figure 3.5: Nucleoid associated proteins contribute to protein occupancy at EPODs that contain prophages. (A) IPODHR occupancy and RNA polymerase occupancy over a known H-NS silenced prophage in WT (blue) and Δ*hns* (red) cells. The quantile normalized robust z scores of the protein occupancy at each 5 bp are represented by the IPODHR occupancy. **(B)** The mean protein occupancy (IPODHR

occupancy) was calculated over WT EPOD locations that contain prophages. The blue dots denote the median and the black line displays the interquartile ranges in each condition. The dashed pink line represents the WT median. (*) indicate the Wilcoxon Rank Sum p value comparing the change in median vs. WT for each condition that has been adjusted using the Benjamini and Hochberg method (against a null hypothesis of no difference in medians). The smaller horizontal line denotes the same comparison between the D.S. conditions. P value < 0.05 = *, <0.005=**, <0.0005=***. **(C)** Mean protein occupancy was calculated across all WT EPOD locations that contain annotated prophages. The change in mean protein occupancy compared to WT was calculated for each condition, where anything negative is a loss in occupancy compared to WT. Hierarchical clustering was performed to examine which genotypes clustered together and were more similar.

The minor loss of occupancy in some genetic backgrounds led us to investigate whether there are certain NAPs that work together to silence specific toxic elements. We examined the change in occupancy from each condition vs WT at WT EPODs that overlap prophages and performed hierarchical clustering analysis (**Fig. 3.5C**) to identify regulators that play similar roles. Interestingly, Δ*hns* and Δ*hfq* are clustered together, as are Δ*stpA* and Δ*fis.* These findings implicate Hfq, a well documented RNA chaperone[50,51] and that has only been recently explored as a protein to compress dsDNA[52,53], as a novel silencer of prophages at the level of protein occupancy across large genomic regions. Since H-NS and StpA are paralogs that bind to similar regions of the chromosome, we speculated that Hfq and Fis might play similarly complementary roles to each other in terms of silencing some prophages despite their lack of structural similarity. In addition, from our HMM analysis, both Fis and Hfq binding are enriched in HMM class 5 (**Table S3.1**), again suggesting a link between the roles and binding locations of Fis and Hfq.

Fis and Hfq are required for cell viability in a prophage dependent manner

We screened the genome for EPODs that contained prophages that lost protein occupancy upon deletion of *fis* or *hfq* individually (e.g., **Fig. 3.6A**). This also aligned with a specific type of class of EPODs associated with high Fis and Hfq binding with an example shown in Fig. 6A. RNA-seq analysis revealed that the genes within the prophage region depicted in Figure 6A were induced in both a Δ*hfq* and Δ*fis* background

(**Fig. 3.6B**). While Fis and Hfq are both NAPs, highly expressed, and bind promiscuously across the genome, they are not known to silence genes via the formation of densely occupied large-scale binding regions. Fis, while known more in the literature as a transcriptional activator [54,55], can inhibit transcription and act as a repressor in certain contexts [56,57]. Hfq is well known as a RNA chaperone[58] and can bind nucleic acids across the faces of the wheel-like homohexamer [52]. Therefore, we were not surprised to find that comparing the log fold-change of expression (relative to WT) of Δ*hfq* and Δ*fis* cells were not highly correlated among all genes (**Fig. 3.6C**). However, when examining only prophage genes, most genes were induced in one or both of the genotypes (**Fig. 3.6D**). We quantified the number of genes that were up in each genotype, both, or down using the quadrant map in **Figure 3.6E**. We applied this map and counted the rate ratios comparing all genes vs prophage genes in each quadrant. The rate ratios of all genes vs prophage genes in the quadrant which represented induced expression in both Δ*hfq* and Δ*fis* was significantly higher in prophage genes (**Fig. 3.6F**). Thus, prophage genes are specifically and significantly enriched among the set of genomic loci that are repressed by both Fis and Hfq, suggesting that Fis and Hfq bind and silence similar prophages.

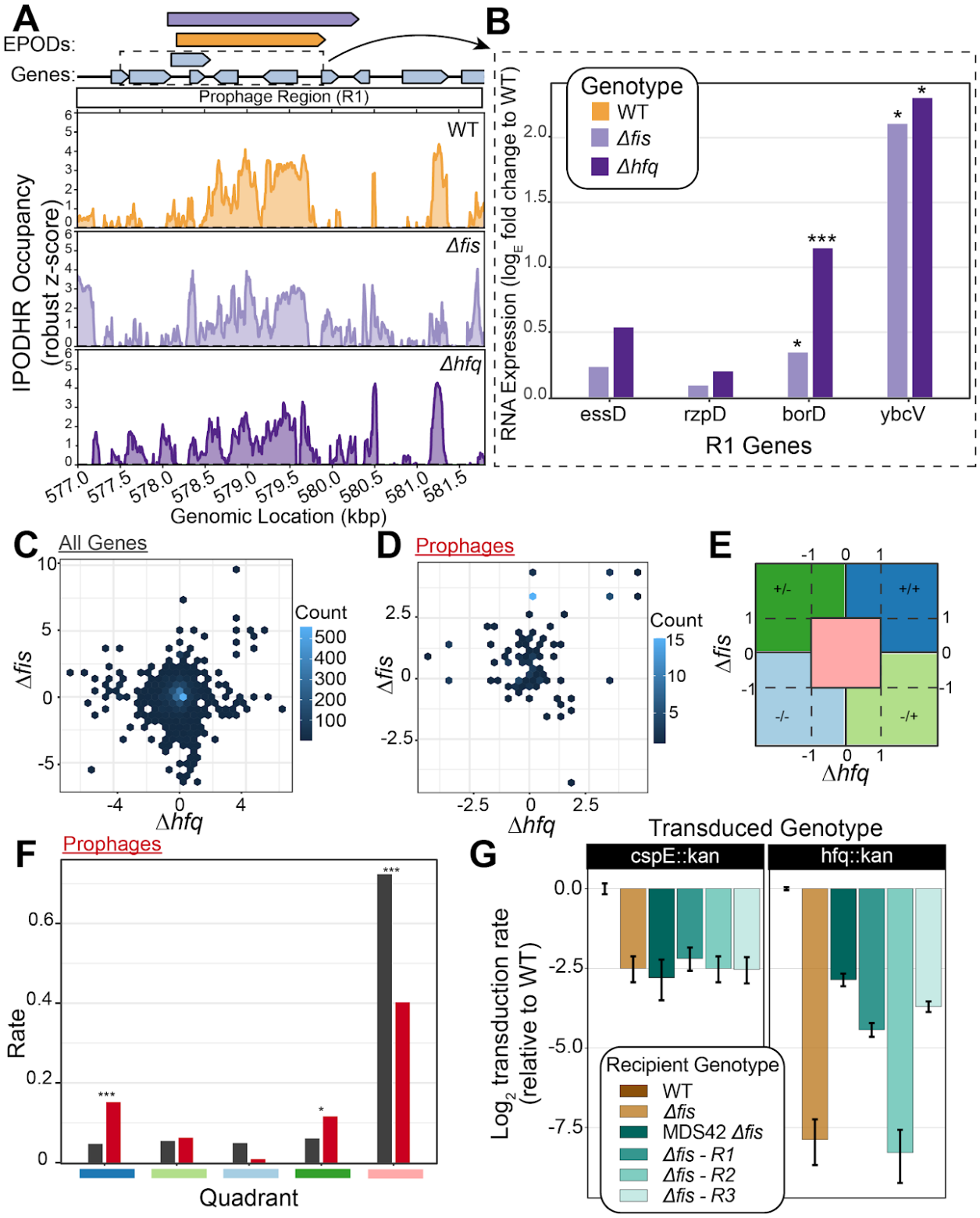

Figure 3.6: Loss of Fis and Hfq is lethal in a prophage-dependent manner. (A) Example prophage region that is annotated with the Fis- and Hfq-associated HMM class 5 in our genome-wide HMM classification. Modest loss of protein occupancy was observed at the same prophage- containing EPOD for *Δfis* (light purple-see color key in

(**B**)) and Δ*hfq* (dark purple-see color key in (**B**)) conditions compared to WT (gold). The quantile normalized robust z scores of the protein occupancy at each 5 bp are represented by the IPODHR occupancy. Prophage genes are highlighted with a red box. **(B)** RNA-seq of WT, Δ*fis,* and Δ*hfq* were performed. The log fold change compared to WT was calculated at prophage genes contained in the dashed box in (**A**). Induction of prophages across the region where loss in occupancy is observed. (*) indicate the sleuth [73] q-value; q value < 0.05 = *, <0.005=**, <0.0005=***. **(C)** The log fold change of all genes for Δ*fis* and Δ*hfq* are shown in a hexbin plot. Counts for each gene transcript contained in one bin are denoted with the counts bar. **(D)** The log fold change of all prophage genes for Δ*fis* and Δ*hfq* are shown in a hexbin plot. **(E)** Outline of quadrant map to calculate the number of genes that fall within each quadrant for (**F**). The symbols represent log fold changes compared to WT in Δ*fis* / Δ*hfq.* For instance, +/+ denotes a positive log fold change in Δ*fis* and Δ*hfq*, -/+ denotes negative log fold change in Δ*fis* and positive in Δ*hfq*. **(F)** Rate ratios of all genes (grey) and prophage genes (red) in each quadrant outline in (**E**), showing a higher rate of genes that resided in the +/+ category, indicating that many prophages are de-repressed in both Δ*fis* and Δ*hfq*. (*) indicate the p-value calculated from testing the null that the rate ratios are the same. P value < 0.05 = *, <0.005=**, <0.0005=***. **(G)** P1 vir transduction experiment to test the viability of Δ*fis* and Δ*hfq.* -Hfq indicates deleting Hfq and -CspE indicates deleting -CspE as a control. Strain identities are indicated in the box. Number of transductions were counted on LB + Kan plates. *-R1* indicates that the prophage region in (**A**) was deleted to test whether the loss of prophages silenced by Fis and Hfq restored viability of a Δ*fis* Δ*hfq* genotype. R2 and R3 were other regions in the genome that contained prophages that appeared to have Fis/Hfq dependent EPODs.

Since the expression of genes from lysogenized bacteriophages can be toxic to the cell even if they are no longer able to form replication-competent virions (particularly if a lytic operon is induced), we asked whether the combined loss of *hfq* and *fis* would more strongly impact cell physiology. We performed P1 transduction experiments in a WT, MG1655 background and a Δ*fis* background, where we attempted to delete the genomic copy of *hfq*. Interestingly, there was a dramatic loss of transduction efficiency in the Δ*fis* background, and we were not able to create the Δ*hfq* Δ*fis* mutant (**Fig 3.6G**). The transduction efficiency for *hfq::kan* dropped more than 100-fold in the Δ*fis* background, and the very small number transductants that did form colonies could not be propagated upon restreaking. To test whether the deletion of *fis* impacts transduction efficiency as a whole, we attempted the same experiment deleting *cspE*, another gene that has not been associated with prophage silencing, and did not observe a similarly dramatic loss in transductants, with simple loss of *fis* leading to only a ~4-fold loss in transduction efficiency. To compliment the experiment in **Figure 3.6G**, we attempted to complement

the combined loss of *fis* and *hfq* using a copy of *hfq* on a temperature sensitive plasmid. We thus cloned *hfq* and its native promoters on a plasmid with a temperature sensitive origin of replication [59]. The plasmid was placed into WT and Δ*fis* cells, and the genomic copy of *hfq* was deleted using P1 vir transduction (See Methods). Cells were grown in a permissive temperature (30° C), and spot titers were performed on LB and LB + chloramphenicol plates to measure CFUs of the culture and presence of the plasmid. Cultures were then shifted to 42° C, which prevents plasmid replication, and thus induces dropping of the plasmid containing *hfq*. After 8hrs of growth at 42° C, spot titers were performed to assess CFUs. We found again that the combination of Δ*hfq* Δ*fis* was not viable (**Fig. S3.6**). We hypothesized that the expression of prophages silenced by Fis and Hfq led to the inviability phenotype. To test this hypothesis, we utilized the strain MDS42 [60] , which lacks the mobile elements and prophages in the *E. coli* K12 genome. We placed the temperature sensitive plasmid containing *hfq* in MDS42 and MDS42 Δ*fis* cells and performed the same temperature shift experiment as described above. Removal of the prophages from the genome, restored viability (**Fig. S3.6**). We also observed a substantial rescue of transduction efficiency, which reached the same level in the Δ*fis* background as that of the control transduction with *cspE::kan* (**Fig. 3G**). Since 42 large regions are deleted from MDS42, we wanted to see if we could identify specific prophage regions that led to inviability. Since there is loss in occupancy in both Δ*hfq* and Δ*fis* in the Figure 6A region (R1:564815-585633; **Table S3.2**), we deleted only R1:564815-585633, and found a partial rescue of transduction efficiency (**Fig. 3.6G**). We found another region that met the same criteria (R3; **Table S3.2**), and also saw rescuing effects (**Fig. 3.6G**). Another region that contained prophages, but did not dip in occupancy in both genotypes did not impact viability (R2; **Table S3.2**) (**Fig. 3.6G**). Thus, we were able to define regions that contribute to viability, and hone in on specific prophages that are silenced by Fis and Hfq and, in the absence of repression by those two NAPs, prevent cell growth.; in particular, loss of either of two prophages (R1 or R3) was sufficient to restore the viability of a *fis*/*hfq* double mutant. This novel interaction defines a new role for NAPs in regulating the expression of prophages, implicating *E. coli* NAPs more broadly in the establishment of defense mechanisms against horizontally acquired DNA.

Heterochromatin domains silence horizontally acquired DNA across diverse species

Because it relies only on elementary physico-chemical principles rather than specific affinity reagents, IPOD-HR is an approach that could be implemented in a wide variety of bacterial species. To further our understanding of conserved features that regulate bacterial genome architecture, we investigated whether a distantly related bacterial species contained EPODs. We performed IPOD-HR on the Gram-positive Firmicute *Bacillus subtilis (B. subtilis)*- a soil dwelling bacterium that has the ability to enter a number of developmental platforms upon nutrient deprivation or other environmental stressors, including the formation of desiccation resistant endospores, biofilm formation, genetic competence, and swimming/swarming motility phenotypes [61]. We performed IPOD-HR in *B. subtilis* strain PY79 and found multi-kb regions of protein occupancy (EPODs) spanning genes that function in a number of metabolic pathways, suggesting a feature conserved with *E. coli* (**Fig. 3.7A**). Many of these pathways are activated in times of nutrient limitation and stress, similarly to the silenced pathways we observe in *E. coli*. As regions of protein occupancy were observed in horizontally acquired DNA in *E. coli,* we proposed that regions of protein occupancy may play a role in horizontally acquired DNA in *B.subtilis.* As *B. subtilis* is naturally competent, the tight regulation of competence development is especially important to regulate and protect against the acquisition of harmful exogenous DNA elements. Surprisingly, we found that many large negative occupancy peaks overlapped annotated prophage genes (**Fig. 3.7B**), appearing similar to EPODs but inverted in sign.

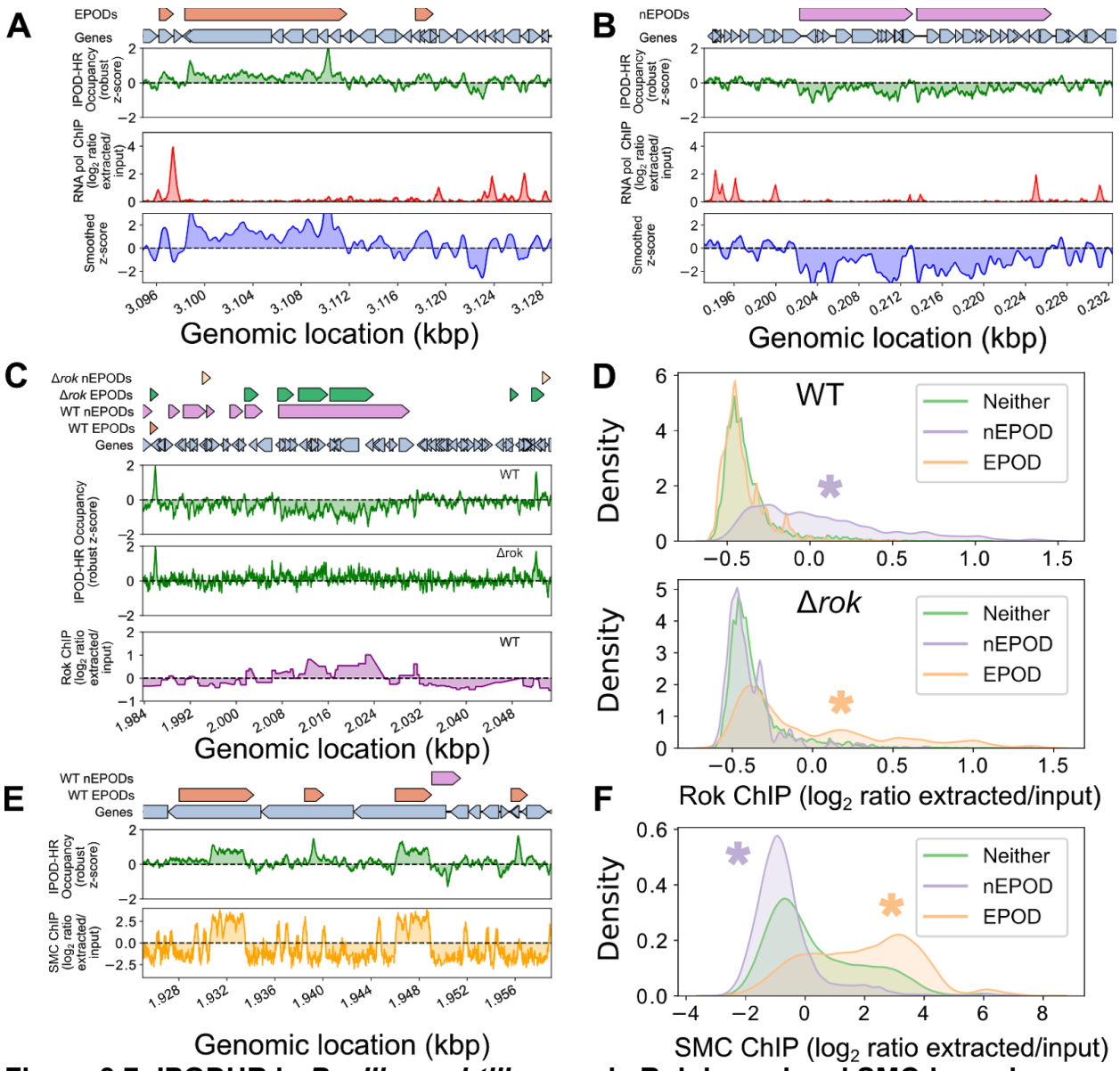

Figure 3.7: IPODHR in *Bacillus subtilis* reveals Rok-bound and SMC-bound domains. (A) IPOD, RNA polymerase ChIP, and IPOD-HR ChIP-subtracted z-scores in the vicinity of a typical extended protein occupancy domain (EPOD). IPOD and ChIP tracks are shown as log2 extracted/input ratios; the z-score is the ChIP-subtracted robust z-score smoothed with a 512 bp rolling median. **(B)** IPOD, RNA polymerase ChIP, and IPOD-HR tracks in the vicinity of a negative EPOD (nEPOD). **(C)** Effects of deletion of *rok* in the vicinity of an nEPOD; Rok ChIP data from[63] shows a strong overlap with the nEPOD boundary, whereas that occupancy region is lost in Δ*rok* cells. **(D)** Distributions of Rok ChIP occupancies (see Methods) in the EPODs and nEPODs called in WT (top) or Δ*rok* (bottom) cells; note that the Rok ChIP occupancy was taken only in WT cells. (*) indicates a significant difference from the 'neither' distribution (that is, genomic sites that are not in an EPOD or nEPOD); $p<0.05$, permutation test. **(E)** Comparison of IPOD occupancy and SMC ChIP occupancy (see Methods) in the vicinity of several typical EPODs. **(F)** Genome-wide distributions of SMC binding in EPODs vs.

nEPODs as assessed in WT cells; (*) indicates a significant difference from the background distribution as in panel **D**.

One of the main regulators of competence in *B. subtilis* is Rok, which acts as a direct repressor of competency genes, regulates several secreted proteins, and is involved in repression of mobile genetic elements [61–63]. Due to the impact of Rok on gene regulation, coupled with the promiscuous binding activity to A+T-rich DNA, we propose that Rok may be a main component of protein occupancy in *B. subtilis.* Surprisingly, while we did not see an enrichment of Rok binding in EPODs using available *rok-myc* ChIP-chip data [63], Rok binding was highly correlated with the negative occupancy peaks (**Fig. 3.7C,D**). To investigate this further, we performed IPOD-HR in Δ*rok* cells and found that indeed, the loss of Rok resulted in an increase in RNA polymerase binding at sites correlated with Rok binding, and loss of negative peaks found in the WT condition (**Fig. 3.7C**). We subsequently performed our analysis pipeline to incorporate negative occupancy and found that these negative EPODs correlated with Rok binding and were enriched for genes known to be regulated by Rok, such as sporulation genes and genes involved in competence activation.

Positive (standard) EPODs were also apparent in the *B. subtilis* data, and we found them to be highly correlated with SMC binding (**Fig. 3.7E,F**), which is known to compact the genome in preparation for chromosomal segregation [64–66]. Our findings align with known datasets [67], and further our understanding of the role of nucleoid binding proteins in defining the genome landscape across species. Without the dependency on antibody-based methods, we can explore unknown protein functions across a variety of developmental platforms and species. Importantly, we did not observe correlations with any suspected nucleoid associated proteins or global regulators in negative peaks in *E. coli*. These findings highlight the broad utility of IPOD-HR, with the ability to detect both conserved and novel genome architecture features in a variety of distantly related bacterial species. In addition, we see that the general pattern of large regions of high protein occupancy apparently silencing genes horizontally acquired DNA is an

extremely widespread feature, occurring in at least two bacterial species separated by more than 2 billion years of evolution.

Discussion

There is increasing evidence of a regulated genome architecture in *E. coli*, both in terms of its three dimensional structure [24,68–70] and in terms of the landscape of protein occupancy on the genome. Both of these classes of features are largely supported by binding of NAPs [11]. IPOD-HR enabled us to study global changes in chromosomal architecture - here defined as highly protein occupied regions of the genome - across species and implicated NAPs as the main component of EPODs. Here, we show the robustness of these large protein domains across media conditions, growth phase, and small genotype differences (**Fig. 3.1**). The maintenance of EPODs across conditions and ancestral strains aligns with the idea that EPODs serve an important regulatory silencing role. A variety of questions emerge: How are EPODs maintained through replication and growth? What recruits proteins to these regions? Further studies are being performed to examine the role of methylation in maintenance and recruitment of protein to EPOD regions. As horizontally acquired DNA and methylation have been shown to be intertwined in *E. coli* [71], and we found that *dam* and *dcm* sites are depleted in EPODs, we believe that methylation plays a role in regulation of EPODs containing prophages.

Our study shows that EPODs are partly composed by NAPs in *E. coli* (**Fig. 3.2**), with the largest contribution clearly made by the major transcriptional silencers H-NS and StpA, but other pairs of NAPs making important contributions at a subset of loci. Due to the wide binding capacity of NAPs across the genome, the question regarding recruitment to EPODs emerges again. IPOD-HR successfully shows losses in occupancy upon deletion of NAPs, however, there may be accessory proteins that facilitate recruitment and maintenance. IPOD-HR may miss subtle changes in proteins that are not as

abundant in the cell, so we have begun to design proteomic analysis of EPODs to define the exact composition of EPODs.

We are able to define the key proteins involved in EPOD regulation of metabolic pathways (**Fig. 3.3,3.4**) and silencing of prophages across the genome (**Fig. 3.5,3.6**). EPODs appear to mediate the formation of transcriptional memory (**Fig. 3.4**), which allows for strong repression of a rarely-used metabolic operon when it has not been transcribed in recent memory, but aids in faster induction of genes important for metabolic response after a single exposure to a relevant nutrient. This type of regulation poses exciting ideas for understanding how architectural proteins facilitate a genome architecture regulation across the genome. Importantly, understanding bacterial genome regulation can be incredibly useful for biotechnology purposes, especially in the case where cells can be grown in a number of conditions that may induce changes in their overall genome architecture that can impact induction of particular genes.

We have also identified here a novel silencing mechanism for prophages and toxic elements across the genome. Together, Hfq and Fis are required to silence some prophages, most notably DLP12 (contained in R1 of **Fig. 3.6**) and Qin prophages (R3 of **Fig. 3.6**). What defines particular prophages to recruit Hfq and Fis remain to be explored. However, these findings contribute to an overarching theme of the genome structure serving as an immune response to a variety of horizontally acquired DNA. As we have previously shown, reports integrate with higher frequency in EPODs and are efficiently silenced[11,21]. We propose that EPODs serve as DNA sinks for foreign DNA, and quickly silence potentially harmful elements. Further investigation into the mechanisms underlying this response will bolster our view of immune responses of bacterial species.

Many NAPs and their functions are conserved across bacterial species, and the overall roles of NAPs in establishing large regions of silencing protein occupancy appear to be conserved even beyond the reach of recognizable homologs of any given NAP. We found that the use of genome architecture as a mode of immunity may also be

conserved in a distantly related species to *E. coli*, *B. subtilis* (**Fig. 3.7**). The exploitation of such a system has a number of promising outcomes. For instance, work established here could inform new antibiotic approaches for pathogenic bacteria by targeting proteins required to suppress toxic elements already in existence in the genome. In addition, understanding how bacteria recruit and build their genome architecture around foreign DNA can inform us of how bacteria interact with the environment. The manipulation of this process may allow us to utilize bacteria in innovative ways, such as novel biosensors or protein engineering.

Materials and Methods

Strain construction

The MG1655 “WT” strain used in all figures was obtained from Hani Goodarzi (Tavazoie Lab, then at Princeton University) in 2009, and is isogenic with ATCC 700926, with the exception of a 9 bp insertion in *dcgJ* [74]. The MDS42 strain was obtained from Alison Hottes in 2009 (Tavazoie Lab, then at Princeton University) and contains a C->T mutation in *ribD* and missing coverage for *lysV*[60].These modifications may be ancestral to MDS42, or specifically present in our MDS42 parental strain. MDS42 deletions were validated using PCR. The MG1655 (2) strain was obtained from the Jakob Lab at the University of Michigan in 2018, and contains a G -> A mutation in *mntP* and C->A mutation in *ybhJ*.

NAP deletion strains:

All nucleoid associated protein (NAP) gene deletions were performed in the same MG1655 “WT” base strain stock discussed above. All NAP deletions were obtained by P1 transduction [75] of the FRT-flanked *kanR* marker from the corresponding knockout strain of the Keio collection[75,76]. With the exception of Δ*ihfA:: Km* Δ*ihfB::Clm*(*ΔihfAB*) and Δ*hupA* Δ*hupB::Km (*Δ*hupAB*),the pCP20 plasmid [77] containing Flp recombinase was used to excise the *kanR*marker, leaving a small scar in the place of the original

open reading frame. Once candidates were isolated for each deletion and contained the pCP20 plasmid, they were grown overnight at 42°C to drop the temperature sensitive pCP20. The overnight cultures were streaked onto LB plates and grown overnight at 37°C. Individual colonies were replica plated onto appropriate selective plates to ensure the loss of both the marker and pCP20 plasmid. The *ΔihfAB* and *ΔhupAB* strains were not cured due to incredibly low efficiency to excise markers via pCP20, and markers were retained to avoid potential suppressor mutations.

Bacillus subtilis strains

The *B. subtilis* PY79 and rok::kan strains came from the Simmons lab at the University of Michigan. Details of the genome sequence can be found in [78] .

Media/culture conditions

LB (Lennox) media (10g/L tryptone, 5g/L yeast extract, 5g/L NaCl) was used for cloning and recovery of cryogenically preserved cells, with addition of 15g/L bacteriological agar for plating.

In the case of physiological experiments, we used appropriately supplemented versions of M9 defined minimal medium (6 g/L Na_2HPO_4, 3 g/L KH_2PO_4, 1 g/L NH_4Cl, 0.5 g/L NaCl, 1 mM $MgSO_4$). Minimal M9 medium (M9/min) contained includes 0.2% (w/v) carbon source (glucose, sodium acetate, glutamine or 5-Keto-D-gluconic acid potassium salt), 0.4 mM CaCl2, 40 μM ferric citrate, and the micronutrient mixture typically incorporated in MOPS minimal media[27]. For all IPOD-HR experiments, we used our M9 rich defined medium (M9/rdm) incorporated with 0.4% (w/v) glucose, MOPS micronutrients, 4 μM CaCl2, 40 μM ferric citrate, and 1x supplements ACGU and EZ as used in MOPS rich defined medium [27].

For *Bacillus subtilis* strains were struck from frozen stocks and grown on LB plates overnight at 37°C. WT and Δ*rok* strains were inoculated into LB and LB supplemented with 5 μg/mL kanamycin, respectively, from a plate wash at a starting an OD_{600} of 0.025

and grown at 37°C with shaking to an OD_{600} between 0.65 and 0.85. Rifampin was added to a final concentration of 150 µg/mL and cultures were incubated for an additional 10 minutes at 37°C with shaking. Sodium phosphate (final concentration 0.01M) and formaldehyde (1% v/v) were added to 30 mL aliquots of culture and cross-linked at room temperature for 5 minutes with shaking. Reactions were quenched by the addition of 0.333M glycine for 10 minutes at room temperature. Cells were collected via centrifugation and watched twice with ice-cold PBS and cell pellets were subsequently flash frozen in liquid nitrogen and stored at -80°C.

Cell growth and harvest for IPOD-HR

Cryogenically preserved cells were streaked onto an LB plate and grown in the media of interest with 1/10th of the carbon source indicated overnight at a temperature of 37°C and shaking at 200 rpm. The culture was back-diluted into fresh, prewarmed media to an OD600 of 0.003 the next day. The culture was grown to the target OD600 (0.2, except in the case of deep stationary phase samples, described below) and treated with a final concentration of 150 µg/mL of rifampin and incubated for 10 minutes under the same growth conditions previously described. The cultures were rapidly poured into falcon tube and mixed with concentrated formaldehyde/sodium phosphate (pH 7.4) buffer sufficient to yield a final concentration of 10 mM NaPO4 and 1% v/v formaldehyde. Crosslinking proceeded for 5 minutes at room temperature, and quenched with an excess of glycine (final concentration 0.333 M) for 10 minutes with shaking at room temperature. The crosslinked cells were chilled on ice for 10 min, and washed twice with 10mL ice cold phosphate buffered saline (PBS). The resulting pellets were carefully dried, remaining media pipetted and discarded, and snap-frozen in a dry ice-ethanol bath and stored at -80°C for no longer than 1 month.

Deep stationary samples followed the same process of being grown in the appropriate media overnight and back diluted to an OD600 0.003. Once the cells reached an OD600 of 0.2, they were grown for an additional 24 hours, treated with rifampin for 20min, and proceeded through the same treatment for crosslinking as described above.

Cell lysis and DNA preparation, IPOD-HR interface extraction, RNA polymerase chromatin immunoprecipitation, and crosslinking reversal and recovery of DNA was performed as previously described [11,27]. Due to the high biomass of deep stationary phase cells, cells were diluted 10x prior to lysis. In the case of the *B. subtilis*, samples were sonicated 4 times for 5s at 25% power with 15s between pulses.

Preparation of next-generation sequencing (NGS) libraries

All DNA samples were prepared for Illumina sequencing using the NEBNext Ultra (or II) Library Prep Kit (NEB product #E7370 or #E7103, respectively). The NEBNext Ultra II Library Prep Kit was used on the *Δhns ΔstpA* and biological replicate 2 of the deep stationary phase samples (*Δdps* and corresponding WT). We consulted with NEB to confirm that there are no differences between the kits that would impact our results. Single index or dual index primers from NEB were used in the prep. The manufacturer's instructions were followed with the same modifications as listed in [11].

All libraries were sequenced on an Illumina NextSeq.

Analysis of NGS data, read quality control and preprocessing, DNA sequencing and protein occupancy calling, and feature calling was performed as previously described[11]. Rescalling was performed on IPOD-HR occupancy analysis described in supplementary text.

5KDG growth experiments

Experimental design for Figure 3:

Cells were grown from a cryogenic stock on LB plates at 37°C, and inoculated into our M9 minimal medium including 0.02% glucose at 37°C. In the morning, cultures were back diluted to an OD600 0.003 in fresh, pre-warmed M9 minimal medium including 0.2% glucose at 37°C. Once cells reached a target OD600 0.2, cells were pelleted and washed twice with 5 mL of warmed PBS. 1 mL of sample was mixed with 1mL DNA /

RNA shield and flash frozen in a dry ice-ethanol bath and stored at -80°C. The remainder of the cells were placed in our M9 minimal medium including 0.2% carbon source (glucose, sodium acetate, or 5-Keto-D-gluconic acid (5-KDG; Sigma Aldrich: Catalog #K4125)) to an estimated OD600 of 0.1 and placed at 37°C. 1 mL samples were taken after 10 minutes, 2 hours, 4 hours, 8 hours, 12 hours, 24 hours, and 28 hours and mixed with 1mL DNA/RNA shield and flash frozen in a dry ice-ethanol bath and stored at -80°C.

Experimental design for Figure 4:

Similar as above, cells were grown from a cryogenic stock on LB plates at 37°C, and inoculated into our M9 minimal medium including 0.02% glucose at 37°C. In the morning, cultures were back diluted to an OD600 0.003 in fresh, pre-warmed M9 minimal medium including 0.2% glucose at 37°C in a total volume 150uL with 100uL of mineral oil in a plate reader. Measurements were taken every 10min at 37C with shaking to calculate lag times. Once cells reached an OD600 ~0.2, they were back diluted to an OD600 of 0.01 in M9 minimal medium with 0.2% 5KDG. Cells were grown to an OD600 ~0.2, then diluted to an OD600 0.003 in M9 minimal medium with 0.2% glucose. The cycle was repeated, where cells were diluted to an OD600 of 0.01 in M9 minimal medium with 0.2% 5KDG after reaching 0.2. The last shift cells were back diluted to an OD600 of 0.003 in M9 minimal medium with 0.2% glucose.

Growth curves (using $\log_2$-scaled optical densities [ODs]) were smoothed using a cubic spline with one knot per five hours (or fraction thereof) in the data; we then identified the maximum value of the slope of the resulting spline as the growth rate. The lag time was calculated by projecting a line with a slope equal to the growth rate through the point at which the most rapid growth was observed, and took the time at which that line intercepted a horizontal line at the initially observed $\log_2$ OD for that culture. Summary statistics were calculated via Bayesian regression as implemented in the brms R package, with population-level effects for the experiment under consideration and default brm values for all other arguments.

RNA isolation and sequencing preparation

All RNA-sequencing samples were collected and prepared for sequencing as follows. Once cells were grown to the appropriate OD600, 2.5mL of culture was mixed with 5mL of RNAprotect (Qiagen: Catalog #76506), vortexed and incubated at room temperature for 5 min. Cells were spun at 4°C for 10 min at 5,000 x *g* in a fixed-angle rotor. The supernatant was removed, and the pellet was flash frozen in a dry ice-ethanol bath and stored at -80°C. For RNA extraction, the pellet was resuspended in 100uL of TE and treated with 177kU (1uL) Ready-lyse lysozyme solution (Lucigen: Catalog #R1804) and 0.2 mg (10uL) proteinase K (Thermo Fisher Scientific: Catalog #EO0492), incubated for 10 min at room temperature with vortexing every 2 min. The RNA was purified using RNA Clean and Concentrator kit-5 (Zymo: Catalog #R1014), treated with 5 units of Baseline-ZERO DNase (Epicentre: Catalog #DB0715K) in the presence of RNAse inhibitor (NEB: Catalog #M0314L) for 30 min at 37°C. RNA was purified again using RNA Clean and Concentrator kit-5 (Zymo: Catalog #R1014). RNA was flash frozen in a dry ice-ethanol bath and stored at -80°C.

rRNA depletion was performed using the bacterial rRNA depletion kit following manufacturer instructions (New England Biolabs (NEB): Catalog #E7850L). The only modification performed was the last step, where instead of a bead clean up, we used the RNA Clean and Concentrator kit-5 (Zymo: Catalog #R1014).

Sequencing preparation was performed using the NEBNext Ultra Directional RNA Library Prep Kit for Illumina following manufacturer instructions (NEB: Catalog #E7420L) for rRNA depleted RNA. Random primers were used and RNA was considered intact for the NEBNext Ultra protocol. Slight modifications to the protocol are as follows. To purify cDNA, the Oligo clean & concentrator was used (Zymo: Catalog #D4061). Following adapter ligation, DNA was purified using DNA Clean & Concentrator-5 (Zymo: Catalog #D4014). Dual index primers from NEB were used in the prep. Libraries were sequenced on an Illumina NextSeq.

RNA-seq analysis began with read preprocessing identical to that described in [11]. Expression quantitation and significance calling for the small-scale dataset considered in **Figure 3.3** was performed using Rockhopper [72] with default parameters. For the more complex set of comparisons between different NAP deletions, we instead used kallisto [79] for read quantitation and sleuth [73] for differential expression calling.

HMM classes

HMM fits were performed using the hmmlearn python package (version 0.2.4) with Gaussian emissions. As input features we used the IPOD-HR robust z scores and RNA polymerase log2(extracted/input) ratios, for a total of 2 features per condition at each of 928,330 sites on the genome (5 bp resolution). We trained a series of HMMs using 20-fold cross validation (dividing the genome into 20 evenly sized blocks), in which we assessed the log-likelihoods for the withheld folds based on an HMM trained on the rest of the genome. After training and evaluating models from 2 to 20 components, we found that the predictive performance increased sharply with component count up to 6 components, and after that increased much more gradually. We thus used a six-component model to provide a balance of interpretability and predictive performance. We fitted 20 final models using the entire genome and selected the one with the highest likelihood to provide the final HMM; state assignments were then obtained using the Viterbi algorithm. Default parameters for hmmlearn were used unless otherwise noted.

Hfq binding was measured by cloning Hfq-PAmCherry from [53] into MG1655 (2) and performing ChIP-seq with a monoclonal mCherry antibody (Thermo Fisher M11217) (*manuscript in preparation*). A 500 bp rolling mean of the $\log_2$ extracted/input ratio was calculated and used for comparison in this paper.

Data visualization and analysis

For data analysis as previously described [11], we made heavy use of numpy [11,80], R version 3.6.3 [81,82], tidyverse [83], and ggplot2 [84].

Data Availability

The raw and processed sequencing data used in this study have been deposited in the Gene Expression Omnibus with study accession GSE164796. Reviewer access is available using the token ozufkmmwddqhrix

Acknowledgments

This work was supported by NIH grants R00-GM097033 and R35-GM128637 (to PLF) and NSF grant MCB 1714539 (to LS).

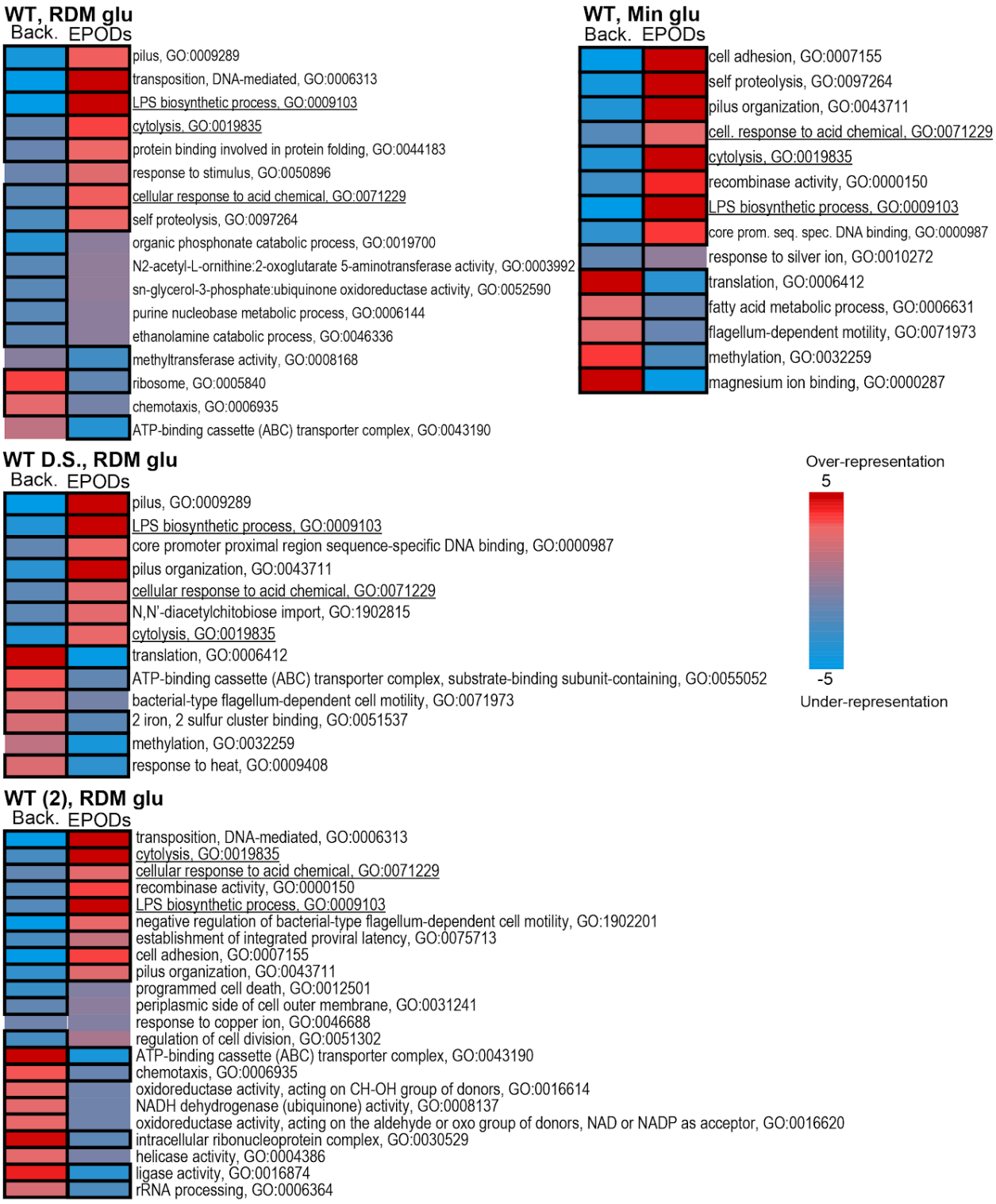

Figure S3.1: Pathway analysis of EPODs across WT conditions. iPAGE analysis revealed key pathways overrepresented in EPODs compared to background that remain across different growth media, harvest growth phase, and parental background

(underlined). Color scale represents over- or under-representation of genes with particular GO term annotations in the EPODs vs. non-EPOD regions (background).

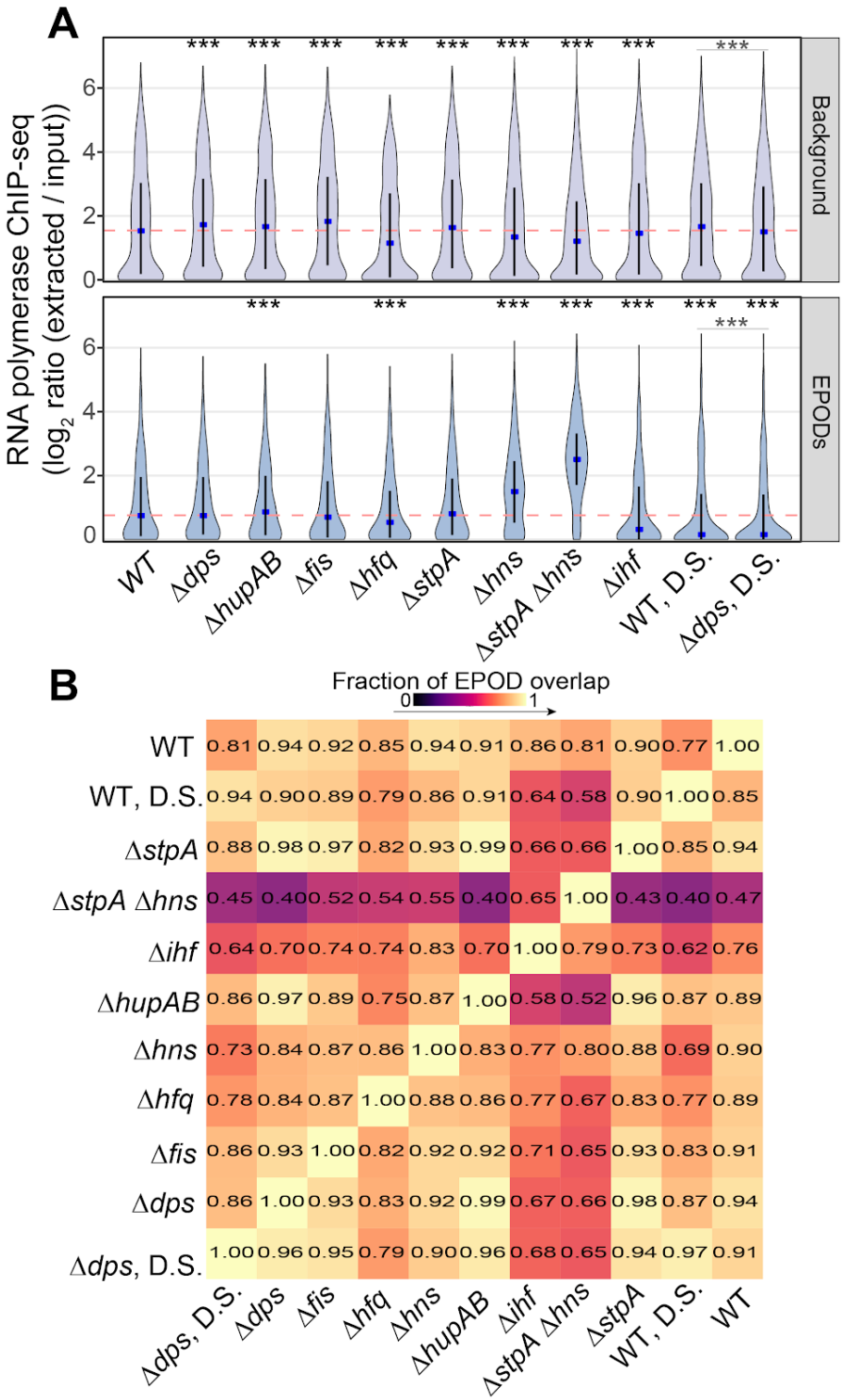

Figure S3.2: Loss of NAPs results in increases in RNA polymerase occupancy and decreases in overlapping EPODs. (A) Average RNA polymerase occupancy was calculated across intergenic regions within WT EPODs and background. Similar to

Figure 2D, The blue dots denote the median and the black line displays the interquartile ranges in each condition. The dashed pink line represents the WT median. (*) indicate the Wilcoxon Rank Sum p value comparing the change in median vs. WT for each condition that has been adjusted using the Benjamini and Hochberg method (against a null hypothesis of no difference in medians). The grey line denotes the same comparison between the D.S. conditions. P value < 0.05 = *, <0.005=**, <0.0005=***. **(B)** Reading left to right: overlap of the relaxed set of EPOD calls (left) over the stringent set of EPOD calls (bottom).

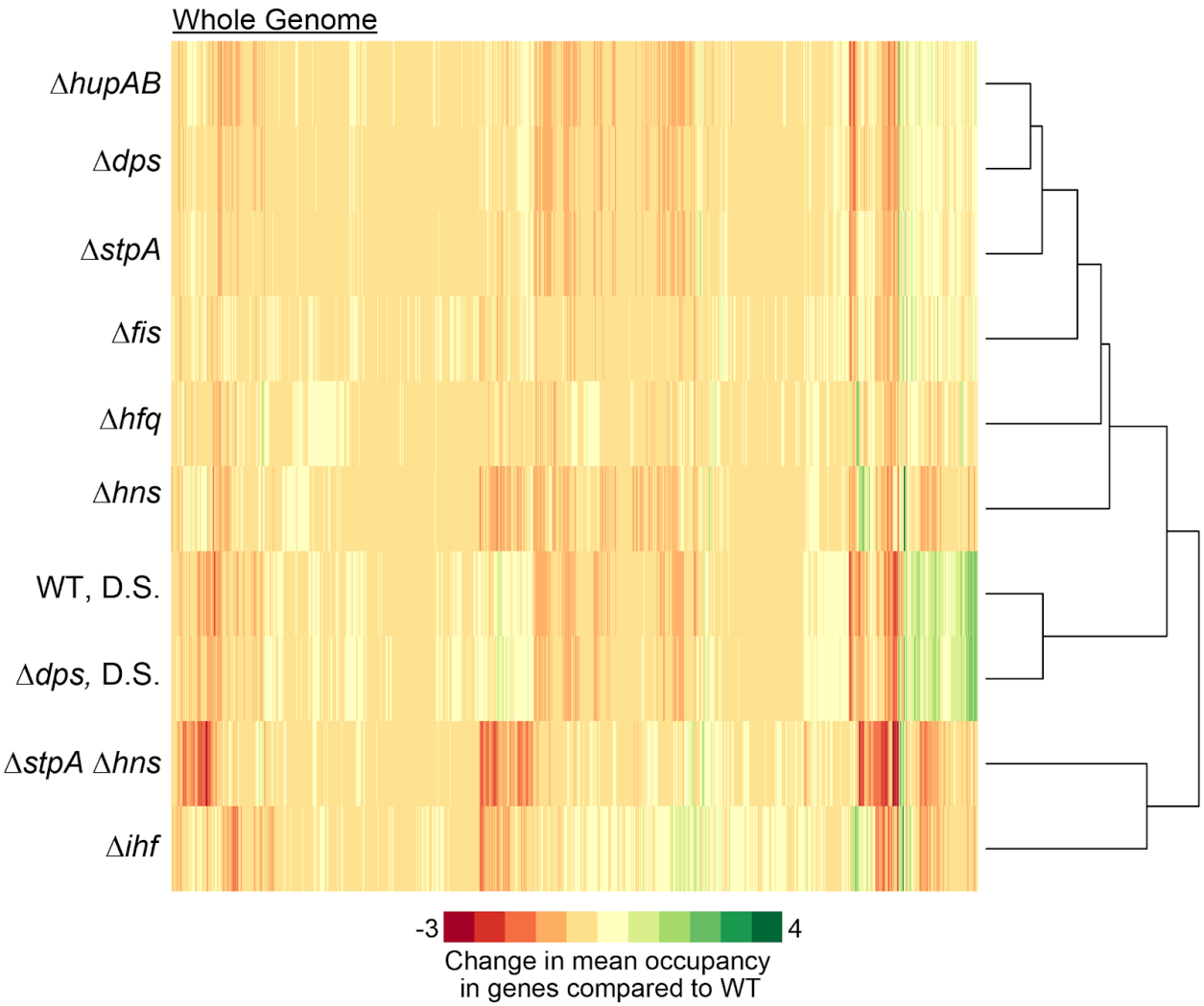

Figure S3.3: Changes in protein occupancy across the genome. The average occupancy was calculated across EPODs and background regions. The change in protein occupancy was calculated by subtracting the WT average at each region for every mutant. A gain in occupancy in the mutant is represented by a positive change in occupancy, while a loss is represented by a negative change in occupancy. Hierarchical clustering distinguished NAPs that have similar impacts on protein occupancy across the genome.

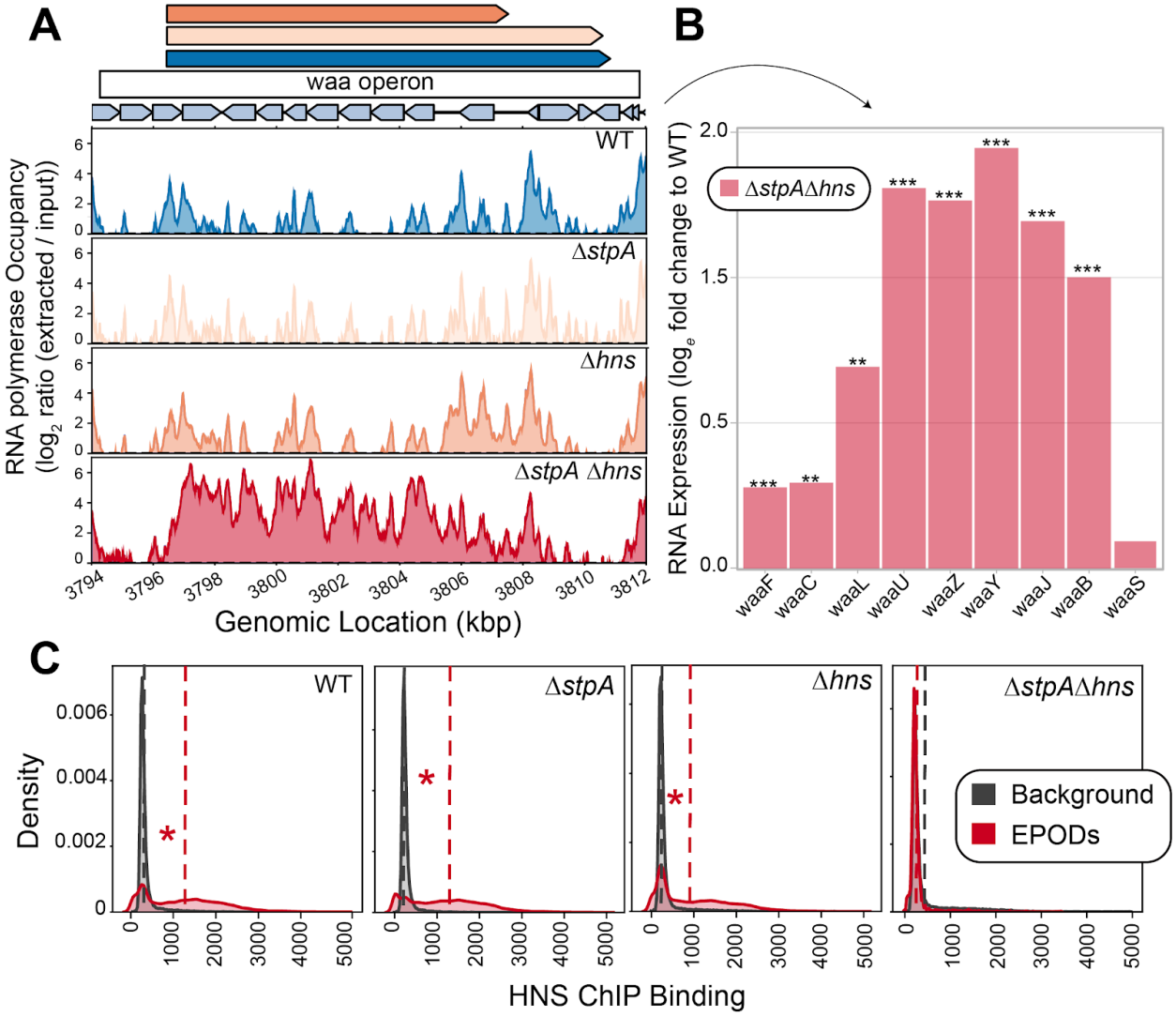

Figure S3.4: Deletion of *hns* and *stpA* impact EPODs across the genome. (A) The loss in protein occupancy shown in Figure 2E leads to increases in RNA polymerase occupancy across the *waa* operon. **(B)** RNA-seq analysis shows log fold change compared to WT of waa operon expression upon deletion of *hns* and *stpA*. (*) indicate sleuth [73] q-value; q value < 0.05 = *, <0.005=**, <0.0005=***. **(C)** Density plots exhibit enrichment of H-NS binding within EPODs that is reduced upon deletion of *hns* and the double deletion of *stpA* and *hns*. Dashed lines are the median for background (grey) and EPODs (red) for each condition. (*) indicates FDR-corrected p<0.005 via permutation test (against a null hypothesis of no difference in medians).

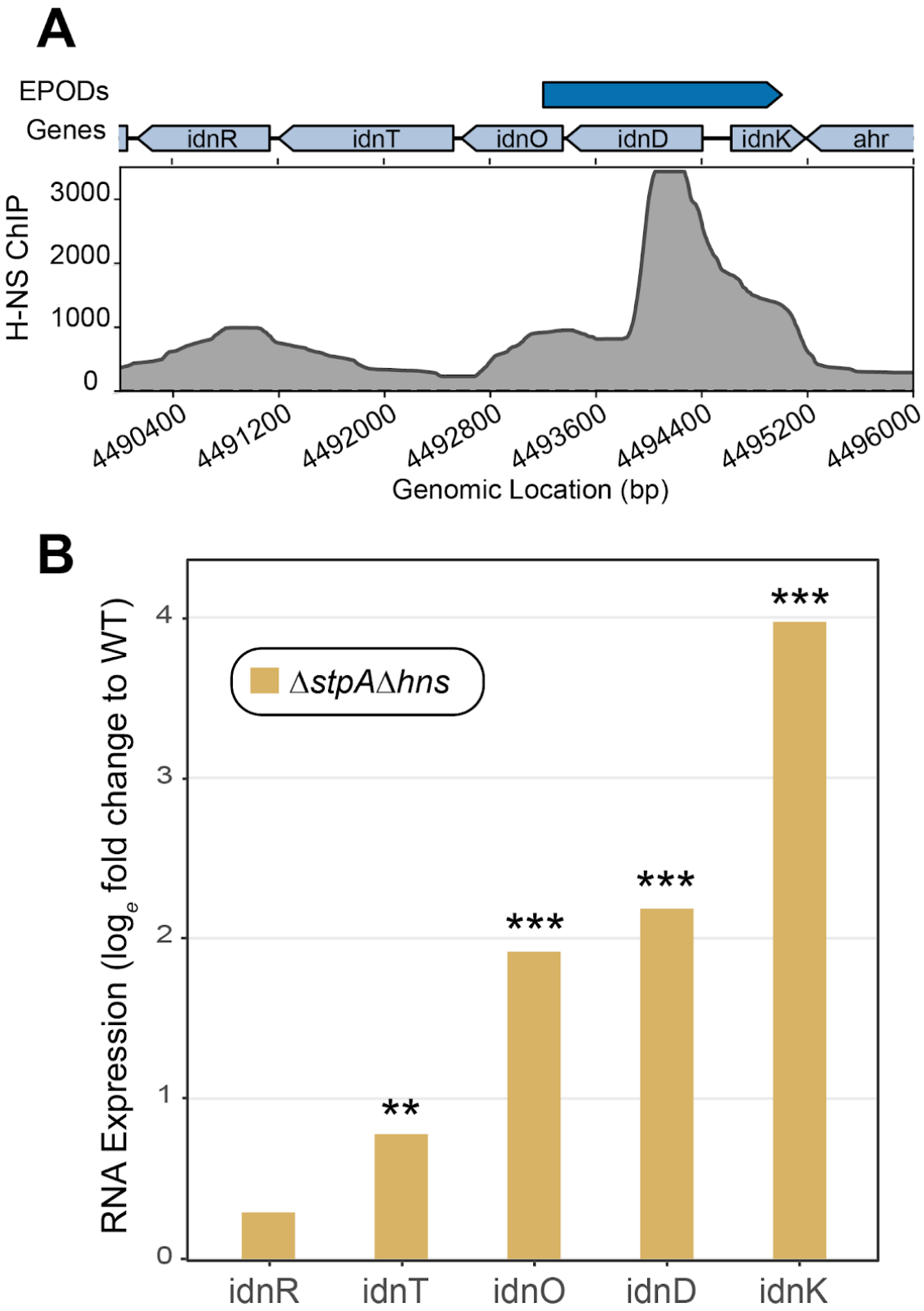

Figure S3.5: H-NS and StpA mediate silencing of the idn operon. (A) The 500bp normalized average of previously published H-NS ChIP-seq [17] exhibits high H-NS binding on the idnD promoter region. **(B)** RNA-seq analysis shows log fold change

compared to WT of idn operon expression upon deletion of *hns* and *stpA*. (*) indicate sleuth q-value [73]; Q value < 0.05 = *, <0.005=**, <0.0005=***.

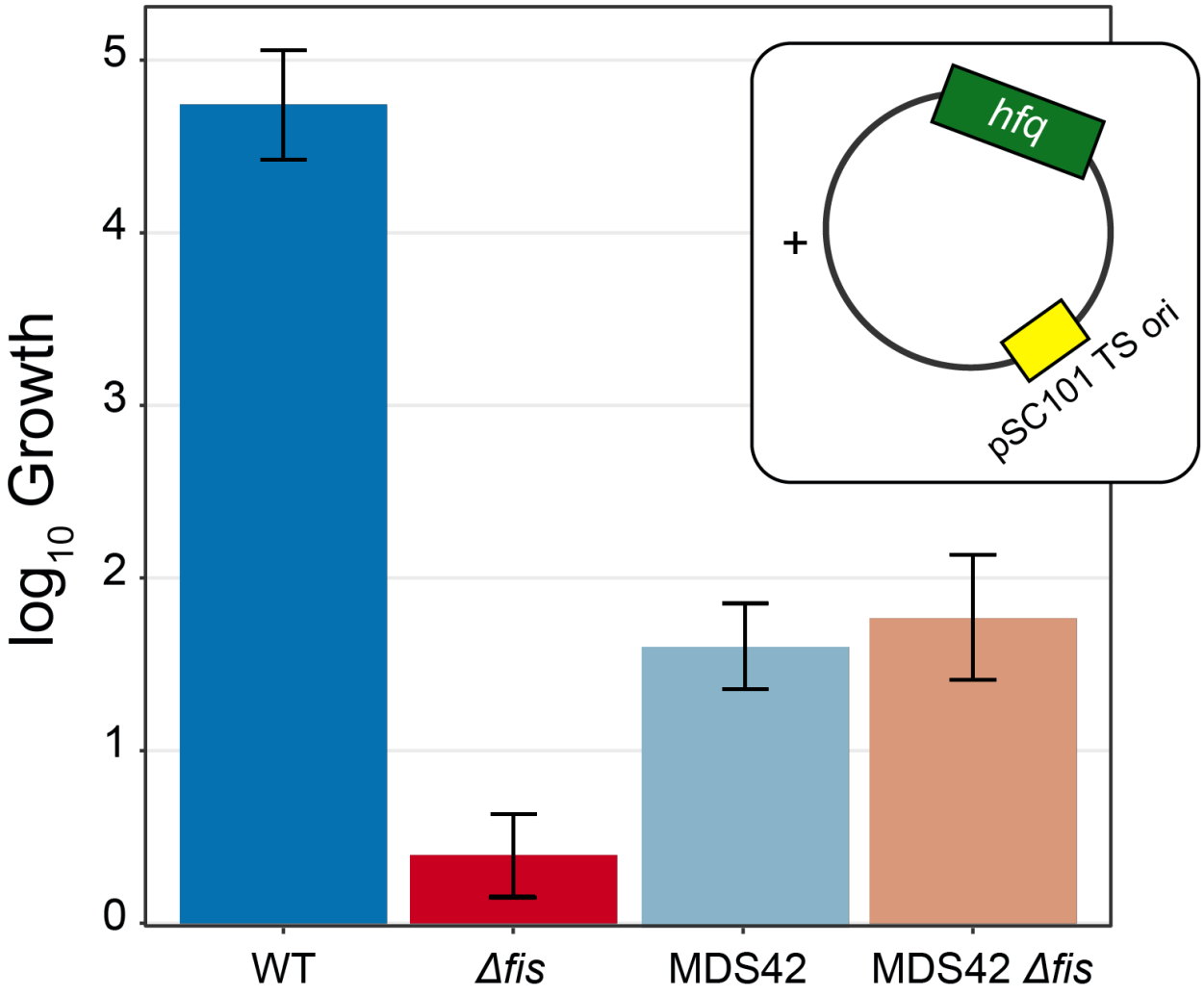

Figure S3.6: Growth deficiency of *Δfis Δhfq* cells. WT, MDS42, *Δfis,* and MDS42 *Δfis* cells containing a temperature sensitive plasmid with *hfq* had their genomic copy of *hfq* deleted. Cells were grown in a permissible temperature (30°C), and then shifted to a non-permissible temperature for plasmid replication (42°C), thus removing *hfq* as the plasmid is dropped. The $\log_{10}$ fold change in CFU is displayed. *Δfis* cells were unable to grow with the loss of the temperature sensitive *hfq* plasmid, however, upon deletion of mobile elements and prophages from the genome (MDS42 strain background), viability was restored.

	Log_2 Ratio Enrichment						Group-Level Mean			
HMM Class	EPODs	TFs	Prom.	IHF	Dam	Dcm	H-NS	Hfq	Fis	HU
0	-0.95	-3.49	-2.31	-1.12	0.16	0.08	276.55	-0.14	227.68	1.02
1	-0.36	-1.10	-1.17	-1.01	0.09	-0.04	340.21	0.19	231.26	1.07
2	1.36	1.22	-0.054	1.83	-0.65	-0.78	1230.42	0.42	153.19	0.88
3	1.66	-2.52	-4.24	-1.42	-0.15	-0.19	608.69	0.06	196.31	0.92
4	-1.53	-4.03	-2.41	-3.68	0.14	0.30	268.33	-0.36	220.40	0.93
5	0.06	2.14	1.70	2.21	-0.39	-0.35	437.99	0.77	238.11	1.08
Genome Average	----	----	----	----	----	----	422.24	0.03	217.63	0.99

Table S3.1: HMM class enrichments. Log_2 Ratio Enrichment: The ratio of the number of EPODs or motifs in a given HMM class to the total number of EPODs or motifs was calculated for each HMM class. A chi-squared test was performed, and all categories were significantly associated with each class; values underlined had a p-value <0.05. Group-Level Mean: The 500-bp rolling mean for the binding of each NAP was used to calculate the group-level means for across each HMM class, and compared with the overall average for the genome. Permutation based p-values were calculated comparing each class vs. the background. The values underlined had a p-value <0.05.

Region #	Coordinates	Genes	Gene Functions
1	564815-585633	essD, ybcS, rzpD, rzoD, borD, ybcV, ybcW, nohB	DLP12 prophage, putative prophage endopeptidase, lysozyme
2	1400247-1482201	lar, recT, ydaQ, ydaC, intR	Rac prophage, recombinase, DNA renaturation
3	1627517-1652838	ynfO, ydfO, gnsB, ynfN, cspI, ydfP, hokD	Qin prophage, cold shock, toxin anti-toxin system

Table S3.2: MDS42 regions containing prophages.

Strain Name	Strain Label	Marker(s)	Relevant Genotype
MG1655	HA04		Freddolino Lab
MDS42	HA01		Posfai et al., 2006
MG1655 (2)	HA69		Jakob Lab
MG1655 *Δhfq*	HA05		this study
MG1655 *Δfis*	HA06		this study
MG1655 *ΔstpA*	HA10		this study
MG1655 ΔR1 *Δfis*	HA25		this study
MG1655 ΔR2 *Δfis*	HA26		this study
MG1655 ΔR3 *Δfis*	HA28		this study
MG1655 *ΔstpA Δhns* (1)	HA27		this study
MG1655 *ΔstpA Δhns* (2)	HA 270		this study
MDS42 *Δfis*	HA30		this study
MG1655 *Δhns* (1)	TG119		this study
MG1655 *Δhns* (2)	HA268		this study
MG1655 *ihfA*::Kan *ihfB*::Clm	TG156	Kan, Clm	this study
MG1655 *ΔhupA hupB*::Kan	TG46	Kan	this study
MG1655 *Δdps*	TG14		this study
PY79			this study
rok::kan			this study

Table S3.3: Strains used in this study.

Name	Marker (s)	Reference	Description
pKD46	Amp	Datsenko and Wanner, 2000	λ Red recombinase
pCP20	Amp	Datsenko and Wanner, 2000	Flp recombinase
pKD4	Kan	Datsenko and Wanner, 2000	Kan Kanamycin resistance casette donor
pKOV-pSC101 (TS ori - hfq)	Clm	Link et al J Bact. 1997 ; cloned in hfq for this study	Temperature sensitive plasmid pKOV, sacB and its promoter replaced with hfq and hfq's promoters

Table S3.4: Plasmids used in this study.

ID	Description	Sequence
P2633	R1_delM12_F	GGTTCGAATCCTGCAGGGCGCGCCATTACAATTCAATCAGgtgtaggctggagctgcttc
P2634	R1_delM12_R	TACATATTCAATCATTAAAACGATTGAATGGAGAACTTTTcatatgaatatcctccttag
P2635	delM12_Check_F	ATTGACTCAGCAAGGGTTGACC
P2636	delM12_Check_R	CGGCCCACGACTTAGAAGTTCC
P2637	delM12_Check_C	GGCCGCCATCAGGAAAGG
P2638	R2_delM2_F	GGCAATTTTTTCTTCAAGTAATCTCAGCATCCGTTCTCTCgtgtaggctggagctgcttc
P2639	R2_delM2_R	TGATGTGGACTGTAGATATTCAGTCCACATCTCAATCCACcatatgaatatcctccttag
P2640	delM2_Check_F	CGACGATGCTGATGGGATTCGATC
P2641	delM2_Check_R	GGTCCGGAAATGGCAGCG
P2642	delM2_Check_C	CGCATAGCAGGTGTCGTATCGC
P2643	R3_delM8_F	GCTATGTTATTGACACACAAAAGCGTTGAGGAACAGTGAGgtgtaggctggagctgcttc
P2644	R3_delM8_R	GTACGCATCTTACCTCTTTTTTAGAGATAACCATTcatatgaatatcctccttag
P2645	delM8_Check_F	GACGCTGGTAACGCGGG
P2646	delM8_Check_R	CCCATACGTTTGATTTCCAGCATGTTGC
P2647	delM8_Check_C	GACGTGAGCAGGCAGCG
P2650	pDK4_dhns_F	CCTCAACAAACCACCCCAATATAAGTTTGAGATTACTACAgtgtaggctggagctgcttc
P2651	pDK4_dhns_R	GCCGCTGGCGGGATTTTAAGCAAGTGCAATCTACAAAAGAcatatgaatatcctccttag
P2652	dhns_check_F	GGGCTATATGCCGCGTC
P2653	dhns_check_C	CGGGTCAATACCGTCAGC
P2654	dhns_check_R	GGGTGAAAGCGTACCGATG
P2655	pDK4_dstpA_F	AATACTTTTTTTGTTTTGGCGTTAAAAGGTTTTCTTTATTgtgtaggctggagctgcttc
P2656	pDK4_dstpA_R	ACGCCGGACGCGCCCTAGCAGCGACATCCGGCCTCAGTAAcatatgaatatcctccttag
P2657	dstpA_check_F	GGTGAGGTAACGCTATAAGCG
P2658	dstpA_check_C	GAGCCATCGCACGGAG
P2659	dstpA_check_R	CTGCCAGGCAGGTAAACG
P3055	dhfq_F_check	ATGTGGTCTTACCTTGAAGGCG
P3056	dhfq_C_check	CGGCGTTGTTACTGTGATGAG
P3057	dhfq_R_check	CGGTCAAACAAGCGTATAACCC
P3058	dfis_F_check	CCATTCCGGTTATCGCGAATG
P3059	dfis_C_check	TGCTCAACTGAATGGTCAGGATG
P3060	dfis_R_check	ACTTTCGGCGGGGATCTTTT
P3409	ihfA_F_check	GAATCCGGCACTGCATCC
P3410	ihfA_C_check	CTTGCTAAGCCCAAGCTTATC
P3411	ihfA_R_check	GCAATCGCACACAGCC
P3412	ihfB_F_check	CTGCAGTTGACGCTAAAGG
P3413	ihfB_C_check	GTGCGATTGCTGGGTG
P3414	ihfB_R_check	CAGTGATCTCAACAATTGCATCC

Table S3.5: Primers used in this study.

Supplementary Text

Rescaling of IPOD-HR occupancy tracks and subsequent EPOD calling

While we have previously shown the robustness of the IPOD-HR analysis pipelines for both deletions of local regulators and substantial changes in physiological conditions [11], we found that for the nucleoid-associated protein deletions considered here, in many cases the assumptions underlying the IPOD-HR normalization methods (that the overall shape of the distribution of occupancy values across the genome would not change substantively between conditions) was violated. We thus modified the EPOD calling scheme to be able to compare EPOD count, coverage, and occupancy across NAP deletion datasets, where some NAP deletions are sufficient to substantively shift the overall score distribution. We rescaled the IPOD-HR occupancy tracks, beginning with the robust z score values indicating occupancy at every five base pairs of the genome, using the following procedure: We found the intersection of EPODs between each NAP deletion and WT with a minimum fractional overlap of 0.2, and plotted the mean occupancy of each overlapped EPOD (NAP deletion vs. WT). We then used a robust linear model with Huber's T for M estimation to estimate the slope of the NAP deletion values as a function of the WT values for the same EPODs, and used the slope to rescale the NAP deletion dataset by dividing the robust z score values by the slope across the genome. These new values were used as input to call EPODs. To call EPODs, we set the cutoff of what would be counted as an EPOD using the WT thresholds, and applied this to all datasets. In summary, these rescaling and thresholding enabled us to make more accurate comparisons between EPODs and occupancy across the NAP deletion datasets. By bringing the observed occupancies at conserved EPODs into register with each other across the different genotypes in our dataset

References

1. Shen BA, Landick R. Transcription of Bacterial Chromatin. Journal of Molecular Biology. 2019. pp. 4040–4066. doi:10.1016/j.jmb.2019.05.041
2. Luijsterburg MS, White MF, van Driel R, Dame RT. The major architects of chromatin: architectural proteins in bacteria, archaea and eukaryotes. Crit Rev Biochem Mol Biol. 2008;43: 393–418.
3. Dillon SC, Dorman CJ. Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. Nat Rev Microbiol. 2010;8: 185–195.
4. Postow L, Hardy CD, Arsuaga J, Cozzarelli NR. Topological domain structure of the Escherichia coli chromosome. Genes Dev. 2004;18: 1766–1779.
5. Deng S, Stein RA, Higgins NP. Organization of supercoil domains and their reorganization by transcription. Mol Microbiol. 2005;57: 1511–1521.
6. Boudreau BA, Hron DR, Qin L, van der Valk RA, Kotlajich MV, Dame RT, et al. StpA and Hha stimulate pausing by RNA polymerase by promoting DNA-DNA bridging of H-NS filaments. Nucleic Acids Res. 2018;46: 5525–5546.
7. Lim CJ, Whang YR, Kenney LJ, Yan J. Gene silencing H-NS paralogue StpA forms a rigid protein filament along DNA that blocks DNA accessibility. Nucleic Acids Res. 2012;40: 3316–3328.
8. Grainger DC, Goldberg MD, Lee DJ, Busby SJW. Selective repression by Fis and H-NS at the Escherichia coli dps promoter. Mol Microbiol. 2008;68: 1366–1377.
9. 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. 1999;181: 6361–6370.
10. Ueguchi C, Mizuno T. The Escherichia coli nucleoid protein H-NS functions directly as a transcriptional repressor. EMBO J. 1993;12: 1039–1046.
11. Freddolino PL, Goss TJ, Amemiya HM, Tavazoie S. Dynamic landscape of protein occupancy across the Escherichia coli chromosome. doi:10.1101/2020.01.29.924811
12. Singh SS, Singh N, Bonocora RP, Fitzgerald DM, Wade JT, Grainger DC. Widespread suppression of intragenic transcription initiation by H-NS. Genes Dev. 2014;28: 214–219.
13. Kotlajich MV, Hron DR, Boudreau BA, Sun Z, Lyubchenko YL, Landick R. Bridged filaments of histone-like nucleoid structuring protein pause RNA polymerase and aid termination in bacteria. Elife. 2015;4. doi:10.7554/eLife.04970
14. Landick R, Wade JT, Grainger DC. H-NS and RNA polymerase: a love-hate relationship? Curr Opin Microbiol. 2015;24: 53–59.
15. van der Valk RA, Vreede J, Qin L, Moolenaar GF, Hofmann A, Goosen N, et al. Mechanism of environmentally driven conformational changes that modulate H-NS DNA-bridging activity. Elife. 2017;6. doi:10.7554/eLife.27369
16. Lucchini S, Rowley G, Goldberg MD, Hurd D, Harrison M, Hinton JCD. H-NS mediates the silencing of laterally acquired genes in bacteria. PLoS Pathog. 2006;2: e81.

17. Kahramanoglou C, Seshasayee ASN, Prieto AI, Ibberson D, Schmidt S, Zimmermann J, et al. Direct and indirect effects of H-NS and Fis on global gene expression control in Escherichia coli. Nucleic Acids Res. 2011;39: 2073–2091.
18. Baumler A. Faculty Opinions recommendation of Selective silencing of foreign DNA with low GC content by the H-NS protein in Salmonella. Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature. 2006. doi:10.3410/f.1032929.492954
19. Navarre WW. Selective Silencing of Foreign DNA with Low GC Content by the H-NS Protein in Salmonella. Science. 2006. pp. 236–238. doi:10.1126/science.1128794
20. Vora T, Hottes AK, Tavazoie S. Protein occupancy landscape of a bacterial genome. Mol Cell. 2009;35: 247–253.
21. Scholz SA, Diao R, Wolfe MB, Fivenson EM, Lin XN, Freddolino PL. High-Resolution Mapping of the Escherichia coli Chromosome Reveals Positions of High and Low Transcription. Cell Syst. 2019;8: 212–225.e9.
22. Nakamura K, Ogura Y, Gotoh Y, Hayashi T. Prophages integrating into prophages: a mechanism to accumulate type III secretion effector genes and duplicate Shiga toxin-encoding prophages in Escherichia coli. doi:10.1101/2020.11.04.367953
23. Remesh SG, Verma SC, Chen J-H, Ekman AA, Larabell CA, Adhya S, et al. Nucleoid remodeling during environmental adaptation is regulated by HU-dependent DNA bundling. Nat Commun. 2020;11: 2905.
24. Verma SC, Qian Z, Adhya SL. Correction: Architecture of the Escherichia coli nucleoid. PLoS Genet. 2020;16: e1009148.
25. Neeli-Venkata R, Martikainen A, Gupta A, Gonçalves N, Fonseca J, Ribeiro AS. Robustness of the Process of Nucleoid Exclusion of Protein Aggregates in Escherichia coli. J Bacteriol. 2016;198: 898–906.
26. Small Things Considered. [cited 25 Dec 2020]. Available: https://schaechter.asmblog.org/schaechter/2009/11/the-limitations-of-lb-medium.html
27. Neidhardt FC, Bloch PL, Smith DF. Culture Medium for Enterobacteria. Journal of Bacteriology. 1974. pp. 736–747. doi:10.1128/jb.119.3.736-747.1974
28. Freddolino PL, Amini S, Tavazoie S. Newly identified genetic variations in common Escherichia coli MG1655 stock cultures. J Bacteriol. 2012;194: 303–306.
29. Goodarzi H, Elemento O, Tavazoie S. Revealing global regulatory perturbations across human cancers. Mol Cell. 2009;36: 900–911.
30. Wang L, Reeves PR. Organization of Escherichia coli O157 O antigen gene cluster and identification of its specific genes. Infect Immun. 1998;66: 3545–3551.
31. D'Souza JM, Wang L, Reeves P. Sequence of the Escherichia coli O26 O antigen gene cluster and identification of O26 specific genes. Gene. 2002. pp. 123–127. doi:10.1016/s0378-1119(02)00876-4
32. Feng L, Han W, Wang Q, Bastin DA, Wang L. Characterization of Escherichia coli O86 O-antigen gene cluster and identification of O86-specific genes. Veterinary Microbiology. 2005. pp. 241–248. doi:10.1016/j.vetmic.2004.12.021

33. Nakao R, Ramstedt M, Wai SN, Uhlin BE. Enhanced biofilm formation by Escherichia coli LPS mutants defective in Hep biosynthesis. PLoS One. 2012;7: e51241.
34. Linkevicius M, Sandegren L, Andersson DI. Mechanisms and fitness costs of tigecycline resistance in Escherichia coli. J Antimicrob Chemother. 2013;68: 2809–2819.
35. Wang Z, Wang J, Ren G, Li Y, Wang X. Deletion of the genes waaC, waaF, or waaG in Escherichia coli W3110 disables the flagella biosynthesis. J Basic Microbiol. 2016;56: 1021–1035.
36. Frenkiel-Krispin D, Levin-Zaidman S, Shimoni E, Wolf SG, Wachtel EJ, Arad T, et al. Regulated phase transitions of bacterial chromatin: a non-enzymatic pathway for generic DNA protection. EMBO J. 2001;20: 1184–1191.
37. Nair S, Finkel SE. Dps protects cells against multiple stresses during stationary phase. J Bacteriol. 2004;186: 4192–4198.
38. Salgado H, Martínez-Flores I, Bustamante VH, Alquicira-Hernández K, García-Sotelo JS, García-Alonso D, et al. Using RegulonDB, the Escherichia coli K-12 Gene Regulatory Transcriptional Network Database. Curr Protoc Bioinformatics. 2018;61: 1.32.1–1.32.30.
39. Bausch C, Ramsey M, Conway T. Transcriptional organization and regulation of the L-idonic acid pathway (GntII system) in Escherichia coli. J Bacteriol. 2004;186: 1388–1397.
40. Gómez KM, Rodríguez A, Rodriguez Y, Ramírez AH, Istúriz T. The subsidiary GntII system for gluconate metabolism in Escherichia coli: alternative induction of the gntV gene. Biol Res. 2011;44: 269–275.
41. Bausch C, Peekhaus N, Utz C, Blais T, Murray E, Lowary T, et al. Sequence analysis of the GntII (subsidiary) system for gluconate metabolism reveals a novel pathway for L-idonic acid catabolism in Escherichia coli. J Bacteriol. 1998;180: 3704–3710.
42. Barh D, Azevedo V. Omics Technologies and Bio-engineering: Volume 1: Towards Improving Quality of Life. Academic Press; 2017.
43. Francis NJ, Kingston RE. Mechanisms of transcriptional memory. Nature Reviews Molecular Cell Biology. 2001. pp. 409–421. doi:10.1038/35073039
44. Lagha M, Ferraro T, Dufourt J, Radulescu O, Mantovani M. Transcriptional Memory in the Drosophila Embryo. Mechanisms of Development. 2017. p. S137. doi:10.1016/j.mod.2017.04.382
45. Palozola KC, Lerner J, Zaret KS. A changing paradigm of transcriptional memory propagation through mitosis. Nat Rev Mol Cell Biol. 2019;20: 55–64.
46. Kundu S, Horn PJ, Peterson CL. SWI/SNF is required for transcriptional memory at the yeast GAL gene cluster. Genes Dev. 2007;21: 997–1004.
47. Moazed D. Mechanisms for the inheritance of chromatin states. Cell. 2011;146: 510–518.
48. Wang X, Kim Y, Ma Q, Hong SH, Pokusaeva K, Sturino JM, et al. Cryptic prophages help bacteria cope with adverse environments. Nat Commun. 2010;1: 147.

49. Hong SH, Wang X, Wood TK. Controlling biofilm formation, prophage excision and cell death by rewiring global regulator H-NS of Escherichia coli. Microb Biotechnol. 2010;3: 344–356.
50. Nagai K. Faculty Opinions recommendation of Hfq: a bacterial Sm-like protein that mediates RNA-RNA interaction. Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature. 2002. doi:10.3410/f.1003709.40104
51. Valentin-Hansen P, Eriksen M, Udesen C. The bacterial Sm-like protein Hfq: a key player in RNA transactions. Molecular Microbiology. 2004. pp. 1525–1533. doi:10.1111/j.1365-2958.2003.03935.x
52. Orans J, Kovach AR, Hoff KE, Horstmann NM, Brennan RG. Crystal structure of an Escherichia coli Hfq Core (residues 2–69)–DNA complex reveals multifunctional nucleic acid binding sites. Nucleic Acids Research. 2020. pp. 3987–3997. doi:10.1093/nar/gkaa149
53. McQuail J, Switzer A, Burchell L, Wigneshweraraj S. The RNA-binding protein Hfq assembles into foci-like structures in nitrogen starved. J Biol Chem. 2020;295: 12355–12367.
54. Bokal AJ 4th, Ross W, Gourse RL. The transcriptional activator protein FIS: DNA interactions and cooperative interactions with RNA polymerase at the Escherichia coli rrnB P1 promoter. J Mol Biol. 1995;245: 197–207.
55. Appleman JA, Ross W, Salomon J, Gourse RL. Activation of Escherichia coli rRNA transcription by FIS during a growth cycle. J Bacteriol. 1998;180: 1525–1532.
56. Chintakayala K, Singh SS, Rossiter AE, Shahapure R, Dame RT, Grainger DC. E. coli Fis protein insulates the cbpA gene from uncontrolled transcription. PLoS Genet. 2013;9: e1003152.
57. Cho B-K, Knight EM, Barrett CL, Palsson BØ. Genome-wide analysis of Fis binding in Escherichia coli indicates a causative role for A-/AT-tracts. Genome Res. 2008;18: 900–910.
58. Updegrove TB, Zhang A, Storz G. Hfq: the flexible RNA matchmaker. Curr Opin Microbiol. 2016;30: 133–138.
59. Link AJ, Phillips D, Church GM. Methods for generating precise deletions and insertions in the genome of wild-type Escherichia coli: application to open reading frame characterization. Journal of bacteriology. 1997. pp. 6228–6237. doi:10.1128/jb.179.20.6228-6237.1997
60. Posfai G. Emergent Properties of Reduced-Genome Escherichia coli. Science. 2006. pp. 1044–1046. doi:10.1126/science.1126439
61. Albano M, Smits WK, Ho LTY, Kraigher B, Mandic-Mulec I, Kuipers OP, et al. The Rok Protein of Bacillus subtilis Represses Genes for Cell Surface and Extracellular Functions. J Bacteriol. 2005;187: 2010.
62. Hoa TT, Tortosa P, Albano M, Dubnau D. Rok (YkuW) regulates genetic competence in Bacillus subtilis by directly repressing comK. Mol Microbiol. 2002;43. doi:10.1046/j.1365-2958.2002.02727.x
63. Wiep Klaas Smits ADG. The Transcriptional Regulator Rok Binds A+T-Rich DNA and Is Involved in Repression of a Mobile Genetic Element in Bacillus subtilis. PLoS Genet. 2010;6. doi:10.1371/journal.pgen.1001207

64. Graumann PL. Bacillus subtilis SMC Is Required for Proper Arrangement of the Chromosome and for Efficient Segregation of Replication Termini but Not for Bipolar Movement of Newly Duplicated Origin Regions. J Bacteriol. 2000;182: 6463–6471.
65. Sullivan NL, Marquis KA, Rudner DZ. Recruitment of SMC to the origin by ParB-parS organizes the origin and promotes efficient chromosome segregation. Cell. 2009;137: 697.
66. Wilhelm L, Bürmann F, Minnen A, Shin H-C, Toseland CP, Oh B-H, et al. SMC condensin entraps chromosomal DNA by an ATP hydrolysis dependent loading mechanism in Bacillus subtilis. 2015 [cited 7 Nov 2020]. doi:10.7554/eLife.06659
67. Al-Bassam MM, Moyne O, Chapin N, Zengler K. Nucleoid openness profiling links bacterial genome structure to phenotype. doi:10.1101/2020.05.07.082990
68. Wasim A, Gupta A, Mondal J. Mapping the Multiscale Organisation of Escherichia Coli Chromosome in a Hi-C-integrated Model. doi:10.1101/2020.06.29.178194
69. Lioy VS, Cournac A, Marbouty M, Duigou S, Mozziconacci J, Espéli O, et al. Multiscale Structuring of the E. coli Chromosome by Nucleoid-Associated and Condensin Proteins. Cell. 2018. pp. 771–783.e18. doi:10.1016/j.cell.2017.12.027
70. Walker DM, Freddolino PL, Harshey RM. A Well-Mixed E. coli Genome: Widespread Contacts Revealed by Tracking Mu Transposition. Cell. 2020;180: 703–716.e18.
71. Shin J-E, Lin C, Lim HN. Horizontal transfer of DNA methylation patterns into bacterial chromosomes. Nucleic Acids Res. 2016;44: 4460–4471.
72. Tjaden B. A computational system for identifying operons based on RNA-seq data. Methods. 2020;176: 62–70.
73. Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. Nat Methods. 2017;14: 687–690.
74. Freddolino PL, Tavazoie S. Beyond Homeostasis: A Predictive-Dynamic Framework for Understanding Cellular Behavior. Annual Review of Cell and Developmental Biology. 2012. pp. 363–384. doi:10.1146/annurev-cellbio-092910-154129
75. Thomason LC, Costantino N, Court DL. E. coli genome manipulation by P1 transduction. Curr Protoc Mol Biol. 2007;Chapter 1: Unit 1.17.
76. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, et al. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol. 2006;2: 2006.0008.
77. Cherepanov PP, Wackernagel W. Gene disruption in Escherichia coli: TcR and KmR cassettes with the option of Flp-catalyzed excision of the antibiotic-resistance determinant. Gene. 1995;158: 9–14.
78. Jeremy W. Schroeder LAS. Complete Genome Sequence of Bacillus subtilis Strain PY79. Genome Announc. 2013;1. doi:10.1128/genomeA.01085-13
79. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34: 525–527.
80. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020;585: 357–362.

81. Verzani J. Getting Started with RStudio: An Integrated Development Environment for R. “O’Reilly Media, Inc.”; 2011.
82. RStudio. [cited 27 Dec 2020]. Available: https://rstudio.com/
83. Tidyverse. [cited 27 Dec 2020]. Available: https://www.tidyverse.org/
84. Create Elegant Data Visualisations Using the Grammar of Graphics. [cited 27 Dec 2020]. Available: https://ggplot2.tidyverse.org

Chapter 4

Interplay of Hfq and Polyphosphate in Bacterial Heterochromatin Formation

Abstract and Introduction

Recent evidence suggests that bacteria contain heterochromatin-like domains, termed extended protein occupancy domains (EPODs), that contribute to gene regulation and protection against foreign DNA. One of the most prominent classes of EPODs, those dependent on the activity of the nucleoid-associated proteins Fis and Hfq, bears a strong enrichment for prophages and mobile elements. Further genetic testing indicates that complete silencing of Fis/Hfq dependent EPODs are also reliant upon normal levels of polyphosphate (polyP), and we find that in fact polyphosphate is essential for appropriate silencing activity of Hfq at integrated prophages. Biochemical results suggest a model in which polyphosphate acts as an Hfq chaperone in order to permit appropriate silencing at EPODs, whereas the well-characterized function of Hfq as an RNA chaperone appears polyphosphate independent. We begin to define a model by which polyP mediates Hfq prophage regulation, where deletion of *ppk* dramatically changes the binding capacity of Hfq to xenogeneic elements. These results provided the first evidence that polyP and Hfq form heterchromatin like regions that suppress the expression of genetic mobile elements and prophages.

The contents of this chapter are in preparation by Francois Beaufay*, Haley M. Amemiya*, Jian Guan, Rishav Mitra, Benjamin Meinen, James C. A. Bardwell, Ursula Jakob and Peter L. Freddolino. Conceptualization: F.B., H.M.A U.J. and P.L.F.; Methodology, F.B., H.M.A U.J. and P.L.F; Investigation, F.B., H.M.A., J.G., R.M., B.M., U.J. and P.L.F., and P.L.F.; Data Analysis and Curation, F.B., H.M.A U.J. and P.L.F.; Writing -- Original Draft, F.B., H.M.A U.J. and P.L.F.; Funding Acquisition, U.J. and P.L.F. (*) indicates co-first authors. I aided in the writing of the manuscript, and wrote all portions pertaining to the RNA-seq analysis and ChIP-seq experiments. I performed analysis on the RNA-seq datasets. I performed the mCherry ChIP experiments and analysis.

Results

Polyphosphate, an extremely simple energy-rich polymer composed of phospho-anhydride bonded phosphates, is one of the most ancient and conserved molecules on earth [1]. Present in every organism tested so far, polyP plays a variety of different biological roles, ranging from a virulence and stress resistance factor in bacteria, to a blood clotting factor and modulator of amyloidogenic processes in eukaryotes[1–8]. The many ascribed activities of polyP have been attributed to its physico-chemical properties, acting as a metal chelator, polyanionic buffer, and, as most recently discovered, protein scaffolding factor, a function of potentially far-reaching physiological consequences[9–12].

Earlier studies in *E. coli* revealed that polyP protects bacteria against the DNA-crosslinking reagent cisplatin [13]. This protection was in part mediated by the ability of polyP to counteract cisplatin-elicited iron stress. In addition to the significant differences in the expression of iron homeostasis genes in *E. coli* mutants lacking the polyP synthesizing polyphosphate kinase (*ppk*) compared to wild-type *E. coli*, however, we observed that the *ppk* deletion strain also showed a pronounced enrichment of upregulated genes related to “genetic mobile elements” (GME) such as prophages and prophage shock genes (**Fig. 4.1A**). Many of these genes, which are associated with transposons and insertion sequence (IS) elements, are known to be induced upon DNA damage, and their mobilization contributes to the lethal consequences of DNA-stress[14–17]. Absence of polyP significantly augmented the cisplatin-induced expression of GMEs and prophages, and, even more surprisingly, also increased their steady-state expression levels in the absence of stress (**Fig. 4.1A**). These results provided the first evidence that polyP might act in DNA damage control by either directly or indirectly suppressing the expression of GMEs and prophages.

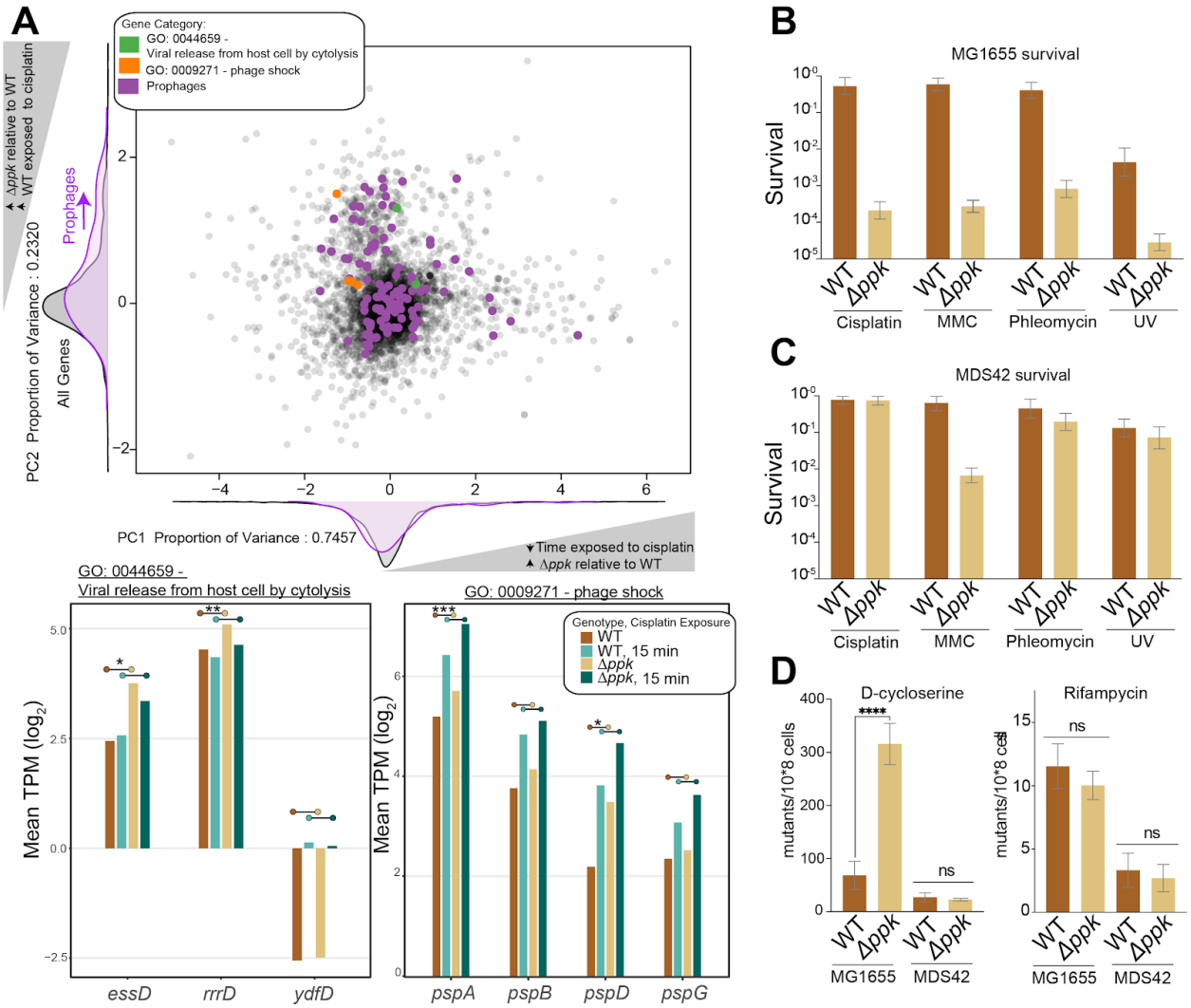

Figure 4.1: Loss of *ppk* leads to an induction of prophages and mobile elements and sensitivity to DNA damaging agents. (A) Mobilization / increased expression of prophages in both dppk and cells exposed to cisplatin. Performed RNA-seq on WT and dppk cells exposed to cisplatin at 0, 5, 15min timepoints. The above plot displays a dimensional reduction of examining the impact of genotype (WT vs *Δppk*) and time exposed to cisplatin. Displayed are all genes (each dot) and highlighted genes that are prophages (purple), phage shock GO term (orange), and viral release from host cell by cytolysis GO term (green). Genes positive in PC1, which comprises 74.57% proportion of variance, decreases expression in time exposed to cisplatin, but up in *Δppk* relative to WT. Genes positive in PC2, 23.20% proportion of variance, increased expression in *Δppk* relative to WT and with exposure to cisplatin. We zoomed in on specific GO terms and analyzed the mean TPM (log2) in each condition. Overall, the deletion of ppk alone induces expression of prophages. **(B)** Survival of WT vs *Δppk* cells exposed to DNA damaging agents. **(C)** Survival of MDS42 (WT) vs MDS42 *Δppk* (*Δppk*) cells exposed to DNA damaging agents. **(D)** Mutagenesis rates in D-cycloserine and Rifampin of WT, *Δppk* (MG1655) and MDS42 (WT), MDS42 *Δppk* (*Δppk*) cells.

To further investigate this idea, we first conducted phenotypical studies on wild-type and the *ppk* deletion strain using three unrelated DNA-damaging reagents, including mitomycin C, phleomycin and UV. We observed that lack of polyP significantly increased the sensitivity of *E. coli* towards all of these DNA damaging reagents and treatments, irrespective of their specific effects on DNA integrity(**Fig. 4.1B**). We hypothesized that the sensitivity was due to the de-repression of mobile elements and prophages due to the loss of *ppk*, leading to an increase in mobilization and lethality with the exposure to DNA damaging agents. The *E. coli* strain MDS42 lacks all identifiable prophages and genetic mobile elements [18], was much less sensitive towards DNA damaging reagents compared to wild-type *E. coli*, and, even more importantly, much less dependent on the presence of polyP for survival. Whereas deletion of *ppk* caused a more than 1,000-fold increase in cisplatin sensitivity in *E. coli* MG1655, absence of polyP in MDS42 led to a less than 10-fold increase in sensitivity compared to the parental strain (**Fig. 4.1C**). As expected, these effects were not restricted to cisplatin treatment but applied to all other tested DNA damaging reagents (**Fig. 4.1C**). These results not only demonstrated that mobilization and transposition of GMEs and prophages significantly contributes to the bacterial killing that is elicited by various DNA damaging reagents, but implied that polyP serves to prevent this mobilization thus protecting bacteria against the lethal effects of DNA damage.

Mobilization of GMEs and prophages inevitably causes chromosomal insertions and deletions. To determine whether polyP's effects on GMEs and prophages influences the mutagenesis rates under non-stress conditions, we compared the bacterial growth of MG1655, MDS42 and the respective *ppk* deletion strains in presence of the inhibitor D-cycloserine or rifampin (**Fig. 4.1D**). Whereas resistance to rifampin is elicited by specific point mutations in the rifampin -binding site of RpoB, an essential *E. coli* protein that tolerates neither insertions nor deletions [19,20], resistance to D-cylcoserine is acquired by loss of function mutations, including insertions or deletions, in the antibiotic transporter CycA[21]. As shown in **Figure 4.1D**, while polyP did not affect rifampin resistance in either strain background, deletion of the *ppk* gene caused a significantly higher rate of resistance to D-cycloserine in wild-type MG1655. In contrast, no

significant increase in D-cycloserine growth was observed in the MDS42 background, consistent with a lack of transposable GMEs and prophages in this strain. Based on these results, we concluded that the protective effect of polyP under DNA damaging conditions is mediated by its ability to either directly or indirectly suppress the expression of GMEs and prophages, a novel and hitherto unknown activity of polyP.

The silencing of GMEs and prophages have been connected to heterochromatin-like domains, termed extended protein occupancy domains (EPODs), composed primary of nucleoid associated proteins (Chapter 2 and Chapter 3). All NAPs have similar qualities in their DNA binding capacity, such as promiscuous binding across the genome and high abundance. H-NS, Hfq and Fis have been shown to silence prophages and mobile elements (Chapter 3 and citations within), however their mechanism of binding remains elusive. In addition, previous reports cited a potential role of polyP in DNA condensation in both *Pseudomonas aeruginosa* [22]and *Cyanobacteria* [23]. These reports led us to consider the possibility that polyP represses mobilization of GMEs and prophages by contributing to nucleoid formation and chromosomal compaction. To genetically interrogate this idea, we individually deleted genes for the six best-characterized nucleoid-associated proteins (NAPs), that is HupA, HupB, StpA, Hfq, Fis, and H-NS, and compared their cisplatin sensitivity to wild-type *E. coli* MG1655 both in the absence and presence of polyP (**Fig. S4.2A**). Neither *hupA*, *hupB* nor *stpA* deletion strains showed any significant increase in the cisplatin sensitivity compared to wild-type *E. coli*, and co-deletion of *ppk* did not lead to any significant further increase in cisplatin sensitivity beyond what we observed in the *ppk* deletion strain (**Fig. S4.2A**).

In contrast, however, deletion of either *hfq* or *fis* increased the sensitivity of MG1655 to an extent that was comparable to the *ppk* deletion strain (**Fig. 4.2A,S4.2A**). To investigate potential genetic interactions between *hfq, fis* and *ppk,* we created double deletions and measured sensitivity. Deletion of the *ppk* gene led to additional sensitization in the *fis* deletion strain background yet did not significantly alter growth or cisplatin sensitivity of the *hfq* deletion strain (**Fig. 4.2A,S4.2A**), suggesting that *hfq* and *ppk* may act in similar capacities and pathways to silence prophages and mobile

elements. Furthermore, a double deletion of *hfq* and *ppk* resulted in no significant changes in mutagenesis rates with exposure to D-cycloserine or Rifampin or additive sensitivity to DNA damaging agents (**Fig. 4.2E, S4.2D**). As a control, we tested whether there was a change in Hfq abundance due to the loss of *ppk*, and found that the deletion of *ppk* does not change Hfq abundance in the cell (**Fig. S4.3**). The epistatic masking of the effects of *ppk* deletion by deletion of *hfq* suggests that the effects of polyphosphate in preventing cisplatin-mediated death occur through Hfq. Indeed, we found that overexpression of Hfq rescued the cisplatin-sensitive phenotype of the *ppk* deletion strain, indicating that increasing the steady state levels of Hfq was sufficient to compensate for the absence of polyP, and this pattern held true for all DNA damaging agents (**Fig. 4.2F, S4.2B**). The converse experiment was not the case and overexpression of PPK was not sufficient to rescue the cisplatin sensitivity of the *hfq* deletion strain (**Fig. 4.2F, S4.2B**). These results suggest that polyP acts upstream of Hfq in terms of resistance to DNA damage.

Similarly to polyP, Hfq has many roles in the cell. Hfq's DNA binding capacity has only been recently described [26, 27], as it is typically considered in the literature to be known for its role as RNA chaperone[24,25]. However, studies have shown a direct role of Hfq in dsDNA binding and compaction [27], highlighting the critical role Hfq plays in silencing specific regions of the chromosome. Furthermore, Hfq form foci in response to starvation stress [26], implicating the importance of Hfq sequestration during physiological changes. We have recently shown that both Hfq and Fis are required silencing prophages across the genome (Chapter 3) and make up the main component of heterochromatin-like domains, EPODs, across these regions. Coupled with our previous findings that Hfq is a critical component of silencing complexes across prophages, and the relationship we have shown between Hfq and polyP, we hypothesized that polyP facilitates Hfq's ability to silence prophages and mobile elements in *E. coli*.

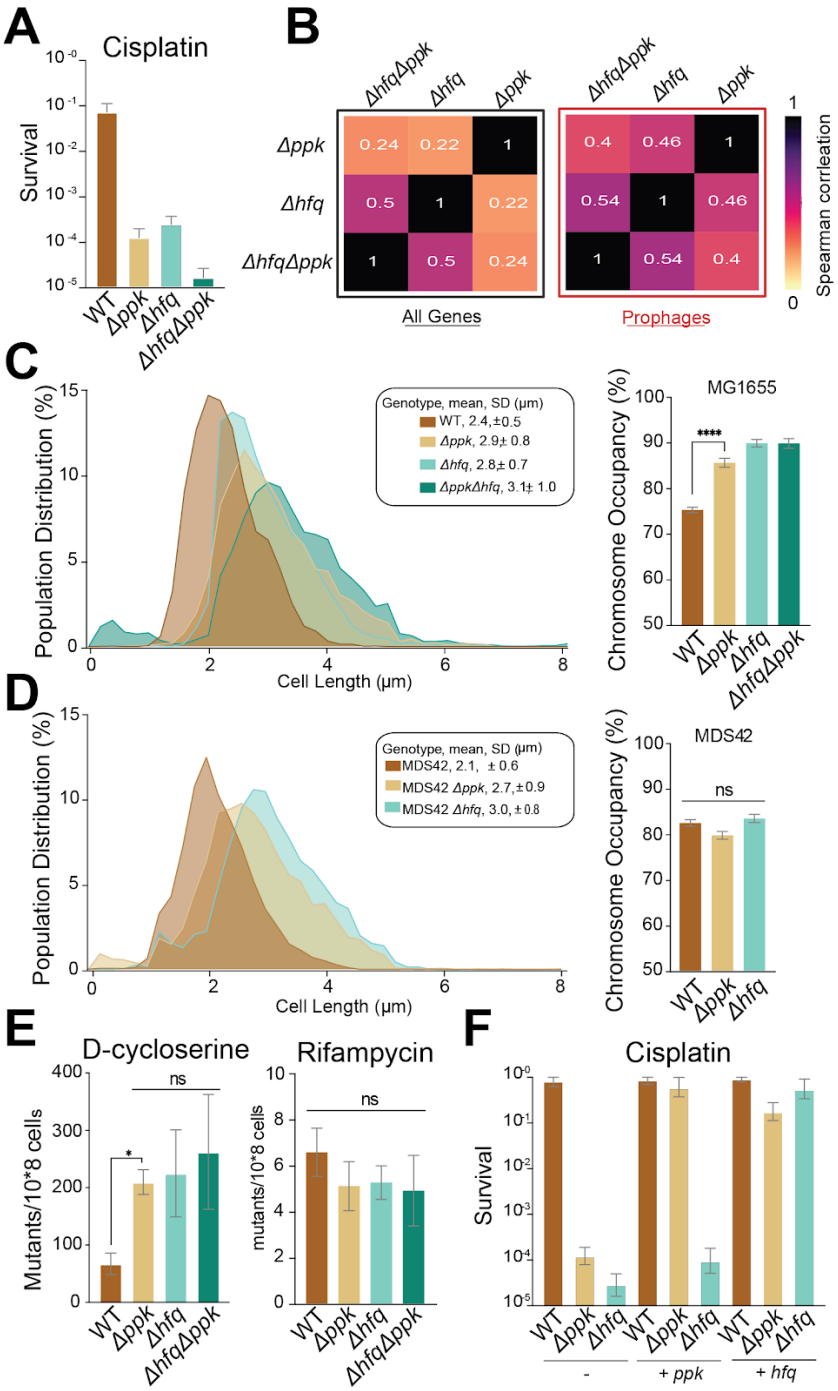

Figure 4.2: polyP and Hfq show evidence of epistasis. (A) Survival assay of WT, *Δppk, Δhfq,* and *ΔppkΔhfq* cells exposed to cisplatin. **(B)** RNA-sequencing differential expression analysis was performed on all genotypes compared to WT. The spearman correlation of these differences was calculated for "All Genes" and "Prophages". There was an increase in correlation specifically at prophage genes, indicating similar induction of toxic elements. **(C)** Cell size distribution of WT (n=8670*), Δppk* (n=8089), *Δhfq* (n=5875), *ΔppkΔhfq* (n=5954) population (left) and the associate space occupied by the nucleoid for each mutant (right) in a MG1655 background under non stress condition in MOPS-Glucose medium. **(D)** Cell size distribution of WT

(n=7329*), Δppk* (n=5577), *Δhfq* (n=4984)*,* population (left) and the associate space occupied by the nucleoid for each mutant (right) in a MDS42 background under non stress condition in MOPS-Glucose medium. **(E)** Resistant mutant frequencies of WT, *Δppk, Δhfq, ΔppkΔhfq* strains to D-cycloserine or rifampycin. **(F)** Epistatic relation between Hfq and polyP to cisplatin exposure (n³3, *, P<0.05; ****, P<0.0001; ns, non-significant, one-way ANOVA).

To ascertain whether polyP impacts Hfq's RNA chaperone activity versus its DNA binding activity (or both), we performed expression analysis of known targets of Hfq's RNA chaperone activity. Expression analysis of genes previously shown to be controlled by the RNA chaperone activity of Hfq, did not reveal any significant differences between wild-type MG1655 and the *ppk* deletion strain (**Fig. S4.1**). These results suggested that absence of polyP does not notably affect the RNA chaperone function of Hfq. In addition to its role as RNA chaperone, about 20% of the cellular Hfq pool has been found to be associated with DNA, where it appears to be involved in bacterial nucleoid formation [26,27]. To directly test the effects of polyP on chromosome condensation in the absence or presence of Hfq, we measured and compared cell lengths and chromosome occupancy using wild-type MG1655 as well as single and double deletions of *ppk* and *hfq*. Cells lacking the *ppk* gene, the *hfq* gene, or both had elongated cell shapes (**Fig. 4.2C**) and showed a near 20% increase in chromosome occupancy compared to wild-type *E. coli* (**Fig. S4.2C**). No significant differences between the three mutant strains (*ppk, hfq,* and *ppk/hfq*) were detected, suggesting that polyP and Hfq cooperate in bacterial nucleoid formation (**Fig. 4.2C**) . We have recently found that in addition to its structural role, Hfq (together with Fis) forms extended protein occupancy domains that play an essential role in silencing several prophages in wild type *E.* coli K12 (Chapter 3). We thus speculated that the prophage-dependent, Hfq-mediated effects of polyp on cisplatin survival might likewise arise due to silencing of prophages by Hfq, which might be enhanced by the presence of polyP. We thus conducted gene expression studies on wild-type *E. coli* lacking *hfq*, *ppk* or *hfq/ppk*. While we observed about 22% of all differentially expressed to be similar in *hfq* and *ppk* deletion strains, we found an almost 50% overlap in prophage genes (**Fig. 4.2B**). With this knowledge, we wanted to see if the chromosomal occupancy phenotype would be rescued in the background MDS42, which lacks many of the mobile elements and

prophages. Indeed, we find that the chromosomal occupancy phenotype is rescued upon deletion of toxic elements, perhaps indicating a new link to prophage expression and phenotypic changes(**Fig. 4.2D**). Given the well-documented capacity of Hfq to bind nucleic acids, and the chemical similarity of polyP to the backbone of DNA/RNA, we hypothesized that DNA-like molecule polyP may act as a Hfq chaperone to specifically silence areas across the genome.

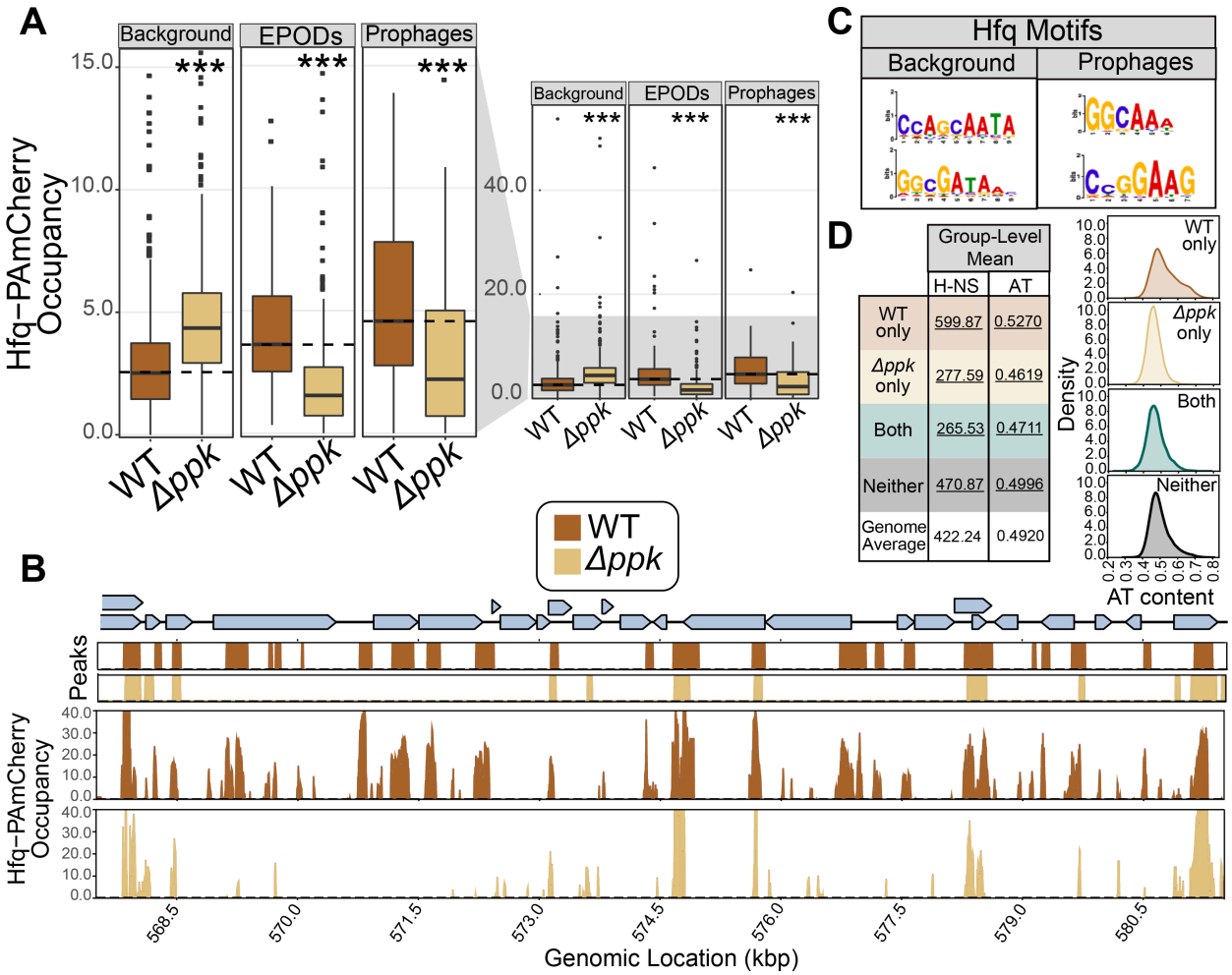

Figure 4.3: Loss of polyphosphate kinase impacts Hfq binding across prophage regions. (**A**) The average Hfq-PAmCherry occupancy was calculated for each genotype. A control WT without any PAmCherry tag was used to subtract any mCherry ChIP-seq signal that is due to noise. (*) indicate the Wilcoxon Rank Sum p value comparing the change in median vs. WT for each condition that has been adjusted using the Bonferroni and Hochberg methods (against a null hypothesis of no difference in medians) (*, P<0.05 ; **,P <0.005 ; ****, P<0.0005). (**B**) Example EPODs that contain prophages and Hfq-PAmCherry occupancy. Peak calls are represented by “Peaks”.

Blue arrows indicate genes. (**C**) Meme chip results of ppk dependent Hfq-PAmCherry peaks in the entire genome (Background) and at EPODs containing prophages (Prophages). (**D**) The 500-bp rolling mean for the binding of H-NS and the AT content of the genome was used to calculate the group-level means for across each peak class, and compared with the overall average for the genome. Permutation based p-values were calculated comparing each class vs. the background. The values underlined had a p-value <0.05.

To examine the impact of the deletion of *ppk* on Hfq binding, we performed mCherry ChIP-seq on Hfq-PAmCherry [28] in WT and *Δppk* strains. To assess binding across the genome, we measured Hfq-PAmCherry binding over three different categories: background, EPODs, and EPODs that contained prophages. Hfq binding significantly decreased with the deletion of *ppk*, with an increase in binding of Hfq to background regions, indicating that Hfq still has the ability to bind DNA if *ppk* is deleted, but less specifically at EPODs (**Fig. 4.3A**). Most dramatically, in EPODs containing prophage genes there was a severe decrease in Hfq binding (**Fig. 4.3A**). An example of one of the regions containing prophages is displayed in **Figure 4.3B**, where much of the Hfq binding is lost. Thus, the deletion of *ppk* impacts Hfq binding across prophages, specifically in Hfq's ability to bind and effectively silence toxic elements. We sought to determine the binding characteristics of Hfq that were specific to its interaction with polyP. Using the peaks called in WT Hfq-PAmCherry and Δppk Hfq-PAmCherry strains, we found the peaks that were ppk dependent to discover Hfq motifs in the entire genome (Background) and at EPODs containing prophages (Prophages) (**Fig. 4.3C**). Furthermore, we found that ppk dependent Hfq peaks (WT only) had an enrichment for H-NS binding association and AT content (**Fig. 4.3D**). In both cases, the loss of ppk shifted Hfq's binding characteristics. Thus, the loss of *ppk* dramatically changes the binding profile of Hfq, with a dramatic loss of binding specifically at prophage regions.

Building upon the genetic evidence of epistasis between *ppk* and *hfq*, and the changes in Hfq binding across the genome with the loss of *ppk*, we investigated the interaction of polyP and Hfq *in vitro*. We found that Hfq has the ability to bind polyphosphate molecules (**Fig. 4.4A**) with a Kd of 2.2uM per phosphate unit of a 300 mer polyP species (compared to the Kd of 6.16uM for a HEX-DNA molecule of a size of ~500bp

(**Fig. 4.4D**)). Strikingly, Hfq forms discrete oligomeric species in the presence of polyphosphate (**Fig. 4.4B,C**), suggesting that polyphosphate mediates higher oligomeric species of Hfq at certain concentrations. We next sought out the interaction between DNA, polyP, and Hfq. Using labeled DNA, we gradually titrated in polyP and found that the high chain length polyP species competes with DNA for binding Hfq (**Fig. 4.4E,F**). This supports a hypothesis that polyP facilitates Hfq reservoirs. However, we investigated whether the higher oligomeric species would still be detected at different chain lengths. Using the same gel shift assay approach in **Figure 4.4B**, we found that the Hfq species was specific to that chain length size (**Fig. 4.4G**). Combined with our findings in **Figure 4.3E**, where we find that the presence of *ppk* is critical for efficient binding of Hfq, we start to construct a model in which polyphosphate, at basal conditions, mediates Hfq binding and silences of prophages and mobile elements (**Fig. 4.4H**). Thus, together, polyP and Hfq mediate a heterochromatin-like structure silences potentially toxic regions across the cell.

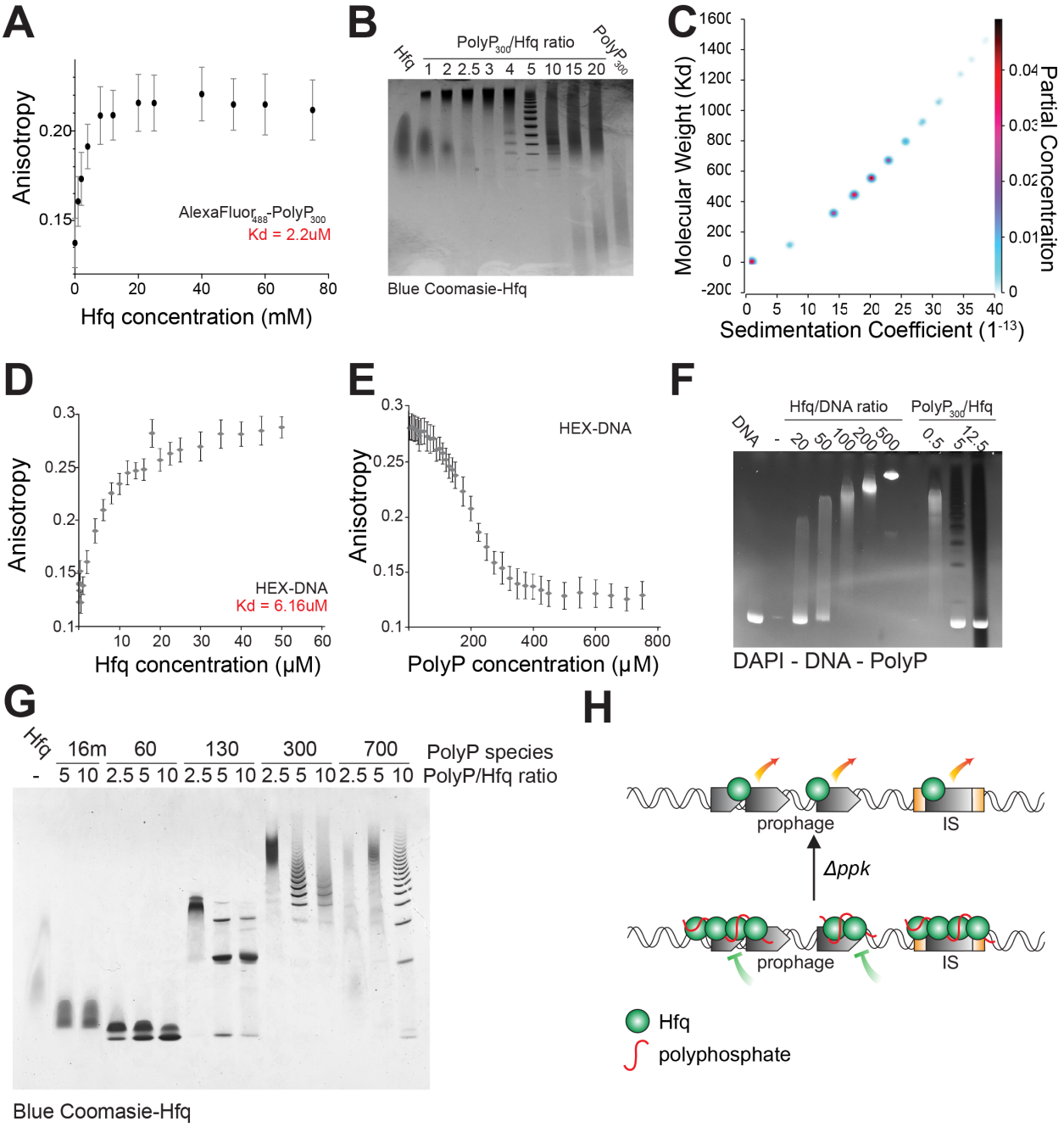

Figure 4.4: Polyphosphate facilitates distinct oligomeric species of Hfq hexamers. **(A)** Fluorescence anisotropy of titrated polyP$_{300}$ with Hfq exhibits a Kd of 2.2uM per Pi. **(B)** Gel shift assay with polyP$_{300}$ and Hfq at different ratios exhibits a distinct laddering pattern of Hfq. **(C)** Sedimentation Velocity Analytical Ultracentrifugation was performed to examine the binding species of Hfq mixed with polyP$_{300}$, and aligned with the findings in **(B)** that show polyP$_{300}$ supported the formation of high oligomeric species of Hfq. **(D)** Binding of HEX-DNA to Hfq was measured using the same approach as **(A)** and has a Kd of 6.61uM per DNA molecule. **(E)** Mixing of HEX-DNA, Hfq, and polyP$_{300}$ exhibits competitive binding for Hfq between polyphosphate and HEX-DNA. **(F)** The competition of Hfq binding was also detected in gel shift assays at different ratios of polyP$_{300}$ and Hfq.

In **(E)** and **(F)**, Hfq is sequestered away from DNA as there is an increase in $polyP_{300}$. **(B)** Different polyP species were mixed at distinct ratios with Hfq, and exhibit discrete hexamer patterns at different chain lengths. **(H)** Model for polyphosphate, Hfq, and DNA interactions, where polyphosphate facilitates silencing at basal conditions.

Discussion

We have discovered a novel mechanism by which Hfq and polyphosphate interact to form heterochromatin-like complexes that silence toxic DNA in *E. coli*. Deletions of *hfq* and *ppk* result in a sensitivity in DNA damaging agents in a prophage dependent manner, and an induction of prophage genes. Assessment of cell physiology, where we find that deletions of *hfq* and *ppk* result in increases in chromosome occupancy, further suggests an epistatic relationship between *hfq* and *ppk*. Sensitivity can be rescued by either deleting prophages or overexpressing Hfq, suggesting the necessity of Hfq to provide the silencing machinery on toxic elements. Our ChIP-seq data reveals that the presence of polyphosphate facilitates Hfq's ability to bind and silence prophages and mobile elements. Coupled with our *in vitro* studies, polyphosphate promotes sequestration of Hfq molecules to DNA across the genome and improves Hfq's ability to impact gene expression. This is the first description of a mechanism by which heterochromatin-like domains silence foreign DNA in *E. coli.*

Materials and Methods

Bacterial strains and growth conditions.

All strains, plasmids, and oligonucleotides used in this study are listed in Table SX in the supplemental material. Null mutations in *E. coli* MG1655 [30] and MDS42 [18] were constructed as previously described [31–33]. *E. coli* and derivative strains were grown at 37°C in lysogenic broth (LB, Fisher) or in MOPS minimal medium (Teknova) supplemented with 0.2% glucose and 1.32 mM KH_2PO_4.

Mutagenesis assay

To detect the rate and spectrum of spontaneous mutations, cells resistant to d-cycloserine or rifampin were selected. For each strain, 4 independent cultures were diluted s into 20 MOPS glucose cultures at 10^3 cells and were grown with agitation to saturation at 37 ◦ C for 24h. Samples from each tube were then spread either on minimal plates containing D-cycloserine or LB agar containing rifampin, incubated respectively 30 h and overnight at 37 ◦ C and CFU were scored. The total number of cells in a tube was determined by spreading dilutions from four tubes onto nonselective plates. Dividing the number of mutations per tube by the average total number of cells in a tube gives the mutational rate [13,21].

Survival assay to mutagens

Survival assays were performed as described previously [13]. For cisplatin, MMC and Phleomycin, E. coli MG1655 and isogenic mutant strains were cultivated in MOPS-G medium until an OD_{600} of 0.5 was reached. Then, the bacteria were 10-fold serially diluted and plated onto M9-G plates containing various concentrations of various mutagen. For survival upon UV treatment, cells were grown in LB medium at 37°C until an OD_{600} of 0.5 was reached, washed in ice-cold water and UVC irradiated ($25J/m^2$) using a spectrolinker XL-1500 UV crosslinker. All radiation experiments were performed on ice. Following irradiation, cells were 10x serial diluted plates onto LB agar plates and incubated overnight at 37°C.

Flurorescent microscopy

All strains were imaged during exponential growth phase after immobilization on 1% agarose pads. Images were taken using a Zeiss Axiobserver.Z1 microscope equipped with an ORCA-Flash 4.0 complementary metaloxide semiconductor (CMOS) camera and filter set 00. Images were processed using the MicrobeJ suite for ImageJ.

Protein purification

Escherichia coli BL21 (DE3) carrying pET-21a-hfq were cultivated in LB ampicillin at 37°C until an OD_{600} of 0.5, induced with IPTG (100ul/l) and shifted at 22°C for 16h. Cells were lysed by sonication at 4°C in lysis buffer (25mM Tris, 300mM NaCl, 5 mg/ml DNase I, Benzonase nuclease (Merck), 100uM MgCl2, cOmplete, pH7.5), incubated first at room temperature for 30min to ensure DNA degradation, then in boiling water for 20min, finally in room temperature water for 15min. Lysates were spun at 30000g, 30min, 4°C and supernatant loaded on a 5ml HisTrap column (Sigma). Column was washed with 20ml of buffer A (25mM Tris, 300mM NaCl, pH 7.5), with buffer B (6M Guanidine HCl, 25mM Tris, 300 mM NaCl, pH7.5), with a gradient of guanidine 0 to 6M (25mM Tris, 300 mM NaCl, pH7.5), then with buffer A and protein were elute by Buffer C (300mM imidazole, 25mM Tris, 300mM NaCl, pH7.5). Fractions containing highly pure HFQ were pooled and incubated for 1h with MgCL2 and Benzonase nuclease before overnight dialysis in buffer A. After dilution of the sample to obtain a NaCl concentration of 150mM. Samples were load on HiTrap SP HP 5ml column (Sigma), washed with 5% of buffer D (25mM Tris, 100mM NaCl, pH7.5) and elute with a gradient 6 to 40% of buffer E (25mM Tris, 1.0 M NaCl, pH 7.5). Purified samples were pooled together in buffer A (25mM Tris, 300mM NaCl, pH 7.5), and stored at -80°C.

Gel shift assay

Hfq, PolyP and DNA were mixed at the indicated concentration in 20mM Hepes, 100mM NaCl, pH8. After adding 10% glycerol (final concentration) and loading dye (NEB), samples were run on a TBE gel on ice (100V 16h for PolyP-HFQ interaction; 100V for 4h for HFQ-DNA). DNA and PolyP were stained by DAPI according to Smith and Morrisey, 2007. Briefly gel was incubated in fixative solution (25% methanol, 5% glycerol 50mM Tris, pH11) containing DAPI (2ug/ml) for 30min, de-stained in fixative solution for 1hr. Gels were then exposed to UV-light for 5min then image with UV-light. Background, DNA, and RNA will remain fluorescent, while polyP will appear as black. Proteins were visualized by classic Coomassie blue (Brunelle and green, 2014) or silver nitrate staining (Thermo) according to the manufacturer protocol.

Fluorescent anisotropy

To determine the dissociation binding constant K_d for Hfq and polyP or DNA, fluorescence anisotropy was performed. In a 1 ml cuvette at 37 °C, either 10 uM labelled polyP or 25 nM labelled dsDNA in 20mM Hepes, 100mM NaCl, pH8, was titrated with a stock (0.5 mM) of Hfq also in 20mM Hepes, 100mM NaCl, pH8. Anisotropy was recorded with a Cary Eclipse Spectrofluorometer (Agilent) using an excitation of 640 nm and an emission of 675 nm when monitoring polyP300-AF647 and using an excitation of 535 nm and an emission of 556 nm for HEX-DNA. Competition binding for Hfq was monitored by fluorescence anisotropy conducted in a similar manner by titrating labelled DNA and unlabeled polyP.

PolyP and DNA labeling

300 Pi chain length polyP was labeled with Alexa Fluor 647 as described in (Lempart et al. 2019). Briefly, $polyP_{300}$ was incubated with Alexa Fluor 647 cadaverine (Life Technologies) and 1-ethyl-3- (3-dimethylaminopropyl) carbodiimide (EDAC) (Invitrogen) at 60°C for 1 h. The reaction was stopped on ice and labeled polyP300-AF647 was separated from free dye and un- labeled polyP using a NAP-5 column (GE Healthcare) equilibrated with 40 mM KPi, pH 7.5. The concentration of polyP was determined via a toluidine blue assay (Mullan et al, 2002) and the 530/630 nm ratio defined the fraction of labelled polyP obtained (Classically between 16.5 to 33% considering one or both extremities labelled). HEX-labelled dsDNA fragments were obtained by PCR using a 5' HEX labelled oligo (Sigma) (**Table 4.3**).

Sedimentation Velocity Analytical Ultracentrifugation (SV-AUC)

Analytical ultracentrifugation was used to determine the behavior of polyphosphate bound to HFQ. The experiment was performed by loading 420 µl of sample into epon-charcoal 2-channel centerpieces with 1.2 cm path-length in an An60Ti rotor in a Beckman Optima Xl-I analytical ultracentrifuge. Measurements were completed at 32,000 rpm for the Hfq-polyphosphate samples and at 48,000 rpm for HFQ alone at 280 nm in intensity mode. All SV-AUC data were analyzed using UltraScan 4 software,

version 4.0 and fitting procedures were completed on XSEDE clusters at the Texas Advanced Computing Center (Lonestar, Stampede, Jetstrean) through the UltraScan Science Gateway (https://www.xsede.org/web/guest/gateways-listing) (Demeler et al., 2016). The partial specific volume (vbar) of Hfq was estimated within UltraScan III based on the protein sequence (Demeler et al., 2009). Raw intensity data were converted to pseudo-absorbance by using the intensity of the air above the meniscus as a reference and edited. Next, 2-dimensional sedimentation spectrum analysis (2DSA) was performed to subtract time-invariant noise and the meniscus was fit using ten points in a 0.05-cm range (Brookes et al., 2010). First arrays with a broad S range were fitted to account for possible large oligomeric states. Final arrays were fit using a broad S range from 1 – 50 for the complex and 1 -10 for Hfq, an f/f0 range of 1–4 with 64 grid points for each, 10 uniform grid repetitions and 400 simulation points. 2DSA was then repeated at the determined meniscus to fit radially invariant and time-invariant noise together using ten iterations. A second approach to fit the data was utilized by fitting a parametric restrained grid to the data (PCSA) (Brookes et al., 2010). In the PCSA analysis the same S range was fitted. The root mean square derivation between the 2DSA-IT method and the PSCA-HL method was comparable low, so that both solutions describe the data well.

Electron microscopy

Cells grown to midlog phase in MOPS glucose were prefixed by 2.5% glutaraldehyde in 0.1 M sodium cacodylate (pH 7.2), and postfixed 1% OsO4 in 0.1 M sodium cacodylate (pH 7.2). Samples are then dehydrated through a series of washes with increasing concentration of acetone and embedded in an epoxy resin. The samples were then sliced by an ultramicrotome into thin sections. Samples were then applied to glow-discharged carbon-coated grids, stained with 2% uranyl acetate for 1 min, washed with a drop of distilled water, blotted, and air-dried. Images were taken at 80 kV on a TECNAI 10 transmission electron microscope with a Gatan 967 slow-scan, cooled CCD camera.

Cell growth and harvest for RNA-sequencing and Hfq-PAmCherry ChIP. Cells (WT, WT-Hfq-PAmCherry, and Δppk-Hfq-PAmCherry; two biological replicates for each genotype) were streaked onto an LB plate from cryogenic storage and grown at 37°C. Individual colonies were used to inoculate MOPS-G medium, and cultures were incubated at 37°C and shaking at 200 rpm. After overnight growth, cells were back diluted to an OD600 of 0.003 and grown to a target OD600 of 0.2. Once the target OD was reached, 2.5 ml of the culture was mixed with 5 ml of RNA protect (Qiagen: Catalog #76506), vortexed, and incubated at room temperature for 5 minutes. The tubes were spun at 4°C for 10 min at 5,000 x g in a fixed-in a dry-ice ethanol bath and stored at -80°C. RNA isolation and sequence preparation is described below. The remaining culture was treated with 150 μg/mL of rifampicin for 10 min at 37°C and 200 rpm. The cultures were then mixed with concentrated formaldehyde / sodium phosphate, pH 7.4 buffer in falcon tubes to achieve a final volume of 10 mM NaPO4 and 1% v/v formaldehyde. Tubes were placed into a shaker for 5 min at room temperature. Excess glycine (final concentration: 0.333 M) was added to quench the crosslinker, and the samples were incubated for 10 min at room temperature with shaking. The tubes were then placed on ice for 10 min, and spun in a fixed-angle rotor for 4 min at 4°C at 5,000 x g. After discarding the supernatant, the respective pellets were washed twice with 10 ml of ice-cold phosphate buffered saline (PBS), dried, snap-frozen in a dry-ice ethanol bath and stored at -80°C.

Preparation of mCherry ChIP (follows the same protocol as the RNA polymerase ChIP procedure in Freddolino et al. 2020). Frozen pellets were resuspended in 600 μL of 1x IPOD lysis buffer (10mM Tris HCl pH 8.0; 50mM NaCl) with 1X protease inhibitors (cOmplete Mini, EDTA-free Protease Inhibitor: Roche) and 52.5 kU/mL of Ready-Lyse (Lucigen), and incubated at 30 C for 15 minutes. After 15 minutes, tubes were placed on ice and sonicated with a Branson digital sonicator with 25% power, 10 seconds on, 10 seconds off, for a total of 30 seconds on at 4C. Tubes were kept in an ice water bath for the entirety of the sonication process. The sonicated lysates were then placed on ice and digested with 6 μL of RNase A (Thermo Fisher), 6μL DNase I (Fisher: Cat #89835), 5.4μL 100mM MnCl2, and 4.5μL 100mM CaCl2, mixed by pipetting, and incubated on

ice for 30 minutes. The reactions were quenched with 50µL of 500 mM EDTA (pH 8.0), thus resulting in 50-200 bp fragments. The digested lysates were placed in a 4C centrifuged and spun for 10 minutes at top speed. To reduce potential noise, we pre-cleared the lysates by mixing with the beads that will be used to pull down protein-antibody complexes. The lysate (600 µL) was mixed 1:1 with 2X IP buffer (200mM Tris, pH 8.0, 600mM NaCl; 4% Triton X-100; 2X protease inhibitors) and 1X molecular grade BSA. We prepared NEB protein G beads (50 µL/ sample) by washing with 1mL of 1X IP buffer without protease inhibitors but including 1X molecular grade BSA and resuspended in the final volume that was started with (50 µL/ sample). Washed beads were distributed with lysates and incubated at 4C with rocking for two hours. Using a magnetic stand, beads were removed, and pre-cleared lysates were placed into fresh tubes.

As an input control, 50µL of pre-cleared lysates were mixed with 450µL of ChIP Elution Buffer (50mM Tris, pH 8.0, 10mM EDTA; 1% SDS) and placed at 65C for no more than 16 hours for crosslink reversal. DNA extraction will be described below. To the remainder of the lysate / IP buffer mixture, we added 5µL of mCherryChIP antibody (mCherry Monoclonal Antibody: ThermoFisher Cat #M11217) and incubated at 4C overnight on a tube rocker. The next morning, we prepared NEB protein G beads (50 µL/ sample) by washing with 1mL of 1X IP buffer without protease inhibitors and resuspended in the final volume that was started with (50 µL/ sample). Washed beads were distributed into antibody and lysate mixtures and incubated at 4C with rocking for two hours. The mixtures were then washed in the following series below with 1 mL washes for each buffer and mixing by inversion. After inversion, tubes were placed on a magnetic stand to remove wash, and new wash was added.

- 1X Wash buffer A (100mM Tris, pH 8.0; 250mM LiCl; 2% Triton X-100; 1mM EDTA)
- 1X Wash buffer B (100mM Tris, pH 8.0; 500mM NaCl; 1% Triton X-100; 0.1% sodium deoxycholate; 1mM EDTA)
- 1X Wash buffer C (10mM Tris, pH 8.0; 500mM NaCl; 1% Triton X-100; 1mM EDTA)

- 1X TE (10mM Tris, pH 8.0; 1mM EDTA)

After the lash wash, beads were resuspended in 500µL of ChIP Elution Buffer (recipe above) and incubated at 65C for 30min with vortexing every 5-10 minutes. The tubes were then placed on a magnetic stand, the supernatant placed in a fresh tube, placed at 65C for no more than 16 hours for crosslink reversal, and processed for DNA extraction as noted below. As a control, mCherry ChIP was performed on MG1655 strains lacking any tag, and the resulting signal from non-specific binding was subtracted from analysis.

DNA extraction after crosslink reversal (is the same as described in Freddolino et al. 2020): Following incubation at 65C, tubes were cooled and 100 μg of RNase A (Thermo-Fisher), incubated 2 hours at 37C, then added 200 μg of proteinase K (Fermentas) and incubated an additional 2 hours at 5C. Phenol-chloroform extraction and ethanol precipitation were performed as described in [cite: Ausubel FM. Current Protocols in Molecular Biology. John Wiley & Sons; 1998]. During the ethanol precipitation, Glycoblue (Ambion) was used as a co-precipitant, NaCl was the precipitating salt, and washes were performed with ice-cold 95% ethanol. Pellets were resuspended in 1XTEe and stored in DNA-Lobind tubes at -20C.

Preparation of next-generation sequencing (NGS) libraries (applies to IPOD-HR, RNA polymerase ChIP-seq, mCherry ChIP-seq) Samples were prepared using the NEBNext Ultra II Library Prep Kit (NEB #E7103) following manufacturer's instructions, with minor modifications: to purify cDNA, the Oligo clean and concentrator kit was used (Zymo #D4061). After the ligation of adapters, the DNA clean and concentrator kit – 5 was used (Zymo #D4014). Dual index primers for NEB were used in the sample preparation and the libraries were sequences on an Illumina NextSeq.

RNA isolation and sequencing preparation

RNA isolation

Frozen pellets (described in previous section) were thawed on ice and resuspended in 100 uL TE and treated with 177kU Ready-lyse lysozyme solution (Lucigen #R1804) and

0.2 mg proteinase K (Thermo Fisher Scientific #EO0492). The mixture was incubated for 10 minutes at room temperature and vortexed every 2 minutes. The RNA was then purified using the RNA Clean and Concentrator kit – 5 (Zymo #R1014) and treated with 5 units of Baseline-ZERO DNase (Epicentre #DB0715K) in the presence of RNAse inhibitor (NEB #M0314L) for 30 minutes at 37C. RNA purification was performed again using the RNA Clean and Concentrator kit – 5, flash frozen in a dry ice-ethanol bath, and stored at -80C. rRNA depletion was performed on the stored RNA using the bacterial rRNA depletion kit following manufacturer instructions (NEB #E7850L), with the only modification being the RNA purification step where we used the RNA Clean and Concentrator kit -5 instead of bead purification.

Sequencing preparation

Purified, rRNA depleted RNA was then put through the NEBNext Ultra Directional RNA Library Prep Kit for Illumina following manufacturer instructions (NEB #E7420L) for rRNA depleted RNA. We used random primers and considered the samples to be “intact” for the protocol specifications. Minor modifications to the protocol were the same as stated in the above NGS library preparation.

Western Blot Analysis

Cells were streaked from cryogenic storage onto LB plates at 37C. A single colony was inoculated into M9 Rich Defined Medium (RDM) in 0.04% glucose and grew overnight at 37C. In the morning, cells were back diluted to an OD600 of 0.003 and grew to an OD600 0.2-0.6 in M9 RDM 0.4% glucose at 37C. Once cells reach the desired OD600, 2mL of cells were pelleted, supernatant removed, flash frozen in a dry ice ethanol bath, and stored at -20C. Cell pellets were resuspended in 150uL of a mixture of Laemmli buffer, DTT, and PBS (1.5mL mix: 375uL 4X Laemmli buffer, 75uL 1M DTT, 1050uL 1X PBS). Resuspended cells were incubated at 99C for 10 minutes. After incubation, 10 uL was run on a stain-free gradient SDS PAGE gel (Bio-Rad Cat #4568086) at 175V for 40 minutes. The gel was then imaged, and subsequently transferred to an Immun-Blot PVDF Membrane (Bio-Red Cat #1620174) for one hour at 60V at 4C. After transfer, the

membrane was imaged to get whole protein abundance. The membrane was soaked in 3% milk in TBST for 30 minutes, milk was removed, and 10mL was then added with 1uL of mCherry antibody (ThermoFisher Cat #M11217) overnight rocking at 4C. The next morning, the milk antibody mixture was removed, and the membrane was washed three times with 3% milk in TBST. 10mL of 3% millk in TBST was added with 2uL of secondary antibody (Goat anti Rat IgG HRP: ThermoFisher Cat #31470) and incubated at room temperature for one hour. Milk was removed, membrane was rinsed with water twice. TBST was added at room temperature for 10 min and removed three times. The washed membrane was then stained with ECL Western Blotting substrate (ThermoFisher Cat #32209) and imaged.

Data visualization and analysis tools

We utilized numpy[35], R version 3.6.3[36,37], tidyverse[38], and ggplot2 for high through-put data analysis and visualization[39].

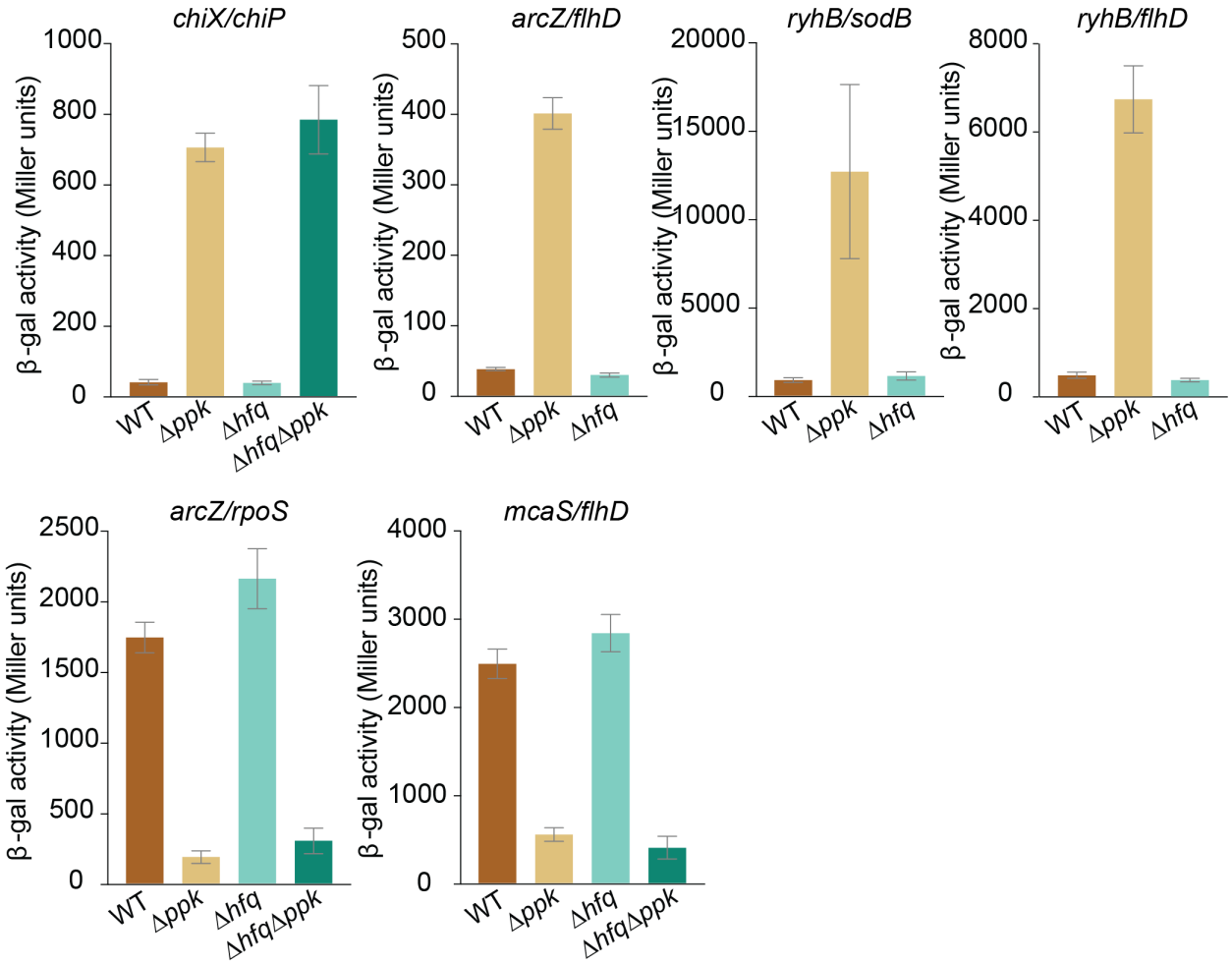

Figure S4.1: Hfq RNA chaperone targets are not impacted by loss of *ppk*.

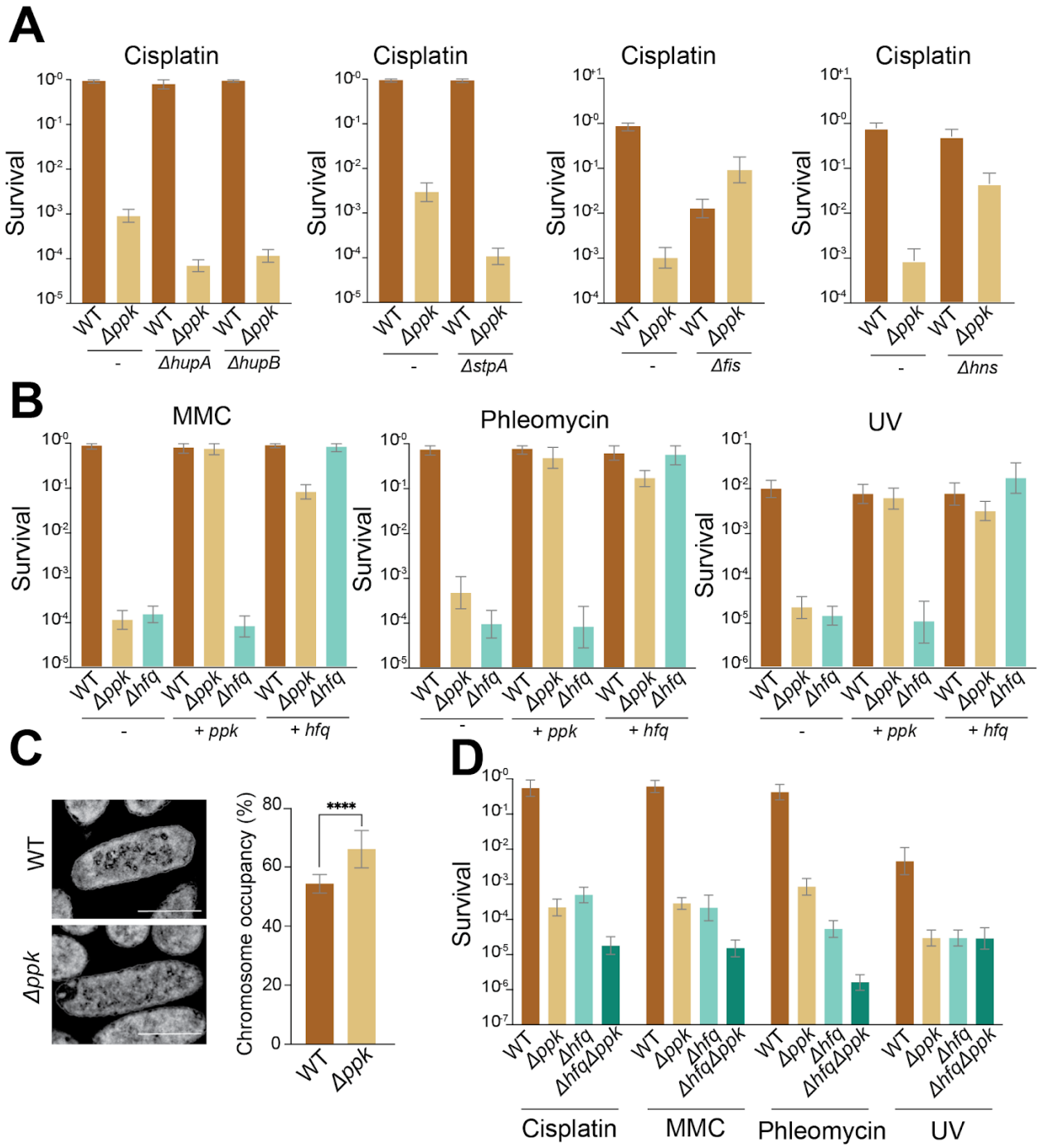

Figure S4.2: DNA damage epistatic response is specific to *ppk* and *hfq*. (A) Cisplatin sensitivity with *ppk* in combination with deletion of other highly abundant nucleoid associated proteins. **(B)** Plasmids containing *ppk* or *hfq* were introduced to *Δppk* and *Δhfq* cells exposed to DNA damaging agents. The same pattern of rescue for *Δppk* and *Δhfq* was observed in the presence of (+*hfq*) similarly to what is observed in **Figure 4.2F**. **(C)** Chromosomal occupancy measurements for WT and *Δppk* with representative images on the left and pooled calculations on the right, show a significant increase in chromosomal occupancy in *Δppk*. **(D)** Survival of WT, *Δppk, Δhfq,* and

ΔhfqΔppk across all DNA damaging agents does not exhibit a complete additive effect with *ΔhfqΔppk*, suggesting an epistatic response.

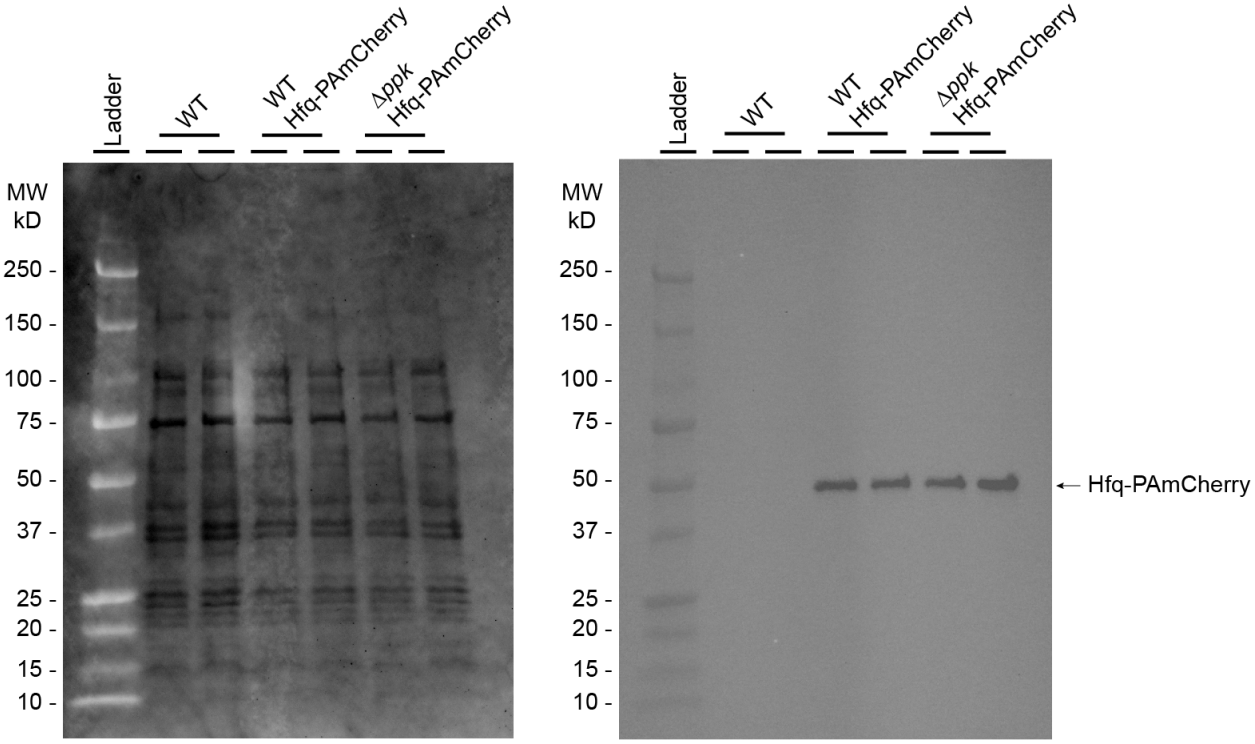

Figure S4.3: Loss of *ppk* does not change the amount of Hfq protein in the cell. WT, WT Hfq-PAmCherry, and *Δppk* Hfq-PAmCherry cells were collected at exponential phase of growth for western blot analysis. Two biological replicates were performed for each genotype. The left image shows the stain-free blot with all protein. The right shows the western blot of Hfq-PAmCherry with mCherry antibody, indicating the absence of signal without the PAmCherry tag, and the equivalent amount of Hfq-PAmCherry in both strains containing the tag.

Strain name	Marker (s)	Relevant Genotype	Source
BL21 (DE3)		F- ompT gal dcm lon hsdSB(rB- mB-) λ(DE3 [lacI lacUV5-T7 gene 1 ind1 sam7 nin5]) ykgD::cat+	invitrogen
MG1655		F-, λ-, rph-1 ilvG- rfb-50	Blattner et al. 1997
MJG224		MG1655 Δppk	Gray et al., 2014
MJG315		MG1655 Δppx	Gray et al., 2014
FB338		MG1655 Δhfq	this study
FB339		MG1655 ΔhfqΔppk	this study
FB340		MG1655 Δfis	this study
FB341		MG1655 ΔfisΔppk	this study
HA307		MG1655 Δhns	this study
HA308		MG1655 ΔhnsΔppk	this study
FB234		MG1655 ΔhupA	this study
FB235		MG1655 ΔhupAΔppk	this study
FB296		MG1655 ΔhupB	this study
FB297		MG1655 ΔhupBΔppk	this study
FB240		MG1655 ΔstpA	this study
FB241		MG1655 ΔstpAΔppk	this study
FB324		MDS42	Posfai et al., 2006
FB336		MDS42 Δppk	this study
HA144		MDS42 Δhfq	this study
FB261	amp	MG1655 pBAD18b	Gray et al., 2014
FB262	amp	MG1655 pBAD18b Δppk	Gray et al., 2014
FB264	amp	MG1655 pBAD18b-ppk	Gray et al., 2014
FB265	amp	MG1655 pBAD18b ppk Δppk	Gray et al., 2014
FB355	amp	MG1655 pBAD18b-hfq	this study
FB356	amp	MG1655 pBAD18b-hfq Δppk	this study
FB358	amp	MG1655 pBAD18b-hfq Δhfq	this study
FB345	kan	MG1655 pWSK129-ppkG688A (PPKD230N)	Rudat et al., 2018
FB343	kan	MG1655 pWSK129	Rudat et al., 2018
FB348	kan	MG1655 pWSK129-ppkG688A (PPKD230N) Δppk	this study
FB346	kan	MG1655 pWSK129 Δppk	this study
FB374	kan	MG1655 pWSK129-ppkG688A (PPKD230N) Δhfq	this study
FB372		MG1655 pWSK129 Δhfq	this study
HA303		MG1655 hfq-PAmcherry	this study
HA306		MG1655 hfq-PAmcherry Δppk	this study
FB361	amp	BL21(DE3) pET21a-hfq	this study
FB396	cm	MG1655 mal::lacIq ΔaraBAD lacI'-PBAD::cat-sacB::lacZ	Zhang et al., 2013
FB397		MG1655 mal::lacIq ΔaraBAD lacI'::PBAD-ompX-lacZ	Zhang et al., 2013
FB398	cm	MG1655 mal::lacIq ΔaraBAD lacI'::PBAD-ompX-lacZ Δhfq::cat-sacB	Zhang et al., 2013
FB399		MG1655 mal::lacIq ΔaraBAD lacI'::PBAD-rpoS-lacZ	Zhang et al., 2013
FB400	cm	MG1655 mal::lacIq ΔaraBAD lacI'::PBAD-rpoS-lacZ Δhfq::cat-sacB purA+	Zhang et al., 2013
FB401		MG1655 mal::lacIq ΔaraBAD lacI'::PBAD-flhD-lacZ	Zhang et al., 2013
FB402	cm	MG1655 mal::lacIq ΔaraBAD lacI'::PBAD-flhD-lacZ Δhfq::cat-sacB purA+	Zhang et al., 2013
FB403		MG1655 mal::lacIq ΔaraBAD lacI'::PBAD -chiP-lacZ	Zhang et al., 2013
FB404	cm	MG1655mal::lacIq ΔaraBAD lacI'::PBAD -chiP-lacZ Δhfq::cat-sacB purA+	Zhang et al., 2013
FB405		MG1655 mal::lacIq ΔaraBAD lacI':: PBAD -sodB-lacZ	Zhang et al., 2013
FB406	cm	MG1655 mal::lacIq ΔaraBAD lacI':: PBAD -sodB-lacZ Δhfq::cat-sacB purA+	Zhang et al., 2013
FB407		MG1655 mal::lacIq ΔaraBAD lacI'::PBAD-219sdhC-lacZ	Zhang et al., 2013
FB408	cm	MG1655 mal::lacIq ΔaraBAD lacI'::PBAD-219sdhC-lacZ Δhfq::cat-sacB purA+	Zhang et al., 2013
FB424	cm kan	MG1655 mal::lacIq ΔaraBAD lacI'-PBAD::cat-sacB::lacZ ppk::kan	this study
FB425	cm kan	MG1655 mal::lacIq ΔaraBAD lacI'::PBAD-ompX-lacZ ppk::kan	this study
FB426	cm kan	MG1655 mal::lacIq ΔaraBAD lacI'::PBAD-ompX-lacZ Δhfq::cat-sacB ppk::kan	this study
FB427	cm kan	MG1655 mal::lacIq ΔaraBAD lacI'::PBAD-rpoS-lacZ ppk::kan	this study
FB428	cm kan	MG1655 mal::lacIq ΔaraBAD lacI'::PBAD-rpoS-lacZ Δhfq::cat-sacB purA+ ppk::kan	this study
FB429	cm kan	MG1655 mal::lacIq ΔaraBAD lacI'::PBAD-flhD-lacZ ppk::kan	this study
FB430	cm kan	MG1655 mal::lacIq ΔaraBAD lacI'::PBAD-flhD-lacZ Δhfq::cat-sacB purA+ ppk::kan	this study
FB431	cm kan	MG1655 mal::lacIq ΔaraBAD lacI'::PBAD -chiP-lacZ ppk::kan	this study
FB432	cm kan	MG1655 mal::lacIq ΔaraBAD lacI'::PBAD -chiP-lacZ Δhfq::cat-sacB purA+ ppk::kan	this study
FB433	cm kan	MG1655 mal::lacIq ΔaraBAD lacI':: PBAD -sodB-lacZ ppk::kan	this study
FB434	cm kan	MG1655 mal::lacIq ΔaraBAD lacI':: PBAD -sodB-lacZ Δhfq::cat-sacB purA+ ppk::kan	this study
FB435	cm kan	MG1655 mal::lacIq ΔaraBAD lacI'::PBAD-219sdhC-lacZ ppk::kan	this study
FB436	cm kan	MG1655 mal::lacIq ΔaraBAD lacI'::PBAD-219sdhC-lacZ Δhfq::cat-sacB purA+ ppk::kan	this study
FB469	amp	FB396 pBR-plac	this study
FB482	amp kan	FB 396 pBR-plac pPPK10k	this study
FB470	amp	FB399 pBRplac-dsrA	this study
FB471	amp cm	FB400 pBRplac-dsrA	this study

FB472	amp	FB399	pBRplac-arcZ	this study
FB473	amp cm	FB400	pBRplac-arcZ	this study
FB474	amp	FB401	pBRplac-arcZ	this study
FB475	amp cm	FB402	pBRplac-arcZ	this study
FB476	amp	FB401	pBRplac-McaS	this study
FB477	amp cm	FB402	pBRplac-McaS	this study
FB478	amp	FB405	pBRplac-RyhB	this study
FB479	amp cm	FB406	pBRplac-RyhB	this study
FB480	amp	FB407	pBRplac-RyhB	this study
FB481	amp cm	FB408	pBRplac-ryhB	this study
FB483	amp	FB399	pBRplac-dsrA pPPK10k	this study
FB484	amp cm kan	FB400	pBRplac-dsrA pPPK10k	this study
FB485	amp kan	FB399	pBRplac-arcZ pPPK10k	this study
FB486	amp cm kan	FB400	pBRplac-arcZ pPPK10k	this study
FB487	amp kan	FB401	pBRplac-arcZ pPPK10k	this study
FB488	amp cm kan	FB402	pBRplac-arcZ pPPK10k	this study
FB489	amp kan	FB401	pBRplac-mcaS pPPK10k	this study
FB490	amp cm kan	FB402	pBRplac-mcaS pPPK10k	this study
FB491	amp kan	FB405	pBRplac-ryhB pPPK10k	this study
FB492	amp cm kan	FB406	pBRplac-ryhB pPPK10k	this study
FB493	amp kan	FB407	pBRplac-ryhB pPPK10k	this study
FB494	amp cm kan	FB408	pBRplac-ryhB pPPK10k	this study
FB495	amp kan	FB424	pBR-plac	this study
FB496	amp kan	FB427	pBRplac-dsrA	this study
FB497	amp cm kan	FB428	pBRplac-dsrA	this study
FB498	amp kan	FB427	pBRplac-arcZ	this study
FB499	amp cm kan	FB428	pBRplac-arcZ	this study
FB500	amp kan	FB429	pBRplac-arcZ	this study
FB501	amp cm kan	FB430	pBRplac-arcZ	this study
FB502	amp kan	FB429	pBRplac-mcaS	this study
FB503	amp cm kan	FB430	pBRplac-mcaS	this study
FB504	amp kan	FB433	pBRplac-ryhB	this study
FB505	amp cm kan	FB434	pBRplac-ryhB	this study
FB506	amp kan	FB435	pBRplac-ryhB	this study
FB507	amp cm kan	FB436	pBRplac-ryhB	this study

Table S4.1: Strains used in this study.

Name	Marker (s)	Description	Reference
pBAD18b	amp	cloning vector with PBAD arabinose-inducible promoter	Guzman et al., 1995
pKD46	amp	λ Red recombinase	Datsenko and Wanner, 2000
pCP20	amp	Flp recombinase	Datsenko and Wanner, 2000
pKD3	cat	cat chloramphenicol resistance casette donor	Datsenko and Wanner, 2000
pKD4	Kan	Kan Kanamycin resistance casette donor	Datsenko and Wanner, 2000
pet21a	amp	IPTG inducible vector for protein purification	Novagen
pBAD18b-hfq	amp	hfq arabinose inducible expressing vector	this study
pBAD18b-ppk	amp	ppk arabinose inducible expression vector	Gray et al., 2014
pWSK129	kan	cloning vector	Rudat et al., 2018
pWSK129-ppkG688A (pPPK10K)	kan	ppkG688A (PPKD230N) under the native ppk promoter	Rudat et al., 2018
pBR-plac	amp	IPTG inducible empty expression vector	Zhang et al., 2013
pBRplac-dsrA	amp	IPTG inducible dsrA sRNA expression vector	Zhang et al., 2013
pBRplac-arcZ	amp	IPTG inducible arcZ sRNA expression vector	Zhang et al., 2013
pBRplac-mcaS	amp	IPTG inducible mcaS sRNA expression vector	Zhang et al., 2013
pBRplac-ryhB	amp	IPTG inducible ryhB sRNA expression vector	Zhang et al., 2013
pet21a-hfq	amp	IPTG inducible hfq vector for protein purification	this study

Table S4.2: Plasmids used in this study.

ID	Description	Sequence
44	pBAD_for	ctgtttctccatacccgtt
45	pBAD_rev	GGCTGAAAATCTTCTCTCAT
245	ppk_lambdared_for	atgGGTCAGGAAAAGCTATACATCGAAAAAGAGCTCGTGTAGGCTGGAGCTGCTTC
246	ppk_lambdared_rev	ttaTTCAGGTTGTTCGAGTGATTTGATGTAGTCATACATATGAATATCCTCCTTA
243	ppk_Up	CGTAATTAAAGCGCCAGCTC
244	ppK_down	ATCTGCATGGCACCATCTAC
	hfq_pet21_for_NdeI	GTGATCATATGGCTAAGGGGCAATCTTTACAAG
	hfq_pet21_rev_XhoI	GatacCTCGAGTTATTCGGTTTCTTCGCTGTCCTGTTGC
156	hupA_up	CTGATTTGTCGTACCTGGAG
157	hupA_down	GACTACAGGCAGTGAGAAGC
162	hupB_up	TGTCTCGCTAAGTTAGATGG
163	hupB_down	CAATTGTCAGCCCACAAGAC
164	stpA_up	GGAATTAGCGAGCAGAGAGC
165	stpA_down	TACTGTTTGCAGGAATCAGC
235	hfq_up	GTATTACAGGTTGTTGGTGC
236	hfq_down	AGACCAGAGATTCAAACTCC
233	hfq_lambdered_for	atgGCTAAGGGGCAATCTTTACAAGATCCGTTCCTGGTGTAGGCTGGAGCTGCTTC
234	hfq_lambdered_rev	ttaTTCGGTTTCTTCGCTGTCCTGTTGCGCGGAAGTCATATGAATATCCTCCTTA
	hns_up	CTCAACAAACCACCCCAATA
	hns_down	TGGCGGGATTTTAAGCAAGT
	hns_lambdered_for	CCTCAACAAACCACCCCAATATAAGTTTGAGATTACTACAgtgtaggctggagctgcttc
	hns_lambdered_rev	GCCGCTGGCGGGATTTTAAGCAAGTGCAATCTACAAAAGAcatatgaatatcctccttag
241	fis_up	GCACATTCAACGCCATTGAG
242	fis_down	GGTCACTCCCTTTGTGACAC
257	fis_lambdared_for	atgTTCGAACAACGCGTAAATTCTGACGTACTGACCGTGTAGGCTGGAGCTGCTTC
258	fis_lambdared_rev	ttaGTTCATGCCGTATTTTTTCAATTTTTTACGCAGCATATGAATATCCTCCTTA
231	hfq_for_BamHI	tcGGATCCGCATATAAGGAAAAGAGAGA
232	hfq_rev_HindIII	tcAAGCTTCCGAAACCttaTTCGGTTTC
237	cycA_up	ACTCTGATGCCGGTAGGTTC
238	cycA_down	gcgccatccagcatgata
311	HEX_nohD_for	tcTCTAGAgaaagggatgctgaaattgag
309	HEX-ipex_for	tcGAATTCaggttgtgcttctaaaggaag
307	HEX-cynR-cynT_for	tcGAATTCggtgaagctgccatgttcag
308	cynR-cynT_rev	tcGGTACCgctgtttaaacaaggcttcc
310	ipex_rev	tcGGTACCcaccttccctaaagcactcg
312	nohD_rev	tcAAGCTTcattcacctcacggatgtag

Table S4.3: Primers used in this study.

References

1. Brown MRW, Kornberg A. Inorganic polyphosphate in the origin and survival of species. Proc Natl Acad Sci U S A. 2004;101: 16085–16087.
2. Chuang Y-M, Belchis DA, Karakousis PC. The polyphosphate kinase gene ppk2 is required for Mycobacterium tuberculosis inorganic polyphosphate regulation and virulence. MBio. 2013;4: e00039–13.
3. Wang L, Yan J, Wise MJ, Liu Q, Asenso J, Huang Y, et al. Distribution Patterns of Polyphosphate Metabolism Pathway and Its Relationships With Bacterial Durability and Virulence. Front Microbiol. 2018;9: 782.
4. Candon HL, Allan BJ, Fraley CD, Gaynor EC. Polyphosphate kinase 1 is a pathogenesis determinant in Campylobacter jejuni. J Bacteriol. 2007;189: 8099–8108.
5. Smith SA, Choi SH, Davis-Harrison R, Huyck J, Boettcher J, Rienstra CM, et al. Polyphosphate exerts differential effects on blood clotting, depending on polymer size. Blood. 2010;116: 4353–4359.
6. Travers RJ, Smith SA, Morrissey JH. Polyphosphate, platelets, and coagulation. Int J Lab Hematol. 2015;37 Suppl 1: 31–35.
7. Zhang C-M, Yamaguchi K, So M, Sasahara K, Ito T, Yamamoto S, et al. Possible mechanisms of polyphosphate-induced amyloid fibril formation of β2-microglobulin. Proc Natl Acad Sci U S A. 2019;116: 12833–12838.
8. Lempart J, Tse E, Lauer JA, Ivanova MI, Sutter A, Yoo N, et al. Mechanistic insights into the protective roles of polyphosphate against amyloid cytotoxicity. Life Sci Alliance. 2019;2. doi:10.26508/lsa.201900486
9. Smith JB, Dwyer SD, Smith L. Cadmium evokes inositol polyphosphate formation and calcium mobilization. Evidence for a cell surface receptor that cadmium stimulates and zinc antagonizes. J Biol Chem. 1989;264: 7115–7118.
10. Kornberg A. Inorganic polyphosphate: toward making a forgotten polymer unforgettable. J Bacteriol. 1995;177: 491–496.
11. Gray MJ, Jakob U. Oxidative stress protection by polyphosphate--new roles for an old player. Curr Opin Microbiol. 2015;24: 1–6.
12. Xie L, Jakob U. Inorganic polyphosphate, a multifunctional polyanionic protein scaffold. J Biol Chem. 2019;294: 2180–2190.
13. Beaufay F, Quarles E, Franz A, Katamanin O, Wholey W-Y, Jakob U. Polyphosphate Functions as an Iron Chelator and Fenton Reaction Inhibitor. MBio. 2020;11. doi:10.1128/mBio.01017-20
14. Foster PL. Stress-induced mutagenesis in bacteria. Crit Rev Biochem Mol Biol. 2007;42: 373–397.
15. Galhardo RS, Hastings PJ, Rosenberg SM. Mutation as a stress response and the regulation of evolvability. Crit Rev Biochem Mol Biol. 2007;42: 399–435.
16. Du Toit A. Phage induction in different contexts. Nature reviews. Microbiology. 2019. pp. 126–127.
17. Herold S, Siebert J, Huber A, Schmidt H. Global expression of prophage genes in Escherichia coli O157:H7 strain EDL933 in response to norfloxacin. Antimicrob Agents Chemother. 2005;49: 931–944.

18. Pósfai G, Plunkett G 3rd, Fehér T, Frisch D, Keil GM, Umenhoffer K, et al. Emergent properties of reduced-genome Escherichia coli. Science. 2006;312: 1044–1046.
19. Severinov K, Soushko M, Goldfarb A, Nikiforov V. RifR mutations in the beginning of the Escherichia coli rpoB gene. Mol Gen Genet. 1994;244: 120–126.
20. Kim C-G. Genetic Studies on the Structure and Function of the B Subunit of Escherichia Coli RNA Polymerase: RpoB Mutations Conferring Rifampicin Resistance. 1988.
21. Fehér T, Cseh B, Umenhoffer K, Karcagi I, Pósfai G. Characterization of cycA mutants of Escherichia coli. An assay for measuring in vivo mutation rates. Mutat Res. 2006;595: 184–190.
22. Rao NN, Gómez-García MR, Kornberg A. Inorganic polyphosphate: essential for growth and survival. Annu Rev Biochem. 2009;78: 605–647.
23. Murata K, Hagiwara S, Kimori Y, Kaneko Y. Ultrastructure of compacted DNA in cyanobacteria by high-voltage cryo-electron tomography. Sci Rep. 2016;6: 34934.
24. Wassarman KM, Repoila F, Rosenow C, Storz G, Gottesman S. Identification of novel small RNAs using comparative genomics and microarrays. Genes Dev. 2001;15: 1637–1651.
25. Moll I, Leitsch D, Steinhauser T, Bläsi U. RNA chaperone activity of the Sm-like Hfq protein. EMBO Rep. 2003;4: 284–289.
26. Malabirade A, Partouche D, El Hamoui O, Turbant F, Geinguenaud F, Recouvreux P, et al. Revised role for Hfq bacterial regulator on DNA topology. Sci Rep. 2018;8: 16792.
27. Orans J, Kovach AR, Hoff KE, Horstmann NM, Brennan RG. Crystal structure of an Escherichia coli Hfq Core (residues 2–69)–DNA complex reveals multifunctional nucleic acid binding sites. Nucleic Acids Res. 2020;48: 3987–3997.
28. McQuail J, Switzer A, Burchell L, Wigneshweraraj S. The assembly of Hfq into foci-like structures in response to long-term nitrogen starvation in Escherichia coli. Cold Spring Harbor Laboratory. 2020. p. 2020.01.10.901611. doi:10.1101/2020.01.10.901611
29. Bondy-Chorney E, Abramchuk I, Nasser R, Holinier C, Denoncourt A, Baijal K, et al. A Broad Response to Intracellular Long-Chain Polyphosphate in Human Cells. Cell Rep. 2020;33: 108318.
30. Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, et al. The complete genome sequence of Escherichia coli K-12. Science. 1997;277: 1453–1462.
31. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, et al. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol. 2006;2: 2006.0008.
32. Datsenko KA, Wanner BL. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc Natl Acad Sci U S A. 2000;97: 6640–6645.

33. Silhavy TJ, Berman ML, Enquist LW. Experiments with Gene Fusions. Cold Spring Harbor Laboratory Press; 1984.
34. Freddolino PL, Goss TJ, Amemiya HM, Tavazoie S. Dynamic landscape of protein occupancy across the Escherichia coli chromosome. Cold Spring Harbor Laboratory. 2020. p. 2020.01.29.924811. doi:10.1101/2020.01.29.924811
35. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020;585: 357–362.
36. Verzani J. Getting Started with RStudio: An Integrated Development Environment for R. “O’Reilly Media, Inc.”; 2011.
37. The R Project for Statistical Computing. [cited 24 Jan 2021]. Available: https://www.R-project.org/
38. Tidyverse. [cited 24 Jan 2021]. Available: https://www.tidyverse.org/
39. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer Science & Business Media; 2009.

Chapter 5

Concluding Remarks and Future Research

Introduction

It is becoming increasingly clear that architectural proteins not only contribute to genome compaction, but gene regulation and organization across all domains of life. However in bacteria, architectural proteins, termed nucleoid associated proteins (NAPs) have been difficult to study, largely due to their promiscuous binding throughout the genome and many modes of transcriptional impact[1]. In this dissertation, I present a tool, *in vivo* protein occupancy display at high resolution (IPOD-HR), that is an antibody-free method to visualize changes in protein binding across the entire genome (Chapter 2 [2]). In *Escherichia coli* (*E. coli*), the utility of IPOD-HR is demonstrated in detecting changing of transcription factors, novel binding motifs, and the presence of extended protein occupancy domains (EPODs), areas of the genome with dense protein occupancy but are transcriptionally silent, similar to heterochromatin in eukaryotes. EPODs overlap a number of different metabolic pathways and annotated prophages, and have an enrichment of H-NS binding, a widely known NAP and silencer of the cell[2]. To understand the individual contribution of NAPs to EPODs, IPOD-HR was performed on NAP deletions, uncovering novel silencers of metabolic pathways and prophages (Chapter 3). I show that EPOD repression can be relieved under metabolic stress, and detect a transcriptional memory response at a specific operon (Chapter 3). This serves as the first indication that EPODs may be present at particular loci to facilitate a memory response, similar to what is seen in heterochromatin in eukaryotes. I identify novel xenogeneic silences, Hfq and Fis that are required for cell viability in a

prophage dependent manner (Chapter 3). Hfq, which is highly conserved across species [3–5], has only begun to be explored for its DNA binding capacity[6], and therefore I show some of the first findings of the role of Hfq in silencing particular prophages via binding to the chromosome (Chapter 3 & Chapter 4). Furthermore, I show a novel mechanism by which Hfq binding occurs through cooperative binding of a poly anion, polyphosphate (Chapter 4). Loss of *hfq* or *ppk* results in an induction of prophage genes and mobile elements, and sensities the cells to DNA damaging agents (Chapter 4). Hfq binding *in vivo* is impacted by the loss of *ppk*, resulting in a decrease in binding at prophage regions (Chapter 4). Together, these reveal a novel mechanism by which Hfq and polyphosphate participate in cooperative binding to silence toxic elements in the genome. Although originally identified in *E. coli,* I show that EPOD-like structures exist in the distantly related Gram-Positive Firmicute *Bacillus subtilis* and are both composed of NAPs and overlap metabolic pathways and annotated prophage regions (Chapter 3). Thus, EPODs may serve conserved functional roles linked to gene regulation and structure. My thesis serves as a strong foundation for understanding NAPs, genome organization, and gene regulation in bacteria. Here, I share proposed follow up methods and experiments to further elucidate bacterial heterochromatin-like domains.

Investigating the impact of methylases on EPOD maintenance.

I show in Chapter 3 the enrichment of Dam and Dcm methylation sites outside of silenced regions, implicating that these two methylases may play a role in establishing sites that remain accessible. DNA adenine methyltransferase (Dam) methylates almost all GATC sequences in *E. coli*, and plays roles in gene expression, DNA replication, mismatch repair, and transposon / mobile element transposition [7–9]. Similarly, DNA cytosine methyltransferase (Dcm), which mehtaltes sites of 5’C-MeC-T 3’, likewise impacts transposition, indicated by a loss in transposition upon *dcm* deletion [10]. It was previously detected that certain areas of the genome remain methylation resistant, perhaps linked to the presence of EPODs[11]. To test whether the presence of

methylases impact EPOD maintenance, I created knockout strains of *dam*, *dcm*, and both *dam dcm* in *E. coli* and performed IPOD-HR. Currently, data analysis and processing of these mutants are currently underway, but early evidence suggests that these mutants do in fact impact RNA polymerase accessibility and EPOD maintenance.

Elucidating the mechanism underlying transcriptional memory response in exotic carbon source exposure.

The findings presented in Chapter 3 reveal a potential functional aspect to EPODs: to promote transcriptional memory. However, the response detected in Chapter 3 could be the result of many different modes of memory (as mentioned in Chapter 3), such as the presence of residual proteins, local regulators that stay bound to the promoter, or even post translational modifications (PTMs). To better understand whether the presence of the EPOD leads to the memory effect, one could attempt to delete the existing EPOD that remains on *idnD* and perform the same set of memory experiments. The expectation would be that there will be no memory detected (no change in lag time) upon a second exposure of the exoctic carbon source. The difficulty of such an experiment is the pleiotropic effects NAP deletions have that lead to growth and replication defects across the cell. In the case of the EPOD that overlaps the *idnD* operon that is mediated by the presence of StpA and H-NS, there is a wide variety of effects by deleting both of these NAPs. I attempted to repeat the experiment in the double deletion of *stpA* and *hns*, and as expected, the strain was too sick to grow in the exotic carbon source even after two weeks (data not shown). Thus, we could try a single deletion of H-NS that may partially destabilize the EPOD, or use the experimental approaches in the next session to see if we can identify any changes in proteins upon the second exposure to the exotic carbon source that may give insight to the mechanisms underlying the response.

Identify the proteins and post-translational modifications that define the structure of EPODs.

EPODs have varying functional roles and NAP configurations, as assessed by loss of occupancy only upon deletion of specific NAP combinations (Chapter 3). In addition, the findings presented and previous RNA-seq datasets [12–14] support a compensatory mechanism in which NAPs maintain proper EPOD distribution in the presence of the loss of any single factor. Based on our NAP deletion IPOD-HR datasets, I have identified a number of EPOD subtypes based on NAP occupancy (Chapter 3). However, the exact composition of EPODs is unknown. Furthermore, I show a transcriptional memory response due to carbon source exposure that may be directly mediated by EPODs, but minimal changes in EPOD locations at the site, indicating that there may be other factors leading to the memory effect such as accessory proteins or post translational modifications (PTMs). PTMs have already been shown to play a role in regulation of global regulators in bacteria[15], further supporting the idea that complex gene regulation can be at play. It is likely that there are also accessory proteins, such as Hha that promotes bridged filaments with H-NS and StpA [16], that may not show dramatic changes in IPOD-HR due to lower abundance, or may be novel binding proteins all together. Therefore, I have begun to develop a method to pull down specific EPODs using a catalytically dead dCas9 [17]and send regions for un-targeted mass spectrometry (MS) analysis (**Fig. 5.1**). The analysis will provide insight into protein composition and their associated PTMs. The unbiased approach will allow us to identify key regulators of silencing. Follow up experiments described in **Figure 5.1** will functionally test the roles of PTMs and accessory proteins in EPOD maintenance. The comprehensive assessment of EPOD formation will allow us to define the mechanisms of maintenance of different EPOD subtypes.

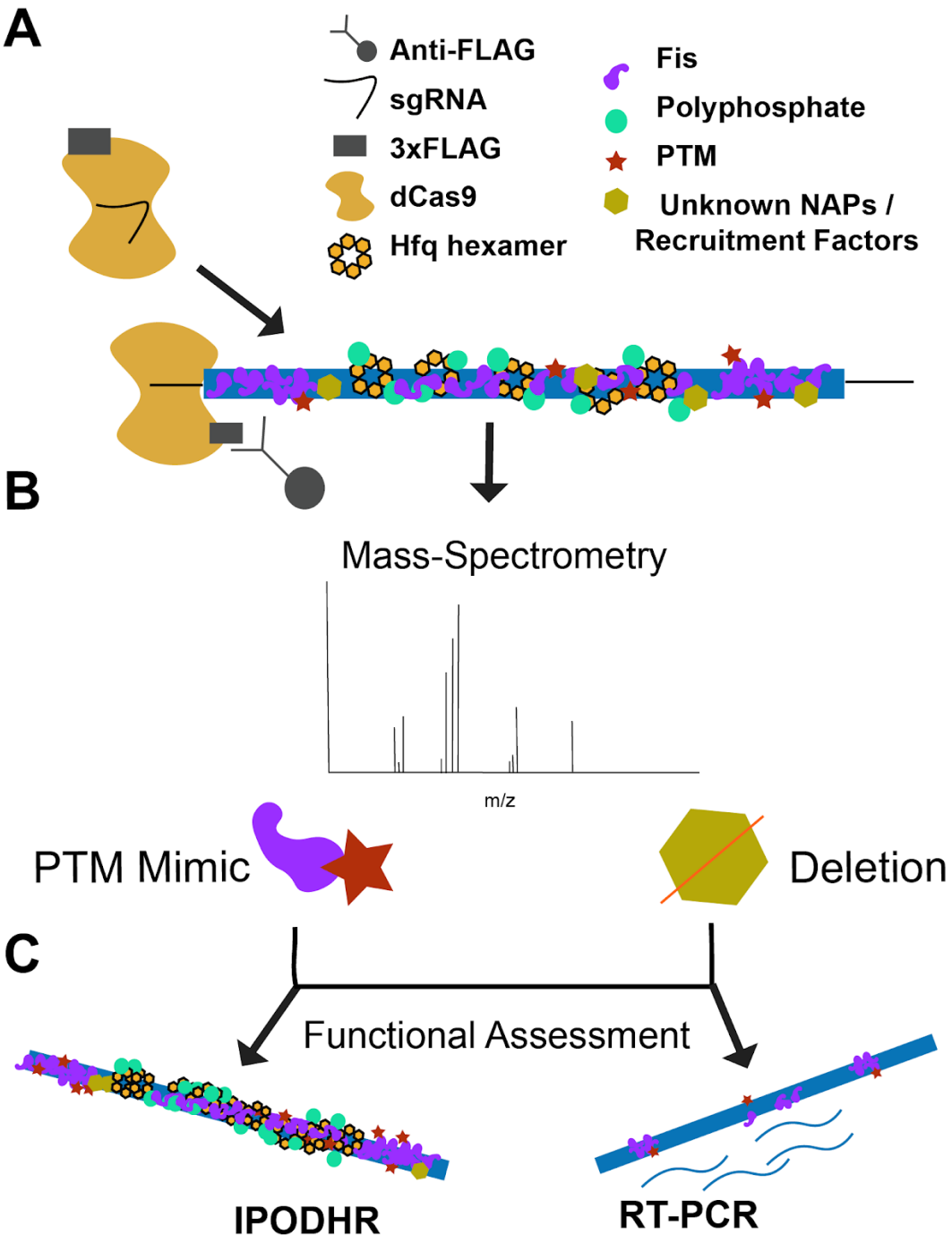

Figure 5.1. Workflow for assessing presence of PTMs and accessory factors. (A) A single guide RNA (sgRNA) dCas9+3xFLAG will be targeted to the region of interest allowing pull down of a specific region (using Anti-FLAG antibody to dCac9+3xFLAG) and its associated proteins. **(B)** Employ un-targeted mass spectrometry to identify bound NAPs, novel factors, and any associated post-translational modifications (PTM). **(C)** Investigation of the roles identified by either introducing PTM mimics or deletion of novel factors. In both instances, examine the impact of EPOD formation and expression of the locus using IPODHR and RT-PCR.

Investigate the mechanism of EPODs in the silencing of harmful genetic elements.

Across the genome, EPOD locations are highly robust and reproducible, indicating specific recruitment of proteins to particular sites. At the same time, recent work from our laboratory has also demonstrated that the chromosomal position itself also has profound effects on the ability of cells to transcribe different genomic regions[18]. It is unclear how EPOD recruitment occurs and if the surrounding genome context impacts silencing. To address this gap in knowledge, it will be important to select and shift EPODs into varying genetic and chromosomal contexts and assess whether normal protein occupancy is established in the transplanted EPODs. We propose that the genome context, such as transcriptional propensity[18], will substantially impact EPOD establishment. To study this, one would insert a Kanamycin (Kan) resistance marker next to each region using lambda red recombineering[19,20], in order to provide a selectable marker for subsequent transplant (**Fig. 5.2A**). Amplification of the region and adjacent marker using PCR on purified genomic DNA would obtain DNA corresponding to the EPOD sequence lacking any existing proteins from the fragment that would be inserted (**Fig. 5.2A**). The fragment would be selectively recomined (**Fig. 5.2B**), here denoted EPOD A, into five regions to determine which characteristics are necessary for silencing to occur. Importantly, a set of combinations in which EPOD A is the H-NS / StpA dependent EPOD and EPOD B is the Fis / Hfq dependent EPOD and vice versa would be performed: 1. EPOD B's location, 2. Into the center of EPOD B, 3. EPOD A's original location: For locations 1, 2, and 3, there are three predicted outcomes: no silencing occurring, silencing of region by original NAPs, and silencing of region by other NAPs. To further examine the universality of EPOD formation across *E. coli* strains, the Fis / Hfq dependent EPOD should also be introduced into UTI89, a pathogenic distantly related *E. coli* strain which lacks the prophage genes encompassed in these EPODs(**Fig. 5.2C**). Follow up for both approaches would be IPOD-HR, ChIP-seq, and RNA-seq of NAPs to understand binding dynamics and expression changes(**Fig. 5.2D**). This approach allows examining whether there are inherent properties of EPOD A that promote specific NAP binding, and whether there is influence

on EPOD B's location, presence of NAPs in promoting silencing, or specific genome features that promote recruitment of silencing proteins.

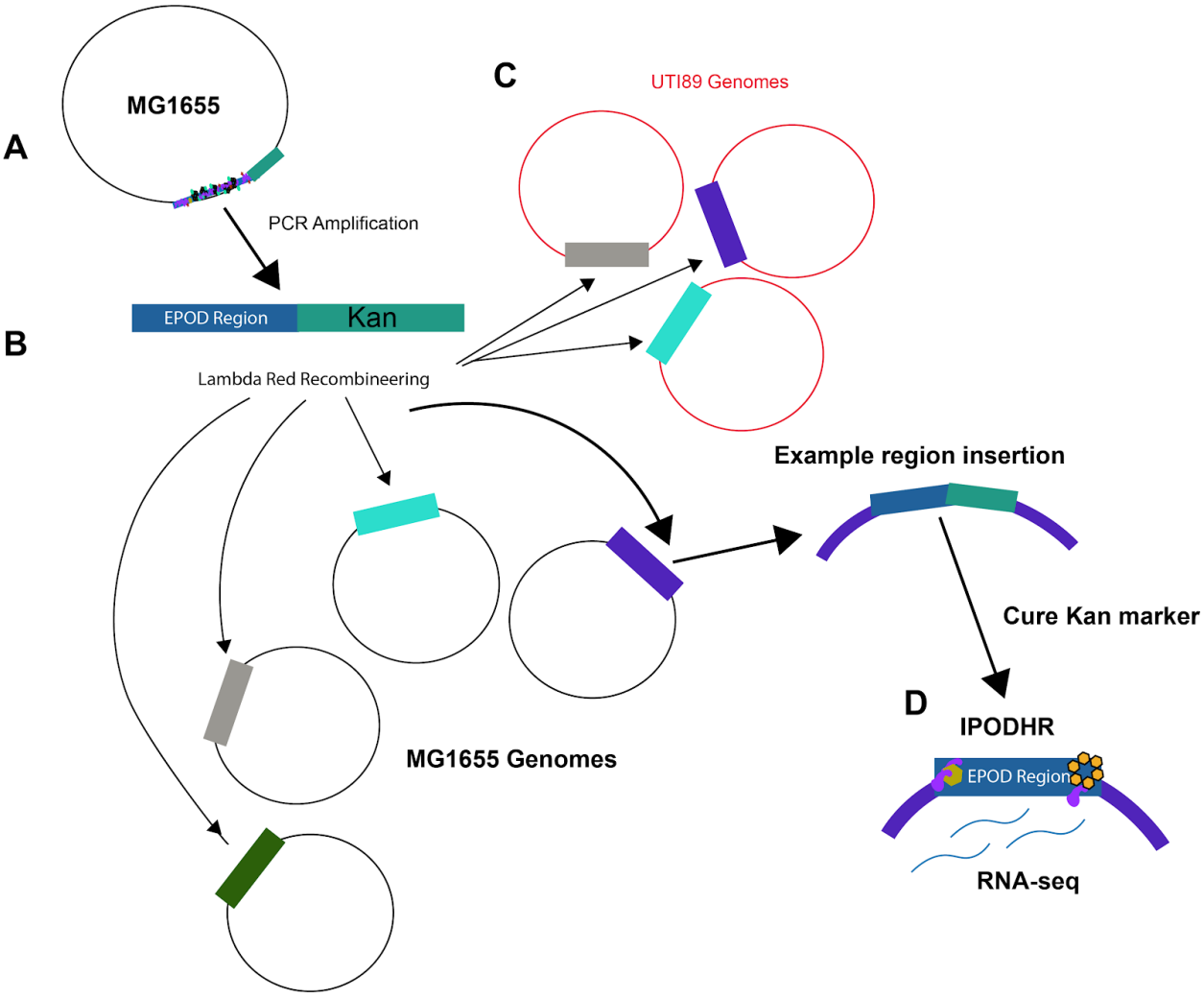

Figure 5.2. **Examining the impact of genome location in EPOD formation**. **A,B,D)** Experimental workflow. Example of EPOD A being inserted in the center of EPOD B. **C)** The process will be identical in UTI89.

Conclusion

In this thesis I explored the impact of genome organization has on gene regulation. In the appendix, I provide a resource for high-throughput data analysis on eukaryotic genomes, where filtering of highly repetitive regions is essential in order to properly process genomics datasets. This resource, the ENCODE Blacklist, further highlights the importance of understanding each step in high-throughput experiments, biology, and

downstream processing[21]. In Chapter 2, I describe a new method, *in vivo* protein occupancy display at high resolution (IPOD-HR) that enables genome wide profiling of protein occupancy across bacterial species without the need for an antibody. IPOD-HR resolves individual changes in transcription factors, discovery of new transcription factor motifs, and the refining of extended protein occupancy domains (EPODs) - highly protein occupied, large, and transcriptionally silent regions of the *E. coli* genome[2]. In Chapter 3, I connect specific NAPs to novel regulatory roles in EPODs. H-NS and StpA are main components of an EPOD overlapping the *idnDOTR* operon, and may mediate a transcriptional memory effect after exposure to an exotic carbon source. I discover that together Fis and Hfq serve essential roles in silencing prophages in *E. coli*. I also show that the silencing functionality of EPODs, mainly regulating metabolism and horizontally acquired DNA, is conserved across distant species, from *E. coli* to *B. subtilis*. Lastly, in Chapter 4, I show the mechanism by which Hfq silences prophages is mediated by a polyphosphate, a poly anion that sequesters Hfq to regions for effective silencing. This is the first example by which polyphosphate has been shown to interact with Hfq, and participate in silencing mobile elements and prophages.

While the key proteins are different across species and are not identical to eukaryotic histones, we find that, overall, organisms have similar strategies to organizing their genomes. All use non-specific nucleic acid binding proteins that facilitate dynamic changes across the DNA, allowing a cell to cope with their environment, pass information, and regulate their metabolism. Much of what we know about how structure impacts gene regulation is in eukaryotes, with many studies in bacteria needing the connection of *in vitro* and *in vivo* studies of gene organizers and regulators. The silencing of mobile elements and prophages via EPODs provides interesting hypotheses to whether heterochromatin-like domains in bacteria play a role in antibiotic resistance, as mobile elements have been linked to increased resistance[22,23]. As many of the NAPs are conserved[3–5], I predict that many of their functions would also be conserved. Further study in defining the maintenance and recruitment of proteins to these regions in bacteria will provide insights into gene regulation and will contribute to the development of novel antimicrobial interventions.

References

1. Shen BA, Landick R. Transcription of Bacterial Chromatin. J Mol Biol. 2019;431: 4040–4066.
2. Freddolino PL, Goss TJ, Amemiya HM, Tavazoie S. Dynamic landscape of protein occupancy across the Escherichia coli chromosome. Cold Spring Harbor Laboratory. 2020. p. 2020.01.29.924811. doi:10.1101/2020.01.29.924811
3. Brown L, Elliott T. Efficient translation of the RpoS sigma factor in Salmonella typhimurium requires host factor I, an RNA-binding protein encoded by the hfq gene. J Bacteriol. 1996;178: 3763–3770.
4. Salgado-Garrido J, Bragado-Nilsson E, Kandels-Lewis S, Séraphin B. Sm and Sm-like proteins assemble in two related complexes of deep evolutionary origin. EMBO J. 1999;18: 3451–3462.
5. Macvanin M, Adhya S. Architectural organization in E. coli nucleoid. Biochim Biophys Acta. 2012;1819: 830–835.
6. Orans J, Kovach AR, Hoff KE, Horstmann NM, Brennan RG. Crystal structure of an Escherichia coli Hfq Core (residues 2–69)–DNA complex reveals multifunctional nucleic acid binding sites. Nucleic Acids Res. 2020;48: 3987–3997.
7. Keseler IM, Mackie A, Santos-Zavaleta A, Billington R, Bonavides-Martínez C, Caspi R, et al. The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. Nucleic Acids Res. 2017;45: D543–D550.
8. Geier GE, Modrich P. Recognition sequence of the dam methylase of Escherichia coli K12 and mode of cleavage of Dpn I endonuclease. J Biol Chem. 1979;254: 1408–1413.
9. Wang MX, Church GM. A whole genome approach to in vivo DNA-protein interactions in E. coli. Nature. 1992;360: 606–610.
10. Yang MK, Ser SC, Lee CH. Involvement of E. coli dcm methylase in Tn3 transposition. Proc Natl Sci Counc Repub China B. 1989;13: 276–283.
11. Ringquist S, Smith CL. The Escherichia coli chromosome contains specific, unmethylated dam and dcm sites. Proc Natl Acad Sci U S A. 1992;89: 4539–4543.
12. Andrade JM, Dos Santos RF, Chelysheva I, Ignatova Z, Arraiano CM. The RNA-binding protein Hfq is important for ribosome biogenesis and affects translation fidelity. EMBO J. 2018;37. doi:10.15252/embj.201797631
13. Kroner GM, Wolfe MB, Freddolino PL. Lrp Regulates One-Third of the Genome via Direct, Cooperative, and Indirect Routes. J Bacteriol. 2019;201. doi:10.1128/JB.00411-18
14. Gawade P, Gunjal G, Sharma A, Ghosh P. Reconstruction of transcriptional regulatory networks of Fis and H-NS in Escherichia coli from genome-wide data analysis. Genomics. 2020;112: 1264–1272.
15. Zamora M, Ziegler CA, Freddolino PL, Wolfe AJ. A Thermosensitive, Phase-Variable Epigenetic Switch: Revisited. Microbiol Mol Biol Rev. 2020;84. doi:10.1128/MMBR.00030-17
16. Boudreau BA, Hron DR, Qin L, van der Valk RA, Kotlajich MV, Dame RT, et al. StpA and Hha stimulate pausing by RNA polymerase by promoting DNA-DNA bridging of H-NS filaments. Nucleic Acids Res. 2018;46: 5525–5546.

17. Tsui C, Inouye C, Levy M, Lu A, Florens L, Washburn MP, et al. dCas9-targeted locus-specific protein isolation method identifies histone gene regulators. Proc Natl Acad Sci U S A. 2018;115: E2734–E2741.
18. Scholz SA, Diao R, Wolfe MB, Fivenson EM, Lin XN, Freddolino PL. High-Resolution Mapping of the Escherichia coli Chromosome Reveals Positions of High and Low Transcription. Cell Syst. 2019;8: 212–225.e9.
19. Yu D, Ellis HM, Lee E-C, Jenkins NA, Copeland NG, Court DL. An efficient recombination system for chromosome engineering in Escherichia coli. Proc Natl Acad Sci U S A. 2000;97: 5978–5983.
20. Datsenko KA, Wanner BL. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc Natl Acad Sci U S A. 2000;97: 6640–6645.
21. Amemiya HM, Kundaje A, Boyle AP. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci Rep. 2019;9: 9354.
22. Brown-Jaque M, Calero-Cáceres W, Muniesa M. Transfer of antibiotic-resistance genes via phage-related mobile elements. Plasmid. 2015;79: 1–7.
23. Partridge SR, Kwong SM, Firth N, Jensen SO. Mobile Genetic Elements Associated with Antimicrobial Resistance. Clin Microbiol Rev. 2018;31. doi:10.1128/CMR.00088-17

Appendix

The ENCODE Blacklist: Identification of Problematic Regions of the Genome

Abstract

Functional genomics assays based on high-throughput sequencing greatly expand our ability to understand the genome. Here, we define the ENCODE blacklist- a comprehensive set of regions in the human, mouse, worm, and fly genomes that have anomalous, unstructured, or high signal in next-generation sequencing experiments independent of cell line or experiment. The removal of the ENCODE blacklist is an essential quality measure when analyzing functional genomics data.

Introduction

The coupling of high-throughput technology with classic genomics assays enables us to study genome-wide architecture and regulation. Assays using high-throughput sequencing as a read-out of a genomic signal rely on an accurate genomic annotation and mapping. Inconsistencies in the underlying annotation exist at regions where assembly has been difficult. For instance, repetitive regions may be collapsed or under-represented in the reference sequence relative to the actual underlying genomic sequence. Resulting analysis of these regions can lead to inaccurate interpretation, as there may be significant enrichment of signal because of amplification of noise [1,2]. Problematic regions such as these have generally been ignored and unfiltered because

The contents of this chapter were published in *Scientific Reports* by Haley M. Amemiya, Anshul Kundaje, and Alan P. Boyle. I wrote the manuscript and prepared the figures. A.P.B. and A.K. conceived of the work. A.P.B. developed the software application. All authors reviewed the manuscript.

they have not been found to affect the signal in the final analyses. However, in functional genomics assays such as chromatin immunoprecipitation followed by genome sequencing (ChIP-seq), accuracy in peak calling and downstream analyses is essential. Alignments in these problematic regions should be identified and filtered before application of any threshold, normalization, or peak calling as they can dramatically bias the results2.

The use of exclusive regions of “blacklists”, or regions where genome assembly results in erroneous signal, to remove signal-artifact regions in ChIP-seq experiments has been employed throughout the ENCODE project production phase [1,3,4]. The original ENCODE blacklist, termed the Duke Excluded Regions (DER), was manually curated on the Homo sapiens (human) genome assembly GRCh37 (hereafter referred to as hg19) to cover a large number of repeat elements in the genome, particularly rRNA, alpha satellites, and other simple repeats. This list was further updated, now referred to as ENCODE Data Analysis Center (DAC) blacklisted regions, to include regions of high signal that presumably represent unannotated repeats in the genome. The removal of these regions eliminated significant background noise that otherwise would have been thought to have been due to biological variation2. While this list was comprehensive, a significant amount of manual annotation was required to generate the final set of regions that would be laborious to apply to updated builds. The affected regions were broad, covering on average 45 kb with the largest being 1.4 Mb. Additionally, artifact regions are not human genome specific, and there was a need for identification of organism-specific regions.

Results

The generation of the ENCODE Blacklists. To generate blacklists in an objective and systematic manner, we developed an automatic procedure to flag regions that appear to have artifact signal. Regions are flagged using uniform criteria applied across a large number of samples. All ENCODE, mouse ENCODE, and modENCODE input ChIP-seq

samples (control data for ChIP-seq) were used for ENCODE (Homo sapiens: hg19 and the updated assembly GRCh38/hg38), mouse ENCODE (Mus musculus: mm9 and mm10), and modENCODE (Caenorhabditis elegans: ce10 and ce11, Drosophila melanogaster: dm3 and dm6) analyses, respectively. To identify regions for inspection, our method searches for regions that provide the signature of existing in multiple copies and are thus overrepresented in control “input” sequences. These “input” datasets were generated as controls for ChIP-seq experiments using randomly sheared DNA regions from non-immunoprecipitated chromatin. We examined all 1 kb windows with 100 bp overlap to identify such regions. Input samples are scored with input read depth and mappability, quantile normalized, and the median signal is selected (See methods). This defines a comprehensive and cell-type agnostic signal across the genome that is unaffected by high signal from a particular cell-line (eg. CNVs) or low signal due to differential processing of input data. Regions with read depths or multi-mapping read rates in the top 1% are considered likely artifacts (**Fig. A.3**). In all cases, the mitochondrial DNA and any reads mapping to these sequences are pre-filtered from analysis and are considered part of the blacklist.

A blacklist was built for the human, mouse, worm, and fly genomes using all reads from input samples. In each organism, only a small portion of the genome was flagged as containing artifact sequence signal (**Fig. A.1 and A.2A**). However, these regions were enriched for ChIP-seq reads in ChIP experiment for transcription factors (**Fig. A.1**) and are particularly enriched for reads and peaks from lower quality experiments. In fact, ENCODE uses this as a quality control metric with some experiments having up to 87% of reads falling into blacklisted regions [5]. In **Figure A.1** we show the distribution of all input reads mapped across chromosome 1, where reads mapping to blacklist regions are represented in red. The signal at blacklist regions are extremely high even though they account for a small fraction of the mappable chromosome (**Fig. A.1B**). For example, this represented 582 million of 2.5 billion uniquely aligning reads mapping to blacklisted regions in the human ENCODE ChIP-seq data in hg19. These findings emphasize the extreme nature of these artifact regions and highlight the importance of filtering these regions to avoid incorrect biological conclusions.

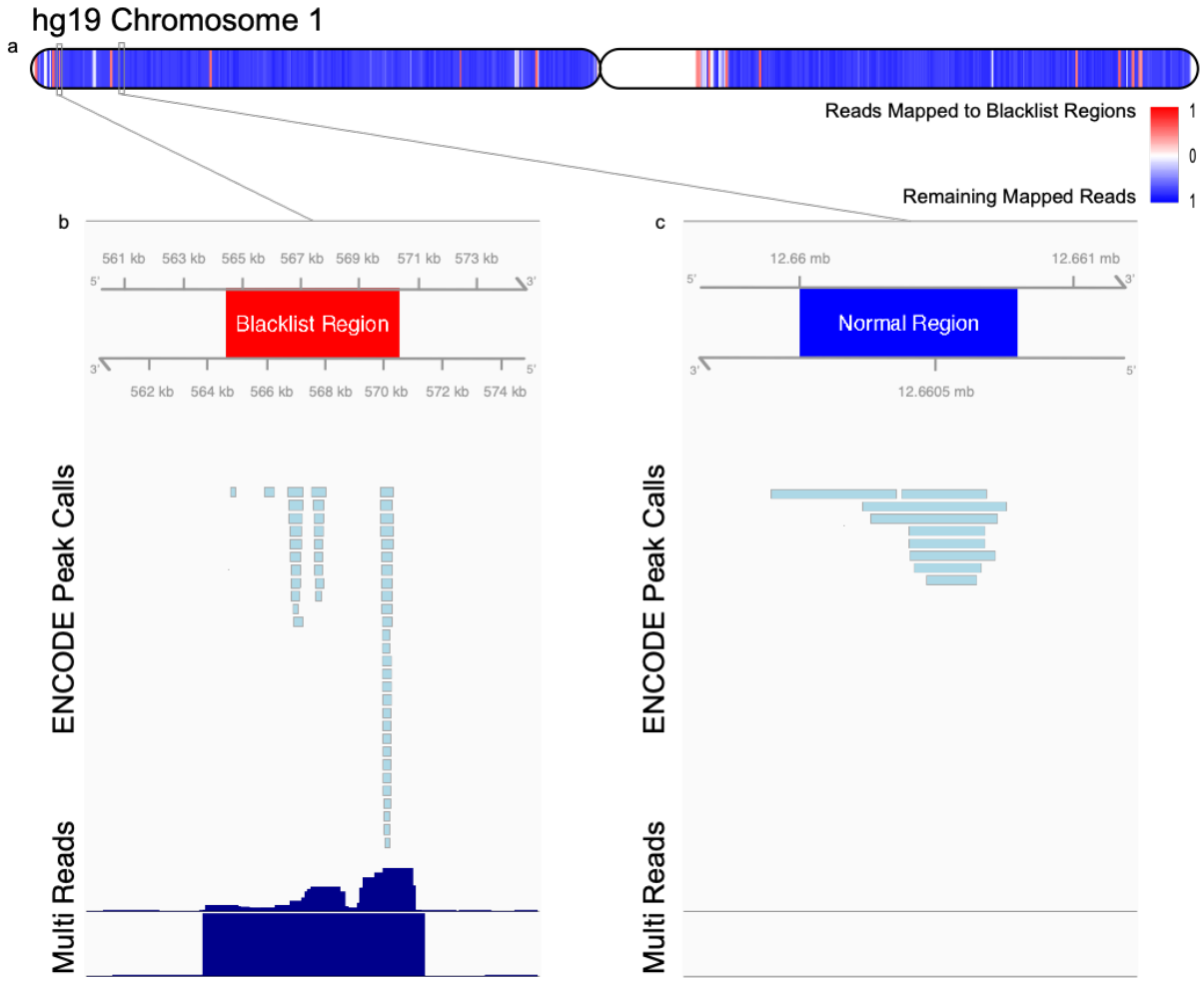

Figure A.1.Blacklist regions are tightly distributed across the chromosome and sequester high read mapping signals. (A) Distribution of mapped reads along human chromosome 1 in hg19. **(B)** An example blacklisted region on chromosome 1. Displayed are pre-filtered ENCODE ChIP-seq peak calls, quantile normalized median read signal (Reads), and quantile normalized median multimapped read signal (Multi). Axes are scaled for illustrative purposes and signal values are truncated at approximately 10-fold enrichment. Signal in these regions are up to 6400× background levels. **(C)** An example "normal" ENCODE ChIP-seq peak region on chromosome 1 selected as a region containing ChIP-seq peaks.

a

	Human	Mouse	Fly	Worm
N	2789	3038	492	122
Average size (bp)	4400	2800	2500	3200
Genomic coverage	0.40%	0.30%	0.70%	0.40%
ChIP Read Coverage	23%	11%	15%	6%

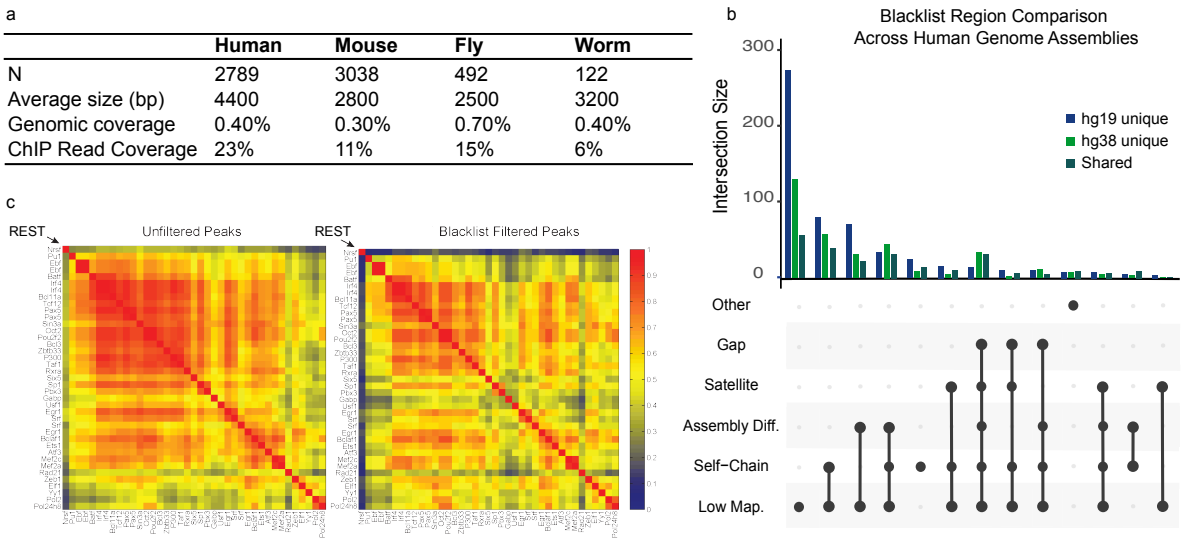

Figure A.2. Blacklist regions account for a significant portion of ChIP-seq reads, are driven by artifacts in genome assemblies, and removal of these regions is essential to removing noise in genomics assays. (A) The number of blacklisted regions across species with their average size, genomic coverage, and input datasets excluding assembly gaps used for hg38, mm10, dm6, and ce11 respectively. **(B)** An UpSet plot displaying the breakdown of uniquely annotated regions in hg19 and hg38, and the shared regions between them. Low-mappability (Low-Map.) regions account for the majority of unique regions in both hg19 and hg38. **(C)** Applying the blacklist to ChIP-seq peaks results in an overall reduced correlation and, in the highlighted example, results in a more biologically meaningful interpretation of the data.

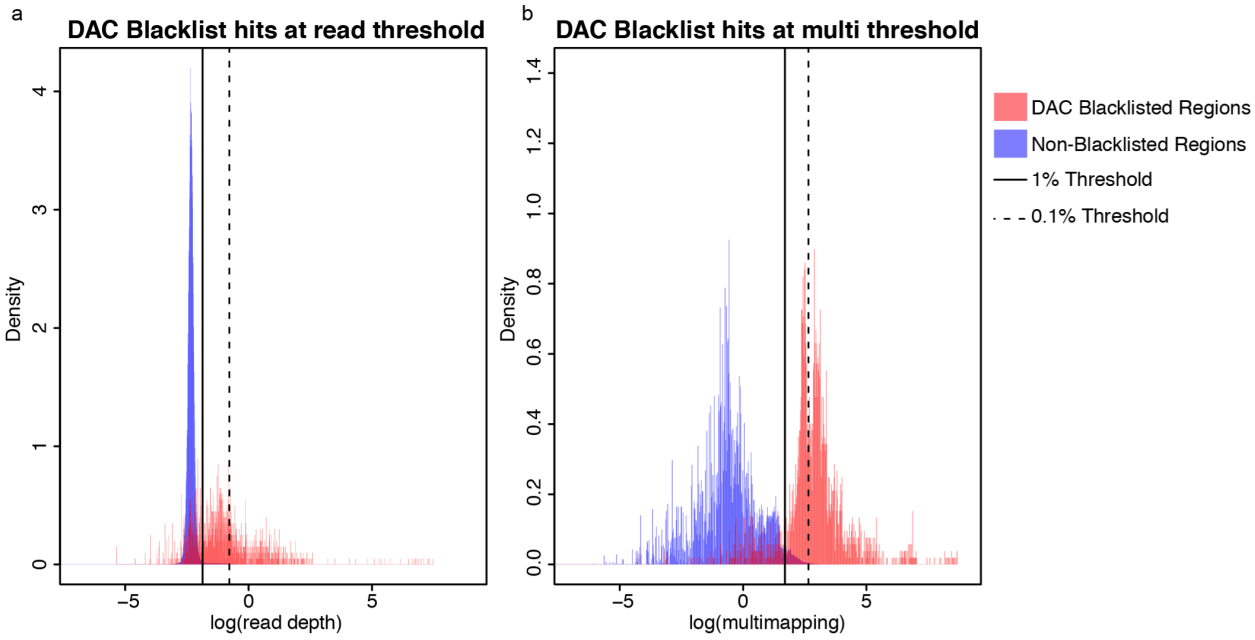

Figure A.3. Justification of thresholds for automated blacklist generation. The initial motivation behind the blacklist was to identify large artifact regions. These regions were envisioned as collapsed repeats in the genome that led to incredibly high numbers of reads. Early manual observations of these regions showed high levels of multimapping reads, high levels of reads, and multiple identical reads, and these manual observations generated what became the DAC blacklisted regions. Often these regions were at signal levels several orders of magnitude higher than the rest of the genome. As a result, in our automated method, we implemented a 1kb window with 100bp overlaps. In an attempt to not significantly overshoot the borders of these regions, this approach maintains a large enough region to identify the high signal in these often multi-kb regions. Here we have generated histograms of all 1kb windows from chromosome 1 and marked the 1% thresholds (black line) used to demonstrate the very long tail and conservative nature of this selection. Blue regions in this plot represent all 1kb windows from chromosome one not annotated by the manual blacklist and red regions represent all windows annotated by the DAC blacklist. Note that these histograms represent overall density in each class so that distinctions can be seen in the sets, but that the blue set represents 2,488,826 windows while the red set represents only 1,671. There is a very clear delineation between the 1kb windows manually identified as artifacts from the rest of the genome, and this transition occurs at the 1% mark. Therefore, this threshold was selected as being optimal for automated genome-wide identification of blacklist regions.

We investigated the underlying characteristics of our automated hg19 blacklisted regions and their agreement with previously published lists [6], which included our manually curated hg19 blacklist (DAC) (**Fig. A.4**). Though satellite repeats were used in the original ENCODE blacklists, they represent a small portion of the automated hg19

blacklist and are generally repeated in the genome annotation (**Fig. A.4C**). Additionally, they are not uniquely mapped with the algorithms used which prevents alignments to these regions. The automated hg19 and DAC blacklists detect the vast majority of regions flagged by a similar and complementary technique [6] (**Fig. A.4B**). Of the regions unique to the automated hg19 blacklist, all cover gaps in the genome assembly (**Fig. A.4C**), lending evidence that these regions of the genome are incomplete in the hg19 assembly. Indeed, a large number of these regions were patched in the next iteration of hg19 or dropped in the GRCh38 assembly (**Fig. A.4C**). Almost all of the flagged regions also included nuclear mitochondrial DNA segments (NUMTs, **Fig. A.4C**), a criterion that was overlooked in the initial manual blacklists. There are many mitochondrial genomes in comparison to the singular nuclear genome, leading to a high read depth of NUMTs. Additionally, NUMT sequences are scattered throughout the genome, contributing to overrepresentation of NUMTs in the input sequence. For these reasons, it is critical to include these sequences in the blacklist. A majority of the regions that were flagged by the DAC blacklist but missed in the automated hg19 blacklist were defined by repeatmasker class annotations as Satellite repeats (**Fig. A.4D**). While many of these repeat regions contain anomalous signal, those that were excluded from the automated hg19 blacklist do not show high signal and are uniquely mapped in hg19. None of the regions unique to the DAC blacklist were patched or removed in the new assembly.

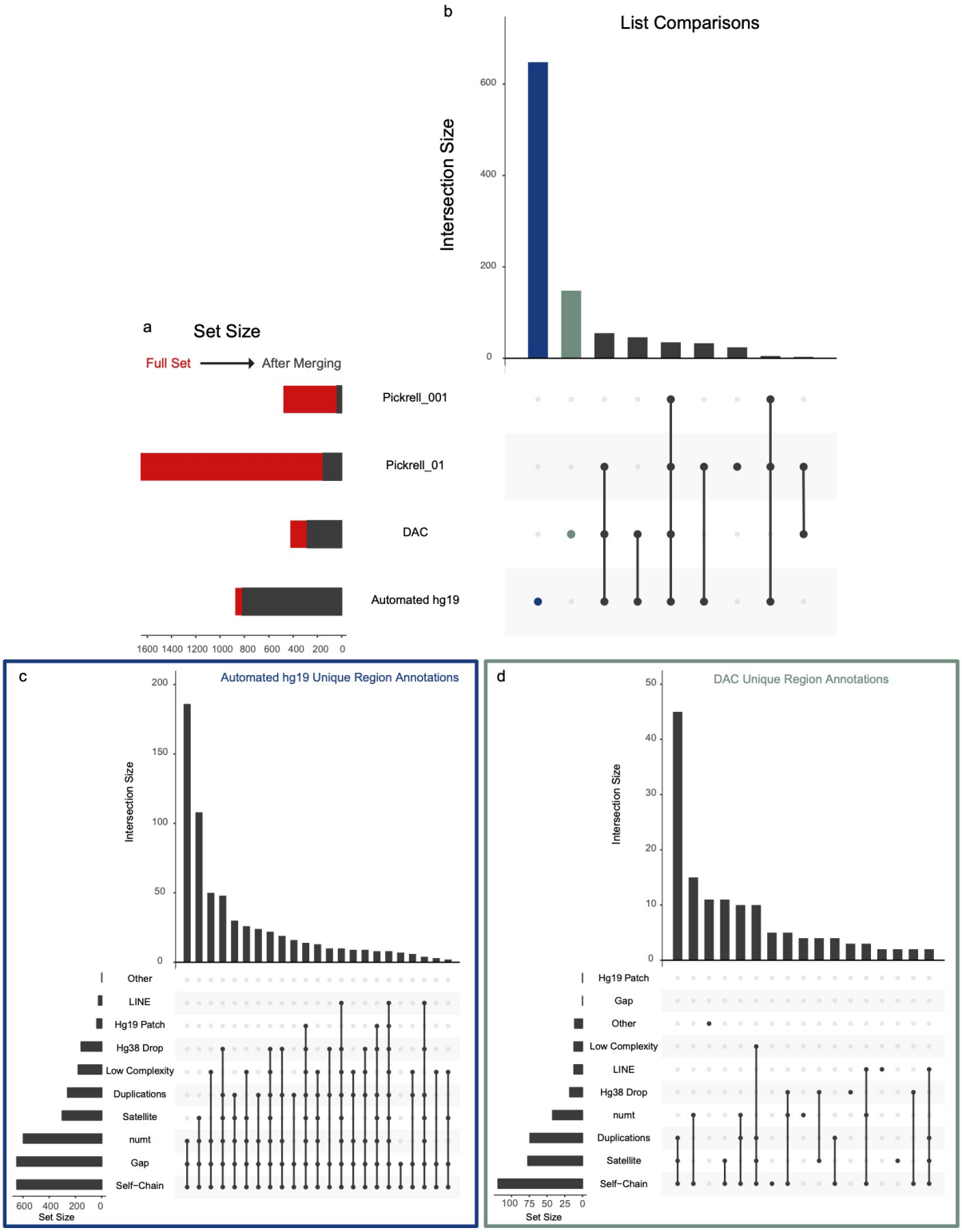

Figure A.4. Comparison across different "blacklists". In order to better understand the types of regions being annotated, we studied the similarities and differences across our automated and manual blacklists in hg19 as well as an analysis done by Pickrell et

al. to identify high-signal sites. **(A)** In order to compare across disparate genomic intervals, we first merged lists of regions. Following merging, many of the lists became shorter due to many small regions overlapping larger annotations. **(B)** An UpSet plot displays the number of unique regions when comparison across the sets. Notably, both the DAC and automated hg19 lists contained the most unique regions which we explored further. **(C)** The automated hg19 unique regions consist of assembly changes and gaps, as well as a large number of nuclear mitochondrial DNA segments. These indicate regions that were problematic in the assembly and were changed in the more updated build of the genome. Furthermore, nuclear copies of mtDNA (numts) were not considered in the initial manual annotation and because of their duplicative nature in the genome are likely to also have high signal. (**D)** The unique regions in the DAC manual blacklist are primarily annotated as satellite repeats. While these regions are repetitive areas, they are mappable in the genome and do not display an aberrant signal. These were likely included from the original DER manual list that was primarily based on satellite annotations and not aberrant signal.

We next sought to characterize the differences between regions blacklisted from the automated pipeline in hg19 compared to the blacklist from the hg38 genome assembly. Generally, the same classes of regions are enriched in both assemblies. Many regions do not overlap due to assembly differences such as the expanded centromere and satellite sequences that are features of the hg38 assembly as well as fixed/new gaps that vary in both builds (**Fig. A.2B**). A large portion of the differences occurs at low-mappability (Low Map.) annotations which represent short repetitive elements in the genome assembly that are poorly mappable and as a consequence do not map well between assemblies (**Fig. A.2B**). Overall, these differences lead to the conclusion that the major differences between blacklists are due to underlying changes in the sequence assemblies. Consequently, this lends to the hypothesis that the driving factor behind artifact regions in the genome are due to issues in the assembly rather than other factors.

Finally, to demonstrate the artificial correlation created by these peak regions, we performed a correlation analysis of the ENCODE peak regions in the human genome using a blacklist-filtered set as well as an unfiltered set of peaks (**Fig. A.2C**). In the unfiltered set of peaks, blacklist regions sequester a large portion of ChIP-seq reads, leading to an illusion of high correlation of these regions with others. After blacklist filtering, the correlation structure is more distinct. As a specific example, the correlation

of REST (a known repressor) auto correlates with other TFs (most of them activating) without the filtering of the blacklist. The removal of the blacklist regions removes spurious correlations seen with REST has essentially disappeared as would be expected given the known biology. We highlight the clear case of removing the noise in REST correlation, but the same standard holds true for the remaining factors in **Figure A.2C**. The ENCODE blacklists have been used to filter all of the ChIP-seq data from the ENCODE project and improvements in data from the application of the blacklist to these data are a key evaluation metric used by the consortium. For a complete list of artifact effects on peaks from all ChIP experiments used in ENCODE, we have provided a reference to the ENCODE quality control metric spreadsheet [5]. Furthermore, another detailed analysis of the detrimental effects of not excluding these artifact regions has been previously described2. Biological validations of the most robust signal regions will likely result in testing of these artifact regions, potentially resulting in incorrect biological conclusions. Therefore, identification of these regions and subsequent filtering lead to more accurate and stable results across experiments.

Discussion

The method implemented here requires a significant amount of input sequencing data from different sources in order to generate an accurate blacklist. For our analyses, we use all available ENCODE ChIP-seq input datasets to estimate the genomic regions that have these artifact properties, and the use of multiple cell-types is important to avoid blacklisting regions that are specific to a single cell-type or tissue. We also caution that the blacklists are specific to each genome assembly and a lift over from an old assembly is not meaningful or valid. Finally, those studying genes in unmappable regions of the genome will find their data filtered by the blacklist. These regions account for ~3% of human protein coding genes that have previously been shown to be unmappable using short-read technologies [7].

We present a resource of genomic regions that should be identified and either filtered from study or analyzed independently for better understanding as to their potential regulatory function. It is important to note that we do not propose a single blacklist that can encapsulate error defined across all NGS based assays. The presented blacklist shows high concordance between chromatin-based filtering (DNase/ATAC-DAC blacklist) and ChIP-seq input based filtering. This is not surprising given that the input DNA for ChIP-seq has been shown to be a proxy for lightly digested open chromatin assays [8]. However, the same criteria cannot be applied for whole genome sequencing (WGS) filtering and RNA-seq filtering. WGS filtering does not result in poorer annotations if there are higher read depths in regions, and therefore this method would be counterproductive to genome assembly. In the case of RNA-seq, more cell-type specific corrections for copy number are appropriate as there is virtually no overlap of coding regions with existing blacklist regions. The method presented is employed by the ENCODE project, as well as many other established analysis pipelines, and allows for a noise filtering on DNase-seq, ATAC-seq, and ChIP-seq datasets to help improve the accuracy of studies using these data. The removal of blacklists differs from the typical removal of signals from duplicate reads since these regions are problematic across different cell types and individual experiments. Ultimately, the removal of blacklists should be integrated within genomic assay analysis pipelines that incorporate high-throughput sequencing in order to assess biologically relevant and true signals.

Materials and Methods

Selection of input datasets. All data were acquired from the ENCODE Data Coordination Center. Using a previously published perl script (https://github.com/Boyle-Lab/ENCODE-API-Apps) [9], we queried the ENCODE DCC API for unfiltered bam files labeled “input” that were released for the correct genome assembly. In the case of humans, these bam files were merged based on ENCODE assigned donor accession numbers to collapse data by cell type or individual. This was performed using ‘samtools

sort' to first sort all samples, 'samtools merge' to merge, and finally 'samtools index' to generate a new index of the resulting bam files [10].

Generation of mappability data. The Umap tool was used to identify all positions on both strands of a target genome for which reads of a desired length starting at that position are uniquely mappable [11].

Building the blacklist. For each input dataset (or merged input for human) from ENCODE, the number of reads per mappable base and the number of multimapping reads per million reads is calculated for each bin of 1 kb with 100 bp overlap across all chromosomes. The values across bins are then quantile normalized and a standard value at the 50% quantile is selected to represent each bin. This threshold was selected to avoid high signal outliers from individual cell types (for example, from copy number variants) and to avoid low signal from failed or incorrectly labeled input datasets. The standard values across the genome are then flagged if they are in the top 0.1% of signal for either read depth or mappability. Neighboring regions are merged if they maintain a signal in the top 1% of all signal or if they have no signal due to no mappability in the genome and any flagged regions within 20 kb were combined. This generates contiguous regions of abnormal signal across the genome.

Data Availability. The blacklist software and called regions for multiple species are made available at https://github.com/Boyle-Lab/Blacklist/ and at the ENCODE DCC for human and mouse (https://www.encodeproject.org/annotations/ENCSR636HFF/).

References

1. ENCODE Project Consortium et al. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
2. Carroll, T. S., Liang, Z., Salama, R., Stark, R. & de Santiago, I. Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. Front. Genet. 5, 75 (2014).
3. Boyle, A. P. et al. Comparative analysis of regulatory information and circuits across distant species. Nature 512, 453–456 (2014).
4. Yue, F. et al. A comparative encyclopedia of DNA elements in the mouse genome. Nature 515, 355–364 (2014).
5. https://docs.google.com/spreadsheets/d/1G4SkqUMiGcUlvR6homc7RW33nSOf4mS9QYJifsd4qo0/.
6. Pickrell, J. K., Gaffney, D. J., Gilad, Y. & Pritchard, J. K. False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. Bioinformatics 27, 2144–2146 (2011).
7. Li, W. & Freudenberg, J. Characterizing regions in the human genome unmappable by next-generation-sequencing at the read length of 1000 bases. Comput Biol Chem 53, 108–117 (2014).
8. Auerbach, R. K. et al. Mapping accessible chromatin regions using Sono-Seq. Proc Natl Acad Sci USA 106, 14926–14931 (2009).
9. Diehl, A. G. & Boyle, A. P. Deciphering ENCODE. Trends Genet 32, 238–249 (2016).
10. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
11. Karimzadeh, M., Ernst, C., Kundaje, A. & Hoffman, M. M. Umap and Bismap: quantifying genome and methylome mappability.Nucleic Acids Research, gky677 (2018)