Systematic determination of a transcription factor/binding site functional interaction landscape

by

Katie Lynn Moravec

B.A. Interdisciplinary Studies
Amherst College, Amehrst MA (2008)

SUBMITTED TO THE GRADUATE PROGRAM IN MICROBIOLOGY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2017

Signature of author:_ Signature redacted
Katie Lynn Moravec
Graduate Program in Microbiology
15 June 2017

Certified by: Signature redacted
Michael T. Laub
Professor of Biology
Thesis supervisor

Certified by: Signature redacted
Kristala Prather
Associate Professor of Chemical Engineering
Co-Director, Microbiology Graduate Program

Systematic determination of a transcription factor/binding site functional interaction landscape

by

Katie Lynn Moravec

Submitted to the Graduate Program in Microbiology
on 15 June 2017 in partial fulfillment of the requirement for the degree of Doctor of Philosophy in Microbiology at the Massachusetts Institute of Technology

ABSTRACT

Cells require their genetic information to be expressed appropriately; the ability to modulate gene expression in a proper spatiotemporal response to external and internal signals is central to survival. Transcription factors are a major class of regulatory proteins that specifically bind DNA to modulate the expression of targeted genes. While they have been extensively studied, major questions remain about the protein-DNA interaction underlying this hub of regulation.
What binding site sequences functionally interact with a given regulator? How does the regulon sample from available functional sequences? How independent is each half of a two part binding site? How do mutations in the regulator impact the regulon?

Using PhoP, the regulator from the *E. coli* magnesium-responsive two-component system PhoPQ, I sought to address these questions. I identified the genomic binding locations for PhoP, verifying and expanding our knowledge of the PhoP regulon. Using two randomized libraries of over 65,000 variants each, I interrogated how changes in DNA sequence impact functional binding of PhoP. Comparing this with genomic binding data showed PhoP regulon members may avoid some sequences based on the dysfunctionality of their neighboring sequences. The functional library sequences reveal context dependence for each half-site and interaction within and across binding site halves. Finally, using an orthogonal PhoP mutant, I found that although these two proteins interacted with very few overlapping promiscuous sequences, there were many single mutations that would switch a promoter from interacting specifically with one protein to the other.

Thesis Supervisor: Michael T. Laub
Title: Professor of Biology

Acknowledgements

Thank you Michael Laub for support, encouragement, perspective, and an enthusiastic scientific outlook.

Thank you to my committee members Alan Grossman and Leona Samson for helping me work through roadblocks, focusing my exploratory project, and asking important questions. Thank you Greg Priebe for serving on my defense committee, and for all your mentorship and guidance as I bridged the gap between my undergraduate and graduate studies.

Thank you to the administrative staff of building 68 and especially to Bonnielee for making sure I always (at least eventually) turned everything in, and the Koch flow cytometry core staff, especially Mervelina.

Thank you to the Laub lab, past and present, for being excellent colleagues. To Anna and Emily for answering all of my questions and helping me define a project. To Chris and Josh for setting the room three wordplay bar high. To Barrett for all the technical help and larger perspective. To Diane, Chris, Christos, Emma, Kristina, and Orr for making the lab such an exciting scientific environment to join. To Mike, Leonor, and Anjana who joined the lab with me and were always up for an enlightening discussion, from challenging scientific areas to world events. To Conor and David for countless talks about the minutia of libraries and various exercise regiments. To Michele, Mathilde, and Monica for being excellent scientists and interesting people at the same time. To James, for keeping me in shape and commiserating about the end of the world. To Peter and Thúy-Lan for puns and garden advice. To Mary, Ben, Chantal, and Kevin—getting to know you all has convinced me the future of the lab is in good hands.

Thank you to my Microbiology classmates for making graduate school such an exciting and interesting time. Thank you to the labs of Building 68, especially the Sauer and Grossman labs for tolerating our Laub Lab antics and always being helpful and generous when I needed anything.

Finally, thank you to my friends and family for your support and efforts to keep me well-rounded these last few years. Thank you especially to my parents for everything (including becoming crazy dog people in the last couple years), my favorite sister Hannah for laughing at all the silliness no one else seems to find quite so funny, and Martin, for being there always.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction: Transcription factor function and specificity

Introduction

Life has adapted to thrive in and transform diverse and extreme environments: from arid deserts to deep ocean vents, arctic ice sheets to seasonal ponds, very little of our planet has not been invaded by living things. While we can easily see the plants and animals present in the world's habitats, it is microscopic bacteria that are the most widely distributed organisms, filling varied niches everywhere from the sea to your skin.

Like all life forms, bacteria owe their success across wide-ranging habitats to their ability to sense the environment and change their behavior accordingly. Transcription factors are a major class of proteins responsible for carrying out these changes. They bind DNA in the promoter regions of genes, and either increase or decrease the rate at which RNA transcripts are produced. Transcription factors are often tightly regulated—through protein modification, degradation, oligomerization, the transcription of their own RNA, or (often) some combination of these factors.

Although they have been extensively studied in the past 50 years, many questions remain about how transcription factors specifically interact with their target sequences. My work has focused on exhaustively identifying and characterizing the set functional sequences that can interact with the *E. coli* transcription factor PhoP. This has allowed me to answer questions including: of all possible 8-mers, which functionally bind PhoP to support gene activation? What genomic sequences are actually bound by PhoP, and how does this set compare to the complete set of functional sequences? What constraints shape the genomic distribution and usage of PhoP binding sites? How do nucleotides interact within and across PhoP binding sites in a given promoter? Before describing the experiments undertook to address these and other questions, I will first present an overview of transcription factors in *E. coli*, and what is currently known about how they specifically interact with their targets. Finally, after presenting and interpreting the data I have assembled during my thesis work, I will lay out some future directions for research related to the study of transcription factor-DNA specificity.

Transcription factor discovery

The first transcription factors described were repressors (such as the repressors of the lac operon or lambda phage) in the 1960s (Jacob and Monod 1961, Ptashne 1967). Around this same time, structural study of a eukaryotic DNA-binding protein suggested the presence of α-helices, and modeling predicted what would eventually be confirmed: many proteins interact in a sequence-specific manner using an α-helix in which certain amino acids contact the bases of the DNA via the major groove (Zubay and Doty 1959). Once DNA-binding repressors were identified, it was several years before transcriptional activators were even proposed. Activators were hypothesized in 1962 (Garen and Echols) and 1965 (Englesberg et al.), but it was several more years before transcription factors were understood to include both activating and repressing proteins. This was partly due to momentum in the study of repressors (and "apparent activators" that were repressors of repressors), and partly due to the more straightforward model of repression, which can be achieved simply by blocking the binding of RNA polymerase to a given promoter. When a loss of function mutation was introduced in what appeared to be a regulatory protein, it would sometimes decrease expression of its target. Although this evidence could be consistent with indirect activation, evidence of direct interaction between some proteins and RNA polymerase made it clear that some transcription factors are indeed activators. For example, direct interaction was established for DsdC with *in vitro* synthesis (Heincz and McFall 1978), CAP with eletrophoretic mobility shift assays (EMSAs) (Fried and Crothers 1983), and AraC with methylation and DNaseI protection assays (Lee et al. 1981). In some cases, researchers found their transcription factors acted as an activator at some promoters and a repressor at others. From this early work, the study of transcription factors has broadened to reveal examples of promoters controlled by multiple factors and gene networks that allow the integration of many signals—all of which result in finely-tuned regulation of transcriptional output specifically optimized for a given set of input conditions.

Transcription in *E. coli*

In *E. coli*, as in all bacteria, transcription is performed by a DNA-dependent RNA polymerase (RNAP). The core enzyme has four different catalytic subunits in the stoichiometry of $\beta\beta'\alpha_2\omega$. With the addition of a sigma subunit (σ) the entire assembly is called the holoenzyme. Although RNAP can initiate transcription *in vitro* without a σ subunit, it cannot find promoters without an associated σ factor (Burgess et al. 1969, Lee et al. 2012). Essential genes and many housekeeping genes require the most common σ factor, σ^{70}, but cells often employ alternate σ factors to efficiently shift the entire transcriptional output of the cell from one set of genes to another (Ishihama 2000, Gruber and Gross 2003).

E. coli promoters have two major shared features at the nucleotide level: the -10 and -35 sites. Each site has a consensus sequence of six nucleotides—and mutations in these regions can increase or decrease promoter strength. The conserved character and spacing of these sites reflects the first step of transcription initiation: specific binding of promoter elements by RNAP. Different regions of the σ factor interface with each promoter element; region two recognizes the -10, while region four recognizes the -35 (Campbell et al. 2002). The C-terminal domain of the α subunit (αCTD) can also directly contact the DNA minor groove at or upstream of position -40 (though due to long flexible linkers in the αCTD domain, this positioning is less strict) (Gourse et al. 2000). Although not studied at a large number of promoters, where it has been explored, it appears RNAP first interacts with more upstream elements and interacts with the -10 sequence last (Feklistov and Darst 2011, Saecker et al. 2011). While these binding elements can be combinatorially tuned to result in a wide range of promoter strengths, transcription factors allow a promoter of one set strength to be responsive to a signal.

RNAP recognizes the core elements of a promoter in a closed conformation, and the DNA of the promoter is double-helical and unmelted. This less-stable complex must undergo isomerization and the promoter DNA must begin to melt into a separated bubble at the -10 element for transcription initiation to proceed. σ mediates strand separation in the transition to the open complex, the more stable isomer competent for transcription initiation (Saecker et al. 2011). While regulators have been found that impact the isomerization step of transcription such as NtrC in *Salmonella typhimurium* (Popham et al. 1989), DksA in *E. coli* (Paul et al. 2004) and GcrA in *Caulobacter crescentus* (Haakonsen et al. 2015), most transcription factors act at the earlier step of RNAP recruitment.

Mechanisms of transcription factor activation

Changing the promoter DNA conformation is one way transcription factors have been shown to regulate downstream genes. This mechanism is employed by MerR family proteins, which themselves change conformation depending on the binding of specific metal ions (Brown et al. 2003). Without bound transcription factor, the conformation of the promoter DNA results in sub-optimal spacing of the canonical promoter elements, so RNAP can find upstream promoter elements, but it cannot proceed to identify the -10 element and initiate transcription. With the distortion introduced by the sequence-specific binding of proteins such as MerR in *Bacillus subtilis* or GrlA in enterohaemorrhagic *E. coli*, the downstream promoter elements can be successfully identified by RNAP and transcription initiation can proceed (Reyes-Caballero et al. 2011, Islam et al. 2011).

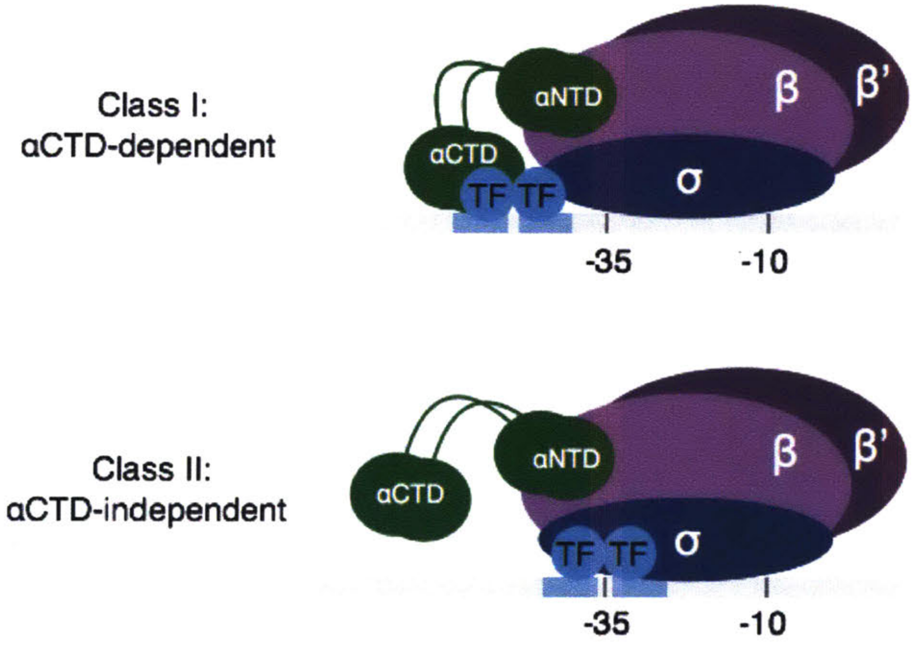
While transcription activation can be achieved simply by distorting promoter DNA, the majority of activators make direct contact with RNAP (Lee et al. 2012). When acting as a bridge between promoter DNA and RNAP, activators do not appear to induce conformational changes in RNAP, but rather to correctly position the complex at DNA elements that may be unrecognized without the regulator (Benoff et al. 2002, Jain et al. 2004). Although it may be attractive to imagine RNAP conformational changes as a necessary step in transcription factor dependent regulation, a series of experiments by Hochschild and colleagues suggests that the minimal elements of a transcription activator are the ability to specifically bind DNA and a region capable of interaction with RNAP (even if that region interacts with a synthetically-introduced tag on RNAP unlikely to be capable of inducing conformational changes) (Dove et al. 1997, Dove and Hochschild 1998, Gregory et al. 2005). To recruit RNAP to a promoter via direct interaction, most transcription factors interact in one of two RNAP locations. Depending on where the transcription factor interacts with the RANP holoenzyme, the transcription factor binding site will appear in a different promoter location (Figure 1.1). A single transcription factor can act on promoters in both classes.

Class I promoters: αCTD-dependent

Class I promoters have an activator binding site upstream of the -35 hexamer, and the activator directly contacts RNAP via the αCTD (Figure 1.1) (Busby and Ebright 1999). These promoters have canonical -35 element sequences and typically have a transcription factor binding site somewhere between -40 and -60. The flexibility in this location reflects the fact that the C-terminal domain of the α subunit has a flexible linker and can bind at variable distances. Class I promoters can be identified by testing for transcription using an αCTD RNAP mutant unable to interact with transcription factors. The *lac* promoter is activated by cyclic AMP receptor protein (CRP) through a Class I promoter (Ebright 1993).

Class II promoters: αCTD-independent

Activator binding sites of class II promoters partially overlap the -35 element, and the activator directly interacts with region 4 of σ^{70} (Figure 1.1) (Busby and Ebright 1997). Generally these promoters also have a functional, canonical -35 element that can interact with σ when the activating transcription factor is not bound. Some promoters, however, are regulated in an αCTD-independent manner and do not appear to have a canonical -35 element. This includes the regulation of the *phoA* promoter by PhoB (Blanco et al. 2011). Unlike Class I promoter binding sites, there is little flexibility in the positioning of Class II promoter binding sites. At some Class II promoters, the activator also contacts the N-terminal domain of the RNAP α-subunit (αNTD) (Busby and Ebright 1997).



Adapted from Busby and Ebright 1997

Figure 1.1: Transcription factor activation classes. Transcription factors can recruit RNA polymerase by interacting with several subunits of the holoenzyme, however a majority use one of two interaction interfaces, which in turn informs the DNA binding site location. Class I promoters have transcription factor binding sites upstream of the -35 element. Transcription factors bound here can interact with the αCTD of RNAP to promote gene expression. The

transcription factor binding site in a class II promoter occludes the -35 site. In this case, the transcription factor interacts with region four of σ.

Repression

There are generally three ways in which transcription factors repress target genes. Steric hindrance is conceptually the simplest way to repress gene expression. In this case, the repressor's binding site overlaps a core promoter element, such as the -10 sequence, preventing RNAP from binding or at least progressing to transcription initiation. This is how the lac repressor acts on the *lac* promoter, for example (Müller et al. 1996). Other proteins appear to use DNA-looping for repression. This is believed to be due to multimers of a transcription factor binding at multiple distal protein binding sites, causing the DNA molecule to double back on itself. This is how GalR is hypothesized to regulate the *gal* promoter (Aki et al. 1996). In still other cases, it appears that repression is achieved by direct binding between repressor and activator. CytR uses this kind of repression by interaction directly with CRP (Shin et al. 2001).

E. coli transcription factor diversity

E. coli has about 150 characterized transcription factors, and is estimated to have a further 150 predicted helix-turn-helix transcription factor genes yet to be experimentally confirmed (Browning and Busby 2004). While some have extensive regulons, about 60 are thought to only control a single promoter. Transcription factors are often grouped into families by sequence similarity. 20 such families have been identified in prokaryotes, and *E. coli* has representatives of each family. The most well-represented family, LysR, has 45 members in the *E. coli* genome (Martínez-Antonio and Collado-Vides 2003). Alm et al. have investigated the origin of some of these transcription factor genes, and show that *E. coli* has expanded its regulatory repertoire using both duplication and divergence, as well as horizontal transfer (2006) Given the large number of DNA-binding proteins and the fact that many are present with close relatives, if signals are to remain insulated and cross-talk minimized, there must be robust specificity determinants in the protein-DNA interaction.

Protein-DNA specificity

Identifying specificity residues

Whether they are activating or repressing downstream genes, transcription factors must use molecular recognition to interact with specific DNA sequences in the promoters they regulate. Early efforts to understand DNA-protein specificity led to rationally mutating, or rewiring, the recognition helix of one protein so that it could bind DNA recognized by a second protein. Wharton and Ptashne were studying two phage repressors with similar overall structures including a conserved helix-turn-helix motif that positions one α-helix into the major groove of B-form DNA during binding. The helix-turn-helix motif used in these phage proteins is a widely distributed DNA-binding motif, found in many transcription factors. They replaced the five "outside" residues of the recognition helix of phage 434 repressor with the corresponding residues from P22, and the hybrid protein interacted with only the P22 operator, not the 434 operator (Wharton and Ptashne 1985). When this hybrid was combined with wild-type 434 repressor, the resulting heterodimer can recognize an asymmetric chimeric operator composed of one 434 binding site and one P22 binding site (Hollis et al. 1988). These studies showed that a small number of specificity determining residues in the recognition helix are necessary and sufficient for recognizing a regulator's target DNA sequence.

Chasing a "recognition code"

At the same time these first promising forays into specificity swapping were being performed, other researchers speculated about eventually deciphering the code for determining a nucleotide binding site based solely on a protein's specificity residues or vice versa. In 1976, Seeman et al. proposed that the hydrogen bonding between the bases in the DNA major groove and the amino acid residues of recognition helices might be conserved across such pairings. For example, they noted that arginine can form two hydrogen bonds only with guanine. If such one-to-one interactions were the basis of protein-DNA interaction specificity, researchers would just need more binding data to fully decode the pattern. Based on the available structural data in 1984 for Cro, lambda repressor, and CAP, however, Pabo and Sauer noted "no one-to-one recognition code is consistent with the current data." They went on to discuss the degeneracy and context dependence of the amino acid to base pair "code" implied by available structural data, pointing out particularly that "the periodicity of an α-helix has no simple relationship to the periodicity of B-DNA," making it impossible to have a predictable set of one-to-one interactions between side chains and bases. In conclusion, however, they predict the problem may eventually be solved with more data: "it is still conceivable that the list of possible interactions will be small enough so that the "code" will have a predictive value." Four years later, with more and higher resolution crystal structure data, Matthews declared "full appreciation of the complexity and individuality of each complex will be discouraging to anyone hoping to find simple answers to the recognition problem" (1988). Indeed, as more structural and affinity data is amassed, the picture has grown more complex. Specific protein-DNA binding can involve the positioning of water molecules (Reddy et al. 2001, Jayaram and Jain 2004), the nucleotide context of the binding site (Nutiu et al. 2011), the spacing between half-sites (Kim and Struhl, 1995), and the binding of other proteins (Wolberger 1999).

While researchers studying protein-DNA interaction specificity may have indeed stopped searching for "simple answers," the advent of computational biology brought new tools to this complex problem. In recent years, computational biologists have made advances in predicting which proteins and regions of proteins are likely candidates for DNA-binding (Si et al. 2015), but taking the further step of predicting recognized sequences has been slower to advance. Structural models and binding simulations have been applied to specific and more general DNA-protein binding scenarios, incorporating a growing amount of structural and binding data (Morozov 2005, Liu and Bradly 2012, Joyce et al. 2015). This computational work has generated new testable hypotheses, provided a starting place for further model refinements, and helped translate large affinity-based datasets into more biologically relevant predictions about binding (Zhou et al. 2015). But the prospect of a recognition code, even a complex algorithmic one, remains elusive.

DNA functional landscapes

The concept of sequence space has been described for both proteins and DNA sequences to explain how mutations might allow proteins or DNA to proceed from one sequence to another along paths that increase or maximize some fitness parameter. A landscape is traditionally envisioned as being made up of nodes that each represent a given sequence and edges that each represent a single mutational step (Figure 1.2).

DNA sequence space landscapes are generally constructed using binding data from *in vitro* experiments that allow the scanning of thousands of unique sequences such as protein-binding microarrays. Sequence affinity is used as a proxy for fitness, with the most tightly bound sequences assumed to be the fittest. Researchers can then choose an affinity threshold based on a combination of known binding sites, biochemical verification, and transcriptional reporter assays to ensure their cutoff has biological relevance.

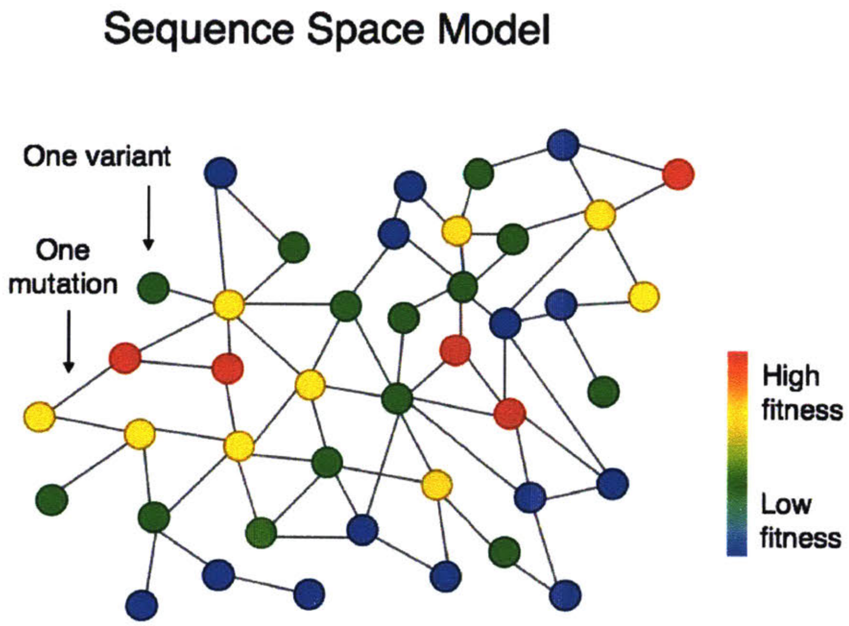


Figure 1.2: Simple model of sequence space. Every node represents one nucleotide sequence and every edge a single mutational step. Nodes are colored based on fitness, where high fitness peaks are red, and lowest fitness valleys are blue. A fitness threshold may make some nodes less accessible. If a threshold is set above, for example, blue nodes, the upper right red node will be less accessible.

While such experiments represent a major breakthrough in the study of protein-DNA specificity, they are not without caveats. *In vitro* assays based on affinity may not reflect the *in vivo* biological function of a given protein—depending on the conditions used, a dataset may reflect binding of a non-biologically relevant oligomeric state, or binding without the presence of other important cofactors or proteins. Even assuming an experimenter has added all necessary cofactors and calibrated the compositions of their buffers to reflect the precise cellular environment they wish to study, they still make the assumption that high affinity corresponds to high fitness. While to some degree this must be true—proteins that have no affinity for a given DNA sequence can't bind and modulate expression at that sequence—it does not necessarily follow that the very highest affinity sites should be preferential for regulatory molecules. Transcription factors must have a wide dynamic range to be useful: if a protein binds its DNA target very tightly under activating conditions but also very tightly under repressed conditions, downstream transcription will not be signal responsive. Furthermore, if DNA sites are present throughout the genome that bind a transcription factor with this very high, constitutive affinity, the protein may be titrated away from functionally interacting elsewhere.

Even with these caveats, recent studies in DNA landscapes have yielded interesting insights. Aguilar-Rodríguez and colleagues have used protein binding microarray data to generate sequence space landscapes for over a thousand transcription factors (2017). They used this large dataset to show that transcription factors with a more smooth, navigable sequence space topology control larger regulons, arguing that evolution may favor landscapes with single, broadly accessible peaks over jagged multipeak landscapes. This is particularly intriguing because although there is evidence that protein specificity landscapes may contain functional regions that are evolutionarily inaccessible, the very nature of DNA molecules—confined generally to a B-form double helical structure made of up only four bases—makes this less likely. Most transcription factor binding sites are either direct or palindromic repeats of only four to six nucleotides long, which is about the maximum "readable" length a single α-helix can interact with in the major groove (Pabo and Sauer 1984). Given how frequently a particular six-mer would occur in a genome by chance (every 4096 bases), it is understandable that transcription factor binding sites exist in pairs (a given 12-mer is expected once every 16 million bases by chance). How sequence space might or might not constrain transcription factor binding sites remains an unanswered question.

Protein-DNA functional landscapes

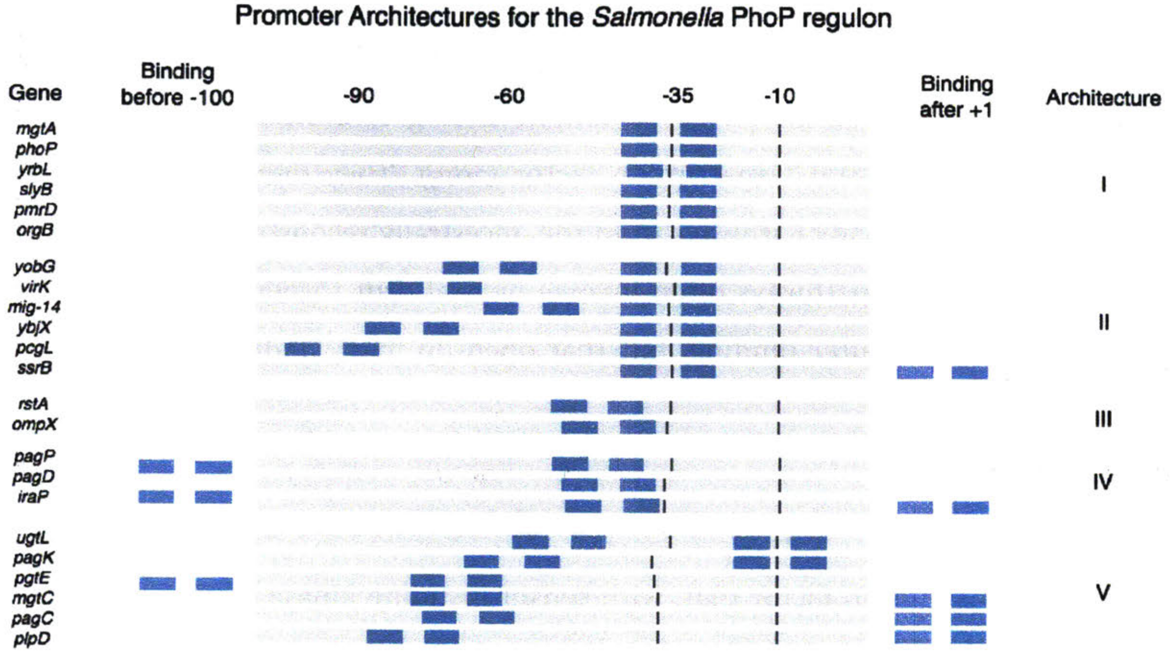
A single protein or DNA functional landscape in isolation only presents a picture of fitness in one static background assuming all other biological factors are held constant. While this is certainly the most tractable way to undergo a deep investigation of a single protein or DNA binding site, it is clearly counter to our understanding of coevolution. Like the long proboscis of a moth predicted by Darwin decades before it was discovered based on the length of an orchid's nectary, the fitness landscape of any one trait is highly interdependent on others (1877).

Anderson et al. tested the 128 combinations between eight protein variants and 16 binding site variants to gain a better understanding of evolutionary trajectories available to an ancestral transcription factor (2015). They found intermolecular epistasis allows for many pathways from the ancestrally reconstructed state to the derived state. These pathways often involve neutral drift through mutations that do not significantly alter binding, but that open up many additional mutational paths.

PhoP regulon and promoter architectures

PhoP is a transcription factor that, with sensor histidine kinase PhoQ, constitutes the magnesium-sensing *E. coli* two-component system PhoPQ (Groisman 2001). Like other canonical two-component systems, PhoPQ is co-operonic and autoregulated. PhoQ autophosphorylates in the presence of low extracellular magnesium and then phosphotransfers intracellularly to PhoP. Phosphorylated PhoP dimerizes and binds to genomic PhoP binding sites. A combination of microarray data, bioinformatic PhoP binding site searching, and occasionally *in-vitro* verification has identified 32 likely regulated operons (Kato et al. 1999, Yamamoto et al. 2002, Minagawa et al. 2003, Monsieurs et al. 2005, Zwir et al. 2005, Ogasawara et al. 2007, Bougdour et al. 2008, Moon and Gottesman 2009, Kaleta et al. 2010, Eguchi et al. 2011, Montero et al. 2011).

PhoP has been studied in both *E. coli* and *Salmonella enterica*. PhoP shares many targets across these two species. In *Salmonella*, the 23 PhoP regulated genes have been subdivided into five groups based on promoter architecture (Figure 1.3) (Zwir et al. 2012). Architecture I promoters have a PhoP binding site straddling the -35 element, as in class II αCTD-independent activation. Indeed, *mgtA*, a gene with a promoter of this architecture, does not require RNAP αCTD for PhoP activation *in vitro* (Perez and Groisman 2009).



Adapted from Zwir et al. 2012

Figure 1.3: Promoter architectures of the *Salmonella* PhoP regulon. Five different promoter architectures are employed in the *Salmonella* PhoP regulon. Blue boxes indicate PhoP binding site position (all promoters were aligned to the -10 element). Promoter architecture impacts how PhoP interacts with RNAP as well as whether the gene is activated or repressed.

Architecture II promoters have a PhoP binding site in the same position as architecture I, but include a second PhoP binding site, usually just upstream. The *virK* promoter (architecture II) was tested for *in vitro* transcription. H-NS was capable of repressing transcription, but PhoP was not simply blocking H-NS binding: when neither H-NS or PhoP were added, transcription did not occur. Thus, it was shown that one PhoP binding site recruits RNAP, while the further upstream site blocks repression by H-NS (Zwir et al. 2012).

Architecture III promoters have a PhoP binding site just upstream of the canonical -35 binding site, in a position that would suggest class I αCTD-dependent activation. This dependence was confirmed for the architecture III promoter *rstA* through in vitro transcription (Zwir et al. 2012).

Architecture IV promoters include a binding site similar to architecture III, as well as a second upstream binding site, and architecture V promoters have PhoP binding sites upstream of the core promoter -35 element as well as binding sites within the gene or at the -10 element. This suggests at least some architecture V genes are repressed by PhoP binding that either prevents transcription initiation or blocks elongation as a downstream roadblock. When compared with *E. coli*, only promoters of architectures I-IV have conserved architectures across the two species. When taken together, this suggests that *E. coli* PhoP likely interacts with different subunits of RNAP depending on promoter architecture.

Conclusion

Transcription is a major point of regulation. Although the first transcription factors discovered were repressors, activators were soon characterized, and some transcription factors were found to be capable of both modes of regulation. In bacteria, transcription can be activated by recruiting RNAP with a transcription factor in a variety of ways, though most commonly by binding near the -35 promoter element and directly interacting with the σ subunit or by binding further upstream and recruiting RNAP in an αCTD-dependent manner. RNAP repression is often achieved with a transcription factor binding site near the -10 promoter element to block transcription initiation or after the transcription start site to block elongation.

Specific protein-DNA interactions facilitate these modes of transcriptional regulation. Molecular recognition between the major groove of promoter DNA and the specificity helix of the regulator is achieved with a small number of amino acids and bases. Theoretical until recently, the affinity-based sequence space for various transcription factors has now been mapped.

In the next chapter, I use the versatile regulator PhoP—shown at least in *Salmonella* to activate through either αCTD-dependent or –independent mechanisms depending on the promoter, to occlude H-NS binding, and to repress gene expression—to explore this concept of sequence space. While affinity landscape models have been built using microarray datasets, it remains unclear whether binding affinity is an appropriate readout of functionality. As deep sequencing and DNA synthesis technology becomes more and more affordable, exploring protein and DNA sequence spaces is now within our grasp. My work provides a starting point for comprehensive *in vivo* study of functional transcription factor landscapes.

References

Aguilar-Rodríguez, J., Payne, J. L., & Wagner, A. (2017). A thousand empirical adaptive landscapes and their navigability. Nature Ecology & Evolution, 1(2), 45.

Aki, T., Choy, H. E., & Adhya, S. (1996). Histone-like protein HU as a specific transcriptional regulator: co-factor role in repression of gal transcription by GAL repressor . Genes to Cells, 1(2), 179–188.

Alm, E., Huang, K., & Arkin, A. (2006). The Evolution of Two-Component Systems in Bacteria Reveals Different Strategies for Niche Adaptation. *PLoS Computational Biology*, 2(11), e143.

Anderson, D. W., McKeown, A. N., & Thornton, J. W. (2015). Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites. eLife, 4.

Benoff, B., Yang, H., Lawson, C. L., Parkinson, G., Liu, J., Blatter, E., ... Ebright, R. H. (2002). Structural basis of transcription activation: the CAP-alpha CTD-DNA complex. Science (New York, N.Y.), 297(5586), 1562–1566.

Blanco, A. G., Canals, A., Bernués, J., Solà, M., & Coll, M. (2011). The structure of a transcription activation subcomplex reveals how σ 70 is recruited to PhoB promoters: Structure of a transcription activation subcomplex. The EMBO Journal, 30(18), 3776–3785.

Bougdour, A., Cunning, C., Baptiste, P. J., Elliott, T., & Gottesman, S. (2008). Multiple pathways for regulation of sigmaS (RpoS) stability in Escherichia coli via the action of multiple anti-adaptors. Molecular Microbiology, 68(2), 298–313.

Brown, N. L., Stoyanov, J. V., Kidd, S. P., & Hobman, J. L. (2003). The MerR family of transcriptional regulators. FEMS Microbiology Reviews, 27(2–3), 145–163.

Browning, D. F., & Busby, S. J. W. (2004). The regulation of bacterial transcription initiation. Nature Reviews Microbiology, 2(1), 57–65.

Burgess, R. R., Travers, A. A., Dunn, J. J., & Bautz, E. K. (1969). Factor stimulating transcription by RNA polymerase. Nature, 221(5175), 43–46.

Busby, S., & Ebright, R. H. (1997). Transcription activation at class II CAP-dependent promoters. Molecular Microbiology, 23(5), 853–859.

Busby, S., & Ebright, R. H. (1999). Transcription activation by catabolite activator protein (CAP). Journal of Molecular Biology, 293(2), 199–213.

Campbell, E. A., Muzzin, O., Chlenov, M., Sun, J. L., Olson, C. A., Weinman, O., … Darst, S. A. (2002). Structure of the bacterial RNA polymerase promoter specificity sigma subunit. Molecular Cell, 9(3), 527–539.

Darwin, Charles (1877) *The various contrivances by which orchids are fertilised by insects*. London: John Murray.

Dove, S. L., & Hochschild, A. (1998). Conversion of the omega subunit of Escherichia coli RNA polymerase into a transcriptional activator or an activation target. Genes & Development, 12(5), 745–754.

Dove, S. L., Joung, J. K., & Hochschild, A. (1997). Activation of prokaryotic transcription through arbitrary protein-protein contacts. Nature, 386(6625), 627–630.

Ebright, R. H. (1993). Transcription activation at Class I CAP-dependent promoters. Molecular Microbiology, 8(5), 797–802.

Eguchi, Y., Ishii, E., Hata, K., & Utsumi, R. (2011). Regulation of acid resistance by connectors of two-component signal transduction systems in Escherichia coli. Journal of Bacteriology, 193(5), 1222–1228.

Englesberg, E., Irr, J., Power, J., & Lee, N. (1965). Positive control of enzyme synthesis by gene C in the L-arabinose system. Journal of Bacteriology, 90(4), 946–957.

Feklistov, A., & Darst, S. A. (2011). Structural basis for promoter-10 element recognition by the bacterial RNA polymerase σ subunit. Cell, 147(6), 1257–1269.

Fried, M. G., & Crothers, D. M. (1983). CAP and RNA polymerase interactions with the lac promoter: binding stoichiometry and long range effects. Nucleic Acids Research, 11(1), 141–158.

Garen, A., & Echols, H. (1962). Genetic control of induction of alkaline phosphatase synthesis in E. coli. Proceedings of the National Academy of Sciences of the United States of America, 48, 1398–1402.

Gourse, R. L., Ross, W., & Gaal, T. (2000). UPs and downs in bacterial transcription initiation: the role of the alpha subunit of RNA polymerase in promoter recognition. Molecular Microbiology, 37(4), 687–695.

Gregory, B. D., Deighan, P., & Hochschild, A. (2005). An artificial activator that contacts a normally occluded surface of the RNA polymerase holoenzyme. Journal of Molecular Biology, 353(3), 497–506.

Groisman, E. A. (2001). The Pleiotropic Two-Component Regulatory System PhoP-PhoQ. Journal of Bacteriology, 183(6), 1835–1842.

Gruber, T. M., & Gross, C. A. (2003). Multiple sigma subunits and the partitioning of bacterial transcription space. Annual Review of Microbiology, 57, 441–466.

Haakonsen, D. L., Yuan, A. H., & Laub, M. T. (2015). The bacterial cell cycle regulator GcrA is a σ 70 cofactor that drives gene expression from a subset of methylated promoters. Genes & Development, 29(21), 2272–2286.

Heincz, M. C., & McFall, E. (1978). Role of the dsdC activator in regulation of D-serine deaminase synthesis. Journal of Bacteriology, 136(1), 96–103.

Hollis, M., Valenzuela, D., Pioli, D., Wharton, R., & Ptashne, M. (1988). A repressor heterodimer binds to a chimeric operator. Proceedings of the National Academy of Sciences of the United States of America, 85(16), 5834–5838.

Ishihama, A. (2000). Functional modulation of Escherichia coli RNA polymerase. Annual Review of Microbiology, 54, 499–518.

Islam, M. S., Bingle, L. E. H., Pallen, M. J., & Busby, S. J. W. (2011). Organization of the LEE1 operon regulatory region of enterohaemorrhagic Escherichia coli O157:H7 and activation by GrlA. Molecular Microbiology, 79(2), 468–483.

Jacob, F., & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. Journal of Molecular Biology, 3(3), 318–356.

Jain, D., Nickels, B. E., Sun, L., Hochschild, A., & Darst, S. A. (2004). Structure of a ternary transcription activation complex. Molecular Cell, 13(1), 45–53.

Jayaram, B., & Jain, T. (2004). The role of water in protein-DNA recognition. Annual Review of Biophysics and Biomolecular Structure, 33, 343–361.

Joyce, A. P., Zhang, C., Bradley, P., & Havranek, J. J. (2015). Structure-based modeling of protein: DNA specificity. Briefings in Functional Genomics, 14(1), 39–49.

Kaleta, C., Göhler, A., Schuster, S., Jahreis, K., Guthke, R., & Nikolajewa, S. (2010). Integrative inference of gene-regulatory networks in Escherichia coli using information theoretic concepts and sequence analysis. BMC Systems Biology, 4, 116.

Kato, A., Tanabe, H., & Utsumi, R. (1999). Molecular characterization of the PhoP-PhoQ two-component system in Escherichia coli K-12: identification of extracellular Mg2+-responsive promoters. Journal of Bacteriology, 181(17), 5516–5520.

Kim, J., & Struhl, K. (1995). Determinants of half-site spacing preferences that distinguish AP-1 and ATF/CREB bZIP domains. Nucleic Acids Research, 23(13), 2531–2537.

Lee, D. J., Minchin, S. D., & Busby, S. J. W. (2012). Activating Transcription in Bacteria. Annual Review of Microbiology, 66(1), 125–152.

Lee, N. L., Gielow, W. O., & Wallace, R. G. (1981). Mechanism of araC autoregulation and the domains of two overlapping promoters, Pc and PBAD, in the L-arabinose regulatory region of Escherichia coli. Proceedings of the National Academy of Sciences of the United States of America, 78(2), 752–756.

Liu, L. A., & Bradley, P. (2012). Atomistic modeling of protein–DNA interaction specificity: progress and applications. Current Opinion in Structural Biology, 22(4), 397–405.

Lynch, M., & Hagner, K. (2015). Evolutionary meandering of intermolecular interactions along the drift barrier. Proceedings of the National Academy of Sciences, 112(1), E30–E38.

Martínez-Antonio, A., & Collado-Vides, J. (2003). Identifying global regulators in transcriptional regulatory networks in bacteria. Current Opinion in Microbiology, 6(5), 482–489.

Matthews, B. W. (1988). Protein-DNA interaction. No code for recognition. Nature, 335(6188), 294–295.

Minagawa, S., Ogasawara, H., Kato, A., Yamamoto, K., Eguchi, Y., Oshima, T., ... Utsumi, R. (2003). Identification and molecular characterization of the Mg2+ stimulon of Escherichia coli. Journal of Bacteriology, 185(13), 3696–3702.

Monsieurs, P., De Keersmaecker, S., Navarre, W. W., Bader, M. W., De Smet, F., McClelland, M., ... Marchal, K. (2005). Comparison of the PhoPQ regulon in Escherichia coli and Salmonella typhimurium. Journal of Molecular Evolution, 60(4), 462–474.

Montero, M., Almagro, G., Eydallin, G., Viale, A. M., Muñoz, F. J., Bahaji, A., ... Pozueta-Romero, J. (2011). Escherichia coli glycogen genes are organized in a single glgBXCAP transcriptional unit possessing an alternative suboperonic promoter within glgC that directs glgAP expression. The Biochemical Journal, 433(1), 107–117.

Moon, K., & Gottesman, S. (2009). A PhoQ/P-regulated small RNA regulates sensitivity of Escherichia coli to antimicrobial peptides. Molecular Microbiology, 74(6), 1314–1330.

Morozov, A. V. (2005). Protein-DNA binding specificity predictions with structural models. Nucleic Acids Research, 33(18), 5781–5798.

Müller, J., Oehler, S., & Müller-Hill, B. (1996). Repression of lac promoter as a function of distance, phase and quality of an auxiliary lac operator. Journal of Molecular Biology, 257(1), 21–29.

Nutiu, R., Friedman, R. C., Luo, S., Khrebtukova, I., Silva, D., Li, R., ... Burge, C. B. (2011). Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. Nature Biotechnology, 29(7), 659–664.

Ogasawara, H., Hasegawa, A., Kanda, E., Miki, T., Yamamoto, K., & Ishihama, A. (2007). Genomic SELEX search for target promoters under the control of the PhoQP-RstBA signal relay cascade. Journal of Bacteriology, 189(13), 4791–4799.

Pabo, C. O., & Sauer, R. T. (1984). Protein-DNA Recognition. Annual Review of Biochemistry, 53(1), 293–321.

Paul, B. J., Barker, M. M., Ross, W., Schneider, D. A., Webb, C., Foster, J. W., & Gourse, R. L. (2004). DksA: a critical component of the transcription initiation machinery that potentiates the regulation of rRNA promoters by ppGpp and the initiating NTP. Cell, 118(3), 311–322.

Perez, J. C., & Groisman, E. A. (2009). Transcription factor function and promoter architecture govern the evolution of bacterial regulons. Proceedings of the National Academy of Sciences, 106(11), 4319–4324.

Popham, D., Szeto, D., Keener, J., & Kustu, S. (1989). Function of a bacterial activator protein that binds to transcriptional enhancers. Science, 243(4891), 629–635.

Ptashne, M. (1967). Isolation of the lambda phage repressor. Proceedings of the National Academy of Sciences of the United States of America, 57(2), 306–313.

Reddy, C. K., Das, A., & Jayaram, B. (2001). Do water molecules mediate protein-DNA recognition? Journal of Molecular Biology, 314(3), 619–632.

Reyes-Caballero, H., Campanello, G. C., & Giedroc, D. P. (2011). Metalloregulatory proteins: Metal selectivity and allosteric switching. Biophysical Chemistry, 156(2–3), 103–114.

Saecker, R. M., Record, M. T., & Dehaseth, P. L. (2011). Mechanism of bacterial transcription initiation: RNA polymerase - promoter binding, isomerization to initiation-competent open complexes, and initiation of RNA synthesis. Journal of Molecular Biology, 412(5), 754–771.

Seeman, N. C., Rosenberg, J. M., & Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. Proceedings of the National Academy of Sciences of the United States of America, 73(3), 804–808.

Shin, M. (2001). Repression of deoP2 in Escherichia coli by CytR: conversion of a transcription activator into a repressor. The EMBO Journal, 20(19), 5392–5399.

Si, J., Zhao, R., & Wu, R. (2015). An Overview of the Prediction of Protein DNA-Binding Sites. International Journal of Molecular Sciences, 16(3), 5194–5215.

Wharton, R. P., & Ptashne, M. (1985). Changing the binding specificity of a repressor by redesigning an alpha-helix. Nature, 316(6029), 601–605.

Wolberger, C. (1999). Multiprotein-DNA complexes in transcriptional regulation. Annual Review of Biophysics and Biomolecular Structure, 28, 29–56.

Yamamoto, K., Ogasawara, H., Fujita, N., Utsumi, R., & Ishihama, A. (2002). Novel mode of transcription regulation of divergently overlapping promoters by PhoP, the regulator of two-component system sensing external magnesium availability. Molecular Microbiology, 45(2), 423–438.

Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R. S., ... Rohs, R. (2015). Quantitative modeling of transcription factor binding specificities using DNA shape. Proceedings of the National Academy of Sciences of the United States of America, 112(15), 4654–4659.

Zubay, G., & Doty, P. (1959). The isolation and properties of deoxyribonucleoprotein particles containing single nucleic acid molecules. Journal of Molecular Biology, 1(1), 1–IN1.

Zwir, I., Latifi, T., Perez, J. C., Huang, H., & Groisman, E. A. (2012). The promoter architectural landscape of the Salmonella PhoP regulon. Molecular Microbiology, 84(3), 463–485.

Zwir, I., Shin, D., Kato, A., Nishino, K., Latifi, T., Solomon, F., ... Groisman, E. A. (2005). Dissecting the PhoP regulatory network of Escherichia coli and Salmonella enterica. Proceedings of the National Academy of Sciences of the United States of America, 102(8), 2862–2867.

Chapter 2

Systematic determination of a transcription factor/binding site functional interaction landscape

Introduction: Protein-DNA specificity and two-component systems

Specific interactions between proteins and DNA sequences are central to the most fundamental processes of life. Because transcription factors are essential to the proper temporal and spatial expression of genes within individual cells and multicellular organisms, they are found in all living things. Without them, cells could not modulate gene expression in response to external environmental signals or internal developmental cues. This ability to tune transcriptional output is a major point of regulation and a hub of adaptation.

Two-component systems are an excellent context for studying transcription factor specificity. *E. coli* has 22 canonical, homologous two-component systems that arose through an evolutionary history of both horizontal transfer and gene duplication and divergence (Alm et al. 2006). Despite the similarities between these proteins' sequences and structures, they nonetheless remain insulated as they process parallel signals (Capra et al. 2012).

Typically, an input signal causes a given histidine kinase to autophosphorylate and then phosphotransfer to its partner response regulator. Most response regulators are transcription factors that dimerize upon phosphorylation and then bind DNA to modulate gene expression. Under repressing conditions, the histidine kinase usually acts as a phosphatase on its cognate response regulator, ensuring that signals can be tuned down when a system no longer needs to be active and that spurious sources of phosphorylation do not activate a regulator inappropriately.

The protein-protein interactions required for phosphotransfer and phosphatase activity have been shown to be specific and insulated within a given organism (Capra et al. 2012). This is achieved primarily through molecular recognition, involving a relatively small number (~4-8) of amino acids on each half of the interface formed by a histidine kinase and response regulator, despite the high sequence and structural similarity across these families of proteins. These interfacial, specificity-determining residues have been bioinformatically inferred by looking for residues that covary between cognate kinases and regulators in large multiple sequence alignments, and verified with structural data when available. Additionally, the necessary and sufficient residues for specificity have been established through rational mutagenesis and the rewiring of kinase-regulator interactions. For instance, the *E. coli* histidine kinase EnvZ, which normally interfaces specifically with OmpR, was rewired to phosphorylate RstA, the target of the kinase RstB, by substituting just three of the key specificity residues of EnvZ with the corresponding residues from RstB (Skerker et al. 2008). Similarly, the entire interface of a *T. maritima* two-component system was rewired by substituting the wild-type specificity residues of both the kinase, HK853, and its cognate regulator, RR468, with the corresponding residues found in *E. coli* PhoR and PhoB, respectively, which required only five mutations in the kinase and four in the regulator (Podgornaia et al. 2013). This work led to a more comprehensive investigation of the specificity residues for the *E. coli* histidine kinase PhoQ in which a library of all 160,000 possible mutants at four key specificity positions was constructed and screened for functionality using a fluorescent reporter of PhoQ activity. The 1,659 functional variants identified in this screen revealed that (i) tremendous functional plasticity exists, i.e. many different combinations of residues support the PhoQ-PhoP interface, (ii) individual substitutions are often highly context-dependent meaning that the effect of a given substitution can often depend on what other substitutions have or have not also been introduced, and (iii) functional proteins exist that are not seen in PhoQ orthologs in nature, likely because reaching them would require deleterious intermediate mutations, making them more isolated in sequence space (Podgornaia and Laub 2015).

Sequence space is a framework for thinking about all possible protein or DNA sequences of some given finite length. Sequence space is often conceptualized as part of a three dimensional fitness landscape where the xy coordinates correspond to the information encoded in the genotype (for proteins, the amino acid sequence) and the z coordinate corresponds to some phenotype-based fitness estimate. Stepwise motion from one coordinate on this landscape to another is thus a modeled evolutionary trajectory; moving from one fitness state to another via mutation. The concept of sequence space was first introduced in 1970 by John Maynard Smith, but the idea was largely theoretical until DNA synthesis and sequencing technologies made it possible to build and assay actual sequence spaces, like the four PhoQ residues described above. The entire sequence space sampled in this case is 160,000. To visually represent the functional sequence space of 1659 variants, a force-directed diagram was used. This is a diagram where each node represents a functional sequence and each connection represents a single mutational step. A force-directed algorithm—where each edge has a spring-like attractive force pulling the nodes at its ends together and nodes themselves have repulsive forces—was applied to the functional sequences, resulting in a visual representation of functional sequence space, where interconnected and thus highly related sequences cluster together. Indeed, when shown on a force-directed diagram, sequences sampled by the specificity residues of the PhoQ protein and its orthologs clustered together while unused functional sequences were isolated with few connecting paths to wild-type.

The sequence space corresponding to the nucleic acids in the promoters activated by a given response regulator, however, may be subject to a very different set of constraints than proteins: DNA does not form the complex folds found in polypeptides, and is limited to only four different bases compared with the 20 highly varied amino acid residues possible at each position of a protein. Both of these factors greatly constrict the number of possible binding surfaces that exist using DNA compared with protein. For example, any randomly chosen ten base DNA sequence has a 99% chance of appearing at least once in a genome the size of *E. coli*'s, whereas the chance of finding a randomly chosen ten amino acid sequence in the protein coding regions of *E. coli* is less than one in ten million. Similarly, while epistasis, or context-dependency, is well-documented in protein mutations, including DNA binding proteins (De Mendoza et al. 2013), DNA positions are often assumed to be largely independent or only capable of short-range interactions (Stormo 2013).

Key questions

Here, I comprehensively investigated the functional DNA interactions possible for the *E. coli* transcription factor PhoP. What promoter sequences functionally interact with PhoP? Do sequences exist that don't look like wild-type binding sites but are nevertheless functional? I find that there are functional sequences that differ from the established consensus binding site. What keeps *E. coli* from employing these sequences—especially given the relatively large genomic sampling of DNA sequence space? We show that avoided sequences exist in an area of sequence space with many constitutively active neighbors, suggesting that evolution may favor sequences robust to single mutations resulting in a loss of function.

Do positions within and across binding sites combine epistatically, or are they largely independent? We find a great deal of interaction between the bases within and across binding sites, as well as context dependence resulting in one binding site position of this direct repeat tolerating different half-sites than the other. And finally, as mutations occur in the transcription factor, what DNA sequences become available and which are lost? Changing specificity in the protein-protein interaction between a toxin and an antitoxin has suggested that promiscuous intermediates may play a key role as protein partners evolve into new areas of sequence space (Aakre et al. 2015). Given the different parameters of DNA, where do promiscuous sequences appear in sequence space, and what does this suggest about evolutionary trajectories? We show that promiscuity is rare and many interconnections exist between two insulated regulons, suggesting promiscuous intermediates are not likely important evolutionary intermediates in this context.

Experimental overview

Early descriptions of transcription factor specificity employed DNase footprinting (Tijan 1978, Ng et al. 1979), followed shortly thereafter by crystal structures of transcription factors bound to DNA (McKay and Steitz 1981, Anderson et al. 1983). These breakthroughs advanced our understanding of how individual transcription factors use molecular recognition to interact specifically with a given stretch of DNA. Such labor intensive methods are not feasible when studying more than a small number of proteins or DNA sequences.

Researchers soon developed higher throughput techniques for studying protein-DNA affinity. ChIP-chip (Blat et al. 1999), and eventually ChIP-seq (Johnson et al. 2007) have been employed to determine where a given protein binds the genome *in vivo*, under varied conditions if desired. Systematic Evolution of Ligands by EXponential enrichment (SELEX) (Tuerk and Gold, 1990), Mechanically Induced Trapping Of Molecular Interactions (MITOMI) (Maerkl and Quake 2007), and then Protein-Binding Microarrays (PBMs) (Bulyk 2007) each can determine the binding preferences of a given protein using large nucleic acid libraries *in vitro*. These powerful tools have greatly expanded the number of proteins whose binding preferences are known, but they are not without caveats.

To begin to answer the questions posed above, I needed a system that could assay large libraries of DNA sequences, like SELEX, MITOMI, or PBMs, but could also, like ChIP, provide functional information in activating and repressing conditions. I used a library system adapted from a previous protein-protein specificity investigation (Podgornaia and Laub, 2015) to determine the *in vivo* functional binding specificity of PhoP.

PhoP belongs to the *E. coli* two-component system PhoPQ. It has been widely studied and is tractable in the lab because, unlike many such systems, activation can be achieved easily—in this case by varying the extracellular magnesium concentration (Groisman, 2001). PhoQ is a membrane-bound histidine kinase that autophosphorylates in the presence of low extracellular magnesium and then phosphotransfers to the transcription factor PhoP (Figure 2.1). PhoP has been previously implicated in the regulation of 32 operons (Kato et al. 1999, Yamamoto et al. 2002, Minagawa et al. 2003, Monsieurs et al. 2005, Zwir et al. 2005, Ogasawara et al. 2007, Bougdour et al. 2008, Moon and Gottesman 2009, Kaleta et al. 2010, Eguchi et al. 2011, Montero et al. 2011).

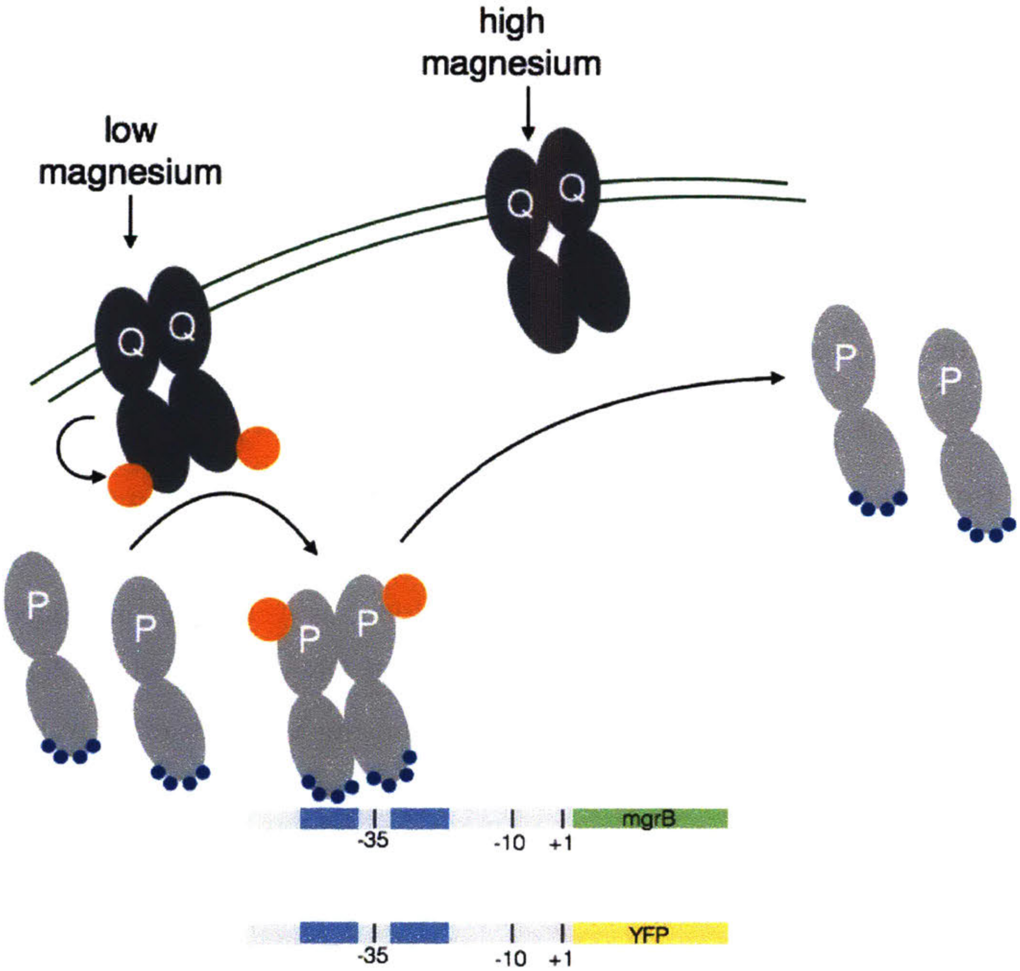


Figure 2.1: The PhoPQ two-component system. The sensor kinase PhoQ is activated in the presence of low magnesium, which shifts the equilibrium of its bifunctional enzymatic activity toward autophosphorylation and phosphotransfer. Phosphorylated PhoP is active and binds to regulatory sequences including the mgrB promoter, here shown in its native architecture and in the context of a transcriptional YFP fusion. If magnesium is returned to the system, PhoQ then acts more, on average, as a phosphatase on PhoP, leading to deactivation of the system.

Results

ChIP seq identifies PhoP consensus sequence and regulon

To identify the genomic binding sites of PhoP and establish its consensus sequence, we performed ChIP-seq. We first tested our 3x-FLAG tagged PhoP construct by performing qPCR on *mgrB* and *mgtA*, two previously-identified PhoP-regulated promoters. Our construct was unable to pull down these positive controls if expressed as the only copy of PhoP in the cell, but could pull down these promoters if the wild-type *phoPQ* operon was also present (Figure 2.2B). Because PhoP and PhoQ are co-operonic, mutations in the C-terminal region of PhoP, such as the coding region for a 3x-FLAG tag, likely impact PhoQ expression necessitating the presence of an additional, wild-type copy of the operon.

Cells expressing *phoP-3xFLAG* were grown in M9 medium containing either low magnesium (10 μM) or high magnesium (50 mM) for six hours prior to ChIP to determine where PhoP binds under activating and repressing conditions, respectively. Two biological replicates of each condition, as well as an untagged PhoP control were then subjected to ChIP, with bound DNA identified by deep sequencing. The resulting ChIP-seq data was analyzed using MACS, leading to the identification of 174 peaks in at least one PhoP-3xFLAG sample (Figure 2.2). While some peaks were found under both activating and repressing conditions, no peak bound in the repressed condition was absent from the activated samples. These data indicate that PhoP is present at some promoters even in non-inducing conditions, with phosphorylation likely influencing the ability of PhoP to activate transcription. At some promoters, PhoP is absent in non-inducing conditions, with the phosphorylation of PhoP upon Mg-limitation likely promoting the binding of PhoP and subsequent activation of transcription.

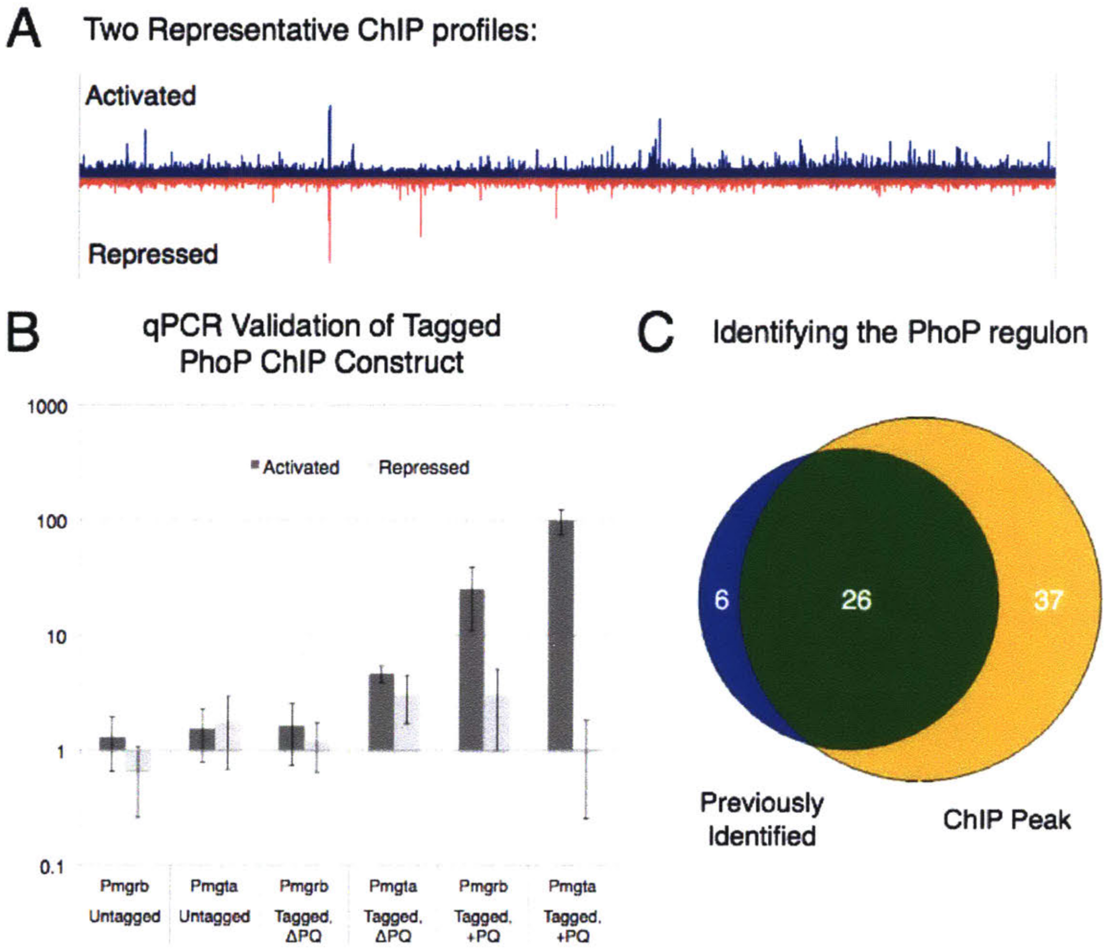


Figure 2.2: PhoP ChIP-seq overview. A: Representative biological replicates for PhoP binding peaks across the *E. coli* genome under activating and repressing conditions. B: qPCR enrichment from tagged and untagged PhoP samples reveals that tagging PhoP directly or indirectly interferes with immunoprecipitation of PhoP regulated genes. This can be circumvented by transforming tagged PhoP into a wild-type (PhoPQ containing) background. C: 69 peaks are called within promoter regions and appear in more than one replicate. 26 called peaks are found in promoters previously predicted or verified to be part of the PhoP regulon. No peak is found at six bioinformatically identified genomic PhoP binding sites. 37 peaks correspond to genes not previously identified as PhoP dependent.

Of the peaks identified, 69 mapped to a predicted intergenic promoter region and were present in at least two of the activated replicates. Of these 69 peaks, 26 corresponded to promoters previously identified or predicted to be under PhoP regulation (Kato et al. 1999, Yamamoto et al. 2002, Minagawa et al. 2003, Monsieurs et al. 2005, Zwir et al. 2005, Ogasawara et al. 2007, Bougdour et al. 2008, Moon and Gottesman 2009, Kaleta et al. 2010, Eguchi et al. 2011, Montero et al. 2011). Six genes previously predicted to have PhoP binding sites by motif search did not have peaks. Of the 69 total peaks, 37 occurred in predicted promoter regions with no previously identified or suggested PhoP binding site. Three of these, however, had been identified as potentially PhoP regulated in DNA microarray studies, but the regulation was assumed indirect because no canonical PhoP binding site was found in their promoters (Figure 2.3) (Minagawa et al. 2003). When analyzed by MEME, seven of the 37 novel peaks spanned predicted PhoP binding sites. No other secondary binding sites were found at these peaks.

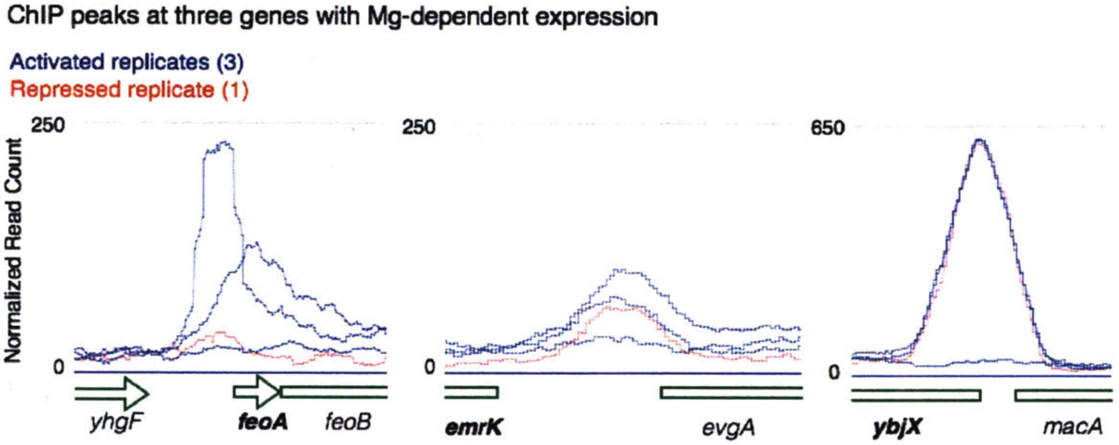


Figure 2.3: ChIP peaks are present at three genes previously identified by microarray. FeoA, EmrK, and YbjX were all identified by microarray as possible PhoP regulon members. No canonical PhoP binding site was found, so the regulation was assumed indirect, however we find evidence of PhoP binding at each of these promoters under activating conditions.

In summary, our ChIP-seq analyses confirmed 26 PhoP-regulated genes and identified 10 peaks that either had a previously unidentified PhoP binding site or controlled genes that were previously reported as PhoP-dependent. We conclude that this set of 36 genes represents the *E. coli* PhoP regulon. No evidence of binding was found for six previously predicted PhoP regulated genes, and 27 peaks were found that may represent noise, an alternative binding site, or PhoP interaction with other proteins. A consensus sequence was identified for the group of 69 peak sequences with MEME (Figure 2.4), and this sequence logo was used to identify the nucleotide positions to target in our randomized libraries, described below.

Randomized Library Design

Although the ChIP-seq data provide some insight into the diversity of sequences that can be bound by PhoP, these naturally occurring instances of PhoP binding sites may not represent the full set of possible binding sites. To comprehensively determine which DNA sequences can bind PhoP, we sought to construct randomized promoter libraries using a known PhoP regulated promoter driving expression of the fluorescent reporter gene YFP. We wanted a reporter with a large dynamic range between on (low Mg^{++}) and off (high Mg^{++}) states. Additionally, because PhoP has been shown to activate and repress genes using a variety of promoter architectures in both RNA polymerase αCTD-dependent and αCTD-independent manners (Perez and Groisman 2009, Zwir et al. 2012), we wanted a straightforward promoter architecture with a single PhoP binding site to randomize. Previously tested transcriptional reporters showed that the *mgrB* promoter has the largest dynamic range of PhoP targets in *E. coli* (Lippa and Goulian 2009). MgrB is a small membrane protein upregulated by PhoP which acts as part of a feedback loop to inhibit PhoQ kinase activity (Lippa and Goulian 2009, Salazar et al. 2016). This promoter has a single PhoP binding site straddling the -35 element. In *Salmonella*, the *mgtA* promoter has the same architecture, and has been shown to be αCTD-independent, suggesting direct interaction between PhoP and sigma-70 of RNA polymerase (Perez and Groisman 2009).

Core binding site library

To investigate the sequence space governing the central core of a PhoP binding site, we constructed a library that randomized the four most highly conserved positions in each half site (Figure 2.4). This "core binding site" library is essential to determining how nucleotides interact within and between binding sites, and how binding site halves functionally combine. By randomizing 8 positions, our library has a theoretical diversity of 65,536, which, as described below, can be fully sampled.

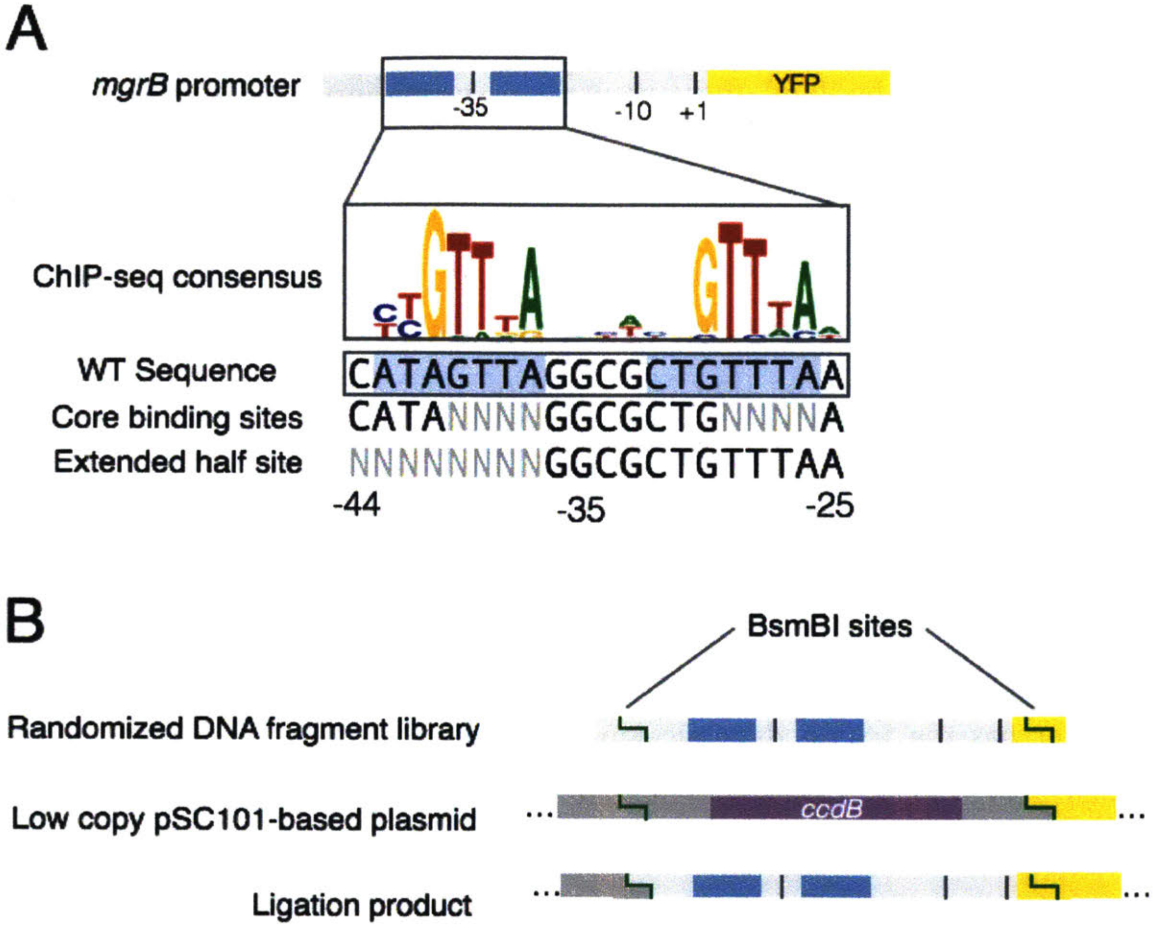


Figure 2.4: Randomized libraries and plasmid library construction. A: Guided by the pattern of conserved residues in the ChIP-seq logo, we designed a "core binding site library" and "extended half-site library" in the mgrB promoter. Nucleotides listed as 'N' were randomized in their respective library. B: The randomized DNA fragment library was made into a plasmid library with two different compatible overhangs cut by BsmBI. Uncut/re-annealed plasmid was counterselected by the toxin CcdB.

Extended half-site library

To determine how nucleotides in the less conserved, extended binding site region are constrained, we built a second library that randomizes eight nucleotides of the left binding site (Figure 2.4A). Traditionally, the PhoP binding site has been described as a direct repeat of 6 to 7 nucleotides (Kato et al. 1999, Zwir et al. 2005), so this "extended binding site" library allows us to determine what constraints are placed on highly conserved core nucleotides compared with their less-conserved neighbors.

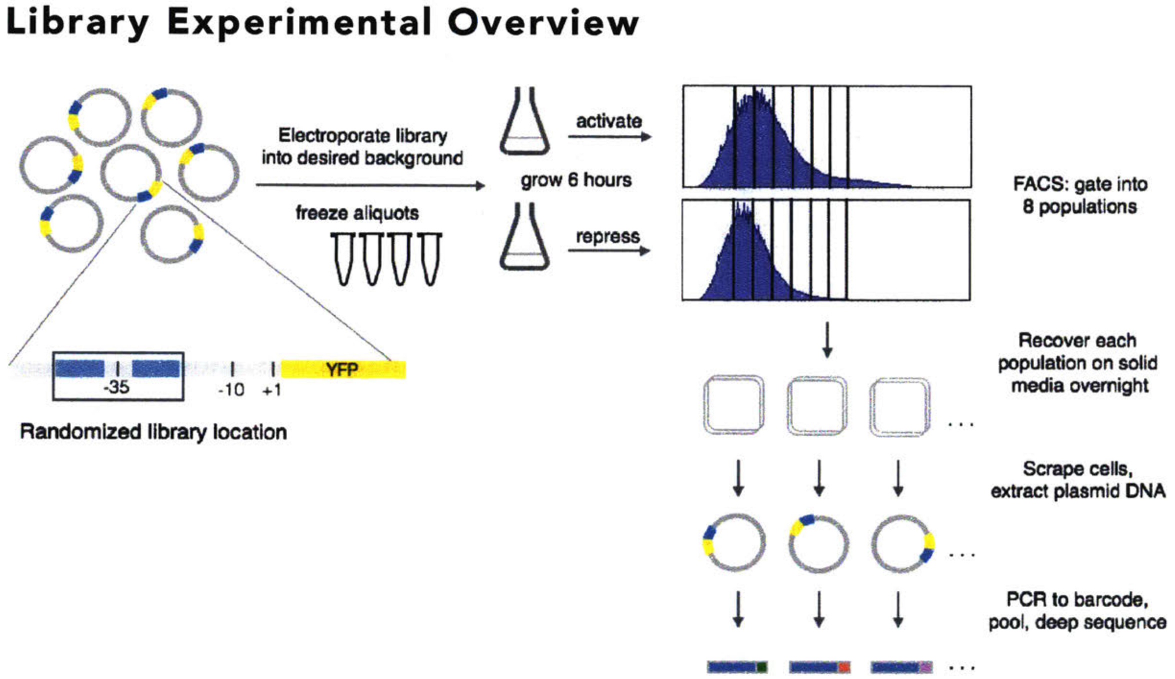

Library Experimental Overview
Electroporate library into desired background
freeze aliquots
activate
grow 6 hours
repress
FACS: gate into 8 populations
-35
-10
+1
YFP
Randomized library location
Recover each population on solid media overnight
Scrape cells, extract plasmid DNA
PCR to barcode, pool, deep sequence

Figure 2.5: Experimental overview. The randomized plasmid libraries were electroporated into the desired genetic background, and aliquots were frozen for later use. After growth for six hours under either activating (low magnesium: 0.01 mM) or repressing (high magnesium: 50 mM) conditions, libraries were sorted via FACS into eight gates (representative gating from the extended binding site library is shown) and recovered on solid media. Plasmids were harvested, and the promoter was amplified by barcoded PCR primers and deep sequenced.

For each library, the *mgrB* promoter, including the randomized nucleotides, was synthesized and then cloned into an expression vector. To ensure that we sampled as many variants as possible, we transformed our plasmid libraries into wild-type cells with > 750X coverage. Frozen aliquots could then be thawed and grown under activating (0.01 mM Mg) or repressing (50 mM Mg) conditions in M9 medium. After six hours in either condition, Fluorescence Activated Cell Sorting (FACS) was performed. To determine the relative fluorescence of each variant, populations of cells were separated into eight subpopulations via gating based on the wild-type sequence's performance in activated and repressed conditions (Figure 2.6A). One million cells per gate were collected and grown on large solid media plates overnight to decrease possible jackpotting events. Colonies were scraped off and pooled, plasmids were isolated, and barcoded primers were used to amplify the randomized library region through PCR. Barcoded PCR products were then deep sequenced. In short, this procedure provided read counts for each variant in each gated subpopulation, effectively recapitulating the distribution of YFP expression levels of each variant in activating and repressing conditions.

To control for PhoPQ-independent activation of individual variants, the plasmid libraries were transformed into a *ΔphoPQ* strain and analyzed as above. However, too few cells were present to harvest and sequence, indicating that very few of the cells observed in the high fluorescence gates in the wild-type background were activated independent of PhoPQ (Figure 2.6B).

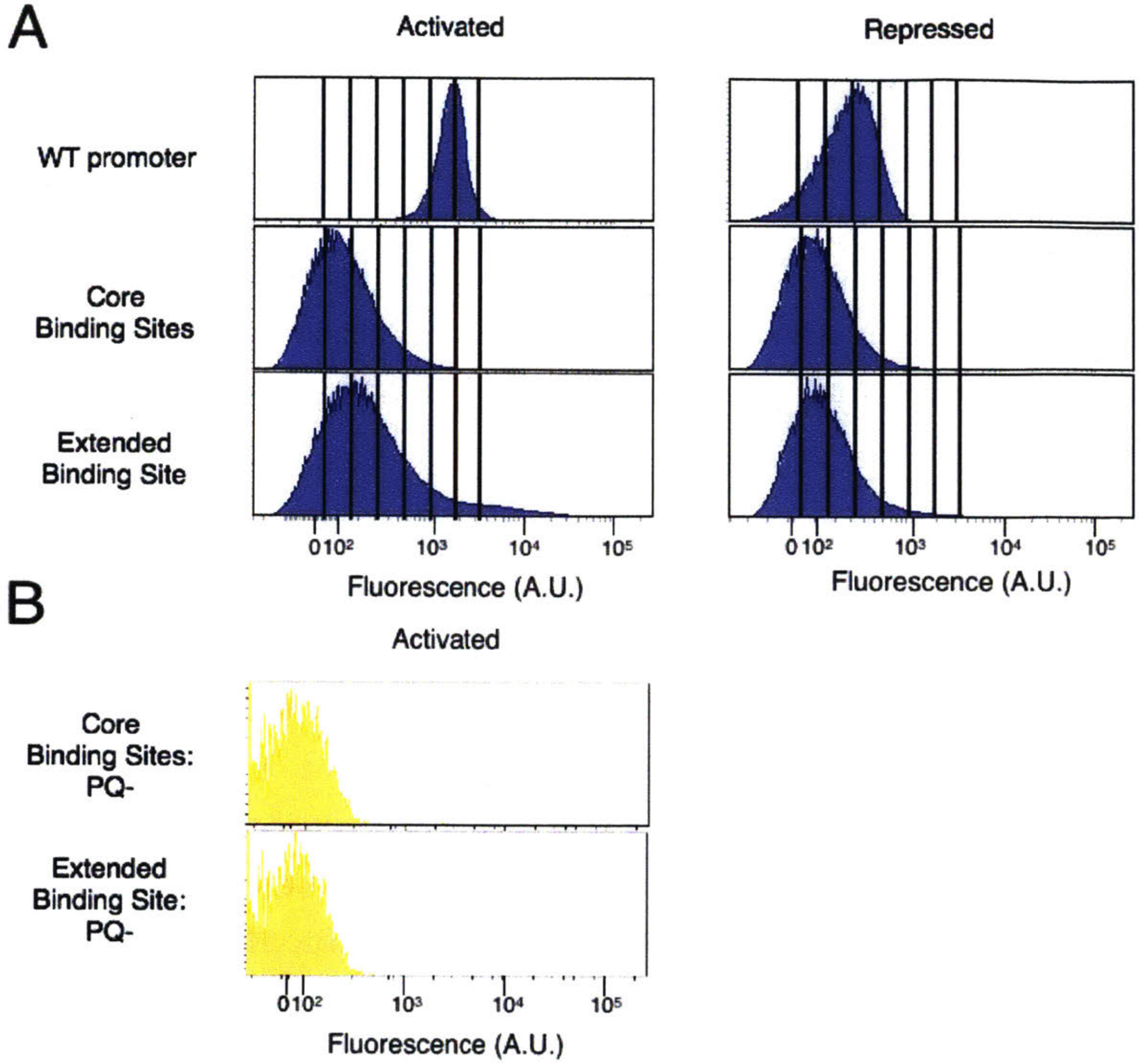


Figure 2.6: Library gating. A: Representative replicates are shown for the core and extended binding site libraries. Gates were chosen to best capture behavior similar to wild-type. B: Each library was transformed into a ΔPhoPQ background to sort for PhoP independent activation, however so few cells have high fluorescence in this context that ΔPhoPQ cells were not collected or sequenced.

Library Data analysis and Validation

To determine how comprehensively we had sampled from all 65,536 variants in each library, we calculated the coverage for a range of read thresholds. This analysis shows that we indeed sampled the vast majority of possible library sequences, with >99% of sequences present above a threshold of 5 counts for each library (Figure 2.8A).

To establish how robust our results were to noise, we examined the correlation of biological and sequencing replicates. The sequencing replicates are highly correlated, with less correlation for the two biological replicates (Figure 2.7). We also examined the set of 256 sequences present in both libraries, observing a high level of agreement across the libraries when biological replicates are combined (Figure 2.9).

As addition validation of our screen and thresholding, we predicted the functionality of 52 sequences and then tested our predictions by measuring the fluorescence of each individual variant under activating and repressing conditions. Of this set of 52, 27 of the 33 sequences predicted to be functional with respect to PhoP binding and activation were indeed functional and 16 of the 19 sequences predicted to be non-functional were, in fact, non-functional. Based on this analysis, we estimated that our screening procedure had a false positive rate of 16% and a false negative rate of 13%. Taken together, these analyses suggest that our dataset of combined biological replicates is comprehensive and fairly robust. Using these combined biological replicates and the thresholds we established, we then constructed models of sequence space.

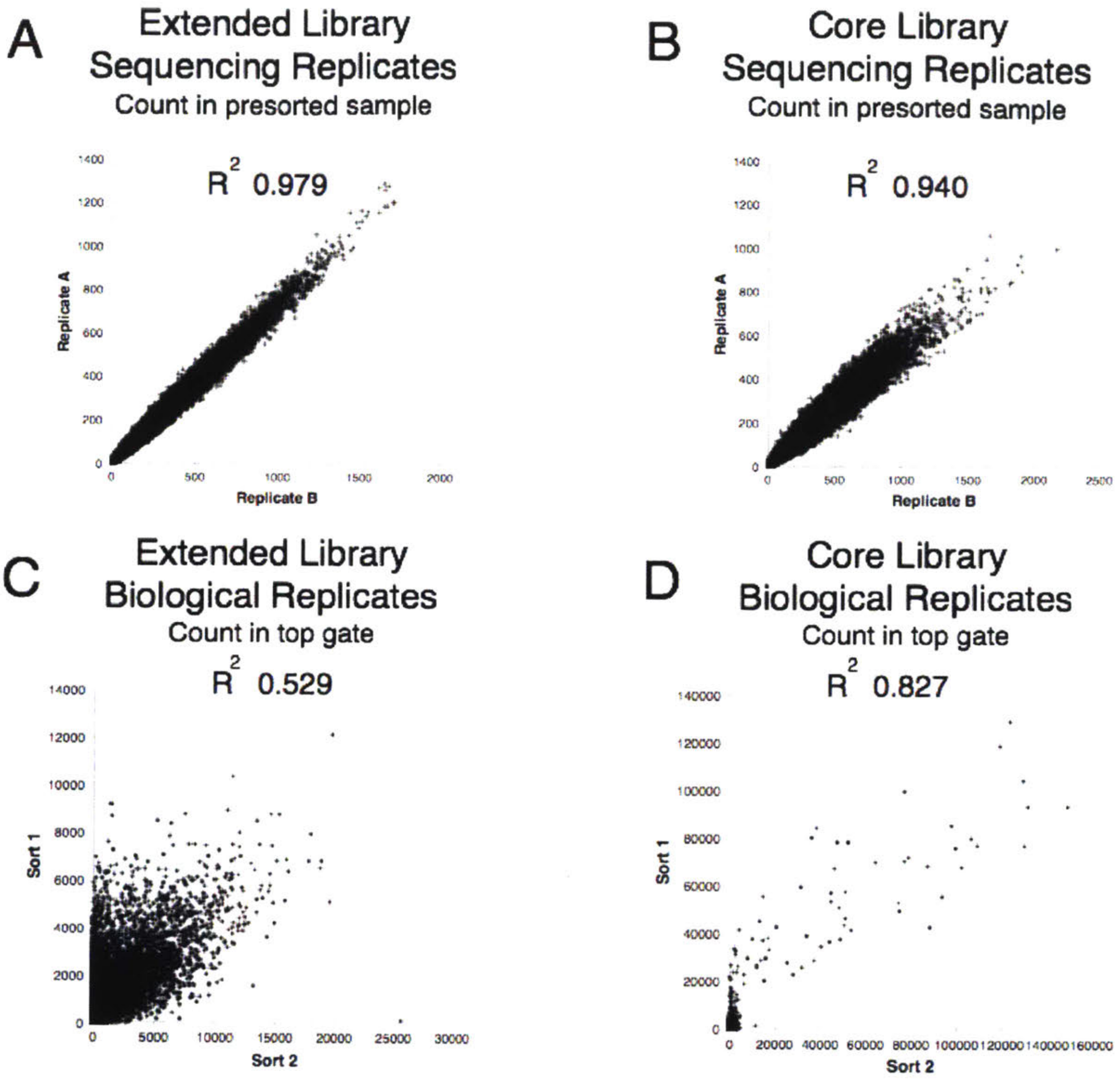


Figure 2.7: Replicates are correlated. Biological replicates are modestly correlated, while sequencing replicates show a high level of agreement. A-B: Sequencing replicates were performed of pre-sort library samples. C-D: Biological replicates were sequenced for the highest fluorescence gate sample.

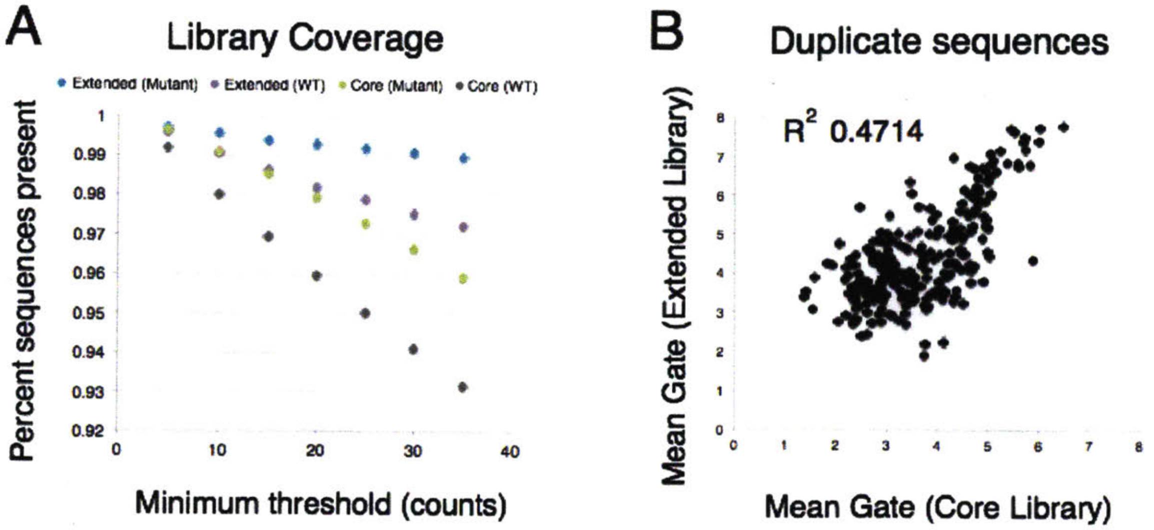


Figure 2.8: Library coverage and inter-library validation. A: Coverage is plotted against minimum threshold; library coverage is above 99% for all library samples with a minimum threshold value of 5, and decreases gradually for all samples as threshold increases. B: The 256 sequences that vary only the left half-site are found in both libraries. These internal controls agree, within some noise, at a correlation coefficient of 0.4714.

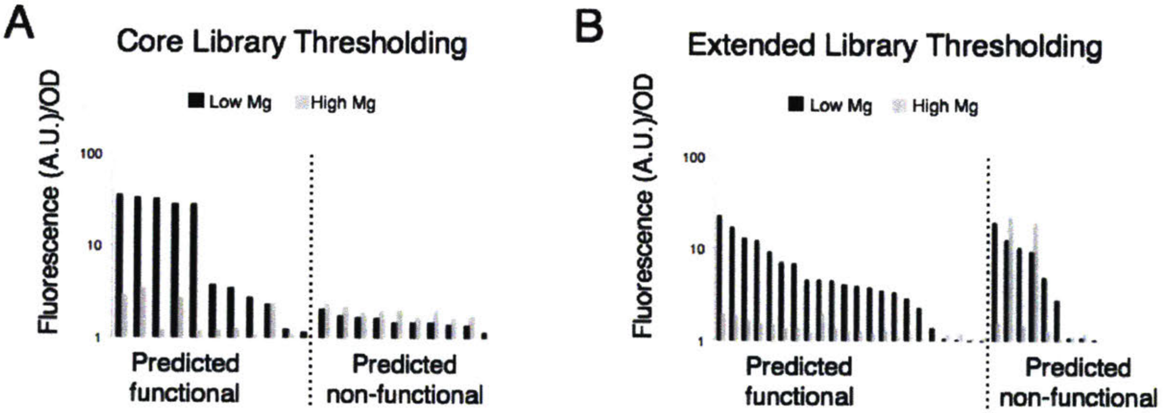


Figure 2.9: Constructed mutants validate chosen thresholds. Activation and dynamic range thresholds were validated by independently building 52 mutants, yielding an overall false positive rate of 16% and a false negative rate of 13%. A: In the core binding site library, an average activated gate threshold of 6.5 and a dynamic range threshold of 1.5 gates yields 2 false positives and no false negatives. B: For the extended library, an average gate threshold of 7.0 and a dynamic range threshold of 3 results in four false positives and four false negatives.

Appropriate activation and repression shape sequence space

Sequence space is often modeled using high-throughput binding data. Transcription factor functionality is assumed to increase with binding affinity. To mimic this kind of analysis, we selected sequences whose average FACS gate under activating conditions was > 6.5 and then built a force-directed graph in which each node represents a single sequence, and each edge represents nodes that differ by one nucleotide. Nodes were colored by their proximity, in terms of mutational steps, to the wild-type *mgrB* promoter (Figure 2.8A). This model reveals one major hairball-style cluster and one smaller satellite cluster. Notably, the wild-type sequence is found in the smaller protruding satellite cluster.

Many of the sequences in this initial graph had high fluorescence in both the activating and repressing conditions. These constitutively ON sequences have very low dynamic ranges and thus would not be useful in regulating genes that need to be transcribed differently depending on the phosphorylation state of PhoP. Appropriate, signal-dependent activation or repression is central to a functional transcription factor/DNA interaction—constitutive effects can be achieved simply by tuning the promoter without using a regulatory protein. Thus, to model functional sequence space, we isolated variants that had an average active gate >6.5 and a dynamic range >1.5, *i.e.* the difference in the mean gate in activating and repressing conditions was > 1.5.

Imposing this additional threshold produced a substantially sparser map of sequence space. Interestingly though, the constitutively ON sequences that were eliminated had not been uniformly distributed; 80% of the sequences from the initial large cluster were eliminated as constitutively ON, but only 12% of the smaller one. This observation means that properly functioning sequences in the initial, larger cluster have many more neighbors that are constitutively ON than do functioning sequences in the smaller cluster. This observation may imply that the functional sequences in the smaller cluster are effectively more robust to mutations in the sense that individual mutations to sequences in the smaller cluster are, probabilistically, more likely to retain function compared to sequences in the larger cluster.

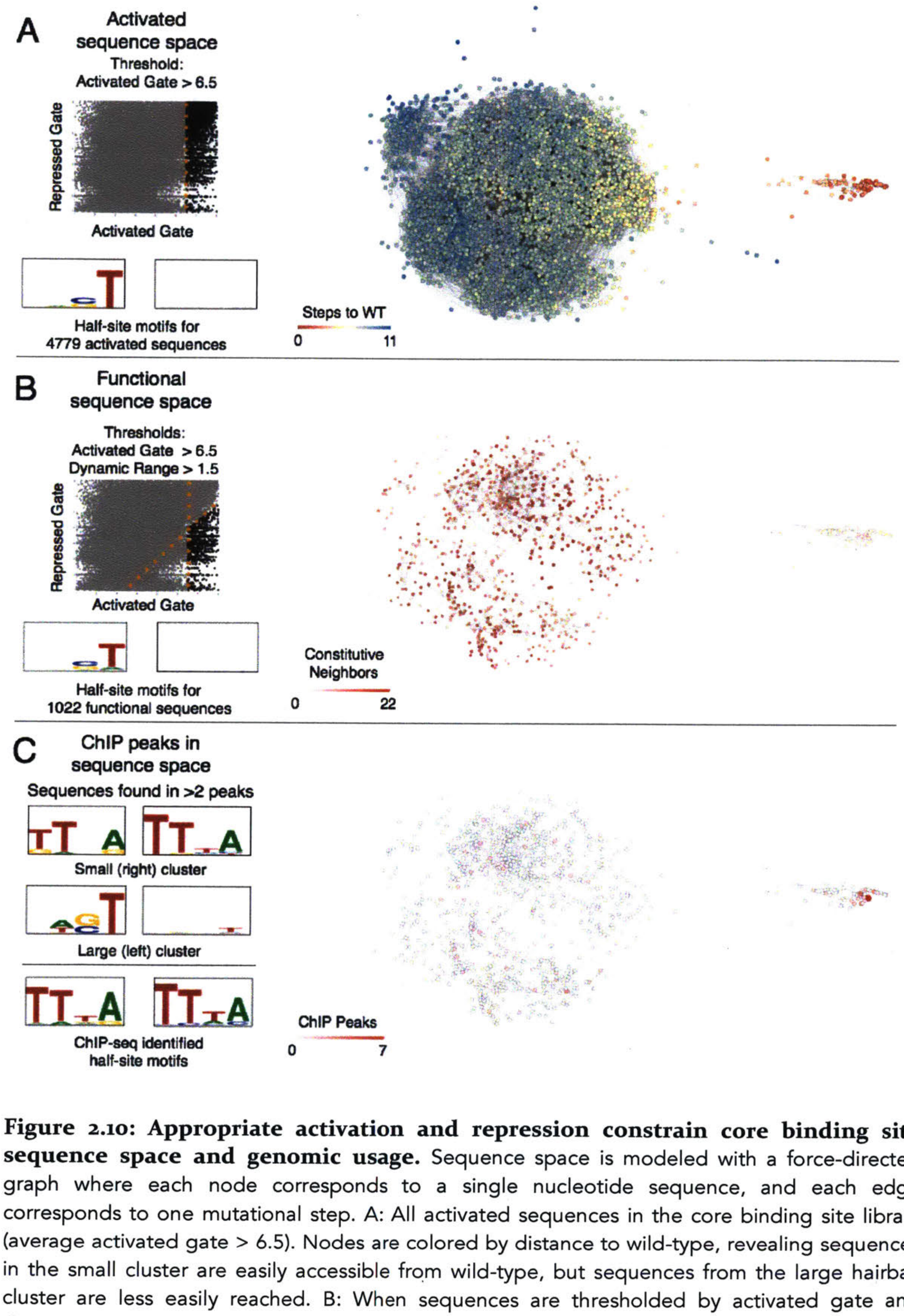


Figure 2.10: Appropriate activation and repression constrain core binding site sequence space and genomic usage. Sequence space is modeled with a force-directed graph where each node corresponds to a single nucleotide sequence, and each edge corresponds to one mutational step. A: All activated sequences in the core binding site library (average activated gate > 6.5). Nodes are colored by distance to wild-type, revealing sequences in the small cluster are easily accessible from wild-type, but sequences from the large hairball cluster are less easily reached. B: When sequences are thresholded by activated gate and dynamic range (1.5 gates), the large cluster becomes much more sparse. Nodes are shaded red

by the number of constitutively active neighbors they have (which are no longer visible on the graph). The sequences of the large cluster have many more constitutive neighbors. Summary sequence logos in A and B were dominated by the larger left cluster. C: Nodes are colored by number of ChIP peaks, and nodes present at two or more ChIP peaks are pooled, by cluster, and analyzed by weblogo. The ChIP-seq logo was determined by MEME. This reveals the right cluster corresponds to the canonical PhoP binding site, while the left cluster has a different sequence character.

Regions of functional sequence space with more constitutively active neighbors are avoided by PhoP regulon

To see if the uneven distribution of constitutively ON sequences had any biological consequences or relevance, we calculated how many times each functional sequence in the force-directed graph appears in a ChIP peak, with node color reflecting this value (Figure 2.10C). This analysis makes it clear that sequences that recruit PhoP *in vivo* are significantly overrepresented in the smaller, wild-type containing cluster (bootstrapping a p-value by randomly sampling from peak usage gives $p < 10^{-5}$). This analysis suggests that PhoP binding sites may have evolved not only to optimize affinity and induction/dynamic range, but that they are potentially optimized for mutational robustness against constitutive activation.

Predicted functionality impacts sequences' genomic distributions

If PhoP binding sites are constrained to avoid mutational proximity to constitutively ON sequences, we wondered if sequences capable of binding PhoP might be distributed differently in the *E. coli* genome. We hypothesized that, outside its targeted promoters, PhoP binding sites would be very infrequent, *i.e.* de-enriched elsewhere in the genome, to reduce spurious binding because this would potentially titrate PhoP away from its intended targets or interfere with the functions of other DNA-binding proteins.

To test this hypothesis, we selected three populations of sequences: OFF/OFF sequences that are constitutively inactive, ON/ON sequences that are constitutively active, and ON/OFF sequences that are functional. We found that, indeed, sequences from each group are distributed differently across the genome. Within promoter regions, ON/OFF and ON/ON sequences are slightly enriched relative to OFF/OFF sequences (Table 1), *i.e.* they occur, on average, more often than OFF/OFF sequences. Outside promoter regions, ON/OFF and ON/ON sequences are significantly de-enriched relative to OFF/OFF sequences by about 10%. These results suggest that not only have we correctly identified sequence functionality through our library screen, but that PhoP binding sites are in fact avoided by non-promoter regions of the genome. This finding, in addition to the distribution of ChIP peaks in sequence space (Figure 2.10C), further suggests that evolution may not only favor binding sites with more functional neighbors in sequence space and fewer constitutively ON neighbors but also de-enrich binding sites outside of promoters.

	Mean Occurrences	P-value (vs. Off/Off)
Promoter Regions		
Off/Off	7.13	-
On/On	7.84	0.0145
On/Off	7.97	0.0269
Non-Promoter Regions		
Off/Off	42.6	-
On/On	37.1	4.09E-10
On/Off	38.6	3.27E-04

Table 2.1: Genomic distribution of PhoP binding sites. Average number of occurrences in either promoter or non-promoter genomic sequences for 8-mers defined as nonfunctional (Off/Off), constitutively on (On/On), or functional (On/Off) in our PhoP core binding site library.

Extended half-site nucleotides form a denser sequence space less constrained by repression

To determine how nucleotides across an entire extended half-site interact and constrain sequence space, I performed similar analyses as above on my extended half-site library. I built force-directed graphs for (i) those sequences identified as functional and constitutively ON (Figure 2.11A) or (ii) only sequences identified as functional (Figure 2.11B). Notably, this library, in which I randomized one entire extended binding site, resulted in many fewer constitutively ON sequences compared to the core binding site library in which both half-sites were randomized. When the 14% of constitutively ON sequences were removed, much of sequence space was still highly connected, and no region had a particularly high enrichment of nonfunctional neighbors (Figure 2.11B).

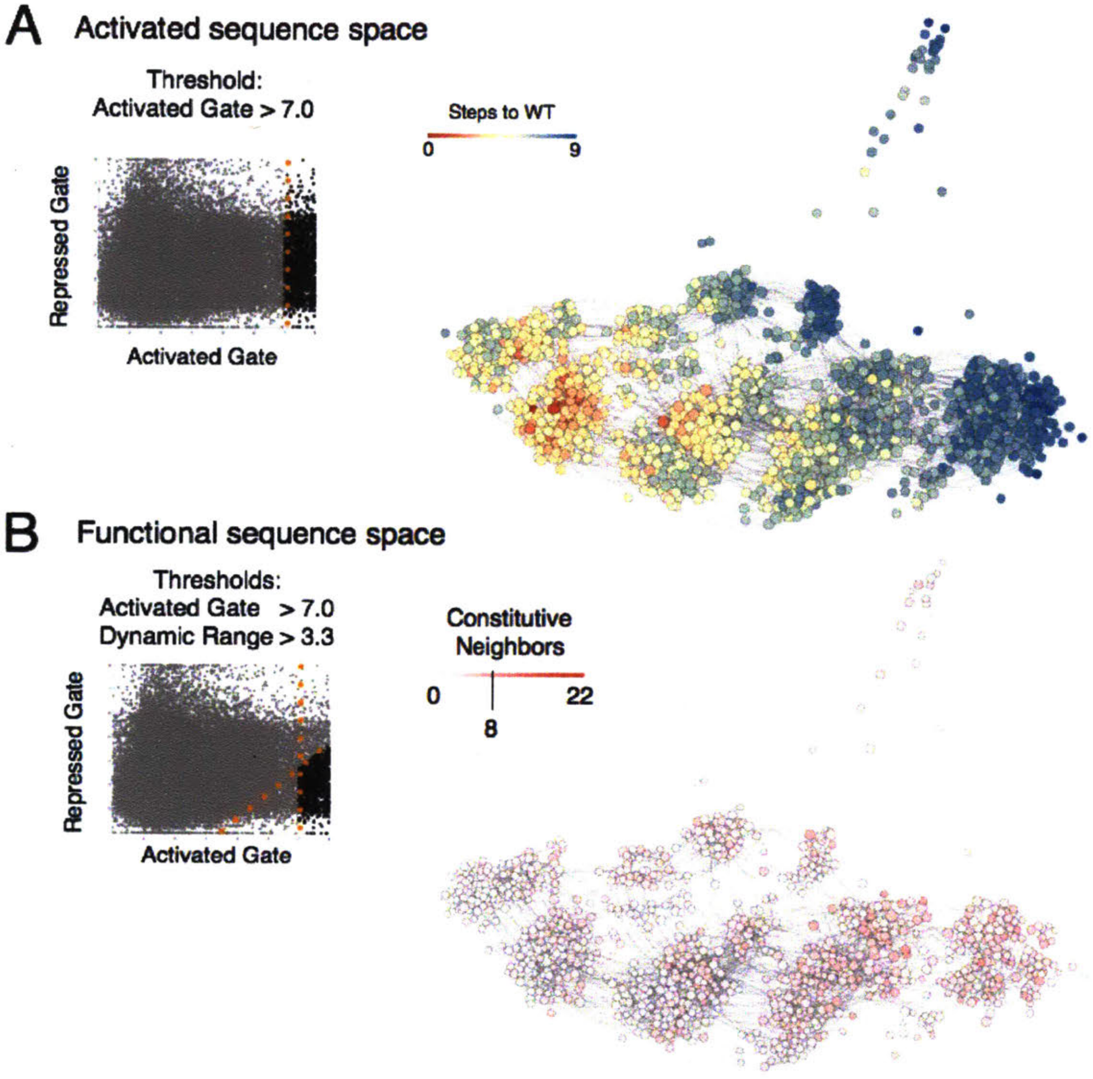


Figure 2.11: Extended binding site nucleotides form a denser sequence space less constrained by repression. Sequence space is again modeled with a force-directed graph where each node corresponds to a single nucleotide sequence, and each edge corresponds to one mutational step. A: Sequences with a mean activated gate greater than 7. Node color corresponds to the number of steps in sequence space from wild-type. B: Sequences whose mean activated gate is greater than 7 and dynamic range is greater than 3.3. Nodes are colored by the number of constitutive neighbors.

Functional sequences from this library formed 8 clusters, which differ from one another in sequence character and the positioning of the characteristic 'TT' repeat or the presence of a terminal A, among other binding-site elements (Figure 2.12).

The emergence of these extended half-site clusters suggests a possible interdependence between nucleotides within the PhoP binding site.

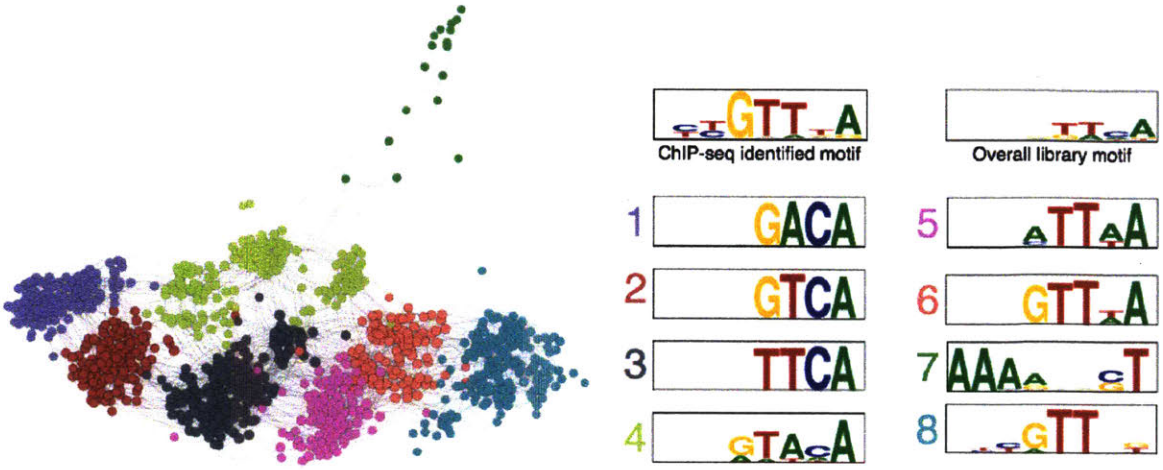


Figure 2.12: Extended binding site clusters reveal functional sequences with different sequence character. Sequences are colored by clusters determined using Gephi's OpenOrd algorithm. The ChIP-seq motif was determined by MEME. Each cluster's sequence logo was determined with weblogo. Several classes of sequences emerge, with either a central TT element (clusters 3, 5, 6, 8) a terminal A (clusters 1, 2, 3, 4, 5, 6) or a penultimate C (clusters 1, 2, 3, 7), for example. This suggests interdependency between nucleotides.

Nucleotides are correlated between and within binding-sites

To determine how mutations at one position in a PhoP binding site might impact mutations at other positions, we calculated the mutual information between each pair of nucleotide positions in our set of functional sequences. These values were normalized by calculating the mutual information for a shuffled list of the respective nucleotides one thousand times to control for differences in nucleotide distribution. The normalized mutual information signal was low for the core binding site library nucleotides, probably in part because functional variants from this library have lower overall diversity, but there was some correlation within and between binding sites (Figure 2.13).

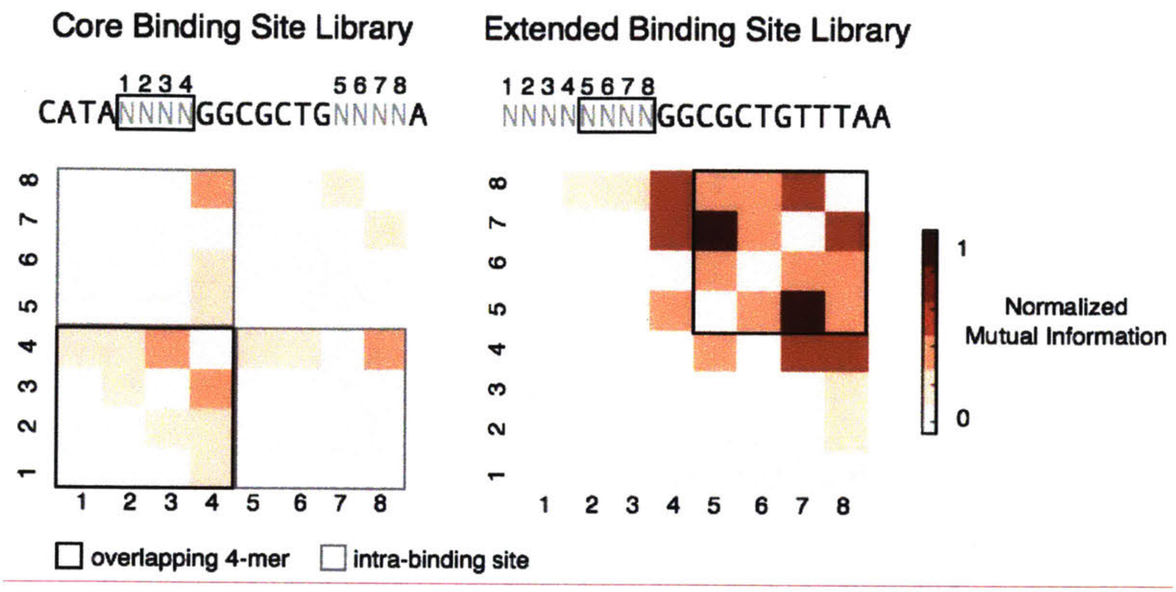


Figure 2.13: Mutual information reveals nucleotides influence positions within and across binding sites. Mutual information was calculated for each pair of positions within the functional sequences of each library. Core binding site mutual information is, overall, lower than extended binding site information. This is likely due to the lower diversity of the core binding site library. Positions 1-4 in the core binding site correspond to the exact same nucleotides as positions 5-8 in the extended binding site (black boxes). The core binding site shows higher mutual information between sites in the same half-site (black box) than across half-sites (grey boxes). This suggests that overall, nucleotides may influence positions within the same binding-site more than across binding-sites. Position 4 in the core binding site library (which is also position 8 in the extended binding site library) has the most impact on nucleotides both within its half-site and across half-sites.

Position four in the left core binding site had the highest correlation to other positions, so we investigated how changes at that position might influence overall sequence characteristics (Figure 2.14D). While position four is dominated by T in functional variants from the core binding site library (875/1022) and A in extended binding site variants (838/1077), sequences with each nucleotide are present and show wide variation in how they restrict the remaining positions. Interestingly, some sites are constrained within the extended half-site library but not the core binding site library. For example, when G is fixed at position 4, the beginning of the core binding site is fairly constrained unless the right binding site is also allowed to vary. This correlation between binding sites led us to further investigate how each binding site half contributes to binding.

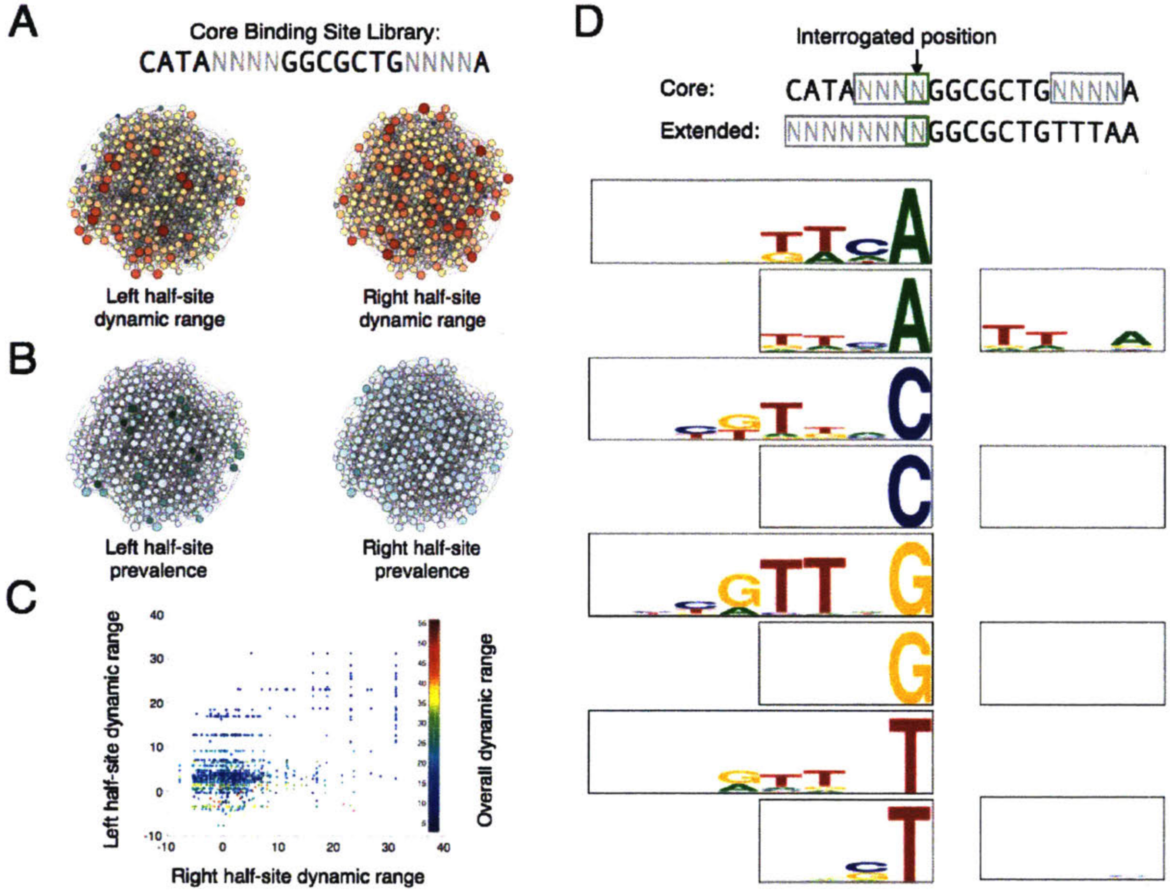


Figure 2.14: Nucleotides are correlated between and within binding-sites, Right and left binding site functionality is context dependent and epistatic. All 256 possible 4-nucleotide sequences (nodes) and their connections through single point mutations (edges) were mapped to a force-directed diagram. A: In the context of a wild-type opposite half-site, each sequence's designated left or right dynamic range determines node size and color. B: Nodes are colored by each half-site's prevalence in the list of overall functional sequences and sized by half-site score as in A. C: Right half-site dynamic range is plotted against left half-site dynamic range, marker color corresponds to overall binding site dynamic range. High dynamic range binding sites are not the combination of two high dynamic range half sites. D: Sequence logos showing how holding the position with the most mutual information (both within and across binding sites) restrains other nucleotides.

Right and left half-site functionality is context dependent and epistatic

The binding sites of dimeric transcription factors like PhoP are often composed of direct repeats: two half-sites of similar sequence character. We established that individual nucleotide positions can impact each other within and across half-sites, but when each half-site is thought of as a unit, how do they functionally combine? To answer this question, we focused on the dynamic range of each half-site in the context of an otherwise wild-type promoter sequence. We then mapped all possible four nucleotide sequences using a force-directed graph, and colored and sized each node according to the dynamic range of the right or left half-site (Figure 2.14A). The distribution of scores revealed that sequences functional in one half-site context are not necessarily functional in the other. Using the same graph topology for all four nucleotide sequences, we then colored sequences by their prevalence within the set of 1,022 functional sequences of the core binding site library (leaving sizes as in panel A) (Figure 2.14B). In this analysis we saw that the left half-site relies on a few sequences more than others, while the right half-site appears less restrictive. Additionally, high-scoring sequences are not enriched.

To better understand how half-sites combine to form functional PhoP binding sites, we plotted the right half-site dynamic range versus the left half-site dynamic range for the 1,022 functional sequences from the core binding site library (Figure 2.14C). Each node was colored to show the overall dynamic range of the combined binding site. Many half-sites with low dynamic ranges in the context of a wild-type partner are nonetheless found among the functional library sequences. Additionally, while all of the sequences plotted are functional, those with the highest overall dynamic range appear to be restricted to left half-sites with low dynamic ranges. Asymmetric variations within the direct repeat might be due to the different nucleotides neighboring each half-site or asymmetry in the way PhoP binds DNA. In either case, our findings point to complex, non-additive inter- and intra-binding-site interactions.

Rare promiscuous variants are not enriched in bridging orthologous protein sequence spaces

Finally, to better understand how transcription factors and their regulons might coevolve, we rationally rewired the specificity of PhoP by mutating DNA-binding residues. This mutant was constructed by changing four DNA-binding specificity residues in the recognition helix of *E. coli* PhoP to match those found at the equivalent positions in *Pseudomonas aeruginosa* PhoP. These positions were chosen based on sequence alignments from known, related proteins that have been structurally characterized in complex with DNA (Figure 2.15A) (Wisedchaisri et al. 2005, Maris et al. 2002, Canals et al. 2012, Narayanan et al. 2014). Of the residues predicted to be involved in molecular recognition of a specific DNA sequence, five differed between *E. coli* and *P. aeruginosa*, four of which were located on the recognition helix.

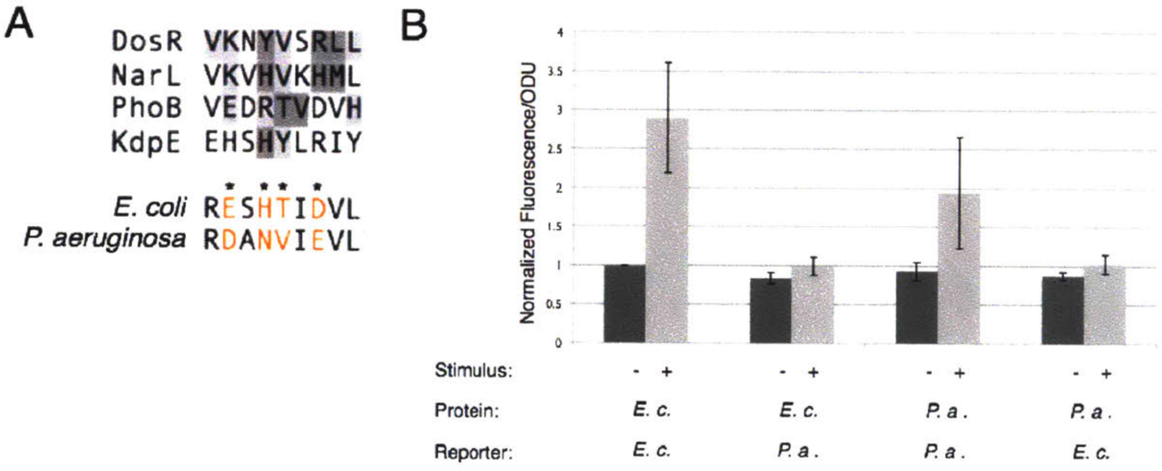


Figure 2.15: Four specificity mutations can functionally rewire PhoP. A: The noncontiguous sequence alignment for specificity determining residues in four winged helix-turn-helix family transcription factors that have been co-crystallized with DNA. Light grey shading indicates an interaction with DNA, darker grey indicates a major groove interaction with a base. *E. coli* and *P. aeruginosa* residues are aligned for the relevant positions, and sites where they differ are highlighted in orange. B: By swapping the four highlighted residues, specificity can be rewired from an *E. coli slyB* transcriptional reporter to a *P. aeruginosa slyB* transcriptional reporter.

The *E. coli* PhoP mutant with four recognition helix residues replaced by their *P. aeruginosa* counterparts could activate a *Pseudomonas slyB* transcriptional fusion in *E. coli*, but was incapable of activating an *E. coli slyB* transcriptional fusion (Figure 2.15B). Although insulated based on these two reporters, the protein was likely not entirely orthogonal, as it can partially rescue the low magnesium growth defects of an *E. coli* strain lacking wild-type PhoP.

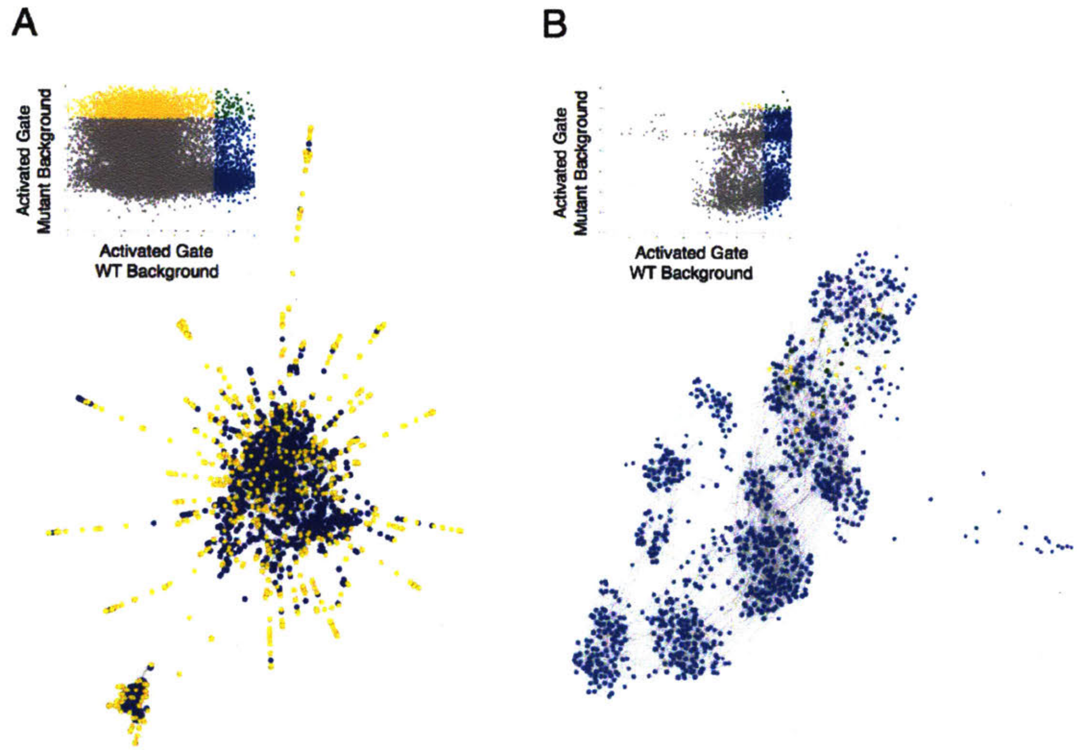


Figure 2.16: Rare promiscuous variants are not enriched in bridging orthologous protein sequence spaces. Sequence space is modeled with a force-directed diagram in Gephi where each node corresponds to a single nucleotide sequence, and each edge corresponds to one mutational step. Nodes are colored by promiscuity: yellow nodes specifically interact with the mutant PhoP, blue nodes specifically interact with wild-type PhoP, and green nodes promiscuously interact with both. A: Core binding site sequences that interact with either wild-type or mutant PhoP, or promiscuously interact with both. Many direct paths exist between blue and yellow nodes. B: Extended binding site sequences that interact with either wild-type or mutant PhoP, or promiscuously interact with both. Very few sequences can interact with the orthogonal PhoP.

To characterize the binding specificity of our quadruple mutant PhoP, we transformed our plasmid-based binding site libraries into a strain harboring the genomically-encoded PhoP mutant. We found that in each library context, very few sequences are responsive to both PhoPs (Figure 2.16). We also found that in the extended binding site library (where the right half-site is held constant) very few sequences can be activated by the orthologous PhoP (Figure 2.16B), suggesting that rewiring promoters to respond to a transcription factor with a different specificity requires changes in both half-sites.

Protein coevolution has been hypothesized to involve important promiscuous intermediates (Aakre et al. 2015). To evaluate the role of promiscuous sequences in the coevolution of a DNA binding site, we divided our sequences into insulated and promiscuous groups, and calculated how many average connections each kind of sequence has. Overall, promiscuous sequences have fewer connections in the core binding site sequence space and about the same number of connections in the extended binding site sequence space (Table 2.2).

Core Binding Site	Average Connections
All	13.33
WT *E.c.* PhoP	13.36
P.a.-based PhoP	10.76
Promiscuous	4.67
Extended Binding Site	
All	9.59
WT *E.c.* PhoP	9.62
P.a.-based PhoP	7.55
Promiscuous	9.70

Table 2.2: Average number of connections in combined orthogonal sequence space. The average number of connections to a given sequence, as categorized by promiscuity, for core and extended binding-site libraries.

For the core binding site library, many variants function with the orthologous PhoP, but they do not cluster together in sequence space. Rather, they form extended chains branching off from the main cluster (Figure 2.16A). Thus, there are many paths between the orthologous DNA sequence spaces, and very few of them use promiscuous variants as a bridge. This suggests that during evolution, as a transcription factor mutates, there may be many paths available to its DNA binding site partners to restore functionality without needing a promiscuous intermediate.

Discussion

By comprehensively mapping the functional sequence space of PhoP, I have shown that this response regulator does not utilize all areas of functional sequence space evenly. Specifically, PhoP's regulon is largely restricted to sequences with few constitutively active neighbors. Unlike previous binding-based datasets, our *in vivo* activation and repression transcription data yields a functional picture of sequence space, which is consistent with the genomic distribution of nonfunctional, functional, and constitutively active sequences. All of this suggests that evolution may not optimize for the highest affinity transcription factor binding site, and in fact may avoid certain functional areas of sequence space because they have too many constitutively active neighbors. This idea has implications for interpreting data from affinity-based transcription factor assays. While many transcription factors' regulons appear in close agreement with the results of high-throughput *in vitro* binding studies, there are many transcription factors that have only been studied using these *in vitro* techniques, so the *bona fide* binding site consensus *in vivo* could be substantially different. The only way to determine if PhoP is typical or more unique in its use of functional sequence space is to apply *in vivo* library-based screens like the one here to more transcription factors.

My library data reveals a complex interplay between nucleotides both within and between half-sites. It is clear that nucleotides are not, as is generally modeled in the position weight matrices used to bioinformatically find binding sites, limited to very short range interactions. Nevertheless, these simple models have proven highly useful in discovering binding sites, including in the case of PhoP. Our ChIP data suggests, however, that a few of these predicted sites may not be biologically functional. Taken together, the results of this study suggest that while position weight matrices and other methods that assume nucleotide independence are an important first pass in the analysis of transcription factor binding sites, more complex analyses might be useful for determining the biological impact of various subtle binding site and binding site context mutations.

My experiments with an orthogonal PhoP suggest that DNA sequences can mutate from the interaction space of one protein to another without requiring promiscuous intermediates, as may be required for some protein-protein interactions (Aakre et al. 2015). While promiscuity was rare, there were many direct paths from the insulated sequences of one hypothetical regulon to another. This could suggest that coevolution between transcription factors and their regulons is more restricted by the mutational landscape of the protein than the DNA binding site.

An overarching theme of my results, in fact, is the contrast between DNA and protein sequence space. Much of the work here reinforces a model of DNA sequence space that is relatively small and crowded compared with previous characterizations of vast, undersampled protein sequence space. Given the size of the genome and the size of transcription factor binding sites, it is highly likely that spurious binding sites simply arise by mutational drift. In fact, we find here that the *E. coli* genome actively avoids PhoP binding sequences (both functional and constitutive) outside of promoter regions—drift like this is being selected against. Similarly, DNA sequence space seems quite crowded when comparing binding profiles of wild-type PhoP and an orthogonal PhoP, where many direct paths appear accessible between two fairly insulated regulons. Vast protein sequence spaces may be constrained by the dearth of mutational paths with functional intermediates, but it appears that for small DNA sequence spaces, genomic oversampling itself leads to constraints in which some regions of sequence space are more or less biologically useful.

Experimental Procedures

Growth media and strains

E. coli was grown in M9 minimal medium (Ix M9 salts, 100 pM CaCl 2 0.2% glucose, and 0.1% casamino acids) with MgSO4 added at the concentrations indicated. Antibiotics were added for plasmid maintenance at the following concentrations: carbenicillin, 50 µg/mL; kanamycin, 30µg/mL; spectinomycin, 50µg/mL.

The wild-type strain used throughout is MG1655. The PhoPQ- strain and the orthogonal mutant strain are genomic mutants of MG1655. PhoPQ- was constructed through viral transduction from TIM175(Podgornaia et al. 2015). The scarless orthogonal mutant strain was constructed using homologous recombination with an inducible Cas9 counterselection against wild-type cells (Reisch and Prather, in preparation). The C-terminally 3x-flag tagged PhoP was constructed on the Psc101-based plasmid bearing the PhoPQ locus under its native promoter through 'round-the-horn PCR.

Library variants and individual point mutants were constructed in a low-copy pSC101-based plasmid containing 250 nucleotides of the mgrB promoter upstream of YFP. Mutations were introduced using PCR based site-directed mutagenesis either by QuickChange (Papworth et al. 1996) or 'round-the-horn.

ChIP-seq

ChIP-seq was performed as described previously (Haakonsen et al. 2015). Briefly, Mid-log cells were harvested, fixed, and washed with PBS. Cells were lysed via sonication, cleared by centrifugation, and supernatants were normalized with a

Bradford assay. Normalized samples were incubated 1 hour with a flag antibody (Sigma Aldrich) and then 1 hour with pre-blocked Protein-A dynabeads. Beads were washed as described previously and eluted. Samples were incubated overnight with RNase, and 2 hours with Proteinas-K. Then samples were phenol-chloroform extracted.

ChIP-seq data analysis

Reads were mapped onto the *E. coli* MG1655 genome sequence. Peaks were analyzed using the MACS software package (Zhang et al. 2008) and called when above 5 standard deviations above the control sample mean. 157 unique peaks were identified. Peaks were then considered if they were present in two out of three activated replicates. ChIP consensus motif was identified with MEME.

Library construction

The core binding site and extended half-site libraries were constructed using the same method as previously published (Podgornaia et al. 2015). The library plasmid was prepared by first inserting *ccdB-camR*, flanked by BsmBI sites, in the promoter region of an mgrB-YFP transcriptional fusion within a pSC101 origin plasmid. DNA fragments containing BsmBI sites and randomized positions of interest were synthesized by DNA 2.0. The fragment and plasmid were digested for 1 hour with BsmBI and ligated in a 1:1 ratio (250 fmol) 16 hours at 16°C. After ligation, the reaction was dialyzed 2 hours on a 0.025um filter (Millipore). Ligated, dialyzed plasmids were transformed into MegaXDH10B ultraelectrocompetent cells (Thermo Fisher) for a total of 10^8 transformants. Cells recovered for 1 hour in SOC and then diluted into LB. OD was monitored until cells had undergone approximately eight doublings, at which time DNA was harvested.

The plasmid library was transformed into electrocompetent Mg1655 , MG1655 PQ-, and orthogonal PhoP cells prepared using density-step centrifugation described previously (Warren 2011).

Flow cytometry and sorting

Library aliquots or individual mutants were thawed, diluted to an OD6000.001 in the media described, and grown for six hours prior to flow cytometry or sorting. After six hours, cultures were diluted to an OD6000.2 in the same growth media and placed on ice when not loaded on the FACS Aria. Eight equally spaced gates were set and re-calibrated each sort to have 2.5% of a population of wild-type sequence control in the highest gate. A million samples were gate were collected, plated on large agar plates, and grown to colonies overnight. Plates were scraped and DNA was extracted.

Illumina sample preparation, sequencing, and analysis

DNA samples from sorted libraries were barcoded using one of 24 unique barcoding primers designed with Illumina adaptor sequences. Barcoding was achieved with 20 PCR cycles. Samples were then gel purified and quantified with a NanoDrop. Samples were multiplexed (8-10 samples per run) and sequenced via Illumina MiSeq with the addition of phiX DNA for diversity.

Sequencing data analysis

Reads were demultiplexed and quality filtered to remove any sequences that did not match a barcode exactly or contained any mismatch in non-randomized nucleotides. Counts were normalized to reads per barcode. Sequences with fewer than five reads were omitted (less than 1% of each library). Average activated gate and average repressed gate were extrapolated and used to calculate a dynamic range for each sequence. Thresholds were chosen for average activated gate and dynamic range based on the behavior of wild type and five constructed point mutants. These thresholds were verified by constructing 52 variants and testing their behavior in fluorescence per OD assays. These thresholds were applied to the whole library dataset, and structural graphs were constructed in MATLAB. These graphs were then imported into Gephi, where the Force Atlas 2 algorithm was used to generate a force-directed diagram. Disconnected nodes were omitted in visualizations.

Fluorescence per OD assay

Assays were performed in a BioTek 96-well plate reader with clear flat-bottomed plates. Cells were grown as for flow cytometry, then diluted 1:100 in their respective media in the 95-well plate. Fluorescence readings were taken using the YFP filter set, and OD readings were taken by measuring scatter at 600nm. All samples were normalized to the same low-fluorescence mutant in the high magnesium (50mM) condition.

References

Aakre, C. D., Herrou, J., Phung, T. N., Perchuk, B. S., Crosson, S., & Laub, M. T. (2015). Evolving New Protein-Protein Interaction Specificity through Promiscuous Intermediates. *Cell, 163*(3), 594-606.

Alm, E., Huang, K., & Arkin, A. (2006). The Evolution of Two-Component Systems in Bacteria Reveals Different Strategies for Niche Adaptation. *PLoS Computational Biology, 2*(11), e143.

Anderson, W. F., Cygler, M., Vandonselaar, M., Ohlendorf, D. H., Matthews, B. W., Kim, J., & Takeda, Y. (1983). Crystallographic data for complexes of the Cro repressor with DNA. *Journal of Molecular Biology, 168*(4), 903–906.

Blat, Y., & Kleckner, N. (1999). Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region. *Cell, 98*(2), 249–259.

Bougdour, A., Cunning, C., Baptiste, P. J., Elliott, T., & Gottesman, S. (2008). Multiple pathways for regulation of sigmaS (RpoS) stability in Escherichia coli via the action of multiple anti-adaptors. *Molecular Microbiology, 68*(2), 298–313.

Bulyk, M. L. (2007). Protein binding microarrays for the characterization of DNA-protein interactions. *Advances in Biochemical Engineering/Biotechnology, 104*, 65–85.

Canals, A., Blanco, A. G., & Coll, M. (2012). σ70 and PhoB activator: getting a better grip. *Transcription, 3*(4), 160–164.

Capra, E. J., Perchuk, B. S., Skerker, J. M., & Laub, M. T. (2012). Adaptive mutations that prevent crosstalk enable the expansion of paralogous signaling protein families. *Cell, 150*(1), 222–232.

de Mendoza, A., Sebe-Pedros, A., Sestak, M. S., Matejcic, M., Torruella, G., Domazet-Loso, T., & Ruiz-Trillo, I. (2013). Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *Proceedings of the National Academy of Sciences, 110*(50), E4858–E4866.

Eguchi, Y., Ishii, E., Hata, K., & Utsumi, R. (2011). Regulation of acid resistance by connectors of two-component signal transduction systems in Escherichia coli. *Journal of Bacteriology, 193*(5), 1222–1228.

Groisman, E. A. (2001). The Pleiotropic Two-Component Regulatory System PhoP-PhoQ. *Journal of Bacteriology, 183*(6), 1835–1842.

Haakonsen, D. L., Yuan, A. H., & Laub, M. T. (2015). The bacterial cell cycle regulator GcrA is a σ 70 cofactor that drives gene expression from a subset of methylated promoters. Genes & Development, 29(21), 2272–2286.

Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science, 316(5830), 1497–1502.

Kaleta, C., Göhler, A., Schuster, S., Jahreis, K., Guthke, R., & Nikolajewa, S. (2010). Integrative inference of gene-regulatory networks in Escherichia coli using information theoretic concepts and sequence analysis. *BMC Systems Biology, 4*, 116.

Kato, A., Tanabe, H., & Utsumi, R. (1999). Molecular characterization of the PhoP-PhoQ two-component system in Escherichia coli K-12: identification of extracellular Mg2+-responsive promoters. *Journal of Bacteriology, 181*(17), 5516–5520.

Lippa, A. M., & Goulian, M. (2009). Feedback Inhibition in the PhoQ/PhoP Signaling System by a Membrane Peptide. *PLoS Genetics, 5*(12), e1000788.

Maerkl, S. J., & Quake, S. R. (2007). A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors. *Science, 315*(5809), 233–237.

Maris, A. E., Sawaya, M. R., Kaczor-Grzeskowiak, M., Jarvis, M. R., Bearson, S. M. D., Kopka, M. L., ... Dickerson, R. E. (2002). Dimerization allows DNA target site recognition by the NarL response regulator. *Nature Structural Biology*, *9*(10), 771–778.

Marniemi, J., & Parkki, M. G. (1975). Radiochemical assay of glutathione S-epoxide transferase and its enhancement by phenobarbital in rat liver in vivo. *Biochemical Pharmacology*, *24*(17), 1569–1572.

Maynard Smith, J. (1970). Natural Selection and the Concept of a Protein Space. *Nature*, *225*(5232), 563–564.

McKay, D. B., & Steitz, T. A. (1981). Structure of catabolite gene activator protein at 2.9 A resolution suggests binding to left-handed B-DNA. *Nature*, *290*(5809), 744–749.

Minagawa, S., Ogasawara, H., Kato, A., Yamamoto, K., Eguchi, Y., Oshima, T., ... Utsumi, R. (2003). Identification and molecular characterization of the Mg2+ stimulon of Escherichia coli. *Journal of Bacteriology*, *185*(13), 3696–3702.

Monsieurs, P., De Keersmaecker, S., Navarre, W. W., Bader, M. W., De Smet, F., McClelland, M., ... Marchal, K. (2005). Comparison of the PhoPQ Regulon in Escherichia coli and Salmonella typhimurium. *Journal of Molecular Evolution*, *60*(4), 462–474.

Montero, M., Almagro, G., Eydallin, G., Viale, A. M., Muñoz, F. J., Bahaji, A., ... Pozueta-Romero, J. (2011). Escherichia coli glycogen genes are organized in a single glgBXCAP transcriptional unit possessing an alternative suboperonic promoter within glgC that directs glgAP expression. *The Biochemical Journal*, *433*(1), 107–117.

Moon, K., & Gottesman, S. (2009). A PhoQ/P-regulated small RNA regulates sensitivity of Escherichia coli to antimicrobial peptides. *Molecular Microbiology*, *74*(6), 1314–1330.

Narayanan, A., Kumar, S., Evrard, A. N., Paul, L. N., & Yernool, D. A. (2014). An asymmetric heterodomain interface stabilizes a response regulator-DNA complex. *Nature Communications, 5*, 3282.

Ng, S. Y., Parker, C. S., & Roeder, R. G. (1979). Transcription of cloned Xenopus 5S RNA genes by X. laevis RNA polymerase III in reconstituted systems. *Proceedings of the National Academy of Sciences of the United States of America, 76*(1), 136–140.

Ogasawara, H., Hasegawa, A., Kanda, E., Miki, T., Yamamoto, K., & Ishihama, A. (2007). Genomic SELEX Search for Target Promoters under the Control of the PhoQP-RstBA Signal Relay Cascade. *Journal of Bacteriology, 189*(13), 4791–4799.

Perez, J. C., & Groisman, E. A. (2009). Transcription factor function and promoter architecture govern the evolution of bacterial regulons. *Proceedings of the National Academy of Sciences, 106*(11), 4319–4324.

Podgornaia, A. I., Casino, P., Marina, A., & Laub, M. T. (2013). Structural Basis of a Rationally Rewired Protein-Protein Interface Critical to Bacterial Signaling. *Structure, 21*(9), 1636–1647.

Podgornaia, A. I., & Laub, M. T. (2015). Protein evolution. Pervasive degeneracy and epistasis in a protein-protein interface. *Science (New York, N.Y.), 347*(6222), 673–677.

Salazar, M. E., Podgornaia, A. I., & Laub, M. T. (2016). The small membrane protein MgrB regulates PhoQ bifunctionality to control PhoP target gene expression dynamics: PhoQ-PhoP dynamics. *Molecular Microbiology, 102*(3), 430–445.

Skerker, J. M., Perchuk, B. S., Siryaporn, A., Lubin, E. A., Ashenberg, O., Goulian, M., & Laub, M. T. (2008). Rewiring the specificity of two-component signal transduction systems. Cell, 133(6), 1043–1054.

Stormo, G. D. (2013). Modeling the specificity of protein-DNA interactions. *Quantitative Biology, 1*(2), 115–130. https://doi.org/10.1007/s40484-013-0012-4

Tarentino, A. L., & Maley, F. (1975). A comparison of the substrate specificities of endo-beta-N-acetylglucosaminidases from Streptomyces griseus and Diplococcus Pneumoniae. *Biochemical and Biophysical Research Communications, 67*(1), 455-462.

Tjian, R. (1978). The binding site on SV40 DNA for a T antigen-related protein. *Cell, 13*(1), 165-179.

Tuerk, C., & Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science (New York, N.Y.), 249*(4968), 505-510.

Warren, D. J. (2011). Preparation of highly efficient electrocompetent Escherichia coli using glycerol/mannitol density step centrifugation. Analytical Biochemistry, 413(2), 206-207.

Wisedchaisri, G., Wu, M., Rice, A. E., Roberts, D. M., Sherman, D. R., & Hol, W. G. J. (2005). Structures of Mycobacterium tuberculosis DosR and DosR-DNA complex involved in gene activation during adaptation to hypoxic latency. *Journal of Molecular Biology, 354*(3), 630-641.

Yamamoto, K., Ogasawara, H., Fujita, N., Utsumi, R., & Ishihama, A. (2002). Novel mode of transcription regulation of divergently overlapping promoters by PhoP, the regulator of two-component system sensing external magnesium availability. *Molecular Microbiology, 45*(2), 423-438.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., ... Liu, X. S. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biology, 9(9), R137.

Zwir, I., Shin, D., Kato, A., Nishino, K., Latifi, T., Solomon, F., ... Groisman, E. A. (2005). Dissecting the PhoP regulatory network of Escherichia coli and Salmonella enterica. *Proceedings of the National Academy of Sciences, 102*(8), 2862-2867.

Chapter 3

Conclusions and future directions

Conclusions

In my thesis work, I have comprehensively mapped the functional landscape of DNA sequences that interact with the transcription factor PhoP. Harnessing the technological power of current DNA synthesis and sequencing technology, I used a functional binding assay based on previous work examining the PhoP PhoQ protein interface (Podgornia and Laub 2015). With this data, I have made a deeply sampled, high resolution map of functional sequence space that deepens our understanding of protein-DNA specificity.

By comprehensively mapping genomic PhoP binding sites, I have provided evidence for the inclusion of previously dismissed regulon members. By comparing my comprehensive library data with this ChIP-seq dataset, I have been able to show how a sequence's functionality shapes its genomic usage. Many functional sequences are not used by the PhoP regulon. My core binding site library shows an entire region of functional sequences that appear to be largely avoided not because the sequences themselves lack any particular ability to induce gene expression, but because they are within a single mutational step of sequences that are constitutively active.

The extended binding site library, which randomizes eight nucleotides encompassing the entire left half-site of the PhoP binding site, shows how nucleotides can interact within a binding site. Mutations at less-conserved nucleotide positions allow for changes at the core, more conserved nucleotides, suggesting longer range epistasis contributes to each half-site's functionality.

Data from the core binding site library, which randomizes conserved nucleotide positions in both half sites, reveals that half sites combine in a highly interdependent manner to produce functional PhoP binding sites. Not only does the context of each PhoP half-site differently constrain functionality, meaning the same four base sequence functional in one half-site may not be in the other, but the functionality of the left half-site is largely dependent on the right half-site and vise versa.

Finally, by transforming my randomized libraries into an orthogonal PhoP strain, constructed based on structural data from related transcription factors and validated with a pair of orthogonal reporters, I have been able to address some questions about how mutations in a regulator impact the sequence space available to the regulon. In the case of my two proteins, there are extremely few promiscuous DNA sequences. Rewiring and promiscuity are for the most part limited to sequences that vary at both half-sites. When this joint sequence space is mapped, we do not find two distinct clusters that interconnect via promiscuous intermediates, as might be expected, but rather many direct connections between sequences that recognize different proteins.

Taken together, this work reveals that although DNA sequence space is smaller, more crowded, and more sampled that protein sequence space, it is nonetheless constrained. While topological constraints may keep certain areas of sequence space unexplored in protein specificity sequence space, usage of DNA sequence space may be constrained more by the proximity of constitutively active sequences.

While this work has expanded our understanding of how transcription factors interact with their regulons, as is often the case in scientific research, new information generates new questions. Does PhoP have a secondary genomic binding site, recruiting it to peaks without a canonical PhoP box? If not, what is the source of these reproducible peaks? Does PhoP, for example, interact with any previously unidentified proteins? How unique is PhoP's interaction with sequence space—specifically, are other proteins' regulons constrained by avoiding sequences with constitutively active neighbors? Is this an issue only for transcription factor binding sites that straddle the -35 in a class II promoter configuration? Are there other constraints when binding sites are further upstream? And how do single and stepwise mutations impact the DNA sequence space that functionally interacts with a protein? While we have information about a quadruple mutant, evolutionary questions may require data about more intermediate steps along the path to an orthogonal protein. In the following section, I propose experiments to further our understanding of the coevolution between a transcription factor and its regulon's promoters.

Future Directions

From an enzyme catalyzing a reaction to a structural protein forming oligomers or a transcription factor binding the proteins it regulates, proteins must interact with other molecules to perform their cellular functions. When these interactions occur between genomically encoded biological molecules, both sides of the interface are subject to random mutational changes. In the case of a transcription factor-DNA interaction, my work has shown that evolution may favor certain areas of DNA sequence space over others because they are more robust to deleterious mutations. When considering mutations in the transcription factor that impact DNA-binding specificity, each member of the regulon may be differentially impacted. As a regulator and regulon traverse mutational trajectories through time, they not only must remain able to interact, but also must remain specific. A comprehensive study of the sequence space available to stepwise mutants of a transcription factor could clarify what intermediate states might be used by evolution.

While some studies have examined the binding landscapes between multiple transcription factor mutants and multiple binding site mutants, the largest number tested is 128 combinations (Anderson et al. 2015). In my thesis work, I generated data for my 65,536-member libraries against both wild-type and mutant PhoP proteins. Here, I show preliminary data about intermediate mutants between wild-type PhoP and my orthogonal quadruple mutant, and suggest experiments that could reveal how single mutational steps in a transcription factor impact which DNA sequences are available for binding.

The *slyB* gene is present in both *E. coli* and *Salmonella*, and in both cases its promoter has one PhoP binding site. By promoter architecture, it is predicted to be activated, like the *mgrB* promoter, in a class II αCTD-independent manner (Zwir et al. 2012). Using plasmid-based fluorescent reporters, I tested every intermediate mutant (on a low copy plasmid) between the *Pseudomonas*-based quadruple mutant, and the wild-type *E. coli* PhoP (Figure 3.1).

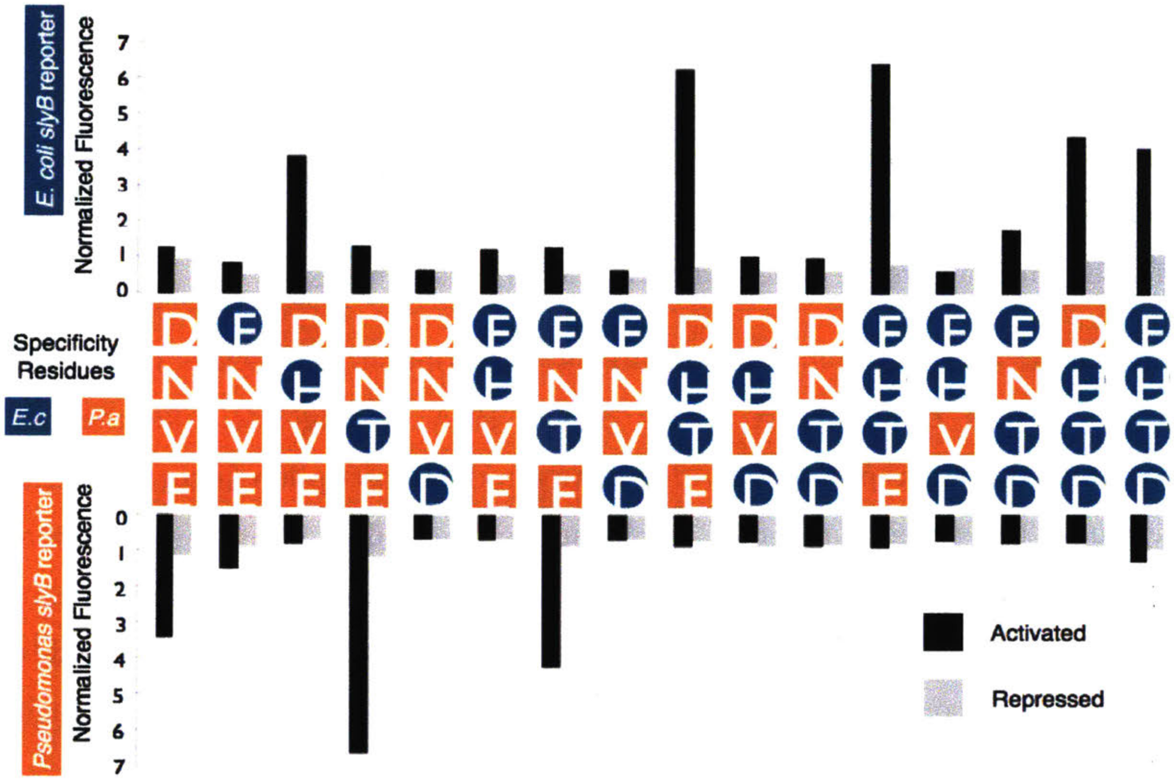


Figure 3.1: Activity of all intermediate PhoP mutants. Each PhoP mutant from the quadruple specificity residue mutant to the wild-type PhoP was tested against the *E. coli* and *Pseudomonas slyB* fluorescence reporters using flow cytometry. PhoP mutants were on a low-copy pSC1010-based plasmid, *slyB* transcriptional reporters were on a medium-copy pBR322-based plasmid.

I found that several intermediate mutants have activity, while some active neither reporter. No intermediate appears to promiscuously activate these two reporters. This is in contrast to research in toxin-antitoxin interactions, which has shown promiscuity to be common when switching from one specificity state to another (Aakre et al. 2015). However, I have only scanned 16 specificity mutants here, so even if specificity is biologically important I may simply not have looked at enough variants to find it. In the toxin-antitoxin study, for example, only about 8% of the tested library's variants were promiscuous (Aakre et al. 2015).

I propose interrogating the four PhoP DNA-binding specificity residues with a randomized library. This library can then be assayed for activity against the two *slyB* reporters. This will reveal any promiscuous variants, which can be further characterized—perhaps against the *mgrB* core binding site library described in chapter 2. Building a map of protein sequence space capable of activating one or both of the *slyB* reporters could reveal whether promiscuity, as in toxin-antitoxin interactions, is a likely pathway to new areas of sequence space, or whether transcription factors are under a different set of constraints.

Ultimately, the most complete picture of transcription factor-DNA interaction would involve combining a library of the DNA binding site, like the one constructed in my thesis work, with a library in the transcription factor specificity residues. The sheer number of combinations involved in such an experiment makes it impossible to achieve using the experimental procedures described in my thesis work: 65,536 DNA variants tested against 160,000 protein variants would be over 10 billion combinations. To be able to sample all of these interactions would require a different technical approach.

Advances using PCR in emulsion droplets point toward methods that could be applied to a library versus library interaction screen (Nakano et al. 2003). Haliburton et al. have performed linkage PCR in single cell containing emulsion droplets and detected the fitness effect of two combined mutations (2017). As a proof of concept experiment, they created a genetic interaction library for amino acid auxotrophy. Six genomic knockouts were made in different amino acid biosynthesis pathways, and then four complementation plasmids were introduced into each background. Encapsulating single cells in droplets with a microfluidic device followed by single-cell linkage PCR revealed that each knockout was only rescued by its corresponding plasmid. Expanding this kind of approach to a combination of promoter and transcription factor libraries would, in essence, give a picture of the elusive "recognition code."