Spring 5-7-2020

Discovery of a DNA-binding Consensus and Potential Genomic Regulatory Binding Sites for the Thermus thermophilus HB8 Transcriptional Regulator TTHA1359

Josiah Teague

Discovery of a DNA-binding Consensus and Potential Genomic Regulatory Binding Sites for the *Thermus thermophilus* HB8 Transcriptional Regulator TTHA1359

by

Josiah Lanier Teague

B.S. in Biochemistry

Kennesaw State University, 2018

Submitted in Partial Fulfillment of the Requirements

For the Degree of Master of Science in the

Department of Chemistry and Biochemistry

Kennesaw State University

2020

Committee Chair	Graduate Program Coordinator
Committee Member	Department Chair
Committee Member	College Dean

ACKNOWLEDGEMENTS

Thank you, Dr. Michael Van Dyke, for your invaluable advice, assistance, and patience throughout the research project. Thank you, Dr. Thomas Leeper and Dr. Melanie Griffin, for serving as members on my research project committee and for your time, support, and input on the project. To the faculty and staff in the Department of Chemistry and Biochemistry, the College of Science and Mathematics, and the Graduate College at Kennesaw State University, thank you for your support. To the National Science Foundation, thank you for financially supporting the project (Grant MCB 1714778).

To my wife, Moriah Hope, and son, Judah Lincoln, thank you both for your love and patience and the sacrifices you have both made during the completion of the research project. I love you both very, very much.

Finally, I thank You, Lord God, for the opportunity to study but a microscopic piece of Your creation, and thank You for Your providence in my life and the life of my family as the research project was pursued.

ABSTRACT

Transcription factor (TF) proteins act as molecular mechanisms that modulate the initiation of the first step in the expression of genes, gene transcription. Currently, knowledge of the DNA-binding specificities and genes regulated by many TFs, including those of well-studied model organisms such as *Escherichia coli* and *Thermus thermophilus*, remains incomplete or lacking which renders gaps in the understanding of the regulatory networks and systems biology of many organisms. Cyclic-AMP receptor protein (CRP) regulators and fumarate and nitrate reduction regulator (FNR) proteins compose the CRP/FNR superfamily of TFs, a diverse subgroup of TFs in bacteria which regulate various gene expression programs. In the present work, a reverse-genetic technique involving the combinatorial selection technique Restriction Endonuclease Protection, Selection, and Amplification (REPSA) has been applied to study TTHA1359, one of the four CRP/FNR superfamily TFs in the model organism *T. thermophilus* HB8. A TTHA1359-binding consensus, 5'-A(T/A)TGT(G/A)A(N_6)T(C/T)ACA(A/T)T-3', was identified using REPSA to select DNA sequences that TTHA1359 preferentially binds, massively parallel sequencing to acquire the sequence information of these selections, and bioinformatics to discover TTHA1359-binding motifs from the acquired sequence information. TTHA1359-binding to the identified consensus was biophysically characterized, and TTHA1359 was found to bind the identified consensus with high affinity, K_D of ~ 3.4 nM. Several potential regulatory binding sites for TTHA1359 were identified bioinformatically by mapping the TTHA1359-binding consensus to the *T.*

thermophilus HB8 genome. The findings of the present work should not only contribute to the knowledge of the DNA-binding specificity and genes regulated by TTHA1359 but also provide insight into the functionality of the applied reverse-genetic technique that should guide its future application to study other TFs.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

TF	Transcription factors
EMSA	Electrophoretic Mobility Shift Assay
PAGE	PolyAcrylamide Gel Electrophoresis
ChIP	Chromatin ImmunoPrecipitation
PBM	Protein-Binding Microarray
bp	Base pair
SELEX	Systematic Evolution of Ligands by EXponential enrichment
REPSA	Restriction Endonuclease Protection, Selection, and Amplification
IISRE	Type IIS restriction endonuclease
cAMP	Cyclic adenosine monophosphate
CRP	cAMP receptor protein
FNR	Fumarate and nitrate reductase regulatory protein
MEME	Multiple Em for Motif Elicitation
BLI	BioLayer Interferometry
FIMO	Find Identified Motif Occurrences

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1. INTRODUCTION

Chapter 1.1 Bacterial Transcription Factors

Gene expression in prokaryotic and eukaryotic organisms begins with gene transcription, and in most prokaryotes and eukaryotes, substantial regulation of gene expression transpires during the initiation, elongation, and termination steps of gene transcription. Although many molecular mechanisms mediate this regulation, transcription factor (TF) proteins act as a major mechanism of regulating transcription initiation, binding to DNA and changing rates of gene transcription initiation from promoters, the sequences positioned upstream of genes where transcription initiation transpires.

Bacterial TFs function as critical constituents of signal transduction networks, sensing cellular and environmental conditions and appropriately altering the transcriptional program. Most bacterial TFs possess two structural domains: a DNA-binding domain that recognizes specific DNA sequences and a second domain that regulates the activity of the DNA-binding domain.[1,2] For some TFs, the secondary domain directly perceives cellular or environmental conditions through the binding of ligands, perhaps small molecules or metabolites, and adjusts the activity of the DNA-binding domain accordingly.[3] For other TFs, the secondary domain indirectly senses this information through an interaction with a partner protein.[3]

Bacterial TFs act as global or local regulators depending on several aspects associated with their activities including the number of genes and the functional diversity

of the genes they regulate.[4] Generally, global TFs regulate the expression of many genes while local TFs regulate the expression of one to a few genes.[4] Some TFs solely function as activators, upregulating gene transcription, or as repressors, downregulating gene transcription. Other TFs function as both, up- or downregulating gene transcription depending on different promoter contexts. Mechanistically, repressors can downregulate transcription by inhibiting RNA polymerase holoenzymes, the transcriptional machinery, from binding at promoters by inducing structural changes in DNA to block access to promoters or by binding promoter sequences overlapping those the holoenzymes recognize such as the -35 and -10 core promoter sequence elements.[5] Activators can upregulate transcription by distorting promoter sequences to improve suboptimal spacing between core promoter elements or by interacting directly with both promoter sequences and RNA polymerase holoenzymes to recruit the transcriptional machinery to promoters.[6] In addition to these mechanisms, repressors can counter activators to downregulate transcription, and conversely, activators can antagonize repressors to upregulate transcription.[7]

Knowledge of the DNA-binding specificities and the genes regulated by bacterial TFs provides insight into not only the regulatory and biological roles of TFs but also the transcriptional regulatory networks and systems biology of bacteria. This knowledge, however, remains incomplete or lacking for many TFs in most, even model, bacteria. For example, ~ 100 of the ~ 300 encoded TFs within the ~ 4,300 genes of the well-studied model bacteria *Escherichia coli* K-12 lack as little evidence as a single TF binding site.[1,8-10] Even for *Thermus thermophilus* HB8, a model thermophilic bacteria with as few as ~ 2,200 genes and ~ 70 encoded TFs, DNA-binding specificities, genes regulated, and

biological functions have been discovered for only a handful.[11-22] Therefore, continuing to study bacterial TFs, discovering the DNA sequences they recognize, genes they regulate, and their regulatory as well as biological functions, is essential if the transcriptional regulatory networks and systems biology of *E. coli*, *T. thermophilus*, and other bacteria is to be fully comprehended.

Chapter 1.2 Functional Characterization of Bacterial Transcription Factors

Genomic screens and bioinformatic analyses initially identify and classify bacterial TFs into families based on primary sequence similarities with previously characterized DNA-binding domains, the helix-turn-helix motif being the most common.[1] Typically, each family is named after its founding family member, and the functions of TFs assigned to a family are inferable from the function of the founding member or other characterized members of the family. However, not every family member behaves or regulates the same biological processes as the founding member or other members of the family. Therefore, an experimental exploration of their DNA-binding specificities and the genes they regulate is required to understand their regulatory and biological roles.

Many *in vivo* and *in vitro* methods have been developed or adapted for such purposes, identifying DNA-binding specificities and regulated genes from various genetic, biological, biochemical, or biophysical perspectives. For example, *in vitro* techniques can identify the DNA-binding affinity and specificity of a TF, how tightly a single DNA sequence is bound and how effectively DNA sequences are differentiated by a TF respectively.[23] This specificity is frequently modeled as a position weight matrix (PWM) which in turn is often graphically depicted as a sequence logo.[24,25] This logo illustrates a TF's relative preference for each base at positions within its binding site and provides a tool to search a genome and identify potential regulatory binding sites and the genes the TF regulates. Alternatively, *in vivo* techniques attempt to identify the genes a TF regulates using more straightforward approaches such as directly isolating TFs bound

at promoters. They are particularly useful when the regulatory activity of a TF is to be studied during defined developmental stages or under various environmental conditions since they can indicate which binding sites are occupied by the TF under these conditions. Here, several classical and contemporary techniques applied to study the DNA-binding specificities and the genes regulated by TFs have been summarized.

The electrophoretic mobility shift assay (EMSA) is one of the oldest techniques applied to study protein-DNA interactions, probing these interactions by taking advantage of the impeded electrophoretic mobility of DNA complexed with protein compared to free DNA.[26,27] In a traditional EMSA, labeled DNA is incubated with protein to attain chemical equilibrium before native, or non-denaturing, polyacrylamide gel electrophoresis (PAGE) is performed followed by detection of the labeled DNA.[28] Binding interactions between the DNA and protein slows DNA movement through the gel which yields an observable gel band shift.[28] The traditional EMSA tolerates a spectrum of experimental parameters, and variations on the traditional EMSA further extend its applicational scope.[28] However, EMSA is not without drawbacks which include the effects of the gel matrix environment on the stability of protein-DNA complexes and the discontinuity of binding equilibrium upon electrophoresis.[28]

Footprinting, another classic *in vitro* technique, investigates the DNA-binding specificities of ligands such as TFs by taking advantage of the local cleavage protection ligands confer to the DNA sequences they bind compared to those they do not.[29-31] In a typical footprinting assay, test and control reactions are performed in which a cleavage agent digests a singly end-labeled DNA probe containing potential ligand-binding sites in the presence and absence of a DNA-binding ligand respectively.[30,31] The DNA probe in

the control reaction is randomly digested, but the DNA probe in the test reaction is selectively cleaved because the ligand shields bound sites from cleavage.[30,31] Consequently, after resolving test and control reaction products by PAGE, bound sites are identified by comparing the gel band pattern of the two reactions for bands that are missing in the test reaction pattern.[30,31] The sequences of these missing bands are then determined by comparing missing band locations with sequencing reaction products.[30,31] In contrast to EMSA, footprinting assays permit ligand-DNA interactions to be studied at binding equilibria, but EMSA is typically less difficult to perform since footprinting requires optimizing both ligand-binding and DNA-cleavage reactions to acquire the footprint signal.[28,30,31] Additionally, gel resolution restricts the size of the footprinting DNA probe, affecting the number of potential binding sites screened during a single assay.[30]

In some bacteria, straightforward genetic approaches involving the manipulation of TF genes can be applied to identify the potential genes regulated by TFs.[11-14,17,32] These approaches rely on gene expression level changes, measured as changes in RNA transcript levels, caused by the overexpression, deletion, or inactivation of TF genes to identify the genes that TFs regulate. Genetic engineering techniques are initially applied to manipulate the genes that encode a TF, creating a mutant bacterial strain.[11,12,32] Mutant and wild-type strains are then cultured, and the RNA transcripts expressed by each strain are extracted.[11,12,32] These transcript samples are reverse transcribed to form complementary DNA (cDNA) sequences and uniquely labeled.[11,12,32] Equivalent concentrations of labeled RNA or cDNA from each strain are exposed to a DNA microarray which has been spotted with single-stranded DNA sequences representing

individual genes from the wild-type strain and are permitted to hybridize with these spotted DNA microarray sequences.[11,12,32] Hybrid cDNA-DNA complexes are then detected and compared for each strain to reveal the changes in gene expression caused by manipulating the genes encoding the TF and the potential genes regulated by the genetically manipulated TF (Figure 1).[11,12,32] These approaches provide *in vivo* views into the regulatory activity of TFs and are applicable to studying the genes regulated by TFs during specific developmental stages or under different environmental growth conditions. However, several factors complicate the results obtained using these approaches. Overexpressing TFs can yield indiscriminate gene regulation due to nonspecific DNA-binding. Deleting or inactivating TF genes can lead to the detection of false positives, resulting from the manipulated TF regulating one or more other TFs, and false negatives, resulting from minor changes in RNA transcript levels or from genes coregulated by multiple TFs. Additionally, not all bacteria have the required genetic tools to apply these approaches, and in bacteria with multiple genome copies, deleting or inactivating all genes encoding TFs can be challenging.

Chromatin immunoprecipitation (ChIP) is an alternative *in vivo* technique that can be applied to identify the binding sites of bacterial TFs and the genes they regulate by directly isolating genomic TF-DNA complexes for analysis.[33,34] In a typical ChIP assay, bacteria are cultivated then chemically treated with formaldehyde to reversibly cross-link genomic TF-DNA complexes.[33] After extracting and fragmenting genomic DNA by sonication, TF-specific antibodies are utilized to immunoprecipitate TF-DNA complexes of interest.[33,34] Genomic DNA sequences bound to the TF are freed by reversing TF-DNA complex cross-linkages and are subsequently sequenced along with genomic control

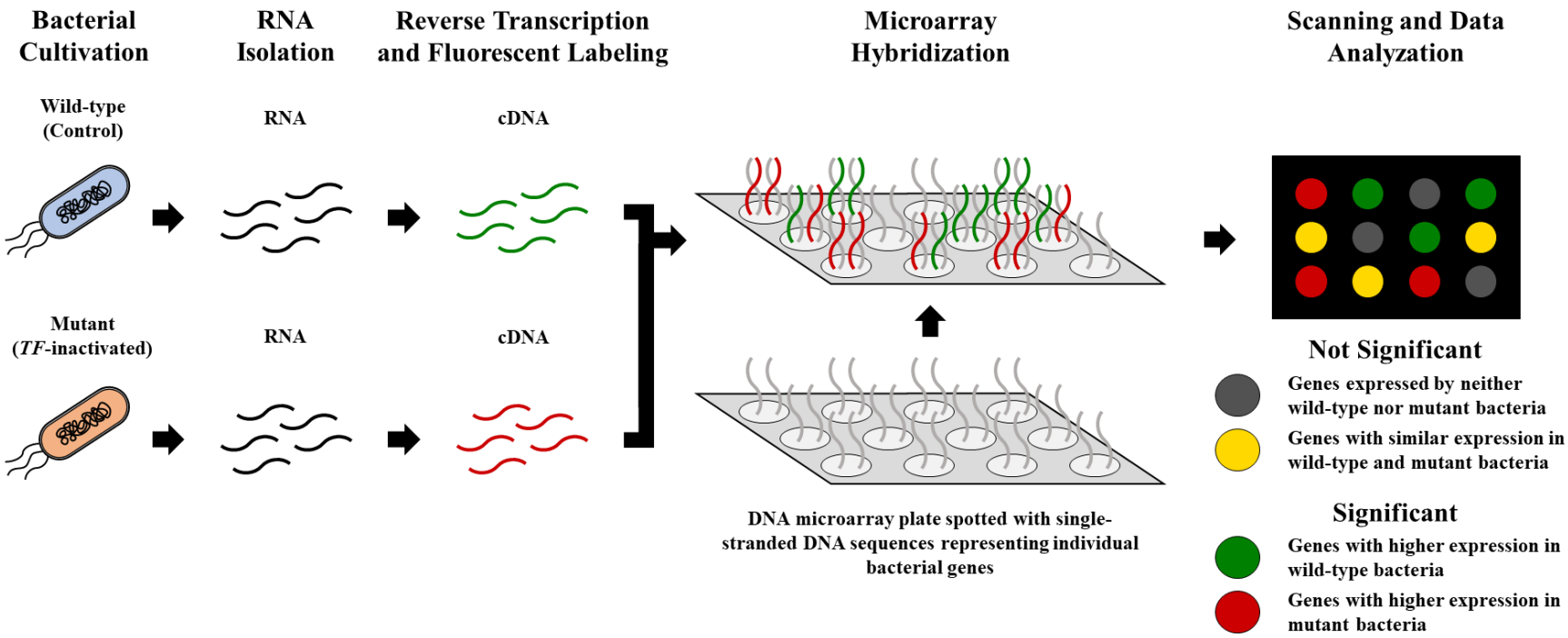


Figure 1. Identifying the genes regulated by a TF through a genetic approach. Steps in a genetic approach involving the inactivation of the genes encoding a TF to identify the genes it regulates have been depicted. These steps include (i) cultivation of bacteria, (ii) isolation of RNA transcripts, (iii) reverse transcription and fluorescent labeling, (iv) microarray hybridization, and (v) scanning and data analysis. Spots considered significant or those with differential gene expression, green and red spots, are used to identify the genes regulated by the TF. Green spots, genes with higher expression in wild-type bacteria, indicate the genes potentially upregulated by the TF, and red spots, genes with higher expression in mutant bacteria, indicate the genes potentially downregulated by the TF.

DNA by microarray (ChIP-chip) or high-throughput sequencing (ChIP-seq).[33,34] Similar to the previous genetic approaches, ChIP assays can capture snapshots of the genes that TFs regulate during distinct developmental stages or under specific environmental conditions, but they can also provide a robust *in vivo* indication of the genes actually regulated by a TF since isolated DNA sequences contain TF-binding sites. While ChIP assays possess limitations, including the required quantity of cells for experiments and the poor resolution of TF-binding sites, ChIP techniques continue to be developed that improve upon the aspects of the standard assay, reducing these experimental limitations.[35,36] However, being an immunoprecipitation technique, ChIP assays require TF-specific antibodies or genetically tagging the genes of TFs with epitopes for which specific antibodies exist which often requires additional time and expense.[33,34]

Protein-binding microarray (PBM) assays offer a high-throughput *in vitro* method for identifying the DNA-binding specificities of TFs and other DNA-binding proteins.[37,38] A PBM assay, as its name suggests, utilizes a DNA microarray spotted with double-stranded DNA sequences. Purified TF, native or epitope-tagged, is exposed to microarray sequences, enabling the formation of TF-DNA complexes.[37,38] These TF-DNA interactions are detected and quantified using TF- or epitope-specific antibodies linked to single fluorophores, revealing the DNA sequences preferentially bound by a TF as a measure of fluorescence intensity.[37,38] Currently, a universal PBM is available for screening TF sequence specificity, encompassing the variation of every predicted 10-base pair (bp) sequence.[38,39] While extensive, some TFs recognize DNA sequences longer than 10 bp, rendering it difficult to comprehensively identify their DNA-binding specificities using PBMs.[39] Additionally, some TFs require cofactors, post-translational modifications,

or interactions with partner proteins to bind DNA specifically which complicates using PBMs to identify their specificities.[39]

Combinatorial *in vitro* selection methods, including systematic evolution of ligands by exponential enrichment (SELEX), selected and amplified binding (SAAB), and cyclic amplification and selection of targets (CASTing), can also be applied to identify the DNA-binding specificities of TFs.[40-43] These techniques typically use ligands such as TFs in repetitive nucleic acid selection, isolation, and amplification cycles to elucidate ligand-nucleic acid specificity.[40-43] In these methods, a ligand is initially incubated with a nucleic acid selection template library containing randomized nucleic acid sequences for probing preferred ligand binding.[40-43] An affinity- or physical-based technique is then applied to separate ligand-nucleic acid complexes formed during the incubation from unbound sequences.[40-43] These complexed nucleic acids are then amplified and utilized as the input nucleic acid for a subsequent cycle.[40-43] Repeating these steps enriches for nucleic acid sequences preferentially bound by the ligand (Figure 2).[40-43] These techniques have yielded useful binding information for an assortment of ligands, and additional adaptations and variations, SELEX serial analysis of gene expression (SELEX-SAGE) and high-throughput SELEX (HT-SELEX) for example, have expanded the utility of these techniques as research tools in biochemistry and molecular biology.[44,45] However, the affinity- or physical-based steps used to separate complexed from free nucleic acid present potential obstacles when studying some ligand-

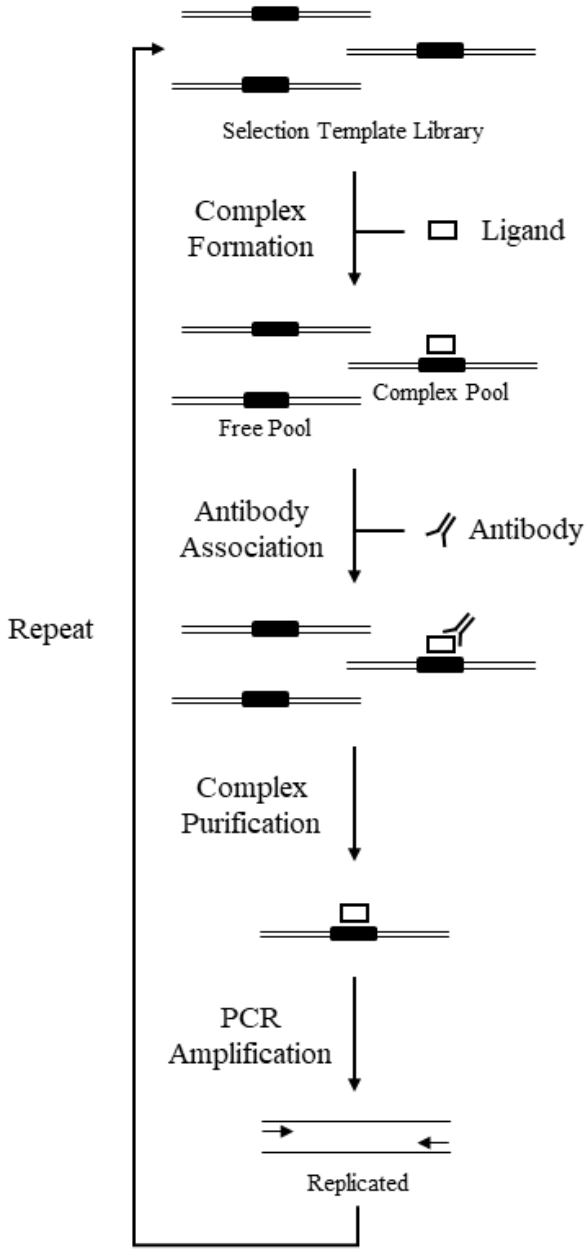


Figure 2. Selecting preferred ligand-binding DNA sequences by cyclic amplification and selection of targets (CASTing). Steps in CASTing have been depicted as a representative illustration of general combinatorial selection technique methodology. These include (i) complex formation, (ii) antibody association, (iii) complex purification, (iv) PCR amplification, and (v) repetition steps to not only select but also enrich for sequences that a ligand preferentially binds.

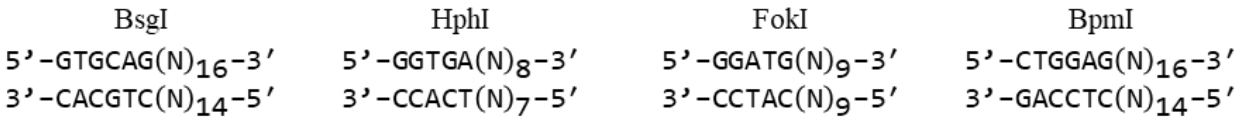
nucleic acid interactions, particularly when minute differences in physical properties exist between complexed and free sequences and when ligands lack the amenability of an affinity-based separation approach.

Chapter 1.3 Identifying Preferred DNA-binding by Restriction Endonuclease Protection, Selection, and Amplification

Pioneered by the Van Dyke Laboratory, restriction endonuclease protection, selection, and amplification (REPSA) is an alternative *in vitro* combinatorial selection method that can be applied to identify the binding specificities of TFs.[46] REPSA selects the DNA sequences a ligand preferentially binds using a methodology like that of other combinatorial selection techniques, but this technique relies on ligand-dependent protection of bound DNA sequences from cleavage by type IIS restriction endonucleases (IISREs) to select preferred DNA-binding sequences and eliminate rejected sequences.[46]

Independent DNA recognition and cleavage domains enable IISREs to recognize DNA with sequence specificity yet non-specifically cleave DNA at a defined distance and, typically, in an asymmetric direction from the recognition site.[47] REPSA exploits these distinctive cleavage properties by utilizing a double-stranded DNA selection template containing a randomized internal DNA core and defined flanking sequences with IISRE recognition sites that are oriented so that cleavage occurs within the randomized core (Figure 3).[46] Thus, unlike other combinatorial selection techniques, REPSA does not require a physical- or affinity-based step to separate selected and rejected template sequences because IISREs are used to eliminate rejected sequences which remain vulnerable to cleavage due to lack of ligand binding.[46]

A



B

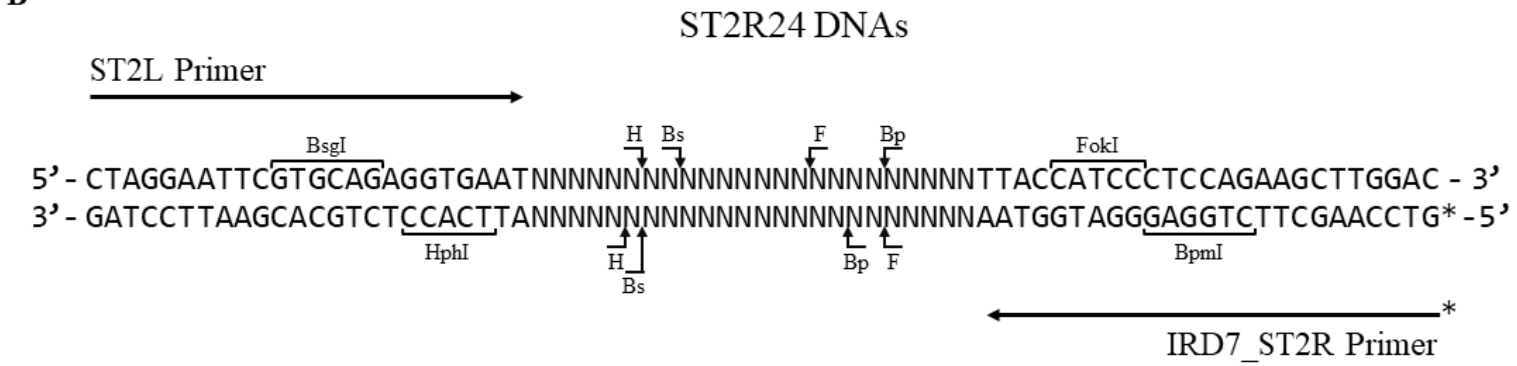


Figure 3. Type IIS restriction endonucleases and selection template ST2R24 structure in REPSA. (**A**) Selected IISREs used in REPSA with respective asymmetric distances to cleavage sites. (**B**) Structure of the selection template ST2R24 used in REPSA containing standard template 2 (ST2) flanks and a randomized 24-bp (R24) core. Brackets indicate respective IISRE recognition sequences within ST2 flanks. Arrows indicate corresponding IISRE cleavage sites within the randomized template core. Horizontal arrows represent primer sequences used to PCR amplify templates. Asterisks indicate the presence of a single IRDye 700 fluorophore label at the 5'-end of primer ST2R and selection template ST2R24. *N* represents a random nucleotide.

A round of REPSA involves several steps: the incubation of a DNA selection template library with a ligand to form ligand-DNA complexes, IISRE cleavage of unbound selection templates, and amplification of ligand-bound templates by PCR for the subsequent round of REPSA (Figure 4).[46] Repetition of REPSA continues until the detected enrichment of ligand-preferred DNA-binding sequences.[46] Highlighting its utility as a research tool, REPSA has been applied to identify the DNA-binding specificities for ligands including small molecules, oligonucleotides, and proteins including TFs.[19-22,48-53] However, selecting low and high affinity sequences as well as intermediate affinity sequences to inclusively identify ligand-binding specificity using REPSA requires the optimization of both ligand-binding and IISRE cleavage reaction parameters.[46] Additionally, REPSA often requires switching from the use of one IISRE to another between rounds to reduce the selection of sequences recognized by IISREs.[46,48] Due to differences in the DNA-binding kinetics and cleavage efficiencies of IISREs, this typically requires optimizing REPSA reaction parameters at least twice if not more.

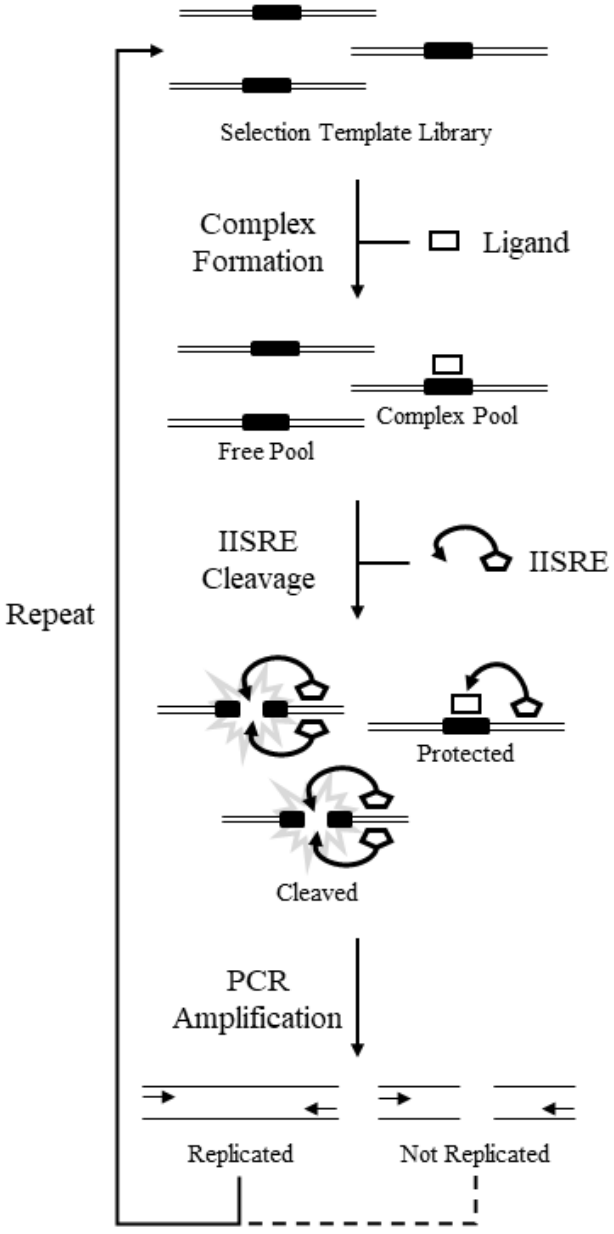


Figure 4. Selecting preferred ligand-binding DNA sequences by restriction endonuclease protection, selection, and amplification (REPSA). REPSA involves (i) complex formation, (ii) IISRE cleavage, (iii) PCR amplification, and (iv) repetition steps to select and enrich for sequences that a ligand preferentially binds.

Chapter 1.4 The CRP/FNR Transcription Factor Superfamily

Cyclic adenosine monophosphate (cAMP) receptor protein (CRP) regulators and fumarate and nitrate reductase regulatory (FNR) proteins together compose the CRP/FNR superfamily of TFs that primarily activate gene expression programs in response to a spectrum of cellular and environmental signals.[54] Members of the superfamily are structurally characterized by a C-terminal domain with a helix-turn-helix DNA-binding motif and large N-terminal domain that typically binds a ligand or cofactor.[54] Currently, most insight into the CRP/FNR superfamily derives from research on the founding and representative members of the superfamily, *E. coli* CRP (CRP_{Ec}) and FNR (FNR_{Ec}).

CRP_{Ec} is a two hundred and nine residue protein with a C-terminal helix-turn-helix DNA-binding domain and an N-terminal domain with a binding site for cAMP.[55,56] This site functions as the primary cAMP-binding site while a site that lies between the N- and C-terminal domains functions as a secondary binding site for cAMP.[55] These binding sites possess respective micromolar and millimolar affinity for cAMP.[57-59] The active DNA-binding form of CRP_{Ec} is a dimer with cAMP bound at the primary binding site of each monomer.[57-59] By comparison, cAMP-free CRP_{Ec} dimer and CRP_{Ec} dimer with cAMP bound at both primary and secondary binding sites have significantly reduced DNA-binding activity.[57-59] CRP_{Ec} acts primarily as a global activator in *E. coli* that binds and directly upregulates the transcription initiation rate of hundreds of genes generally associated with the movement and metabolism of alternative carbon sources.[4,60,61]

Currently, two hundred and seventy-one binding sites are listed for CRP_{Ec} in RegulonDB, a database repository for information on gene expression regulatory networks in *E. coli*.[10] Extracting and submitting these sequences to the online motif analysis program Multiple Em for Motif Elicitation (MEME) yields the DNA-binding consensus motif of CRP_{Ec} (Figure 5A1).[10,62] CRP_{Ec} recognizes 22-bp stretches of DNA and exhibits high affinity for its 22-bp consensus sequence (Figure 5A2).[63]

FNR_{Ec}, is a two hundred and fifty residue protein with a C-terminal helix-turn-helix DNA-binding domain and an N-terminal domain that senses intracellular redox conditions through interactions with iron-sulfur clusters.[64-68] FNR_{Ec} switches between an inactive conformation as a monomer and an active conformation as a homodimer depending on the oxidative state of *E.coli*.[65-68] Aerobic conditions promote the inactive monomer conformation while anaerobic conditions promote the active homodimer conformation by reducing and oxidizing the state of the associated iron-sulfur cluster respectively.[65-68] Active FNR_{Ec} functions as a global regulator that primarily activates the transcription of genes associated with anaerobic energy metabolism and to a lesser extent represses genes associated with aerobic energy metabolism.[4,69-72] Currently, eighty-eight binding sites are listed in RegulonDB for FNR_{Ec}.[10] Extracting and submitting these sequences to MEME for motif analysis as with CRP_{Ec} yields the DNA-binding consensus motif of FNR_{Ec} (Figure 5B1).[10,62] Its minimal 14-bp consensus sequence shares similarities with that of CRP_{Ec} (Figure 5B2).[73,74]

As mentioned previously, TFs can mechanistically activate transcription at promoters by binding promoter DNA and directly engaging RNA polymerase

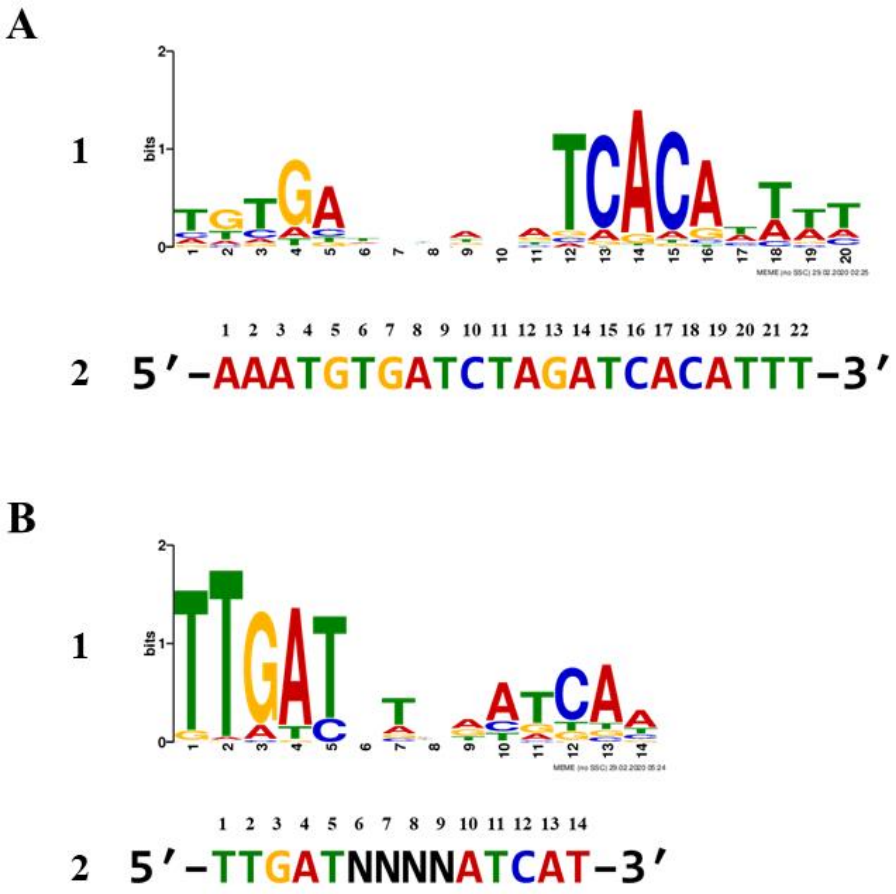


Figure 5. DNA-binding consensus motifs and sequences for the transcription factors CRP_{Ec} and FNR_{Ec}. (**A**) CRP_{Ec} and (**B**) FNR_{Ec} consensus motifs (**1**) and consensus sequences (**2**). Consensus motifs, represented as sequence logos, were obtained from MEME analysis of the currently listed binding sites for CRP_{Ec} and FNR_{Ec} in RegulonDB.[10]

holoenzymes to recruit the transcriptional machinery.[75] The two simplest forms of the mechanism, termed Class I and Class II transcriptional activation, require only one TF bound at a single site for activation.[75] This terminology was initially applied to describe simple CRP_{Ec}-dependent activation mechanisms at promoters and now extends to describe the activation mechanisms of other TFs.[6,75]

A TF that promotes transcription from a binding site positioned upstream of the -35 core promoter element activates through the Class I mechanism, engaging the C-terminal domain of one or both RNA polymerase holoenzyme α-subunits to upregulate transcription initiation.[75,76] Alternatively, a TF that upregulates transcription from a binding site that overlaps the -35 core promoter element activates through the Class II mechanism, primarily contacting the RNA polymerase holoenzyme σ subunit to upregulate transcription initiation.[75,77] Class I and Class II activation efficiency appears dependent to a significant extent on the spacing of binding sites for TFs relative to core promoter elements due to the helical nature of duplex DNA.[78,79] Studies on TF binding site spacing have demonstrated that an optimal TF binding site aligns a TF with core promoter elements along the double helix of DNA and that displacing an optimal TF binding site by even 1 bp is sufficient to reduce the activation efficiency of a TF.[78,79] For Class I activation, a TF bound at a site centrally positioned approximately 51.5, 61.5, 72.5, 82.5, or 92.5 bp upstream, sites separated by integral turns of the helix, can activate transcription with the greatest activation efficiency transpiring at positions 61.5 and 72.5 bp upstream.[78,79] For Class II activation, the optimal TF binding site corresponds to an upstream central position at 41.5 bp.[78]

Chapter 1.5 CRP/FNR Superfamily Transcription Factors in *T. thermophilus* HB8

Thermophilic organisms, thriving in environments of extreme heat, constitute a subgroup of extremophilic organisms, flourishing in environments of at least one extreme. Characteristically, proteins and enzymes expressed by thermophilic organisms, such as those expressed by *T. thermophilus* HB8, possess relatively higher intrinsic thermostability and function optimally at elevated temperatures, and as a result, they are of significant interest to biotechnology and industry as well as structural biology.[80]

Discovered in Japanese thermal spas, gram negative *T. thermophilus* HB8 can grow at temperatures as relatively cold as 47° to as hot as 85°C with an optimal growth temperature range of 65°-72°C.[81,82] As a polyploid bacterium, *T. thermophilus* HB8 maintains multiple copies of a 1.85 mega-bp circular chromosome (TTHA); a 257 kilo-bp megaplasmid, pTT27 (TTHB); and a 9.3 kilo-bp miniplasmid, pTT8 (TTHC).[83] This bacterium is the subject of the Structural and Functional Whole-Cell Project undertaken by the research institute RIKEN in Japan (http://thermus.org/).[84] This structural genomics project pursues the structural and functional characterization of all highly ordered *T. thermophilus* HB8 proteins with the aim of understanding all the biological phenomena of this bacterium (see previous link).[84] Due to its smaller proteome size, the thermostability of its proteins, the increased aptitude of its proteins to crystalize, and the ability to genetically manipulate its genome, *T. thermophilus* HB8 was selected as the subject of the project.[84]

Screening of the *T. thermophilus* HB8 genome has identified proteins TTHA1359, TTHA1437, and TTHB099 as CRP family members and protein TTHA1567 as an NtcA family member, a family of TFs within the CRP/FNR superfamily functioning in global nitrogen assimilation in cyanobacteria.[54] Currently, regulatory and biological functions have been assigned to only the CRP family members.[11-13,16] A review of these roles for these proteins has been provided here.

TTHA1437 was the first *T. thermophilus* HB8 CRP family member studied and was designated *T. thermophilus* HB8 CRP due to its primary sequence and biochemical attributes.[11-13,16] It is a two hundred and sixteen residue protein that forms a homodimer in solution and binds cAMP based on the findings of gel filtration chromatography, light scattering photometry, and protease sensitivity experiments.[11] Inactivating the *TTHA1437* genes in *T. thermophilus* HB8 has no apparent effect on *T. thermophilus* HB8 growth, and therefore, the *TTHA1437* gene is considered dispensable for the train.[11] Six *T. thermophilus* HB8 gene promoters are activated by TTHA1437 in a mechanism dependent on cAMP.[11] These gene promoters were initially identified using a genetic approach involving the inactivation of *TTHA1437* genes in *T. thermophilus* HB8 and the comparison of gene expression levels for wild-type and *TTHA1437*-inactivated *T. thermophilus* HB8 strains.[11] They were then validated using in vitro run-off transcription assays which demonstrated they were upregulated in a TTHA1437-dependent, cAMP-dependent mechanism.[11] Genes associated with these promoters were found to encode clustered regularly-interspaced short-palindromic repeat (CRISPR)-associated (Cas) proteins, proteins with potential roles in DNA replication and transcriptional regulation, and several proteins with no known function.[11] The DNA-binding consensus 5'-

(C/T)NNG(G/T)(G/T)C(A/C)N(A/T)NNTCACAN(G/C)(G/C)-3' was predicted for TTHA1437 from the sequences of these six gene promoters.[11]

Currently, TTHA1359 is the most extensively studied *T. thermophilus* HB8 CRP family member.[11-13,16] It is a two hundred and two residue protein with the alternative name of stationary phase-dependent regulatory protein (SdrP) due to its elevated gene transcript level in the stationary phase of *T. thermophilus* HB8 growth and its function as a transcriptional regulator.[12] Its conformation in solution is a homodimer, and its overall ligand-free dimeric crystal structure closely resembles that of CRP_{Ec} with cAMP complexed at both primary and secondary cAMP-binding sites.[12,55] Inactivating the genes encoding TTHA1359 in *T. thermophilus* HB8 slightly impairs the growth of *T. thermophilus* HB8 and its ability to cope with oxidative stress.[12]

TTHA1359 activates sixteen *T. thermophilus* HB8 promoters whose genes encode proteins that function in redox control, transcriptional regulation, RNA metabolism, DNA and protein repair, and the supply of energy and nutrients.[12,13] Eight of these gene promoters were initially identified using a genetic approach involving the inactivation of *TTHA1359* genes in *T. thermophilus* HB8 and the comparison of gene expression levels for the wild-type and *TTHA1359*-inactivated *T. thermophilus* HB8 strains.[12] They were then validated using in vitro run-off transcription assays which demonstrated that they were upregulated in a TTHA1359-dependent, cAMP-independent mechanism.[12] Because similarities were observed for structures of TTHA1359 and CRP_{Ec}, TTHA1359-binding sites were predicted for these eight promoters based on the consensus DNA-binding properties of CRP_{Ec} and the positions of CRP_{Ec}-binding sites associated with Class I and Class II activation mechanisms.[12,55,56,75,85] From these predicted binding sites, the putative

DNA-binding consensus 5'-(A/T)(A/T)GTGA(N_{5-7})ACAC(T/A)(T/A)-3' was derived for TTHA1359.[12] The other eight gene promoters regulated by TTHA1359 were identified using an alternative genetic approach.[13] In this approach, gene expression levels were collected from experiments with wild-type and gene-inactivated *T. thermophilus* HB8 strains in which cultivation conditions such as duration, culture medium, and chemical treatment were varied.[13] Expression levels were then analyzed for individual genes whose expression correlated with the expression of *THA1359*.[13] Of those whose expression strongly positively correlated with that of *TTHA1359*, eight genes possessed promoter sequences resembling the putative TTHA1359-binding consensus site derived previously.[12,13] These eight promoters were upregulated *in vitro* in a cAMP-independent, TTHA1359-dependent mechanism as with the eight gene promoters identified previously.[13] From the binding sites of all sixteen promoters, the putative TTHA1359-binding consensus 5'-TTGTG(N_9)CNC-3' was derived.[12,13]

TTHB099 is the most recently studied *T. thermophilus* HB8 CRP family member.[11-13,16] It is a one hundred and ninety-five residue protein that forms a homodimer in solution, and its overall ligand-free dimeric crystal structure resembles that of TTHA1359 and that of CRP_{Ec} with cAMP complexed at both primary and secondary cAMP-binding sites.[12,16,55] It is considered to function biologically as a transcriptional activator of carotenoid biosynthesis in *T. thermophilus* HB8 based on the function of the nearly identical protein TT_P0055 encoded in the megaplasmid pTT27 of the closely related *T. thermophilus* HB27 strain.[16]

Chapter 1.6 Objectives of the Present Work

REPSA has been previously applied in the Van Dyke laboratory as part of a reverse-genetic technique to study the potential regulatory and biological functions of TFs.[19-22] Initially, REPSA is used in combination with next-generation sequencing and bioinformatic research tools to discover the DNA-binding specificities of TFs. These specificities are then bioinformatically mapped to genomes and bioinformatically characterized to identify potential regulatory binding sites for TFs. These potential sites are examined for binding by TFs *in vitro*, and validated binding sites are used to predict the genes regulated by TFs. Possible biological roles are then described for these TFs with respect to these validated genes. An end goal in the Van Dyke laboratory is the implementation of this reverse-genetic technique as an effective alternative approach to study putative TFs that lack characterization beyond mere family categorizations. Achieving this end goal requires a series of studies to demonstrate the effectiveness of this technique regarding the accurate determination of potential regulatory binding sites for TFs within subject genomes and the possible biological functions of TFs. Several of these studies have already been performed in which the reverse-genetic technique was implemented to identify the potential regulatory binding sites and possible biological roles of four transcriptional repressors from *T. thermophilus* HB8.[19-22] For each of these TFs, a level of regulatory and biological characterization had already been established, allowing the accuracy of the reverse-genetic approach to be assessed by comparing its findings with those previously reported for these repressors.[14,15,17-22]

These studies not only corroborated what had been established previously but also reported additional findings that have contributed to a greater functional understanding of these TFs.[14,15,17-22] In the present work, this reverse-genetic technique has been implemented to study TTHA1359-binding specificity and the genes that TTHA1359 potentially regulates. This protein is an excellent next-step study subject for evaluating the effectiveness of the reverse-genetic for several reasons. TTHA1359 has been previously studied which permits the findings of the present work to be evaluated.[12,13] As with the four previously studied TFs, TTHA1359 does not require a ligand to bind DNA.[12,13,19-22] However, unlike the TFs studied previously, TTHA1359 functions as a transcriptional activator.[12-15,17-22] Studying TTHA1359 should yield further insight into the functionality of the reverse-genetic technique and guide its future application to study other previously characterized TFs and eventually putative TFs.

CHAPTER 2. MATERIALS AND METHODS

Chapter 2.1 Nucleic Acids and Stock Selection Template Library Preparation

RIKEN kindly supplied plasmid pET11a-*TTHA1359* encoding TTHA1359 protein for the project. Oligonucleotide precursors and primers were synthesized and purified by Integrated DNA Technologies and are listed in Appendix A (Table A1). Stock library selection template ST2R24 DNAs were prepared by assembling, PCR amplifying, and subsequently combining the PCR products of five individual 25 μL PCR reactions containing 1X Standard *Taq* Reaction Buffer (New England Biolabs®, NEB), 50 μM each dNTPs, 600 nM each ST2L and IRD7_ST2R primers, 1 ng single-stranded ST2R24 template DNAs, and 25 U *Taq* DNA Polymerase. Cycling conditions consisted of 5 cycles of 95°C for 30 sec, 60°C for 30 sec, and 68°C for 1 min; 1 cycle of 95°C for 30 sec, 60°C for 30 sec, and 68°C for 1.5 min; and 1 final cycle of 95°C for 30 sec, 60°C for 30 sec, and 68°C for 2 min.

Chapter 2.2 Expression and Purification of TTHA1359 Protein

Expression and purification of TTHA1359 protein followed a protocol similar to that previously established.[19] Transformation of competent *E. coli* strain BL21 (DE3) bacteria with plasmid pET11a-*TTHA1359* was followed by cultivation in 50 mL LB medium containing 100 μg/mL ampicillin at 37°C/250 rpm to an OD_{A600} of 0.6 before induction by supplementation of the growth medium with 1 mM IPTG. Induced bacteria were cultivated 5 hrs further then isolated by centrifugation (4000 x g, 15 min, 4°C). The obtained bacterial pellet was suspended in 1 mL cold 50 mM Tris-Cl (pH 8.1 at 25°C), 200 mM NaCl, 0.2 mM EDTA, 2 mM DTT, and 1 mM PMSF. Suspended bacteria were lysed by incubation at 0°C with 0.3 mg/mL supplemented lysozyme for 15 min followed by sonication (5 cycles, 3 W, 10 sec on/ 10 sec off, 0°C). After centrifugation (21,000 x g, 15 min, 4°C), soluble cell lysate was aspirated and incubated at 70°C for 15 min before centrifugation a second time (21,000 x g, 15 min, 4°C). Aspiration and dilution of the resulting supernatant in an equivalent volume of glycerol, ensuring homogenization by rocking at 4°C for 1 hr, yielded the purified TTHA1359 stock utilized in downstream experiments. The purified stock was stored between experiments at -20°C. The concentration of purified TTHA1359 protein prior to dilution in glycerol was determined by densitometric quantification against a BSA standard curve and estimated at 2.45 mg/mL, corresponding to a purified TTHA1359 stock of 27.5 μM monomer or 55 μM dimer as calculated with the 22,317 Dalton molecular weight of TTHA1359 protein obtained from the UniProtKB database, entry Q5SIL0.[86] Samples obtained during the

expression and purification of TTHA1359 were analyzed alongside PageRuler™ Prestained Protein Ladder (Thermo Fisher Scientific™) by 12% SDS-PAGE analysis (Figure 6).

Chapter 2.3 Identification of a DNA-binding Consensus for TTHA1359

Rounds of REPSA were performed following a formerly outlined protocol.[19] REPSA selections were performed in 10 μL reaction volumes containing 1X CutSmart® Buffer (NEB), 4.515 ng (10 nM) selection template ST2R24 DNAs, and 34 nM monomeric or 17 nM dimeric purified TTHA1359 protein stock. TTHA1359 protein was diluted in cold 50 mM Tris-Cl (pH 8.1 at 25°C), 200 mM NaCl, 0.2 mM EDTA, 2 mM DTT, and 200 μg/mL BSA, and 0.2% wt/vol Tween 20 on ice prior to its addition to reactions. The initial REPSA selection reaction contained stock selection template ST2R24 DNAs while each subsequent round of REPSA contained the selection template ST2R24 DNAs that were selected and amplified during the previous REPSA round. The 4.515 ng stock selection template ST2R24 DNAs (100 fmoles or 6.023×10^{10} molecules) used as the input for round one of REPSA was not considered sufficient to encompass all random 24-bp sequences possible for the selection template ST2R24 ($4^{24}/2 = 1.4 \times 10^{14}$), but it was considered sufficient to identify preferred TTHA1359-binding sequences based on REPSA selections with the four previously studied *T. thermophilus* HB8 TFs which were initiated with less than half the input of ST2R24 DNAs reported in the present work.[19-22] During REPSA cycles, selection reactions were incubated at 55°C for 20 min then 37°C for 10 min with a IISRE, 0.8 U FokI or 2 U BpmI for the initial two or final three rounds respectively, added to the reactions at the 37°C incubation midpoint. Selection reactions were terminated by incubation below 0°C. Alongside selection reactions, two selection control reactions were run which were assembled without

TTHA1359 protein and IISRE (DNA control) or solely without TTHA1359 (IISRE control). REPSA selections were PCR amplified for 6, 9, and 12 cycles as before except the PCR reaction input for each was 1 μL REPSA-selected ST2R24 template DNAs, the concentrations of ST2L, IRD7_ST2R, and NEB *Taq* DNA Polymerase were doubled, and the thermocycling annealing and elongation temperatures were adjusted to 58°C and 68°C respectively.[19] REPSA rounds each ended with the assessment of selection and amplification reactions, performed by native PAGE and IR fluorescence after the protocol described previously.[19] Equivalent 2 μL volumes of reactions and 6X loading solution prepared with 20% wt/vol dextrose, 0.9% wt/vol Orange G dye, 0.1% wt/vol SDS, and 66 mM EDTA then loaded on 1X Tris-borate-EDTA (TBE, pH 8.3 at 25°C), 10% wt/vol polyacrylamide (19:1 acryl:bis) gels and electrophoresed at 5 V/cm for 10 min then 15 V/cm for 30 min. REPSA selection and amplification reactions were visualized by LI-COR instrument IR fluorescence imaging, and quantification of amplification reactions determined by a Qubit double-stranded DNA high-sensitivity assay kit (Thermo Fischer Scientific) and Qubit 3 Fluorometer (Thermo Fisher Scientific) as before.[19] REPSA was terminated after round 5 as an apparent population of TTHA1359-dependent, IISRE cleavage-resistant selection template ST2R24 DNAs was detected during the assessment of round 5 selection results (Figure 7).

Amplicon library DNAs were prepared from REPSA-selected round five amplified ST2R24 DNAs for massively parallel semiconductor sequencing by modification of the method outlined previously.[19] A 2 μL aliquot of REPSA round 5 DNAs was added to a 25 μL PCR reaction containing 1X Standard *Taq* Buffer, 150 μM each dNTPs, 200 nM each trP1_ST2L and A_BC06_ST2R primers, and 0.08 U *Taq*

DNA Polymerase (NEB). Cycling conditions were seven cycles of 95°C for 30 sec, 54°C for 30 sec, and 68°C for 1 min. Deviating from the previous method, 2 μL aliquots of PCR product were added to three individual 25 μL PCR reactions containing 1X Standard *Taq* Buffer, 150 μM each dNTPs, 200 nM each trP1_uni and A_uni primers, and 0.08 U *Taq* DNA Polymerase. Cycling conditions consisted of 6, 9, and 12 cycles of 95°C for 30 sec, 55°C for 30 sec, and 68°C for 1 min. From each reaction, 2 μL aliquots were removed and assessed as described during rounds of REPSA. Results were visualized by UV exposure using a Gel Doc™ EZ (BIO-RAD) instrument after staining in 20 mL water containing 50 μg ethidium bromide. The resulting library of amplicon DNAs yielded by the 9-cycle PCR reaction, possessing the highest visible concentration and double-stranded to single-stranded ratio of amplicon DNAs, was selected and purified using a DNA Clean & Concentrator kit (Zymo Research). Massively parallel sequencing of the prepared and purified library of amplicon DNAs proceeded as described.[19]

Raw sequencing data was initially refined by the Sequencing1.java program as described while the first one thousand unique sequences of the resulting refined sequence set were manually identified due to the apparent removal from the web of the program applied previously to remove duplicate sequences, DuplicatesFinder v 1.1.[19] The unique sequence set was uploaded to MEME v 5.1.1 for non-palindromic and palindromic consensus discovery.[62] Motif discovery parameters were set at default except the maximum width of discovered motifs was restricted to 24 bp, corresponding to the width of the randomized core of selection template ST2R24 DNAs, and the palindromic filter was applied for the palindromic motif search. Resulting motifs with statistical

significance values, E-values, < 0.01 were analyzed. Data from the discovered, selected motifs such as statistical significances and number of sites contributing to the motifs were recorded, and graphical representations of the motifs were exported (Figure 9A, 9B). From the obtained consensus motifs, a TTHA1359-DNA binding consensus was derived (Figure 9C).

Chapter 2.4 TTHA1359-DNA Binding Assays

An EMSA was performed to validate the authenticity of the emergent REPSA-selected TTHA1359-binding ST2R24 DNAs which generally followed the protocol that has been described previously.[19] EMSA binding reactions were performed in 5 μL reaction volumes containing 0.5X Tris-acetate-EDTA (TAE, pH 8.4 at 25°C) supplemented to 1 mM EDTA, 0.2% wt/vol Tween 20, 2% wt/vol glucose, 0.5 μg BSA, and 10 nM IRD7-labeled REPSA round one and five PCR-amplified selections incubated with 0, 1, 10, 100, or 1000 nM TTHA1359 protein dimer. Binding reactions were assembled and equilibrated at 55°C for 5 min prior to the addition of TTHA1359 protein. Binding reactions were then incubated at 55°C for 30 min before they were directly loaded onto a 0.5X TAE, 10% wt/vol polyacrylamide (19:1 acryl:bis) gel. Loaded reactions were electrophoresed in 0.5X TAE buffer at 10 V/cm for 30 min. Results were visualized by IR fluorescence as described previously (Figure 8).[19]

Binding kinetics were determined between TTHA1359 and biotinylated DNAs by biolayer interferometry (BLI) essentially as described previously but with additional control reactions.[19] TTHA1359-binding control, consensus, mutant consensus, and CRP_{Ec} consensus DNAs were PCR biotinylated and amplified for BLI assays in 50 μL reaction volumes containing 1X Standard *Taq* Reaction Buffer, 50 μM each dNTPs, 350 nM ST2L primer, 300 nM Bio_ST2R primer, 2 ng DNA template precursor, and 2 U *Taq* DNA Polymerase (Table A1). Cycling conditions consisted of 30 cycles of 95°C for 30 sec, 58°C for 30 sec, and 68°C for 1 min. BLI assays were performed with an

Octet^{QK} (FortéBio) instrument at 37°C in 25 mM Tris-Cl (pH 8.1 at 25°C), 100 mM NaCl, 1 mM EDTA, and 0.05% Tween 20 buffer in 96-well microplates. For each assay, 2 nM biotinylated DNA sequences were immobilized during a 20 min incubation on Streptavidin Dip and Read Biosensors (FortéBio) which were hydrated in buffer 5-10 minutes at 25°C prior to each assay. Loaded biosensors were rinsed in fresh buffer for 5 min to establish immobilized DNA baselines, then immersed for 5 min in buffer containing different dimeric TTHA1359 protein concentrations (5.7, 17, 51, 153, and 461 nM), and finally washed for 10 min in fresh buffer. Biosensor data for each assay, recorded in wavelength shift (nm), was obtained every 1.6 sec. Besides the PCR amplified biotin-labeled control DNA sequence tested for nonspecific or weak binding, ST2_1359_ctrl, each assay included controls, assembled without TTHA1359 protein dimer and biotinylated DNA sequences (biosensor control) or TTHA1359 protein dimer (biosensor load control). Data from the control was subtracted point-by-point from the experimental biosensor data prior to data analysis. Binding kinetics were determined for each assay by non-linear regression analysis of association and dissociation biosensor data collected using protein concentrations of 5.7, 17, 51, and 153 nM TTHA1359 dimer using GraphPad Prism software as before (Table 1).[19] Representative raw associative and dissociative step BLI data and calculated best-fit line traces for TTHA1359 and tested DNA sequences have been provided in the text (Figure 10) while the remaining plots have been provided in Appendix B (Figure B1).

Chapter 2.5 Identification of Potential Regulatory TTHA1359-binding Sites in the *T. thermophilus* HB8 Genome

Since the top non-palindromic motif discovered by MEME was a truncation of the consensus suggested by the motif, the motif was not directly imported into the program Find Identified Motif Occurrences (FIMO) as previously described.[19,87] Instead, position-dependent letter-probability matrix data from positions 6-16 of the top non-palindromic motif discovered by MEME was initially utilized to derive an extended 22-bp position-dependent letter-probability matrix motif. A text file suitable for upload and utilization by FIMO was then written in MEME minimal motif format, containing the targeted version number of MEME, the extended motif alphabet and strand information, and the extended position-dependent letter-probability matrix motif (http://meme-suite.org/doc/meme-format.html?man_type=web) (data not shown). The file was uploaded to FIMO v 5.1.1 and utilized to scan the GenBank *Thermus thermophilus* HB8 universal identifier 13202 version 210 database for potential binding sites with statistically significant p-values less than 0.0001. Potential binding sites selected for further bioinformatic analysis were limited to the identified potential binding sites with p-values less than 6×10^{-5} (Table 2).

Genomic sequence contexts, encompassing -210/+310 bp of these potential binding sites, were obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database.[88] As explained in the next paragraph, some of these genomic sequences contexts have been provided in the text (Figure 11) while others have been provided in Appendix C (Figure C1). Potential binding sites located between or near

bidirectionally oriented genes were examined individually, extracting the genomic sequence context in both orientations. Genes positioned near potential TTHA1359-binding sites were noted (Table 2), and transcription unit information for these genes, whether they were transcribed as single units or as one of several genes in an operon, was ascertained through the BioCyc database collection (Table 2).[89] Potential core promoter elements within the obtained sequence contexts were predicted by sequence context submission to the online bacterial promoter prediction program BPROM by Softberry (Figure 11, Figure C1).[90]

Initially, FIMO-identified potential binding sites with no identifiable core promoter elements were removed from further consideration. Potential binding sites were also excluded by limiting the distance between predicted transcription and known translation start sites of local genes to 100 bp. This was an arbitrary restriction as no comprehensive transcription start site data is currently available for *T. thermophilus* HB8. Of the remaining potential binding sites, those positioned downstream of predicted core elements were excluded from further consideration. Finally, potential binding sites with central positions greater than 100 bp upstream of predicted transcription start sites were excluded. Potential TTHA1359-binding sites which were excluded from further analysis have been provided in Appendix C (Figure C1). Potential TTHA1359-binding sites which were investigated for regulatory potential have been provided in the text (Figure 11).

Distances were calculated from transcription start sites to the center of remaining binding sites overlapping or upstream of predicted -35 core promoter elements (Table 3). Calculated distances were used to postulate binding site roles in transcriptional activation and corresponding activation mechanisms, either Class I or Class II, based on CRP_{Ec}-

binding site mutational studies with regard to transcriptional activation.[78,79] Roles in transcriptional repression were also postulated based on potential binding sites for TTHA1359 overlapping predicted core promoter elements.

CHAPTER 3. RESULTS

Chapter 3.1 Expression and Purification of TTHA1359

Recombinant expression of TTHA1359 protein from *T. thermophilus* HB8 by induction of cultivated *E. coli* strain BL21 (DE3) transformed with plasmid pET11a-*TTHA1359* yielded significant TTHA1359 protein production as evidenced by analyzed culture samples harvested pre- and post-induction (Figure 6, Lanes 2 and 3). Expressed TTHA1359 protein was initially recovered in the cell lysate together with *E. coli* proteins after sonoporation of harvested cells treated with lysozyme (Figure 6, Lane 4). TTHA1359 protein was then purified from most contaminant *E. coli* proteins by subjection of the recovered cell lysate to high-temperature incubation (Figure 6, Lane 5). The identities of the few contaminant *E. coli* proteins that remained soluble during heat treatment have yet to be determined (Figure 6, Lane 5). Further purification to remove these remaining proteins was deemed unnecessary because previous applications with *T. thermophilus* HB8 proteins, expressed and purified similarly to TTHA1359, appeared unaffected by remaining contaminant proteins.[19-22] This likely resulted from the dilutions of purified *T. thermophilus* HB8 proteins utilized in previous experiments effectively rendering remaining contaminant proteins as non-existent.[19-22] The identity of the expressed and purified protein as TTHA1359 monomer was substantiated by the ~ 22 kilodalton molecular weight of the presented protein band, consistent with the molecular weight of TTHA1359 monomer as determined its UniProtKB database entry, Q5SIL0 (Figure 6, Lanes 3-5).[86]

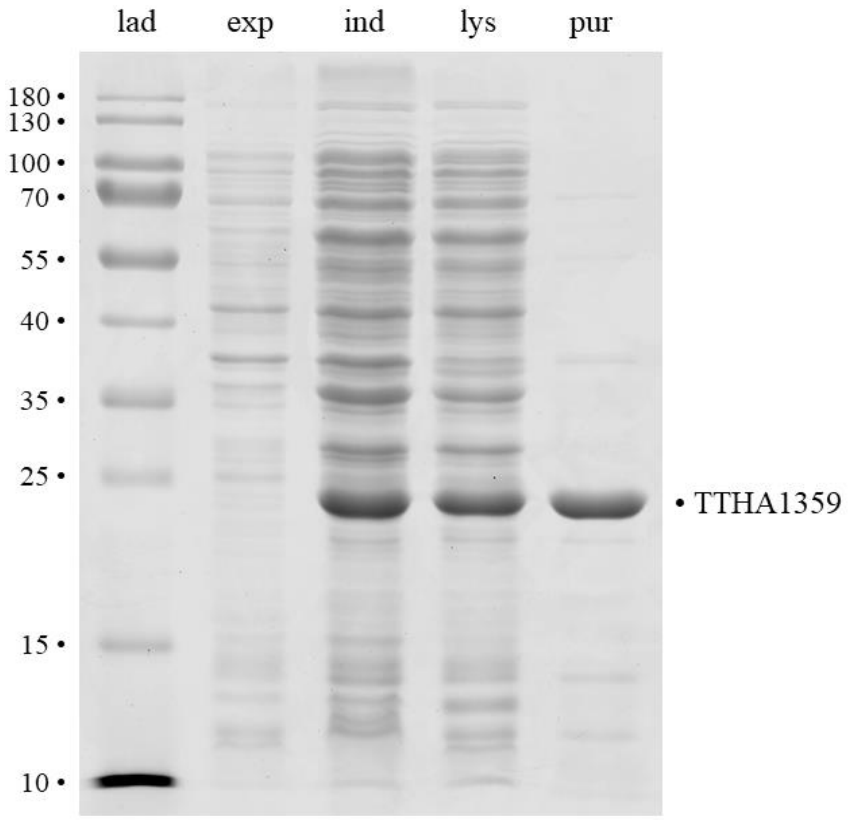


Figure 6. Expression and purification of TTHA1359 protein. Image depicts samples obtained during TTHA1359 protein expression and purification as analyzed by 12% SDS-PAGE. Lanes were loaded with samples of protein ladder (lad), cultivated transformant *E. coli* in the exponential growth phase (exp), induced cultivated transformant *E. coli* post-incubation (ind), recovered cell lysate following harvested bacterial cell membrane disruption (lys), and recovered soluble cell content after high temperature purification (pur). Molecular weights of the protein ladder are indicated to the left of the figure. The protein band corresponding to TTHA1359 monomer is indicated to the right of the figure. Additional *E. coli* proteins that remained soluble during heat treatment (pur) have yet to be determined.

Chapter 3.2 Identification of a DNA-binding Consensus for TTHA1359

REPSA selections of TTHA1359-binding selection template ST2R24 DNAs were performed for a total five rounds before the detected emergence of an apparent population of TTHA1359-dependent, IISRE cleavage-resistant ST2R24 template DNAs (Figure 7). Rounds of REPSA initially relied on the IISRE FokI to interrogate TTHA1359-binding specificity, but selection of a minor population of ST2R24 DNAs with cleavage resistance against FokI, observed in the cleavage control and selection reactions during the assessment of REPSA round two selection results through image intensification (data not shown), required the switch to BpmI, an alternative IISRE, for the third round of REPSA. Unlike rounds one through four of REPSA, approximately half the input ST2R24 template DNAs in the selection reaction of REPSA round five were not digested by IISRE in the presence of 17 nM TTHA1359 dimer (Figure 7, Round 5, Lane 3), suggestive of an enriched population of ST2R24 template DNAs to which TTHA1359 preferentially binds. Titration of REPSA round one and five amplified selections of ST2R24 template DNAs with TTHA1359 in an EMSA validated the enriched presence of preferred TTHA1359-binding DNAs (Figure 8). No mobility shift was observed for REPSA round one amplified DNAs across the range of TTHA1359 dimer titrated (Figure 8, Lanes 1-5). However, an increasing mobility shift was observed for round five amplified DNAs beginning as low as equivalent concentrations of TTHA1359 dimer and DNAs (Figure 8, Lanes 6-10). Consequently, the cleavage

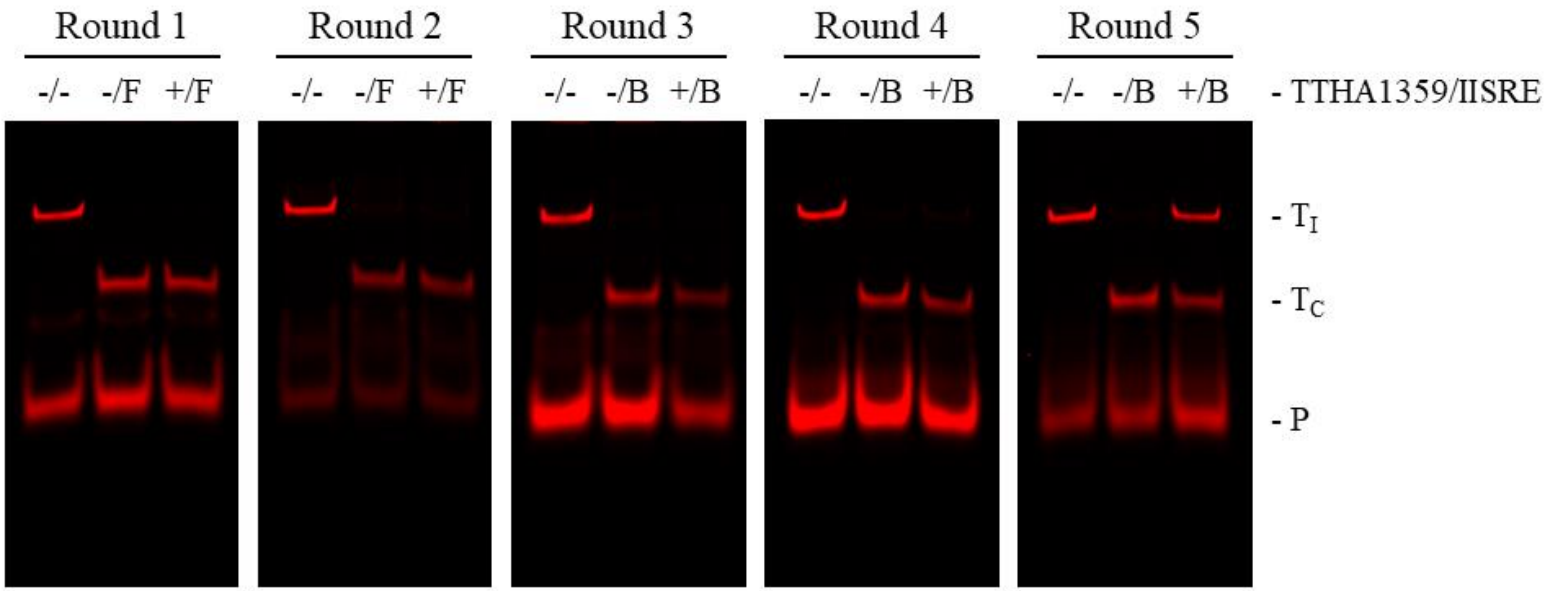


Figure 7. REPSA selection of a population of TTHA1359-binding 73-bp template ST2R24 DNAs. Each IR fluorescence image depicts selection reaction results from individual rounds of REPSA. Reactions in lanes 1-3 in each figure correspond to DNA controls prepared without TTHA1359 or IISRE (-/-), cleavage controls including only IISRE (-/F or -/B representing FokI and BpmI IISRE inclusion respectively), and the selection reaction prepared with 17 nM dimeric TTHA1359 protein and IISRE (+/F or +F). Band designations: T_I, intact ST2R24 template DNAs; T_C, cleaved ST2R24 template DNAs; P, remnant 5'-labeled fluorescent ST2R primer.

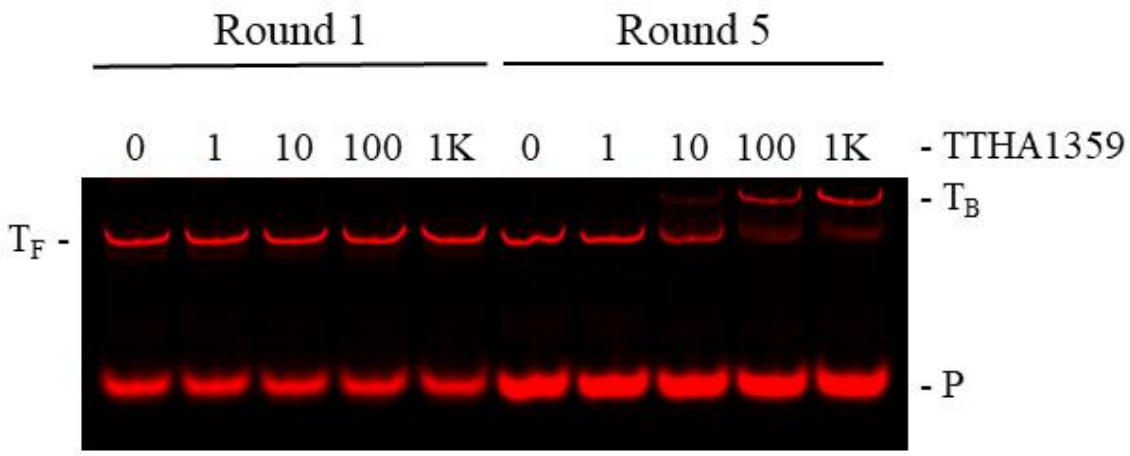


Figure 8. Titration of amplified REPSA-selected 73-bp template ST2R24 DNAs from rounds one and five with TTHA1359. Reactions in lanes 1-5 and 6-10 contained 10 nM amplified selections from REPSA rounds 1 and 5 respectively. Reactions were incubated with either 0, 1, 10, 100, or 1,000 (1K) nM dimeric TTHA1359 protein. Band designations: T_F, Free ST2R24 template DNAs; T_B, bound ST2R24 template DNAs; P, remnant 5'-labeled fluorescent ST2R primer.

resistance of the intact population of ST2R24 template DNAs at REPSA round five was attributed to TTHA1359-binding as a result of the ST2R24 template DNAs possessing core sequences preferentially bound by TTHA1359.

Massively parallel sequencing of the synthesized amplicon library DNAs yielded 9,516,545 total base reads with an incorrect base calling quality score, ≥ *Q*20, of 8,631,131 and 158,313 reads of 60 bp average length. Sequencing runs with the amplicon library DNAs prepared from the REPSA-selected DNAs of the four previously studied *T. thermophilus* HB8 transcriptional repressors yielded 2-5 million total base reads, 1.5-3.5 million ≥ *Q*20 quality scores, and 40,000-100,000 reads of average length 48-51 bp.[19-22] The improved sequencing results reported here were considered to be the result of the method in which the amplicon library DNAs were prepared for sequencing which differed than that used in the previous studies.[19-22] The obtained 158,313 reads were reduced to 61,754 sequences of appropriate quality for further investigation by the applied Sequencing1.java refinement program. Manual examination of the first one thousand and five refined sequences yielded the first one thousand unique sequences as five sequences were identified as duplicates and removed. Submission of the identified unique sequence set to MEME yielded the top three non-palindromic and palindromic motifs each within the unique sequence set as determined by the program with the top two motifs discovered during each search possessing E-values, measurements of statistical significance, < 0.01 (Figure 9A, 9B). The top non-palindromic and palindromic motifs, corresponding to a 19-bp motif within a lengthier 24-bp pseudo- motif (Figure 9A1) and a 24-bp motif (Figure 9B1), were discovered respectively within seven hundred

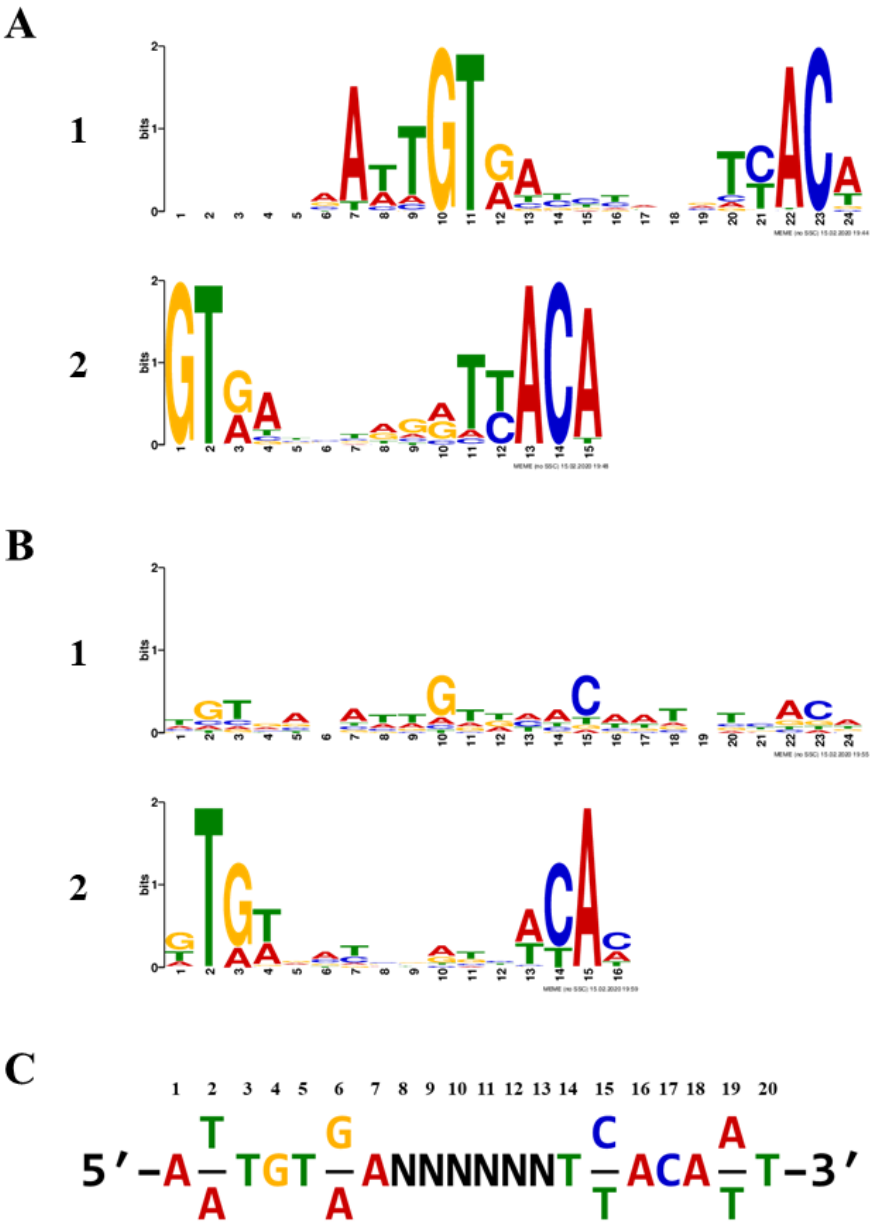


Figure 9. Discovered and derived TTHA1359-DNA binding motifs. The top two (**A**) non-palindromic and (**B**) palindromic motifs discovered by MEME analysis with statistically significant E-values < 0.01, represented as position-dependent letter-probability matrices. (**C**) The minimal TTHA1359-DNA binding consensus sequence derived from MEME discovered motifs.

and eighty-nine and eight hundred and ten of the unique one thousand sequence set input with remarkably high statistical significances as indicated by associated E-values of $1.1x10^{-2,996}$ and $4.5x10^{-1,153}$. Top secondary non-palindromic and palindromic motifs, corresponding to a 15-bp (Figure 9A2) and a 16-bp (Figure 9B2) motif, were respectively identified in one hundred sixty-seven and one hundred thirty-nine of the unique sequence set input with much lower statistical significances with E-values of $4.2x10^{-264}$ and $1.5x10^{-51}$. A minimal 20-bp inverted-repeat TTHA1359-binding consensus was derived primarily from the non-palindromic top motif (Figure 9C). Positions 1-7 and 14-20 of this THA1359-binding consensus (Figure 9C) were found to share significant similarities with positions 2-8 and 15-21 of the CRP_{Ec} consensus which correspond to twelve of the fourteen positions that CRP_{Ec} demonstrates strong sequence preferences (Figure 5A2).[63,91-93] Since TTHA1359 is a homodimer and its derived DNA-binding consensus is a palindrome, each TTHA1359 dimer subunit is considered to bind half of the consensus.[12]

Chapter 3.3 Biophysical Characterization of TTHA1359-DNA Binding

BLI assays were performed to characterize the DNA-binding affinity of TTHA1359 for its identified consensus and CRP_{Ec} consensus sequences and to investigate its consensus-binding specificity. In these BLI experiments, real-time binding kinetics for select TTHA1359-DNA interactions were measured by recording wavelength shifts in biosensor-reflected, white-light interference patterns brought about by the interactions of TTHA1359 with biosensor-immobilized DNA.[94] Raw wavelength shift data collected during these experiments at concentrations of 5.7, 17, 51, and 153 nM TTHA1359 dimer was used to calculate the DNA-binding kinetics between TTHA1359 and the DNA sequences tested (Table 1). Raw data recorded at concentrations of 461 nM TTHA1359 dimer was excluded from these calculations because no significant or detectable binding was observed between TTHA1359 and the TTHA1359-binding control sequence except at 461 nM TTHA1359 dimer (data not shown). Raw associative and dissociative step BLI data and calculated best-fit line trace plots for TTHA1359 and its consensus, the CRP_{Ec} consensus, and control sequences have been provided here (Figure 10). The remaining plots for TTHA1359 and other DNA sequences have been provided in Appendix C (Figure C1).

High-affinity binding was observed between TTHA1359 and its consensus sequence, having a quick on rate, slow off rate, and a calculated apparent K_D of ~ 3.4 nM with an R^2 value > 0.95 (Figure 10A, Table 1). Individual single-point mutants were then introduced to positions 1-7 of the TTHA1359-binding consensus sequence (Figure 9C) to

Table 1. TTHA1359-DNA binding parameters

Name	Sequence	k_{on}($M^{-1}\cdot s^{-1}$)	k_{off}(s^{-1})	K_D(M)	R^2
con_wt	ATTGTGACACACATCACAAT	457555	0.001577	$3.447x10^{-9}$	0.9572
con_p1	cTTGTGACACACATCACAAT	——— Ambiguous ———			
con_p2	AgTGTGACACACATCACAAT	317173	0.01166	$36.78x10^{-9}$	0.9820
con_p3	ATgGTGACACACATCACAAT	215769	0.01867	$86.53x10^{-9}$	0.9793
con_p4	ATTtTGACACACATCACAAT	——— Ambiguous ———			
con_p5	ATTGgGACACACATCACAAT	——— Ambiguous ———			
con_p6	ATTGTtACACACATCACAAT	288341	0.02153	$74.67x10^{-9}$	0.9766
con_p7	ATTGTGgCACACATCACAAT	267033	0.006236	$23.35x10^{-9}$	0.9843
con _s5	ATTGTGAcacacTCACAAT	——— Ambiguous ———			
con _s7	ATTGTGAcacacacTCACAAT	——— Ambiguous ———			
CRP_Ec	AAATGTGATCTAGATCACATTT	726387	0.0007341	$1.011x10^{-9}$	0.8979
ctrl	ATACGAAAAACACACAC	——— Ambiguous ———			

Lowercase nucleotides indicate single-point or spacing mutations from the wild-type consensus. K_D, k_{on}, and k_{off} indicate the equilibrium dissociation, association rate, and dissociation rate constants respectively. *Ambiguous* indicates the absence of a significant observable or detectable binding event by BLI.

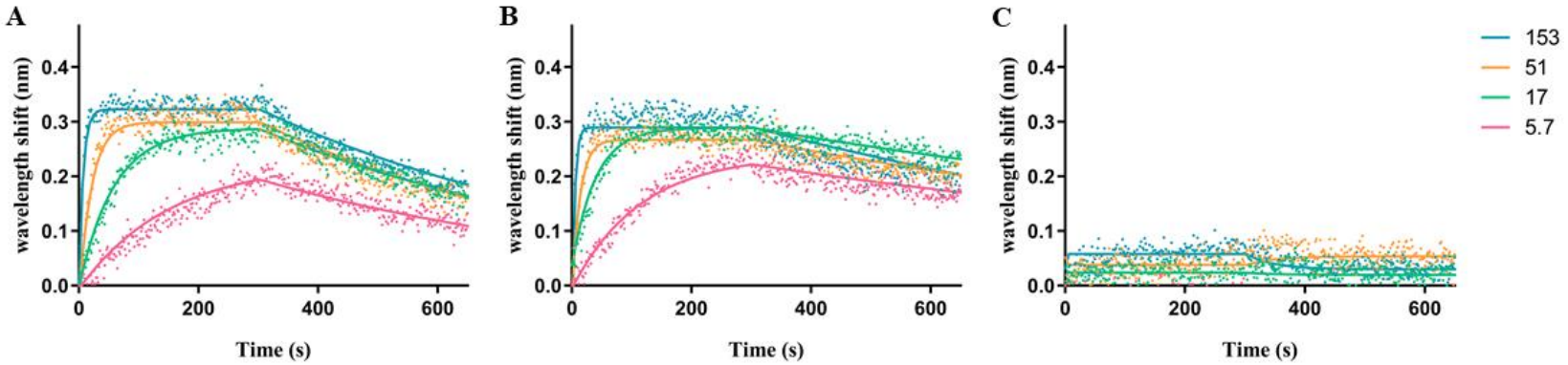


Figure 10. Representative BLI association and dissociation step raw data plots of TTHA1359-DNA binding. Graphs depict raw association and dissociation step data measured during BLI experiments with TTHA1359 and (**A**) ST2_1359_con_wt, (**B**) ST2_CRP_Ec, and (**C**) ST2_1359_ctrl biotinylated DNAs. These DNAs correspond respectively to wild-type TTHA1359-binding consensus, CRP_{Ec}-binding consensus, TTHA1359-binding control sequences. Dots represent raw data points for these TTHA359-binding assays, and solid lines depict calculated best-fit lines for the raw data points. Line colors pink, green, orange, and blue indicate concentrations of TTHA1359 protein corresponding to 5.7, 17, 51, and 153 nM TTHA1359 dimer.

characterize consensus sequence binding specificity, having mutated individual base positions to either the least represented or the non-complementary partner of the most highly represented base as identified by the top non-palindromic motif (Figure 9A1). Single-point mutant binding kinetics varied dramatically with single-point mutants introduced at positions 1, 4, and 5 yielding ambiguous binding under the applied experimental conditions (Table 1), suggesting a strong sequence preference at these positions. Results of mutations introduced at positions 2, 3, 6, and 7 were less detrimental as binding kinetics were actually attainable with observed binding affinity for the wild-type TTHA1359-binding consensus ranging from ~ 6.5- to ~ 25-fold higher than binding affinities for single point mutants by calculated apparent K_D value comparison (Table 1). Additional mutant consensus sequences, possessing either a truncated 5-bp or extended 7-bp spacer sequence instead of the observed 6-bp spacer at positions 8-13 within the consensus sequence (Figure 9C), were tested and yielded ambiguous binding results, suggestive of a strict consensus sequence recognition spacing requirement under the reaction conditions specified (Table 1). Finally, as TTHA1359 and CRP_{Ec} were found to recognize similar sequences, TTHA1359-binding to the CRP_{Ec} consensus was investigated (Figure 5A2, Figure 9C, Figure 10B).[63] TTHA1359 was found to have greater than 3-fold higher affinity for the CRP_{Ec} consensus over its identified consensus, having an apparent K_D of ~ 1 nM with an R^2 value of ~ 0.9 (Figure 10B, Table 1). This increased affinity was attributed to a slight difference in the recognition sequences, corresponding to positions 2 and 19 within the TTHA1359 and 3 and 20 within the CRP_{Ec} consensus sequences, as well as differences in the 6-bp composition of the interior spacer regions of the two sequences (Figure 5A2, Figure 9C, Table 1).[63]

Chapter 3.4 Identification of Potential Regulatory TTHA1359-binding Sites in the *T. thermophilus* HB8 Genome

Having investigated TTHA1359-binding to its REPSA-selected consensus to ascertain the validity of its selected consensus, an extended TTHA1359 consensus motif which was derived from the top non-palindromic MEME motif discovered (Figure 9A1, positions 6-16) was submitted to FIMO to identify potential regulatory binding sites. The FIMO search yielded eighty-eight potential THA1359-binding sites within the *T. thermophilus* HB8 genome with p-values less than 0.0001 (data not shown). Selection of identified potential binding sites with p-values less than 6×10^{-5} narrowed the list of potential binding sites to the first fifty-nine for further analysis (Table 2). Cursory inspection of the selected potential sites using the KEGG database extended the list of sites to investigate to seventy-four, as fifteen of the fifty-nine selected sites were positioned near or within bidirectionally-oriented genes which were analyzed individually.

Initially, eleven of the seventy-four FIMO-identified potential binding sites for TTHA1359 were excluded from further analysis based on the absence of a predicted promoter, corresponding to potential sites upstream of genes *TTHA0843*, *TTHB243*, *TTHA0583*, and *TTHA1326* (FIMO result nos. 5, 16, 36a, and 44 in Figure C1 and Table 2) and downstream of genes *TTHA1836*, *TTHA1236*, *TTHA1158*, *TTHA0486*, *TTHA1301*, *TTHA1352*, and *TTHA0030* (FIMO result nos. 18, 25, 26, 27, 32, 34a, and 53 in Figure C1 and Table 2). In the absence of available comprehensive transcription start

site data for *T. thermophilus* HB8, distances between predicted transcription and known translation start sites of genes located proximally to potential TTHA1359-binding sites was restricted to 100 bp. Based on this restriction, potential binding sites downstream of genes *TTHA0974*, *TTHC001*, *TTHA0910*, *TTHA0644*, *TTHA0030*, *TTHA0943*, *TTHA1848*, *TTHA0507*, *TTHB177*, *TTHA1632*, *TTHA0910*, and *TTHA0036* (FIMO result nos. 1, 3, 12, 13, 20, 23, 33, 35b, 38, 47, 48, and 54 in Figure C1 and Table 2) and upstream of genes *TTHA0784*, *TTHA1730*, *TTHA1538*, *TTHB113*, *TTHA0055*, *TTHB088*, *TTHB177*, *TTHA1854*, *TTHA0303*, *TTHA0544*, *TTHA1398*, and *TTHA0205* (FIMO result no. 7, 21b, 29b 30, 31a, 37a, 38, 39, 42a, 45, 56a, 57a in Figure C1 and Table 2) were not investigated further. The limited genomic sequence context prevented the visual determination of predicted transcription start site distances to nearby translation start sites for potential binding sites positioned downstream of genes *TTHA0910*, *TTHA0030*, and *TTHA0036* (FIMO result nos. 12, 20, and 54 in Figure C1 and Table 2). Returning to the KEGG database revealed there were no translation start sites within the distance restraint, and thus, these binding sites were excluded from further analysis. To the resulting thirty-seven, only those potential TTHA1359-binding sites centrally positioned upstream or overlapping predicted core promoter elements were considered further which lead to the elimination of ten additional potential binding sites. These ten sites included potential sites identified upstream of genes TTHA0942, TTHA0936, TTHA1223, TTHA0084, TTHA0643, TTHA0643, TTHA0532 (FIMO result nos. 19, 22, 24, 41, 46, 49, and 52 in Figure C1 and Table 2) and downstream of genes TTHA1729, TTHA0344, TTHA1073, (FIMO result nos. 21a, 28, and 50 in Figure C1 and Table 2). For the remaining sites, an upstream limit of 100 bp from the central position of potential TTHA1359-binding sites

to predicted start sites of transcription was imposed based on the previously determined spacing requirements for CRP_{Ec}-dependent transcriptional activation at promoters in *E. coli*.[78,79] This final restriction excluded potential binding sites upstream of genes *TTHA1537*, *TTHA0056*, *TTHA1353*, *TTHA0304*, *TTHA1135*, and *TTHA1133* (FIMO result nos. 29a, 31b, 34b, 42b, 43, and 59 in Figure C1 and Table 2) from further analysis. These refinement steps excluded fifty-one potential TTHA1359-binding sites (Figure C1), yielding a total twenty-three FIMO-identified TTHA1359-binding sites which were analyzed for potential regulatory functions (Figure 11).

Of the remaining potential TTHA1359-binding sites with central positions overlapping or upstream predicted -35 core promoter elements, five sites were optimally positioned for TTHA1359 to bind and upregulate transcription based on calculated distances to predicted transcription start sites and information on binding site positions associated with simple CRP_{Ec}-dependent activation mechanisms (Table 3).[78,79] These sites corresponded to potential sites centered 41.5 bp from the predicted transcription start sites of *TTHA0425*, *TTHA0584*, and *TTHA0770*; 61.5 bp from the predicted transcription start site of *TTHA0506*; and 82.5 bp from the predicted transcription start site of *TTHA0987* (FIMO result nos. 4, 36b, 58, 35a, and 6 in Figure 11 and Table 3). Potential TTHA1359-binding sites upstream of *TTHA0425*, *TTHA0584*, and *TTHA0770* gene promoters were found to correspond to optimal TF-binding positions associated with transcriptional upregulation by a Class II mechanism (Table 3).[75,78,79] Potential binding sites upstream of *TTHA0506* and *TTHA0987* gene promoters were observed to correspond to optimal TF-binding positions associated with transcriptional upregulation by a Class I mechanism (Table 3).[75,78,79] An additional potential regulatory binding site

for TTHA1359 was noted, centered at 93.5 bp upstream of *TTHB089* (Table 3). However, since the transcriptional activity CRP_{Ec} when centrally bound 92.5 bp upstream of a transcription start site has been demonstrated to be minor, substantial upregulation of *TTHB089* transcription by TTHA1359 at this potential binding site was considered unlikely to occur (Table 3).[79] Potential binding sites were also noted upstream of *TTHA0953*, *TTHA0446*, *TTHA0447*, *TTHA0533*, *TTHA0606* twice, and *TTHA0206* genes, but were considered unlikely to be regulatory due to their orientation on the DNA double-helix with respect to predicted core promoter elements (Table 3).[78,79] Interestingly, a considerable number of remaining potentially binding sites overlapped predicted core promoter elements (Table 3). These binding sites included those upstream of genes *TTHA0080*, *TTHA0081*, *TTHA0954*, *TTHA0534*, *TTHA0109*, *TTHA0607*, *TTHA0607* again, *TTHA1026*, *TTHA1399*, and *TTHA1133* (FIMO result nos. 9a, 9b, 10, 14b, 17, 40b, 51b, 55, 56b, and 59 in Figure 11 and Table 3). Although a potential role as transcriptional repressor was not suggested by previous work with TTHA1359, the positioning of these sites with respect to predicted transcription start sites suggests a potential role for TTHA1359 as a repressor of these genes by a simple steric hindrance mechanism.[12,13]

Table 2. FIMO-identified potential TTHA1359-binding sites in the *T. thermophilus* HB8 genome

FIMO Result	Loci	Start	End	p-Value	*Q*-value	Sequence	Position	Local Gene(s)	Operon
1	Chromosome	917,761	917,782	1.01×10^{-8}	0.0125	AAATGTGAACATATTCACTTTC	Intragenic	*TTHA0974* *TTHA0973*	1/1 1/6
2	Chromosome	898,965	898,986	1.48×10^{-8}	0.0125	AATCGTGAAGTTTATCACATAT	Intergenic	*TTHA0952* *TTHA0953*	1/1 1/1
3	Miniplasmid	1,503	1,524	1.78×10^{-8}	0.0125	GAAAGTGAGATAACTCACATAT	Intergenic	*TTHC001*	1/1
4	Chromosome	402,440	402,461	1.01×10^{-7}	0.0532	TAAAGTGCTTTATTTCACAAAA	Intergenic	*TTHA0424* *TTHA0425*	1/1 1/1
5	Chromosome	809,120	809,141	2.64×10^{-7}	0.104	AATTGTGCTGGGCCACACAAAT	Intragenic	*TTHA0843* *TTHA0842*	1/3 2/3
6	Chromosome	931,166	931,187	2.95×10^{-7}	0.104	ATTAGTGAAACTTTTCACGATT	Intragenic	*TTHA0986* *TTHA0987*	1/1 1/1
7	Chromosome	752,613	752,634	3.71×10^{-7}	0.105	GAAGGTAACTTCAAACACTTTC	Intragenic	*TTHA0785* *TTHA0784*	6/6 1/1
8a	Chromosome	418,352	418,373	3.97×10^{-7}	0.105	AATTGTCAACGGGATTACGTAT	Intergenic	*TTHA0446*	1/1
8b	Chromosome	418,352	418,373	3.97×10^{-7}	0.105	ATACGTAATCCCGTTGACAATT	Intergenic	*TTHA0447*	1/5
9a	Chromosome	81,405	81,426	5.35×10^{-7}	0.113	CCATGTGTTTTAGTTTACTTTA	Intragenic	*TTHA0080*	1/2
9b	Chromosome	81,405	81,426	5.35×10^{-7}	0.113	TAAAGTAAACTAAAACACATGG	Intragenic	*TTHA0081*	1/3
10	Chromosome	899,588	899,609	5.82×10^{-7}	0.113	AATCGTGAATAAAATCACTAGG	Intergenic	*TTHA0953* *TTHA0954*	1/1 1/2
11	Chromosome	932,531	932,552	5.88×10^{-7}	0.113	TCTTGTACTTTTATTCACGATT	Intergenic	*TTHA0987*	1/1

12	Chromosome	871,755	871,776	1.69×10^{-6}	0.298	ACTTGTCAGCAAAATTACGATG	Intergenic	*TTHA0910*	3/3
13	Chromosome	613,187	613,208	2.35×10^{-6}	0.383	CAATGTCCTTTTAAGCTCAATT	Intragenic	*TTHA0644* *TTHA0643*	2/3 3/3
14a	Chromosome	496,704	496,725	3.11×10^{-6}	0.471	GAAAGAGAATGTTAGCACATTT	Intergenic	*TTHA0533*	1/2
14b	Chromosome	496,704	496,725	3.11×10^{-6}	0.471	AAATGTGCTAACATTCTCTTTC	Intergenic	*TTHA0534*	1/2
15	Chromosome	1,507,780	1,507,801	5.01×10^{-6}	0.706	ATTTGGACGCAATTTCACTTTT	Intragenic	*TTHA1584*	3/3
16	Megaplasmid	249,380	249,401	5.72×10^{-6}	0.757	AAATGTGGCTTAGGGCGCAAAA	Intragenic	*TTHB244* *TTHB243*	5/5 1/7
17	Chromosome	109,940	109,961	6.09×10^{-6}	0.758	TTTCGTGCCCTAGTTCACTAAG	Intragenic	*TTHA0110* *TTHA0109*	3/3 1/1
18	Chromosome	1,718,950	1,718, 971	9.54×10^{-6}	1	CAACGTGTTTGAGCTCACCAAG	Intragenic	*TTHA1836* *TTHA1837*	1/1 1/1
19	Chromosome	890,892	890,913	1.05×10^{-5}	1	TATTGTGCTCAATTTCATACTG	Intragenic	*TTHA0943* *TTHA0942*	1/7 2/7
20	Chromosome	32,701	32,722	1.33×10^{-5}	1	GGCTGTGTACGAAATTACAACT	Intragenic	*TTHA0030* *TTHA0031*	1/2 2/2
21a	Chromosome	1,622,533	1,622,554	1.35×10^{-5}	1	AAAGGTGGGCAAAAAGACGTTT	Intragenic	*TTHA1729*	1/1
21b	Chromosome	1,622,533	1,622,554	1.35×10^{-5}	1	AAACGTCTTTTTGCCCACCTTT	Intragenic	*TTHA1730*	1/3
22	Chromosome	885,131	885,152	1.37×10^{-5}	1	AACAGTAACCCATCATGCACTT	Intergenic	*TTHA0937* *TTHA0936*	7/7 1/1
23	Chromosome	891,322	891,343	1.42×10^{-5}	1	GAATGTTGGCAGATGTACTGTT	Intragenic	*TTHA0943* *TTHA0942*	1/7 2/7

24	Chromosome	1,161,988	1,162,009	$1.49x10^{-5}$	1	GACTGTAACGGCGTGCACCATC	Intragenic	*TTHA1222* *TTHA1223*	1/1 1/1
25	Chromosome	1,174,082	1,174,103	$1.53x10^{-5}$	1	CAAGGTGGTCCTGGACACCTTT	Intragenic	*TTHA1236* *TTHA1235*	8/9 9/9
26	Chromosome	1,104,298	1,104,319	$1.67x10^{-5}$	1	CATCGTGGGCATCGTCACCATG	Intragenic	*TTHA1158* *TTHA1159*	1/3 2/3
27	Chromosome	456,182	456,203	$1.69x10^{-5}$	1	CTGCGTGAGCTACCTCACCATT	Intragenic	*TTHA0486*	2/2
28	Chromosome	326,380	326,401	$1.72x10^{-5}$	1	CAAGGTGAGCCTAGGCACCATC	Intragenic	*TTHA0344* *TTHA0343*	1/2 2/2
29a	Chromosome	1,464,992	1,465,013	$1.74x10^{-5}$	1	TAATGTGCGCTATTTCCCAACT	Intergenic	*TTHA1537*	1/1
29b	Chromosome	1,464,992	1,465,013	$1.74x10^{-5}$	1	AGTTGGGAAATAGCGCACATTA	Intergenic	*TTHA1538*	1/1
30	Megaplasmid	101,237	101,258	$1.96x10^{-5}$	1	GACTTTGCCCTAAGGTACAAAC	Intergenic	*TTHB114* *TTHB113*	2/2 1/2
31a	Chromosome	54,523	54,544	$2.08x10^{-5}$	1	ATCAGTGACATACGTGCCATTT	Intergenic	*TTHA0055*	1/2
31b	Chromosome	54,523	54,544	$2.08x10^{-5}$	1	AAATGGCACGTATGTCACTGAT	Intergenic	*TTHA0056*	1/4
32	Chromosome	1,241,335	1,241,356	$2.22x10^{-5}$	1	CACTGGGACGTCATCTACATTG	Intragenic	*TTHA1301* *TTHA1300*	8/10 9/10
33	Chromosome	1,733,129	1,733,150	$2.32x10^{-5}$	1	GATTGTGAGCGGAGCCACGAGC	Intragenic	*TTHA1848* *TTHA1849*	1/1 1/1
34a	Chromosome	1,292,459	1,292,480	$2.35x10^{-5}$	1	CATGGTGAGCATGATGACCTTG	Intragenic	*TTHA1352*	1/3
34b	Chromosome	1,292,459	1,292,480	$2.35x10^{-5}$	1	CAAGGTCATCATGCTCACCATG	Intragenic	*TTHA1353*	1/2

35a	Chromosome	472,200	472,221	$2.44x10^{-5}$	1	CCCTGTTTTTCAAGATACAAAA	Intragenic	*TTHA0506*	1/1
35b	Chromosome	472,200	472,221	$2.44x10^{-5}$	1	TTTTGTATCTTGAAAAACAGGG	Intragenic	*TTHA0507*	1/1
36a	Chromosome	550,760	550,781	$2.62x10^{-5}$	1	AAAAGGGGGCCAGGGCACTTTT	Intergenic	*TTHA0583*	1/1
36b	Chromosome	550,760	550,781	$2.62x10^{-5}$	1	AAAAGTGCCCTGGCCCCCTTTT	Intergenic	*TTHA0584*	1/1
37a	Megaplasmid	79,603	79,624	$2.84x10^{-5}$	1	CAGGGTGAGGGGGAGCACATTC	Intergenic	*TTHB088*	1/1
37b	Megaplasmid	79,603	79,624	$2.84x10^{-5}$	1	GAATGTGCTCCCCCTCACCCTG	Intergenic	*TTHB089*	1/3
38	Megaplasmid	171,682	171,703	$2.86x10^{-5}$	1	AACTTTATCCCTTTATACAGAT	Intragenic	*TTHB177* *TTHB176*	1/3 2/3
39	Chromosome	1,507,336	1,507,357	$3.3x10^{-5}$	1	AAGAGTATGTCAAAACCCTTTA	Intragenic	*TTHA1583* *TTHA1584*	2/3 3/3
40a	Chromosome	574,514	574,535	$3.32x10^{-5}$	1	TAAACTAAGAAAGTTTACGAAA	Intergenic	*TTHA0606*	1/2
40b	Chromosome	574,514	574,535	$3.32x10^{-5}$	1	TTTCGTAAACTTTCTTAGTTTA	Intergenic	*TTHA0607*	1/1
41	Chromosome	83,733	83,754	$3.42x10^{-5}$	1	CTTCGTGAAGAAAGGCACGAAA	Intergenic	*TTHA0083* *TTHA0084*	3/3 1/17
42a	Chromosome	290,819	290,840	$3.58x10^{-5}$	1	TATAGTGAGGTATGGTCCGGTT	Intragenic	*TTHA0303*	1/1
42b	Chromosome	290,819	290,840	$3.58x10^{-5}$	1	AACCGGACCATACCTCACTATA	Intragenic	*TTHA0304*	1/3
43	Chromosome	1,075,934	1,075955	$3.68x10^{-5}$	1	GATCGTCTTCAAGATCACGAGC	Intragenic	*TTHA1134* *TTHA1135*	2/3 3/3

44	Chromosome	1,265,643	1,265,664	3.76×10^{-5}	1	CTTTGTGGGCCAGCTCCCTATA	Intergenic	*TTHA1325* *TTHA1326*	1/1 1/2
45	Chromosome	508,691	508,712	3.79×10^{-5}	1	CACCGTGGACTACGACACCATC	Intragenic	*TTHA0543* *TTHA0544*	1/1 1/2
46	Chromosome	612,470	612,491	3.94×10^{-5}	1	TTTGGTAATCTTTGTCTTAATT	Intragenic	*TTHA0644* *TTHA0643*	2/3 3/3
47	Chromosome	1,546,227	1,546,248	4×10^{-5}	1	CATTGACTTTATGCTCACCATC	Intragenic	*TTHA1632*	2/2
48	Chromosome	871,788	871,809	4.26×10^{-5}	1	CTTTGGGCTTTGAAAAACAAAG	Intergenic	*TTHA0910*	3/3
49	Chromosome	612,463	612,484	4.63×10^{-5}	1	ATCTTTGTCTTAATTTACATCG	Intragenic	*TTHA0644* *TTHA0643*	2/3 3/3
50	Chromosome	1,022,206	1,022,227	4.69×10^{-5}	1	CATTGAGACCTACTCCACCAAG	Intragenic	*TTHA1074* *TTHA1073* *TTHA1072*	1/1 1/1 1/2
51a	Chromosome	574,529	574,550	4.74×10^{-5}	1	GTTCGTAAACCAAAATAAACTA	Intergenic	*TTHA0606*	1/2
51b	Chromosome	574,529	574,550	4.74×10^{-5}	1	TAGTTTATTTTGGTTTACGAAC	Intergenic	*TTHA0607*	1/1
52	Chromosome	495,754	495,775	4.89×10^{-5}	1	GTTTGACATCATAGTTACGAAC	Intragenic	*TTHA0533* *TTHA0532*	1/2 2/2
53	Chromosome	32,413	32,434	4.94×10^{-5}	1	CAAGGTGAAGGACGTCACCAAG	Intragenic	*TTHA0030*	1/2
54	Chromosome	37,707	37,728	4.94×10^{-5}	1	CAAGGTGAAGGACGTCACCAAG	Intragenic	*TTHA0036*	1/3
55	Chromosome	976,524	976,545	4.95×10^{-5}	1	GTTGGTGAACTTCAGTAAAATC	Intragenic	*TTHA1025* *TTHA1026*	2/2 1/1
56a	Chromosome	1,332,409	1,332,430	5.15×10^{-5}	1	CCATGTGCCCAAGTATACTTAG	Intragenic	*TTHA1398*	1/1

56b	Chromosome	1,332,409	1,332,430	5.15×10^{-5}	1	CTAAGTATACTTGGGCACATGG	Intragenic	*TTHA1399*	1/1
57a	Chromosome	203,014	203,035	5.48×10^{-5}	1	TTGTGTAAAAATGCGCACCATG	Intragenic	*TTHA0205*	1/3
57b	Chromosome	203,014	203,035	5.48×10^{-5}	1	CATGGTGCGCATTTTTACACAA	Intragenic	*TTHA0206*	1/3
58	Chromosome	738,907	738,928	5.71×10^{-5}	1	CATTGTGCCCCGGGGTGCCTTT	Intergenic	*TTHA0770*	1/2
59	Chromosome	1,075,413	1,075,434	5.75×10^{-5}	1	AAATGTCCCAAAAAGTCCCTTT	Intergenic	*TTHA1133*	1/3

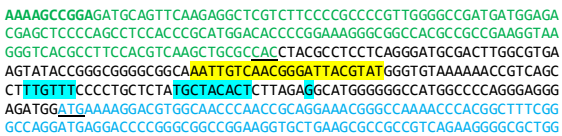
P-value indicates the FIMO calculated probability of random sequence of input motif length fitting the identified potential binding site sequence equally well or better than the input motif. *Q-value* indicates the rate of false discovery for potential binding sites considered important. *Operon* indicates the position of the gene within its transcribed unit.

2. Chromosome (898755...899296); TTHA0952 and TTHA0953

4. Chromosome (402230...402771); TTHA0424 and TTHA0425

6. Chromosome (930956...931497); TTHA0986 and TTHA0987

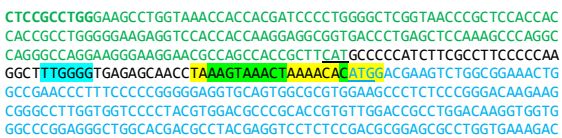
8a. Chromosome, compl. (418042...418583); TTHA0447 and TTHA0446



8b. Chromosome (418142...418683); TTHA0446 and TTHA0447

9a. Chromosome, compl. (81095...81636); TTHA0081 and TTHA0080

9b. Chromosome (81195...81736); TTHA0080 and TTHA0081



10. Chromosome (899378...899919); TTHA0953 and TTHA0954

14a. Chromosome, compl. (496394...496935); TTHA0534 and TTHA0533

14b. Chromosome (496494....497035); TTHA0533 and TTHA0534

17. Chromosome, compl. 109630...110171); TTHA0110 and TTHA0109

35a. Chromosome, compl. (471890...472431); TTHA0507 and TTHA0506

36b. Chromosome (550550...551091); TTHA0583 and TTHA0584

37b. Megaplasmid (79393...79934); TTHB088 and TTHB089

40a. Chromosome, compl. (574204...574745); TTHA0607 and TTHA0606

TGGGCAGAAGCCGCTTTTAACGTTGAGAAGACGGACCAGCTTGAGCCGCCGTCCGAAAAAGTA
CTCCTTGACCCGGATGGCCGCCTCCGCCAAAGCAGGCAGGGCCTCGGGTGGGGCCTCGAGGAC
AAAAAGGGCTTCCTGTAGGCTCAAGGGCCTTTCGGCCAGGCGAGGAGGGTCCAGATCGGGCCA
CGCCATGTTCGTAAACCAAAATAAACTAAGAAAGTTTACGAAAGCAAGGGGCGCGTTGCCCAA
ATCACGGCCGTGCCCTAAACTTAGACCTCGTGAAGAGGGTGATGTTCCACGCCAAGATCCACC
GGGCCACCGTGACCCAGGCCGACCTCCACTACGTGGGCTCGGTGACGGTGGATCAGGACCTCC
TGGACGCCGCCGGGATCCTTCCTTTTGAGCAGGTGGACATCTACGACATCACCAACGGGGCCC
GGCTCACCACCTACGCCCTTCCCGGGGAACGGGGGTCCGGCGTCATCGGGATCAACGGGGCCG
CGGCCCACCTGGTGAAGCCCGGGGACCT**GGTTATCCTC**

40b. Chromosome (574304...574845): TTHA0606 and TTHA0607

CGGGCCCCGTTGGTGATGTCGTAGATGTCCACCTGCTCAAAAGGAAGGATCCCGGCGGCGTCC
AGGAGGTCCTGATCCACCGTCACCGAGCCCACGTAGTGGAGGTCGGCCTGGGTCACGGTGGCC
CGGTGGATCTTGGCGTGGAACATCACCCTCTTCACGAGGTCTAAGTTTAGGGCACGGCCGTGA
TTTGGGCAACGCGCCCCTTGCTTTCGTAAACTTTCTTAGTTTATTTTGGTTTACGAACATGGC
GTGGCCCGATCTGGACCCTCCTCGCCTGGCCGAAAGGCCCTTGAGCCTACAGGAAGCCCTTTT
TGTCCTCGAGGCCCCACCCGAGGCCCTGCCTGCTTTGGCGGAGGCGGCCATCCGGGTCAAGGA
GTACTTTTTCGGACGGCGGCTCAAGCTGGTCCGTCTTCTCAACGTTAAAAGCGGCTTCTGCCC
AGAGGACTGCGCCTACTGCGCCCAGTCCGCTAGATCCCAAGCCGCCATCGCCCGGTACCCCCT
GCTCTCCTTGGAGGAGATCTTGGAGAGG**GCCGAGGAGG**

51a. Chromosome, compl. (574219...574760); TTHA0607 and TTHA0606

GTAGGCGCAGTCCTCTGGGCAGAAGCCGCTTTTAACGTTGAGAAGACGGACCAGCTTGAGCCG
CCGTCCGAAAAAGTACTCCTTGACCCGGATGGCCGCCTCCGCCAAAGCAGGCAGGGCCTCGGG
TGGGGCCTCGAGGACAAAAAGGGCTTCCTGTAGGCTCAAGGGCCTTTCGGCCAGGCGAGGAGG
GTCCAGATCGGGCCACGCCATGTTCGTAAACCAAAATAAACTAAGAAAGTTTACGAAAGCAAG
GGGCGCGTTGCCCAAATCACGGCCGTGCCCTAAACTTAGACCTCGTGAAGAGGGTGATGTTCC
ACGCCAAGATCCACCGGGCCACCGTGACCCAGGCCGACCTCCACTACGTGGGCTCGGTGACGG
TGGATCAGGACCTCCTGGACGCCGCCGGGATCCTTCCTTTTGAGCAGGTGGACATCTACGACA
TCACCAACGGGGCCCGGCTCACCACCTACGCCCTTCCCGGGGAACGGGGGTCCGGCGTCATCG
GGATCAACGGGGCCGCGGCCCACCTGGT**GAAGCCCGGG**

51b. Chromosome, (574319..574860), TTHA0606 and TTHA0607

ATGTCGTAGATGTCCACCTGCTCAAAAGGAAGGATCCCGGCGGCGTCCAGGAGGTCCTGATCC
ACCGTCACCGAGCCCACGTAGTGGAGGTCGGCCTGGGTCACGGTGGCCCGGTGGATCTTGGCG
TGGAACATCACCCTCTTCACGAGGTCTAAGTTTAGGGCACGGCCGTGATTTGGGCAACGCGCC
CCTTGCTTTCGTAAACTTTCTTAGTTTATTTTGGTTTACGAACATGGCGTGGCCCGATCTGGA
CCCTCCTCGCCTGGCCGAAAGGCCCTTGAGCCTACAGGAAGCCCTTTTTGTCCTCGAGGCCCC
ACCCGAGGCCCTGCCTGCTTTGGCGGAGGCGGCCATCCGGGTCAAGGAGTACTTTTTCGGACG
GCGGCTCAAGCTGGTCCGTCTTCTCAACGTTAAAAGCGGCTTCTGCCCAGAGGACTGCGCCTA
CTGCGCCCAGTCCGCTAGATCCCAAGCCGCCATCGCCCGGTACCCCCTGCTCTCCTTGGAGGA
GATCTTGGAGAGGGCCGAGGAGGCCCAA**CGCCTTTCTG**

55. Chromosome, (976314...976855); TTHA1025 and TTHA1026

AGGCCCTCTGGGGCCTCCTGGTCGAGGCCATCGCCGACCGCCTCGTGGACGGTTGGGACCGGTA
CGGGGCCCCCAGCGCGGCCCGGGATCCCGAGGGGGGATTCGTGGCCAGCGCCGAGGGGCCAGGG
GAGCCCATCGTGGTTCGGGCCTCCACGAAGCGCGAGGCCTACCGGGAGGCTCGCCGGGAGTGGG
TGAAGCGGCTCCTAAACGGTTGGTGAACTTCAGTAAAATCCCTCTGCTCCGTACCCCTCTTTGG
GGTTCCTTGATACTCTCGAGTCGTGCAAGGAGAAGTCCTGACCGTGGCGGAGGTGGCCAAGCTC
CTCCGGGTTAGCCGCAAGACGGTGGAAAAGCTCATCTACGCCAAGAAGCTCCACGCGGTGAAGG
TGGGCCGGGTTTGGCGCATCCCGAGGGCCGCCGTGGAGGCGTTCTTGGAGGGGAAGGAGTACCG
TGAACCTTAGGAAGCTCTGGTCCTGGCTTCGCTTCGCCAAAGCCCTCTCCCGGGGCCGGGTGGG
CAAGCGGTGGGCCCTACGCC**GGGCCCGGTG**

56b. Chromosome, (1332199...1332740); TTHA1398 and TTHA1399

CTGAGCCTCCCCTCCTCCAGCCTCCCCGAGACCTTCACCTTCAAGGCGTCGTAGAGGCCCGTG
ACCCGGACGGGGTACTCCGCCTCCCCCGTGTACTTCCAGTAGAAGACGGGGAGGAAGGGGGGG
TAGCCCACGGCCAGGCCGTACCGCTCGGCGTACCAGTAGAGGCCCAGGGCCGCCAGGGCCAGA
AAGAGCAGGGTCCGCATTGGCCTAAGTATACTTGGGCACATGGCGGGCACCGGCCACACCCCG
GAAGAGGCCCTGGCCCTCCTCAAGCGGGGGGCCGAGGAGATCGTCCCCGAGGAAGAGCTTCTC
GCCAAGCTCAAGGAGGGGCGGCCCCTCACGGTCAAGCTCGGAGCCGACCCCACGAGGCCCGAC
CTGCACCTGGGCCACGCGGTGGTCCTGAGGAAGATGCGCCAGTTCCAGGAGCTCGGCCACAAG
GTGGTCCTCATCATCGGCGACTTCACCGGGATGATCGGGGACCCTTCGGGCCGTTCCAAGACC
CGGCCCCCCCTCACCCTGGAGGAGACCC**GGGAGAACGC**

57b. Chromosome, (202804 .. 203345); TTHA0205 and TTHA0206

GGGCAGCCCCGCCCTGCCCTCCAAGGGCCGCGTCCAGCCGGGGGCGAGGAGGAGGGCGGCCAAG
GCCAGGCGCCCCGCGCTCATCACCAGGCTGAAGGCGAGGCTTTGGTCCCCGGAGGCCCGAAGGG
CCAGGCGCACCAGGATGCTCCCGAAGCTTATGGCCAGGATGCCCACGAGAAGGACGGCGGCGAT
CTTCCACCCCGGGGGGGCGCATGGTGCGCATTTTTACACAAGGGAAGCCTTCCGGGAAGCGTCAA
AGTAGTGTAGCCCAAGGGAAAACCCGGGGACGGACGCCCCACCCCGCGCTGGTATACTTCCCCC
GAAGGAGGCGCCTATGGTGACCGTCGCGGTTCCCAAGGAAAGGGCCCCAGGGGAAAGGAGGGTG
GCCCTGGTGCCCGAGGTGGTGGCCCGCCTGGTGAAAGGCGGGGCCCGGGTGCGGGTGGAGCGGG
GCGCGGGGGAGGGCGCCTACCACCCCGACGAGGCCTACCAGGAGGCCGGGGCCGAGGTGGTGGA
GCGAGGGGAGCTCCTAAAGG**GCGCCCACCT**

58. Chromosome, compl. (738597...739138); TTHA0771 and TTHA0770

GACCACGAGTTTCTACTCCGCCCGGGACAGTACCTACTAGAGCTAGGGACCCCAGGGGTAGTC
TCCCTACGCCTACAAGCCTGGGAGACCGCTCCAGGAAACCAAGAAATCCTCGTACCCCTAGAG
ATGACCGAGAGCCTCTAAGCGTTCTCCCGTAGGGACCAATCCCCTCCGCTGGCGGCCGCGGTC
CCCGGCCAAAGCCTGAACGCCCATTGTGCCCCGGGGTGCCTTTGTTTCCTGAAAGGGGGGTAT
CATAGGGGCAGATGAAGGACTTTCTGCGCCTAGAACTCCCCGTGCTCCCTCTCCGGAACACCG
TGGTCCTCCCCCACACCACCACCGGGGTGGACGTGGGGCGCCTCAAGAGCAAGCGGGCGGTGG
AGGAGGCCCTCTCCGCCGACCGGCTCCTCTTCTTGGTGACCCAGAAGGACCCGGAGGTGGACG
ACCCCGCCCCGGAGGACCTTTACGCCGTGGGCACCCTGGCCGTGGTCAAGCAGGCCATGCGCC
TGCCGGACGGCACCCTCCAGGTGATGGT**GGAGGCGAGA**

59. Chromosome, (1075203...1075744); TTHA1132, TTHA1133, and TTHA1134

ATGGTCCGGGCCGTGGTGGACCCCGAGGGGAAGTTCCCGGTCCTCGAGGCGGTGGCCACCGAGG
GCAAGGGGGTCTTTGAGACCCTCAAGGAGGTGAGCCGCCTGGTCCTCGCCCGCGTGGCGGGCGG
GTCTTAAGCGGCCCACGCGCCCGGGGGCGGGGCGACCCGCCCCCTTTTACTCCCCACCCGTAT
TCCGGACTTTATAGGGACAAATGTCCCAAAAAGTCCCTTTGCCAAGGGACAAATGCCCCTCACC
GCAAAGCCCGGTTTCCGCTACCATGGAGGCCGTATACGGCACCTGCCGTTAGGGAGGAGGGGTA
TGGAAGAAAAGCCCAAAGGCGCACTGGCGGTCATCCTGGTCCTGACCCTCACCATCCTGGTCTT
CTGGCTGGGAGTGTACGCCGTCTTCTTCGCTAGGGGGTAGGTATGGTGGACGAGCACAAGGCCC
ACAAGGCGATCCTGGCCTACGAGAAGGGGTGGCTCGCCTTCTCCTTGGCGATGCTCTTCGTCTT
CATCGCCCTCATCGCCTACA**CCCTGGCCAC**

Figure 11. Bioinformatically characterized TTHA1359-binding sites in the *T. thermophilus* HB8 genome with regulatory potential. Sequences depict the genomic sequence context, -210/+310 bp, of the refined twenty-three FIMO-identified TTHA1359-binding sites analyzed for potential regulatory functions. Blue, red, green, and black nucleotides represent genes oriented 5'→ 3', overlapping genes oriented 5'→ 3', genes oriented 5'→ 3' on the complementary strand, and intergenic regions respectively. Cyan, yellow, and green highlighting represent predicted core promoter elements, potential TTHA1359-binding sites, and potential TTHA1359-binding sites overlapping core promoter elements respectively. Underlined nucleotides indicate translation start sites.

Table 3. Identification of potential TTHA1359-binding transcriptional activation sites

FIMO Result	Candidate Regulated Gene(s)	Potential Binding Site Distance to Predicted Transcription Start Site	Phase (bp)	Regulatory	Mechanism
2	*TTHA0953*	-74.5	3	N	-
4	*TTHA0425*	-41.5	0	Y	Activating, Class II
6	*TTHA0987*	-82.5	0	Y	Activating Class I
8a	*TTHA0446*	-65.5	4	N	-
8b	*TTHA0447*	-38.5	3	N	-
14a	*TTHA0533*	-44.5	3	N	-
35a	*TTHA0506*	-61.5	0	Y	Activating, Class I
36b	*TTHA0584*	-41.5	0	Y	Activating, Class II
37b	*TTHB089*	-93.5	1	Y	Activating, Class I
40a	*TTHA0606*	-58.5	3	N	-
51a	*TTHA0606*	-73.5	2	N	-
57b	*TTHA0206*	-99.5	3	N	-
58	*TTHA0770*	-41.5	0	Y	Activating, Class II

Phase indicates the alignment of potential binding sites in relation to predicted core promoter elements. A phase of 0 indicates sites are aligned. Other phase numbers indicate sites are unaligned and by how many bp they are out of alignment. Potential sites were considered likely (Y) or unlikely (N) to function as regulatory sites based on phase alignment.[78,79] Potential mechanisms associated with identified potential regulatory TTHA1359-binding sites were described as *Activating, Class I* if sites were positioned upstream of predicted -35 elements and *Activating, Class II* if sites overlapped predicted -35 elements.

CHAPER 4. DISCUSSION

In the Van Dyke laboratory, a reverse-genetic technique involving the combinatorial selection technique REPSA has been applied to study the possible regulatory and biological roles of TFs by identifying their DNA-binding specificities and the potential genes they regulate.[19-22] Since many TFs lack a level of characterization beyond a mere family categorization, an end goal in the Van Dyke laboratory is to implement this reverse-genetic technique as an effective alternative to other methods used to study putative TFs. Previously, this reverse-genetic technique had been applied to study the DNA-binding specificities and potential genes regulated by four *T. thermophilus* HB8 transcriptional repressors.[19-22] These TFs were specifically selected to study using this reverse-genetic technique because they possessed a pre-existing level of regulatory and biological characterization.[14,15,17,18] This allowed the findings from the reverse-genetic studies of these TFs to be compared to previously reported findings to evaluate the effectiveness of the reverse-genetic technique.[14,15,17-22] In general, the reverse-genetic studies with these four TFs not only corroborated previous findings but also reported additional findings which further characterized these TFs.[14,15,17-22] In the present work, several steps in the application of this reverse-genetic technique to identify the possible regulatory and biological roles of the previously studied *T. thermophilus* HB8 transcriptional activator TTHA1359 have been described as part of yet another study to evaluate the effectiveness of this reverse-genetic technique.[12,13]

The DNA-binding consensus 5'-A(T/A)TGT(G/A)A(N_6)T(C/T)ACA(A/T)T-3' was identified for TTHA1359 using REPSA to select sequences preferentially bound by TTHA1359, massively parallel sequencing to acquire the sequence information of these selections, and MEME to elucidate TTHA1359-binding motifs from obtained sequence information. Previous studies with TTHA1359 initially identified sixteen gene promoters which were upregulated in the presence of TTHA1359.[12,13] Eight of these promoters were initially identified through differential gene expression levels obtained from wild-type and *TTHA1359*-inactivated *T. thermophilus* HB8 strains.[12] The other eight gene promoters were initially identified by analyzing gene expression levels obtained from various experiments with *T. thermophilus* HB8 strains for individual genes whose expression showed strong positive correlations with *TTHA1359* expression.[13] After validating that these sixteen gene promoters were upregulated in a TTHA1359-dependent mechanisms in run-off transcription assays *in vitro*, putative TTHA1359-binding sites were assigned to these sixteen gene promoter sequences from which the putative consensus TTHA1359-binding site 5'-TTGTG(N_9)CNC-3' was derived.[12,13] Comparing this consensus to the consensus identified for TTHA1359 in the present work reveals a partial overlap between consensus sequences, particularly in the consensus composition along the 5'-end of both consensuses.[13] While the consensus from the previous studies was derived from sixteen predicted TTHA1359-binding sites, the TTHA1359-binding consensus in the present work was derived from a motif, primarily the top non-palindromic motif, discovered in seven hundred and eighty-nine experimentally selected unique sequences.[12,13] Therefore, greater confidence is placed in

the REPSA-selected TTHA1359-binding consensus as an accurate representation of the DNA-binding specificity of TTHA1359.

Prior to bioinformatically applying the TTHA1359-binding consensus identified in the present work to identify potential regulatory binding sites for TTHA1359 in the *T. thermophilus* HB8 genome, the consensus was evaluated biophysically by BLI to assess its true reflection of intrinsic TTHA1359-DNA binding specificity. This evaluation included testing not only TTHA1359-binding to the identified consensus but also to single-point and spacing mutants of the consensus to probe this specificity. TTHA1359 was observed to bind with high affinity, apparent K_D of ~ 3.4 nM, to the identified consensus, and single-point mutations to the consensus were found to lower this high binding affinity (Table 1). Of the single-point mutations to the consensus tested, some were found to dramatically reduce this binding affinity with no significant or detectable binding having transpired under the reaction conditions applied while others were found to reduce this binding affinity less radically with apparent K_D values ranging from ~ 23 to ~ 86 nM (Table 1). These single-point mutations were found to reduce TTHA1359-binding affinity in a method that reflected the significance of the positions in the TTHA1359-binding consensus as indicated by the top MEME non-palindromic motif (Figure 9A1, Table 1). Additionally, consensus spacing mutations were observed as having no significant or observable binding which suggests that TTHA1359 possesses a stringent consensus spacing requirement for recognition (Table 1). Because BLI assays were limited by the instrument to temperatures considered non-native for the protein, 37°C as opposed to at least 47°C, this consensus spacing requirement might merely reflect an experimental constraint.[81,82] Higher temperatures might introduce a level of

flexibility within the TTHA1359 dimer which might allow TTHA1359 to recognize shorter or lengthier consensus spacer sequences. With an apparent K_D of ~ 1 nM, the CRP_{Ec} consensus was the only sequence tested that exhibited higher binding affinity than that of the identified TTHA1359-binding consensus (Table 1). Because of a significant degree of consensus conservation between the tested TTHA1359-binding and CRP_{Ec}-binding consensus sequences, the higher binding affinity observed for the CRP_{Ec}-binding consensus was considered the cumulative result of differences between the two sequences (Figure 5A2, 9C, Table 1).[63] These differences included the 6-bp spacer sequences of these consensuses, 5'-TCTAGA-3' in the CRP_{Ec} consensus sequence as opposed to 5'-CACACA-3' in the TTHA1359 consensus sequence, and single bp differences in the half-sites between consensus sequences (Figure 5A2, 9C, Table 1).[63] Together, these kinetic experiments support an accurate reflection of preferred TTHA1359-DNA binding in the REPSA-selected TTHA1359-binding consensus.

Curiously, *in vitro* run-off transcription assays which were performed to demonstrate TTHA1359-dependent transcriptional upregulation at the promoters identified in previous studies utilized a concentration of 1 μM TTHA1359 dimer.[12,13] Results of these run-off transcription experiments were either black or white with an observed upregulation of transcription observed in reactions containing TTHA1359.[12,13] The relatively high concentration of TTHA1359 protein in these reactions solicited initial concern as BLI experiments in the present work suggest that TTHA1359 possesses a low level of binding specificity.[12,13] This was suggested by individual single-point mutant consensus experiments which found that TTHA1359-binding to its consensus sequence had an apparent K_D that was minimally ~ 8-fold higher than that of mutant

consensus sequences (Table 1). Additionally, a binding event between TTHA1359 and a control sequence was detected by BLI at 461 nM TTHA1359 dimer, the highest concentration of TTHA1359 protein used in BLI experiments (data not shown). Whether this interaction was specific or non-specific was not ascertained. Previously, the expression level of *TTHA1359* in the stationary phase had been reported measuring minimally ~ 8.5-fold higher than that of *TTHA1359* in the exponential phase of wild-type *T. thermophilus* HB8 growth during cultivation in different growth media.[12] However, a high concentration of expressed *TTHA1359* transcripts might not necessarily correlate with a high concentration of translated TTHA1359 protein. Based on the present work, concentrations of 1 μM TTHA1359 dimer used in previous studies to validate TTHA1359-dependent transcriptional regulation were initially considered excessive, resulting in false positive indications of regulation by TTHA1359 from non-specific binding driven by high protein concentrations.[12,13]

However, results of run-off transcription assays performed with TTHA1359 and increasingly shorter promoter sequences upstream of genes *TTHA0337* and *TTHA0634* suggested the presence of specific promoters sequences essential to the TTHA1359-dependent transcriptional upregulation of these gene promoters.[12] These reported results were baffling considering the predicted binding sites for TTHA1359 upstream of these genes, 5'-AGGGGAACCCACGGCACAC-3' and 5'-TTGTGCCTTTTACCCCT-3' respectively, were very different from the REPSA-selected TTHA1359-binding consensus.[12] As a result of the observed phenomenon, a hypothesis was formulated with regard to TTHA1359 transcriptional activity. If TTHA1359-binding conveys insufficient specificity to discriminately upregulate transcription, then perhaps a second mechanism

conveys further specificity. Since similarities have been reported for structures of TTHA1359 and CRP_{Ec} and drastically different binding affinities have been reported for CRP_{Ec} and its consensus compared to one of its most well-understood regulatory binding sites, studies with CRP_{Ec} were investigated for a potential answer.[12,63]

As it happens, simple CRP_{Ec}-dependent transcriptional activation at promoters has been reported to be dependent upon the positioning of binding sites for CRP_{Ec} relative to core promoter elements.[78,79] Studies in which a binding site for CRP_{Ec} was displaced at increasingly lengthier distances from the -35 core promoter element suggested a loss followed by a return of transcriptional activity depending on the distance introduced from the initial site of activity.[78,79] Spacing mutants for which transcriptional activity was lost corresponded to spacer sequences resulting in partial or half-turns of the DNA double-helix while activity was returned for spacer sequences resulting in full or nearly-full turns of the DNA, positioning CRP_{Ec} interactions with RNA polymerase holoenzyme out of or into alignment with core promoter elements respectively.[75,78,79]

Provided the importance of spacing for simple CRP_{Ec}-dependent transcriptional activation, a spacing requirement was imposed during the search for potential regulatory TTHA1359-binding sites based on distances from predicted core promoter elements. Initially, eighty-eight potential TTHA1359-binding sites within the *T. thermophilus* HB8 genome were identified. Of these, the first fifty-nine were selected for further investigation, totaling seventy-four potential binding sites when sites positioned near bidirectionally-oriented promoters with dual regulatory potential were accounted (Table 2). These potential sites were narrowed down to twenty-three TTHA1359-binding sites

with regulatory potential based on several criteria including the presence of predicted core promoter elements, the positioning of these promoter elements relative to translation start sites of local genes, and positioning of potential TTHA1359-binding sites relative to these promoter elements (Figure 11). Since the criteria for reducing the number of potential regulatory TTHA1359-binding sites revolved to a large extent around predicted core promoter elements, several potential regulatory binding sites for TTHA1359 might have been missed due perhaps to the inability to predict core promoter elements or the inaccurate prediction of core promoter element positioning. This is a current limitation of the reverse-genetic technique resulting from the orientation of promoter prediction software to predict core promoter elements based on those recognized by the housekeeping RNA polymerase holoenzyme in *E. coli* and not that of *T. thermophilus*.[90] While attempts by the Van Dyke laboratory are being made to more accurately predict and map *T. thermophilus* core promoter elements, the promoter prediction software used in this study was currently considered the best option to predict promoters with a user-defined restriction as to the allowable distance between predicted transcription and translation start sites.

Of the twenty-three refined TTHA1359-binding sites with regulatory potential, thirteen of the twenty-three were positioned upstream of predicted -35 core promoter elements with the potential to function as transcriptional activation binding sites for TTHA1359 (Table 3). However, only five of these thirteen potential TTHA1359-binding sites were positioned optimally upstream to activate transcription at gene promoters (Table 3). These five potential binding sites corresponded to those found upstream of genes *TTHA0425*, *TTHA0506*, *TTHA0584*, *TTHA0770*, and *TTHA0987*. A sixth potential

activating site for TTHA1359 was found upstream of *TTHB089*, but this site was considered unlikely to lead to substantial upregulation of *TTHB089* by TTHA1359 (Table 3). The eight remaining upstream potential sites were positioned poorly, oriented along different or opposite faces of the DNA double-helix as core promoter elements (Table 3). These potential TTHA1359-binding sites were considered unlikely to upregulate transcription of local downstream gene promoters. As a point of interest, the remaining ten potential binding sites overlapped the core promoter elements in a manner suggestive of transcriptional repression. Because previous studies with TTHA1359 only reported its activity as a transcriptional activator, the role of these sites has yet to be determined.[12,13]

In contrast to the absence of previous findings supporting a role for TTHA1359 as a transcriptional repressor, ample findings have been reported supporting the role of TTHA1359 as a transcriptional activator.[12,13] TTHA1359 has been demonstrated to upregulate transcription from promoters upstream of the sixteen genes *TTHA0425*, *TTHA0654*, *TTHA0337*, *TTHA0986*, *TTHA0770*, *TTHA1028*, *TTHA0634*, *TTHA0570*, TTHA0557, *TTHA0029*, *TTHA1128*, *TTHA1215*, *TTHA1625*, *TTHA1635*, *TTHA1892*, and *TTHB132* in run-off transcriptional assays *in vitro*.[12,13] Comparing these genes to potentially regulated genes identified in the present work, *TTHA0425*, *TTHA0506*, *TTHA0584*, *TTHA0770*, and *TTHA0987*, suggests the overlapping identification of only the two genes *TTHA0425* and *TTHA0770*.[12,13] A third gene, *TTHA0987*, would have overlapped the genes identified in both studies had it not failed to be validated in a previous study.[13] Previous TTHA1359 studies predicted TTHA1359-binding sites within ~ 30-50 bp of transcription start sites, positions suggestive of Class II transcription activation binding sites for TTHA1359.[12,13] However, in the case of the *TTHA0987* gene

promoter as determined in the current study, the TTHA1359 binding site maps to ~ 82 bp upstream of the transcription start site, indicative of a Class I transcriptional activation binding site for TTHA1359. Analyzing the run-off transcription template promoter sequence utilized to test *TTHA0987* promoter regulation by TTHA1359 in the previous study suggested the TTHA1359-binding site identified in the present study had been occluded by 20 bp.[13] This was considered the likely reason why the *TTHA0987* gene promoter failed to be validated in the previous study.[13] Interestingly, the *TTHA0987* gene promoter might be coregulated by TTHA1359 and TTHB032, a previously studied transcriptional repressor in the Van Dyke laboratory, based on potential regulatory binding sites for both TFs identified in the *TTHA0987* gene promoter.[17,22]

Of the five genes potentially activated by TTHA1359 based on optimally positioned potential regulatory binding sites in their promoters, the genes *TTHA0506* and *TTHA0584* appear to be novel genes which could have been missed in previous studies.[12,13] These sites could have been missed in previous genetic studies due to occlusion of these binding sites by chromosome organization or the activity of transcriptional repressors interfering with the activation of *TTHA0506* and *TTHA0584* gene promoters.[12,13] Regarding the remaining fourteen TTHA1359-regulated genes identified in previous studies with TTHA1359, these sites could have been missed using the reverse genetic technique due to promoter prediction limitations mentioned in a preceding paragraph or the limited number of FIMO-identified results investigated.[12,13] To test the latter possibility, the extended TTHA1359-binding consensus motif created should be applied using FIMO to search the promoter sequences of these fourteen genes used in run-off transcription assays in previous studies.[12,13] Identified TTHA1359 motif

matches within these promoter sequences should then be searched within the list of potential TTHA1359-binding sites in the *T. thermophilus* HB8 genome initially identified by FIMO using the extended TTHA1359-binding consensus motif. The initial FIMO scan could be expanded, if necessary, by resubmitting the initial search with a lower p-value threshold for matches to the TTHA1359-binding consensus. Identifiable since on the list would support that the missed genes reported in previous studies could have been identified as genes potentially activated by TTHA1359 using the reverse-genetic technique if more FIMO-identified sites would have been investigated.[12,13]

As the present work served as an additional test as to the effectiveness of a reverse-genetic technique involving REPSA to study TFs, a conclusion of the present work without evaluating the effectiveness of this reverse-genetic technique would leave the present work incomplete. Based on the findings of the current study, applying a reverse-genetic methodology involving REPSA alone to study global transcriptional activators appears inadequate to the task. A strength of the genetic approach previously applied to study TTHA1359 was the apparent *in vivo* identification of genes whose promoters were potentially regulated by TTHA1359 which allowed run-off transcription assays to be performed *in vitro* to validate TTHA1359-dependent regulation of the identified promoters.[12,13] In some instances, a genetic approach such as the approaches used to study TTHA1359 might miss potentially regulated genes.[12,13] Such a case is demonstrated by the *T. thermophilus* HB8 gene *TTHA0987* which was identified as potentially regulated by TTHA1359 but not validated for TTHA1359-dependent transcriptional upregulation previously.[13] Additionally, genes, potentially including *TTHA0506* and *TTHA0584* identified in the present work, might be missed due to other

factors complicating genetic approaches to studying TFs which have been mentioned previously. However, the genetic approaches used to study TTHA1359 identified more genes under the control of TTHA1359 than the reverse-genetic technique used in the present work.[12,13] Strengths of the reverse-genetic technique include the ability to characterize the intrinsic DNA-binding specificity of TFs and biophysically characterize TF-DNA interactions.[19-22] However, as suggested in the present work with TTHA1359, identifying potential regulatory binding sites within genomes based on statistically significant matches to the DNA-binding consensuses of TFs does not account for additional mechanisms such as binding site spacing from core promoter elements which might prove more critical to the regulatory roles of TFs than these matches. A combinatorial approach, understanding the DNA-binding specificity of TFs and application of such understanding to validate potentially regulated gene promoters identified by a genetic approach might prove a more effective method. However, since TTHA1359 functions as a global activator, it might represent a special case.[12,13]

Finally, a few future directions for the project have been described. Future work should seek to validate whether TTHA1359 upregulates the transcription of *TTHA0506* and *TTHA0584* which were genes not previously identified in previous studies with TTHA1359.[12,13] *In vitro* run-off transcription assays could be used to validate these genes as with the genes in previous TTHA1359 studies.[12,13] Additionally, an *in vitro* run-off transcription assay with the *TTHA0987* gene promoter should be performed incorporating both TTHA1359 and TTHB023 to determine whether coregulation of this gene occurs as suggested by prior work and the present study.[22] The fourteen TTHA1359-regulated genes reported in previous studies but missed in the present work

should be investigated for potential TTHA1359-binding sites using FIMO and the extended TTHA1359-binding motif to identify whether these genes would have been identified if more FIMO-identified matches were investigated.[12,13] Potential regulatory binding sites for TTHA1359 which were found to overlap predicted core promoter elements should be examined more closely to identify the potential regulatory roles of these sites. This might be accomplished by obtaining the TTHA1359-inactivated and wild-type *T. thermophilus* HB8 gene expression levels collected previously for genes downstream of these sites to provide insight into their potential regulatory roles.[12,13]

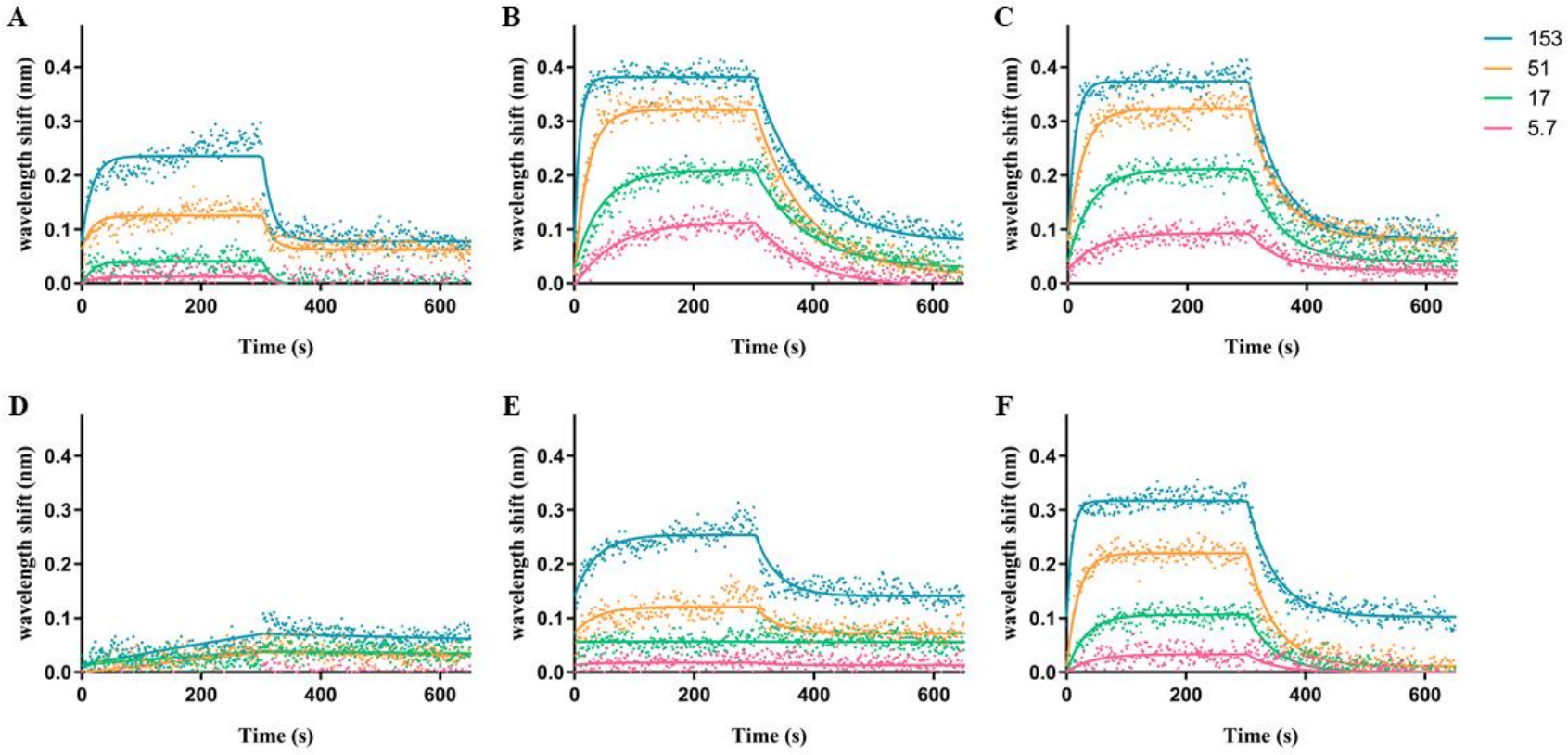
APPENDIX A: OLIGONUCLEOTIDES

Table A1. Oligonucleotides

Name	Sequence	Length	Purification	Application
ST2R24	CTAGGAATTCGTGCAGAGGTGAATNNNNNNNNNNNNNNNNNNNNNNNNNTTACCATCCCTCCAGAAGCTTGGAC	73	PAGE	Selection Template ST2R24 Precursor
ST2L	CTAGGAATTCGTGCAGAGGTGAAT	24	Desalt	PCR Primer
ST2Ls	CTAGGAATTCGTGCAGAGGTGA	22	Desalt	PCR Primer
IRD7_ST2R	/5IRD700/GTCCAAGCTTCTGGAGGGATGGTAA	25	HPLC	5'-labeled IRDye700 PCR Primer
A_BC06_ST2R	CCATCTCATCCCTGCGTGTCTCCGACTCAGCTGCAAGTTCGATGTCCAAGCTTCTGGAGGGATG	64	PAGE	Fusion PCR Primer
trP1_ST2L	CCTCTCTATGGGCAGTCGGTGATCTAGGAATTCGTGCAGAGGTGA	45	PAGE	Fusion PCR Primer
A_uni	CCATCTCATCCCTGCGTG	18	Desalt	PCR Primer
trP1_uni	CCTCTCTATGGGCAGTCGG	19	Desalt	PCR primer
Bio_ST2R	/5BiodT/GTCCAAGCTTCTGGAGGGATG	22	HPLC	5'-biotinylated PCR primer
ST2_1359_ctrl	CTAGGAATTCGTGCAGAGGTGAATACGAAAAACACACACCATCCCTCCAGAAGCTTGGAC	60	Desalt	TTHA1359 Control Probe Precursor
ST2_1359_con_wt	GGAATTCGTGCAGAGGTGAATTGTGACACACATCACAATCATCCCTCCAGAAGCTTGG	58	Desalt	Wild-type TTHA1359 Consensus Probe Precursor
ST2_1359_con_p1	GGAATTCGTGCAGAGGTGACTTGTGACACACATCACAATCATCCCTCCAGAAGCTTGG	58	Desalt	TTHA1359 Consensus Position 1 Point Mutation Probe precursor
ST2_1359_con_p2	GGAATTCGTGCAGAGGTGAAGTGTGACACACATCACAATCATCCCTCCAGAAGCTTGG	58	Desalt	TTHA1359 Consensus Position 2 Point Mutation Probe Precursor
ST2_1359_con_p3	GGAATTCGTGCAGAGGTGAATGGTGACACACATCACAATCATCCCTCCAGAAGCTTGG	58	Desalt	TTHA1359 Consensus Position 3 Point Mutation Probe Precursor
ST2_1359_con_p4	GGAATTCGTGCAGAGGTGAATTTTGACACACATCACAATCATCCCTCCAGAAGCTTGG	58	Desalt	TTHA1359 Consensus Position 4 Point Mutation Probe Precursor
ST2_1359_con_p5	GGAATTCGTGCAGAGGTGAATTGGGACACACATCACAATCATCCCTCCAGAAGCTTGG	58	Desalt	TTHA1359 Consensus Position 5 Point Mutation Probe Precursor
ST2_1359_con_p6	GGAATTCGTGCAGAGGTGAATTGTTACACACATCACAATCATCCCTCCAGAAGCTTGG	58	Desalt	TTHA1359 Consensus Position 6 Point Mutation Probe Precursor
ST2_1359_con_p7	GGAATTCGTGCAGAGGTGAATTGTGGCACACATCACAATCATCCCTCCAGAAGCTTGG	58	Desalt	TTHA1359 Consensus Position 7 Point Mutation Probe Precursor
ST2_1359_con_s5	GGAATTCGTGCAGAGGTGAATTGTGACACACTCACAATCATCCCTCCAGAAGCTTGG	57	Desalt	TTHA1359 Shortened Spacer Consensus Mutation Probe Precursor
ST2_1359_con_s7	GGAATTCGTGCAGAGGTGAATTGTGACACACACTCACAATCATCCCTCCAGAAGCTTGG	59	Desalt	TTHA1359 Lengthened Spacer Consensus Mutation Probe Precursor
ST2_CRP_Ec	GGAATTCGTGCAGAGGTGAAATGTGATCTAGATCACATTTCATCCCTCCAGAAGCTTGG	59	Desalt	E. coli CRP Consensus Probe Precursor

Length is reported in nucleotides (nt).

APPENDIX B: ADDITIONAL BLI ASSOCIATION AND DISSOCIATION STEP RAW DATA PLOTS OF TTHA1359-DNA BINDING



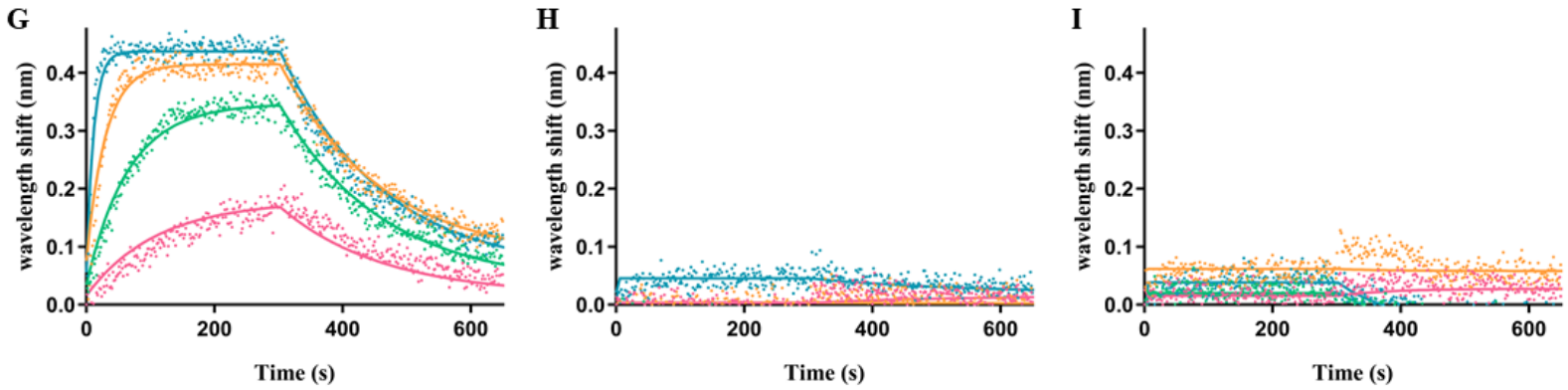
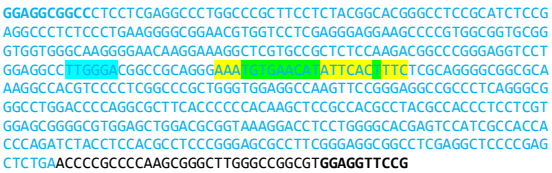


Figure B1. Additional BLI association and dissociation step raw data plots of TTHA1359-DNA binding. Graphs depict raw association and dissociation step data measured during BLI experiments with TTHA1359 and (**A**) ST2_1359_con_p1, (**B**) ST2_1359_con_p2, (**C**) ST2_1359_con_p3, (**D**) ST2_1359_con_p4, (**E**) ST2_1359_con_p5, (**F**) ST2_1359_con_p6, and (**G**) ST2_1359_con_p7 biotinylated DNAs. These DNAs correspond to TTHA1359-binding consensus sequences containing single point mutations at position 1, 2, 3, 4, 5, 6, and 7 respectively. Graphs also depict the raw step data measured during BLI experiments with TTHA1359 and (**H**) ST2_1359_con_s5 and (**I**) ST2_1359_con_s7 biotinylated DNAs. These DNAs correspond to TTHA1359-binding consensus sequences containing 5- and 7-bp spacer sequences respectively instead of the 6-bp spacer of the wild-type TTHA1359-binding consensus. Dots represent raw data points, and solid lines depict the calculated best-fit lines for these raw data points. Line colors pink, green, orange, and blue indicate different concentrations of TTHA1359 protein corresponding to 5.7, 17, 51, and 153 nM TTHA1359 dimer.

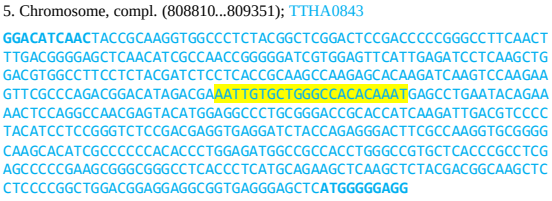
APPENDIX C: BIOINFORMATICALLY CHARACTERIZED TTHA1359-BINDING SITES WITHIN THE *T. THERMOPHILUS* HB8 GENOME EXCLUDED FROM ANALYSIS

1. Chromosome, compl. (917451...917992); TTHA0974



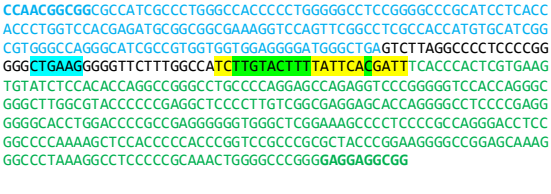
3. Miniplasmid pTT8 (1293...1834); TTHC001

5. Chromosome, compl. (808810...809351); TTHA0843



7. Chromosome, compl. (752303...752844); TTHA0785 and TTHA0784

11. Chromosome (932321....932862); TTHA0987 and TTHA0988



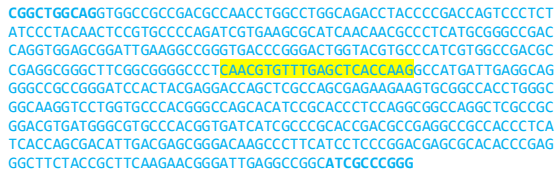
12. Chromosome (871545...872086); TTHA0910 and TTHA0911

13. Chromosome, compl. (612877...613418); TTHA0644

15. Chromosome (1507570...1508111); TTHA1584

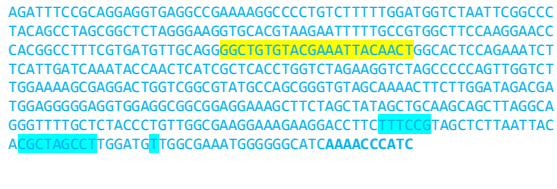
16. Megaplasmid, compl. (249070...249611); TTHB244 and TTHB243

18. Chromosome (1718740...1719281); TTHA1836



19. Chromosome, compl. (890582...891123); TTHA0943 and TTHA0942

20. Chromosome (32491...33032); TTHA0030

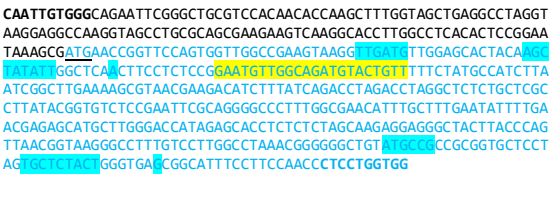


21a. Chromosome, compl. (1622223...1622764); TTHA1730 and TTHA1729

21b. Chromosome (1622322...1622864); TTHA1729 and TTHA1730

22. Chromosome, compl. (884821...885362); TTHA0937 and TTHA0936

23. Chromosome, compl. (891012...891553); TTHA0943



24. Chromosome (1161778...1162319); TTHA1222 and TTHA1223

25. Chromosome, compl. (1173772...1174313); TTHA1237 and TTHA1236

26. Chromosome (1104088...1104629); TTHA1158

27. Chromosome, compl. (455872...456413); TTHA0486

28. Chromosome, compl. (326070...326611); TTHA0345 and TTHA0344

29a. Chromosome, compl. (1464682...1465223); TTHA1538 and TTHA1537

29b. Chromosome (1464782...1465323); TTHA1537 and TTHA1538

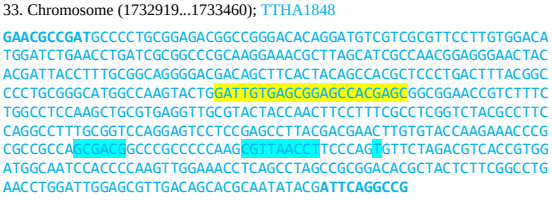
30. Megaplasmid, compl. (100927...101468); TTHB114 and TTHB113

31a. Chromosome, compl. (54213...54754); TTHA0056 and TTHA0055

31b. Chromosome (54313...54854); TTHA0055 and TTHA0056

32. Chromosome, compl. (1241025...1241566); TTHA1301

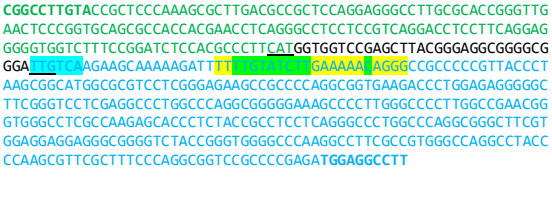
33. Chromosome (1732919...1733460); TTHA1848



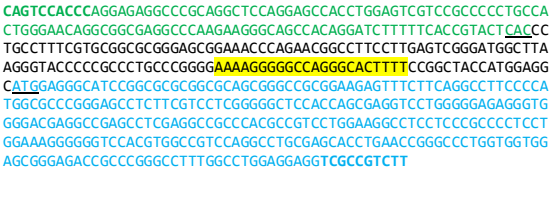
34a. Chromosome, compl. (1292149...1292690); TTHA1352

34b. Chromosome (1292249...1292790); TTHA1352 and TTHA1353

35b. Chromosome (471990...472531); TTHA0506 and TTHA0507

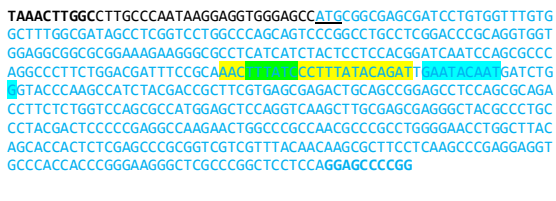


36a. Chromosome, compl. (550450...550991); TTHA0584 and TTHA0583



37a. Megaplasmid, compl. (79293...79834); TTHB089 and TTHB088

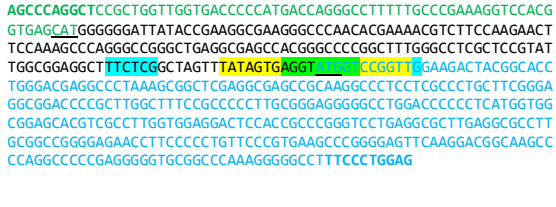
38. Megaplasmid, compl. (171372...171913); TTHB177



39. Chromosome (1507126...1507667); TTHA1583 and TTHA1584

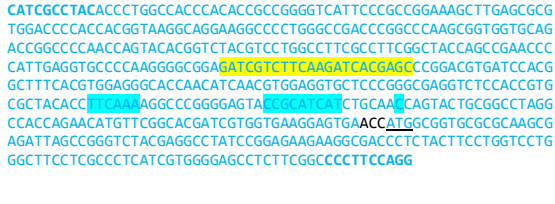
41. Chromosome (83523...84064); TTHA0083 and TTHA0084

42a. Chromosome, compl. (290509...291050); TTHA0304 and TTHA0303

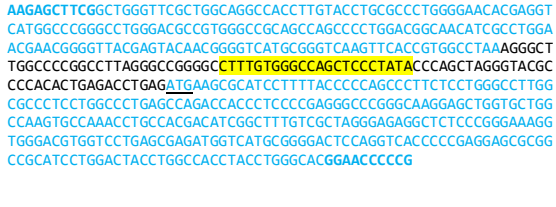


42b. Chromosome (290609...291150); TTHA0303 and TTHA0304

43. Chromosome (1075724...1076265); TTHA1134 and TTHA1135



44. Chromosome (1265433...1265974); TTHA1325 and TTHA1326



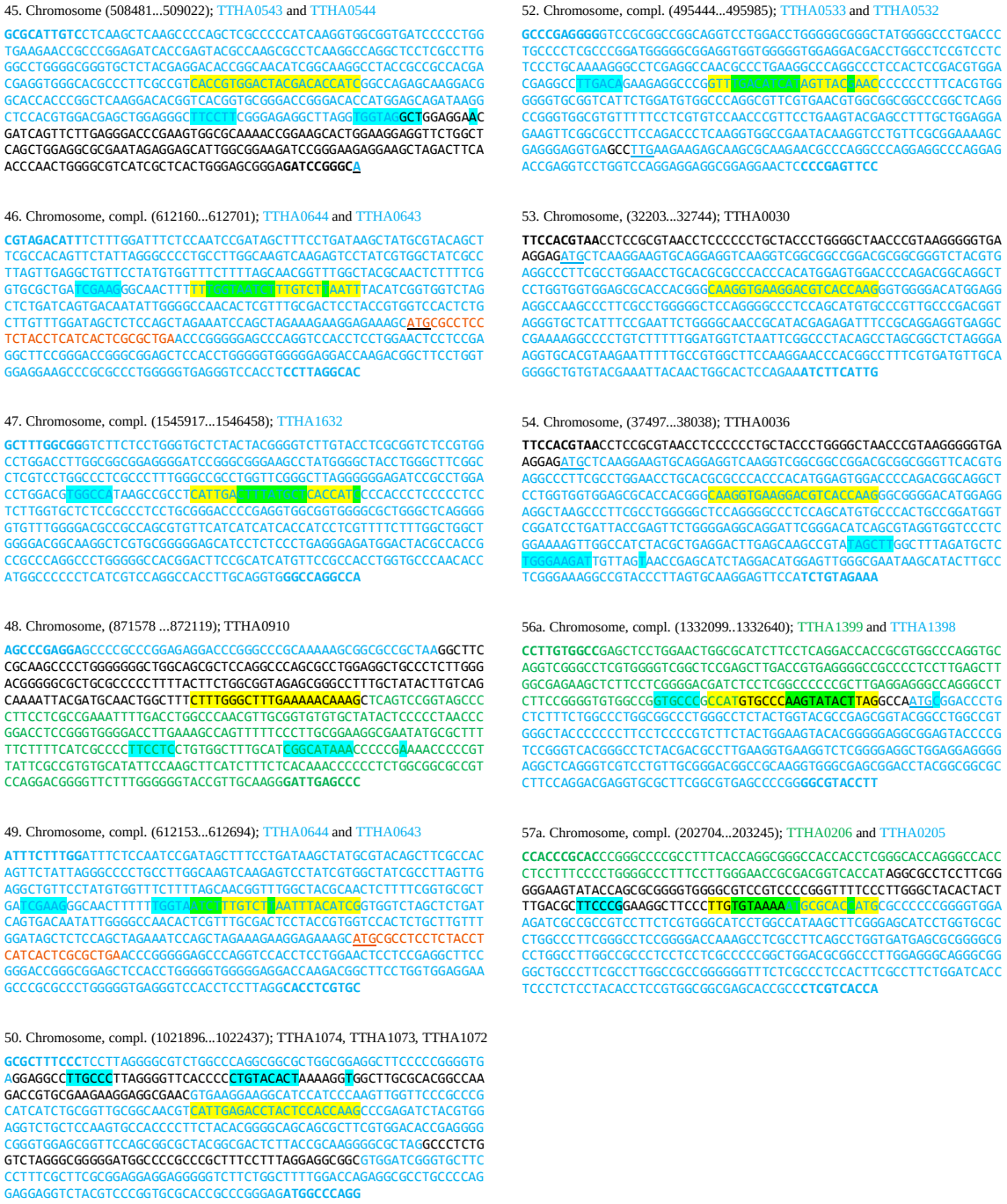


Figure C1. Bioinformatically characterized potential TTHA1359-binding sites within the T. thermophilus HB8 genome excluded from analysis. Sequences depict the genomic context of potential binding sites, -210/+310 bp, identified by FIMO which were excluded from further analysis. Blue, red, green, and black nucleotides represent genes oriented 5'→ 3', overlapping genes oriented 5'→ 3', genes oriented 5'→ 3' on the complementary strand, and intergenic regions respectively. Cyan, yellow, and green highlighting represent predicted core promoter elements, potential TTHA1359-binding sites, and potential TTHA1359-binding sites overlapping core promoter elements respectively. Underlined nucleotides indicate translation start sites.

REFERENCES

1. Pérez-Rueda, E.; Tenorio-Salgado, S.; Huerta-Saquero, A.; Balderas-Martínez, Y. I.; Moreno-Hagelsieb, G. The functional landscape bound to the transcription factors of Escherichia coli K-12. *Computational Biology and Chemistry* **2015**, *58*, 93-103. DOI:10.1016/j.compbiolchem.2015.06.002.

2. Perez-Rueda, E.; Hernandez-Guerrero, R.; Martinez-Nuñez, M. A.; Armenta-Medina, D.; Sanchez, I.; Ibarra, J. A. Abundance, diversity and domain architecture variability in prokaryotic DNA-binding transcription factors. *PloS ONE* **2018**, *13*, e0195332. DOI:10.1371/journal.pone.0195332.

3. Ulrich, L. E.; Koonin, E. V.; Zhulin, I. B. One-component systems dominate signal transduction in prokaryotes. *Trends in Microbiology* **2005**, *13*, 52-56. DOI:10.1016/j.tim.2004.12.006.

4. Martínez-Antonio, A.; Collado-Vides, J. Identifying global regulators in transcriptional regulatory networks in bacteria. *Current Opinion in Microbiology* **2003**, *6*, 482-489. DOI:10.1016/j.mib.2003.09.002.

5. Browning, D. F.; Busby, S. J. W. The regulation of bacterial transcription initiation. *Nature Reviews Microbiology* **2004**, *2*, 57-65. DOI:10.1038/nrmicro787.

6. Lee, D. J.; Minchin, S. D.; Busby, S. J. W. Activating Transcription in Bacteria. *Annual Review of Microbiology* **2012**, *66*, 125-152. DOI:10.1146/annurev-micro-092611-150012.

7. Browning, D. F.; Busby, S. J. W. Local and global regulation of transcription initiation in bacteria. *Nature reviews. Microbiology* **2016**, *14*, 638-650. DOI:10.1038/nrmicro.2016.103.

8. Blattner, F. R.; Plunkett III, G.; Bloch, C. A.; Perna, N. T.; Burland, V.; Riley, M.; Collado-Vides, J.; Glasner, J. D.; Rode, C. K.; Mayhew, G. F.; Gregor, J.; Davis, N. W.; Kirkpatrick, H. A.; Goeden, M. A.; Rose, D. J.; Mau, B.; Shao, Y. The Complete Genome Sequence of Escherichia coli K-12. *Science* **1997**, *277*, 1453-1462. DOI:10.1126/science.277.5331.1453.

9. Pérez-Rueda, E.; Collado-Vides, J. The repertoire of DNA-binding transcriptional regulators in Escherichia coli K-12. *Nucleic Acids Research* **2000**, *28*, 1838-1847. DOI:10.1093/nar/28.8.1838.

10. Santos-Zavaleta, A.; Salgado, H.; Gama-Castro, S.; Sánchez-Pérez, M.; Gómez-Romero, L.; Ledezma-Tejeida, D.; García-Sotelo, J. S.; Alquicira-Hernández, K.; Muñiz-Rascado, L. J.; Peña-Loredo, P.; Ishida-Gutiérrez, C.; Velázquez-Ramírez, D. A.; Del Moral-Chávez, V.; Bonavides-Martínez, C.; Méndez-Cruz, C.; Galagan, J.; Collado-Vides, J. RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. *Nucleic Acids Research* **2019**, *47*, D212-D220. DOI:10.1093/nar/gky1077.

11. Akeo Shinkai; Satoshi Kira; Noriko Nakagawa; Aiko Kashihara; Seiki Kuramitsu; Shigeyuki Yokoyama Transcription Activation Mediated by a Cyclic AMP Receptor Protein from Thermus thermophilus HB8. *Journal of Bacteriology* **2007**, *189*, 3891-3901. DOI:10.1128/JB.01739-06.

12. Agari, Y.; Kashihara, A.; Yokoyama, S.; Kuramitsu, S.; Shinkai, A. Global gene expression mediated by Thermus thermophilus SdrP, a CRP/FNR family transcriptional regulator. *Molecular Microbiology* **2008**, *70*, 60-75. DOI:10.1111/j.1365-2958.2008.06388.x.

13. Agari, Y.; Kuramitsu, S.; Shinkai, A. Identification of novel genes regulated by the oxidative stress-responsive transcriptional activator SdrP in Thermus thermophilus HB8. *FEMS Microbiology Letters* **2010**, *313*, 127-134. DOI:10.1111/j.1574-6968.2010.02133.x.

14. Agari, Y.; Agari, K.; Sakamoto, K.; Kuramitsu, S.; Shinkai, A. TetR-family transcriptional repressor Thermus thermophilus FadR controls fatty acid degradation. *Microbiology* **2011**, *157*, 1589-1601. DOI:10.1099/mic.0.048017-0.

15. Sakamoto, K.; Agari, Y.; Kuramitsu, S.; Shinkai, A. Phenylacetyl Coenzyme A is an Effector Molecule of the TetR Family Transcriptional Repressor PaaR from Thermus thermophilus HB8. *Journal of Bacteriology* **2011**, *193*, 4388-4395. DOI:10.1128/JB.05203-11.

16. Agari, Y.; Kuramitsu, S.; Shinkai, A. X-ray crystal structure of TTHB099, a CRP/FNR superfamily transcriptional regulator from Thermus thermophilus HB8, reveals a DNA-binding protein with no required allosteric effector molecule. *Proteins: Structure, Function, and Bioinformatics* **2012**, *80*, 1490-1494. DOI:10.1002/prot.24049.

17. Yoshihiro Agari; Keiko Sakamoto; Seiki Kuramitsu; Akeo Shinkai Transcriptional Repression Mediated by a TetR Family Protein, PfmR, from Thermus thermophilus HB8. *Journal of Bacteriology* **2012**, *194*, 4630-4641. DOI:10.1128/JB.00668-12.

18. Agari, Y.; Sakamoto, K.; Yutani, K.; Kuramitsu, S.; Shinkai, A. Structure and function of a TetR family transcriptional regulator, SbtR, from Thermus thermophilus HB8. *Proteins: Structure, Function, and Bioinformatics* **2013**, *81*, 1166-1178. DOI:10.1002/prot.24266.

19. Van Dyke, M. W.; Beyer, M. D.; Clay, E.; Hiam, K. J.; McMurry, J. L.; Xie, Y. Identification of Preferred DNA-Binding Sites for the Thermus thermophilus Transcriptional Regulator SbtR by the Combinatorial Approach REPSA. *PLOS ONE* **2016**, *11*, e0159408. DOI:10.1371/journal.pone.0159408.

20. Lee, M.; Um, H.; Van Dyke, M. W. Identification and characterization of preferred DNA-binding sites for the Thermus thermophilus transcriptional regulator FadR. *PLOS ONE* **2017**, *12*, e0184796. DOI:10.1371/journal.pone.0184796.

21. Cox, J. S.; Moncja, K.; Mckinnes, M.; Van Dyke, M. W. Identification and Characterization of Preferred DNA-Binding Sites for the Thermus thermophilus HB8 Transcriptional Regulator TTHA0973. *International Journal of Molecular Sciences* **2019**, *20*, 3336. DOI:10.3390/ijms20133336.

22. Cox, J. S.; Van Dyke, M. W. General and Genomic DNA-Binding Specificity for the Thermus thermophilus HB8 Transcription Factor TTHB023. *Biomolecules* **2020**, *10*, 94. DOI:10.3390/biom10010094.

23. Stormo, G. D.; Zhao, Y. Determining the specificity of protein–DNA interactions. *Nature Reviews Genetics* **2010**, *11*, 751-760. DOI:10.1038/nrg2845.

24. Schneider, T. D.; Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research* **1990**, *18*, 6097-6100. DOI:10.1093/nar/18.20.6097.

25. Stormo, G. D. Modeling the specificity of protein-DNA interactions. *Quantitative Biology* **2013**, *1*, 115-130. DOI:10.1007/s40484-013-0012-4.

26. Garner, M. M.; Revzin, A. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic Acids Research* **1981**, *9*, 3047-3060. DOI:10.1093/nar/9.13.3047.

27. Fried, M. G.; Crothers, D. M. Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Research* **1981**, *9*, 6505-6525. DOI:10.1093/nar/9.23.6505.

28. Hellman, L. M.; Fried, M. G. Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nature Protocols* **2007**, *2*, 1849-1861. DOI:10.1038/nprot.2007.249.

29. Galas, D. J.; Schmitz, A. DNAse Footprinting: A Simple Method for the Detection of protein-DNA Binding Specificity. *Nucleic Acids Research* **1978**, *5*, 3157-3170. DOI:10.1093/nar/5.9.3157.

30. Hampshire, A. J.; Rusling, D. A.; Broughton-Head, V. J.; Fox, K. R. Footprinting: A Method for Determining the Sequence Selectivity, Affinity and Kinetics of DNA-binding ligands. *Methods* **2007**, *42*, 128-140. DOI:10.1016/j.ymeth.2007.01.002.

31. Jain, S. S.; Tullius, T. D. Footprinting protein-DNA complexes using the hydroxyl radical. *Nature Protocols* **2008**, *3*, 1092-1100. DOI:10.1038/nprot.2008.72.

32. Zheng, M.; Wang, X.; Templeton, L. J.; Smulski, D. R.; LaRossa, R. A.; Storz, G. DNA Microarray-Mediated Transcriptional Profiling of the Escherichia coli Response to Hydrogen Peroxide. *Journal of Bacteriology* **2001**, *183*, 4562-4570. DOI:10.1128/JB.183.15.4562-4570.2001.

33. Ren, B.; Robert, F.; Wyrick, J. J.; Aparicio, O.; Jennings, E. G.; Simon, I.; Zeitlinger, J.; Schreiber, J.; Hannett, N.; Kanin, E.; Volkert, T. L.; Wilson, C. J.; Bell, S. P.; Young, R. A. Genome-wide Location and Function of DNA Binding Proteins. *Science* **2000**, *290*, 2306-2309. DOI:10.1126/science.290.5500.2306.

34. Johnson, D. S.; Mortazavi, A.; Myers, R. M.; Wold, B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science* **2007**, *316*, 1497-1502. DOI:10.1126/science.1141319.

35. Park, P. J. ChIP-seq: Advantages and Challenges of a Maturing Technology. *Nature Reviews Genetics* **2009**, *10*, 669-680. DOI:10.1038/nrg2641.

36. Furey, T. S. ChIP–seq and Beyond: New and Improved Methodologies to Detect and Characterize protein–DNA Interactions. *Nature Reviews Genetics* **2012**, *13*, 840-852. DOI:10.1038/nrg3306.

37. Mukherjee, S.; Berger, M. F.; Jona, G.; Wang, X. S.; Muzzey, D.; Snyder, M.; Young, R. A.; Bulyk, M. L. Rapid Analysis of the DNA-binding Specificities of Transcription Factors with DNA Microarrays. *Nature Genetics* **2004**, *36*, 1331-1339. DOI:10.1038/ng1473.

38. Berger, M. F.; Philippakis, A. A.; Qureshi, A. M.; He, F. S.; Estep III, P. W.; Bulyk, M. L. Compact, Universal DNA Microarrays to Comprehensively Determine Transcription-Factor Binding Site Specificities. *Nature Biotechnology* **2006**, *24*, 1429-1435. DOI:10.1038/nbt1246.

39. Berger, M. F.; Bulyk, M. L. Universal Protein-Binding Microarrays for the Comprehensive Characterization of the DNA-binding Specificities of Transcription Factors. *Nature Protocols* **2009**, *4*, 393-411. DOI:10.1038/nprot.2008.195.

40. Ellington, A. D.; Szostak, J. W. In Vitro Selection of RNA Molecules that Bind Specific Ligands. *Nature* **1990**, *346*, 818-822. DOI:10.1038/346818a0.

41. Blackwell, T. K.; Weintraub, H. Differences and Similarities in DNA-binding Preferences of MyoD and E2A Protein Complexes Revealed by Binding Site Selection. *Science* **1990**, *250*, 1104-1110. DOI:10.1126/science.2174572.

42. Tuerk, C.; Gold, L. Systematic Evolution of Ligands by Exponential Enrichment: RNA Ligands to Bacteriophage T4 DNA Polymerase. *Science* **1990**, *249*, 505-510. DOI:10.1126/science.2200121.

43. Wright, W. E.; Binder, M.; Funk, W. Cyclic Amplification and Selection of Targets (CASTing) for the Myogenin Consensus Binding Site. *Molecular and Cellular Biology* **1991**, *11*, 4104-4110. DOI:10.1128/mcb.11.8.4104.

44. Roulet, E.; Busso, S.; Camargo, A. A.; Simpson, A. J. G.; Mermod, N.; Bucher, P. High-throughput SELEX-SAGE method for quantitative modeling of transcription-factor binding sites. *Nature Biotechnology* **2002**, *20*, 831-835. DOI:10.1038/nbt718.

45. Zhao, Y.; Granas, D.; Stormo, G. D. Inferring Binding Energies from Selected Binding Sites. *PLoS Computational Biology* **2009**, *5*, e1000590. DOI:10.1371/journal.pcbi.1000590.

46. Van Dyke, M. W.; Van Dyke, N.; Sunavala-Dossabhoy, G. REPSA: General combinatorial approach for identifying preferred ligand–DNA binding sequences. *Methods* **2007**, *42*, 118-127. DOI:10.1016/j.ymeth.2006.09.008.

47. Szybalski, W.; Kim, S. C.; Hasan, N.; Podhajska, A. J. Class-IIS restriction enzymes — a review. *Gene* **1991**, *100*, 13-26. DOI:10.1016/0378-1119(91)90345-C.

48. Hardenbol, P.; Van Dyke, M. W. Sequence Specificity of Triplex DNA Formation: Analysis by a Combinatorial Approach, Restriction Endonuclease Protection Selection and Amplification. *Proceedings of the National Academy of Sciences of the United States of America* **1996**, *93*, 2811-2816. DOI:10.1073/pnas.93.7.2811.

49. Hardenbol, P.; Wang, J. C.; Van Dyke, M. W. Identification of preferred hTBP DNA binding sites by the combinatorial method REPSA. *Nucleic Acids Research* **1997**, *25*, 3339-3344.

50. Hardenbol, P.; Wang, J. C.; Van Dyke, M. W. Identification of Preferred Distamycin-DNA Binding Sites by the Combinatorial Method REPSA. *Bioconjugate Chemistry* **1997**, *8*, 617-620.

51. Shen, J.; Wang, J. C.; Van Dyke, M. W. Identification of Preferred Actinomycin–DNA Binding Sites by the Combinatorial Method REPSA. *Bioorganic & Medicinal Chemistry* **2001**, *9*, 2285-2293.

52. Vashisht Gopal, Y. N.; Van Dyke, M. W. Combinatorial Determination of Sequence Specificity for Nanomolar DNA-Binding Hairpin Polyamides. *Biochemistry* **2003**, *42*, 6891-6903.

53. Sunavala-Dossabhoy, G.; Van Dyke, M. W. Combinatorial Identification of a Novel Consensus Sequence for the Covalent DNA-Binding Polyamide Tallimustine. *Biochemistry* **2005**, *44*, 2510-2522.

54. Körner, H.; Sofia, H. J.; Zumft, W. G. Phylogeny of the Bacterial Superfamily of Crp-Fnr Transcription Regulators: Exploiting the Metabolic Spectrum by Controlling Alternative Gene Programs. *FEMS Microbiology Reviews* **2003**, *27*, 559-592. DOI:10.1016/S0168-6445(03)00066-4.

55. Passner, J. M.; Steitz, T. A. The Structure of a CAP-DNA Complex Having Two cAMP Molecules Bound to Each Monomer. *Proceedings of the National Academy of Sciences of the United States of America* **1997**, *94*, 2843-2847. DOI:10.1073/pnas.94.7.2843.

56. Parkinson, G.; Wilson, C.; Gunasekera, A.; Ebright, Y. W.; Ebright, R. H.; Berman, H. M. Structure of the CAP-DNA Complex at 2.5 Å Resolution: A Complete Picture of the Protein-DNA Interface. *Journal of Molecular Biology* **1996**, *260*, 395-408. DOI:10.1006/jmbi.1996.0409.

57. Heyduk, T.; Lee, J. C. Escherichia coli cAMP Receptor Protein: Evidence for Three Protein Conformational States with Different Promoter Binding Affinities. *Biochemistry* **1989**, *28*, 6914-6924. DOI:10.1021/bi00443a021.

58. Heyduk, T.; Lee, J. C. Application of fluorescence energy transfer and polarization to monitor Escherichia coli cAMP receptor protein and lac promoter interaction. *Proceedings of the National Academy of Sciences of the United States of America* **1990**, *87*, 1744-1748. DOI:10.1073/pnas.87.5.1744.

59. Pyles, E. A.; Lee, J. C. Mode of Selectivity in Cyclic AMP Receptor Protein-Dependent Promoters in Escherichia coli. *Biochemistry* **1996**, *35*, 1162-1172. DOI:10.1021/bi952187q.

60. Grainger, D. C.; Hurd, D.; Harrison, M.; Holdstock, J.; Busby, S. J. W. Studies of the Distribution of Escherichia coli cAMP-Receptor Protein and RNA Polymerase along the E. coli Chromosome. *Proceedings of the National Academy of Sciences of the United States of America* **2005**, *102*, 17693-17698. DOI:10.1073/pnas.0506687102.

61. Shimada, T.; Fujita, N.; Yamamoto, K.; Ishihama, A. Novel Roles of cAMP Receptor Protein (CRP) in Regulation of Transport and Metabolism of Carbon Sources. *PLoS One* **2011**, *6*, e20081. DOI:10.1371/journal.pone.0020081.

62. Bailey, T. L.; Elkan, C. In *In Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Biopolymers;* AAAI Press: Menlo Park, California, 1994; , pp 28-36.

63. Ebright, R. H.; Ebright, Y. W.; Gunasekera, A. Consensus DNA site for the Escherichia coli catabolite gene activator protein (CAP): CAP exhibits a 450-fold higher affinity for the consensus DNA site than for the E. coli lac DNA site. *Nucleic Acids Research* **1989**, *17*, 10295-10305. DOI:10.1093/nar/17.24.10295.

64. Trageser, M.; Spiro, S.; Duchêne, A.; Kojro, E.; Fahrenholz, F.; Guest, J. R.; Unden, G. Isolation of Intact FNR Protein (Mr 30,000) of Escherichia coli. *Molecular Microbiology* **1990**, *4*, 21-27. DOI:10.1111/j.1365-2958.1990.tb02011.x.

65. Lazazzera, B. A.; Beinert, H.; Khoroshilova, N.; Kennedy, M. C.; Kiley, P. J. DNA Binding and Dimerization of the Fe-S-containing FNR Protein from Escherichia coli Are Regulated by Oxygen. *The Journal of Biological Chemistry* **1996**, *271*, 2762-2768. DOI:10.1074/jbc.271.5.2762.

66. Khoroshilova, N.; Popescu, C.; Münck, E.; Beinert, H.; Kiley, P. J. Iron-sulfur Cluster Disassembly in the FNR Protein of Escherichia coli by O2: [4Fe-4S] to [2Fe-2S] Conversion with Loss of Biological Activity. *Proceedings of the National Academy of Sciences of the United States of America* **1997**, *94*, 6087-6092. DOI:10.1073/pnas.94.12.6087.

67. Crack, J.; Green, J.; Thomson, A. J. Mechanism of Oxygen Sensing by the Bacterial Transcription Factor Fumarate-Nitrate Reduction (FNR). *Journal of Biological Chemistry* **2004**, *279*, 9278-9286. DOI:10.1074/jbc.M309878200.

68. Sutton, V. R.; Mettert, E. L.; Beinert, H.; Kiley, P. J. Kinetic Analysis of the Oxidative Conversion of the [4Fe-4S]2+ Cluster of FNR to a [2Fe-2S]2+ Cluster. *Journal of Bacteriology* **2004**, *186*, 8018-8025. DOI:10.1128/JB.186.23.8018-8025.2004.

69. Salmon, K.; Hung, S.; Mekjian, K.; Baldi, P.; Hatfield, G. W.; Gunsalus, R. P. Global gene expression profiling in Escherichia coli K12. The effects of oxygen availability and FNR. *The Journal of Biological Chemistry* **2003**, *278*, 29837-29855. DOI:10.1074/jbc.M213060200.

70. Kang, Y.; Weber, K. D.; Qiu, Y.; Kiley, P. J.; Blattner, F. R. Genome-Wide Expression Analysis Indicates that FNR of Escherichia coli K-12 Regulates a Large Number of Genes of Unknown Function. *Journal of Bacteriology* **2005**, *187*, 1135-1160. DOI:10.1128/JB.187.3.1135-1160.2005.

71. Constantinidou, C.; Hobman, J. L.; Griffiths, L.; Patel, M. D.; Penn, C. W.; Cole, J. A.; Overton, T. W. A Reassessment of the FNR Regulon and Transcriptomic Analysis of the Effects of Nitrate, Nitrite, NarXL, and NarQP as Escherichia coli K12 Adapts from Aerobic to Anaerobic Growth. *The Journal of Biological Chemistry* **2006**, *281*, 4802-4815. DOI:10.1074/jbc.M512312200.

72. Myers, K. S.; Yan, H.; Ong, I. M.; Chung, D.; Liang, K.; Tran, F.; Keleș, S.; Landick, R.; Kiley, P. J. Genome-scale Analysis of Escherichia coli FNR Reveals Complex Features of Transcription Factor Binding. *PLoS Genetics* **2013**, *9*, e1003565. DOI:10.1371/journal.pgen.1003565.

73. Eiglmeier, K.; Honoré, N.; luchi, S.; Lin, E. C.; Cole, S. T. Molecular Genetic Analysis of FNR-dependent Promoters. *Molecular Microbiology* **1989**, *3*, 869-878. DOI:10.1111/j.1365-2958.1989.tb00236.x.

74. Scott, C.; Partridge, J. D.; Stephenson, J. R.; Green, J. DNA target sequence and FNR-dependent gene expression. *FEBS Letters* **2003**, *541*, 97-101. DOI:10.1016/s0014-5793(03)00312-0.

75. Busby, S.; Ebright, R. H. Transcription Activation by Catabolite Activator Protein (CAP). *Journal of Molecular Biology* **1999**, *293*, 199-213. DOI:10.1006/jmbi.1999.3161.

76. Liu, B.; Hong, C.; Huang, R. K.; Yu, Z.; Steitz, T. A. Structural basis of bacterial transcription activation. *Science* **2017**, *358*, 947-951. DOI:10.1126/science.aao1923.

77. Feng, Y.; Zhang, Y.; Ebright, R. H. Structural basis of transcription activation. *Science* **2016**, *352*, 1330-1333. DOI:10.1126/science.aaf4417.

78. Gaston, K.; Bell, A.; Kolb, A.; Buc, H.; Busby, S. Stringent Spacing Requirements for Transcription Activation by CRP. *Cell* **1990**, *62*, 733-743. DOI:10.1016/0092-8674(90)90118-x.

79. Ushida, C.; Aiba, H. Helical Phase Dependent Action of CRP : Effect of the Distance between the CRP Site and the −35 Region on Promoter Activity. *Nucleic Acids Research* **1990**, *18*, 6325-6330. DOI:10.1093/nar/18.21.6325.

80. Cava, F.; Hidalgo, A.; Berenguer, J. Thermus thermophilus as Biological Model. *Extremophiles* **2009**, *13*, 213-231. DOI:10.1007/s00792-009-0226-6.

81. Oshima, T.; Imahori, K. Isolation of an Extreme Thermophile and Thermostability of its Transfer Ribonucleic Acid and Ribosomes. *The Journal of General and Applied Microbiology* **1971**, *17*, 513-517. DOI:10.2323/jgam.17.513.

82. Oshima, T.; Imahori, K. Description of Thermus thermophilus (Yoshida and Oshima) comb. nov., a Nonsporulating Thermophilic Bacterium from a Japanese Thermal Spa. *International Journal of Systematic Bacteriology* **1974**, *24*, 102-112. DOI:10.1099/00207713-24-1-102.

83. Ohtani, N.; Tomita, M.; Itaya, M. An Extreme Thermophile, Thermus thermophilus, Is a Polyploid Bacterium. *Journal of Bacteriology* **2010**, *192*, 5499-505. DOI:10.1128/JB.00662-10.

84. Yokoyama, S.; Hirota, H.; Kigawa, T.; Yabuki, T.; Shirouzu, M.; Terada, T.; Ito, Y.; Matsuo, Y.; Kuroda, Y.; Nishimura, Y.; Kyogoku, Y.; Miki, K.; Masui, R.; Kuramitsu, S. Structural genomics projects in Japan. *Nature Structural Biology* **2000**, *7 Suppl*, 943-5. DOI:10.1038/80712.

85. Schultz, S. C.; Shields, G. C.; Steitz, T. A. Crystal Structure of a CAP-DNA Complex: The DNA Is Bent by 90 Degrees. *Science* **1991**, *253*, 1001-1007. DOI:10.1126/science.1653449.

86. The UniProt Consortium UniProt: A Worldwide Hub of Protein Knowledge. *Nucleic Acids Research* **2019**, *47*, D506-D515. DOI:10.1093/nar/gky1049.

87. Grant, C. E.; Bailey, T. L.; Noble, W. S. FIMO: Scanning for Occurrences of a Given Motif. *Bioinformatics (Oxford, England)* **2011**, *27*, 1017-1018. DOI:10.1093/bioinformatics/btr064.

88. Kanehisa, M.; Sato, Y.; Furumichi, M.; Morishima, K.; Tanabe, M. New Approach for Understanding Genome Variations in KEGG. *Nucleic Acids Research* **2019**, *47*, D590-D595. DOI:10.1093/nar/gky962.

89. Karp, P. D.; Billington, R.; Caspi, R.; Fulcher, C. A.; Latendresse, M.; Kothari, A.; Keseler, I. M.; Krummenacker, M.; Midford, P. E.; Ong, Q.; Ong, W. K.; Paley, S. M.; Subhraveti, P. The BioCyc Collection of Microbial Genomes and Metabolic Pathways. *Briefings in Bioinformatics* **2019**, *20*, 1085-1093. DOI:10.1093/bib/bbx085.

90. Solovyev, V.; Salamov, A. Automatic Annotation of Microbial Genomes and Metagenomic Sequences. In *Metagenomics and Its Applications in Agriculture, Biomedicine and Environmental Studies*; Li, R. W., Ed.; Nova Science Publishers: 2011; pp 61-78.

91. Gartenberg, M. R.; Crothers, D. M. DNA Sequence Determinants of CAP-induced Bending and Protein Binding Affinity. *Nature* **1988**, *333*, 824-829. DOI:10.1038/333824a0.

92. Dalma-Weiszhausz, D. D.; Gartenberg, M. R.; Crothers, D. M. Sequence-dependent Contribution of Distal Binding Domains to CAP Protein-DNA Binding Affinity. *Nucleic Acids Research* **1991**, *19*, 611-616. DOI:10.1093/nar/19.3.611.

93. Gunasekera, A.; Ebright, Y. W.; Ebright, R. H. DNA Sequence Determinants for Binding of the Escherichia Coli Catabolite Gene Activator Protein. *Journal of Biological Chemistry* **1992**, *267*, 14713-14720.

94. Concepcion, J.; Witte, K.; Wartchow, C.; Choo, S.; Yao, D.; Persson, H.; Wei, J.; Li, P.; Heidecker, B.; Ma, W.; Varma, R.; Zhao, L.; Perillat, D.; Carricato, G.; Recknor, M.; Du, K.; Ho, H.; Ellis, T.; Gamez, J.; Howes, M.; Phi-Wilson, J.; Lockard, S.; Zuk, R.; Tan, H. Label-free Detection of Biomolecular Interactions Using BioLayer Interferometry for Kinetic Characterization. *Combinatorial Chemistry & High Throughput Screening* **2009**, *12*, 791-800. DOI:10.2174/138620709789104915.