Kennesaw State University
DigitalCommons@Kennesaw State University

Master of Science in Chemical Sciences | Chemistry & Biochemistry

Summer 7-29-2015

REPSA Directed Assessment of Native Cleavage Resistance of DNA to Type IIS Restriction Endonucleases and Modification of REPSA for High Temperature Application

Matthew Beyer
Kennesaw State University

Follow this and additional works at: http://digitalcommons.kennesaw.edu/mscs_etd

 Part of the Biochemistry Commons

Recommended Citation
Beyer, Matthew, "REPSA Directed Assessment of Native Cleavage Resistance of DNA to Type IIS Restriction Endonucleases and Modification of REPSA for High Temperature Application" (2015). *Master of Science in Chemical Sciences.* Paper 5.

This Thesis is brought to you for free and open access by the Chemistry & Biochemistry at DigitalCommons@Kennesaw State University. It has been accepted for inclusion in Master of Science in Chemical Sciences by an authorized administrator of DigitalCommons@Kennesaw State University. For more information, please contact digitalcommons@kennesaw.edu.

REPSA Directed Assessment of Native Cleavage Resistance of DNA to Type IIS Restriction Endonucleases and Modification of REPSA for High Temperature Application

by

Matthew D. Beyer

BS

Emory University, 2010

Submitted in Partial Fulfillment of the Requirements

For the Degree of Master of Science in the

Department of Chemistry and Biochemistry

Kennesaw State University

2015

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

Acknowledgments

I would like to than Dr. Michael Van Dyke for being an invaluable mentor. I would also like to thank my committee for their guidance and support as well as all of the faculty and staff of the Kennesaw State University graduate program. I would like to thank Dr. Kirk and Mrs. Manly for their invaluable support.

Finally, I would like to thank my wife, Nicole Davis, as well as my family and friends for dealing with and supporting me when times were tough.

REPSA Directed Assessment of Native Cleavage Resistance of DNA to Type IIS Restriction Endonucleases and Modification of REPSA for High Temperature Application

Abstract

We have modified the combinatorial selection method Restriction Endonuclease Protection and Selection Assay (REPSA) to work in high temperature conditions for the discovery of new DNA-binding proteins in thermophiles (HT-REPSA). We utilized *Thermus thermophilus* (HB-8/ATCC 27634/DSM 579) as a test organism due to its amenable nature in a laboratory setting and current status as a model thermophilic organism. We used a TetR Family (TFR) transcription factor SbtR as the model protein for optimization of HT-REPSA protocols, as data had previously been obtained regarding SbtR physical characteristics and DNA-binding properties. REPSA was conducted until a cleavage resistant species arose after 7 rounds. Massively parallel sequencing of the selected DNAs and bioinformatics analysis yielded a consensus binding sequence of 5'-GA(t/c)TGACC(c/a)GC(t/g)GGTCA(g/a)TC, a 20base pair palindromic site comparable to that described in the literature. Taken together, our data provide a proof-of-concept that HT-REPSA can be successfully used to identify the preferred DNA-binding sequences of transcription factors from extreme thermophilic organisms.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Introduction

Currently, there is an information lag between genome sequencing and genomic understanding. Due to rapid increases in computing power and concomitant automation, sequencing has become a relatively inexpensive and straightforward process. GOLD (Genome Online Database) has approximately ~63,000 genomes stored from a multitude of sequencing projects, probing biomes spread across the entirety of the planet (Reddy et al., 2014). However, the actual understanding of the information encoded in these genomes, as well as the network of proteins (transcription factors) which control information access, has significantly lagged behind (Hanson et al., 2009). Large portions of these genomes contain genes of unidentified or unknown function. These largely unknown segments of the genome that have no known function are often referred to as orphan proteins in the literature (Hanson et al., 2009).

Understanding which open reading frames (the genomic area between a start codon and a stop codon which often corresponds to a gene) are regulated by which transcription factors is a slow process. Advances have been made, adapting traditional methods such as PBM (**P**rotein **B**inding **M**icroarray) (Berger et al., 2009), SELEX (**S**ystematic **E**volution of **L**igands by **EX**ponential enrichment) (Djordjevic, 2007), or ChIP (**Ch**romatin **I**mmuno**P**recipitation) (Collas, 2010) for higher throughput application, though the assay rate is still limited and prone to error (Dey et al., 2012). These methods, in addition to several others described later, are generally reliable for currently known transcription

factors (TFs) and such methods can be used to rapidly, if expensively, obtain consensus sequences for known TFs (Dey et al. 2012). However, these methods fade in applicability when applied to the discovery of unknown TFs. The predominant high-throughput methodologies for obtaining valid consensus sequences rely on physical separation methods, which explicitly depends on previous knowledge of the TFs and its physical characteristics (Van Dyke et al., 2007).

Complementary to our lack of understanding of TF and ORF (Open Reading Frame) assignment is the gap in understanding how that genomic information is regulated, often referred to as the regulome (Cordero& Hogeweg, 2009; Vicente & Mingorance, 2008). The majority of ORFs for many organisms, and the genes described therein, are currently appraised as having no known biological function (Hanson et al., 2009). For example, *Escherichia coli*, the biotechnological workhorse species that has been under extensive study and manipulation for over 60 years, has less than half of the total number of putative genes annotated in any fashion (Baba et al. 2006; Yamamoto et al., 2009; Gama-Castro et al., 2010).

Thus we are left with a question of how to accurately and quickly obtain consensus sequences for unknown transcription factors in a high throughput manner. REPSA, detailed in section 2.3, is a methodology that has generated valid consensus sequences for native DNA-binding proteins in addition to being sensitive enough to generate relevant consensus sequences from complex solutions (Hardenbol & Van Dyke, 1996; Hardenbol & Van Dyke, 1997; Tonhat et al., 2010). Due to these two properties, as well as its overall simplicity, it holds the potential for use in high throughput discovery of heretofore

unknown transcription factors, especially those that possess unknown structural motifs and thus remain unannotated (Van Dyke et al., 2007).

Notably, not all organisms, particularly those that occupy extreme environments from a human point of view (extremophiles), are amenable for REPSA experimentation in its current form. This is specifically due to temperature limitations of IISREs (Type IIS Restriction Endonucleases) currently used by this method. Organisms that occupy these niches are generally of substantial biotechnological interest (Cavicchioli et al., 2002; Demirjian et al. 2001; Frock et al., 2012; Kumar et al., 2011, Stan-Lotter & Fendrihan, 2012), being the source of a number of now widely utilized enzymes that have revolutionized biotechnology. In order to assess transcription factors from organisms that occupy high temperature ecological niches, the REPSA procedure must be adjusted.

Thermus thermophilus HB8, a polyextremeophile bacterium, is one such organism for which our genomic understanding is lacking (section 2.5). It is a model thermophilic species, a minimal model of life, as well as a putative model for early life on Earth (Cava et al., 2009). Due in part to this model status, there is a need to understand how it regulates its genome, and thus a current project, the Structural-Biological Whole Cell Project, is being undertaken by the RIKEN Institute of Japan. Data from this project is stored in the Whole Cell Project Database (WCPDB). This project is utilizing current technologies to predicted transcription factors at a rate of roughly one per year (Agari et al., 2008; Agari et al., 2011; Sakamoto et al., 2011; Agari et al., 2012; Agari et al., 2013). However, via other means, there are approximately 70 predicted TFs (Table A2, Appendix A), of which 12 have been elucidated in detail and had their consensus sequences obtained since *T. thermophilus* was first identified in 1974 (Ohsima &

Imahori, 1974; Sanchez et al; 2000; Agari et al., 2008; Sevostynova & Arsimovitch, 2010; Agari et al., 2011; Sakamoto et al., 2011; Agari et al., 2012; Agari et al., 2013; Iwanaga et al., 2014)). This organism, given the extreme conditions it must endure while continuing to process information at the genome level, may well possess DNA-binding protein motifs that have yet to be identified. The first step we can contribute towards better understanding its TFs and regulome is a modified REPSA for high temperature application, as described in this thesis.

CHAPTER 2: BACKGROUND

Transcription Factors

Protein-DNA interactions, particularly with regard to transcriptional regulators, are key to understanding how an organisms regulates information accession, stored via genes, in response to environmental stimuli. Organisms require specificity in these interactions, otherwise such interactions would not possess the means to respond to stimuli in a meaningful or efficient manner. A subset of transcriptional regulators are transcriptions factors (TF), a broad class of proteins with a common purpose, regulating transcription of open reading frames (ORFs) within an organism in response to stimuli (Babu & Teichmann, 2003). To accomplish this regulation, TFs utilize various motifs to bind to specific DNA elements. These motifs are broad and varied, and a TF may utilize multiple motifs, some extremely well represented within all domains of life and some restricted to use within families (Charoensawan et al., 2010; Lohse et al., 2013).). Motifs may be highly stereotyped, varying little structurally across the domains, while others are structurally vague, hidden in the generic group of intrinsically disordered proteins (IDP) (Charoensawan et al., 2010).

Current Methods of Obtaining Consensus Sequences for Transcription Factors

Many methods, both computational and experimental, have been developed to assess the DNA and TF interactions in order to ascertain probable regions of specific

interaction, a consensus sequence, both *in vivo* and *in vitro*, with varying degrees of bias, cost, and ease of use. The primary experimental methods of ascertaining consensus sequences for DNA-binding proteins such as transcription factors are described following. For further review, see also Dey et al., 2012.

DNA Footprinting. DNA footprinting is a cleavage protection assay that identifies the site of transcription factor binding by way of site-specific inhibition of DNA cleavage. DNA footprinting relies on the DNA cleavage properties of the desired cleavage agent and the cleavage inhibitory properties of the DNA-binding ligand (Hampshire et al., 2007). Optimally, the cleavage pattern of the agent should not display bias, presenting all bands in equal intensity when visualized. Thus far, the "perfect" neutral cleaver has remained elusive. At present, the most widely utilized agents are DNase I and hydroxyl radicals [H_2O_2.Fe(II)] although others such as MPE (methidiumpropyl-EDTA.Fe(II)) have been used (Van Dyke et al., 1982; Van Dyke & Dervan, 1983; Hampshire et al., 2007; Jain & Tullius, 2007). Radiolabeling or fluorescent labeling are preferred for visualization of the reaction products, though fluorescent tagging is not as sensitive as radiolabeling (Hampshire et al., 2007)

Footprinting is powerful and quite sensitive to DNA-ligand interactions. However, it is limited with regards to the number of sites that can be assayed due to both acrylamide gel electrophoresis limitations (50-200 base pairs/reaction) and the reality of effectively assessing the large number of potential sequences, given the equation $4^n/2$ where n is the binding sequence length and the 2 accounts for DNA complementarity (Hampshire et al., 2007). For example, a DNA binding site that is four nucleotides long gives 128 potential combinations, five gives 256, and six gives 512. The current assessable binding site size

survey is thus realistically limited to six base pairs for footprinting methods. This limitation can be partially circumvented by use of nested sequences (de Bruijn sequence construction) as in the case of MS-1 and MS-2 for tetra nucleotide size binding sites, though this method is not currently feasible for probing of binding sites larger than six base pairs (Hampshire et al., 2007). Thus, if the size of the binding site is small (four bp or less), then footprinting can be reliably used to assay all potential sequences and extract a consensus from analysis of the highest affinity sites. Addition of capillary gel electrophoresis to the method has increased throughput as well as increasing assayable DNA length to 400 bp (Yindeeyoungyeon & Schell, 2000). However, these advances are insufficient to propel footprinting methods to identify consensus sequences much greater than 6 base pairs.

PBM: Protein Binding Microarrays. PBM is a combinatorial methodology that relies on fluorescence and antibody affinity to determine transcription factor interactions with DNA. A strand of DNA is anchored to a microarray plate and then probed with a protein ligand of interest. An antibody, usually tagged with a fluorescent molecule for visualization, is then applied that is specific to that protein. This limits PBMs to either proteins that have specific antibodies available or to recombinant proteins with highly utilized epitope or fluorescent tags, e.g., FLAG-tag, GFP, etc. PBM, like other combinatorial methodologies, is primarily limited by the size of the binding site it can effectively probe for in a reasonable and cost effective manner. At present, a universal 10-mer oligomer group, containing all possible 10 nt long sequences utilizing de Brujin stacking, is the longest assay group yet synthesized (Berger et al. 2009).

Protein binding microarrays require large quantities of protein in order to accurately assess binding motifs, which can be problematic for recombinant proteins that

are difficult to manufacture in quantity at full length or those that require post-translational modifications. Furthermore, PBMs tend to reveal only the highest affinity binding sequences, due to multiple wash steps required to remove non-specific interactions, which may not necessarily be biologically relevant DNA motifs (Berger et al. 2009).

SELEX, CASTing and *in vitro* selection. SELEX (Systematic Evolution of Ligands by EXponential Enrichment) is a combinatorial method that involves physical isolation of DNA-protein complexes from a larger pool of DNA followed by selective amplification of those isolated DNA-protein complexes (Tuerk & Gold, 1990). Also referred to as either CASTing (Cyclic Amplification and Selection of Templates) (Wright et al. 1991), or *in vitro* selection (Ellington & Szostak 1990), it incredibly useful and robust for analyzing single protein-DNA/interactions. They have been utilized to ascertain a variety of consensus sequences for many proteins with high degrees of accuracy and biological validity (Jolma et al., 2013; Nitta, et al., 2015). The modification of SELEX to genomic SELEX (Zimmermann et al., 2010), using a ligand probe against a genomic DNA library for a given organism, further increased the biological relevance of such sequences though the sequences lacked the diversity of standard SELEX. Further enhancements were made in 2002 (SELEX-SAGE) and again in 2009 (High Throughput (HT)-SELEX) with the automation of the SELEX process, thereby allowing for high throughput analysis of protein-DNA interactions (Roulet et al., 2002; Zhao et al., 2009).

SELEX was originally utilized to probe and analyze high-affinity small molecule, protein, RNA, and DNA interactions with a heterogeneous DNA oligonucleotide pool (Tuerk & Gold, 1990; Wright et al., 1991; Ellington & Szostak 1990). However, it readily

became apparent that though SELEX was powerful, the aptamers that were generated were not always biologically relevant in terms of the consensus sequences obtained (Djordjevic, 2007; Stormo & Zhao, 2010). Natural selection does not necessarily select for the highest possible affinity interactions within the interactome. Rather, the types of interactions that are selected for are those that increase the overall survivability and adaptability of the organism and these sequences may vary in kinetic and thermodynamic parameters (Stormo & Zhao, 2010). What may be suboptimal in terms of raw binding affinity may be highly specific and optimal for the functioning of the metabolic network as a whole and would thus be selected for (Stormo & Zhao, 2010).

ChIP: Chromatin Immunoprecipitation. ChIP is an *in vivo* method that covalently fixes DNA bound protein followed by DNA shearing, and subsequent physical separation of the DNA-protein complex from the larger chromatin pool. ChIP (Chromatin Immuno-precipitation), in all its myriad variations, is a fairly robust and common methodology for determination of protein-DNA interaction sites *in vivo* (Gade & Kalvakolanu, 2012). ChIP obtained consensus sequences are often reliable, providing insights into regulatory network shifts due to tissue and temporal differences in chromatin structure, though the practice is not without difficulty (Stormo & Zhao, 2010). ChIP is a cumbersome technology, requiring extensive time and large material input for a single protein assay, especially with regard to transcription factors as only a few copies are likely to be manufactured per cell. Binding site resolution is poor, 100bp being an average length for a co-precipitated DNA fragment. Immuno-precipitated samples are only enriched for proteins of interest. As a result, they are not pure with regards to a

given proteins interaction, and there is often significant background noise from these outliers in a given ChIP experiment (Gade & Kalvakolanu, 2012; Collas, 2010).

To better identify potential binding sites by ChIP, proteins can be manipulated to be produced at higher than wild-type levels. However, this runs the risk of obtaining off-target sites (Gade & Kalvakolanu, 2012). In addition, unless an antibody already exists for the protein under investigation, either recombinant proteins with a common epitope tag and corresponding antibody must be used, or an antibody has to be generated that is specific to the protein of interest and is also competent for protein-DNA complex immunoprecipitation. This leads to increased experimental time and cost per protein assayed (Collas, 2010).

REPSA Methodology

REPSA, in contrast to the previously mentioned methodologies, lacks the need to apply physical separation methods to obtain viable consensus sequences for proteins of interest (Van Dyke et al., 2007). The general REPSA protocol (Figure 1A), relies instead on the putative neutral cleavage characteristics of type IIS restriction endonucleases (IISRE), a multi-domain restriction endonuclease with separate recognition and cleavage domains tethered by a short linker, that cleaves DNA in a sequence independent manner at a fixed location relative to the recognition site (Szybalski et al., 1991). During a REPSA round, a combinatorial (heterogeneous) pool of DNA oligomers is incubated with the transcription factor of interest under physiological temperatures for a given time determined during optimization. These oligomers contain a random core region flanked on either side by defined regions containing recognition sites for IISREs, oriented so that

they cleave within the random core, some, likely very small portion of the oligomer pool will interact with and bind specifically with the transcription factor. The complex pool will then be incubated with a IISRE, resulting in cleavage for unprotected (unbound) oligomers and cleavage protection with the transcription factor bound oligomers (Van Dyke et al., 2007). Cleavage protected, intact templates will amplify, whereas cleaved templates will not (Figure 1A) (Van Dyke et al., 2007). This leads to preferential amplification of ligand-binding templates such that these templates will increase in representation relative to those templates that cannot bind ligands post-amplification. Thus the template pool at the end of a given round will be greatly enriched for templates that specifically bound to the TF. Repeated rounds of binding, selection and amplification will eventually yield a population that is enriched and largely composed of templates which contain specific, high-affinity binding sites for the TF (Figure 1B and Figure 1C) (Van Dyke et al., 2007).

Type IIS Restriction Endonuclease (IISRE). Type IISREs are key to REPSA application. There are dozens of known IISREs (Roberts et al., 2014). Like most all restriction endonucleases, they are not classed based on phylogenetic relationships but rather on their relatively fluid cleavage behaviors (Szybalski et al., 1991; Pingoud et al., 2014). As such, they are a rather enigmatic class and their behavior is assumed to follow that of the class archetype, FokI (Szybalski et al., 1991; Wah et al., 1998)

(IISRE) are a unique subset of a broad class of type II restriction endonucleases. They preferentially cleave at a relatively fixed distance outside of their recognition site, which is usually short (4-8 bp) and asymmetric. The distance varies depending on the tether length as well as its slippage characteristics, i.e. the likelihood of non-stereotyped

cleavage products by movement of the enzyme cleavage domain (Lundin et al., 2015). The 'S' designate for this subgroup refers to this distal **s**hift between the recognition site and the cleavage site (Szybalski et al., 1991). They, like most type II restriction endonucleases possess a $PDx_n(D/E)xK$ catalytic center, in which water molecules and a divalent metal ion, usually magnesium, are coordinated in such a manner that they act to stabilize the phosphate backbone for cleavage (Kosinski et al., 2005).

The cleavage is usually asymmetric, with the cleavage occurring on one side of the recognition site, though some Type IISREs are symmetric cleavers, e.g. BpuJ, which cleaves on both sides of the recognition site (Roberts et al., 2014). They usually require some divalent metal ion (Fe^{2+}, Mn^{2+}, Zn^{2+}, Co^{2+}, or Mg^{2+}) for their function and like most other type II REs, primarily use Mg^{2+} to yield the highest enzymatic activity (Szybalski et al., 1991). Calcium seems to inhibit most known type IISREs, likely by disrupting the coordination of the backbone in the active site due to its large ionic radius (~1Å) relative to the other divalent cations previously listed (Szybalski et al., 1991). However, the IISRE BifI, which possess a catalytic domain similar to phosolipase D (PLD), has been shown to not require metal ion cofactors for catalysis and possesses only a single catalytic center (Sasnauskas et al., 2003). Most important, IISREs appear to display no sequence preference for cleavage of the phosphate backbone, though DNA modifications (e.g. methylation), and superstructure modification (e.g. supercoiling or histone condensation) can hinder their activity (Szybalski et al., 1991; Pingoud et al., 2014).

FokI is the archetypal IISRE as it is the most studied of the class. Of the 68 restriction endonuclease structures stored in REBASE, either whole or fragment, 18 are classed into type IIS (Roberts et al., 2014).

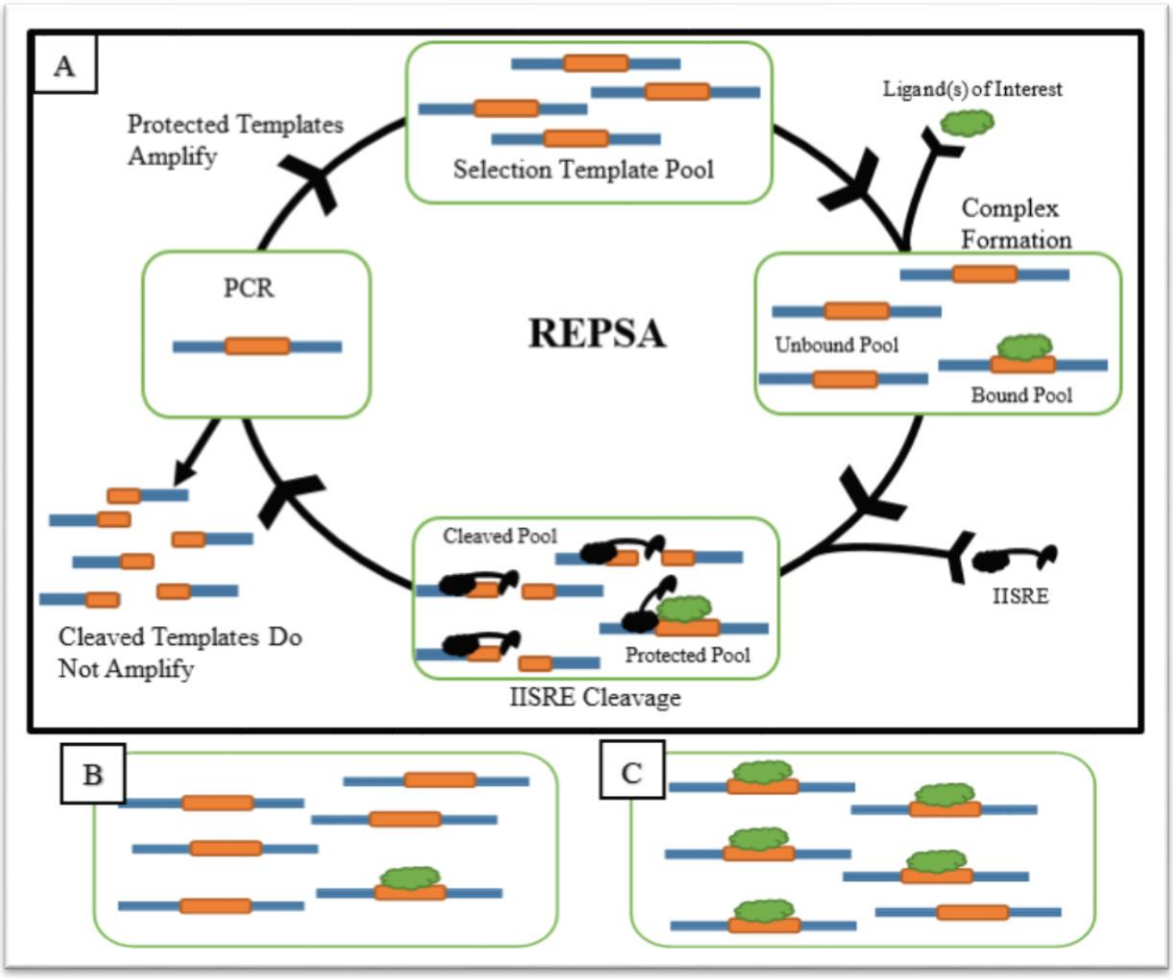

Figure 1. (**A**) Basic REPSA Protocol. A template pool is incubated with a ligand or complex ligand mixture for a set time to permit ligand association with binding sites on oligomers. A IISRE is added and cleavage of unbound templates occurs. After the IISRE is disabled, PCR is performed, with protected templates amplifying while cleaved templates do not. Templates are then purified before a subsequent round of REPSA is performed. (**B, C**): Expected enrichment of ligand binding pool between initial, poorly cleavage resistant template pool, **B**, and highly cleavage resistant pool, **C**, enriched for ligand binding sites after several rounds of REPSA.

Given the similarity displayed with asymmetric type IISREs, they are assumed to behave in a similar manner to FokI (Pingoud et al., 2014). FokI is a 61-kDa protein composed of three distinct domains: a 41-kDa recognition domain, a 20-kDa cleavage domain, and a short, rigid tether domain connecting the two (Wah et al., 1998). FokI binds DNA at its recognition site as a monomer and this interaction seems to prompt the extension of the cleavage domain (Pernstich et al., 2012). Cleavage requires homo-dimerization, mediated by the dimerization domain on the cleavage head (Bitinaite et al., 1998). This dimerization is thought to require the other FokI monomer to also be bound to DNA, at a different recognition site, bound to DNA either in *cis* (same DNA molecule) or in *trans* (separate DNA molecule) (Catto et al., 2006; Laurens et al., 2012).

Previous Applications of REPSA

REPSA has been successfully applied in probing for ligand binding sites for small molecule drugs, DNA and RNA triplexes, polyamide hairpins, and proteins since it was first put forward by Hardenbol and Van Dyke as a viable combinatorial selection methodology in 1996 (Hardenbol & Van Dyke, 1996). Previous applications of REPSA to elucidate preferred ligand binding sites on duplex DNA are described in following subsections, organized by the class of DNA-binding ligand investigated.

Nucleic Acids. The first reported use of REPSA was in the probing for triplex nucleic acid species in the first generation selection template (ST1-R19). BsgI was the selecting IISRE used. ODN1, a potential triplex-forming oligomer was selected on ST1 until a cleavage resistant population arose from the initial random pool after 11 rounds of REPSA. ODN1 was found to have a preference for a subset of the 19-mer target 5′-

$AG_3AG_4AG_4AG_3A$-3′, which was confirmed by DNase I footprinting (Hardenbol & Van Dyke, 1996).

More recently, REPSA was applied to the study of modified triplex forming oligonucleotides, specifically a bis-aminoU containing 9-mer (Cardew et al, 2012). REPSA application yielded A_6 tracts as a probable binding site. Reaction conditions were acidic, pH 5.0 for triplex formation and pH 5.6 for FokI selection, to allow for cytosine protonation; potentially improving putative binding interactions. Validation with footprinting proved difficult as oligopurine tracts, the suspected preferred motif for BAU containing synthetic oligomers, are resistant to DNase I nicking due to minor groove rigidity (Cardew et al, 2012).

Small Molecules. The first small molecule analyzed by REPSA was the oligopyrrole antibiotic distamycin A (Hardenbol et al., 1997). REPSA application with distamycin yielded cleavage resistant species after 12 rounds. Subsequent analysis of this pool of cleavage resistant sequences indicated an enrichment of AT base pairs relative to the starting pool. Several potential binding site motifs, later validated by DNase I footprinting, were isolated from the AT rich, cleavage resistant pool. These motifs were also AT rich e.g (5′-ATAAATTAT-3′) (Hardenbol et al., 1997). This result is consistent with previous data indicating that distamycin preferred AT rich sequences (Luck et al., 1974; Zimmer & Wähnert, 1986).

Actinomycin D, a phenoxazine polypeptide anticancer antibiotic, was selected for and yielded a cleavage resistant species after 10 rounds. Analysis of the selected sequences revealed that the sequence 5′-(T/A)GC(A/T)-3′ was associated with cleavage resistance and often occurred in nested multiples of this sequence e.g. 5′- TGCTGCTGC-

3′ (Shen et al., 2001). This sequence is consistent with previous data concerning Actinomycin GC intercalation and flanking sequence preferences (Sobell, 1985).

In a later study, minor groove-interacting hairpin polyamides, ImPyPyPy-γ-PyPyPyPy-β-Dp and ImPyPyPy-γ-ImPyPyPy-β-Dp were investigated by REPSA and an ApT/TpA degenerate consensus sequence was obtained (Gopal & Van Dyke, 2003). This sequence, despite being degenerate, was the only sequence isolated indicating that the sequence likely was a binding site for the polyamide hairpin. DNase I footprinting confirmed this result in keeping with Dervan polyamide selection rules, which predicted specific degenerate interactions for the polyamides tested (Pitch et al., 1996).

Tallimustine, a distamycin-derivative anticancer agent with covalent binding activity to DNA, had its consensus sequence determined by REPSA and previously identified consensus sequences were recovered (5′-TTTTTTTC-3′ and 5′-AAATTTC-3′) (Sunavala-Dossabhoy & Van Dyke, 2005). In addition a third, novel sequence 5′-TAGAAC-3′ was identified. N7 alkylation of guanines specifically flanking the binding sites was noted and not previously observed with tallimustine. These were hypothesized to be a cooperative effect occurring at high tallimustine concentrations. N3 alkylation of adenine had been previously observed and seems to be the predominant alkylation pattern observed in literature (Broggini et al., 1995). N7 alkylation of guanine by tallimustine has not been confirmed outside of this report. This data is consistent with tallismustine consensus sequences derived from other methods (Herzig et al., 1999).

Proteins. The first protein to be directly assayed with REPSA was the human TATA binding protein (hTBP), a subunit of the general human transcription factor TFIID (Hardenbol & Van Dyke, 1997). TBP is critical to TFIID recognition of TATA boxes and

TFIID is a key protein in the nucleation and regulation of RNA polymerase complexes (Bieniossek et al., 2013). Human TBP had previously had its consensus sequence derived by crystallographic and by DNase I footprinting so it served as a test case for REPSA application with proteins (Juo et al., 1996; Nikolov et al., 1996). It was assayed against ST2-R14 and the IISRE FokI. Of the 57 sequences isolated, 47 contained a simple TATA box. Expansion of search parameters yielded variations of the basic TATA motif such as 5′-TATAAAATA-3′, 5′-TATAAATA-3′, 5′-TATAATA-3′ and 5′-TATATA-3′. These sequences were validated by REPA (restriction endonuclease protection assay) and by transcriptional runoff assays (Hardenbol & Van Dyke, 1997).

In addition to the ODN1 consensus sequence determined (discussed above), two other consensus sequences were determined within the same sample pool, hinting at the sensitivity of REPSA. BsgI, the type IIS restriction endonuclease utilized during the assay of triplex oligonucleotide ODN1, unexpectedly selected for a higher affinity binding site for itself, closely matching the consensus sequence that had been previously determined (Hardenbol & Van Dyke, 1996, Szybalski et al., 1991). The presence of a *Bacillus sphaericus* (or *Lysinibacillus sphaericus*/ATCC 14577/UniProt Taxon ID 1421) contaminant DNA binding protein in the BsgI preparation was also revealed in this study and appeared to have a consensus sequence of 5′-TGGGA($N_{7/8}$)GTCCCA-3′. Presently, no known DNA-binding proteins from *L. sphaericus* has been identified that possesses this consensus sequence nor has a transcriptional regulator been identified for ORFs flanked by this consensus sequence in its promoter region.

The most recent protein to be assayed by this method was SlmA, a nucleoid occlusion factor in *E. coli* (Bernhardt & Boer, 2005). REPSA was applied as other

methods of determining its recognition site failed to yield binding data (Tonthat et al., 2011). Application of REPSA produced a consensus sequence: 5′-GTGAGTACTCAC-3′ which was validated with a series of mutant DNAs and fluorescence polarization experiments. The *E. coli* genome was analyzed via ChIP-Seq, to deduce the biological frequency and genomic locations of SlmA binding sites (Tonhat et al., 2011). These data reinforced the indications that SlmA was not a transcriptional regulator but rather was involved in a process downstream of DNA replication (Bernhardt & Boer, 2005; Cho et al., 2011).

***T. thermophilus*: a Model Extreme Thermophilic Organism**

Model extremophile species have been put forth to focus efforts and to gain greater predicting power in determining the range of conditions conducive to the formation of life. *Thermus thermophilus*, strain HB8, is one such species that is being considered for this role within the extremophile and early Earth community. The ease of use of *T. thermophilus* in a laboratory setting as well as the general stability of its proteome, short generational separation from wild-type strains, persistent natural competence, and its broad geographic distribution have made it an ideal candidate species (Cava et al., 2009).

T. thermophilus is a marine, non-motile, non-sporulating, yellow pigmented, polyploid, facultatively aerobic, Gram negative, heterotrophic obligate thermophile organism initially discovered in the hot springs of Izu Prefecture in Mine, Japan (Oshima & Imahori et al., 1974; Henne et al., 2004; Ohtani et al., 2010). Various strains have since been found across the planet in geothermally active, marine biomes (Stan-Lotter & Fendihan, 2012).

T. thermophilus is a polyextremeophile. It is acidotolerant, surviving down to pH 4, alkaliolerant, surviving up to pH 9.5, and halotolerant, demonstrating growth in up to 5% (w/v) NaCl by way of heavy osmolyte production, similar to other halotolerant microbes (Nunes et al., 1995). A pH of 7.0-7.5 seems to be optimal for growth. It undergoes growth in the temperature range 45-80 °C with optimal growth at 70 °C (Cava et al., 2009). Below 45 °C it appears to enter into a state similar to hibernation not to be confused with anhyrdobiosis as, like other members of *Thermus*, it is acutely sensitive to desiccation (Omelchenko et al., 2005). Above 80 °C, its growth is severely retarded and above 85 °C death occurs (Cava et al., 2009, Stan-Lotter & Fendrihan, 2012).

T. thermophilus strain HB8 has a genome size of 3.01 Mb housed in four primary structures: a single chromosome (TTA) of 1.85 Mb containing 1,973 postulated ORFs (open reading frames) and three plasmids, pTT27/TTB (256.992 kb/251 ORFs), pVV8 (81.151 kb/91 ORFs), and pTT8/TTC (9.322 kb/14 ORFs). (NCBI accession numbers: NC_006461, NC_006462, NC_017767, and NC_006463 respectively) (The UniProt Consortium, 2014). Each has a G/C content of at least 69%. TTA, pTT27, and pTT8 were sequenced in 2004 by Henne et al. with 69.5%, 69.4% and 69.0% G/C content respectively. A third plasmid, pVV8, with 68% GC content, has recently been reported and analysis revealed that this plasmid is not present in the RIKEN strain, explaining its absence in the initial stored genome (Ohtani et al., 2012).

In toto, the HB8 genome appears to contain 2,324 putative genes. At present 414 have been reviewed and are in the Swiss-prot database (high quality manual curation) and 1,910 not reviewed in TrEMBL (computational annotation or not yet reviewed by a curator) (Uniprot Consortium, 2015). Approximately 756 of these putative proteins are

uncharacterized within Uniprot, having only minimal levels of annotation regarding their putative function within *T. thermophilus*. These 756 uncharacterized proteins were identified by searching UniprotKB for "taxonomy: "*Thermus thermophilus* (strain HB8 / ATCC 27634 / DSM 579) [300852]" uncharacterized".

At present, there are approximately 70 potential transcription factors or regulators, annotated in the *T. thermophilus* HB8 genome stored in UniprotKB (The Uniprot Consortium, 2015). 84 ORFs were initially identified to be involved in transcription in some fashion (20 Swiss-prot and 64 TreEMBL), either putatively or experimentally. They were found by searching Uniprot for "taxonomy: *Thermus thermophilus* (strain HB8 / ATCC 27634 / DSM 579) [300852]" transcription". 14 of these proteins are not transcription factors, but are rather involved in the transcription/translation process in some other manner. Putative or described transcription factors are listed in Table A2 in Appendix A.

Despite its small genome, 82% or 1910 ORFs are unreviewed, lacking manual curation by Swiss-Prot. Approximately 18%, 414, of the total determined ORFS have undergone manual curation, indicating that the data concerning those proteins or RNAs coded by those ORFs is generally consistent and reinforcing (Magrane & The Uniprot Consortium, 2011; Poux et al., 2014). Of the total number of ORFs, 33%, 762, are uncharacterized, indicating they lack functional assignment. Of these, 761 are unreviewed and one is reviewed but has no known function.

SbtR: A High Temperature REPSA Proof-of-Concept Test Subject

In order to effectively optimize REPSA protocols for assaying high temperature proteins, a model type protein was needed that had been well characterized via other methodologies. SbtR, (intermolecular di**S**ulfide **B**ridge-containing **T**etR family **R**egulator /TTHA0167/NCBI accession number YP_143433.1) had previously had its 14 base pair palindromic binding sequence, 5′-TGACCCNNKGGTCA-3′, ascertained via genomic SELEX at 55 °C, within the preferred temperature range of *T. thermophilus* HB8 and validated by SPS (Surface Plasmon Resonance) (Agari et al., 2013). It is a homo-dimeric protein, consistent with both its palindromic recognition site, a hallmark of homo-dimeric DNA binding proteins, as well as findings regarding other TetR proteins (Cuthbertson & Nodwell, 2013).

TetR type proteins are one of the more characterized transcription factor families (Cuthbertson & Nodwell, 2013). All currently identified TetR type proteins have been shown function as homo-dimeric repressors in their native state, requiring no additional modifications for adherence to DNA. They are highly similar across eubacteria and archaea, consisting of 9-10 α helices, with the C-terminus being near the dimerization domain and a helix turn helix (HTH) "foot" that houses the DNA recognition domain being N-terminal (Cuthbertson & Nodwell, 2013). The HTH recognition domain is generally conserved across species with the dimerization and small molecule interaction domains being highly variable.

The consensus DNA-binding sequences of TetR-family proteins tend to be large, 10-30 base pairs (bp), and are usually palindromic. SbtR's preliminary palindromic binding site is 14 bp long. These proteins intrinsically inhibit transcription initiation and their repressive activity can be removed by interaction of some cognate small molecule that usually binds the protein near the dimerization domain (Cuthbertson & Nodwell, 2013). It is thought that this ligand induces a conformation change that shifts the dimerization and/or recognition domains into a geometry unfavorable for binding, increasing the likelihood of DNA dissociation (Cuthbertson & Nodwell, 2013). SbtR's cognate molecule is currently unknown though a putative binding pocket has been identified near the dimerization domain of the protein.

At present, the function of SbtR's disulfide bridge is not known, it has been postulated that it might increase thermostability or act as a cognate ligand gate due to is location at the "mouth" of the ligand site or both, (Agari et al., 2013). Putative ORFs controlled by SbtR (TTHA0027, TTHA0785, TTHA0786, TTHA0787, TTHA1818, TTHA1819, TTHA1820, TTHA1821, TTHA1822, and TTHA1823) have been shown *in vitro* to be repressed by increasing concentrations of this protein (Agari et al., 2013).

Based on analysis of the genes it has been shown to repress, including Tth-RecA and CinA, it appears that SbtR may be involved in the bacterial SOS response, a DNA-damage response involving increased presence of single-stranded DNA (Kato & Kuramitsu, 1993). A LexA homologue was recently discovered, housed in pVV8, in *T. thermophilus* HB8 (Ohtani et al., 2012). Assessment of this genes function in other members of the Deinococcus-Thermus family, *Deinococcus radiodurans*, hint that LexA may not be necessary for SOS response in this taxon, though further study is required

(Narumi et al., 2001). This suggests that SbtR may be a part of a SOS response that differs from the canonical RecA-LexA variant.

SbtR had previously had its denaturation temperature assayed as 98.5 °C in its native dimeric form by Differential Scanning Calorimetry (DSC) (Agari et al., 2013).Several factors seem to result in this unusually high denaturation temperature. Consistent with the general trend observed in thermostable proteins, SbtR appears to rely on hydrophobic moieties to maintain its high temp stability (Agari et al., 2011; Sakamoto et al., 2011; Agari et al., 2012). The aforementioned interprotein disulfide bridge, at Cys164, also appears to play a significant role in maintaining SbtR's thermostability. Cys164, located on the face of the dimerization domain, appears to form an interprotein disulfide bridge with a partner Cys164 on an adjacent SbtR monomer SbtR C164A mutants displayed a significantly lower denaturation temperature of 90.4 °C though this did not seem to interfere with its repressive ability (Agari et al., 2013)

Findings, Aims and Objectives

At present, there are a number of methods available to ascertain a probable consensus DNA sequence for a given ligand. There is a consistent drive to move towards high throughput methodologies, either by modification of existing methods, or by *de novo* creation of new methods. HT-SELEX and PBM have proven incredibly useful for high throughput assessment of known transcription factors and seem consistent with each other (Orenstein & Shamir, 2014). However, all of the technologies presented here rely on some physical modification of the transcription factor(s) in question to pursue identification of its consensus sequence (Stormo & Zhao, 2010; Gade & Kalvakolanu,

2012; Dey et al., 2012). This entails prerequisite knowledge of a transcription factor's physical properties, which is not amenable to discovery of heretofore unknown or less well characterized transcription factors, the subject of our studies. These methods focus on isolation of DNA-transcription factors complexes, relying on differing physical properties between TF-bound compared to TF-unbound DNA. Physical separation methods have to be selective for the bound complex as well as maintaining the integrity of the complex through the isolation procedure, hampering assessment for transcription factors with weak or moderate affinities (Collas, 2010) Application of these methods for discovery of new transcription factors remains elusive.

While other methodologies' selection and isolation methods implicitly rely on the physical separation of TF-DNA complexes, which can be challenging depending on the physical properties of the TF-DNA complex, REPSA does not (Van Dyke et al., 2007). Instead, REPSA selection depends on the preferential amplification of protected, and thus intact, templates during routine PCR compared to cleaved templates (Van Dyke et al., 2007). The lack of physical separation methods for REPSA allow for the use of routine, kit-based DNA purification methods.

REPSA is able to generate biologically relevant consensus sequences that are consistent with other methodologies for a variety of ligand types. REPSA is sufficiently sensitive to resolve small molecule DNA binding sites as well as complex solutions, generating multiple consensus sequences for DNA-binding ligands present in such solutions (Hardenbol & Van Dyke, 1996; Hardenbol & Van Dyke 1997; Tonhat et al., 2011). For example, the first application of REPSA resolved not only a consensus sequence for ODN1 but also two additional sequences, the recognition site of the IISRE

used, BsgI, as well as an unknown contaminant DNA binding protein present in solution with supposedly pure BsgI solution (Hardenbol & Van Dyke, 1996).

Despite the overall strengths of REPSA, it has limitations. It is unable to assay with biological validity TFs of organisms outside the active range of FokI, BsgI, and BpmI (25-40 °C), due primarily to IISRE temperature limitations. This temperature limitation unreasonably isolates REPSA application to mesophilic transcription factors and subsequently excludes large portions of the biosphere from analysis, particularly psychrophiles (cold-loving organisms) and thermophiles (heat-loving organisms). These extremophilic organisms have proven to be immensely important for biotechnological innovation and development, and have been an invaluable source of an array of enzymatic products (Stan-Lotter & Fendrihan, 2012; Seckbach et al., 2010). However, psychrophilic IISREs have yet to be made commercially available, whereas there are several thermophilic IISREs that are (Roberts et al., 2014). Thus REPSA could be adapted for study of thermophilic transcription factors, potentially yielding insight into how thermophiles regulate their metabolic networks.

Thermus thermophilus HB8, in line with its assignment as a model organism, is an ideal candidate species for future applications of REPSA. Approximately 33% of its total genome is of unknown function, presenting large gaps in the proteome. It is amenable to lab culturing, is naturally competent, its proteins are highly stable under a variety of conditions, and its genome is relatively small. Such proteins could prove industrially useful if their putative functions can be ascertained, if indirectly, by way of REPSA application.

SbtR is a relatively well defined, stable protein with preliminary data supporting a transcriptional repressor function, in keeping with other TetR members. Its consensus sequence and binding kinetics have been determined. Its high native melting temperature, 98.5 °C, allows for heat purification of the recombinant protein when expressed in a mesophilic organism (e.g., *E. coli*). It is a natively repressing protein, requiring no cognate molecules to function and is thus an ideal proof-of-concept candidate protein for high-temperature REPSA method development.

The primary aim of this work was to ascertain if REPSA could be successfully modified for use at high temperatures to obtain a *T. thermophilus* transcription factor consensus sequence followed by bioinformatics assessment for validity. Chapter 3 presents data with regard to the primary aim. Appendix A lists supplemental data. Appendix B lists the materials and methods utilized for the experiments described in Chapter 3. The current study assesses the validity of the primary aim.

CHAPTER 3: RESULTS I

DETERMINATION OF SELECTION TEMPLATE 4 (ST4) THERMOSTABILITY

Introduction

In order to effectively probe for *T. thermophilus* HB8 TFs and obtain biologically valid consensus sequences, these high temperature transcription factors require testing under thermophilic physiological conditions. Mesophilic IISREs, FokI, BsgI and BpmI, which had been previously utilized for REPSA experimentation, would be denatured at the relevant physiological temperature range 50-85°C for *T. thermophilus* HB8 (Szybalski et al., 1991; Van Dyke et al., 2007; Cava et al., 2009). Previous selection templates (ST 1-3) were designed for exclusive use with these mesophilic IISREs. Thus it was essential to develop reaction conditions and design new selection templates suitable for high-temperature REPSA investigations with high temperature IISREs.

High temperature IISREs. A new template, ST4, was designed to accommodate our need for thermophilic IISREs and is described in detail in the sections below. The IISREs met the prerequisites of being active in the optimum temperature range of *T. thermophilus*, 65-72 °C, in addition to having their cleavage site be a minimum of 8 bp from their recognition site, thereby allowing them to effectively probe into the central region of the ST4-R20 template.

IISRE BseXI (Thermo Fisher/ER1451/Lot: 0019126), isolated from *Bacillus stearothermophilus* Ra 3-212), has the shortest reach, cleaving 12 bp 3′ of its recognition

site, 5′-GCAGC-3′. It has an optimum temperature of 65 °C for DNA cleavage. BsmFI (NEB/R0572S/Lot: 0241310), isolated from *Bacillus stearothermophilus* F (ER2683), cleaves 14 bp from its recognition site 5′-GGGAC-3′, and has an optimum temperature of 65 °C. BtgZI, (NEB/R0703S/Lot: 0051311) isolated from *Bacillus thermoglucosidasius*, cleaves 14 bp from its recognition site 5′-GCGATG-3′ and has the lowest optimum temperature at 60 °C. BtgZI, although it did not meet the requirements of being active within the optimum temperature range of *T. thermophilus* HB8, was investigated. It was the next most thermostable IISRE among all other commercially available high temperature IISREs at the time of template design with the longest cleaving head reach.

Selection template design. A new template was designed, ST4, that contained high temperature IISRE recognition sites for BseXI, BtgZI, and BsmFI. These were located within either 20-base pair defined flanks, immediately adjoining the random core region, and positioned such that their cleavage domains cleave within the random core (Figure 2). The defined flanks serve two functions: acting as IISRE recognition and as primer annealing sites, thereby allowing for controlled IISRE cleavage as well as reliable PCR amplification. A FokI recognition site was also included bordering the BsmFI recognition site. This provided us with the potential to compare the transcription factor binding and cleavage protection under both thermophilic and mesophilic conditions should it become necessary (*e.g.*, in the presence of transcription factor independent cleavage resistance).

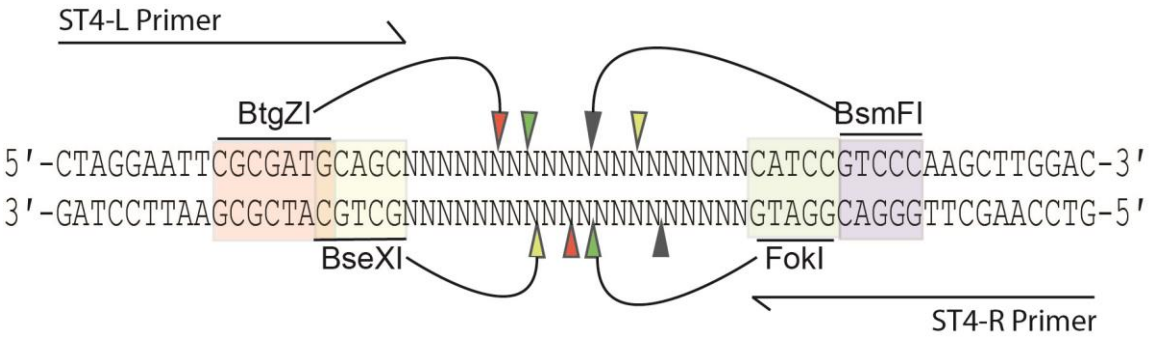

Figure 2: ST4-R20 design for HT-REPSA selections. Represented is the core ST4-R20 template, with 20 bp defined flanking regions containing the recognition sites [colored boxes] for four IISREs (BtgZI, BsmFI, BseXI, and FokI), and a 20 bp central random region with arrows [red = BtgZI, yellow = BseXI; blue = BsmFI, green = FokI] indicating the cleavage positions for each IISRE. N represents random nucleotide.

The ST4-R20 variant possesses a central randomized region of 20 base pairs when in duplex DNA form. Per the equation, $(4^n)/2$, where n = the length of the random region and two accounts for the degeneracy of dsDNA, yields approximately 550 billion ($\sim 5.5 \times 10^{11}$) different sequence combinations for this selection template. A second variant, ST4-SbtR, was designed to be a specific control for SbtR binding. It contained the SbtR consensus sequence 5′-TGACCCNNKGGTCA-3′ in its core region (Agari et al., 2013). We chose to use the specific sequence 5′-TGACCCTAGGGTCA-3′ in our ST4-SbtR control template, to eliminate degeneracy and negate potential issues that may arise in a heterogeneous template pool, *e.g.* improper annealing.

To be practical for HT-REPSA, the ST4-R20 template requires its minimum melting temperature to be above 70 °C, the optimum temperature for *T. thermophilus* HB8. Initial optimization testing high temperature IISREs BtgZI, BsmFI, and BseXI with a standard length ST4-R20-S yielded unexpected cleavage resistance following a mock SbtR incubation step, 70 °C for 10 minutes with SbtR vehicle buffer. This is thought to

occur due to the formation of "bubbles", regions of single-stranded DNA within the randomized region, when two imperfectly complementary DNA strands anneal (Figure 3). REPSA, to accurately assess binding motifs for TFs, requires an equilibration between the template and the TF prior to cleavage selection by the IISRE.

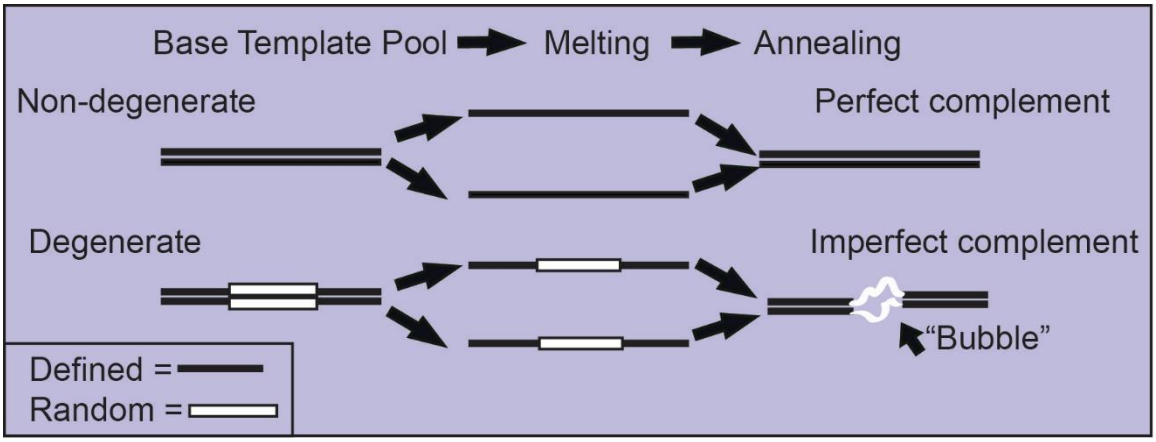

Figure 3: "Bubble" formation upon melting and annealing of ST4-Random template (degenerate) vs. ST4-Specific template (non-degenerate). Non-degenerate (transcription factor-specific selection template, ST4-SbtR) has only one sequence combination. Thus all ssDNA strands have a perfect complement after melting and annealing. The degenerate random selection template (ST4-R20) has many potential combinations, resulting in imperfect annealing. These can create ssDNA "bubbles" post-melting in the central random region, which are refractory to IISRE cleavage.

For HT-REPSA to accurately select for biologically relevant DNA motifs, the TF needs to be incubated at the physiologically relevant conditions for the organism whose TF is under study. For *T. thermophilu*s HB8, this range is 50-85 °C with 65-72 °C being the optimum temperature range (Cava et al., 2009). Additionally, 65 °C is the optimal operating temperature of two of the three IISREs utilized for HT-REPSA, BseXI and BsmFI. In order to produce experimentally valid data, the template should be minimally thermostable at the highest IISRE probing temperature. Thus modifications to our

standard ST2-R20 were necessary to maintain its integrity under optimal high temperature conditions

Results

Early optimization attempts with IISREs and ST4-R20-Short resulted in unexpected cleavage resistance. The IISREs were being incubated with the ST4 oligomer pool following a heating step intended to replicate SbtR incubation. It was hypothesized that this cleavage resistance was due to the random template was partially melting during the SbtR equilibration step, 70 °C for 10 minutes, forming cleavage resistant "bubble" or "looped" species. This improper annealing of the 20 bp random core, previously termed as "looping" by Hardenbol and Van Dyke in 1996, results in mismatched and single stranded "bubbles" that are refractory to IISRE cleavage as IISREs have no demonstrated ability to cleave ssDNA. (Figure 3). As the ST4-R20 template has 550 billion potential combinations it is highly prone to improper annealing following melting; the probability of a strand finding its perfect complement are essentially nil. Thus to minimize bubble formation, it is necessary to minimize template denaturation under our standard high-temperature reaction conditions.

Type IISREs have not yet been demonstrated to have effective ssDNA cleavage capability, so these single stranded regions are generally resistant to cleavage under our reaction conditions (Szybalski et al., 1991). In addition, if the template is melted into ssDNA strands during incubation with the IISRE or with the TF, then its selection and survival each round will not be dependent on either. These sequences would be selected for based on their reduced thermal stability alone. As TF-dependent cleavage resistance is

the sought outcome during REPSA, it is not optimal nor is it experimentally valid for cleavage resistance to result from ssDNA bubbles or from melted templates.

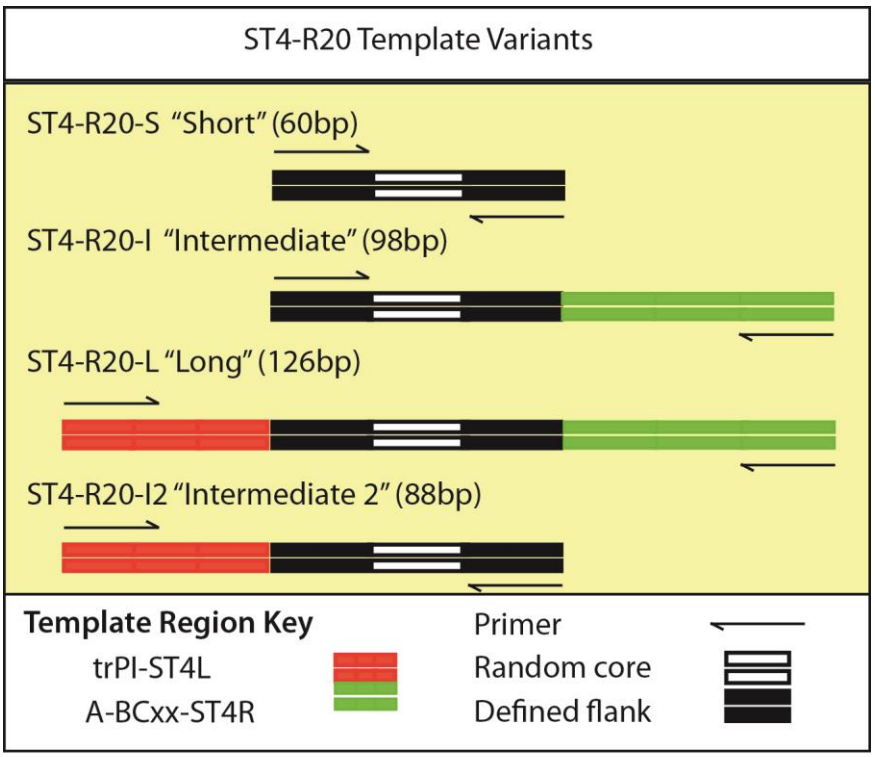

Figure 4: ST4 template length variants manufactured. trPI-ST4L and A-Bxx-ST4R (xx = 11-14) are extenders for ST4 use with the Ion Torrent PGM. The defined flanks house the IISRE recognition sites oriented so that they cleave in the 3′ direction, within the random core.

To test for unwanted melting and subsequent bubble formation, the thermostability of the ST4-R20 and ST4-SbtR templates were modified by increasing the overall length of the templates, thereby increasing their overall melting temperatures. To this end, additional ST4 variants for both ST4-SbtR and ST4-R20 were manufactured by appending extender regions, Ion Torrent PGM (trP1-ST4L) marker to the left side of the 60 bp core (ST4-R20 or ST4-SbtR) template and a barcode marker (A-BCxx-ST4R) to the right side, in three possible additional combinations (Figure 4). These extensions are

significantly longer than general PCR primer design rules allow for (18-22 bases), with A-BCxx-ST4R (xx = 11- 14) being 63 bases in length and Trp1-ST4L being 44 bases in length. PCR cycling for these templates was limited to six cycles to both prevent formation of "bubble" or "looped" species, in keeping with established REPSA methodology, as well as to limit unwanted product formation, e.g. primer dimers, that are more likely to result with such large primers (Van Dyke et al., 2007).

Figure 5 demonstrates the cleavage resistance expected from generation of bubble species as a result of undesirable melting of the shorter, standard ST4-R20 template. As expected, the non-degenerate ST4-SbtR was cleaved by BtgZI with high efficiency compared to a ST4-SbtR control regardless of ST4-SbtR's heat treatment prior to BtgZI application. The ST4-R20 65 °C I and L DNAs are comparable in cleavage efficiency to the ST4-SbtR 65 °C group when compared to the ST4-R20 negative control ST4-R20-S 65 °C appears to be less efficiently cleaved when compared to ST4-SbtR-S 65 °C, likely due to bubble formation. ST4-R20 70°C group display a similar cleavage resistance pattern to the ST4-R20 65°C group, with both the ST4-R20-I (intermediate) and ST4-R20-L (long) variations (Lanes 3 & 4) displaying comparable cleavage resistance to ST4-SbtR 65 °C. As with the ST4-R20 65 °C group, the ST4-R20-S variant also exhibits increased cleavage resistance compared to the ST4-SbtR 70 °C in addition to greatly increases cleavage resistance compared to ST4-R20-S 65 °C. Taken together, these data reveal that the ST4-R20-S partially melts at the reaction temperatures required for application of HT-REPSA, with greater melting occurring at 70°C incubation as compared to 65 °C incubation. To reduce the likelihood of template melting, the longest

template ST4-R20-L (126 bp) was determined to be the most suitable for application with HT-REPSA, as it produced the least denaturation-related cleavage resistance.

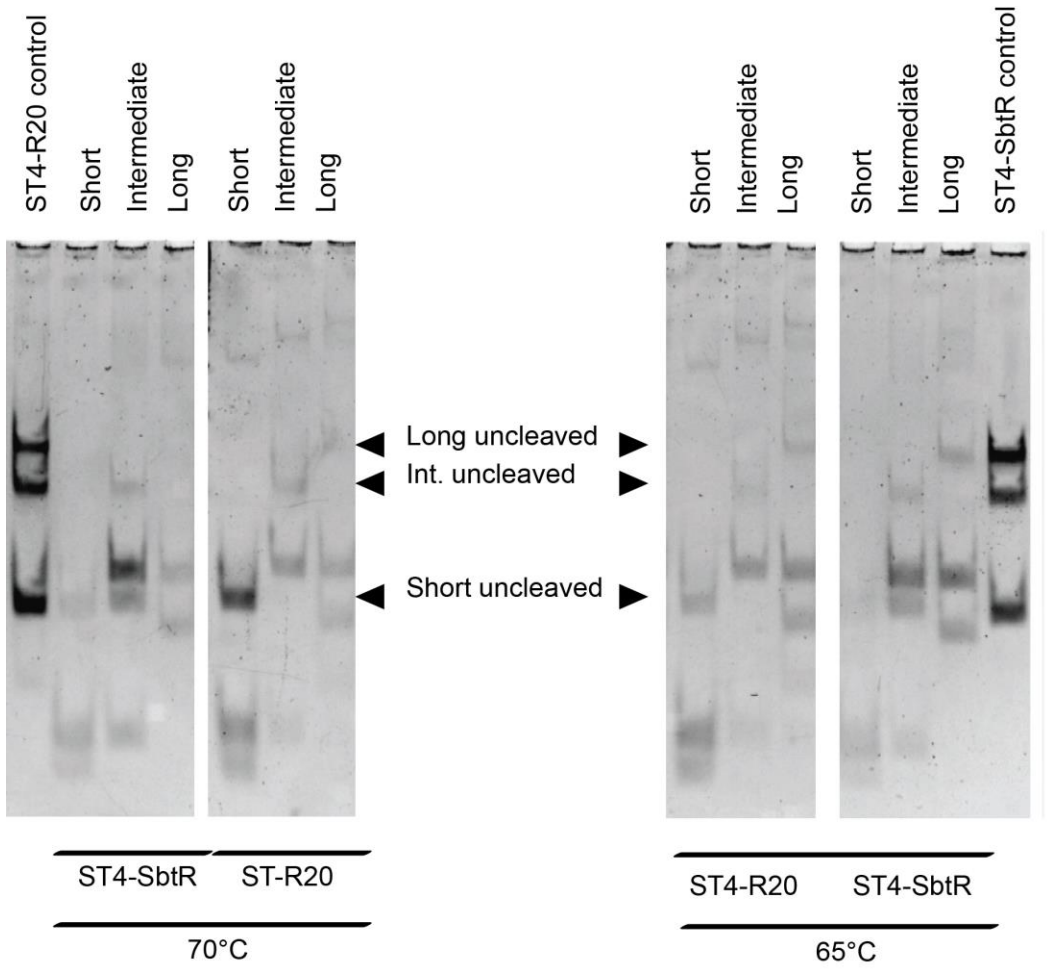

Figure 5: Bubble dependent cleavage resistance as a result of undesirable melting of standard-length ST4-R20 template under high temperature reaction conditions. Both SbtR specific (ST4-SbtR) as well as the random 20 (ST4-R20), Short(60 bp), Intermediate=Int (98 bp), and Long (126 bp) DNAs were incubated at 65 °C or 70 °C for 10 minutes, cooled to 60 °C to encourage annealing, and subsequently incubated with BtgZI (1U) for 6 minutes at 60 °C to probe for cleavage resistant bubbles.

CHAPTER 3: RESULTS II

SBTR PRODUCTION AND ACTIVITY ASSESSMENT

Introduction

Thermophilic transcription factor production was among the first steps in the process of adapting REPSA to high temperature conditions. A TetR type protein is ideal for the optimization as they are highly studied, natively homodimeric repressors that intrinsically bind specific palindromic duplex DNA sequences and are usually deactivated by small molecules that bind near their dimerization domain (Cuthbertson et al., 2013).

Among the 70 putative transcription factors in *T. thermophilus* HB8, at least four are TetR type transcription factors: FadR, PaaR, PfmR, and SbtR. (Agari et al., 2011; Sakamoto et al., 2011; Agari et al., 2012; Agari et al., 2013). SbtR was chosen from this group as both a His-tagged and native protein variant were gifted by the RIKEN Institute, whereas only native variants were gifted for the other three *T. thermophilus* HB8 TetR proteins. The His-tagged form should have allowed for easier, column based-purification in case highly purified protein were required for subsequent assays. However, we investigated the native form, which allows for assessment of SbtR in its unmodified state.

Results

SbtR production. Plasmid pET-SbtR (pET21a), with SbtR under the control of a T7 promotor, was introduced into competent BL21(DE3) *E. coli* cells. Production of SbtR was

driven by IPTG (Figure 6, IPTG+). A control group was also grown under the same conditions and timeframe without IPTG (Figure 6, IPTG–). Proteins consistent in mass for both the monomer, ~22 kDa (Figure 6, lower arrow), and the dimeric form, ~44kD (Figure 6, upper arrow), of SbtR are strongly present in the IPTG-induced group as compared to the uninduced group (Figure 6).

SbtR had previously been found to resist dissociation into its component monomers during SDS-PAGE under reducing conditions (50 mM DTT), likely due to the presence of an intermolecular disulfide bond at the dimerization interface (Agari et al., 2013). They found that increasing the concentration of DTT (dithiothreitol) reduced the disulfide bridge, driving SbtR to its monomeric form during SDS-PAGE (Agari et al. 2013). A DTT concentration of 50mM was utilized during SDS-PAGE to observe the formation of both the monomeric and dimeric bands as this concentration seemed to result in the strong presence of both bands (Agari et al., 2013). However, the monomeric species seems to be favored here, with it displaying a far stronger band than the dimeric species.

Given that the melting point of SbtR in its covalently dimeric form is 98.5°C, it should be possible to denature most *E. coli* proteins following heat treatment of soluble bacterial extracts at 80 °C for 20 minutes (Agari et al. 2013). This permits the purification of SbtR from heat-denatured *E. coli* proteins following the latter's aggregation and separation by centrifugation (Agari et al. 2013).

As shown in Figure 6, heat treatment (Heated Lane/IPTG +) appears to denature the bulk of *E. coli* proteins present in whole cell extracts. The heat-purified SbtR fraction likely contains other proteins, though they are likely to be a small fraction of the total protein

load. Thus the preparation containing SbtR is a complex protein mixture, primarily composed of SbtR, of unknown activity, and a small fraction of various *E. coli* proteins.

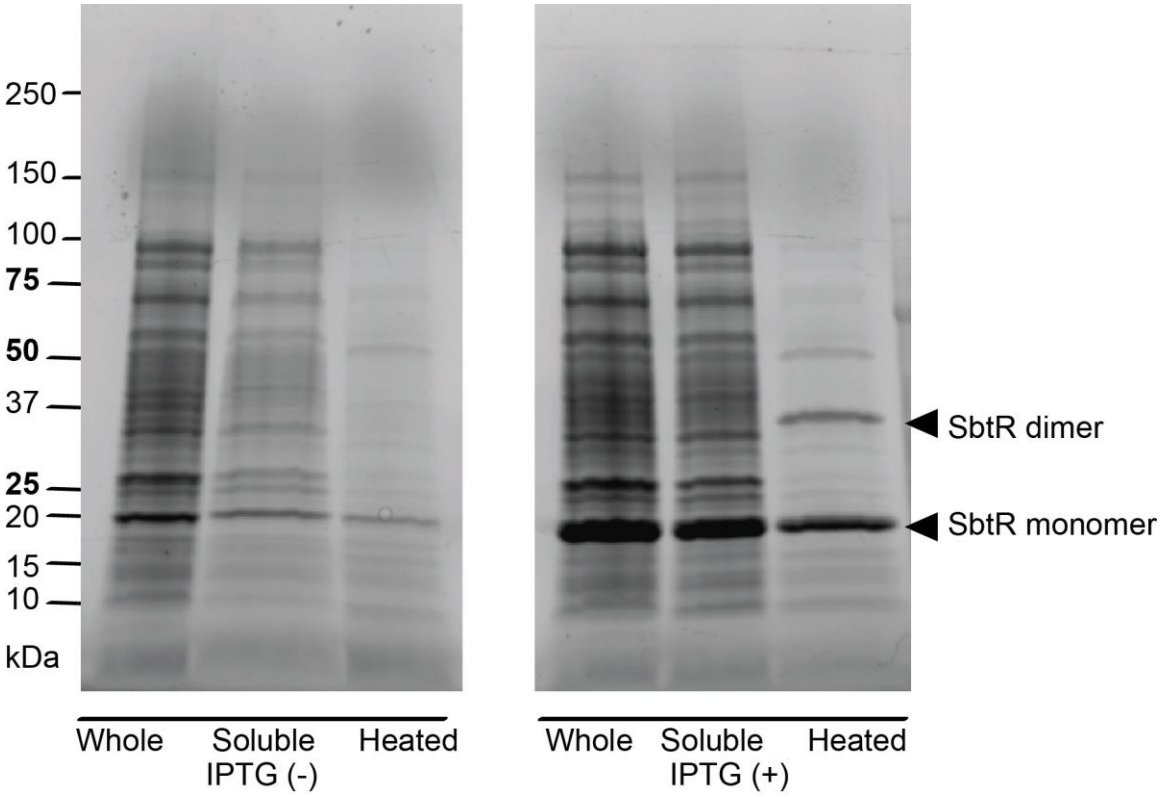

Figure 6: Production of thermophilic TetR transcription factor SbtR by pET-SbtR transformed *E. coli* BL21(DE3). Shown is a Coomassie Brilliant Blue R-250-stained SDS-PAGE 4-12%. Above are indicated fractions from IPTG-induced (+) and uninduced (-) bacteria. Whole = Whole cell fraction. Soluble = Soluble fraction. Heated = Heat-treated (80°C for 20 min.) soluble fraction. Sample buffer contained 50mM DTT.

SbtR is a complex mixture. As the heat-treated SbtR fraction is a partially purified cellular lysate, it is likely that there are other cellular components present in addition to proteins. This suspension is known to be complex and contain proteins or protein fragments other than SbtR (Figure 6). However, whether additional macromolecular components are present, e.g. nucleic acids such as RNA or DNA, was unknown. Such nucleic acids could potentially interfere with REPSA analysis of SbtR-DNA binding. Thus 1% agarose gel electrophoresis and ethidium bromide staining was performed to assess the nucleic acid content of this fraction (Figure 7).

As shown in Figure 7, the nucleic acid species observed were primarily small species, approximately 300 bp in apparent length. These species are likely tRNA as tRNA molecules are small, highly stable, and quite abundant in cells. Larger fragments, e.g. denatured rRNA in the apparent kilobase range, appear as smears and are preferentially present in the IPTG-induced samples. Treatment of the heat-purified SbtR solution with RNase A at 37 °C for 30 min. versus untreated SbtR fraction confirmed that these fragments are RNA (data not shown). The bulk of the nucleic acid fragments that are present appear to occur only in the induced species indicating the fragments likely tied in some way tied to the induction process, lending further credence to the notion that these are tRNA and rRNA fragments. Thus, the SbtR preparation is not only a complex protein mixture, but a complex mixture containing nucleic acid species as well.

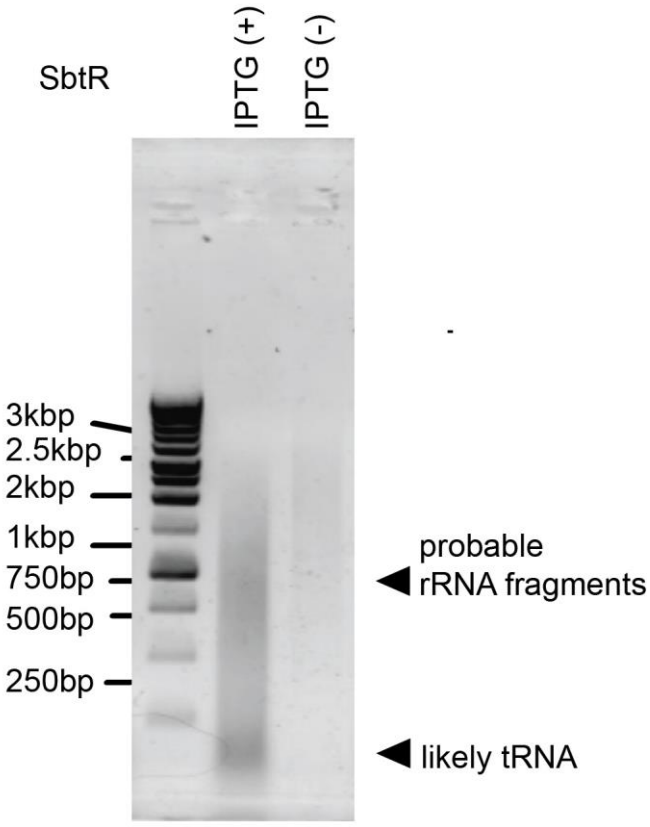

Figure 7: Determination of nucleic acid content of SbtR sample preparation. Indicated are likely fragments of various RNA species. Shown is a 1% agarose gel stained with ethidium bromide (EtBr). L\ "+" indicates IPTG induction. "-"indicates no IPTG induction.

SbtR activity and stability: The protein produced in Figure 6 was consistent with SbtR's previously established physical characteristics and behaviors (Agari et al. 2013). However, the activity of the SbtR produced was undetermined. To test this, ST4-SbtR-S and ST4-R20-S were incubated at 65 °C with a 5-fold titration of SbtR solution as part of an Electrophoretic Mobility Shift Assay (EMSA). An example EMSA (Figure 8) is included for experimental clarity. In our EMSA analysis of SbtR, the ST4-R20-S

template (Figure 9, SDS (-)) did not produce a noticeable band shift at any SbtR concentration compared to the ST4-R20-S template alone

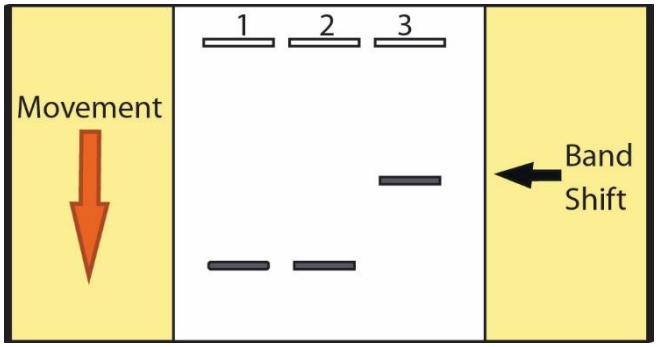

Figure 8: Example EMSA. Lane 1 is a negative control, containing the TF-binding specific template only. Lane 2 contains a TF nonspecific template and the TF. No band shift should be observed in this lane. Lane 3 contains a TF specific template and the TF. Under the right experimental conditions, the TF is bound to DNA on the gel, causing a shift in the banding pattern compared to lanes 1 and 2. The TF: DNA complex moves more slowly down the gel then the unbound template during electrophoresis, thus appearing to "shift" upwards, indicated by the arrow, on the gel when visualized.

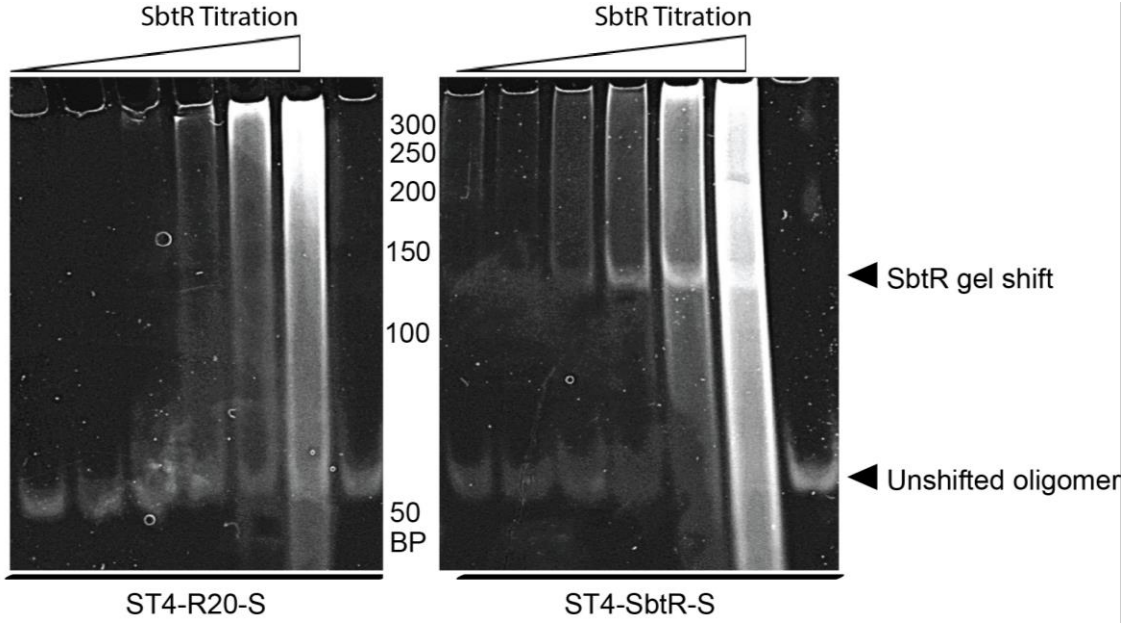

Figure 9: EMSA determination of SbtR activity. 10% PAGE, EtBr stain, positive image. SbtR was titrated against the templates, ST4-R20 (nonspecific) and ST4-SbtR (specific), in five-fold steps, starting with undiluted SbtR sample solution to SbtR 1/3,125 dilution.

(Figure 9, lane 7). The ST4-SbtR-S template (Figure 9, lanes 9-14) did produce a band shift (Figure 9) that decreased with decreased SbtR concentration compared to both the ST4-SbtR-S template alone (lane15) and the random, nonspecific template. This shift indicates that the protein is likely SbtR and is active with regards to its sequence-specific DNA binding.

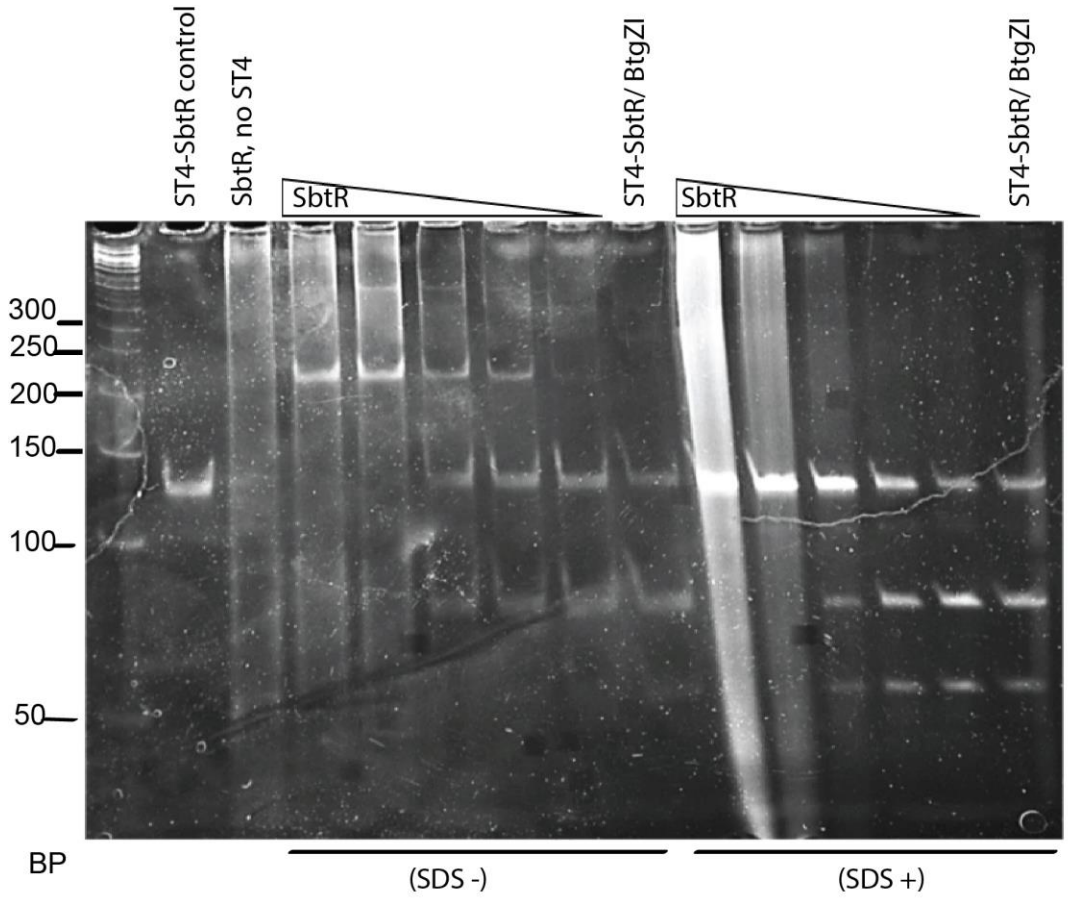

Figure 10: SbtR blocks BtgZI cleavage in a concentration dependent manner. 10% native PAGE, EtBr stained, positive image. SDS was included in lanes 10-15 to better observe template cleavage. BtgZI was incubated with templates under standard reaction conditions.

SbtR blocks IISRE cleavage. The next step in preparing for REPSA was to ensure that the DNA-binding protein reliably inhibited IISRE cleavage of the template. To assess SbtR protection of the template as well as to assess the optimal SbtR concentration to be used during REPSA, differing amounts of SbtR were incubated at 65 °C for 10 minutes with the SbtR-specific ST4 template to allow for SbtR equilibration. Afterwards, 0.25 U BtgZI IISRE was added and cleavage allowed to ensue at 60 °C for 6 minutes before the mixture was cooled to 4 °C to halt the reaction. The solution was then mixed with either NEB Loading Buffer Blue (LBB) (SDS +) (Lanes 10-15), which contains a low concentration of SDS (0.017% final) to allow for more precise quantitation of cleaved species, or a LBB lacking SDS (SDS-) (Lanes 4-9) to allow for gel shifting behaviors to be observed.

Figure 10 demonstrates that SbtR blocks cleavage of BtgZI in a concentration dependent manner, with cleavage protection decreasing with decreasing concentrations of SbtR. In addition to the expected band shift, super shifts were observed in Lanes 5 and 6, indicating that SbtR and BtgZI were likely binding to the same template strand and remained during electrophoresis.

CHAPTER 3: RESULTS III

HIGH TEMPERATURE REPSA WITH SBTR

Introduction

With the production of active SbtR and the development of a suitable high temperature selection template, it was then feasible to pursue identification of preferred SbtR-DNA binding sites by high temperature REPSA (HT-REPSA). In general, the protocol that was followed closely matched previously described REPSA protocol, with changes dictated by HT-IISREs used, choice of DNA binding protein, and use of high throughput sequencing technologies replacing subcloning (Van Dyke et al., 2007). Transcription factor dependent cleavage resistance (TFDCR) is the prime goal. However, during REPSA rounds, all three HT-IISREs utilized displayed transcription factor independent cleavage resistance (TFICR), requiring repeated interchange of each IISRE to overcome the TFICR of the previously utilized IISRE.

Results

HT-REPSA rounds were initiated by incubating 30 ng (360 femtomoles, less than 39.4% of the 550 billion potential sequence combinations) of ST4-R20-L with a 1/3,125 dilution (five serial five-fold dilutions) of SbtR stock sample (10 minutes, 65 °C) in 20 μL of a buffer appropriate for HT-IISRE cleavage. After SbtR binding, an HT-IISRE (0.5 units for BtgZI, 0.25 units for BseXI, and 0.5 units for BsmFI) was added and incubated for 6 minutes at the optimum temperature for the IISRE being used. In a typical series of

REPSA selections, IISREs were rotated once they displayed approximately 20% transcription factor independent cleavage, which usually occurs after three rounds of selection. After IISRE application, 10 μL is set aside for cleavage analysis and visualization with PAGE. A variable aliquot of the remaining 10 μL was utilized, depending upon the percentage of template cleaved, to seed a PCR reaction for generation of the next round input material.

Figure 11 demonstrates that round 1 of HT-REPSA conducted with BsmFI showed no cleavage discrimination between lane 2 (-/+), no SbtR present, and lane 3 (+/+), SbtR present. REPSA should ideally produce a discriminatory cleavage pattern between the cleavage control lane (-/+) and the REPSA selection lane (+/+) once DNAs containing transcription factor binding sequences become more abundant in the population. The template pool is thus expectedly poor in SbtR recognition sites after only one round of selection. A cleavage-resistance selection assay (CRSA), where the IISRE itself is the selecting agent, was also initiated with this round to select for sequences that may be intrinsically cleavage resistant to the IISRE BsmFI. CRSAs were conducted for each high temperature enzyme to assess the potential for this behavior.

After seven rounds of REPSA, cleavage discrimination was observed. (Figure 11). BsmFI (Rounds 1-3) and BseXI (Rounds 4-6) have appeared to select for intrinsically cleavage resistance sequences. BsmFI passed the threshold cleavage resistance after Round 3 and BseXI was subsequently utilized for rounds 4-6. However, BseXI also passed the threshold cleavage resistance during Round 6 REPSA selection. No evidence for SbtR-dependent cleavage resistance was observed with BsmFI and BseXI in Rounds 1-6. To progress further in the REPSA rounds, BtgZI was utilized for Round 7. BsmFI

resistance remained even after three rounds of selection with BseXI. The material that was natively resistant to BsmFI and BseXI was not resistant for BtgZI in this round. This allowed for SbtR's contribution to cleavage resistance to become evident and SbtR dependent cleavage discrimination was observed in this round (Figure 11; Round 7 lane 2 v lane 3).

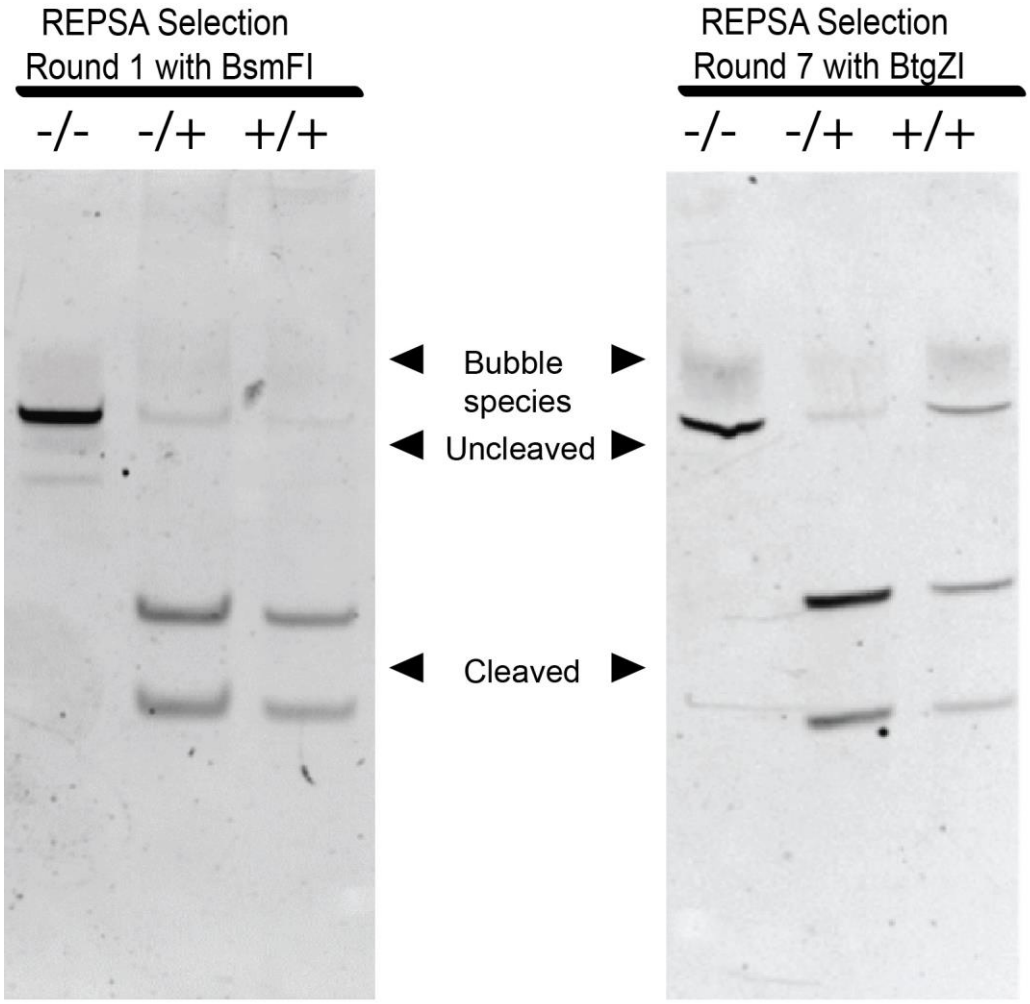

Figure 11: SbtR dependent cleavage resistance is evident by Round 7. Round 1 REPSA with SbtR demonstrates no discriminatory cleavage and Round 7 REPSA with SbtR shows discriminatory cleavage. Native PAGE, 10% gel, SDS-containing loading buffer, EtBr staining, negative image. -/- = Template only control. -/+ = Template with the designated IISRE. +/+ = Template w/BsmFI and SbtR. Improperly annealed bubble species are observed above the uncleaved ST4-R20-L bands.

CHAPTER 3: RESULTS IV

SEQUENCE ANALYSIS AND DETERMINATION OF OLIGOMER IDENTITY OF NON-RANDOM ROUND 7 SEQUENCE POOL

Sequencing. HT-REPSA selections that displayed cleavage resistance had aliquots taken post selection and were PCR amplified for 30 cycles to generate sufficient material for both dideoxy big-dye and Ion Torrent sequencing. The remainder of the selection round was utilized as previously described to seed subsequent REPSA selection rounds. Sequence pools were sent for conventional dideoxy big dye sequencing. The Round 7 BtgZI selection (Figure 12B) seemed to be non-random in sequence composition relative to the Round 0 ST4-R20-L pool (Figure 12A). This strongly suggests that selection had likely occurred. As the Round 7 pool is still heterogeneous in composition, it becomes necessary to obtain sequence information on individual sequences. While this has historically been done through subcloning and conventional sequencing, with the availability of massively parallel sequencing, more expedient means were employed. We used the proton-detection sequencer Ion PGM as the primary method of identifying the sequences of individual strands present within the cleavage-resistant Round 7 pool. The Ion PGM determines sequences based on small changes in pH due to proton release upon nucleotide addition to the elongating strand. Round 7 selection DNAs were assessed by this method using a 100,000 well chip with four other barcoded experiments, resulting in 10,422 sequences for review.

Sequence sorting. Sequence data from the Ion Torrent PGM was sorted via Excel 2013. These sequences were subsequently analyzed for accuracy as the machine is prone to read errors, especially in polyhomonucleotide runs, and not all templates will be accurately amplified during PCR, producing both truncated and elongated strands. To sort strands for read accuracy, strands were first separated on overall length. With the trP1 and barcode regions removed, the core sequence will be exactly 60 bp in length. 1,597 sequences of the 10,422 obtained met this criterion. Approximately 7,100 sequences fell within +/- 2bp of this length and cursory examination of these sequences revealed that the difference in length was likely due to machine error in polyhomonucleotide runs, a common fault for proton detection sequencing technologies. The sequences were further sorted based on the defined regions adjacent to the central random core. Both the left and right defined regions are 20 bp long and only those sequences that perfectly matched the expected sequences were retained for analysis. Thus only 190 of the original 10,422 remained after the final screen.

Subsequent analysis of these 190 sequences by MEME (Multiple Em for Motif Elicitation) analysis with a palindromic sequence filter resulted in the discovery of the sequence 5'-GA(T/C)TGACC(C/A)GC(T/G)GGTCA(G/A)TC-3' (Figure 13A) with a statistical significance (e value) of 2.1e-109 (Bailey et al., 2009). It is unknown why the palindrome is extended beyond the TGACCNNNNGGTCA motif that was previously discovered (Agari et al., 2013). Figure 13B denotes the consensus sequence for all 4 sequences previously assessed to have SbtR-dependent repressive characteristics in the *T. thermophilus* HB8 genome, all on chromosome A. These sequences were utilized to generate the previous consensus sequence 5'-TGACCCNNKGGTCA-3' (Agari et al.,

2013). . In keeping with the 20 bp palindrome ascertained by HT-REPSA, a genome search was made with the expanded 20 bp palindrome.

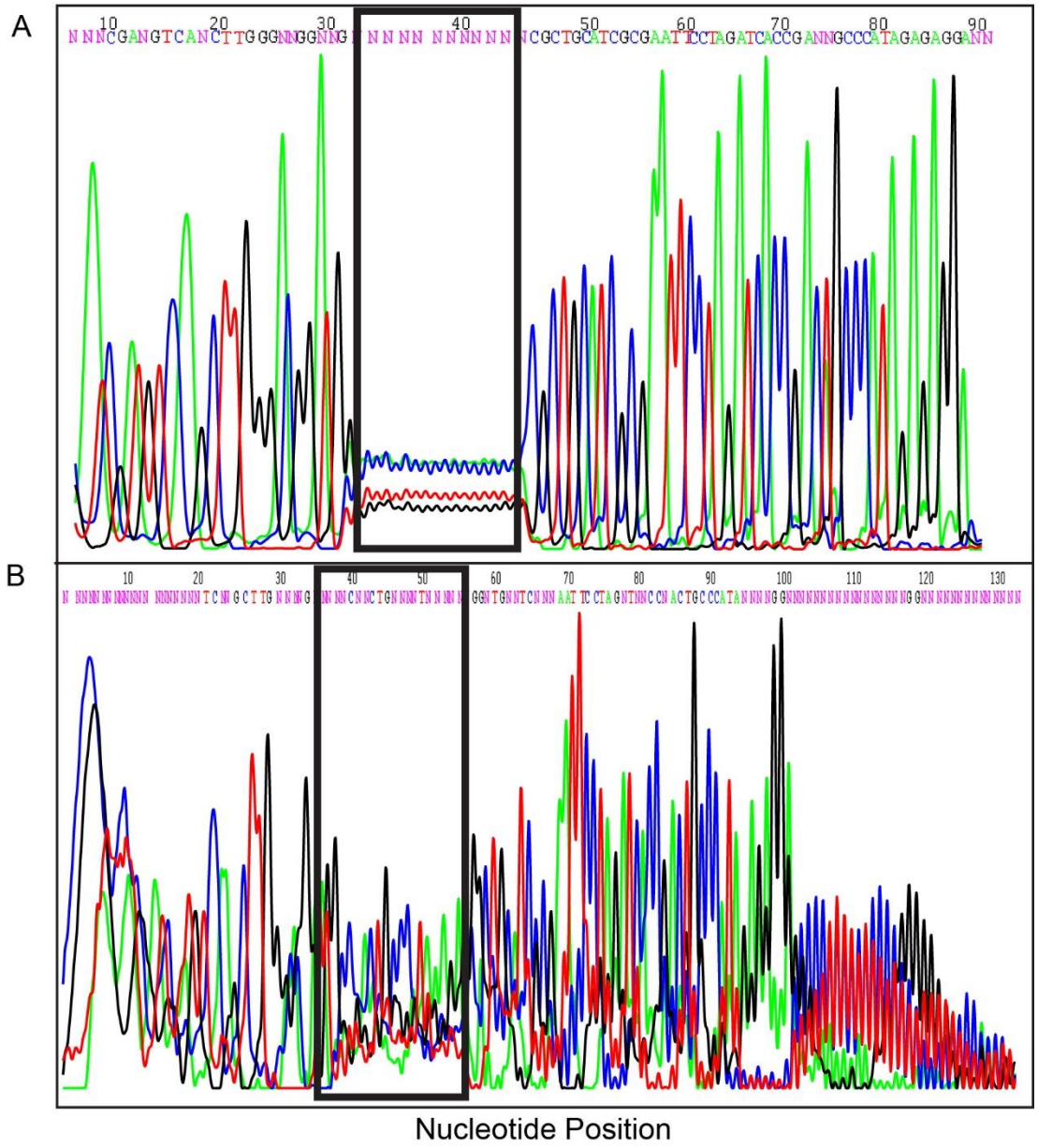

Figure 12: Sequencing of Round 7 HT-REPSA SbtR selection pool indicates the presence of nonrandom sequences compared to origin material, pre-selection ST4 template. **(A)** ST4-R20-L Round 0 pool sequence composition. **(B)** ST4-R20-L Round 7 BtgZI REPSA pool sequence composition. Blue = C, Red = T, Green = A, Black = G. Box indicates original randomized sequence region. Sequences were determined by big dye dideoxy sequencing.

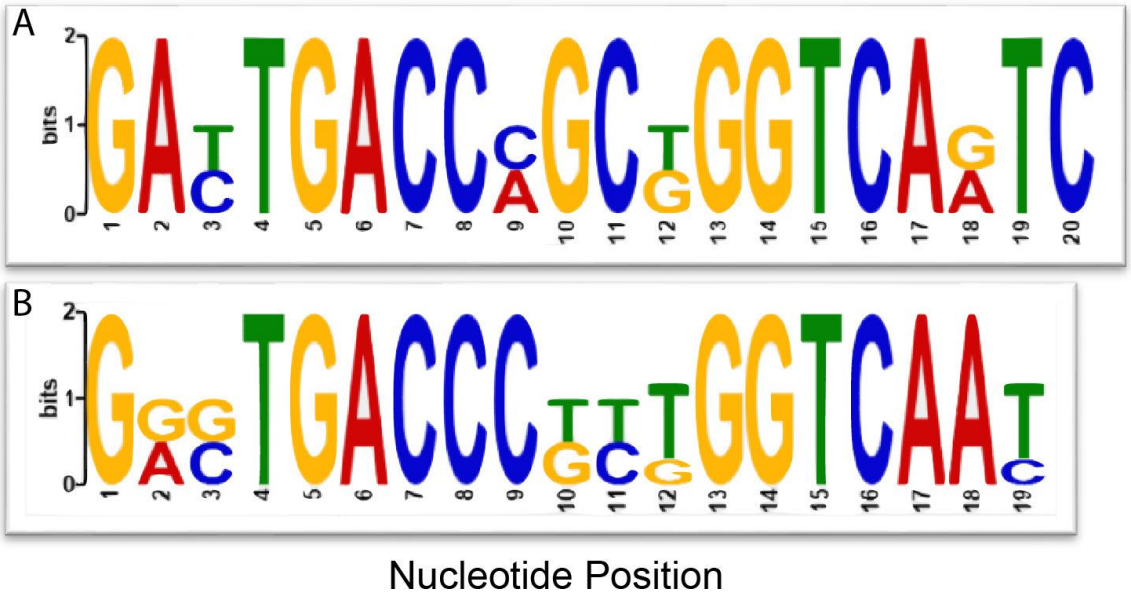

Figure 13: Comparison of sequence logos for SbtR. **(A)** Round 7 SbtR cleavage resistant REPSA selection pool. **(B)** Four sequences for which there is preliminary data for repressive activity (Agari et al., 2013). Y-axis (Bits) indicates information content and degeneracy of the nucleotide position. Bit equates to more statistical significance and less likely to be background noise. Sequence logos were determined by MEME analysis (Bailey et al., 2009).

Eight total sequences were identified when *T. thermophilus* HB8 was searched for NNNTGACCNNNNGGTCANNN, keeping the core palindrome but allowing for retrieval of sequences matching the length of the extended recognition site. Four of the sequences (Table A3) are located within ORFs TTHA0579, TTHA1325, TTHA1342, and TTHA1851, and are thus not likely a component of a gene promoter, although SbtR's ability to interfere with the transcription of these ORFs has not been fully excluded. One was shown by Agari et al. 2013 to lack repressive ability, so it was likely not a part of the promoter even though it was upstream of TTHA1330 (Table A4). Comparison of Round 7 material (Figure 13A) to genome derived sequences previously identified to display repressive ability (Figure 13B) show remarkable sequence similarity.

CHAPTER 4: DISCUSSION

These experiments, taken together, demonstrate that REPSA can be utilized to identify binding sites of DNA-binding proteins from the high temperature extremophile *Thermus thermophilus* HB8. REPSA application to unmodified SbtR in a complex mixture yielded consensus binding sequence in close agreement with previous findings (Agari et al. 2013).

The physical properties of unmodified SbtR seem to adhere closely to what was previously observed. It appears to possess a disulfide bridge within its dimerization domain, covalently linking two monomers of SbtR. The maintenance of a small dimeric band, consistent with the dimeric form of SbtR, even under reducing SDS PAGE conditions, lends credence to this assessment. Its high temperature of denaturation is also consistent with previous data as discussed in Chapter 2. Intriguingly, EMSA assessment of SbtR required no modification to native PAGE to achieve visible gel shifting with EtBr staining. It is likely that there is a caging factor to this gel shifting perhaps as a result of gel matrix confinement and localized concentration increases. However, it may be more likely a result of the solution cooling, locking SbtR onto the ST4-SbtR template by decreasing its movement range. SbtR evolved to function in high temperature conditions. In its active conformation(s) it must remain tightly bound to DNA under those high energy conditions. When subjected to cooler, lower energy conditions, it may be less likely to shift to its inactive set of states, remaining strongly bound to DNA even when subjected to the sieving effect of the gel matrix, resulting in the intense banding pattern

observed here (Figure 9 and 10). Such an effect may also explain the supershift observed in figure 10 when both BtgZI and higher concentrations of SbtR are present.

HT-REPSA application to unmodified SbtR resulted in a highly palindromic consensus sequence. The sequence found here,

5′-GA(T/C)TGACC(C/A)GC(T/G)GGTCA(G/A)TC-3′ is extended beyond the 14 base pair consensus sequence determined by Agari et al. in 2013, 5′-TGACCCNNNGGTCA-3′ by six base pairs (Agari et al., 2013) to 20 base pairs in length. The previous consensus sequence was determined at 55 °C in a heat and column purified solution, and the new consensus sequence was determined at 65 °C in a heat purified, complex solution. The core palindrome TGACCNNNNGGTCA seems to be largely maintained, confirming the previous findings. However, the palindrome that was determined by REPSA selection is sufficiently different from the SELEX derived material to warrant further study. It is unknown why the additional base pairs are so strongly selected for binding in this case, and requires further investigation as to what part they may play in protein-DNA binding. MEME analysis indicates that the extended portions of the palindrome seem to be highly preferred, with no demonstrated degeneracy in either the two base pair ending or in the two base pair center (bolded):

5′-**GA**(T/C)TGACC(C/A)**GC**(T/G)GGTCA(G/A)**TC**-3′

The remaining base pairs outside of the original TGACC-GGTCA motif are only partially degenerate, indicating that these are still highly preferred in these locations.

No DNA-SbtR co-crystal is currently available for assessment of SbtR-DNA interactions so the range of mechanisms SbtR may utilize to bind DNA over the active

temperature range of *T. thermophilus*, 45-80 °C, is unknown. The extended palindrome may be of biological import, potentially providing additional stability to the binding site, likely indirectly, allowing for selective binding, and greater transcriptional control *in vivo*.

HT-REPSA has been demonstrated, in agreement with previous findings, to be able to determine likely binding sites for unmodified transcription factors in a complex mixture (Hardenbol & Van Dyke, 1996). The SbtR heat purified sample mixture, was likely a complex mixture of RNA fragments, soluble *E. coli* protein fragments, and SbtR proteins, both active and denatured.

In addition, during execution of these experiments, all three HT-IISREs utilized here, BtgZI, BsmFI, and BseXI, seem to have some set of sequences for which they are cleavage refractory. They all demonstrated strong selective preference for transcription factor independent cleavage resistance after only a few (three to six) rounds of REPSA selection. The root cause of this cleavage resistance is unknown. This resistance may be due to selection of a second binding site for each IISRE in the random core as was observed for BsgI (Hardenbol & Van Dyke, 1997). However, the more intriguing possibility, for which there is little data in the literature, is that there is some sequence or set of sequences that are refractory to IISRE cleavage (Lundin et al., 2015). IISREs are primarily modeled around the behavior of FokI, which has not been demonstrated to display sequence specificity in its cleavage ability, however it does not mean that these enzymes have no sequence specificity. The possibility of selectivity in cleavage may provide greater insight into how IISREs function. Studying these three in particular may

yield insight into how IISREs have adapted to cleave DNA under high temperature conditions.

Conclusion and future directions. In summary, we have established that REPSA has the potential to be applicable for combinatorial selection of transcription factor consensus binding sequences found in high temperature organisms. HT-IISREs utilized here seemed to hide SbtR's contribution to cleavage resistance due to selection of transcription factor independent cleavage resistance. Understanding their cleavage behavior or modifying the technique to make use of better understood HT-IISREs, or other IISREs in general, may allow for more reliable use of HT-REPSA.

Though the ST4-Long template variant proved useful in these studies, its flaw became apparent when sequencing needed to be performed. Manufacture of the extended length ST4 templates required both the Ion Torrent identifier sequence, trP1, as well as a long, A_BC barcode region. The barcode region was preferred for high temperature application due to its length. However, cleavage resistant rounds all contained the same barcode when separated from the REPSA selection pool, hampering effective Ion Torrent analysis. As a result, future experimentation should utilize either an elongated core ST4 template, extended by 10 to 15 base pairs on either side of the template, or they should only utilize the trP1 sequence, allowing for custom barcoding and more efficient use of the Ion Torrent.

Agari et al., 2013 did not assess five of the putative SbtR binding sites, limiting the potential validity of their determined consensus sequence (Table A3 and A4). These remaining sites, in addition to the briefly mentioned TGACCGGTCA containing sites, need to be assessed for SbtR repressive control. Alternatively, the large palindrome may

have resulted due to a high concentration of SbtR. REPA (Restriction Endonuclease Protection Assay) optimization experiments were carried out prior to REPSA rounds to achieve the lowest experimentally viable concentration to limit such an effect. More thorough kinetics analysis needs to be conducted to provide a more complete understanding of SbtR's binding characteristics.

APPENDIX A

Table A1

Identifier	Experiment	Barcode
ST4-R0-000-00	ST4 R20 L TEMPLATE	ABC11
ST4-R4-F1C-01	BsmF1 CRSA R4	ABC11
ST4-R4-Z1C-02	BstgZ1 CRSA R4	ABC11
ST4-R6-X1C-03	Bsex1 CRSA R6	ABC11
ST4-R5-Z1R-04	BtgZ1 REPSA Attempt 1 R5	ABC11
ST4-R7-Z1R-05	BtgZ1 REPSA Attempt 2 R7	ABC11
ST4-R8-F1R-06	BsmF1 REPSA Attempt 2 R8	ABC11
ST4-R8-X1R-07	Bsex1 REPSA Attempt 2 R8	ABC11
ST4-R11-F1R-08	BsmF1 REPSA Attempt 2 R11	ABC11
ST4-R11-Z1R-09	BtgZ1 REPSA Attempt 2 R11	ABC11
ST4-R12-K1R-10	FokI REPSA Attempt 2 R12	ABC11
ST4-R13-X1R-11	BsexI REPSA Attempt 2 R13	ABC11
ST4-R5-F1R-12	BsmF1 REPSA Attempt 2 R5	ABC11

Table A1: Cleavage Resistant Populations Obtained from CRSA and REPSA.
CRSA=Cleavage Resistance Selection Assay. Identifier is broken down into four parts: part 1 denotes the selection template used (ST4=ST4R20). Part 2 denotes the round, e.g. R3=round 3, in which discriminate cleavage resistance was observed. Part 3 denotes the IISRE used in that experiment (F1=BsmF1, K1=Fok1, X1=BseX1, Z1=BtgZ1) and the type of selection (C=CRSA and R=REPSA). Part 4 is a unique identifier number indicating the order in which cleavage resistance was obtained. The base template is listed as R0 and is given all 0 identifiers as it is not a cleavage resistant species. Attempt indicates REPSA experimental group. Attempt 1 covered optimization of REPSA protocol for high temperature IISREs. Attempt 2 was a true REPSA experiment, incorporating information garnered from Attempt 1. Barcode indicates identifier code, of which there are four for ST4, for sequencing on Ion Torrent. Note: Sequence 04 was

misidentified as a cleavage resistant species obtained during REPSA. Sequence analysis revealed it to be ST4-SbtR.

Table A2

Entry	Status	Protein names	ORF Identifier	Length
Q5SGM2	unreviewed	Anti-toxin-like protein	TTHC012	70
Q5SI21	reviewed	Arginine repressor	argR TTHA1559	164
Q5SK65	reviewed	Bifunctional protein PyrR	pyrR TTHA0783	181
Q5SLW8	unreviewed	Cold shock protein	TTHA0175	73
Q5SLD4	unreviewed	Cold shock protein	TTHA0359	68
Q5SLN8	unreviewed	Ferric uptake regulation protein	TTHA0255	147
Q5SLE9	unreviewed	Ferric uptake regulatory protein	TTHA0344	131
Q5SM85	unreviewed	Heat-inducible transcription repressor HrcA	hrcA TTHA0058	300
G9MB63	unreviewed	HicB family protein	TTHV009	155
Q53W62	reviewed	HTH-type transcriptional repressor CarH	carH TTHB100	285
G9MB68	unreviewed	LacI-family transcriptional regulator	TTHV015	325
Q5SKK9	unreviewed	Magnesium chelatase related protein	TTHA0634	464
Q5SJ59	unreviewed	Mercuric resistance operon regulatory protein (MerR)	TTHA1155	142
Q5SIS2	unreviewed	Metal uptake regulation protein, putative	TTHA1292	122
Q5SJ93	reviewed	N utilization substance protein B homolog (Protein NusB)	nusB TTHA1121	151
Q5SM86	unreviewed	Nitrogen regulatory protein P-II	TTHA0057	116
Q5SLZ8	unreviewed	Phosphate regulon transcriptional regulatory protein PhoB	TTHA0145	223
Q5SH54	unreviewed	Probable repressor, phenylacetic acid catabolic pathway	TTHA1876	260
Q5SLV7	unreviewed	Probable transcriptional regulator	TTHA0186	285
Q53W30	unreviewed	Probable transcriptional regulator, CopG family	TTHB136	96
Q5SK31	reviewed	Probable transcriptional regulatory protein TTHA0821	TTHA0821	244
P38383	reviewed	Protein translocase subunit SecE	secE TTHA0249	60
Q5SHK8	unreviewed	Putative response regulator	TTHA1722	225

Q53VW8	unreviewed	Putative RNA polymerase sigma factor	TTHB211	193
Q53W36	unreviewed	Putative transcriptional regulator	TTHB130	112
Q53VY3	unreviewed	Putative transcriptional regulator	TTHB186	329
Q5SHS3	reviewed	Redox-sensing transcriptional repressor Rex	rex TTHA1657	211
Q5SMC3	unreviewed	Response regulator	TTHA0020	223
Q5SJK6	unreviewed	Response regulator	TTHA1002	240
Q5SJH8	unreviewed	Response regulator	TTHA1030	192
Q5SIL7	unreviewed	Response regulator	TTHA1352	215
Q5SIK7	unreviewed	Response regulator	TTHA1362	227
Q5SI72	unreviewed	Response regulator	TTHA1502	227
Q53VZ7	unreviewed	Reverse gyrase	rgy TTHB172	1116
Q5SK01	reviewed	Ribosomal RNA small subunit methyltransferase B	rsmB TTHA0851	398
Q5SKW1	unreviewed	RNA polymerase sigma factor SigA	sigA TTHA0532	423
Q5SKM1	unreviewed	Transcription elongation factor GreA	greA TTHA0622	155
Q5SJG6	reviewed	Transcription inhibitor protein Gfh1	gfh1 TTHA1042	156
Q5SID7	unreviewed	Transcription regulator, Crp family	TTHA1437	216
Q5SJE9	unreviewed	Transcription termination factor Rho	rho TTHA1065	426
P48514	reviewed	Transcription termination/antitermination protein NusA	nusA TTHA0701	387
P35872	reviewed	Transcription termination/antitermination protein NusG	nusG TTHA0248	184
Q5SKY6	unreviewed	Transcriptional regulator	TTHA0507	274
Q5SK45	unreviewed	Transcriptional regulator	TTHA0807	344
Q53W89	unreviewed	Transcriptional regulator	TTHB073	258
Q5SLX6	unreviewed	Transcriptional regulator (TetR/AcrR family)	TTHA0167	189
Q5SKB5	unreviewed	Transcriptional regulator MarR family	TTHA0733	144
Q5SJD9	reviewed	Transcriptional regulator MraZ	mraZ TTHA1075	144
Q53W63	unreviewed	Transcriptional regulator, Crp family	TTHB099	195

Q5SIL0	unreviewed	Transcriptional regulator, FNR/CRP family	TTHA1359	202
Q5SI00	unreviewed	Transcriptional regulator, GntR family	TTHA1580	220
Q53VT7	unreviewed	Transcriptional regulator, IclR family	TTHB248	283
Q53W81	unreviewed	Transcriptional regulator, lacI family	TTHB081	330
Q5SM20	unreviewed	Transcriptional regulator, LysR family	TTHA0123	317
Q5SKY5	unreviewed	Transcriptional regulator, MerR family	TTHA0508	233
Q5SJN5	unreviewed	Transcriptional regulator, TetR family	TTHA0973	203
Q53WD9	unreviewed	Transcriptional regulator, TetR family	TTHB023	191
Q5SI13	unreviewed	Transcriptional regulatory protein	TTHA1567	207
Q53VZ6	unreviewed	Transcriptional regulatory protein	TTHB173	217
Q5SK94	unreviewed	Transcriptional repressor	TTHA0754	219
Q5SM09	reviewed	Transcriptional repressor NrdR	nrdR TTHA0134	153
Q5SKD8	unreviewed	Transcriptional repressor SmtB	TTHA0705	123
Q5SM42	unreviewed	Transcriptional repressor, TetR family	TTHA0101	205
Q5SJW3	unreviewed	Transcription-repair-coupling factor (TRCF) (EC 3.6.4.-)	mfd TTHA0889	978
Q5SLY7	unreviewed	Uncharacterized protein	TTHA0156	98
Q5SLX5	unreviewed	Uncharacterized protein	TTHA0168	164
Q5SKI8	unreviewed	Uncharacterized protein	TTHA0655	200
Q5SJM7	unreviewed	Uncharacterized protein	TTHA0981	107
Q5SJJ9	unreviewed	Uncharacterized protein	TTHA1009	104
Q5SHY7	unreviewed	Uncharacterized protein	TTHA1593	216
Q53WC5	unreviewed	Uncharacterized protein	TTHB037	875
Q53VU5	unreviewed	Uncharacterized protein	TTHV057 TTHB234	76
G9MBB2	unreviewed	Uncharacterized protein	TTHV060	81

Table A2: Potential Transcription Factors Annotated within T. thermophilus HB-8 Genome Cached in NCBI Database. Putative transcription factors, annotated by homology either automatically or manually. Accessed by searching for ""taxonomy: *Thermus thermophilus* (strain HB-8 / ATCC 27634 / DSM 579) [300852]" transcription" in Uniprot KB. Ribosomal subunits and RNA/DNA polymerase components were removed from the original list to reduce list to likely transcription factors.

Table A3

ORF Identifier	Putative Function	Recognition Site
TTHA0027	Postassium Channel Subunit Beta	5′-GACTGACCCGCTGGTCAATC-3′
TTHA0785	Putative Sulfie Exporter (TauE)	5′-GACTGACCCGCGGGTCAACC-3′
TTHA0786	Glycerate dehydrogenase/Glyoxylate Reductase	
TTHA0787	Hypothetical Protein	
TTHA1818	Recombinase A	
TTHA1819	2'-5' RNA Ligase (ligT)	
TTHA1820	CinA (competence inducibale protein)	
TTHA1821	Folate Bindin Aminomethyltransferase	5′-GGGTGACCCTTTGGTCAATA-3′
TTHA1822	Putative Transporter	5′-TATTGACCAAAGGGTCACCC-3′
TTHA1823	Putative Hydrolase	

Table A3: Putative ORFs Controlled by SbtR by Genomic Analysis by (Agari et al. 2013)
Colors indicate operon. There are 5 additional putative sites found via genomic searching for TGACCNNNNGGTCA. However, 1 site failed to elicit a repressive response, likely due to lying to far from the promotor sequences. The remaining 4 reside within ORFS and are thus not likely to be bacterial promoters.

Table A4

ORF	Recognition Site	Reason for Repression Failure and Putative Function of ORF
TTHA0579	5'-CCTTGACCCGCCGGTCAATC-3'	Inside ORF Sugar ABC Transporter
TTHA1325	5'-TTATGACCTCTTGGTCAGCC-3'	Inside ORF Sulfite Oxidase
TTHA1330	5'-CCCTGACCCGTTGGTCACGC-3'	Outside of promoter region. Peptide ABC Transporter/Permease
TTHA1342	5'-CGCTGACCGACCGGTCATGC-3'	Inside ORF ABC Transporter ATP Binding Protein
TTHA1851	5'-TTCTGACCGGCGGGTCAGGT-3'	Inside ORF Unknown Protein

Table A4: Additional putative SbtR binding sites with unknown biological function.

APPENDIX B: MATERIALS AND METHODS

Template Purification

Either Qiagen Minelute PCR purification kit or Zymogen DNA Clean Kits were used to purify DNA from REPSA selection rounds and PCR-amplified templates. Standard protocols were used with the following modifications: (1) modified Qiagen protocol incorporated a 1 minute drying step to remove residual alcohol. (2) modified Zymogen protocol incorporated a 600 μL wash step and 30 s drying step to reduce potential contamination. Zymogen DNA Clean Kits were preferentially utilized due to greater apparent template yield and higher purity using a simpler protocol with more stable columns.

Assessment of DNA Concentration

A Nanodrop 2000 (Fischer-Thermo Scientific) was utilized to do a rough assessment of DNA purity with "pure" DNA having a 260/280 ration greater than 1.8 and a 230/260 ratio greater than 2. 230/260 and 260/280 measurements have provided good general indications of the success of the amplification step as well as the level of potential contamination. However, it does not reveal the concentration of dsDNA, only the 260 absorbance of the sample so these results should be checked via other methodologies if available. Qubit was later utilized and preferred for this role due to its ability to detect dsDNA in a complex solution, allowing for relatively quick analysis of unpurified post PCR solutions. With the Nanodrop, DNA concentrations were tested post PCR, following a purification step to eliminate stronger background 260nm noise from dNTPs and ssDNA primers. With Qubit, the DNA concentration was tested post PCR,

ignoring the purification step, as its method of measurement is dependent on an intercalating fluorophore, allowing for ready detection of only dsDNA.

Nanodrop Protocol. The Nanodrop 2000, set to the "nucleic acids" setting, is blanked with a vehicle buffer that is to contain the analyzed samples. Either Qiagen's Buffer TE or Zymogen's Elution Buffer, are used for this purpose until a flat spectra is obtained after blanking. At least 3 separate aliquots of the sample, either 1µl or 2 µl depending on how much sample is available, are analyzed to obtain a more accurate spectra. Ideally, the least amount of sample should be used to obtain readings so 1µl aliquots are preferred.

Denaturing PAGE

Denaturing page was conducted utilizing standard protocols obtained from Bio-Rad. Protein gels, denaturing and 1.0mm thickness, were 4-12% (37.5:2) stacking gels (SDS PAGE) and were run at 100V until bromophenol blue (BPB) indicator dye front had run off of gel. Gels were stained with 0.1% Coomassie Blue R-250 (10% acetic acid, 50% methanol, 40% H_2O) four 4 hours and destained in 5:1:5 MeOH:HOAc:water. Gels were electrophoresed in Tris-Glycine Buffer [5x stock (1 L=15.1g Tris base, 94g glycine, 50 ml of 10% SDS)]. Sample buffer was NuPAGE LDS 4X (Thermo-Fisher Scientific).

General REPSA Buffers

Table B1. General REPSA Buffers

Buffer	Storage Concentration and Conditions	Working Concentration	Use
TBE	5x, Refrigerator	90mM Tris-borate, 2mM EDTA pH 8.3	Polyacrylamide gel running buffer
TAE	50x, Refrigerator	40mM Tris, 20mM Acetic acid, 1mM EDTA	Agarose gel running buffer
TE	1x, -20 °C	10mM Tris-Cl pH 8.0, 1mM EDTA	DNA dilution
TEN 100 (TthA)	10x, -20 °C	1mM Tris-Cl pH 8.0, 1mM EDTA pH 8.0, 10mM NaCl, 50% glycerol.	Cell resuspension for storage
TthB	10x, -20 °C	10mM Tris-Cl pH 8.0, 1mM EDTA pH 8.0, 10mM NaCl.	Sample Buffer
NEB Cutsmart Buffer (CSB)	10x, -20 °C	50mM KAc, 20mM Tris-Ac, 10mM $MgAc_2$, 100µg/ml BSA, pH 7.9	BtgZ1 and BsmFI IISRE reaction buffer
Thermoscientific BseX Buffer	10x, -20 °C	50mM Tris-HCl, pH 7,5 @ 37 °C, 2mM $MgCl_2$, 100mM NaCl, 100µg/ml BSA	BseXI reaction buffer
NEB Diluent A	5x, -20 °C	10mM Tris-Cl pH7.4, 1mM DTT, 0.1mM EDTA, 200µg/mL, 50mM KCl, 5% glycerol.	Glycerol containing dilution buffer
IISRE Buffer	5x, -20 °C	10mM Tris-Cl pH7.4, 1mM DTT, 0.1mM EDTA, 200µg/mL, 50mM KCl	Glycerol free IISRE dilution buffer

Unless otherwise noted, pH for buffers was obtained at 25 °C.

Native PAGE

Native PAGE gels, 1.0mm thickness, 10% (19:1 acrylamide:bis-acrylamide) were run at 100V until BPB dye front ran off of gel. Running buffer was 1x TBE. They were stained with EtBr for 15 minutes, destained in ddH_2O for 5 minutes, and visualized on a UV plate reader with a 2 minute exposure time. Sample buffer was NEB Loading Buffer Blue (6X).

Agarose Gel Electrophoresis

Agarose gels were 1% unless otherwise noted as they were primarily utilized for plasmid size and restriction digest testing. Smaller nucleic acid species were assessed via PAGE instead of 2 or 3% agarose gel electrophoresis.

Transformation and Plasmid Purification

Plasmids containing the proteins of interest were obtained from the Riken Institute Whole Cell Project, a large collaborative project among Japanese Universities and private institutes, which currently works on *Thermus thermophilus* HB8 genome analysis. Upon receipt, the plasmid vectors were transfected into ***E. coli*** JM109 cells for the purpose of plasmid amplification. JM109 cells were ampicillin selected for transformation on a streak plate. Transformed JM109 cells were incubated overnight in 5mL of 50µg/ml ampicillin containing LB Miller media. Plasmids were purified using an Omega Bio-Tek E.Z.N.A Plasmid Spin Kit and standard protocol.

Chemical Competence Protocol. Cells, ***E. coli*** JM109 and BL21DE3, were made chemically competent by a protocol obtained from OpenWetWare (Chemically Competent Cells 2015).

Transformation Protocol. Promega's Quick Protocol for ***E. coli*** Transformation was used to transform ***E. coli***. The heat shock step was extended to 30s from 20s as this seemed to yield a greater number of viable transformants.

Cell Culture

Unless otherwise noted, all cell work was performed in a laminar flow hood with aseptic technique to prevent contamination of cell cultures. OD_{600} was measured for with HD-BioTek plate reader with blank LB media controls.

Escherichia coli. E. coli cultures JM109 and BL21DE3 were seeded from freeze down cultures obtained as a generous gift from Dr. Glen Meades, Kennesaw State University, Department of Chemistry and Biochemistry. ***E. coli*** cultures were incubated at 37 ºC during growth phases and 25 ºC post IPTG induction.

Thermus thermophilus HB8. *T. thermophilus* cells were ordered and subsequently obtained from ATCC in a double vial. The vial was opened via heat shock method in a laminar flow hood. The pellet within the vial was resuspended in 4 mL DSMZ-74 medium, gently mixed by pipette, and transferred to 3 high temperature (ATCC 697 Medium) agar plates and 2 agar slants by sterile metal loop. Plates and slants were placed within an aerated plastic box, agar down, with a 100 mL beaker full of water to lessen dehydration of the agar during growth as cell incubators utilized, New Brunswick Innova 44, lack humidity controls. Plates were incubated at 72 ºC for 24 hours. The remainder of

resuspended *T. thermophilus* cells were used to seed a 50 mL liquid cultures in a narrow neck flask with DSMZ-74 Thermus broth and grown for 24 hours at 72 ºC at 200 rpm. Both culture methods were performed to reduce the chance of loss of the strain due to incubation failure/interruption. After cells had incubated for 24 hours, six 0.5mL aliquots were taken from the liquid culture, mixed with 0.5 mL 40% glycerol cryo-storage medium, flash frozen in liquid nitrogen, and stored at -80 ºC.

Protein Expression

E. coli BL21DE3 cells were utilized for expression of proteins. SbtR, as well as other TFs obtained from RIKEN, are under lac operon control so they may be induced by IPTG induction. Neither variant of SbtR appeared to be toxic to ***E. coli*** BL21DE3 cells.

Protein purification

Both SbtR and SbtR-His were heat purified at 80 ºC for 20 minutes. This is well below SbtR's demonstrated denaturation temperature of 98.5 ºC (Agari et al., 2013). SbtR denaturation and refolding tests indicate that SbtR lacks the ability to spontaneously refold under our reaction conditions so care should be taken to keep the proteins below their melting point. Heat purification was performed either in a thermocycler for small scale testing and optimization, or in a water loaded hot block for large scale production.

PCR Protocols

ST4 modified template manufacture. ST4-long variations were manufactured by appended the Ion Torrent (Trp1_ST4L) marker to the 5′ end of the top strand (Figure 2) of the core template and a barcode marker (A-BCxx-R) to the 5′ end of the bottom strand

(primers detailed in Figure B1). This made the template of suitable length for HT-REPSA. However, the appending of these long sections of DNA can go no more than 6 cycles of PCR without serious risk of overproducing primer-dimers as well as undesirable tertiary species from off-target annealing and extension. The products from this reaction should always be column purified before amplifying further. Due to the truncated nature of the cycling, only 100nM primer concentration is necessary and encouraged to reduce the likelihood of off-target products. Annealing temperature should be 62 °C and elongation should be 72 °C. Annealing temperatures below 62 °C seems to result in greater primer dimer formation and less affirming to the ST4 core template.

General PCR amplification. Short selection template universal, for all current REPSA templates, primers (Trp1-L-universal) and (A-R-universal) were utilized in all subsequent amplifications with ST4-L templates. As before, 100nM primers are preferred here to reduce the likelihood of amplification of undesirable off-target products that may have survived purification. Annealing temperature should be no higher than 55 °C and elongation phase temperature should be 68°C as the universal primers are less than ideal in Tm and this temperature set seems to reliably amplify ST4 templates.

The core ST4 template has its own set of primers tailored for its amplification with 58 °C being the optimum annealing temperature and 72 °C being the optimum elongation temperature. The specific template is amplified with 200nM L and R primers as it is usually cycled for 35 rounds. The random template cannot be cycled for more than 6 rounds so no more than 100nM of L and R primers should be used to reduce waste and reduce off-target and tertiary species formation within the random core region.

Figure B1. Primers utilized in PCR for ST4 variants

WORKS CITED

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.

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.

Agari, Y.; Sakamoto, K.; Kuramitsu, S.; Shinkai, A. Transcriptional Repression Mediated By a TetR Family Protein, PfmR, from Thermus Thermophilus HB8. *Journal of Bacteriology*. **2012**, *194*, 4630–4641.

Agari, Y.; Sakamoto, K.; Yutani, K.; Kuramitsu, S.; Shinkai, A. Structure And Function of a TetR Family Transcriptional Regulator, SbtR, fromHB8. *Proteins: Structure, Function, and Bioinformatics*. **2013**, *81*, 1166–1178.

Baba, T.; Ara, T.; Hasegawa, M.; Takai, Y.; Okumura, Y.; Baba, M.; Datsenko, K. A.; Tomita, M.; Wanner, B. L.; Mori, H. Construction Of Escherichia Coli K-12 in-Frame, Single-Gene Knockout Mutants: the Keio Collection. *Molecular Systems Biology*. **2006**, *2*.

Babu, M. M.; Teichmann, S. A. Evolution Of Transcription Factors and the Gene Regulatory Network in Escherichia Coli. *Nucleic Acids Research*. **2003**, *31*, 1234–1244.

Bailey, T. L.; Boden, M.; Buske, F. A.; Frith, M.; Grant, C. E.; Clementi, L.; Ren, J.; Li, W. W.; Noble, W. S. MEME SUITE: Tools for Motif Discovery and Searching. *Nucleic Acids Research*.**2009**, *37*, W202–W208.

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.

Bernhardt, T. G.; de Boer, P. A. SlmA, a Nucleoid-Associated, FtsZ Binding Protein Required for Blocking Septal Ring Assembly over Chromosomes in E. Coli. *Molecular Cell*. **2005**, *18*, 555–564.

Bieniossek, C.; Papai, G.; Schaffitzel, C.; Garzoni, F.; Chaillet, M.; Scheer, E.; Papadopoulos, P.; Tora, L.; Schultz, P.; Berger, I. The Architecture of Human General Transcription Factor TFIID Core Complex. *Nature*. **2013**, *493*, 699–702.

Bitinaite, j.; Wah, D. A.; Schildkraut, I.; Aggarwal, A. K. Structure Of FokI Has Implications for DNA Cleavage. *Proceedings of the National Academy of Sciences*. **1998**, *95*, 10564–10569.

Blattner, F. R.; Plunkett, 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.

Broggini, M.; Coley, H. M.; Mongelli, N.; Pesenti, E.; Wyatt, M. D.; Hartley, J. A.; Dlncaici, M. DNA Sequence-Specific Adenine Alkylation by the Novel Antitumor Drug Tallimustine (FCE 24517), a Benzoyl Nitrogen Mustard Derivative of Distamycin. *Nucleic Acids Research*. **1995**, *23*, 81–87.

Cardew, A. S.; Brown, T.; Fox, K. R. Secondary Binding Sites for Heavily Modified Triplex Forming Oligonucleotides. *Nucleic Acids Research*. **2012**, *40*, 3753–3762.

Catto, L. E.; Ganguly, S.; Milsom, S. E.; Welsh, A. J.; Halford, S. E. Protein Assembly and DNA Looping by the FokI Restriction Endonuclease. *Nucleic Acids Research*. **2006**, *34*, 1711–1720.

Cava, F.; Hidalgo, A.; Berenguer, J. Thermus Thermophilus as Biological Model. *Extremophiles*. **2009**, *13*, 213–231.

Cavicchioli, R.; Siddiqui, K. S.; Andrews, D.; Sowers, K. R. Low-Temperature Extremophiles and Their Applications. *Current Opinion in Biotechnology*. **2002**, *13*, 253–261.

Charoensawan, V.; Wilson, D.; Teichmann, S. A. Lineage-Specific Expansion of DNA-Binding Transcription Factor Families. *Trends in Genetics*. **2010**, *26*, 388–393.

Cho, H.; Mcmanus, H. R.; Dove, S. L.; Bernhardt, T. G. Nucleoid Occlusion Factor SlmA Is a DNA-Activated FtsZ Polymerization Antagonist. *Proceedings of the National Academy of Sciences*.**2011**, *108*, 3773–3778.

Collas, P. The Current State Of Chromatin Immunoprecipitation. *Molecular Biotechnology*. **2010**, *45*, 87–100.

Cordero, O. X.; Hogeweg, P. Regulome Size in Prokaryotes: Universality and Lineage-Specific Variations. *Trends in Genetics*. **2009**, *25*, 285–286.

Cuthbertson, L.; Nodwell, J. R. The TetR Family Of Regulators. *Microbiology and Molecular Biology Reviews*. **2013**, *77*, 440–475.

Daum, L. T.; Rodriguez, J. D.; Worthy, S. A.; Ismail, N. A.; Omar, S. V.; Dreyer, A. W.; Fourie, P. B.; Hoosen, A. A.; Chambers, J. P.; Fischer, G. W. Next-Generation Ion Torrent Sequencing Of Drug Resistance Mutations in Mycobacterium Tuberculosis Strains. *Journal of Clinical Microbiology*. **2012**, *50*, 3831–3837.

Demirjian, D. C.; Morís-Varas Francisco; Cassidy, C. S. Enzymes From Extremophiles. *Current Opinion in Chemical Biology*. **2001**, *5*, 144–151.

Dey, B.; Thukral, S.; Krishnan, S.; Chakrobarty, M.; Gupta, S.; Manghani, C.; Rani, V. DNA–Protein Interactions: Methods for Detection and Analysis. *Molecular and Cellular Biochemistry*.**2012**, *365*, 279–299.

Djordjevic, M. SELEX Experiments: New Prospects, Applications and Data Analysis in Inferring Regulatory Pathways. *Biomolecular Engineering*. **2007**, *24*, 179–189.

Ellington, A. D.; Szostak, J. W. In Vitro Selection of RNA Molecules That Bind Specific Ligands. *Nature*. **1990**, *346*, 818–822.

Frock, A. D.; Kelly, R. M. Extreme Thermophiles: Moving beyond Single-Enzyme Biocatalysis. *Current Opinion in Chemical Engineering*. **2012**, *1*, 363–372.

Fujiwara, K.; Tsubouchi, T.; Kuzuyama, T.; Nishiyama, M. Involvement Of the Arginine Repressor in Lysine Biosynthesis of Thermus Thermophilus. *Microbiology*. **2006**, *152*, 3585–3594.

Gade, P.; Kalvakolanu, D. V. Chromatin Immunoprecipitation Assay As a Tool for Analyzing Transcription Factor Activity. *Methods in Molecular Biology*. **2014**, *809*, 85–104.

Gama-Castro, S.; Salgado, H.; Peralta-Gil, M.; Santos-Zavaleta, A.; Muniz-Rascado, L.; Solano-Lira, H.; Jimenez-Jacinto, V.; Weiss, V.; Garcia-Sotelo, J. S.; Lopez-Fuentes, A.; Porron-Sotelo, L.; Alquicira-Hernandez, S.; Medina-Rivera, A.; Martinez-Flores, I.; Alquicira-Hernandez, K.; Martinez-Adame, R.; Bonavides-Martinez, C.; Miranda-Rios, J.; Huerta, A. M.; Mendoza-Vargas, A.; Collado-Torres, L.; Taboada, B.; Vega-Alvarado, L.; Olvera, M.; Olvera, L.; Grande, R.; Morett, E.; Collado-Vides, J. RegulonDB Version 7.0: Transcriptional Regulation of Escherichia Coli K-12 Integrated within Genetic Sensory Response Units (Gensor Units). *Nucleic Acids Research*. **2010**, *39*, D98–D105.

Gold, L.; Polisky, B.; Uhlenbeck, O.; Yarus, M. Diversity Of Oligonucleotide Functions. *Annual Review of Biochemistry*. **1995**, *64*, 763–797.

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

Hahn, S.; Buratowski, S.; Guarente, L. Yeast TATA-Binding Protein TFIID Binds To TATA Elements with Both Consensus and Nonconsensus DNA Sequences. *Proceedings of the National Academy of Sciences*. **1989**, *86*, 5718–5722.

Hampshire, A.; Rusling, D.; Broughtonhead, V.; Fox, K. Footprinting: A Method for Determining the Sequence Selectivity, Affinity and Kinetics of DNA-Binding Ligands. *Methods*. **2007**, *42*, 128–140.

Hanson, A. D.; Pribat, A.; Waller, J. C.; Crécy-Lagard, V. D. 'Unknown' Proteins and 'Orphan' Enzymes: the Missing Half of the Engineering Parts List – and How to Find It. *Biochemical Journal*. **2010**, *425*, 1–11.

Hardenbol, P.; Van Dyke, M. W. Identification Of Preferred HTBP DNA Binding Sites by the Combinatorial Method REPSA. *Nucleic Acids Research*. **1997**, *25*, 3339–3344.

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*. **1996**, *93*, 2811–2816.

Hardenbol, P.; Wang, J. C.; Dyke, M. W. V. Identification Of Preferred Distamycin−DNA Binding Sites by the Combinatorial Method REPSA. *Bioconjugate Chemistry*. **1997**, *8*, 617–620.

Hellman, L. M.; Fried, M. G. Electrophoretic Mobility Shift Assay (EMSA) for Detecting Protein–Nucleic Acid Interactions. *Nature Protocols*. **2007**, *2*, 1849–1861.

Henne, A.; Brüggemann, H.; Raasch, C.; Wiezer, A.; Hartsch, T.; Liesegang, H.; Johann, A.; Lienard, T.; Gohl, O.; Martinez-Arias, R.; Jacobi, C.; Starkuviene, V.; Schlenczeck, S.; Dencker, S.; Huber, R.; Klenk, H.-P.; Kramer, W.; Merkl, R.; Gottschalk, G.; Fritz, H.-J. The Genome Sequence of the Extreme Thermophile Thermus Thermophilus. *Nature Biotechnology*. **2004**, *22*, 547–553.

Herzig, M. C. S.; Trevino, A. V.; Arnett, B.; Woynarowski, J. M. Tallimustine Lesions In Cellular DNA Are AT Sequence-Specific but Not Region-Specific. *Biochemistry*. **1999**, *38*, 14045–14055.

Hu, P.; Janga, S. C.; Babu, M.; Díaz-Mejía, J. J.; Butland, G.; Yang, W.; Pogoutse, O.; Guo, X.; Phanse, S.; Wong, P.; Chandran, S.; Christopoulos, C.; Nazarians-Armavil, A.; Nasseri, N. K.; Musso, G.; Ali, M.; Nazemof, N.; Eroukova, V.; Golshani, A.; Paccanaro, A.; Greenblatt, J. F.; Moreno-Hagelsieb, G.; Emili, A. Global Functional Atlas Of Escherichia Coli Encompassing Previously Uncharacterized Proteins. *PLoS Biology*. **2009**, *7*, e96.

Iwanaga, N.; Ide, K.; Nagashima, T.; Tomita, T.; Agari, Y.; Shinkai, A.; Kuramitsu, S.; Okada-Hatakeyema, M.; Kuzuyama, T.; Nishiyama, M. Genome-Wide Comprehensive Analysis

of Transcriptional Regulation by ArgR in Thermus Thermophilus. *Extremophiles*. **2014**, *18*, 995–1008.

Jain, S. S.; Tullius, T. D. Footprinting Protein–DNA Complexes Using the Hydroxyl Radical. *Nature Protocols*. **2008**, *3*, 1092–1100.

Jolma, A.; Yan, J.; Whitington, T.; Toivonen, J.; Nitta, K. R.; Rastas, P.; Morgunova, E.; Enge, M.; Taipale, M.; Wei, G.; Palin, K.; Vaquerizas, J. M.; Vincentelli, R.; Luscombe, N. M.; Hughes, T. R.; Lemaire, P.; Ukkonen, E.; Kivioja, T.; Taipale, J. DNA-Binding Specificities Of Human Transcription Factors. *Cell*. **2013**, *152*, 327–339.

Juo, Z. S.; Chiu, T. K.; Leiberman, P. M.; Baikalov, I.; Berk, A. J.; Dickerson, R. E. How Proteins Recognize The TATA Box. *Journal of Molecular Biology*. **1996**, *261*, 239–254.

Karuppiah, V.; Thistlewaithe, A.; Dajani, R.; Warwicker, J.; Derrick, J. P. Structure And Mechanism of the Bifunctional CinA Enzyme from Thermus Thermophilus. *Journal of Biological Chemistry*. **2014**, *289*, 33187–33197.

Kato, R.; Kuramitsu, S. RecA Protein From an Extremely Thermophilic Bacterium, Thermus Thermophilus HB8. *The Journal of Biochemistry*. **1993**, *114*, 926–929.

Kim, Y.; Babnigg, G.; Jedrzejczak, R.; Eschenfeldt, W. H.; Li, H.; Maltseva, N.; Hatzos-Skintges, C.; Gu, M.; Makowska-Grzyska, M.; Wu, R.; An, H.; Chhor, G.; Joachimiak, A. High-Throughput Protein Purification and Quality Assessment for Crystallization. *Methods*. **2011**, *55*, 12–28.

Kosinski, J.; Feder, M.; Bujnicki, J. M. The PD-(D/E)XK Superfamily Revisited: Identification of New Members among Proteins Involved in DNA Metabolism and Functional Predictions for Domains of (Hitherto) Unknown Function. *BMC Bioinformatics*. **2005**, *6*.

Kumar, L.; Awasthi, G.; Singh, B. Extremophiles: A Novel Source Of Industrially Important Enzymes. *Biotechnology(Faisalabad)*. **2011**, *10*, 121–135.

Laurens, N.; Rusling, D. A.; Pernstich, C.; Brouwer, I.; Halford, S. E.; Wuite, G. J. L. DNA Looping by FokI: the Impact of Twisting and Bending Rigidity on Protein-Induced Looping Dynamics.*Nucleic Acids Research*. **2012**, *40*, 4988–4997.

Lee, D. K.; Horikoshi, M.; Roeder, R. G. Interaction Of TFIID in the Minor Groove of the TATA Element. *Cell*. **1991**, *67*, 1241–1250.

Lohse, M. B.; Hernday, A. D.; Fordyce, P. M.; Noiman, L.; Sorrells, T. R.; Hanson-Smith, V.; Nobile, C. J.; Derisi, J. L.; Johnson, A. D. Identification And Characterization of a Previously Undescribed Family of Sequence-Specific DNA-Binding Domains. *Proceedings of the National Academy of Sciences*. **2013**, *110*, 7660–7665.

Luck, G.; Triebel, H.; Waring, M.; Zimmer, C. Conformation Dependent Binding of Netropsin and Distamycin to DNA and DNA Model Polymers. *Nucleic Acids Research*. **1974**, *1*, 503–530.

Lundin, S.; Jemt, A.; Terje-Hegge, F.; Foam, N.; Pettersson, E.; Kaller, M.; Wirta, V.; Lexow, P.; Lundeberg, J. Endonuclease Specificity And Sequence Dependence of Type IIS Restriction Enzymes. *PLos One*. **2015**, *10*, e0117059.

Magrane, M.; Consortium, U. UniProt Knowledgebase: a Hub of Integrated Protein Data. *Database*. **2011**, *2011*, bar009–bar009.

Mcdonald, J. H. Patterns Of Temperature Adaptation in Proteins from the Bacteria Deinococcus Radiodurans and Thermus Thermophilus. *Molecular Biology and Evolution*. **2001**, *18*, 741–749.

Medina-Rivera, A.; Abreu-Goodger, C.; Thomas-Chollier, M.; Salgado, H.; Collado-Vides, J.; Helden, J. V. Theoretical And Empirical Quality Assessment of Transcription Factor-Binding Motifs.*Nucleic Acids Research*. **2011**, *39*, 808–824.

Narumi, I.; Satoh, K.; Kikuchi, M.; Funayama, T.; Yanagisawa, T.; Kobayashi, Y.; Watanabe, H.; Yamamoto, K. The LexA Protein From Deinococcus Radiodurans Is Not Involved in RecA Induction Following Gamma Irradiation. *Journal of Bacteriology*. **2001**, *183*, 6951–6956.

Nikolov, D. B.; Chen, H.; Halay, E. D.; Hoffman, A.; Roeder, R. G.; Burley, S. K. Crystal Structure of a Human TATA Box-Binding Protein/TATA Element Complex. *Proceedings of the National Academy of Sciences*. **1996**, *93*, 4862–4867.

Nitta, K. R.; Jolma, A.; Yin, Y.; Morgunova, E.; Kivioja, T.; Akhtar, J.; Hens, K.; Toivonen, J.; Deplancke, B.; Furlong, E. E. M.; Taipale, J. Conservation Of Transcription Factor Binding Specificities across 600 Million Years of Bilateria Evolution. *eLife*. **2015**, *4*, e04837.

Nunes, O. C.; Manaia, C. M.; Da Costa, M. S.; Santos, H. Compatible Solutes In the Thermophilic Bacteria Rhodothermus Marinus and ‘‘Thermus Thermophilus.’’ *APPLIED AND ENVIRONMENTAL MICROBIOLOGY*. **1995**, *61*, 2351–2357.

Ogawa, N.; Biggin, M. D. High-Throughput SELEX Determination of DNA Sequences Bound by Transcription Factors in Vitro. *Methods in Molecular Biology*. **2012**, *786*, 51–63.

Ohtani, N.; Tomita, M.; Itaya, M. An Extreme Thermophile, Thermus Thermophilus, Is a Polyploid Bacterium. *Journal of Bacteriology*. **2010**, *192*, 5499–5505.

Ohtani, N.; Tomita, M.; Itaya, M. The Third Plasmid pVV8 from Thermus Thermophilus HB8: Isolation, Characterization, and Sequence Determination. *Extremophiles*. **2012**, *16*, 237–244.

Omelchenko, M. V.; Wolf, Y. I.; Gaidamakova, E. K.; Matrosova, V. Y.; Vasilenko, A.; Zhai, M.; Daly, M. J.; Koonin, E. V.; Makarova, K. S. Comparative Genomics of Thermus Thermophilus and Deinococcus Radiodurans: Divergent Routes of Adaptation to Thermophily and Radiation Resistance. *BMC Evolutionary Biology*. **2005**, *5*, 57.

Orenstein, Y.; Shamir, R. A Comparative Analysis of Transcription Factor Binding Models Learned from PBM, HT-SELEX and ChIP Data. *Nucleic Acids Research*. **2014**, *42*, e63.

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.

Ouellette, M. M.; Wright, W. E. Use Of Reiterative Selection for Defining Protein—Nucleic Acid Interactions. *Current Opinion in Biotechnology*. **1995**, *6*, 65–72.

Pan, Y.; Tsai, C.-J.; Ma, B.; Nussinov, R. Mechanisms Of Transcription Factor Selectivity. *Trends in Genetics*. **2010**, *26*, 75–83.

Pernstich, C.; Halford, S. E. Illuminating The Reaction Pathway of the FokI Restriction Endonuclease by Fluorescence Resonance Energy Transfer. *Nucleic Acids Research*. **2012**, *40*, 1203–1213.

Phan, A. T.; Kuryavyi, V.; Patel, D. J. DNA Architecture: from G to Z. *Current Opinion in Structural Biology*. **2006**, *16*, 288–298.

Pilch, D. S.; Polskar, N.; Gelfand, C. A.; Law, S. M.; Breslauer, K. J.; Baird, E. E.; Dervan, P. B. Binding Of a Hairpin Polyamide in the Minor Groove of DNA: Sequence-Specific Enthalpic Discrimination. *Proceedings of the National Academy of Sciences*. **1996**, *93*, 8306–8311.

Pingoud, A.; Wilson, G. G.; Wende, W. Type II Restriction Endonucleases--a Historical Perspective and More. *Nucleic Acids Research*. **2014**, *42*, 7489–7527.

Poux, S.; Magrane, M.; Arighi, C. N.; Bridge, A.; O'Donovan, C.; Laiho, K.; The Uniprot Consortium. Expert Curation in UniProtKB: a Case Study on Dealing with Conflicting and Erroneous Data. *Database*. **2014**, *2014*.

Preparing chemically competent cells. OpenWetWare RSS, http://openwetware.org/wiki/preparing_chemically_competent_cells (accessed Jul 2, 2015).

Quail, M.; Smith, M. E.; Coupland, P.; Otto, T. D.; Harris, S. R.; Connor, T. R.; Bertoni, A.; Swerdlow, H. P.; Gu, Y. A Tale of Three next Generation Sequencing Platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq Sequencers. *BMC Genomics*. **2012**, *13*, 341.

Ramos, J. L.; Martinez-Bueno, M.; Molina-Henares, A. J.; Teran, W.; Watanabe, K.; Zhang, X.; Gallegos, M. T.; Brennan, R.; Tobes, R. The TetR Family Of Transcriptional Repressors.*Microbiology and Molecular Biology Reviews*. **2005**, *69*, 326–356.

Reddy, T. B. K.; Thomas, A. D.; Stamatis, D.; Bertsch, J.; Isbandi, M.; Jansson, J.; Mallajosyula, J.; Pagani, I.; Lobos, E. A.; Kyrpides, N. C. The Genomes OnLine

Database (GOLD) v.5: a Metadata Management System Based on a Four Level (Meta)Genome Project Classification. *Nucleic Acids Research*. **2014**.

Roberts, R. J.; Vincze, T.; Posfai, J.; Macelis, D. REBASE__a Database for DNA Restriction and Modification:Enzymes, Genes, and Genomes. *Nucleic Acids Research*. **2014**, *43*, D298–D299.

Robinson, K.; McGuire, A. M.; Church, G. M. A Comprehensive Library of DNA-Binding Site Matrices for 55 Proteins Applied to the Complete Escherichia Coli K12 Genome. *Nucleic Acids Research*. **1998**, *284*, 241–254.

Rothberg, J. M.; Hinz, W.; Rearick, T. M.; Schultz, J.; Mileski, W.; Davey, M.; Leamon, J. H.; Johnson, K.; Milgrew, M. J.; Edwards, M.; Hoon, J.; Simons, J. F.; Marran, D.; Myers, J. W.; Davidson, J. F.; Branting, A.; Nobile, J. R.; Puc, B. P.; Light, D.; Clark, T. A.; Huber, M.; Branciforte, J. T.; Stoner, I. B.; Cawley, S. E.; Lyons, M.; Fu, Y.; Homer, N.; Sedova, M.; Miao, X.; Reed, B.; Sabina, J.; Feierstein, E.; Schorn, M.; Alanjary, M.; Dimalanta, E.; Dressman, D.; Kasinskas, R.; Sokolsky, T.; Fidanza, J. A.; Namsaraev, E.; Mckernan, K. J.; Williams, A.; Roth, G. T.; Bustillo, J. An Integrated Semiconductor Device Enabling Non-Optical Genome Sequencing. *Nature*. **2011**, *475*, 348–352.

Roulet, E.; Busso, S.; Camargo, A. A.; Simpson, A. J.; Mermod, N.; Bucher, P. High-Throughput SELEX–SAGE Method for Quantitative Modeling of Transcription-Factor Binding Sites. *Nature Biotechnology*. **2002**, *20*, 831–835.

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*, 4388–95.

Salgado, H.; Peralta-Gil, M.; Gama-Castro, S.; Santos-Zavaleta, A.; Muniz-Rascado, L.; Garcia-Sotelo, J. S.; Weiss, V.; Solano-Lira, H.; Martinez-Flores, I.; Medina-Rivera, A.; Salgado-Osorio, G.; Alquicira-Hernandez, S.; Alquicira-Hernandez, K.; Lopez-Fuentes, A.; Porron-Sotelo, L.; Huerta, A. M.; Bonavides-Martinez, C.; Balderas-Martinez, Y. I.; Pannier, L.; Olvera, M.; Labastida, A.; Jimenez-Jacinto, V.; Vega-Alvarado, L.; Moral-

Chavez, V. D.; Hernandez-Alvarez, A.; Morett, E.; Collado-Vides, J. RegulonDB v8.0: Omics Data Sets, Evolutionary Conservation, Regulatory Phrases, Cross-Validated Gold Standards and More. *Nucleic Acids Research*. **2012**, *41*, D203–D213.

Sanchez, R.; Roovers, M.; Glansdorff, N. Organization And Expression of a Thermus Thermophilus Arginine Cluster: Presence of Unidentified Open Reading Frames and Absence of a Shine-Dalgarno Sequence. *Journal of Bacteriology*. **2000**, *182*, 5911–5915.

Sasnauskas, G.; Halford, S. E.; Siksnys, V. How The BfiI Restriction Enzyme Uses One Active Site to Cut Two DNA Strands. *Proceedings of the National Academy of Sciences*. **2003**, *100*, 6410–6415.

Seckbach, J.; Oren, A.; Stan-Lotter, H. *Polyextremophiles: Life under Multiple Forms of Stress*.

Sevostyanova, A.; Artsimovitch, I. Functional Analysis of Thermus Thermophilus Transcription Factor NusG. *Nucleic Acids Research*. **2010**, *38*, 7432–7445.

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.

Sobell, H. M. Actinomycin And DNA Transcription. *Proceedings of the National Academy of Sciences*. **1985**, *82*, 5328–5331.

Stan-Lotter, H.; Fendrihan, S. *Adaption Of Microbial Life to Environmental Extremes: Novel Research Results and Application*; Springer: Wien, 2012.

Starr, D.; Hawley, D. K. TFIID Binds in the Minor Groove of the TATA Box. *Cell*. **1991**, *67*, 1231–1240.

Stormo, G. D.; Zhao, Y. Determining The Specificity of Protein–DNA Interactions. *Nature Reviews Genetics*. **2010**.

Structual-Biological Whole Cell Project. Structual-Biological Whole Cell Project, http://www.thermus.org/e_index.htm (accessed Apr 12, 2015).

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.

Szybalski, W.; Kim, S. C.; Hasan, N.; Podhajska, A. J. Class-IIS Restriction Enzymes — a Review. *Gene*. **1991**, *100*, 13–26.

Takayama, G.; Kosuge, T.; Maseda, H.; Nakamura, A.; Hoshino, T. Nucleotide Sequence of the Cryptic Plasmid pTT8 from Thermus Thermophilus HB8 and Isolation and Characterization of Its High-Copy-Number Mutant. *Plasmid*. **2004**, *51*, 227–237.

Taylor, A.; Webster, K. A.; Gustafson, T. A.; Kedes, L. The Anti-Cancer Agent Distamycin A Displaces Essential Transcription Factors and Selectively Inhibits Myogenic Differentiation.*Molecular and Cellular Biochemistry*. **1997**, *169*, 61–72.

Thieffry, D.; Huerta, A. M.; Pérez-Rueda, E.; Collado-Vides, J. From Specific Gene Regulation to Genomic Networks: a Global Analysis of Transcriptional Regulation in Escherichia Coli.*BioEssays*. **1998**, *20*, 433–440.

Tonthat, N. K.; Arold, S. T.; Pickering, B. F.; Dyke, M. W. V.; Liang, S.; Lu, Y.; Beuria, T. K.; Margolin, W.; Schumacher, M. A. Molecular Mechanism by Which the Nucleoid Occlusion Factor, SlmA, Keeps Cytokinesis in Check. *The EMBO Journal*. **2010**, *30*, 154–164.

Tuerk, C.; Gold, L. Systematic Evolution of Ligands by Exponential Enrichment: RNA Ligands to Bacteriophage T4 DNA Polymerase. *Science*. **1990**, *249*, 505–510.

Tullius, T. D.; Greenbaum, J. A. Mapping Nucleic Acid Structure by Hydroxyl Radical Cleavage. *Current Opinion in Chemical Biology*. **2005**, *9*, 127–134.

The UniProt Consortium. UniProt: a Hub for Protein Information. *Nucleic Acids Research*. **2015**, *43*, D204–D212.

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.

Van Dyke, M. W.; Dervan, P. Footprinting With MPE{Middle Dot}Fe(II). Complementary-Strand Analyses of Distamycin- and Actinomycin-Binding Sites on Heterogeneous DNA. *Cold Spring Harbor Symposia on Quantitative Biology*. **1983**, *47*, 347–353.

Van Dyke, M. W. .; Hertzberg, R. P.; Dervan, P. B. Map Of Distamycin, Netropsin, and Actinomycin Binding Sites on Heterogeneous DNA: DNA Cleavage-Inhibition Patterns with Methidiumpropyl-EDTA.Fe(II). *Proceedings of the National Academy of Sciences*. **1982**, *79*, 5470–5474.

Vicente, M.; Mingorance, J. Microbial Evolution: the Genome, the Regulome and Beyond. *Environmental Microbiology*. **2008**, *10*, 1663–1667.

Wah, D. A.; Hirsch, J. A.; Dorner, L. F.; Schildkraut, I.; Aggarwal, A. K. Structure Of the Multimodular Endonuclease FokI Bound to DNA. *Nature*. **1997**, *388*, 97–100.

Wright, W. E.; Binder, M.; Funk, W. Cyclic Amplification and Selection of Targets (CASTing) for the Myogenic Consensus Site. *Molecular and Cellular Biology*. **1991**, *11*, 4104–4110.

Wright, W. E.; Funk, W. D. CASTing For Multicomponent DNA-Binding Complexes. *Trends in Biochemical Sciences*. **1993**, *18*, 77–80.

Yamamoto, N.; Nakahigashi, K.; Nakamichi, T.; Yoshino, M.; Takai, Y.; Touda, Y.; Furubayashi, A.; Kinjyo, S.; Dose, H.; Hasegawa, M.; Datsenko, K. A.; Nakayashiki, T.; Tomita, M.; Wanner, B. L.; Mori, H. Update On the Keio Collection of Escherichia Coli Single-Gene Deletion Mutants. *Molecular Systems Biology*. **2009**, *5*.

Yindeeyoungyeon, W.; Schell, M. A. Footprinting With an Automated Capillary DNA Sequencer. *Biotehniques*. **2000**, *29*, 1034–6, 1038, 1040–1.

Zhao, Y.; Granas, D.; Stormo, G. D. Inferring Binding Energies From Selected Binding Sites. *PLoS Computational Biology*. **2009**, *5*, e1000590.

Zhou, J.; Richardson, A. J.; Rudd, K. E. EcoGene-RefSeq: EcoGene Tools Applied to the RefSeq Prokaryotic Genomes. *Bioinformatics*. **2013**, *29*, 1917–1918.

Zhou, J.; Rudd, K. E. EcoGene 3.0. *Nucleic Acids Research*. **2012**, *41*, D613–D624.

Zimmer, C.; Wähnert, U. Nonintercalating DNA-Binding Ligands: Specificity of the Interaction and Their Use as Tools in Biophysical, Biochemical and Biological Investigations of the Genetic Material. *Progress in Biophysics and Molecular Biology*. **1986**, *47*, 31–112.

Zimmermann, B.; Bilusic, I.; Lorenz, C.; Schroeder, R. Genomic SELEX: A Discovery Tool for Genomic Aptamers. *Methods*. **2010**, *52*, 125–132.