Title
Regulation of transcription in pathogens, yeast and people

Author
Nelson, Christopher Steven

Publication Date
2012

Peer reviewed|Thesis/dissertation

Regulation of transcription in pathogens, yeast and people

by

Christopher Steven Nelson

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

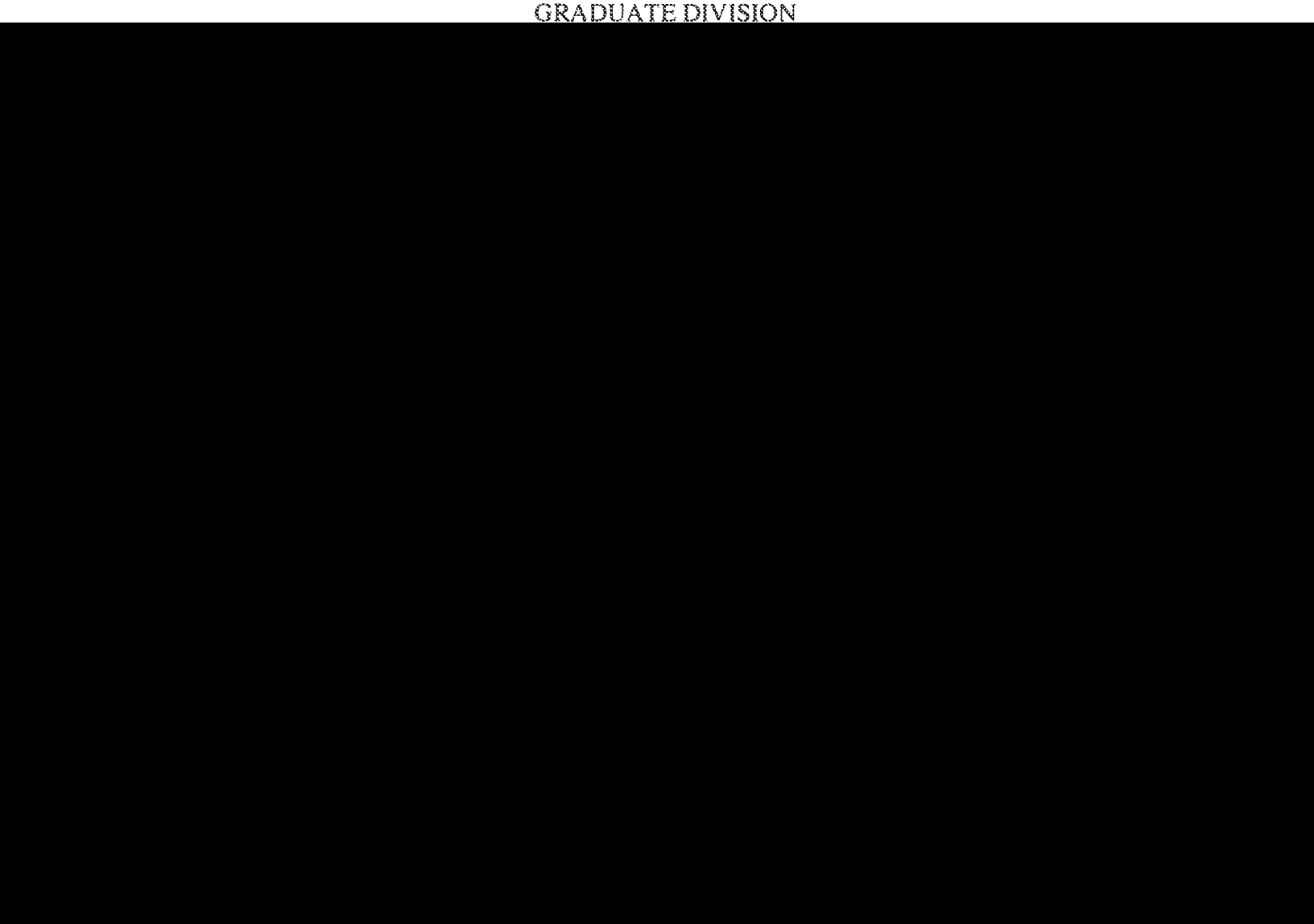
DOCTOR OF PHILOSOPHY

in

Biomedical Sciences

in the

GRADUATE DIVISION



Dedicated to Kemi Mustapha

Table of Contents

Regulation of transcription in pathogens, yeast and people

by

Christopher Steven Nelson

Abstract

In the post genomic and high throughput era we have a wealth of sequence and expression data, yet we are still learning to understand the punctuation and syntax of the genome. This thesis presents case studies in gene regulation for a range of organisms, addressing whole transcriptome pattern changes in *Plasmodium falciparum* and *Bartonella quintana*, and then more focused descriptions of the binding sites of transcription factors from Mycobacterium tuberculosis, yeast, chimps, and humans. Early on in my thesis, I studied the immunology of *Plasmodium* infections. Later I profiled the transcriptomes of *Plasmodium* parasites after exposure to artemisinin, currently the front-line standard of care, and uncovered evidence for a developmental stall. The parasite was not previously thought to undergo any cell cycle arrest, and this may explain observed clinical recrudescence of infection after artemisinin monotherapy treatment and could represent a means for the parasite to adapt to drug pressure and select for resistance. After these studies I turned to a technique, called MITOMI2.0 (for mechanical trapping of molecular interactions), a means of measuring the specificity and energetic affinities of DNA binding proteins. The primary data produced with a lab-on-a-chip is the amount of DNA binding to a given protein, over a library of DNA sequences. We improved made the

technique more robust and were able to apply it to several systems. We found that the yeast unfolded protein response factor Hac1 bound two DNA sites with distinct sequences. For another yeast stress regulator, Msn2, we measured its absolute affinity for regulated sites in target promoters and confirmed that it was a low affinity DNA binder, capable of a linear induction of its targets, as opposed to a frequently observed more binary induction response. ChIP-seq and MITOMI analysis revealed that FOXP2 the best-studied example of a protein involved in the development of human language, has had a conserved binding site preference, yet the complement of available binding sites has changed in humans. We confirmed the sequence specificity of a *Mycobacterium tuberculosis* factor that controls virulence in macrophage infection, and studied the way the protein interacts with DNA both as a dimer and a monomer.

List of Tables

List of Figures or Illustrations

Chapter 1…….

Introduction

Around the turn of the millennium the maturation of genomics made monitoring whole transcriptome dynamics feasible and greatly expanded the amount we knew about the noncoding portions of the genome. This opened new avenues of studying how transcriptional regulators read the punctuation marks in our genome to coordinate transcriptome dynamics. This thesis presents case studies in gene regulation for a range of organisms, describing whole transcriptome dynamics in *Bartonella quintana* and *Plasmodium falciparum*, and describing the binding site preferences of transcription factors from *Mycobacterium tuberculosis*, yeast, chimpanzees, and humans.

Early in my thesis I was primarily interested in transcriptional regulation at the genomic level in the context of infectious organisms. Along with Charlie Kim, I studied the induction of interferon gene cascades in *Plasmodium chabaudi* infection, a mouse model of malaria. After that I made custom gene expression microarrays for studying responses to drug pressure or host temperature whole transcriptome pattern changes in the pathogens *Plasmodium falciparum* and *Bartonella quintana.* (These transcriptome studies were performed as deep-sequencing technology was being adopted, and will probably be the last expression microarray studies from the DeRisi lab, which had microarray technology as its foundational platform.) The general principle that we learned is that pathogens frequently exhibit large genome-wide transcriptional changes in response to the host niche

After studying changes in whole transcriptome patterns, I became interested in the some of the molecular drivers of these patterns, sequence specific DNA-binding transcription factors. After working to improve the MITOMI2.0 technique, along with Polly Fordyce, I studied how transcription factors choose the DNA that they bind and how different codes lead to different strengths of binding. It became apparent to me that transcription factors present a wide spectrum of behaviors and specificities. For example, we found that the structurally simple transcription factor Hac1 has at least two non-overlapping binding site preferences, we found the stress regulator Msn2 coordinates a graded induction of the environmental stress response through a low-affinity short DNA motif, and we found that the FOXP2 transcription factor involved in the development of human speech has conserved it's binding specificity but the underlying *cis* binding sequences. Below are fuller summaries of each of these projects.

Genome-wide transcriptome changes in response to infection

*Splenic Red Pulp Macrophages Produce Type I Interferons as Early Sentinels of Malaria Infection but are Dispensable for Control***.** Beginning my interest in pathogenesis and transcription regulation I assisted Charlie Kim in studying mouse models of malaria immunity. Malaria continues to be the most important parasitic infection in man with over 225 million new infections and just under one million deaths annually (WHO 2010). The basis for natural clearance of malaria infection and long-term immunity to malaria is very poorly understood. Accordingly, current malaria vaccines have disappointing efficacy and their

effects tend to be very short-lived. Currently the RTS,S vaccine is the closest candidate vaccine to adoption, yet it has only 30% efficacy for one year after a three dose vaccination schedule (NEJM 2012). Hopes for eradication of this malaria would be greatly aided by a better understanding of how the host can detect the parasite and deploy immune defenses.

Type I interferons (T1IFNs) are among the earliest cytokines produced during infections due to their direct regulation by innate immune signaling pathways. Reports have suggested that T1IFNs are produced during malaria infection, but little is known about the *in vivo* cellular origins of T1IFNs or their role in protection. We employed experimental malaria infection in mice to study the source and mechanism of induction of T1IFNs. We found that plasmacytoid dendritic cells, and splenic red pulp macrophages (RPMs) can generate significant quantities of T1IFNs in response to *P. chabaudi* infection in a TLR9-, MYD88-, and IRF7-dependent manner. (TLR9, MYD88, and IRF7 form a potential innate immune signaling pathway for the sensing of parasite nucleic acids or hemazoin crystals.) Furthermore, T1IFNs regulate expression of interferon-stimulated genes redundantly with IFNG, and this translates into redundancy in resistance to experimental malaria infection. Despite their role in sensing and promoting immune responses to infection, we observe that RPMs are dispensable for control of parasitemia. Our results reveal that RPMs are early sentinels of malaria infection, but that effector mechanisms previously attributed to RPMs are not essential for control.

Bartonella quintana deploys host and vector temperature-specific transcriptomes. Using some of the transcriptome profiling techniques that I learned in from the experimental malaria infections, I turned to studying bacterial pathogen *Bartonella quintana. B. quintana* has adapted to both the human host and body louse vector, producing persistent infection with high titer bacterial loads in both the host (up to 10^5 CFU / ml) and vector (over 10^8 CFU / ml). The *B. quintana* is passed between humans by body lice. Using a novel custom microarray platform, we analyzed bacterial transcription at temperatures corresponding to the host (37°C) and vector (28°C), to probe for temperature-specific and growth phase-specific transcriptomes.

We observed that transcription of 7% (93 genes) of the *B. quintana* genome is modified in response to growth phase, and that 5% (68 genes) of the genome is temperature-responsive. Among these changes were the induction of known *B. quintana* virulence genes and several previously unannotated genes in response to temperature and growth phase changes. Hemin binding proteins, secretion systems, and genes for invasion and cell attachment were prominent among the differentially-regulated *B. quintana* transcriptional responses. This study represents the first analysis of global transcriptional responses by *B. quintana* and provides insight into the niche-specific gene expression involved in the transition of *B. quintana* between the human host and body louse vector.

Artemisinin induces transcriptome arrest in drug susceptible and Plasmodium falciparum. Returning to malaria, I led a project that investigated a novel

phenotype in development of the *Plasmodium falciparum*. When treated with artemisinins. The artemisinin class of drugs is the most important tool for treating multi-drug resistant malaria globally, yet is associated with frequent recrudescence of disease unless used in combination with another established antimalarial drug. Furthermore, recent evidence suggests that clinical resistance to artemisinin has emerged in Cambodia. Despite the importance of the problem there are no validated molecular markers of resistance.

In collaboration with the Dennis Kyle's lab at the University of South Florida, we found that artemisinin drugs induce dormancy in the earliest stage of development in the erythrocyte and this is associated with transcriptional arrest in both drug susceptible and resistant parasites clones. In this study, we conducted temporal parasite morphology and transcriptome analysis experiments of naive and drug-selected *Plasmodium falciparum* parasite strains to assess the effects of dihydroartemisinin (DHA) on parasite development. Our results showed that following artemisinin treatment, both sensitive and resistant ring-stage parasites pause in a dormant state characterized by small rounded morphology and a transcriptional state of 8-11 h post-invasion rings. Transcriptional analysis identified genes that are differentially expressed during DHA-induced transcriptome arrest, and between the artemisinin sensitive and resistant clones. These data provided the first evidence for a long-lasting transcriptome arrest in response to an antimalarial drug and have implications for recrudescence following treatment with artemisinins and for the study of emerging artemisinin resistance in the field.

After the Artemisinin studies, I wanted to see whether the huge developmental dynamics in *Plasmodium* gene expression could be tied to a core group of transcription factors. I attempted to use genetic techniques to disrupt putative stage-specific by double homologous recombination. I transfected parasites with knockout vector constructs for 10 putative *Plasmodium* transcription factors: PFF1100c, PF11_0347, PF11_0404, PF13_0097, PFL0815w, PF10_0075, PFL1075w, PFF0200c, PF11_0442, and PF11_0477. Unfortunately, instead of integrating the knockout vector, the parasites disabled its negative-selection marker, which would otherwise force the gene knockout event. This result was discouraging because it came after many months of daily cell culture and it was inconclusive, only suggestiing that the targeted genes were essential. Luckily, I had started on a parallel *in vitro* means to studying transcription factors, MITOMI, which ended up being fruitful in other organisms.

Using MITOMI to study the diversity of transcription factor behavior

*MITOMI 2.0 development***.** Mechanically Induced Trapping of Mechanical interactions (MITOMI) was initially developed by Sebastian Maerkl in the Quake laboratory to study the energetic effects of binding a transcription factor against a known binding site. These microfluidic devices measure the binding of transcription factors to different DNA sequences and are controlled by pressurized valves. Subsequently, Polly Fordyce extended the platform to enable motif discovery for transcription factors without known binding sites, calling the

extended technique MITOMI2.0. These new devices were more than 6 times larger than the original devices to incorporate a larger random DNA library.

Besides the DNA library problems, the expansion of the MITOMI platform had resulted in robustness problems. The original Quake Lab MITOMI microfluidic device was already at the leading edge of density and complexity for a silicone rubber labs-on-a-chip. Overall, some early MITOMI 2.0 devices performed properly and generated data less than 30% of the time. The valves on the device would often fail to close fully, rendering them inoperable, or mixing reagents when inappropriate.

Through trial and error, we tried different valve design concepts, aiming at increased closure tightness and overall device robustness. We improved device performance by building larger footprint valves and wiring valves in parallel to avoid inoperability caused by a single defect. Defects are often caused in microfluidic devices by small pieces of dust. Given the complexity of our devices and their vulnerability to defects, we moved our whole fabrication process into a local clean room and had increasing success. Additionally, we spaced out the high-pressure control lines from each other to avoid the chance that a piece of dust or missing feature could short-circuit the device. Probably the single idea that made the most difference was the reduction in the footprint and complexity of the devices. This was enabled by the reduction of the DNA library size. Once we arrived at devices that performed properly more than 90% of the time we were ready to return from engineering to biology. With our improved MITOMI2.0 devices, we first turned to a deceptively simple factor Hac1.

The Basic Leucine Zipper Transcription Factor Hac1 Binds DNA In Two Distinct Modes as Revealed by Microfluidic Analyses. In collaboration with Polly Fordyce and the Walter lab, we used Hac1, a well-characterized basic leucine zipper (bZIP) transcription factor involved in the Unfolded Protein Response (UPR), as a model to investigate interactions between bZIP transcription factors and their target sites. During the UPR, the accumulation of unfolded proteins leads to unconventional splicing and subsequent translation of *HAC1* mRNA, followed by transcription of UPR target genes. Initial candidate-based approaches identified a canonical *cis*-acting Unfolded Protein Response Element (UPRE-1) within target gene promoters; however, subsequent studies identified a large set of Hac1 target genes lacking this UPRE-1 and containing a different motif (UPRE-2). Using a combination of unbiased and directed microfluidic DNA binding assays, we established that Hac1 binds in two distinct modes: i) to short (6-7 bp) UPRE-2-like motifs, and ii) to significantly longer (11-13 bp) extended UPRE-1-like motifs. Using a library of Hac1 mutants, we demonstrate that a region of extended homology N-terminal to the basic DNA binding domain is required for this dual site recognition. These results establish Hac1 as the first bZIP transcription factor known to adopt more than one binding mode and unify previously conflicting and discrepant observations of Hac1 function into a cohesive model of UPR target gene activation. Our results also suggest that even structurally simple transcription factors can recognize multiple divergent

target sites of very different lengths, potentially enriching their downstream target repertoire.

Microfluidic affinity and ChIP-seq analyses converge on a deeply conserved FOXP2 binding motif that enables the detection of evolutionarily novel regulatory targets. FOXP2 has generated broad interest in the literature because it has a focal phenotype on the development of language, with evidence of recent evolution. People with mutations in the DNA binding domain of FOXP2 in humans are of average intelligence but have jerky, nonfluent speech with poor syntax and recall of words. In addition, phylogenetic and human population studies have suggested that the human FOXP2 gene sequence has undergone recent evolutionary selection.

There is still a gulf in our understanding of the functional properties of FOXP2 and its role in the development of human language. Previous studies have disagreed both about the identity of the FOXP2 target binding site and also with how FOXP2 binding interactions have changed during evolution. In particular, it has remained unclear whether changes in these binding interactions have been driven by *trans*-regulatory changes in protein binding specificity or *cis*-regulatory changes in target binding sites.

Using a combination of *in vitro* MITOMI microfluidic affinity assays and analysis of *in vivo* ChIP-seq data, we have profiled the binding site specificity of human and chimp FOXP2 orthologs and identified candidate target sites within the two lineages. We produced a single nucleotide resolution model of the binding

affinities to the FOXP2 binding sites and all its variants. In an independent ChIP-seq analysis approach, we found an identical optimal motif, confirming our *in vitro* results. This motif turns out to be distinct from all previously suggested FOXP2 binding sites but is consistent with family level consensus motifs. Furthermore, we found that the DNA binding site affinities have been largely conserved between humans. However, we find evidence in our ChIP-seq analysis that some genomic target sites are uniquely human. This suggests that the FOXP2 involvement in human language may be due to the evolution of the targeted sites. We provide candidate examples of such *cis* evolution in FOXP2 target genes with roles in neural development, in neural plasticity, and potentially in language.

Graded and Co-linear Regulation from the Stress Responsive Factor Msn2. To thrive in challenging and rapidly changing circumstances, cells tightly regulate the production of cytoprotective proteins. In the budding yeast *S. cerevisiae,* a wide range of stresses evoke the Environmental Stress Response (ESR), which results in the association of the homologous transcription factors Msn2/4 to stress responsive genes. In work with Jacob Stewart-Ornstein, we show that Msn2 activates gene expression in a graded and uniform manner across transcriptional targets. The stress response system generates a linear relationship between Msn2 activity and target gene expression through low affinity binding of Msn2 to target genes and an excess of binding sites relative to the quantity of Msn2 protein. These features provide a simple and general

mechanism for co-linear activation of target genes, allowing proportionate response to different magnitudes of stress and maintaining stoichiometry within the Msn2/4-responsive program across a wide range of conditions.

MITOMI analysis of the tuberculosis virulence regulator EspR as a monomer and dimer. Due to our mutual interest in pathogenesis and gene regulation, I collaborated with the Cox lab at UCSF in studying the *Mycobacterium tuberculosis* virulence regulator EspR. *Mycobacterium tuberculosis*, of course, is a disease of historic and global impact, causing 14 million chronic infections and over a million deaths every year (WHO 2009). It is incredibly successful in invading and replicates within lung macrophages, while evading innate and acquired immunity. Once inside its target cells, *M. tuberculosis,* secrete effector proteins to repurpose the endocytic valcuole of macrophages to suit bacterial growth. The Cox lab has a long-standing interest in such secretion systems in tuberculosis.

The secretion system ESX-1 avoids long term immune surveillance by only turning itself on within the cell and for short burst of time when appropriate. The transcriptional regulator EspR accomplishes this regulation. EspR appears to be an atypical helix-turn helix (HTH) DNA binding protein, adopting a unique DNA binding mode. Additionally, EspR has been suggested to interact with DNA in a sequence-nonspecific mode, as ChIP-seq experiments suggest that it can bind wide areas of the genome in clusters. My MITOMI results suggest that EspR binding *in vitro* is indeed sequence-specific but has several degeneracies in its

short binding site, and that similar half-site specificities can be derived from both monomeric and dimeric forms of the protein. In conclusion, it seems best to consider EspR as a factor that may function either as a monomer or dimer with moderate specificity.

Chapter 2…….

Splenic Red Pulp Macrophages Produce Type I Interferons as Early Sentinels of Malaria Infection but are Dispensable for Control

Charles C. Kim, Christopher S. Nelson, Emily B. Wilson, Baidong Hou, Anthony L. DeFranco, Joseph L. DeRisi

This chapter is a reprint from the following reference:
Kim CC, Nelson CS, Wilson EB, Hou B, Defranco AL, Derisi JL. Splenic red pulp macrophages produce type I interferons as early sentinels of malaria infection but are dispensable for control. PLoS One. 2012;7(10):e48126.

Author contributions:

Charles Kim bred most of the mice strains and performed all expression analyses, and conceived and designed the experiments. Christopher Nelson performed long-course infection trials. Emily Wilson assisted long-course infection trials. Anthony DeFranco conceived and designed experiments and contributed mice strains. Baidong Hou contributed a mouse strain. Joseph DeRisi analyzed data, conceived and designed experiments, and helped write the paper. Joseph L. DeRisi, Thesis Advisor

Abstract

Type I interferons (T1IFNs) are among the earliest cytokines produced during infections due to their direct regulation by innate immune signaling pathways. Reports have suggested that T1IFNs are produced during malaria infection, but little is known about the *in vivo* cellular origins of T1IFNs or their role in protection. We have found that in addition to plasmacytoid dendritic cells, splenic red pulp macrophages (RPMs) can generate significant quantities of T1IFNs in response to *P. chabaudi* infection in a TLR9-, MYD88-, and IRF7-dependent manner. Furthermore, T1IFNs regulate expression of interferon-stimulated genes redundantly with IFNG, and this translates into redundancy in resistance to experimental malaria infection. Despite their role in sensing and promoting immune responses to infection, we observe that RPMs are dispensable for control of parasitemia. Our results reveal that RPMs are early sentinels of malaria infection, but that effector mechanisms previously attributed to RPMs are not essential for control.

Introduction

Early recognition of infection by innate immune defenses initiates a complex cascade of intra- and intercellular signaling events that ultimately leads to the generation of a systemic immune response. Although detailed analysis of early innate immune events is underway for model organisms such as *Listeria* [1], relatively little is understood about early detection and responses to *Plasmodium sp.,* the leading parasitic cause of infectious mortality and morbidity in the world. This is despite growing evidence that innate immune responses,

particularly of monocytes and macrophages, play a vital role in the control of malaria infection. For example, inflammatory monocytes contribute to elimination of parasites in *P. chabaudi* infection [2], and in humans, a subset of peripheral monocytes is associated with control of infection in *ex vivo* assays [3]. Additionally, adoptive transfer of a recently discovered progenitor cell that primarily generates monocytes enhances clearance of malaria infection [4]. In contrast, B cells are required for elimination of persistent infection but are dispensable for control of the primary parasitemia [5–8]. Similarly, $CD8^+$ T cells are not essential for control of blood stage infection [9]. The dispensability of the major effector arms of adaptive immunity highlights the importance of innate mechanisms of anti-parasitic recognition and clearance.

Detection of the offending organism is the critical first step in activating innate immune mechanisms. Many microbes are recognized by innate immune sensors such as toll-like receptors (TLRs), cytosolic nucleic acid sensors such as RIG-I and MDA5, and NOD-like receptors, which can activate downstream production of immunomodulatory cytokines such as the type I interferons alpha and beta (T1IFNs, IFNA, IFNB), tumor necrosis factor (TNF), and interleukin 12 (IL12). In the case of malaria, TLR9 has emerged as a major sensor of infection, although the identity of the ligand remains controversial [4,10–13]. These studies were conducted using *in vitro*-differentiated plasmacytoid dendritic cells (pDCs), suggesting that pDCs may play a role in *in vivo* recognition of *Plasmodium* infection. This was recently demonstrated to be the case in a report of TLR9-dependent expression of *Ifna* in pDCs during *P. chabuadi* infection of mice [14].

However, it is well known that other innate leukocyte populations such as conventional dendritic cells (cDCs) and macrophages also express and signal through TLR9, but the role of these innate leukocyte subsets in recognition of malaria infection remains largely unexplored.

Although it is clear that detection of malaria infection occurs through TLRs and likely also by other innate immune receptors, the mechanisms through which innate cells contribute to defense against *Plasmodium* parasites are poorly characterized. During viral and bacterial infections, signaling through TLRs and other innate sensing pathways frequently results in the immediate downstream production of cytokines such as T1IFNs. With regard to malaria, *Plasmodium* ligands have been reported to stimulate T1IFN production in *in vitro* systems [10,13,15], experimentally infected mice [14], and *Plasmodium*-infected individuals [10,16]. However, in contrast to IFNG, which has been shown to be an important activator of anti-malarial mechanisms, the role of T1IFNs in protection against malaria infection is not well characterized.

In order to address these gaps in our knowledge, we conducted a systematic investigation of T1IFN production during malaria infection using the *P. chabaudi* model of uncomplicated malaria. Here we present evidence that in addition to pDCs, splenic red pulp macrophages (RPMs) are an important contributor to systemic T1IFN during early malaria infection. Additionally, we have found that T1IFNs regulate gene expression and contribute to control of infection in a manner that is largely redundant with IFNG signaling. However, despite the role of RPMs in T1IFN production, mice lacking RPMs exhibit no

deficiencies in their ability to control infection. Our findings demonstrate that T1IFNs play an important immunomodulatory role during *in vivo* malaria infection and provide us with a basic understanding of the molecular and cellular machinery involved in innate immune recognition of malaria parasites. We also demonstrate that RPMs are not essential for control of infection despite their role in early sensing of infection and their key location in contact with circulating parasites.

Results

T1IFNs and IFNG mediate the early inflammatory response to Plasmodium infection

We previously reported that genes stimulated as a result of interferon signaling constitute the most extensive gene expression module during the early whole blood response of mice to *P. chabaudi* [17]. In order to identify a highly reproducible signature of early gene expression, we conducted multiple independent gene expression profiling experiments of whole blood of mice infected or mock-infected with *P. chabaudi* at 24 h post-infection. Statistical analysis of the two groups revealed a set of 117 probes (103 unique genes) that were reproducibly increased in relative abundance at 24 h after *P. chabaudi* infection (Supplementary Table 1). As previously observed, these genes were significantly enriched for known interferon-stimulated genes (ISGs; PANTHER biological process "response to interferon-gamma" $p = 10^{-9}$), including classical markers of interferon signaling such as *Cxcl10*, *Il6*, and multiple members of the

Gbp, *Ifi*, *Ifit*, *Oas,* and *Slfn* gene families (representative genes shown in Fig. 1; complete list available in Supplementary Table 1).

Members of the two well-characterized classes of inflammatory interferons, T1IFNs and type II interferon (the sole member being interferon gamma, IFNG) can stimulate cells to induce transcription of ISGs. In order to assess the role of T1IFNs and IFNG in ISG induction in response to *P. chabaudi*, we examined whole blood gene expression signatures in mice deficient in components required for T1IFNs and IFNG signaling. In *Ifnar1*$^{-/-}$ mice (deficient in the receptor for T1IFNs), we observed that ISG expression was still induced in response to *P. chabaudi* infection, suggesting that IFNG signaling was a significant mediator of the ISG response. Similarly, *P. chabaudi* infection of *Ifngr1*$^{-/-}$ mice (deficient in IFNG receptor) also resulted in increased ISG transcript abundance compared to mock-infected animals, implying that T1IFNs signaling was also contributing to ISG expression during the early response to infection. To determine whether these genes were being induced in a redundant manner, we generated mice doubly deficient in both interferon receptors, and also examined mice deficient in the downstream transcription factor STAT1, which is required for both T1IFN and IFNG signaling. The ISG response in both *Ifnar1*$^{-/-}$ *Ifngr1*$^{-/-}$ and *Stat1*$^{-/-}$ animals was completely abolished, demonstrating that stimulation of ISG expression was occurring as a result of redundancy in the classical interferon signaling pathways and not a result of signaling through other pathways.

Although both *Ifnar1*$^{-/-}$ and *Ifngr1*$^{-/-}$ animals were capable of mounting an ISG response, the magnitude of the response in wild type animals appeared to

be greater than in either of the immunodeficient strains (41% and 32% average reductions in fold induction by *P. chabaudi* in *Ifnar1*$^{-/-}$ and *Ifngr1*$^{-/-}$ mice, respectively; Fig. S1A). We therefore assessed whether the magnitude of the responses to T1IFNs and IFNG was independent (additive) or redundant (sub-additive). We observe that the sum of the magnitudes of the ISG response in the *Ifnar1*$^{-/-}$ and *Ifngr1*$^{-/-}$ animals was on average greater than the magnitude of the wild type ISG response (slope = 0.7; Fig. S1B), indicating that the T1IFNs and IFNG pathways are inducing the ISG response in a partially redundant manner. Additionally, some redundancy is exhibited even by ISGs that show a degree of preferential induction by T1IFNs or IFNG (Fig. S1C). Although T1IFNs and IFNG are generally thought to mediate different aspects of immune activation, these results demonstrate that at least in the context of early malaria infection, the majority of genes regulated by one type of interferon are also regulated by the other.

In order to directly demonstrate T1IFN production, we performed quantitative reverse transcription PCR (qRT-PCR) to estimate relative transcript abundance for *Ifna* and *Ifnb* in the spleens of mice infected with *P. chabaudi*. Examination of splenic transcripts every 3 h for the first 30 h post-infection revealed that both *Ifna* and *Ifnb* transcripts, as well as *Ifng*, exhibited a peak of increased abundance centered around 24 h (Fig. 2A). Upon return to baseline levels, splenic T1IFN transcripts were not induced again within the first three days of infection (measured in 6 h intervals after 30 h). Detection of elevated *Ifna* and *Ifnb* in spleens of infected animals at 24 h post-infection was highly

reproducible across independent experiments (Fig. 2B), and IFNA and IFNB were reproducibly detected in the plasma of infected animals (Fig. 2C). Together, these findings provide evidence that a burst of T1IFNs is produced during the early response to *P. chabaudi* infection and contributes to induction of ISG expression.

T1IFNs and IFNG redundantly promote control of parasitemia

Our results show early production of T1IFNs during *P. chabaudi* infection, but the contribution of T1IFNs to the control of malaria parasite replication is not well characterized. The normal course of *P. chabaudi* infection in C57BL/6 mice develops as an exponentially increasing load of parasites in the blood that typically peaks at 7-10 days post-infection followed by control and resolution of the primary peak over the next 2-4 days (Fig. 3). A recent study reported a slight increase in the magnitude of peak *P. chabaudi* parasitemia in *Ifnar1*$^{-/-}$ mice, but resolution occurred with kinetics identical to wild type (129Sv) animals [14]. In contrast, we observed no significant differences in the magnitude or times to manifestation of any of the ascending, descending, or clearance phases of parasitemia in *Ifnar1*$^{-/-}$ animals as compared to infection of C57BL/6 mice (Fig. 3). The discrepancy between our findings and those of Voisine *et al.* could be a result of the different backgrounds used, since 129Sv mice produce higher levels of T1IFNs (Fig. S3 and [18]).

Although our results would appear to suggest that T1IFNs do not contribute to control of malaria infection, we considered the possibility that the

redundancy between T1IFNs and IFNG in the regulation of ISG expression could confer redundancy in control of infection. We therefore examined the course of parasitemia in *Ifngr1*$^{-/-}$ animals as compared to *Ifnar1*$^{-/-}$ *Ifngr1*$^{-/-}$ animals in order to assess the function of T1IFNs in the absence of IFNG signaling. We observed that *Ifngr1*$^{-/-}$ animals exhibited defects in their ability to resolve parasitemia as compared to wild type animals – although most animals controlled the primary and secondary peaks, peak parasitemias were higher in *Ifngr1*$^{-/-}$ animals, and a tertiary peak of parasitemia occurred in most animals (Fig. 3). Despite the increased severity of infection in *Ifngr1*$^{-/-}$ mice, parasites were controlled in all mice by 40 days post-infection. In contrast, *Ifnar1*$^{-/-}$ *Ifngr1*$^{-/-}$ animals exhibited mortality, multiple late peaks of high parasitemia, and an inability to completely clear parasites from the bloodstream within the duration of the 70 day study, indicating that T1IFNs and IFNG signaling exhibit redundancy in the regulation of anti-parasitic mechanisms that are essential to the control of malaria infection.

Plasmacytoid dendritic cells and red pulp macrophages produce T1IFNs in response to P. chabaudi

Previous *in vitro* studies have reported mixed results in production of IFNA by pDCs after stimulation with malaria ligands [10–13]. Another study recently reported *Ifna* expression in pDCs during *P. chabaudi* infection [14], but this observation was made at a time after the peak of C57BL/6 T1IFN production and did not assess other potential cellular sources. We therefore took an unbiased

approach to identify the cellular origins of T1IFN production in response to physiologically relevant stimuli during *in vivo* infection with *P. chabaudi*.

In order to achieve single-cell resolution of T1IFN expression, transgenic *Ifnb-Yfp* reporter mice [19] were mock- or *P. chabaudi*-infected and analyzed for splenic *Ifnb* expression by flow cytometry. Animals infected with *P. chabaudi* exhibited a small but highly reproducible population of YFP^+ cells, whereas no YFP^+ events were detected in any of the spleens of mock-infected animals (Fig. 4A). Lineage marker analysis of the YFP^+ populations demonstrated that approximately 75% of the YFP^+ events were $CD11c^{int}$ Siglec-H^+, consistent with markers of pDCs (Fig. 4B). In contrast, conventional dendritic cells (cDCs; $CD11c^{hi}$ Siglec H^-) and $CD11b^{hi}$ $F4/80^{int\text{-}hi}$ SSC^{lo} monocytes (Mono) constituted none of the YFP^+ events. Interestingly, a small but reproducible fraction (~15%) of the total YFP^+ events was $F4/80^{hi}$ $CD11b^{lo/-}$, consistent with markers of splenic RPMs. Similar frequencies of YFP^+ and lineage markers were observed using *Ifna6-Gfp* reporter mice (Fig. S2) [20]. Notably, the pDCs and RPMs together account for nearly all the YFP^+ and GFP^+ cells, indicating that, together, they are the major populations responsible for splenic T1IFN induction during *P. chabaudi* infection.

Because T1IFN can be produced at low levels by other cell types, we assessed whether pDCs and RPMs measurably contribute to systemic T1IFN levels. In order to examine the role of RPMs in T1IFN production, we employed *SpiC*$^{-/-}$ mice [21], which lack a transcription factor required for development of RPMs but not other myeloid populations (Fig. S3A). pDCs were depleted 18 h

pre-infection with *P. chabaudi* using the anti-mPDCA-1 antibody, which reproducibly depleted 85% of splenic pDCs with no impact on RPM frequency (Fig. S3B). After 24 h infection with *P. chabaudi*, *SpiC*$^{-/-}$ mice exhibited roughly half the splenic IFNB of *SpiC*$^{+/-}$mice (Fig. 4C), with similar results also observed in plasma (Fig. S3C-D). pDCs were also required for T1IFN production, with depletion resulting in over 80% reduction of splenic and plasma IFNB levels in *SpiC*$^{+/-}$ mice (Fig. 4C and S4D) and C57BL/6 mice (Fig. S4E). The absence of both populations resulted in over 90% reduction of splenic IFNB (Fig. 4C), with the residual levels likely reflecting incompletely depleted pDCs. Together with the reporter data, these results demonstrate that pDCs and RPMs are the primary sources of T1IFN during experimental malaria infection.

Red pulp macrophages are the primary source of splenic T1IFN transcripts during P. chabaudi *infection*

In order to corroborate our observations with the *Ifnb* reporter mice, we isolated the same splenic leukocyte subsets by FACS and assessed T1IFN transcriptional induction by qRT-PCR. Consistent with our observations in *Ifnb-Yfp* reporter mice, RPMs strongly induced both *Ifna* and *Ifnb* transcript post-infection with *P. chabaudi* (Fig. 5A). Similarly, microarray analysis of isolated RPMs from mock- and *P. chabaudi*-infected mice demonstrated induction of multiple members of the *Ifna* family along with a variety of other cytokines and chemokines including *Tnf, Il1b, Il6, Il10, Cxcl1,* and *Cxcl2* (Fig. S4A), and RPM-deficient mice exhibited decreased plasma levels of TNFA and IL12p70 (Fig.

S4B). These results demonstrate that RPMs activate a diverse repertoire of immunomodulatory products, including T1IFNs, during the early response to *P. chabaudi* infection.

In contrast to RPMs, splenic pDCs surprisingly did not demonstrate any significant induction of T1IFN transcript as measured by qRT-PCR. This was not a result of elevated baseline T1IFN transcript levels as observed in other studies (Fig. S5) [22]. We also did not detect elevated T1IFN in pDCs at earlier time points, which is consistent with our observation that splenic T1IFN transcript abundance peaks at 24 h with no other earlier peaks (Fig. 2A). Because YFP has a very long half life, we speculate that YFP^+ splenic pDCs have become activated at a slightly earlier time and in a different compartment before migrating to the spleen, consistent with their role as sentinel cells that can migrate to sites of inflammation [23]. The results also suggest that RPMs are the population primarily responsible for induction of T1IFN transcription in the spleen, although both RPMs and pDCs contribute to production of circulating plasma T1IFN.

To further assess the role of pDCs in T1IFN induction in *P. chabaudi*-infected mice, we depleted pDCs as above and measured splenic T1IFN transcript at 24 h post-infection. Animals depleted of pDCs were intact in their ability to induce splenic transcription of either *Ifna* or *Ifnb* as compared with IgG2b-treated control animals (Fig. 5B). These results indicate that pDCs do not contribute to the splenic T1IFN transcript pool at 24 h post-infection, despite the fact that they are responsible for the majority of circulating T1IFN protein.

In a complementary approach, we examined the role of RPMs in transcriptional induction of T1IFNs using *SpiC*$^{-/-}$ mice. Wild type 129Sv and *SpiC*$^{+/-}$ mice both strongly induced splenic T1IFN transcription in response to *P. chabaudi* infection (Fig. 5C). In contrast, T1IFN transcriptional induction was reduced by over 90% in *SpiC*$^{-/-}$ animals, which we note is similar to the extent of RPM depletion in *SpiC*$^{-/-}$ mice [21]. Together, these results confirm that transcriptional induction of T1IFNs in the spleens of *P. chabaudi*-infected mice is primarily occurring in RPMs, but that systemic T1IFN is produced by both RPM and pDCs.

In a final approach to characterizing the cellular origins of splenic T1IFN transcriptional induction, we employed mice homozygous for a floxed allele of *Myd88* (*Myd88*$^{fl/fl}$), which encodes an adaptor molecule required for TLR signaling. *Myd88*$^{fl/fl}$ mice that are hemizygous for the *CD11c-Cre* or *Lyz2-Cre* transgene efficiently delete *Myd88* from the genomes of dendritic cells and macrophages, respectively [22]. Consistent with the model that RPMs are responsible for transcriptional induction of T1IFNs in spleens, we observe that *CD11c-Cre* animals exhibit normal levels of splenic *Ifna* and *Ifnb* transcription, whereas *Lyz2-Cre* animals exhibit a nearly 90% reduction in T1IFN transcription (Fig. 5D-E). These data indicate that *Myd88* is required for transcriptional induction of splenic T1IFNs, but that this requirement only occurs in macrophages and not in dendritic cells. Taken together, our data lead us to conclude that RPMs, and not pDCs, are the primary source of splenic T1IFN

transcripts at 24 h after *P. chabaudi* infection, despite the fact that both populations contribute IFNB to the circulating plasma pool.

TLR9-dependent signaling is required for full induction of T1IFNs

Activation of TLR9 by A-type CpG leads to generation of IFNA in pDCs through a MYD88- and IRF7-dependent mechanism [24], and previous work similarly found TLR9-dependence of IFNA production in pDCs during *in vitro* infection with malaria parasites [14]. To characterize the molecular mechanisms by which RPMs respond to *Plasmodium* infection, we took advantage of the fact that splenic T1IFN transcript is derived from RPMs to examine the role of several signaling molecules. We measured the induction of splenic T1IFN transcript by qRT-PCR in wild type and knockout mice infected with *P. chabaudi* for 24 h. In *Tlr9*$^{-/-}$ mice, we consistently observed a two- to four-fold decrease in production of *Ifna* and *Ifnb* transcript compared to wild type mice (Fig. 6A). Consistent with an important role for TLR9 signaling in T1IFN production, *Ifna* and *Ifnb* transcripts failed to be induced to wild type levels in *Myd88*$^{-/-}$ animals, similar to our results in *Lyz2-Cre Myd88*$^{fl/fl}$ mice (Fig. 6A and 5D). Mice harboring the *Ifnb-Yfp* reporter gene in addition to a deficiency in either *Myd88* or the *Unc93b1* lesion (required for TLR3, TLR7, and TLR9 signaling [25]) also failed to induce the *Ifnb-Yfp* reporter in response to *P. chabaudi* infection (Fig. S6). These results indicate that TLR9 and MYD88 contribute to the *P. chabaudi*-induced transcription of T1IFNs in RPMs, similar to *Ifna* and *Ifnb* expression in pDCs ([14] and Fig. S6, respectively). We note that T1IFN induction in *Myd88*$^{-/-}$ animals as measured by

qRT-PCR is not completely abrogated, suggesting the existence of a MYD88-independent pathway for T1IFN induction.

In pDCs, IFNA production as a result of TLR9 activation by CpG is strongly dependent on the transcription factor IRF7 [24]. We observe that RPM from *Irf7*$^{-/-}$ mice also exhibit decreased splenic T1IFN transcription in response to *P. chabaudi* (Fig. 6B). We also assessed the role of the transcription factor IRF3 in T1IFN production since TLR signaling in macrophages can generate an early wave of T1IFN that is dependent on IRF3-mediated initiation of an IFNAR1- and IRF7-dependent autocrine amplification loop [26]. RPM from *Irf3*$^{-/-}$ mice demonstrated no defect in *Ifnb* induction in response to *P. chabaudi*. In contrast, *Irf3*$^{-/-}$ mice exhibited a diminished (two- to three-fold) capacity for *Ifna* expression compared with wild type animals. In addition to implicating IRF3 activation in *Ifna* production, these results indicate that the regulatory mechanisms of *Ifna* and *Ifnb* induction in response to *P. chabaudi* have distinct requirements, as also observed in macrophages responding to West Nile Virus [27].

Finally, we assessed the possibilities that a T1IFN amplification loop or crosstalk with IFNG could influence T1IFN induction by *P. chabaudi*. We observed that *Ifnar1* is required for wild type levels of expression of *Ifna* but not *Ifnb*, suggesting that *Ifna* (but not *Ifnb*) may be amplified through an amplification loop (Fig. 6C). In contrast, *Ifngr1*$^{-/-}$ animals exhibit no defects in T1IFN induction, ruling out the possibility of crosstalk from IFNG signaling. Taken together, the results suggest a model in which IFNA expression in RPMs is dependent on

IRF3 and on an amplification loop requiring IFNAR1, whereas IFNB induction is independent of IRF3 and the T1IFN amplification loop.

Red pulp macrophages are dispensable for control of parasitemia

Previous work using 120G8-derived antibodies to deplete pDCs has demonstrated that these cells are dispensable for control of *P. chabaudi* infection [14], and we have made similar observations using the anti-mPDCA-1 pDC-depleting antibody (Fig. S7A). In contrast, it is generally believed that RPM play an important role in parasite control given their association with circulating malaria parasites [28], their ability to phagocytose “pitted” parasites and parasitized erythrocytes [29], their expansion during infection [30,31], the exacerbation of experimental malaria infection upon phagocyte depletion [32,33], and the contribution of monocyte-derived splenic leukocytes to parasite elimination [2,4]. Together with our own observation that RPMs are early sentinels of infection, the above evidence led us to hypothesize that mice lacking RPMs would exhibit increased susceptibility to infection. Surprisingly, the course of rising and falling parasitemia in *SpiC*$^{-/-}$ mice occurred with kinetics essentially identical to the course in *SpiC*$^{+/-}$ animals (Fig. 7A) and wild type C57BL/6 animals (Fig. 3A and 3B). In order to explore the possibility that another myeloid population was compensating for the lack of RPMs, we enumerated major myeloid populations in the blood and spleen of *SpiC*$^{+/-}$ and *SpiC*$^{-/-}$ mice over the course of *P. chabaudi* infection. No consistent trends were observed in neutrophils (CD11b^{hi} Ly6g^{+}), cDCs, pDCs, or splenic marginal zone

macrophages (CD11b^{-} F4/80^{-} MARCO^{+}) (Fig. S7B-C). In contrast, Ly6c^{lo} monocyte frequencies were increased in the blood of *SpiC*$^{-/-}$ mice during resolution of peak parasitemia (days 9 and 12), and were significantly higher in spleens of *SpiC*$^{-/-}$ mice throughout infection (Fig. 7B*; p* = 0.02, Wilcoxon matched pairs signed-rank test). Detailed examination of monocyte frequencies on day 12 post-infection confirmed that Ly6c^{lo} monocytes were significantly elevated in both blood and spleens of *SpiC*$^{-/-}$ mice (Fig. 7C). We therefore conclude that although RPM contribute to early immune infection recognition and activation and are well positioned to interact with parasites, they are ultimately dispensable for control of infection, possibly as a result of compensation by Ly6c^{lo} monocytes.

Discussion

We previously found that *P. chabaudi* infection of mice induced robust expression of an interferon-induced gene signature as the earliest detectable expression response in blood [17]. Here, we demonstrate that this ISG response is the combined result of T1IFNs and IFNG, acting in a largely redundant fashion. Although the prominent involvement of IFNG in responses to malaria infection was well established, much less is understood about production of T1IFNs. Studies have shown that malaria extracts can induce IFNA by human pDCs *in vitro* [10,13], and studies have observed IFNA induction in *P. chabaudi*- [14] and *P. berghei*-infected mice [34]. Using a variety of approaches, we have demonstrated that T1IFNs are indeed produced during *in vivo* infection with *P.*

chabaudi, and that both pDCs and RPMs are the key cellular sources that contribute to the systemic T1IFN pool.

Although the protective role of T1IFNs in viral infections is well established, in some bacterial infections and autoimmune disorders, T1IFNs appear to exacerbate disease [35]. Similar to viral infections, our functional studies indicate that T1IFNs act redundantly with IFNG to activate mechanisms that protect against malaria disease. Together, our findings in concert reveal a remarkable multi-layered system of redundancy: first, at the level of multiple molecular sensing pathways in RPMs feeding into T1IFN production; second, at the level of multiple leukocyte populations generating systemically available T1IFNs; and finally, at the level of T1IFNs conferring protection that is redundant with IFNG. It is likely that this tiered redundancy design is widespread in immunological systems but has been overlooked due to absent or mild phenotypes in organism-level assays.

The cellular origin of inflammatory T1IFNs is frequently pDCs, which are also known as “interferon producing cells” due to their ability to produce more T1IFNs than any other cell type in human blood [36]. Our observation that pDCs produce IFNA and IFNB during malaria infection is in line with the general function of pDCs and similar findings from Voisine *et al.* [14]. However, we have demonstrated that RPMs also contribute significantly to total T1IFN production during the response to *P. chabaudi*, indicating that these macrophages play a role in early immune activation during malaria infection. We estimate that roughly 3000 pDC and 1000 RPM per spleen produce high levels of T1IFN, and the

comparable fluorescence levels of these populations in *Ifnb-Yfp* reporter animals suggest that pDC and RPM are capable of transcribing similar levels of *Ifnb*. Whether or not this corresponds to similar levels of IFNB production on a per-cell basis remains to be determined; regardless, our findings contribute to the increasing body of literature indicating that macrophages and other non-pDC populations are significant sources of T1IFNs *in vivo* [27,37–41].

It is likely that the localization of the infections at the tissue, cellular, and sub-cellular levels in part defines which leukocytes respond and in what manner. This is likely to be the case for T1IFN production by RPMs in malaria infection. Ultrastructural studies have demonstrated that macrophages of the red pulp are capable of phagocytosis of both whole infected erythrocytes and parasites that have been “pitted” from infected erythrocytes in the spleen [29], and trafficking studies using stained infected erythrocytes have demonstrated localization to the splenic red pulp [28]. Although these studies only examined splenic organization during the time of peak parasitemia, it was reasonable to expect that RPMs would also function as early detectors of malaria parasites due to their inherent role in filtering parasites from the blood. We have demonstrated that this is indeed the case despite the low parasite load during early sub-patent infection, and that RPMs respond by producing T1IFNs and a host of additional chemokines and cytokines. To the best of our knowledge, this is the first demonstration of production of an immunomodulatory cytokine by RPMs during early malaria infection.

We have found that TLR9-MYD88-IRF7 signaling is required for full T1IFN expression in RPMs, similar to the role of this pathway in pDC [14]. This is at odds with the fact that no *in vitro* studies of TLR9 activation have reported IFNA production in mouse pDCs or macrophages, but it is possible that malaria ligands may be less potent than synthetic ligands and therefore require additional activating signals from other leukocyte populations present *in vivo*. Obvious candidates for such signals include cytokines that signal through MAPK and NFKB pathways, which participate in *Ifnb* induction through the heterodimeric transcription factors ATF-2/c-Jun and p50/RelA [42]. Consistent with this possibility, inhibition of NFKB signaling in mice infected with West Nile Virus decreases IFNB production [27]. Further studies will be required to understand the relative contributions of these pathways *in vitro* and *in vivo*, and also to identify the pathway(s) responsible for residual levels of T1IFN production in the absences of TLR9 and MYD88.

T1IFNs can augment their own expression through a feed-forward signaling loop, but for *P. chabaudi*, only *Ifna,* not *Ifnb,* induction appears to rely on IFNAR1-dependent amplification. This result is similar to observations from *Listeria* infection, in which IFNB generation is essentially unaffected by the absence of IFNAR1 whereas IFNA production is severely diminished [40]. Similarly, expression of *Ifna* by cDCs during West Nile Virus infection was diminished in mice lacking IFNAR1, whereas *Ifnb* expression was not [27]. The consistency between these observations taken from viral, bacterial, and protozoan infection models suggests that the paradigm of the T1IFN amplification

loop may apply primarily to IFNA, but not to IFNB, production during *in vivo* infections. With regard to *Ifna* induction by *P. chabaudi*, we observe that *Ifnar1*$^{-/-}$and *Irf3*$^{-/-}$ mice exhibit similar levels of reduction, consistent with observations from other systems that these molecules are both required for T1IFN amplification [26].

Although RPMs are produce T1IFNs and other cytokines during early infection, mice lacking RPMs clear parasites with kinetics identical to control animals. This result was surprising given the general belief that RPMs contribute to control of parasitemia through phagocytic mechanisms [28,29,33,43]. Furthermore, we observe that RPMs act as early sentinels of infection and produce cytokines that ultimately contribute to elimination of infection. Given our observation that pDCs also produce T1IFNs, it is possible that all of the important functions of RPMs are redundant with other leukocyte subsets. For example, splenic monocytes are capable of phagocytosis of *P. chabaudi* [2], and this population undergoes expansion near the time of peak parasitemia in both *SpiC*$^{+/-}$ and *SpiC*$^{-/-}$ mice (Fig. S7C). Our data indicate that the Ly6c^{lo} monocytes are also significantly increased in frequency in RPM-deficient mice, suggesting the possibility that this subset could be providing redundancy with RPMs. Although the exact mechanism requires further investigation, our data indicate that the important role of the spleen in clearance of malaria infection is due to functions that are not specific to RPMs.

In summary, our results demonstrate that T1IFNs play a redundant but important protective role during experimental malaria infection. This T1IFN is

derived from both pDCs and RPMs, identifying the major populations responsible for early innate recognition of malaria infection. Future work will reveal how these innate populations and T1IFNs promote the development of an integrated immune response that can ultimately resolve malaria infection.

Materials and Methods

Mice. C57BL/6 9-14 week old female mice (Jackson Laboratories or National Cancer Institute) were maintained on a 12 h light cycle (on from 0600 to 1800 h). All mice used in this study (*Ifnar1*$^{-/-}$, *Ifngr1*$^{-/-}$, *Ifnar1*$^{-/-}$ *Ifngr1*$^{-/-}$, *Stat1*$^{-/-}$, *Ifnb-Yfp*$^{+/+}$, *Ifna6-gfp*$^{+/-}$, *Tlr9*$^{-/-}$, *Myd88*$^{-/-}$, *Irf7*$^{-/-}$) were > 95% C57BL/6 by microsatellite genotyping at 94 loci (UCSF genomics core) with the exceptions of *Irf3*$^{-/-}$ (80% C57BL/6) and *SpiC* mice (129Sv). This study was conducted in strict accordance with the guidelines of the Office of Laboratory Animal Welfare and with the approval of the UCSF Institutional Animal Care and Use Committee.

Parasites. P. chabaudi AS (MRA-429) was maintained in C57BL/6 mice. Blood was harvested by cardiac puncture from an infected mouse just prior to peak parasitemia and 10^6 infected erythrocytes introduced by *i.p.* injection. All infections were initiated at 1400 h. Blood was harvested by cardiac puncture, and spleens were harvested for analysis at specified times.

RNA. Samples for RNA preparation were immersed in RNAlater (Ambion) upon harvest and stored at -80^{o}C. RNA from blood was isolated by using the Mouse

Ribopure-Blood kit (Ambion) and amplified in a single round using the Amino Allyl MessageAmp II aRNA Amplification Kit (Ambion). RNA from spleens was isolated using Trizol as per the manufacturer's protocol, followed by two rounds of treatment with Turbo DNase (Ambion). RNA from FACS-sorted leukocyte subsets was isolated and treated with DNase using the RNAqueous Micro Kit (Ambion).

Microarrays. All microarray methods used in this study were as previously described [17]. Further details are provided as supplementary material. Data are available through the Gene Expression Omnibus (GSE23565).

qRT-PCR. For splenic RNA analysis by qRT-PCR, 3 ug of RNA was reverse transcribed, diluted, and amplified with the Quantitect SYBR Green (Qiagen) on an Opticon thermal cycler (MJ Research). Sorted leukocyte RNA was processed similarly except the entire RNA sample was used in the RT. "Universal" primers were designed to target multiple *Ifna* variants (GTGAGGAAATACTTCCACAG, GGCTCTCCAGACTTCTGCTC). Primers for *Act* (GGCTGTATTCCCCTCCATCG, CCAGTTGGTAACAATGCCATGT) and *Ifnb* (CAGCTCCAAGAAAGGACGAAC, GGCAGTGTAACTCTTCTGCAT) were from PrimerBank [44]. T1IFN transcript levels were normalized to beta-actin levels and fold-inductions calculated using the Pfaffl method.

ELISA. Assays for IFNA and IFNB were performed as per the manufacturer's instructions (Pestka Biomedical Laboratories) on K_2EDTA plasma or spleens homogenized in PBS with a protease inhibitor cocktail (Roche) using a TissueLyzer II (Qiagen).

Flow cytometry. Spleens were mechanically homogenized in FACS buffer. Erythrocytes were lysed in 1x RBC lysis solution. Fc receptors on the leukocytes were blocked with anti-CD16/CD32 antibody (2.4G2; UCSF hybridoma core), stained with specific antibodies, and analyzed/sorted on a LSR II or FACSAria II. Antibodies used for leukocyte subset identification included those targeting Siglec H (eBio440c), Ly6c (HK1.4), CD11c (N418) , and rat IgG1 staining control (eBioscience); F4/80 (BM8), CD11b (M1/70), Ly6g (1A8), and rat IgG2a staining control (2A3) (UCSF hybridoma core); and MARCO (ED31; Thermo Fisher).

Acknowledgements

We thank Shizuo Akira, Tadatsugu Taniguchi, Richard Locksley, Ruslan Medzhitov, Ken Murphy, Jon Clingan, Mehrdad Matloubian, Laura Lau, Greg Barton, and Russell Vance for providing mice; members of the Innate Immunity P01 AI063302 for advice and technical support; Lewis Lanier and Mehrdad Matloubian for discussions; Kaman Chan and Alyssa Baccarella for technical assistance; Sarah Elmes and the UCSF Laboratory for Cell Analysis for flow cytometry support; and the UCSF Center for Advanced Technology for microarray support. This work was supported by the Howard Hughes Medical

Institute (JLD), the Giannini Family Foundation (CCK), and NIAID K99 AI085035 (CCK).

References

1. Kang S-J, Liang H-E, Reizis B, Locksley RM (2008) Regulation of hierarchical clustering and activation of innate immune cells by dendritic cells. Immunity 29: 819–833. doi:10.1016/j.immuni.2008.09.017.

2. Sponaas A-M, Freitas do Rosario AP, Voisine C, Mastelic B, Thompson J, et al. (2009) Migrating monocytes recruited to the spleen play an important role in control of blood stage malaria. Blood 114: 5522–5531. doi:10.1182/blood-2009-04-217489.

3. Chimma P, Roussilhon C, Sratongno P, Ruangveerayuth R, Pattanapanyasat K, et al. (2009) A distinct peripheral blood monocyte phenotype is associated with parasite inhibitory activity in acute uncomplicated Plasmodium falciparum malaria. PLoS Pathog 5: e1000631. doi:10.1371/journal.ppat.1000631.

4. Belyaev NN, Brown DE, Diaz A-IG, Rae A, Jarra W, et al. (2010) Induction of an IL7-R(+)c-Kit(hi) myelolymphoid progenitor critically dependent on IFN-gamma signaling during acute malaria. Nat Immunol 11: 477–485. doi:10.1038/ni.1869.

5. Meding SJ, Langhorne J (1991) CD4+ T cells and B cells are necessary for the transfer of protective immunity to Plasmodium chabaudi chabaudi. Eur J Immunol 21: 1433–1438. doi:10.1002/eji.1830210616.

6. van der Heyde HC, Huszar D, Woodhouse C, Manning DD, Weidanz WP (1994) The resolution of acute malaria in a definitive model of B cell deficiency, the JHD mouse. J Immunol 152: 4557–4562.

7. von der Weid T, Honarvar N, Langhorne J (1996) Gene-targeted mice lacking B cells are unable to eliminate a blood stage malaria infection. J Immunol 156: 2510–2516.

8. van der Heyde HC, Batchelder JM, Sandor M, Weidanz WP (2006) Splenic gammadelta T cells regulated by CD4+ T cells are required to control chronic Plasmodium chabaudi malaria in the B-cell-deficient mouse. Infect Immun 74: 2717–2725. doi:10.1128/IAI.74.5.2717-2725.2006.

9. Süss G, Eichmann K, Kury E, Linke A, Langhorne J (1988) Roles of CD4- and CD8-bearing T lymphocytes in the immune response to the erythrocytic stages of Plasmodium chabaudi. Infect Immun 56: 3081–3088.

10. Pichyangkul S, Yongvanitchit K, Kum-arb U, Hemmi H, Akira S, et al. (2004) Malaria blood stage parasites activate human plasmacytoid dendritic cells and murine dendritic cells through a Toll-like receptor 9-dependent pathway. J Immunol 172: 4926–4933.

11. Coban C, Ishii KJ, Kawai T, Hemmi H, Sato S, et al. (2005) Toll-like receptor 9 mediates innate immune activation by the malaria pigment hemozoin. J Exp Med 201: 19–25. doi:10.1084/jem.20041836.

12. Parroche P, Lauw FN, Goutagny N, Latz E, Monks BG, et al. (2007) Malaria hemozoin is immunologically inert but radically enhances innate

responses by presenting malaria DNA to Toll-like receptor 9. Proc Natl Acad Sci USA 104: 1919–1924. doi:10.1073/pnas.0608745104.

13. Wu X, Gowda NM, Kumar S, Gowda DC (2010) Protein-DNA complex is the exclusive malaria parasite component that activates dendritic cells and triggers innate immune responses. J Immunol 184: 4338–4348. doi:10.4049/jimmunol.0903824.

14. Voisine C, Mastelic B, Sponaas A-M, Langhorne J (2010) Classical CD11c+ dendritic cells, not plasmacytoid dendritic cells, induce T cell responses to Plasmodium chabaudi malaria. Int J Parasitol 40: 711–719. doi:10.1016/j.ijpara.2009.11.005.

15. Newman KC, Korbel DS, Hafalla JC, Riley EM (2006) Cross-talk with myeloid accessory cells regulates human natural killer cell interferon-gamma responses to malaria. PLoS Pathog 2: e118. doi:10.1371/journal.ppat.0020118.

16. Ojo-Amaize EA, Salimonu LS, Williams AI, Akinwolere OA, Shabo R, et al. (1981) Positive correlation between degree of parasitemia, interferon titers, and natural killer cell activity in Plasmodium falciparum-infected children. J Immunol 127: 2296–2300.

17. Kim CC, Parikh S, Sun JC, Myrick A, Lanier LL, et al. (2008) Experimental malaria infection triggers early expansion of natural killer cells. Infect Immun 76: 5873–5882. doi:10.1128/IAI.00640-08.

18. Seeds RE, Gordon S, Miller JL (2009) Characterisation of myeloid receptor expression and interferon alpha/beta production in murine plasmacytoid

dendritic cells by flow cytomtery. J Immunol Methods 350: 106–117. doi:10.1016/j.jim.2009.07.016.

19. Scheu S, Dresing P, Locksley RM (2008) Visualization of IFNbeta production by plasmacytoid versus conventional dendritic cells under specific stimulation conditions in vivo. Proc Natl Acad Sci USA 105: 20416–20421. doi:10.1073/pnas.0808537105.

20. Kumagai Y, Takeuchi O, Kato H, Kumar H, Matsui K, et al. (2007) Alveolar macrophages are the primary interferon-alpha producer in pulmonary infection with RNA viruses. Immunity 27: 240–252. doi:10.1016/j.immuni.2007.07.013.

21. Kohyama M, Ise W, Edelson BT, Wilker PR, Hildner K, et al. (2009) Role for Spi-C in the development of red pulp macrophages and splenic iron homeostasis. Nature 457: 318–321. doi:10.1038/nature07472.

22. Hou B, Reizis B, DeFranco AL (2008) Toll-like receptors activate innate and adaptive immunity by using dendritic cell-intrinsic and -extrinsic mechanisms. Immunity 29: 272–282. doi:10.1016/j.immuni.2008.05.016.

23. Randolph GJ, Ochando J, Partida-Sánchez S (2008) Migration of dendritic cell subsets and their precursors. Annu Rev Immunol 26: 293–316. doi:10.1146/annurev.immunol.26.021607.090254.

24. Honda K, Yanai H, Negishi H, Asagiri M, Sato M, et al. (2005) IRF-7 is the master regulator of type-I interferon-dependent immune responses. Nature 434: 772–777. doi:10.1038/nature03464.

25. Tabeta K, Hoebe K, Janssen EM, Du X, Georgel P, et al. (2006) The Unc93b1 mutation 3d disrupts exogenous antigen presentation and signaling via Toll-like receptors 3, 7 and 9. Nat Immunol 7: 156–164. doi:10.1038/ni1297.

26. Sato M, Suemori H, Hata N, Asagiri M, Ogasawara K, et al. (2000) Distinct and essential roles of transcription factors IRF-3 and IRF-7 in response to viruses for IFN-alpha/beta gene induction. Immunity 13: 539–548.

27. Daffis S, Suthar MS, Szretter KJ, Gale M, Diamond MS (2009) Induction of IFN-beta and the innate antiviral response in myeloid cells occurs through an IPS-1-dependent signal that does not require IRF-3 and IRF-7. PLoS Pathog 5: e1000607. doi:10.1371/journal.ppat.1000607.

28. Yadava A, Kumar S, Dvorak JA, Milon G, Miller LH (1996) Trafficking of Plasmodium chabaudi adami-infected erythrocytes within the mouse spleen. Proc Natl Acad Sci USA 93: 4595–4599.

29. Schnitzer B, Sodeman T, Mead ML, Contacos PG (1972) Pitting function of the spleen in malaria: ultrastructural observations. Science 177: 175–177.

30. Krücken J, Mehnert LI, Dkhil MA, El-Khadragy M, Benten WPM, et al. (2005) Massive destruction of malaria-parasitized red blood cells despite spleen closure. Infect Immun 73: 6390–6398. doi:10.1128/IAI.73.10.6390-6398.2005.

31. Stevenson MM, Kraal G (1989) Histological changes in the spleen and liver of C57BL/6 and A/J mice during Plasmodium chabaudi AS infection. Exp Mol Pathol 51: 80–95.

32. Couper KN, Blount DG, Hafalla JCR, van Rooijen N, de Souza JB, et al. (2007) Macrophage-mediated but gamma interferon-independent innate immune

responses control the primary wave of Plasmodium yoelii parasitemia. Infect Immun 75: 5806–5818. doi:10.1128/IAI.01005-07.

33. Stevenson MM, Ghadirian E, Phillips NC, Rae D, Podoba JE (1989) Role of mononuclear phagocytes in elimination of Plasmodium chabaudi AS infection. Parasite Immunol 11: 529–544.

34. Haque A, Best SE, Ammerdorffer A, Desbarrieres L, de Oca MM, et al. (2011) Type I interferons suppress $CD4^+$ T-cell-dependent parasite control during blood-stage Plasmodium infection. Eur J Immunol 41: 2688–2698. doi:10.1002/eji.201141539.

35. Trinchieri G (2010) Type I interferon: friend or foe? J Exp Med 207: 2053–2063. doi:10.1084/jem.20101664.

36. Colonna M, Trinchieri G, Liu Y-J (2004) Plasmacytoid dendritic cells in immunity. Nat Immunol 5: 1219–1226. doi:10.1038/ni1141.

37. Eloranta ML, Alm GV (1999) Splenic marginal metallophilic macrophages and marginal zone macrophages are the major interferon-alpha/beta producers in mice upon intravenous challenge with herpes simplex virus. Scand J Immunol 49: 391–394.

38. Dalod M, Salazar-Mather TP, Malmgaard L, Lewis C, Asselin-Paturel C, et al. (2002) Interferon alpha/beta and interleukin 12 responses to viral infections: pathways regulating dendritic cell cytokine expression in vivo. J Exp Med 195: 517–528.

39. Ciavarra RP, Taylor L, Greene AR, Yousefieh N, Horeth D, et al. (2005) Impact of macrophage and dendritic cell subset elimination on antiviral immunity,

viral clearance and production of type 1 interferon. Virology 342: 177–189. doi:10.1016/j.virol.2005.07.031.

40. Stockinger S, Kastner R, Kernbauer E, Pilz A, Westermayer S, et al. (2009) Characterization of the interferon-producing cell in mice infected with Listeria monocytogenes. PLoS Pathog 5: e1000355. doi:10.1371/journal.ppat.1000355.

41. Swiecki M, Colonna M (2010) Unraveling the functions of plasmacytoid dendritic cells during viral infections, autoimmunity, and tolerance. Immunol Rev 234: 142–162. doi:10.1111/j.0105-2896.2009.00881.x.

42. Panne D, Maniatis T, Harrison SC (2007) An atomic model of the interferon-beta enhanceosome. Cell 129: 1111–1123. doi:10.1016/j.cell.2007.05.019.

43. Engwerda CR, Beattie L, Amante FH (2005) The importance of the spleen in malaria. Trends Parasitol 21: 75–80. doi:10.1016/j.pt.2004.11.008.

44. Spandidos A, Wang X, Wang H, Seed B (2010) PrimerBank: a resource of human and mouse PCR primer pairs for gene expression detection and quantification. Nucleic Acids Res 38: D792–799. doi:10.1093/nar/gkp1005.

Figure Legends

Fig. 1. T1IFN and IFNG signaling redundantly regulate early gene expression responses to *P. chabaudi* infection. A representative set of ISG is shown for the gene expression response in whole blood from animals infected for

24 h with *P. chabaudi* in C57BL/6 knockout mice. Each column represents an individual mouse.

Fig. 2. T1IFNs are produced during *P. chabaudi* infection. (A) Kinetics of early *Ifna, Ifnb*, and *Ifng* transcription using whole spleen qRT-PCR. Fold mRNA induction represents the ratio of transcript in infected- over mock-infected C57BL/6 mice. (B) Reproducibility of T1IFN transcript induction as detected by whole spleen qRT-PCR. Each point represents an independent experiment with 4-6 animals, with horizontal bars displaying the geometric mean. (C) Plasma IFNA and IFNB at 24 h post-infection. Data are combined from two independent experiments with each point representing one animal. N.D. = not detected (*n* = 8).

Fig. 3. T1IFNs contribute to control of *P. chabaudi* infection. Infected mice were monitored for parasitemia by thin blood smear and survival. Wild type C57BL/6 and congenic knockout mice were infected and monitored for percent parasitemia (*n* = 5 per strain), which are represented as geometric means with standard deviations and Mann-Whitney *p*-value. A representative experiment of two is shown. Crosses indicate deaths due to parasitemia.

Fig. 4. *P. chabaudi* infection induces IFNB production in pDCs and RPMs. (A) No splenocytes are YFP^+ in mock-infected samples, but some splenocytes become YFP^+ 24 h after *P. chabaudi* infection of C57BL/6 *Ifnb* reporter mice. 2.5 x 10^6 events are depicted in each dot plot. (B) pDCs and RPMs constitute over 90% of YFP^+ events in C57BL/6 mice. (C) Both pDCs and RPMs contribute to splenic IFNB production in 129Sv mice. pDCs were depleted with a single 500

mg dose of anti-mPDCA1 antibody 18 h before infection with *P. chabaudi*. Grey dots represent individual mice, with horizontal bars representing the mean (B) or geometric mean (C).

Fig. 5. Cellular requirements for splenic T1IFN transcriptional induction. (A) RPMs, but not other macrophage or dendritic cell subsets, induce *Ifna* and *Ifnb* at 24 h post-infection with *P. chabaudi* as detected by qRT-PCR in C57BL/6 mice. Fold mRNA induction represents fold induction of transcript in infected versus mock-infected normalized to beta-actin. Grey dots represent independent experiments conducted on different days; black bars denote the geometric means of the fold inductions. (B) pDCs are not required for splenic *Ifna* or *Ifnb* transcriptional induction in response to *P. chabaudi* in C57BL/6 mice. (C) Genetic deletion of RPMs in 129Sv mice results in diminished T1IFN transcriptional induction. (D) Genetic deletion of *Myd88* from dendritic cells does not impact transcriptional induction of T1IFNs in spleens of C57BL/6 mice. (E) Genetic deletion of *Myd88* from macrophages/neutrophils decreases transcriptional induction of T1IFNs by an order of magnitude in C57Bl/6 mice. Grey bars represent geometric means with 95% confidence intervals. Asterisks represent $p < 0.05$ in a Student's *t*-test against control samples.

Fig. 6. Molecular requirements for splenic T1IFN transcriptional induction. (A) *Tlr9* and *Myd88* are required for full transcriptional induction of T1IFNs in spleens of *P. chabaudi*-infected C57Bl/6 mice. Grey dots represent the means of independent experiments using 4-6 total mice, with T1IFN fold mRNA induction in knockout mice represented as a percentage of induction in wild type animals.

Black bars represent means; asterisks represent $p < 0.05$ as compared to wild type induction in a two-tailed Student's *t*-test assuming unequal variances. (B) *Irf7* is required for full T1IFN induction, and *Irf3* is required for full *Ifna* but not *Ifnb* induction. (C) *Ifnar1* is required for full *Ifna* induction but dispensable for *Ifnb* induction, whereas *Ifngr1* is dispensable for all T1IFN induction.

Fig. 7. Mice lacking RPMs exhibit wild type infection kinetics. (A) Parasitemia courses in 129Sv *SpiC*$^{+/-}$ (*n* = 4) and *SpiC*$^{-/-}$ (*n* = 5) mice are represented as geometric means with standard deviations and Mann-Whitney *p*-value. (B) Ly6c^{lo} monocyte (CD11b^{+} F4/80^{+} Ly6g^{-} SSC^{lo} Ly6c^{lo}) frequencies in blood and spleen during the course of infection. Days depicted in blue and orange represent a 1.5-fold decrease or increase, respectively, in Ly6c^{lo} monocyte frequencies in blood and spleen of mice infected with *P. chabaudi*. *SpiC*$^{+/-}$ are depicted in white and *SpiC*$^{-/-}$ are depicted in black bars (C) Ly6c^{lo} monocyte frequencies on day 12 post-infection. Means are presented with standard errors; *p*-values represent a two-tailed *t*-test assuming unequal variances. Data represent three independent experiments (*n* = 6-7 mice per group total).

FIGURES

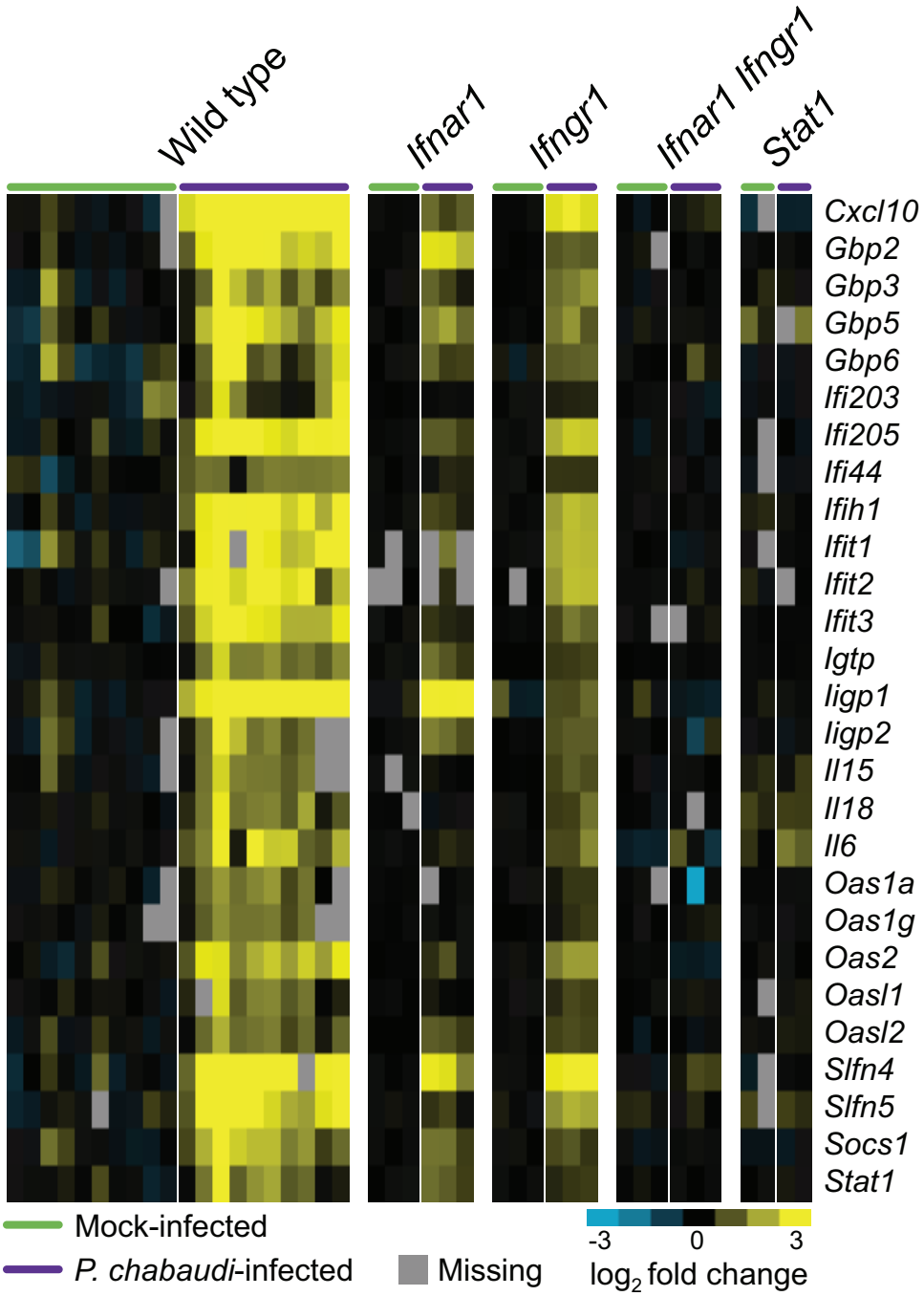


Figure 1

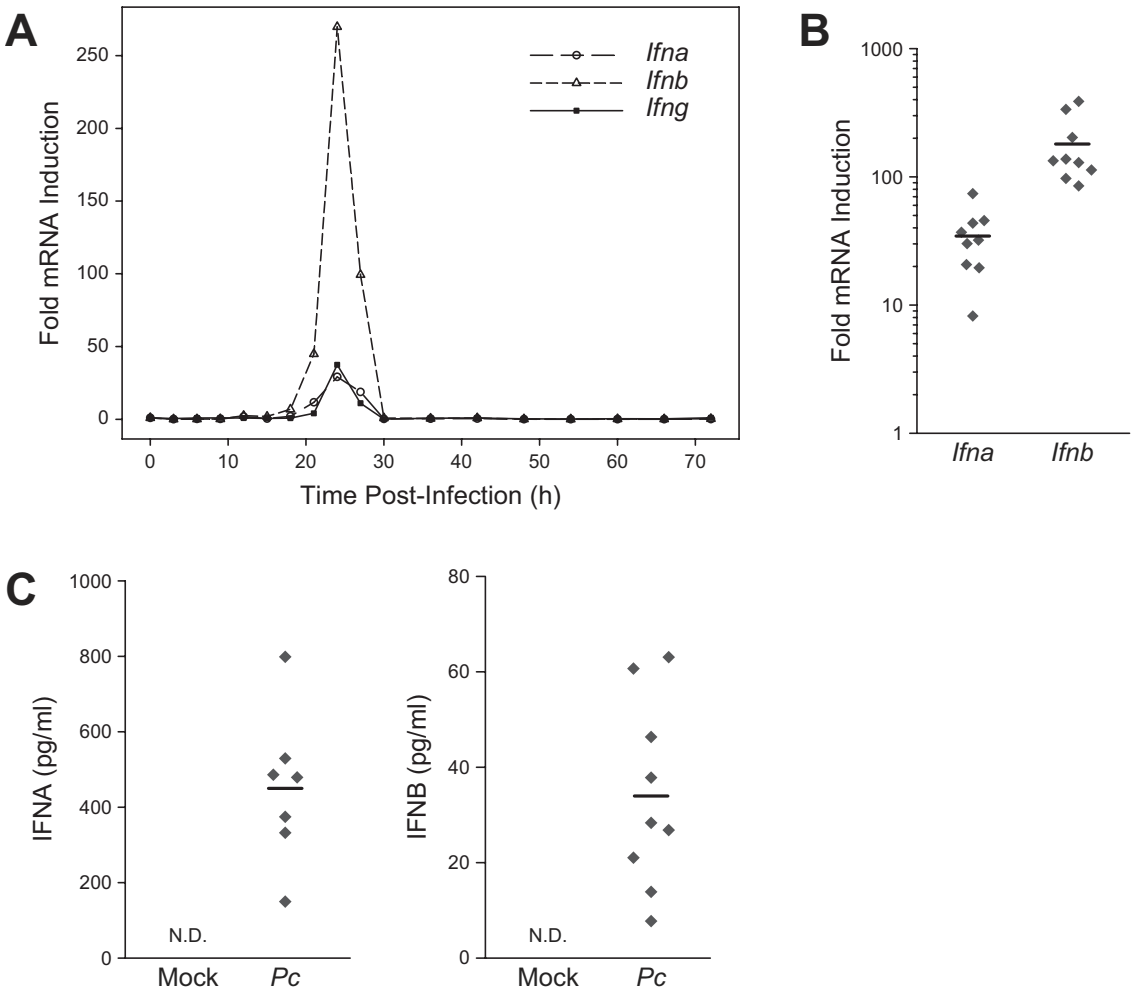

A
Fold mRNA Induction
250
200
150
100
50
0
0
10
20
30
40
50
60
70
Time Post-Infection (h)
Ifna
Ifnb
Ifng
B
Fold mRNA Induction
1000
100
10
1
Ifna
Ifnb
C
IFNA (pg/ml)
1000
800
600
400
200
0
N.D.
Mock
Pc
IFNB (pg/ml)
80
60
40
20
0
N.D.
Mock
Pc

Figure 2

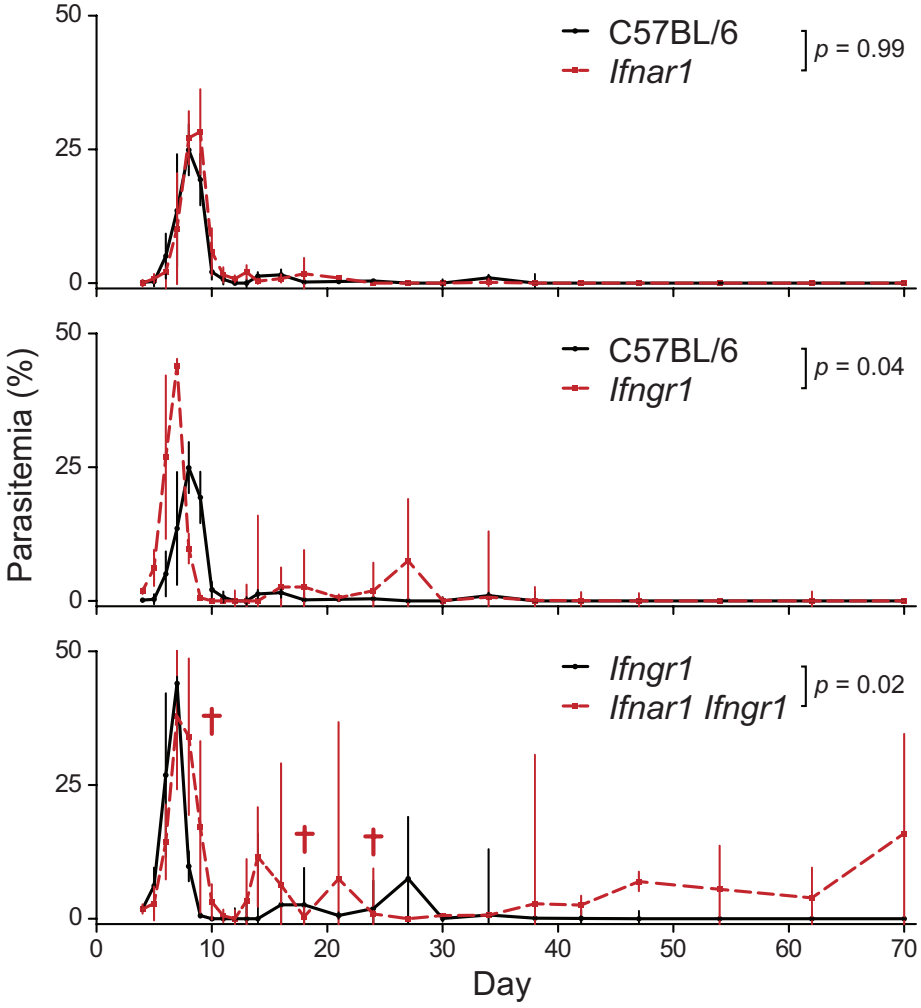


Figure 3

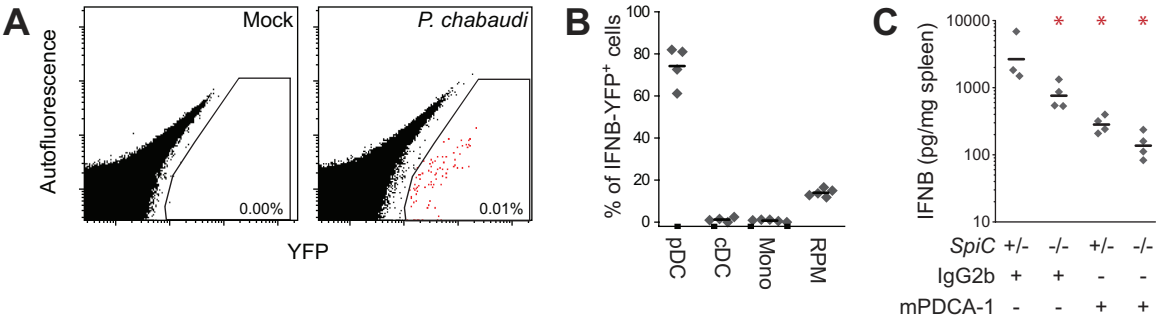


Figure 4

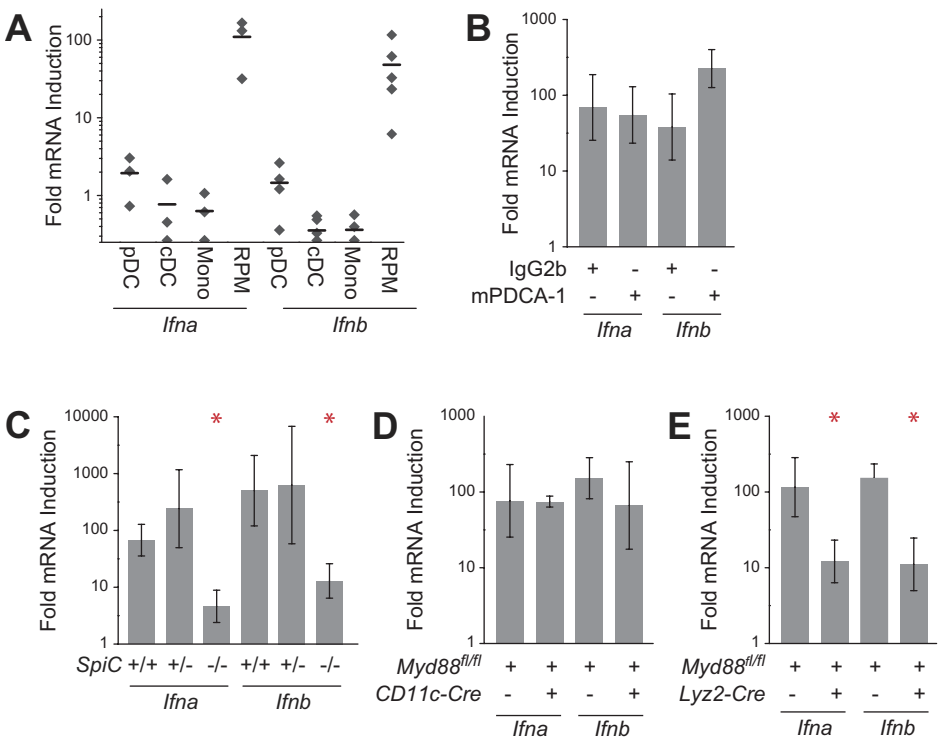


Figure 5

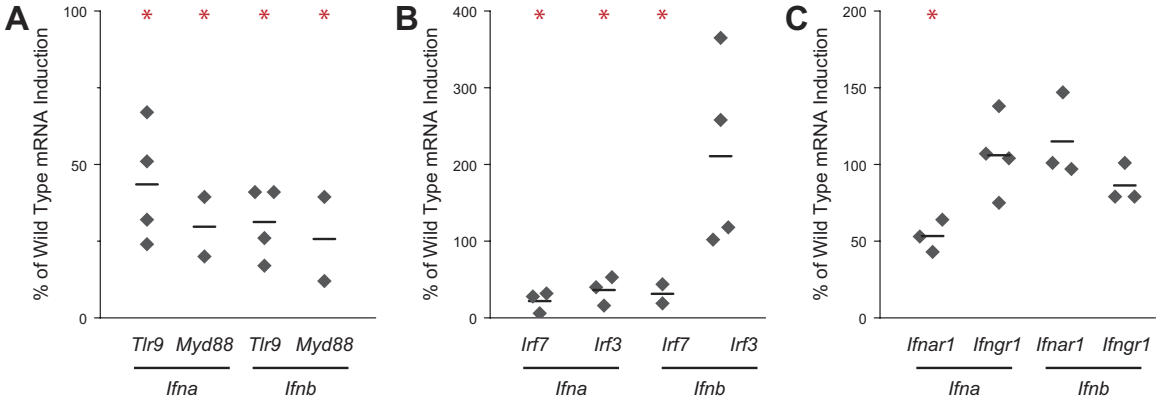


Figure 6

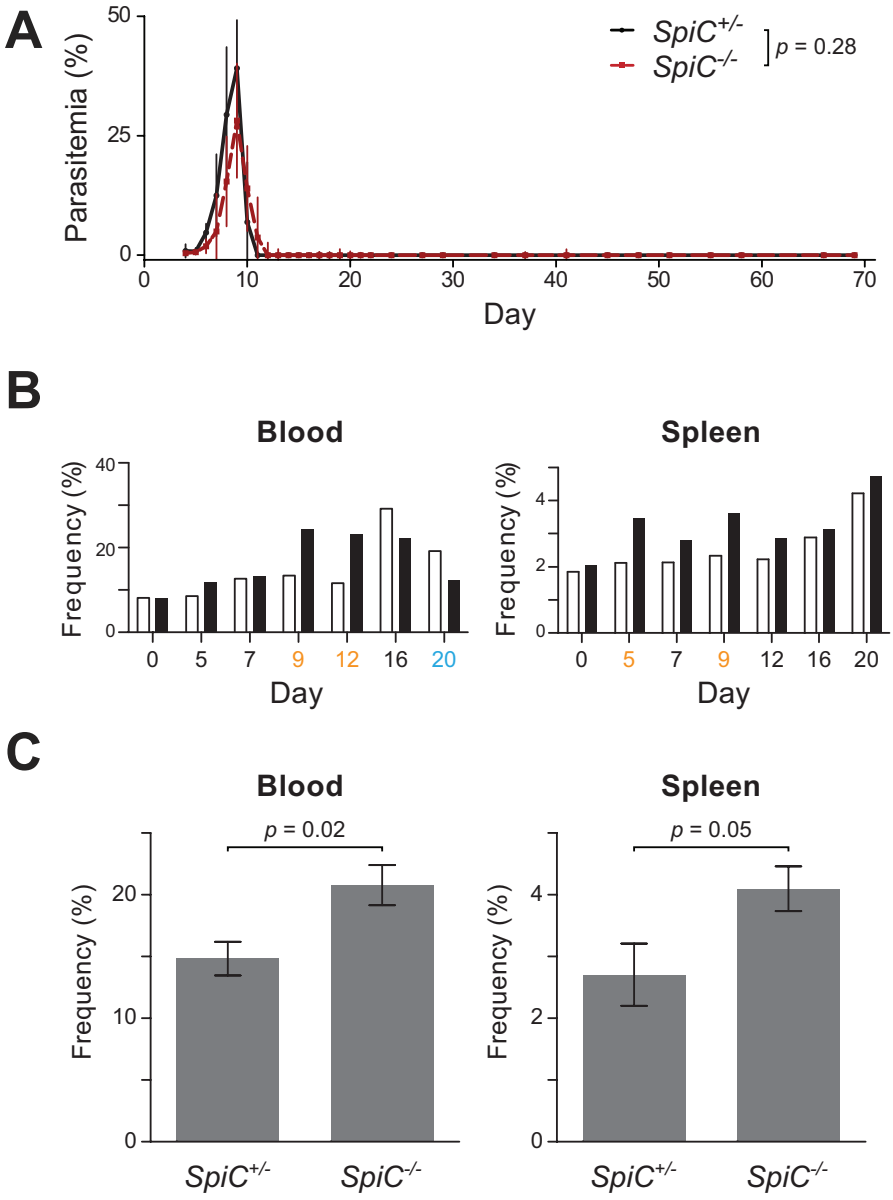


Figure 7

Supplemental Experimental Procedures

Microarray hybridization. All whole blood samples were hybridized against a reference of pooled samples supplemented with amplified Universal Mouse Reference RNA (Stratagene, 740100). Red pulp macrophage hybridizations were

performed against a common aRNA pool from amplified red pulp macrophage samples. In all experiments, Cy5 was used to label the experimental sample and Cy3 was used to label the reference. Sample and reference amplified RNA were combined and mixed with hybridization solution (polyA, yeast tRNA, HEPES, SSC, SDS), boiled, and hybridized to mouse whole-genome microarrays for 15-17 h at 65°C under MAUI AO mixing chambers. The MEEBO microarray probe set was utilized due to its high genome coverage and constitutive exonic design (1).

Microarray Analysis. Image data were extracted in Genepix 6 (Molecular Devices) and normalized and filtered in Acuity 4 (Molecular Devices). Data were ratio-normalized, and control and poor quality features (as determined by both visual examination and application of quantitative filters for saturation, feature diameter, and variance) were removed. Data were further filtered for spots that did not exhibit signal in any of the sample channels (based on percentage of pixels in the feature above background and feature intensity) and for excessive missing data. The ratio of medians of the remaining data was transformed to $\log_2$ space, median centered by array, and median centered by gene.

The normalized and filtered data were analyzed for differential gene expression by hierarchical clustering using Xcluster (2) and statistical analysis in Significance Analysis of Microarrays (SAM) (3). All SAM analyses were conducted as two-class unpaired analyses using a *t*-statistic, K-nearest neighbor imputation of missing values, and cutoffs set at a 1% false discovery rate.

Microarray data were visualized in Java Treeview (4). Significant gene sets were subject to functional analysis using gene ontology analysis with DAVID (5, 6).

References

1. Verdugo RA, Medrano JF (2006) Comparison of gene coverage of mouse oligonucleotide microarray platforms. *BMC Genomics* 7:58.

2. Gollub J, Sherlock G (2006) Clustering microarray data. *Meth. Enzymol* 411:194-213.

3. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U.S.A* 98:5116-5121.

4. Saldanha AJ (2004) Java Treeview--extensible visualization of microarray data. *Bioinformatics* 20:3246-3248.

5. Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37:1-13.

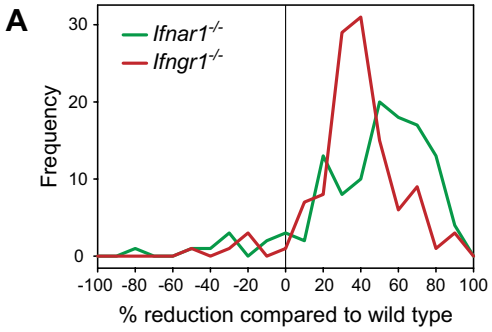
6. Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4:44-57.

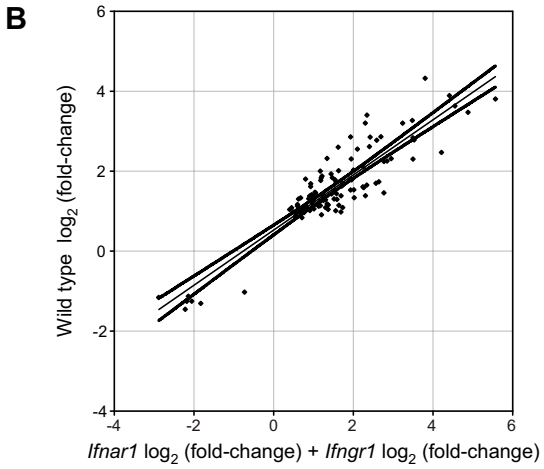
Supplemental Table 1: expression array results with interferon signaling mutants

UNIQID	NAME	Fold Change (Pc/Mock) B6	FC Ifnar1	FC Ifngr1	FC Ifnar1 Ifngr1	FC Stat1
mMC009244	STEAP family member 4 (Steap4)	2.611719574	3.89061979	1.270150983	0.841382276	1.790050142
mMC100078	adhesion molecule, interacts with CXADR antigen 1 (Amica1)	1.856390217	1.104454001	1.3441244	0.964263938	1.305859787
mMC024950	leukocyte immunoglobulin-like receptor, subfamily B, member 4 (Lilrb4)	1.965641197	1.042465761	1.529789694	1.002313162	1.986184991
mMC009749	transcription factor EC (Tcfec)	1.952063522	1.101905116	1.540430222	0.986232704	1.484523571
mMC004932	bone marrow stromal cell antigen 1 (Bst1)	2.522754818	2.231994285	1.821339667	0.934651216	1.602139755
mMC022593	chemokine (C-X-C motif) ligand 2 (Cxcl2)	2.45055939	0.838955775	2.383915887	0.77781556	2.361985323
mMC011595	interleukin 6 (Il6)	3.572663409	1.295342252	2.143546925	0.687770909	1.945309895
mMC100426	zinc finger protein 213 (Zfp213)	3.323031572	2.453769955	1.205807828	1.252664439	1.189207115
mMC018907	ubiquitin specific peptidase 18 (Usp18)	3.561948522	1.351910833	1.693490625	0.934651216	0.787307977
mMC004775	lectin, galactose binding, soluble 9 (Lgals9)	1.785919022	1.164733586	1.404444876	1.098092814	1.057018041
mMA033055	lectin, galactose binding, soluble 9 (Lgals9)	1.823164129	0.996540263	1.460706845	1.079228237	1.01395948
mMC020533	2'-5' oligoadenylate synthetase 1G|2'-5' oligoadenylate synthetase 1A (Oas1g Oas1a)	2.186902481	1.053361036	1.433955248	0.488579984	0.923382311
mMC018541	lymphocyte antigen 6 complex, locus C1|lymphocyte antigen 6 complex, locus C2 (Ly6c1 Ly6c2)	1.975201723	0.98851402	1.427344254	0.755236293	0.795536484
mMC101123	mMC101123	2.744734621	1.064370182	1.949809711	0.801069878	1.04608494
mMR026533	poly (ADP-ribose) polymerase family, member 14 (Parp14)	2.122846418	1.071773463	1.662475792	1.310393404	0.904379378
mMR030870	RIKEN cDNA A530023O14 gene (A530023O14Rik)	2.436585945	1.337927555	1.844632387	1.047899238	1.109569472
mMR026646	RIKEN cDNA A530064D06 gene (A530064D06Rik)	2.694000279	1.087991033	2.516932745	1.121166078	2.006943497
mMC003934	interleukin 1 receptor, type II (Il1r2)	2.744734621	2.566851795	2.651238884	1.060687741	2.150988781
mMC022606	MAS-related GPR, member A2 (Mrgpra2)	3.240575395	1.43893358	2.310705394	1.104454001	1.892115293
mMC002484	immunoresponsive gene 1 (Irg1)	4.776652748	2.224272172	3.234030609	1.160703914	1.205807828
mMC011115	nucleotide-binding oligomerization domain containing 2 (Nod2)	2.07915887	1.203025036	1.674039226	0.820741609	1.510472586
mMC021817	interleukin 18 (Il18)	2.515769944	0.845767679	1.887748625	0.890384263	1.071773463
mMC022890	membrane-spanning 4-domains, subfamily A, member 6D (Ms4a6d)	4.034108817	1.093030254	2.834969734	1.150690623	1.301341855
mMC017690	membrane-spanning 4-domains, subfamily A, member 6D (Ms4a6d)	3.810551992	1.286394669	2.114036081	1.0334725	1.257013375
mMC020632	DNA segment, Chr 14, ERATO Doi 668, expressed (D14Ertd668e)	2.635359903	1.233992225	1.883392035	1.038859103	0.832198735
mMR030864	DEAD (Asp-Glu-Ala-Asp) box polypeptide 58 (Ddx58)	2.512671766	1.248330549	1.515716567	1.08047573	0.901250463
mMA033726	signal transducer and activator of transcription 1 (Stat1)	2.212994706	2.051482243	1.515716567	1.050930065	0.982820599
mMC024858	mMC024858	5.535639528	10.17295333	1.823444977	0.957603281	1.064370182
mMR028612	interferon inducible GTPase 1 (Iigp1)	11.11164109	14.79119456	1.990779358	0.904379378	0.885767519
mMC021241	interferon inducible GTPase 1 (Iigp1)	12.41871886	12.69920842	1.844632387	1.347233577	0.867538687
mMC011991	T-cell specific GTPase (Tgtp)	4.936349163	4.367073058	2.584705661	0.987943197	1.068065408
mMC007052	guanylate nucleotide binding protein 5 (Gbp5)	4.982949662	3.003545807	2.572789339	1.077359696	1.510472586
mMC019577	CD274 antigen (Cd274)	4.081218637	2.018569602	1.986184991	0.944419673	1.031683179
mMC024746	guanylate nucleotide binding protein 2 (Gbp2)	6.889715663	5.540437872	2.080119868	1.025741121	1.057018041
mMC025349	interferon gamma induced GTPase (Igtp)	2.609909895	1.981601227	1.63202897	1	0.972654947
mMC000979	interleukin 18 binding protein (Il18bp)	3.247008387	2.828427125	2.094588246	0.918594468	1.160703914
mMR028602	suppressor of cytokine signaling 1 (Socs1)	3.446184639	2.193649959	1.769489662	0.882702996	1.003471749
mMR027974	poly (ADP-ribose) polymerase family, member 9 (Parp9)	2.265549655	1.353473524	1.874708993	1.038259208	0.771105413
mMC025225	interferon inducible GTPase 2 (Iigp2)	3.006728171	2.340256506	2.046747784	0.882702996	0.955945318
mMC003495	guanylate nucleotide binding protein 3 (Gbp3)	3.048289661	1.68568309	2.80240733	0.898132373	0.843815796
mMC016938	serine (or cysteine) peptidase inhibitor, clade A, member 3F (Serpina3f)	3.327513327	3.379164695	1.844632387	0.982253063	1.152686347
mMC007427	chemokine (C-C motif) ligand 12 (Ccl12)	6.137447969	1.252664439	4.228072162	1.148698355	1.035264924
mMC001992	2'-5' oligoadenylate synthetase-like 1 (Oasl1)	2.162116775	1.091768265	1.605845764	1.121166078	1.128964405
mMC024156	chemokine (C-C motif) receptor-like 2 (Ccrl2)	2.490179949	1.295342252	1.85746282	1.105730653	1.038859103
mMC013715	chemokine (C-X-C motif) ligand 10 (Cxcl10)	20.05498371	2.070529848	6.742732358	1.103816227	1.136816973
mMC020923	torsin family 3, member A (Tor3a)	2.599078125	0.972654947	1.918528239	0.987943197	1.068065408
mMC023596	2'-5' oligoadenylate synthetase 1G (Oas1g)	2.244275831	1.081724666	1.397969934	1.066216194	0.91383145
mMR027562	myeloid cell nuclear differentiation antigen|interferon activated gene 205 (Mnda Ifi205)	9.208715931	1.936341392	4.868014055	0.893475454	0.946057647
mMC012881	receptor transporter protein 4 (Rtp4)	1.879417665	1.222640278	1.879045498	1.028113827	1.049716684
mMC010618	schlafen 5 (Slfn5)	7.236149538	1.420763739	3.775497251	1.20163605	0.888842681
mMC025156	cDNA sequence BC013672 (BC013672)	5.86649985	1.505246747	2.854688508	1.038859103	1.00695555
mMA035547	mMA035547	7.255122631	1.474269217	2.584705661	1.136816973	0.959264119
mMC019312	interferon activated gene 205 (Ifi205)	10.59271128	1.853176124	2.738400258	0.952637998	0.817902059
mMC024032	interferon-induced protein with tetratricopeptide repeats 3 (Ifit3)	6.067070181	1.391524844	2.208908001	1.114708637	0.955945318
mMC012417	interferon-induced protein with tetratricopeptide repeats 2 (Ifit2)	6.854783548	1.424050196	4.218314518	1.081100018	0.929804943
mMC010553	interferon induced with helicase C domain 1 (Ifih1)	7.280310593	1.551144762	4.169863043	0.902813565	0.812252396
mMC025695	myxovirus (influenza virus) resistance 2 (Mx2)	3.449835995	1.251218139	2.326777621	1.022782939	0.885767519
mMC020260	cDNA sequence BC013672 (BC013672)	4.738546005	1.695448152	4.018526698	1.298338588	0.982820599
mMC017279	schlafen 4 (Slfn4)	13.99116746	4.981798489	9.51365692	1.18304094	1.469168633
mMC005508	thymidylate kinase family LPS-inducible member (Tyki)	9.658869326	2.143546925	5.205367422	1.058851301	0.969289817
mMC013848	interferon-induced protein with tetratricopeptide repeats 1 (Ifit1)	7.146977931	2.620786808	4.306950273	0.901250463	1.172834949
mMC013685	ring finger protein 213 (Rnf213)	2.772139771	1.109569472	1.84889932	1.008701984	0.969289817
mMR030989	poly (ADP-ribose) polymerase family, member 12 (Parp12)	2.869236058	1.609560345	1.785919022	1.055187954	1.017479692
mMC016180	poly (ADP-ribose) polymerase family, member 12 (Parp12)	3.103962464	1.378723669	2.394957409	0.977724561	0.826450318
mMC009824	immunity-related GTPase family, M (Irgm)	3.140512475	2.60870414	1.936341392	1.058851301	1.086734863
mMC000953	2'-5' oligoadenylate synthetase 2 (Oas2)	4.301977562	1.464085696	3.25652507	0.815072332	0.882702996
mMC019234	transcription factor B2, mitochondrial|gene model 1818, (NCBI) (Tfb2m Gm1818)	2.039195366	1.05457863	1.565546833	0.934651216	0.907519155
mMC009139	2'-5' oligoadenylate synthetase-like 2 (Oasl2)	2.128740365	1.879045498	1.769489662	0.857078944	1.101905116
mMC009424	interleukin 15 (Il15)	2.541280377	1.079228237	1.887748625	0.926588062	0.976031761
mMR030098	signal transducer and activator of transcription 1 (Stat1)	2.060507907	1.733074092	1.487957514	1.026333784	0.901250463
mMC007217	RIKEN cDNA A630077B13 gene (A630077B13Rik)	2.480370361	1.38831345	1.325619442	0.955945318	1.010451446
mMC012433	guanylate binding protein 6 (Gbp6)	2.895876345	1.883392035	2.143546925	1.250495616	1.010451446
mMC020765	deltex 3-like (Drosophila) (Dtx3l)	1.9495094	1.180992661	1.2397077	1.010451446	1.053361036
mMR026790	cDNA sequence BC057170 (BC057170)	2.709449955	1.936341392	1.540430222	0.904379378	0.920187651
mMC026094	SAM domain and HD domain, 1 (Samhd1)	2.200757219	1.547564994	1.457335791	0.943329267	1.105730653
mMC006893	SAM domain and HD domain, 1 (Samhd1)	2.225300241	1.591072968	1.624504793	0.976031761	1.098092814
mMC006622	Fc receptor, IgG, low affinity IV (Fcgr4)	2.020902893	1.301341855	1.440596862	0.982820599	1.071773463
mMC025310	PHD finger protein 11 (Phf11)	4.00277355	1.140763716	1.972465409	0.907519155	1
mMC024437	interferon activated gene 205|interferon activated gene 203|myeloid cell nuclear differentiation ant	2.468554791	0.968170696	1.591072968	0.991373087	0.989656656
mMC021322	expressed sequence AI451617 (AI451617)	2.034959384	1.231144413	1.378723669	1.019244379	0.986232704
mMC024994	tripartite motif protein 30 (Trim30)	3.491063664	1.477679441	2.566851795	0.939522749	1.042465761
mMC001101	placenta-specific 8 (Plac8)	2.368543224	1.143402487	1.967913307	0.907519155	1.090507733
mMC019705	membrane-spanning 4-domains, subfamily A, member 4C (Ms4a4c)	3.40817638	0.950439478	2.38943027	1.098092814	1.117287138
mMC024853	cDNA sequence BC094916 (BC094916)	3.063587851	0.890898718	2.133664486	0.924983798	1.017479692
mMA032762	interferon activated gene 205|myeloid cell nuclear differentiation antigen|interferon activated gene	3.218129143	0.899170536	2.118926189	0.870550563	0.939522749
mMA033534	interferon activated gene 203 (Ifi203)	2.056227653	0.963707118	1.369200129	0.90909313	1.278985581
mMA035185	interferon activated gene 203 (Ifi203)	2.119905567	1.025741121	1.331759279	0.832198735	0.969289817
mMR026842	RIKEN cDNA I830012O16 gene|interferon-induced protein with tetratricopeptide repeats 3 (I8300	4.991674321	1.658639092	1.526259209	1.219114248	0.969289817
mMC015428	MAX dimerization protein 1 (Mxd1)	2.472550522	1.547564994	1.84889932	1.226884977	2.056227653
mMC021173	GTPase, very large interferon inducible 1 (Gvin1)	2.450370664	2.03731162	1.477679441	1.066216194	1.003471749
mMA033083	interferon activated gene 203 (Ifi203)	1.959444139	0.963707118	1.643380629	0.984525173	0.879649076
mMC006436	RIKEN cDNA 2700019D07 gene (2700019D07Rik)	2.112571251	0.892959511	1.967913307	0.87206042	1.301341855
mMC018950	lipase, endothelial (Lipg)	9.225042486	1.494849249	3.294364069	1.066832243	1.296839555
mMC012328	mMC012328	4.955394791	1.717130873	2.244924097	0.976031761	0.817902059
mMC100640	chemokine (C-C motif) ligand 12 (Ccl12)	3.673468382	1.21981864	1.903076206	1.170804341	1.082975046
mMC100490	F-box protein 39 (Fbxo39)	2.407440263	1.410949807	1.681792831	1.250495616	1.021012126
mMC004204	transmembrane protein 67 (Tmem67)	2.08384793	1.069299999	1.453972517	1.024556823	0.976031761
mMC015136	eukaryotic translation initiation factor 2-alpha kinase 2 (Eif2ak2)	2.657371628	1.035264924	1.866065983	0.858565436	0.870550563
mMC003293	complement factor B (Cfb)	3.489450823	1.533328345	1.1341934	1.024556823	1.231144413
mMC020628	RIKEN cDNA A530045L16 gene (A530045L16Rik)	2.887858391	2.37841423	1.602139755	0.702222438	1.144724161
mMC018948	ganglioside-induced differentiation-associated-protein 10 (Gdap10)	2.130216407	1.729074463	1.427344254	1.079228237	1.098092814
mMC009638	caspase 4, apoptosis-related cysteine peptidase (Casp4)	2.07915887	1.148698355	1.450617005	1.04608494	1.283425898
mMC012064	Fanconi anemia, complementation group F (Fancf)	14.84826257	10.82778819	1.967913307	0.999422544	1.31494276
mMC010121	SLAM family member 8 (Slamf8)	2.105708241	1.885569072	NA	0.799682931	1.057018041
mMR100086	predicted gene, EG634650 (EG634650)	2.015656777	2.271008858	1.295342252	0.917004043	1.176906737
mMA031507	caspase 4, apoptosis-related cysteine peptidase (Casp4)	2.177994031	1.316462719	1.662475792	0.979420298	1.00695555
mMC010522	pre-B-cell colony-enhancing factor 1 (Pbef1)	1.970946873	1.733074092	1.861759432	1	1.094293701
mMC015201	heat shock protein 1B (Hspa1b)	2.620786808	1.337927555	1.453972517	0.915416372	1.370782805
mMC007031	interferon-induced protein 44 (Ifi44)	2.319799309	1.265756594	1.536875181	0.974341891	1.057018041
mMC009969	MAD homolog 7 (Drosophila) (Smad7)	0.491410299	0.780966913	0.770215111	0.806641759	1.602139755
mMC012296	spinster homolog 2 (Drosophila) (Spns2)	0.364502345	0.476318999	0.450625231	0.957603281	1.664397469
mMC003600	brain expressed myelocytomatosis oncogene (Bmyc)	0.403880389	0.556068043	0.50580972	0.833642075	0.873572896
mMC013349	complement receptor 2 (Cr2)	0.447911532	0.341510064	0.395934403	0.920187651	1.125058485
mMC100872	X Kell blood group precursor related X linked (Xkrx)	0.287494676	NA	NA	0.772442795	1.064370182
mMC013531	RIKEN cDNA D230012E17 gene (D230012E17Rik)	0.419413279	0.427303587	0.562529242	0.876605721	1.021012126
mMC017935	Ral GEF with PH domain and SH3 binding motif 2 (Ralgps2)	0.457232545	0.38778619	0.583713902	0.891928519	1.296839555
mMR029994	hydrogen voltage-gated channel 1 (Hvcn1)	0.418413121	0.495400307	0.444421341	0.931417569	0.986232704

SUPPLEMENTAL FIGURES

Figure S1. **Redundancy and specificty in interferon signaling.** (A) The distribution of percent reduction in fold-induction for individual ISG in IFN receptor knockout mice. (B) The sum of the average magnitudes of T1IFN and IFNG gene induction amount to more than the whole observed in wild type mice, indicating redundancy in gene expression. Each point represents a different probe, and lines represent the linear regression and 95% confidence interval. (C) A subset of ISG exhibit preferential induction by either T1IFN or IFNG. The $\log_2$ fold induction of the 117 early response genes are plotted for *Ifnar1*$^{-/-}$ and *Ifngr1*$^{-/-}$ mice to identify preferentially induced genes. Residuals from identity (x = y) were calculated, and an arbitrary cutoff of 1.4 was chosen to highlight the most distant genes (*i.e.* the most preferentially induced genes). Green points represent genes preferentially induced by IFNG, and red points denote genes preferentially induced by T1IFN.





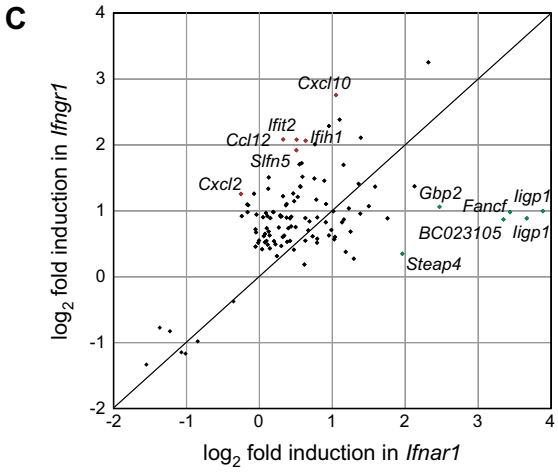
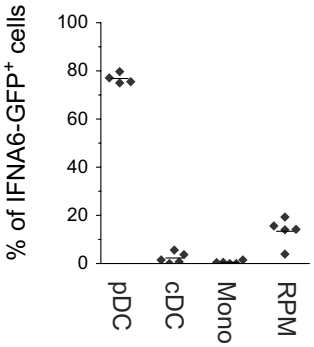


Figure S2. Induction of *Ifna6-Gfp* expression in splenic leukocytes by *P. chabaudi*. GFP$^+$ events were analyzed for lineage markers 24 h after infection.



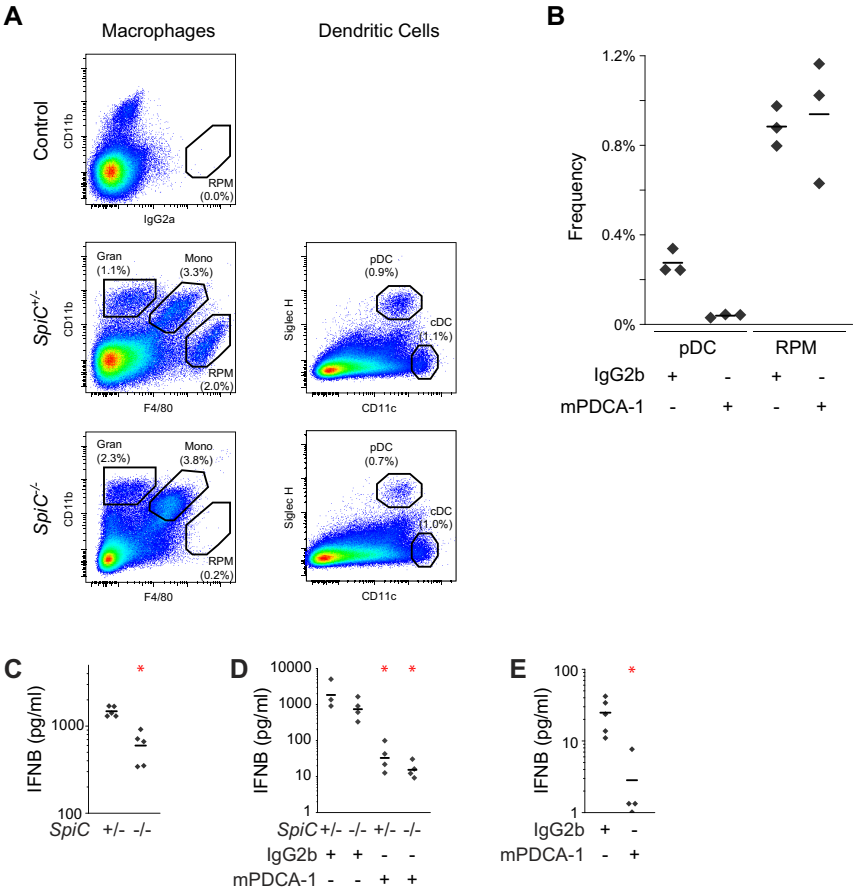


Figure S3. (A) Live singlet cells were subjected to lineage marker analysis for myeloid populations, demonstrating that *SpiC*$^{-/-}$ mice exhibit reduced RPM frequency compared to *SpiC*$^{+/-}$ animals, but otherwise have intact splenic macrophage and dendritic cell populations. (B) Treatment of C57BL/6 mice with mPDCA-1 antibody depletes splenic pDC populations but does not affect red pulp macrophages. Data represents frequencies measured after 18h depletion plus 24h infection with *P. chabaudi*. (C) Plasma IFNB levels are diminished in 129 *SpiC*$^{-/-}$ compared to 129 *SpiC*$^{+/-}$ mice. (D) Deficiencies in RPM and pDCs diminish the plasma IFNB response to *P. chabaudi* in 129 *SpiC*$^{-/-}$ mice. (E) Depletion of pDCs in C57BL/6 mice decreases the plasma IFNB response to *P. chabaudi*. Asterisks represent $p < 0.05$ in a two-tailed *t*-test assuming unequal variances compared with intact controls.

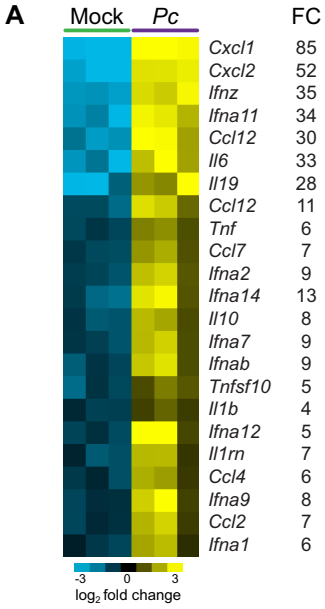


Figure S4. RPMs induce expression of *Ifna* and other cytokines and chemokines in response to *P. chabaudi* infection. (A) RNA was harvested from FACS-isolated RPMs from mock- or *P. chabaudi*-infected C57BL/6 animals, amplified, and hybridized to microarrays. A representative set of cytokines and chemokines induced upon infection are shown with fold change in transcript abundance. (B) Plasma cytokines of 129 *SpiC*$^{+/-}$ and *SpiC*$^{-/-}$ mice infected for 24 h with *P. chabaudi* were measured using Milliplex analysis (Millipore) on a MagPix instrument (Luminex). Differences between *SpiC*$^{+/-}$ and *SpiC*$^{-/-}$ mice are significant by a two-tailed *t*-test assuming unequal variances ($\alpha = 0.05$; red asterisks).

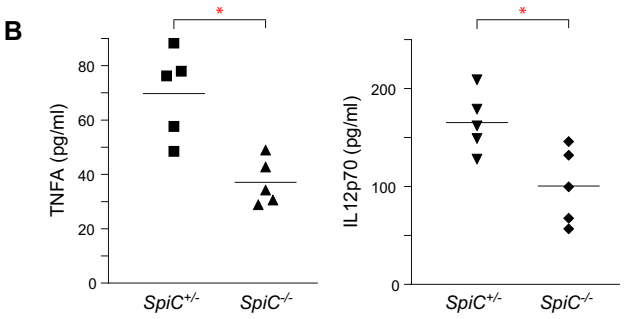


Figure S5. Basal C(t) values for leukocyte subsets. FACS-sorted populations from mock-infected animals were subjected to qRT-PCR for T1IFN transcripts. The data were aggregated from 4 independent experiments, with means and 95% confidence intervals represented. No significant differences were observed for any populations.

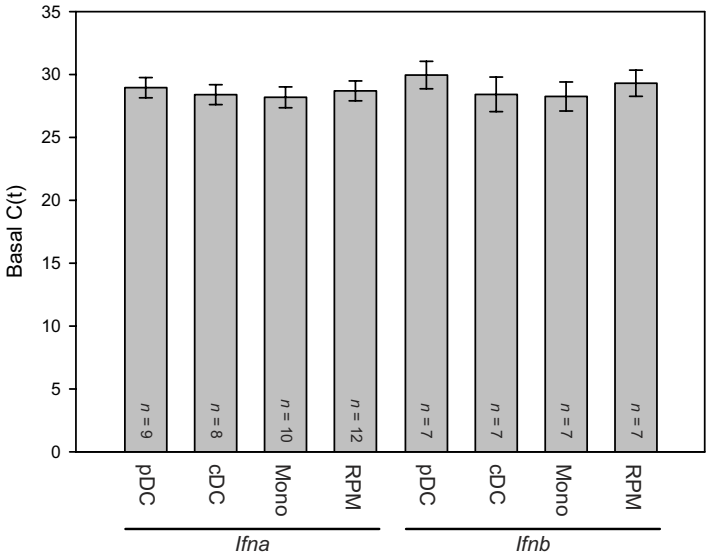
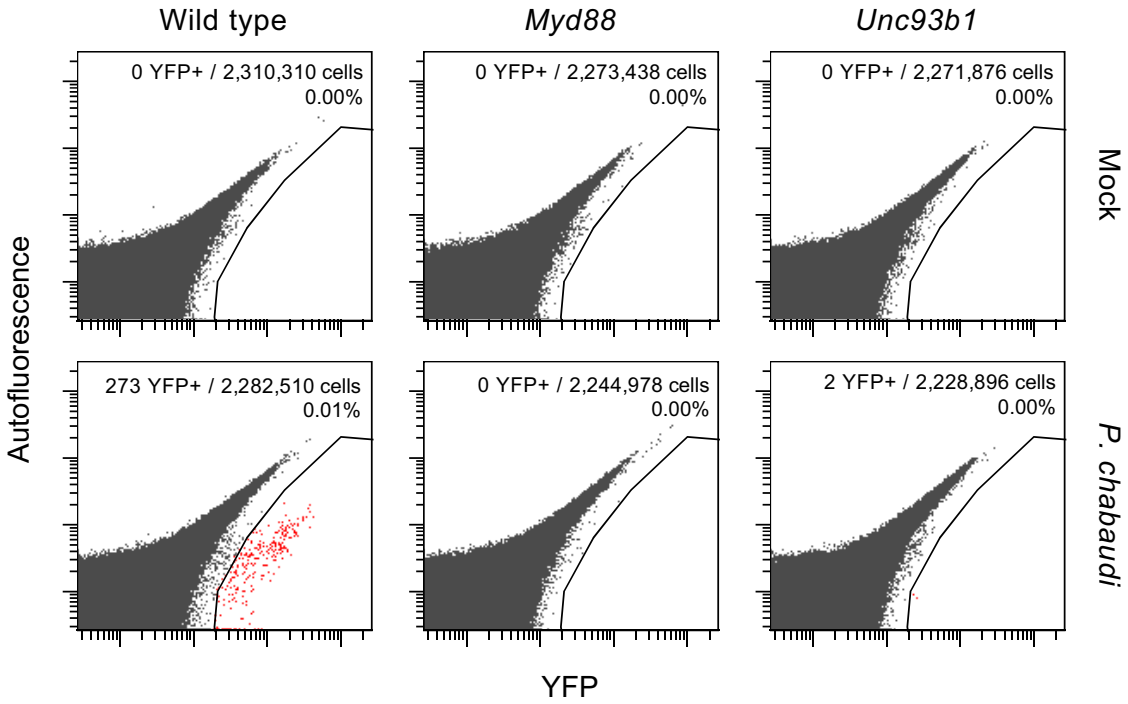


Figure S6. MYD88 is required for *Ifnb-Yfp* induction. Mice were inoculated with 10^6 infected erythrocytes or mock-infected with uninfected erythrocytes and spleens were harvested and processed for flow cytometry 24 h later.



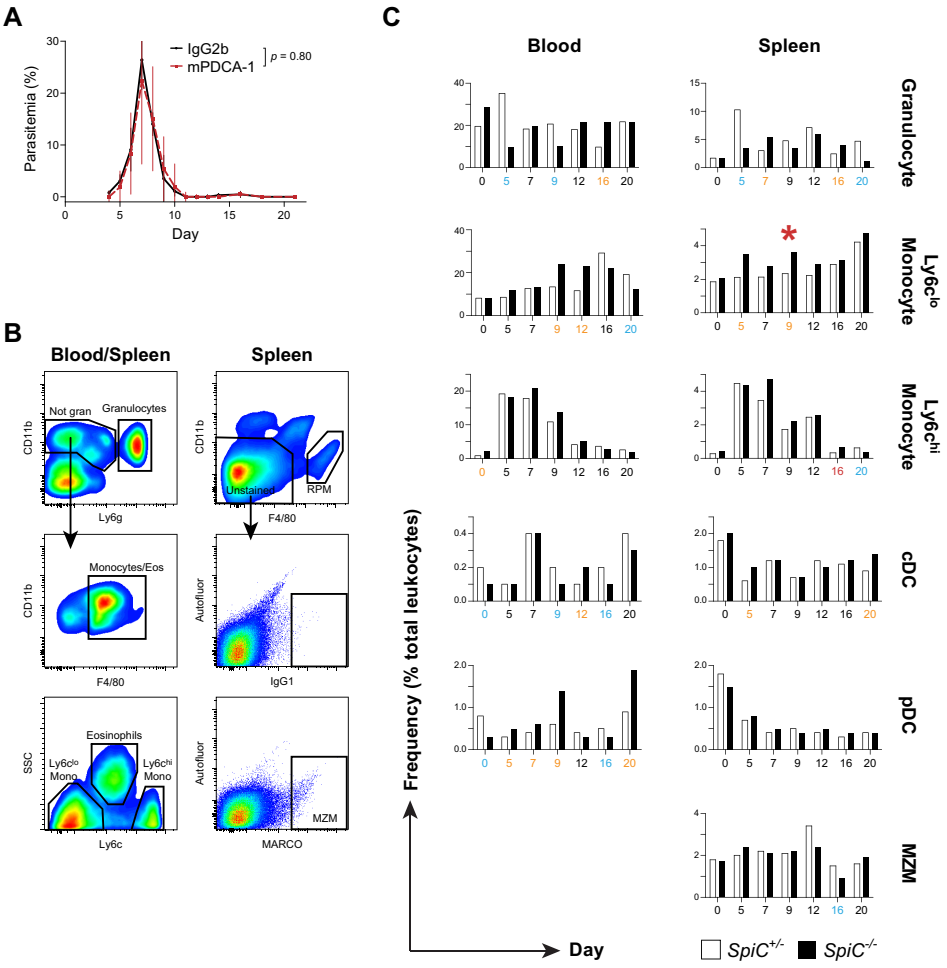


Figure S7. pDC and RPM are both dispensable for the control of *P. chabaudi* parasitemia. (A) C57BL/6 mice were intraperitoneally infected with 10^6 parasites. On day 4 post-infection, 500 ug of anti-mPDCA-1 antibody (Miltenyi Biotec) or IgG2b isotype control antibody (LTF2, UCSF hybridoma core) was administered intraperitoneally. Parasitemias are presented as geometric means with standard deviations and Mann-Whitney *p*-value. (B) Gating strategy for identification of myeloid populations in blood and spleen. Live singlet cells (not shown) were subjected to lineage marker analysis. MZM = marginal zone macrophages. (C) Myeloid population frequencies in blood and spleens of 129 *SpiC*$^{+/-}$ and *SpiC*$^{-/-}$ mice infected with *P. chabaudi* for 20 days. Days depicted in blue and orange represent a 1.5-fold decrease or increase, respectively, in frequency in *SpiC*$^{-/-}$ mice compared to *SpiC*$^{+/-}$ mice; red asterisks represent a significant difference over the entire infection course (Wilcoxon matched pairs signed rank test, $\alpha = 0.05$).

Chapter 3…….

***Bartonella quintana* deploys host and vector temperature-specific transcriptomes**

Stephanie Abromaitis*, Christopher S. Nelson*, Domenic Previte, Kyong S.Yoon, J. Marshall Clark, Joseph L. DeRisi, Jane E. Koehler

* = these two authors contributed equally

Author contributions:

Stephanie Abromaitis* conceived and designed experiments,cultured *Bartonella* , harvested and processed samples performed qPCR and wrote the paper. Christopher S. Nelson* conceived and designed experiments, designed the microarray, harvested and processed samples, performed the microarray analysis and wrote the paper. Domenic Previte, Kyong S.Yoon, and J. Marshall Clark performed lice culture of *Bartonella*, Joseph L. DeRisi, and Jane E. Koehler conceived and designed experiments, and helped write the paper

Joseph L. DeRisi, Thesis Advisor

Abstract

The bacterial pathogen *Bartonella quintana* is passed between humans by body lice. *B. quintana* has adapted to both the human host and body louse vector, producing persistent infection with high titer bacterial loads in both the host (up to 10^5 CFU / ml) and vector (over 10^8 CFU / ml). Using a novel custom microarray

platform, we analyzed bacterial transcription at temperatures corresponding to the host (37°C) and vector (28°C), to probe for temperature-specific and growth phase-specific transcriptomes. We observed that transcription of 7% (93 genes) of the *B. quintana* genome is modified in response to growth phase, and that 5% (68 genes) of the genome is temperature-responsive. Among these changes were the induction of known *B. quintana* virulence genes and several previously unannotated genes in response to temperature and growth phase changes. Hemin binding proteins, secretion systems, and genes for invasion and cell attachment were prominent among the differentially-regulated *B. quintana* transcriptional responses. This study represents the first analysis of global transcriptional responses by *B. quintana* and provides insight into the niche-specific gene expression involved in the transition of *B. quintana* between the human host and body louse vector.

Introduction

In the last three decades, there has been a resurgence of *Bartonella quintana* infections, with the most severe illness occurring among immunocompromised people [1]. *B. quintana* is a vector-borne Gram-negative bacterium; the vector is the human body louse (*Pediculus humanus humanus)* [2]. In a recent analysis, 33.3% of body lice recovered from infested homeless individuals in California were PCR positive for *B. quintana*, underscoring the high prevalence of this potentially fatal bacterium in the human environment [3]. The *B. quintana* bacteria colonize the louse alimentary tract and establish a life-long commensal relationship within the gut of the body louse, enabling a

single louse to infect multiple humans [4]. Upon introduction into the human host, *B. quintana* can persist in the normally sterile bloodstream for weeks or months [5]. This remarkable, prolonged persistance in the host bloodstream demonstrates the ability of *B. quintana* to avoid clearance by the host immune defenses [6]. Persistent *B. quintana* infections manifest in humans as relapsing fever, endocarditis, and potentially fatal vascular proliferative lesions.

During the infectious cycle, *B. quintana* alternates between two environmental niches: the bloodstream of the human host and the gut of the body louse vector. One important difference between these two niches is the ambient temperature: 37°C in the human bloodstream, and approximately 28°C within the louse gut [7]. To maintain the transmission cycle, *B. quintana* must rapidly deploy the appropriate growth programs to survive and proliferate in the two different environments of host and vector. Modification of the bacterial transcriptome in response to temperature change has been documented in the vector-borne human pathogens *Borrelia burgdorferi* [8], *Yersinia pestis* [9], *Francisella tularensis* [10], and *Rickettsia* spp. [11,12,13]. Temperature shift experiments have provided powerful insight into the adaptation of virulence and metabolic programs necessary for niche-specific infection with these vector-borne pathogens [8,9,10,12].

The reponse and adaptation of *B. quintana* to the distinct niches it occupies has not been studied using global transcriptional analysis. To define the *B. quintana* host- and vector-specific transcriptomes, we designed the first *B. quintana* whole genome DNA microarray (printed by Agilent Technologies, Santa Clara, CA). The array contains 60-mer oligos representing protein coding regions, intergenic regions, and RNA genes. The

coverage is approximately six oligos per gene, yielding highly sensitive transcriptional analysis.

In this study, we used the *B. quintana* array to identify growth phase-specific genes and genes that are differentially expressed at host (37°C) and vector (28°C) temperatures. We determined that transcription of 7% (93 genes) of the *B. quintana* genome is modified in response to entry into stationary/death phase, and that 5% (68 genes) of the genome is temperature-responsive. Additionally, analysis of *B. quintana* transcription in infected body lice demonstrated that genes that are highly transcribed at 28°C *in vitro* also were highly transcribed *in vivo*, in the body louse. The temperature-specific genes we identified included type 4 secretion system (T4SS) components, members of the hemin binding protein family, and several previously unannotated genes. Collectively, these temperature-specific genes provide a model of the transcriptional program of the *B. quintana* transition from arthropod vector to human host.

Materials and Methods

Bacterial strains and growth conditions

For all experiments, low-passage *B. quintana* strain JK31 was used. The JK31 strain was isolated from a patient co-infected with *B. quintana* and HIV [14]. JK31 *B. quintana* bacteria were streaked onto fresh chocolate agar plates [14] from frozen stock and were grown for 6-7 days in candle extinction jars at 35°C prior to passage and use in experiments. The liquid media used for *B. quintana* was M199S, which consists of M199 supplemented with 20% fetal bovine serum, glutamine, and sodium pyruvate [15]. For microarray transcription profiling

experiments, *B. quintana* JK31 strain bacteria that had been passed once since plating from frozen stock were harvested from 2 confluent chocolate agar plates and resuspended in M199S to a final concentration of 0.6 at OD_{600}. A total of 12 biological samples were profiled from two independent timecourses. 100 μl of the bacterial suspension was plated on each chocolate agar plate. Plates were grown in candle extinction jars at 37°C for 48 h, and then a portion of the jars were shifted to 28°C to grow for the remainder of the temperature shift experiment.

B. quintana genomic DNA, RNA, and cDNA preparation from bacteria grown on chocolate agar plates

B. quintana genomic DNA was isolated using the Qiagen Purgene Core Kit B following the manufacturer's instructions. For RNA isolation, *B. quintana* were harvested from confluent plates into 1 ml stop solution (M199, 45% EtOH, 5% water-saturated phenol) to prevent RNA degradation [16]. Bacteria were then pelleted by centrifugation at 4,000 x g at 4°C. The bacteria pellet was stored at -80°C until RNA was isolated. Bacterial cells were lysed by incubating in fresh lysozyme (0.4 mg ml^{-1} in 10 mM Tris, 1 mM EDTA) for 5 min at room temperature. The RNA was extracted using TRIzol reagent (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions. Total RNA was RQ1 DNase (Promega, Madison, WI) treated for 3 h and then further purified using the RNeasy® Mini Kit (Qiagen, Valencia, CA). For reverse transcriptase-quantitative PCR (RT-qPCR) analysis, cDNA was generated from 0.5 μg of total RNA using

random hexamer primers (Invitrogen) and SuperScript™ III (Invitrogen) following the manufacturer's instructions. Reverse transcription reactions without Superscript™ III were performed as negative controls and to evaluate DNase treatment efficiency.

B. quintana cDNA generation and labeling for microarray hybridization

For microarray analysis, cDNA was generated from 15 μg of total RNA. RNA was combined with 15 μg random nonamer primers (Integrated DNA Technology, San Jose, CA) and 1.8 μl of A/T biased amino-allyl mix in a total of 30 μl. A/T biased amino-allyl mix consisted of 100 mM dATP, 100 mM dGTP, 100 mM dCTP, 100 mM dTTP, and 50 mM amino allyl-dUTP at a ratio of 5:2.5:2.5:1:8. cDNA reactions were incubated 5 min at 65°C and then incubated for a minimum of 1 min on ice. 30 μl of reverse transcription mix consisting of Invitrogen reagents (12 μl 5x reverse transcription buffer, 3 μl 0.1 M DTT, 3 μl RNaseOUT, 3 μl SuperScript™ III, and 4.2 μl H_2O) was added to each reaction. The reactions were incubated for 12 min at 25°C and then for 8 h at 46°C. An additional 3 μl of SuperScript™ III were added to each reaction and the reactions were incubated for an additional 8 h at 46°C. cDNA generation was terminated by a 5 min incubation at 85°C. Reactions were chilled on ice and then treated with RNaseA to degrade remaining RNA. The cDNA was purified using Zymo Research (Irvine, CA) DNA Clean & Concentratrator™-25 Kit, following the manufacturer's instructions. Amino-allyl cDNA aliquots were coupled to Cy5 and Cy3 (GE Health Sciences, Piscataway, NJ) in 1.0M pH 9.0 sodium bicarbonate

for 1 h, and then cleaned up with Zymo Research DNA Clean & Concentratrator™-25 Kit.

B. quintana genome-wide transcriptional profiling

A custom microarray with 15,744 probes was designed using the *B. quintana* Toulouse genomic sequence deposited at NCBI (NC_005955.1), and the annotations found at the Microbial Genome Database for Comparative Analysis (http://mbgd.genome.ad.jp/htbin/MBGD_gene_list.pl?spec=bqu) and JCVI (http://cmr.jcvi.org/cgi-bin/CMR/GenomePage.cgi?org=ntbq01) [17]. Gene sequences were extracted from the genomic sequence with nibFrag Software (http://hgwdev.cse.ucsc.edu/~kent/src/unzipped/utils/nibFrag/, Jim Kent, University of California, Santa Cruz). 60mer probes were chosen with ArrayOligoSelector software (http://arrayoligosel.sourceforge.net/) with up to 10 oligos per gene. Arrays were ordered in 8 x 15K format from Agilent Technologies (Santa Clara, CA) (design amaID 025396).

Hybridization mix was comprised of 10 μl of Cy5 labeled sample, 10 μl of Cy3 labeled pooled reference sample, 5 μl blocking buffer, and 25 μl of Agilent Gene Expression Buffer. Hybridizations were carried out at 65°C for 16-19 h in Agilent hybridization chambers rotating at 10 rpm in a convection oven. After hybridization, the arrays were washed with Agilent wash buffers following the manufacturer's instructions. Image data were acquired taking care to balance the observed total fluorescence in the Cy3 and Cy5 channels. Images were scanned on a Genepix 4000B scanner (Molecular Devices, Union City, CA) and

data were extracted using Genepix6.0 software in the Center for Advanced Technology (CAT) at University of California, San Francisco.

Microarray analysis

Raw array data were uploaded to Nomad v2.0 (http://ucsf-nomad.sourceforge.net/), where the data were normalized in bins of pixel intensity R^2, and then filtered to remove spots with "bad" or "missing" manual flags added during gridding, and spots with sum of median intensities less than 1000. The resulting ratio Cy5/Cy3 intensity tables were $\log_2$ transformed and re-centered at 0. The $\log_2$ transformed data were then analyzed using cluster 3.0 (Eisen laboratory, University of California, Berkeley) and SAM (SAM version 3.0, http://www-stat.stanford.edu/~tibs/SAM/) [18]. SAM results were reported as a ranked list of d-scores that represent the difference between two groups of array data compared to a background of randomly shuffled data with associated false discovery rates (fdr%). The GEO array data accession number is GSE42685, and the array design record is GPL16349.

Annotation of unannotated temperature-responsive *B. quintana* genes

We translated the ORFs of gene models with annotations of "hypothetical gene" and ran a blastp query against the nr database with expect threshold 1 and word size of 3. We submitted the same sequences to pHMMER (http://hmmer.janelia.org/search/phmmer), querying against the nr database with

sequence E-value cutoff 0.01 and hit E-value cutoff of 0.01, and the default gap-open penalty of 0.02 and gap extension penalty of 0.4.

Motif search upstream of temperature-regulated *B. quintana* genes

To identify potential *cis* elements involved in the observed temperature response, motif searches with MEME were performed on the list of regulated genes. The promoters of all the genes in the differentially regulated lists were taken from the *B. quintana* Toulouse strain genome using Mochiview (http://johnsonlab.ucsf.edu/mochi.html) [19], and then submitted to the motif search algorithm MEME. MEME (http://meme.sdsc.edu/meme/cgi-bin/meme.cgi) searches looked for any number of repetitions of motifs within a sequence for motifs from 6 to 11 bases in length within all of the genes, 37°C genes, 28°C genes, and the top 11 28°C genes. A significant motif was returned only for the search of promoter regions down-regulated at 28°C.

Quantitative gene expression analysis from *B. quintana* grown *in vitro*

For verification of microarray results, reverse transcriptase-quantitative PCR (RT-qPCR) was performed using a MX3000P machine (Stratagene/Agilent Technologies, Santa Clara, CA) to determine the relative abundance of specific mRNA. cDNA was diluted 1:19 for use in reactions. The reaction mixture included: 10 μL SYBR Fast qPCR master mix (Kapa Biosystems, Woburn, MA), 0.4 μl ROX low (Kapa Biosystems), 7.6 μl template, and 2 μl 1 pmol μl^{-1} primer. The reaction conditions were: 95°C for 10 min, 40 cycles of 95°C for 15 s and

60°C for 60 s, with dissociation protocol. Threshold fluorescence was determined during the geometric phase of logarithmic gene amplification; from this, the quantification cycle (C_q) was set. Standard curves for each primer set were generated by plotting log genomic DNA vs. C_q. These plots were used to ensure that equivalent reaction efficiency was obtained with all primer sets. Primers used are listed in Table S1. The relative level of gene transcript in samples was determined by converting transcript level into genomic copy number using standard curves. This value was divided by the genomic copy number of the constitutively expressed *B. quintana* reference gene *purA* (adenylosuccinate synthetase) or 16S rRNA, to obtain a relative level of transcription for each gene. Data from three independent experiments were used for statistical analysis by Student's *t* test and to determine average gene transcription values.

Body lice infection with *B. quintana*

The body louse (*Pediculus humanus humanus*) strain SF was collected in San Francisco, CA. Collected lice were maintained on human blood using an *in vitro* rearing system [20]. The lice were maintained under conditions of 30 °C, 70–80% relative humidity, and 16L:8D light-dark cycles in a rearing chamber. The human blood (American Red Cross, Dedham, MA) used for feeding was comprised of 25 ml fresh red blood cells (blood type A+) and 25 ml plasma (blood type A+), supplemented with 25 µl of a penicillin plus streptomycin antibiotic

mixture (10,000U penicillin and 10mg streptomycin per ml, in 0.9% NaCl) [21]. Prior to infection, lice were fed blood without antibiotic supplement for 2-3 days.

B. quintana strain JK31 was used in the body lice infections. Bacteria were harvested from chocolate agar plates, washed with PBS, and then resuspended in human blood without antibiotics for the infection, at a final concentration of $5.77 \times 10^8 \pm 1.20 \times 10^8$ bacteria per ml blood. Female SF strain lice were starved 8 h prior to infection, to ensure feeding on the *B. quintana*-inoculated blood. The lice were fed for 24 h on infected blood or control blood, to which PBS without bacteria had been added. Throughout the remainder of the experiment, lice were fed on uninfected human blood. Populations of uninfected and infected lice were removed from the colony and snap frozen in liquid nitrogen immediately after the 24 h feeding on the *B. quintana*-containing blood (1 day post-inoculation [dpi]), 5 days after the commencement of feeding (5 dpi), or 9 days after the commencement of feeding (9 dpi).

B. quintana genomic DNA, RNA, and cDNA preparation from body lice

For genomic *B. quintana* DNA isolation, lice were homogenized in ATL buffer (Qiagen) using a glass Dounce homogenizer. The homogenate was digested with Proteinase K for 16 h at 56°C and then treated with RNaseA. The DNA was then isolated using a Qiagen DNeasy Blood and Tissue kit following the manufacturer's instructions. The number of *B. quintana* bacteria per louse was determined using the isolated genomic DNA as template for RT-qPCR. The C_q value was used to calculate the DNA copy number by comparison to standard

curves. The number of amplified DNA copies was converted into the number of *B. quintana* bacteria assuming 1 attomole gDNA = 3.01 x 10^5 cells [22]. Primers used for bacterial quantification are listed in Table S1.

For *B. quintana* RNA isolation, lice were homogenized in RLT buffer using a glass Dounce homogenizer and treated with lysozyme. The sample was then further homogenized using QIAshredder™ columns (Qiagen) following the manufacturer's instructions. RNA was purified from the homogenate using Qiagen RNeasy® Mini Kit following the manufacturer's instructions. The purified RNA was used as template for cDNA synthesis following the protocol above.

Results and Discussion

A cluster of growth-phase specific genes is identified in *B. quintana* grown at either 37°C or 28°C. To ensure that *B. quintana* cultures were in the same phase of growth at 37°C and 28°C, it was first necessary to develop a reproducible and growth stage-matched experimental scheme. Agar-grown cultures of *B. quintana* were synchronized at the two different temperatures, as shown in Figure 1A. Agar media was used for the analysis because we found insufficient growth of *B. quintana* in liquid culture at 28°C. To identify *B. quintana* growth phase-specific genes, bacteria grown at 37°C or 28°C were harvested 3 to 7 days or 3 to 9 days after plating, respectively (Figure 1A). At each time point, replicate plates were harvested for colony-forming unit (CFU) enumeration.

At 28°C, *B. quintana* demonstrated a brief period of exponential growth on agar, followed by a prolonged stationary phase; death phase was not observed

over the 9 days of growth at 28°C (Figure 1B). At 37°C, *B. quintana* exhibited active growth (log phase) 3, 4, and 5 days after plating on solid agar; this was followed by a rapid death phase (Figure 1B). We did not observe a sustained stationary phase at 37°C. Prior to our analysis, *B. quintana* growth dynamics had not been analyzed at the vector temperature of 28°C in any culture medium, but growth of *B. quintana* in liquid media at 37°C or 35°C had been reported by several groups [22,23,24]. Similar to our results for agar-grown *B. quintana* at 37°C, cultivation of *B. quintana* in liquid media at 35°C or 37°C resulted in a rapid decrease in CFU per ml following the exponential growth phase, with no detectable stationary phase [22,23,24].

Analysis of the *B. quintana* transcriptional profile over time at 28°C and 37°C identified both growth stage- and temperature-responsive *B. quintana* genes. We determined that transition from active growth to stationary or death phase elicits a specific transcriptional profile, independent of the temperature at which the *B. quintana* is cultivated (Figure 2). In stationary/death phase, global SAM analysis of transcription identified 10 genes with significantly increased transcription and 83 genes with decreased transcription, (changes over 2 fold displayed in Table S2). Growth phase-specific virulence gene regulation has been well documented in a number of bacteria [25, 26, 27], and our *B. quintana* cultures exhibited a robust phase-specific response encompassing 93 significantly altered transcripts. Several of the stationary/death phase responsive genes we identified are associated with *Bartonella* virulence (Figure 2). Among the *B. quintana* virulence genes that were up-regulated during logarithmic phase

relative to stationary phase are components of the Trw T4SS (Figure 2). The Trw T4SS in *B. henselae* has been implicated in mediating host-specific erythrocyte adhesion [28], and is likely important for initial colonization of the mammalian host bloodstream by *Bartonella*. A cue provided by the growth phase could prepare the *B. quintana* bacteria for interaction with host erythrocytes after introduction into the host.

***B. quintana* has unique transcriptional profiles when grown at human host (37°C) compared with arthropod vector (28°C) temperature**. During the infectious cycle, *B. quintana* occupies the bloodstream of the human host and the alimentary tract of the body louse vector. Global transcription in *B. quintana* cultivated at either the human host temperature (37°C) or the body louse vector temperature (28°C) was analyzed to identify *B. quintana* niche-specific genes. For this analysis, bacterial transcription was evaluated during the logarithmic phase of growth at both temperatures (Figure 3). When bacteria were harvested for transcriptional profiling, CFU were enumerated from replicate plates to ensure that the bacteria were in the logarithmic growth phase (Figure 3).

Sixty-eight genes were differentially expressed at 37°C versus 28°C by SAM analysis, from replicate time courses (Table 1). Of the temperature-responsive genes, 56 had increased transcription at 28°C, and 12 had decreased transcription at 28°C, compared to 37°C. The results of the microarray transcriptional profiling experiments were validated by RT-qPCR. Three replicate temperature shift experiments were performed, and transcription of eight

temperature-regulated genes was analyzed. The RT-qPCR analysis corroborated the findings of the microarray experiments, with the level of transcription of each gene being significantly different at 28°C compared to 37°C (Figure 4).

We classified the temperature-responsive genes identified by our microarray analysis into functional categories, based on the classification scheme of the Cluster of Orthologous Groups (COG) database [29] (Figure 5). The temperature-regulated genes within COG functional category P (Inorganic ion transport and metabolism) were of particular interest because of their potential role in *B. quintana* hemin metabolism and detoxification. Hemin and hemoglobin are the only iron sources that *Bartonella* can metabolize [30], making acquisition and metabolism of these nutrients essential for bacterial survival. A major difference between the host and vector environments is the amount of free hemin available. The human bloodstream is extremely hemin restricted, whereas toxic levels of hemin are present in the body louse alimentary tract. Hemin can produce reactive oxygen molecules that are potentially toxic [31]. *Bartonella* is unique in its ability to survive exposure to hemin concentrations that are typically bactericidal (>1 mM) [30,32,33]. We identified four hemin-related proteins in COG functional category P that are up-regulated at 28°C: hemin binding protein A (*hbpA*), hemin binding protein C (*hbpC*), and heme exporter protein A and B (*ccmA, ccmB*) (Table 1).

As their name suggests, the hemin binding proteins (Hbp) bind hemin [34]. Previous analysis of temperature-specific transcription of the five *hbp* family

genes in *B. quintana* by Battisti *et al.* [35] identified *hbpC* as temperature-responsive. Similar to what is observed in *B. quintana,* in *B. henselae hbpC* displays increased expression at 28°C versus 37°C when cultivated on chocolate agar [36]. In *B. henselae*, up-regulated expression of *hbpC* at arthropod temperature ameliorates the antibacterial toxicity of the concentrated hemin in the arthropod gut [36]. Thus, the significant up-regulation of *hbpC* appears to be part of the critical hemin detoxification response in *Bartonella* species during adaptation to the arthropod niche.

Also prominent among the temperature regulated genes are some components of the VirB T4SS, a second T4SS (in addition to Trw) in *B. quintana*. The VirB T4SS apparatus is involved in the injection of effector proteins into host cells [37,38], and appears to have a different function from the Trw T4SS [39]. Of note, *virB2*, *virB3*, *virB4*, and *virB6* are highly up-regulated at 28°C; in contrast, *virB8-11* are growth-phase regulated (d-score 3.3-3.8) but not temperature regulated, so perhaps multiple environmental cues are integrated before producing the fully functional VirB secretion complex encoded on adjacent but distinct operons [17]. The Trw T4SS components are growth-phase regulated, supporting the differential function and responsiveness to environmental cues for these two *B. quintana* T4SS.

The expression of two response regulators (COG functional category T), encoded by *B. quintana phyR* (BQ10980) and *ompR* (BQ03390), were found to be temperature-responsive. Expression of *phyR* was increased 4-fold at 28°C versus 37°C (Table 1, Figure 4). *B. quintana* PhyR is a positive regulator of

RpoE (unpublished data); *B. quintana* RpoE is an alternative sigma factor that is involved in transcription of genes necessary for survival in the high hemin environment of the body louse gut (unpublished data). As predicted, we found that *phyR*, the positive regulator of *rpoE*, is one of the most highly transcribed genes at body louse temperature. Expression of the response regulator *ompR* was increased at 28°C (Table 1). In *B. henselae, ompR* is transcribed in response to contact with human endothelial cells [39], and OmpR has been shown to be involved in *B. henselae* invasion of human endothelial cells [40]. Our observation that *ompR* transcription is temperature regulated suggests that OmpR is involved in priming human endothelial cell invasion by *B. quintana* during the transition from body louse to mammalian host. Our data also suggest niche-specific roles for other, less-studied transcriptional regulators in *B. quintana* such as BQ08990 and BQ06490. BQ08990 has homology to the ArsR family of transcriptional regulators, which has a role in sensing environmental metal concentrations, and in the induction of pathogenicity in *Bacillus anthracis* and *Streptococcus mutans* [41,42,43]. BQ06490 has homology to the AsnC transcriptional regulators that are typically involved in environmentally-cued induction of alternative amino acid metabolic pathways [44]. In summary, it appears that ambient temperature drives niche adaptation by controlling expression of several transcriptional regulators (4 genes out of 37 genes annotated as transcriptional regulators).

Annotation of unannotated, temperature-responsive *B. quintana* genes reveals potential niche-specific virulence genes. Many of the genes identified as temperature-responsive were unannotated. We reevaluated the annotation of these genes using homology searches. The full-length peptide sequences of the temperature-responsive, unannotated *B. quintana* genes were evaluated using blastp and pHMMER search engines against nr database (http://hmmer.janelia.org/search/phmmer). We annotated 18 genes with E-values of 4.00E-08 or less as putative *B. quintana* homologs. These genes are shown in Table 2. In most cases, these improved gene annotations were corroborated by both the pHMMER and blastp search results.

One previously unannotated gene of particular interest was gene BQ00450, which was up-regulated at 37°C (Table 1). We classified this gene as a putative zinc metalloprotease (Table 2). Zinc metalloproteases are found in several pathogenic bacteria and have been implicated in bacterial invasion and pathogenicity in *Pseudomonas aeruginosa, Vibrio cholerae,* and *Bacillus anthracis* [45]. These metalloproteases act to cleave immune effector proteins and to remodel the niche for bacterial attachment. It is possible that BQ00450 has a similar role in *B. quintana* colonization of the human host.

We annotated the genes BQ10280 and BQ10290 as putative autotransporters, and identified orthologous genes in many other *Bartonella* spp (all give blastp hits with E-value <1E-129) (Table 2). Both of these putative autotransporter genes were highly up-regulated at 28°C (Table 1; Figure 4), and their genomic placement suggests that they could be co-transcribed as an

operon. Autotransporters serve a number of functions in bacteria; of particular interest, they are involved in adhesion [46,47,48] and in biofilm formation [49]. *B. quintana* adheres to body louse gut epithelial cells [50], and the bacteria form a biofilm-like structure within the louse feces [21], but the molecular mechanisms underlying both of these processes are unknown. These autotransporters, BQ10280 and BQ10290, which are highly expressed at the vector temperature, could be involved in *B. quintana* adhesion or biofilm formation in the body louse gut. VompD (BQ01390), a member of the variably-expressed outer membrane protein family of adhesins, also is upregulated at 28°C, and is another candidate for mediating adhesion to the louse gut epithelium [14].

A purine-rich, temperature-responsive, putative promoter motif is identified for genes up-regulated at body louse temperature (28°C). We analyzed the upstream intergenic sequences of the differentially-regulated, temperature-responsive genes to identify motifs that correlate with temperature-dependent changes in expression, using the MEME algorithm. The temperature-responsive genes up-regulated at 37°C did not produce any significant MEME results. MEME analysis of all the 28°C-specific genes, or just the upstream noncoding regions of the eleven genes most highly transcribed at 28°C, returned a single motif with E-value <0.1. This 8-mer motif was purine-rich ("AGRGRRRA"), with an E-value of 8.3×10^{-3}. Additionally, variants of this motif repeatedly scored well over a range of motif-length input parameters, from 6-mers to 12-mers (Figure 6A). The identified motif was present 35 times in 10 of the 11 upstream regions.

For example, this motif was repeated three times upstream of *hbpC* and four times upstream of genes in the *virB* T4SS operon (Figure 6b).

Quantification of *in vivo* transcripts in *B. quintana*-infected body lice corroborates up-regulated genes identified *in vitro* at 28°C by microarray.

From the *in vitro* microarray analysis, we identified a number of genes whose transcription was increased at 28°C and thus could represent genes critical for *B. quintana* colonization of the body louse vector. *In vivo* analysis of *B. quintana* transcription was performed to corroborate our *in vitro* microarray data. Female lice in a colony established from body lice removed recently from an infested person were used for the *in vivo* experiments. These lice were fed only human blood through an artificial membrane-rearing system [20]. This system provides a more natural model of infection than the Culpepper body louse laboratory strain that was adapted decades ago to feed only on rabbits [51]. The lice were infected by feeding for 24 hours on a *B. quintana*-inoculated human blood meal, and then were subsequently fed on uninfected human blood.

We first established that the number of *B. quintana* per louse increased over the course of the infection, by performing quantitative analysis of *B. quintana* proliferation in the infected body lice. Figure 7 documents infection of the body lice with viable, replicating *B. quintana*. Similar rates of *B. quintana* replication were observed in our study and the previous work by Seki *et al*. [21].

RNA was isolated from the lice 24 hours after feeding on the *B. quintana*-containing human blood meal and at five and nine dpi for transcriptional analysis.

B. quintana transcription of two genes (*hbpC* and BQ10280) that were highly expressed at 28°C by microarray analysis was evaluated. The relative level of transcription of *hbpC* and BQ10980 in lice was similar to that observed when the *B. quintana* were cultivated *in vitro* on chocolate agar plates at 28°C, and was greater than that observed when the bacteria were cultivated on chocolate agar plates at 37°C (Figure 8). Transcription of both genes was greatest at 1 dpi, suggesting that HbpC and BQ10280 could have an important role in initial vector colonization. This is the first reported quantification of *B. quintana* transcription within the body louse environment.

Conclusions

B. quintana must survive and proliferate within the body louse vector as well as the human host during the course of the infectious cycle. Each of these niches presents the *B. quintana* bacteria with unique nutritional and environmental conditions. To begin to understand how *B. quintana* adapts to each environment, we analyzed global transcription in bacteria grown at temperatures corresponding to the human host (37°C) or the body louse vector (28°C). We observed unique patterns of gene expression at each of these two niche-associated temperatures. These genes included temperature-specific virulence factors with known or predicted roles in secretion, iron binding and transport, and regulation of transcription. For some of the genes that were only described as encoding "hypothetical proteins," we improved the annotation and identified additional, potential virulence genes whose expression is temperature-regulated.

Upstream of some of the genes that were up-regulated at 28°C, we found a conserved purine-rich motif that could permit coordinate transcription of temperature-regulated, niche-specific *B. quintana* genes.

Our *in vitro* data were corroborated *in vivo* using a novel model for body louse infection that recapitulates the natural route of infection of body lice with *B. quintana*. The louse infection model utilizes an artificial membrane-feeding system [20] that enabled us to feed lice on human blood inoculated with *B. quintana.* From the perspective of transcriptional regulation, we found that the transition from mammalian host to arthropod vector temperature principally involves deployment of different hemin binding systems and the preparation of export systems to adapt to the new niche. In the course of this work, we have developed important tools (*in vitro* whole genome *B. quintana* DNA microarray and *in vivo* body louse infection with *B. quintana*) that provide a new understanding of *B. quintana* host and vector adaptation and will allow further study of the host-vector relationship.

Acknowledgments

The authors thank Bela Cuperstein and the Center for Advanced Technologies at UCSF for technical support, and Kaman Chan and Charlie Kim for technical advice.

References

1. Greub G, Raoult D (2004) *Bartonella* Infections Resurgence in the New Century In: Fong IW, Drlica K, editors. Emerging Infectious Diseases of the 21st Century: Springer US. pp. 35-68.
2. Raoult D, Roux V (1999) The body louse as a vector of reemerging human diseases. Clin Infect Dis 29: 888-911.
3. Bonilla DL, Kabeya H, Henn J, Kramer VL, Kosoy MY (2009) *Bartonella quintana* in body lice and head lice from homeless persons, San Francisco, California, USA. Emerg Infect Dis 15: 912-915.
4. Vinson JW, Varela G, Molina-Pasquel C (1969) Trench fever. 3. Induction of clinical disease in volunteers inoculated with *Rickettsia quintana* propagated on blood agar. Am J Trop Med Hyg 18: 713-722.
5. Foucault C, Barrau K, Brouqui P, Raoult D (2002) *Bartonella quintana* bacteremia among homeless people. Clin Infect Dis 35: 684-689.
6. Pulliainen AT, Dehio C (2012) Persistence of *Bartonella* spp. stealth pathogens: from subclinical infections to vasoproliferative tumor formation. FEMS Microbiol Rev 36: 563-599.
7. Wigglesworth VB (1941) The sensory physiology of the human louse *Pediculus humanus corporis* de Geer (Anoplura). Parasitology 33: 67-109.
8. Revel AT, Talaat AM, Norgard MV (2002) DNA microarray analysis of differential gene expression in *Borrelia burgdorferi*, the Lyme disease spirochete. Proc Natl Acad Sci U S A 99: 1562-1567.

9. Han Y, Zhou D, Pang X, Song Y, Zhang L, et al. (2004) Microarray analysis of temperature-induced transcriptome of *Yersinia pestis*. Microbiol Immunol 48: 791-805.

10. Horzempa J, Carlson PE, Jr., O'Dee DM, Shanks RM, Nau GJ (2008) Global transcriptional response to mammalian temperature provides new insight into *Francisella tularensis* pathogenesis. BMC Microbiol 8: 172.

11. Audia JP, Patton MC, Winkler HH (2008) DNA microarray analysis of the heat shock transcriptome of the obligate intracytoplasmic pathogen *Rickettsia prowazekii*. Appl Environ Microbiol 74: 7809-7812.

12. Dreher-Lesnick SM, Ceraul SM, Rahman MS, Azad AF (2008) Genome-wide screen for temperature-regulated genes of the obligate intracellular bacterium, *Rickettsia typhi*. BMC Microbiol 8: 61.

13. Ellison DW, Clark TR, Sturdevant DE, Virtaneva K, Hackstadt T (2009) Limited transcriptional responses of *Rickettsia rickettsii* exposed to environmental stimuli. PLoS One 4: e5612.

14. Zhang P, Chomel BB, Schau MK, Goo JS, Droz S, et al. (2004) A family of variably expressed outer-membrane proteins (Vomp) mediates adhesion and autoaggregation in *Bartonella quintana*. Proc Natl Acad Sci U S A 101: 13630-13635.

15. Koehler JE, Quinn FD, Berger TG, LeBoit PE, Tappero JW (1992) Isolation of *Rochalimaea* species from cutaneous and osseous lesions of bacillary angiomatosis. N Engl J Med 327: 1625-1631.

16. Gaynor EC, Cawthraw S, Manning G, MacKichan JK, Falkow S, et al. (2004) The genome-sequenced variant of *Campylobacter jejuni* NCTC 11168 and the original clonal clinical isolate differ markedly in colonization, gene expression, and virulence-associated phenotypes. J Bacteriol 186: 503-517.

17. Uchiyama I, Higuchi T, Kawai M (2010) MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity. Nucleic Acids Res 38: D361-365.

18. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 98: 5116-5121.

19. Homann OR, Johnson AD (2010) MochiView: versatile software for genome browsing and DNA motif analysis. BMC Biol 8:49.

20. Yoon KS, Strycharz JP, Gao J-R, Takano-Lee M, Edman JD, et al. (2006) An improved *in vitro* rearing system for the human head louse allows the determination of resistance to formulated pediculicides. Pesticide Biochemistry and Physiology 86: 195–202

21. Seki N, Kasai S, Saito N, Komagata O, Mihara M, et al. (2007) Quantitative analysis of proliferation and excretion of *Bartonella quintana* in body lice, *Pediculus humanus*. Am J Trop Med Hyg 77: 562-566.

22. Lynch T, Iverson J, Kosoy M (2011) Combining culture techniques for *Bartonella*: the best of both worlds. J Clin Microbiol 49: 1363-1368.

23. Maggi RG, Duncan AW, Breitschwerdt EB (2005) Novel chemically modified liquid medium that will support the growth of seven *Bartonella* species. J Clin Microbiol 43: 2651-2655.

24. Riess T, Dietrich F, Schmidt KV, Kaiser PO, Schwarz H, et al. (2008) Analysis of a novel insect cell culture medium-based growth medium for *Bartonella* species. Appl Environ Microbiol 74: 5224-5227.

25. Navarro Llorens JM, Tormo A, Martinez-Garcia E (2010) Stationary phase in gram-negative bacteria. FEMS Microbiol Rev 34: 476-495.

26. Mangan MW, Lucchini S, Danino V, Croinin TO, Hinton JC, et al. (2006) The integration host factor (IHF) integrates stationary-phase and virulence gene expression in *Salmonella enterica* serovar Typhimurium. Mol Microbiol 59: 1831-1847.

27. Bachman MA, Swanson MS (2001) RpoS co-operates with other factors to induce *Legionella pneumophila* virulence in the stationary phase. Mol Microbiol 40: 1201-1214.

28. Vayssier-Taussat M, Le Rhun D, Deng HK, Biville F, Cescau S, et al. (2010) The Trw T4SS of *Bartonella* mediates host-specific adhesion to erythrocytes. PLoS Pathog 6: e1000946.

29. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, et al. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29: 22-28.

30. Sander A, Kretzer S, Bredt W, Oberle K, Bereswill S (2000) Hemin-dependent growth and hemin binding of *Bartonella henselae*. FEMS Microbiol Lett 189: 55-59.

31. Graca-Souza AV, Maya-Monteiro C, Paiva-Silva GO, Braz GR, Paes MC, et al. (2006) Adaptations against heme toxicity in blood-feeding arthropods. Insect Biochem Mol Biol 36: 322-335.

32. Myers WF, Cutler LD, Wisseman CL, Jr. (1969) Role of erythrocytes and serum in the nutrition of *Rickettsia quintana*. J Bacteriol 97: 663-666.

33. Myers WF, Osterman JV, Wisseman CL, Jr. (1972) Nutritional studies of *Rickettsia guintana*: nature of the hematin requirement. J Bacteriol 109: 89-95.

34. Minnick MF, Sappington KN, Smitherman LS, Andersson SG, Karlberg O, et al. (2003) Five-member gene family of *Bartonella quintana*. Infect Immun 71: 814-821.

35. Battisti JM, Sappington KN, Smitherman LS, Parrow NL, Minnick MF (2006) Environmental signals generate a differential and coordinated expression of the heme receptor gene family of *Bartonella quintana*. Infect Immun 74: 3251-3261.

36. Roden JA, Wells DH, Chomel BB, Kasten RW, Koehler JE (2012) Hemin binding protein C is found in outer membrane vesicles and protects *Bartonella henselae* against toxic concentrations of hemin. Infect Immun 80: 929-942.

37. Schulein R, Dehio C (2002) The VirB/VirD4 type IV secretion system of

Bartonella is essential for establishing intraerythrocytic infection. Mol Microbiol 46: 1053-1067.

38. Eicher, S. C. and Dehio, C (2012) *Bartonella* entry mechanisms into mammalian host cells. Cell Microbiol 14: 1166-1173.
39. Quebatte M, Dehio M, Tropel D, Basler A, Toller I, et al. (2010) The BatR/BatS two-component regulatory system controls the adaptive response of *Bartonella henselae* during human endothelial cell infection. J Bacteriol 192: 3352-3367.
40. Gillaspie D, Perkins I, Larsen K, McCord A, Pangonis S, et al. (2009) Plasmid-based system for high-level gene expression and antisense gene knockdown in *Bartonella henselae*. Appl Environ Microbiol 75: 5434-5436.
41. O'Rourke KP, Shaw JD, Pesesky MW, Cook BT, Roberts SM, Bond JP, Spatafora GA (2010) Genome-wide characterization of the SloR metalloregulome in *Streptococcus mutans*. J Bacteriol 192: 1433-1443.
42. Zhao H, Volkov A, Veldore VH, Hoch JA, Varughese KI (2010) Crystal structure of the transcriptional repressor PagR of *Bacillus anthracis*. Microbiology 156(Pt 2): 385-91.
43. Wu J, Rosen BP. Metalloregulated expression of the ars operon. J Biol Chem. 1993 Jan 5;268(1):52-8.
44. Knoten CA, Hudson LL, Coleman JP, Farrow JM 3rd, Pesci EC (2011) KynR, a Lrp/AsnC-type transcriptional regulator, directly controls the kynurenine pathway in *Pseudomonas aeruginosa*. J Bacteriol 193: 6567-6575.

45. Miyoshi S, Shinoda S (2000) Microbial metalloproteases and pathogenesis. Microbes Infect 2: 91-98.

46. Benz I, Schmidt MA (1992) AIDA-I, the adhesin involved in diffuse adherence of the diarrhoeagenic *Escherichia coli* strain 2787 (O126:H27), is synthesized via a precursor molecule. Mol Microbiol 6: 1539-1546.

47. Emsley P, Charles IG, Fairweather NF, Isaacs NW (1996) Structure of *Bordetella pertussis* virulence factor P.69 pertactin. Nature 381: 90-92.

48. Lindenthal C, Elsinghorst EA (1999) Identification of a glycoprotein produced by enterotoxigenic *Escherichia coli*. Infect Immun 67: 4084-4091.

49. Sherlock O, Dobrindt U, Jensen JB, Munk Vejborg R, Klemm P (2006) Glycosylation of the self-recognizing *Escherichia coli* Ag43 autotransporter protein. J Bacteriol 188: 1798-1807.

50. Ito S, Vinson JW (1965) Fine Structure of *Rickettsia Quintana* Cultivated in Vitro and in the Louse. J Bacteriol 89: 481-495.

51. Culpepper GH (1946) Factors influencing the rearing and maintenance of a laboratory colony of the body louse. J Econ Entomol 39: 472-474.

Figure Legends

Figure 1. *B. quintana* were enumerated to select time points for microarray analysis of growth stage-regulated genes. (A) The diagram depicts the experimental design utilized in cultivation of *B. quintana* for *in vitro* transcriptome profiling at early *vs*. late stage growth. *B. quintana* were plated on chocolate agar and grown at 37ºC for 2 days, at which point half of the cultures were shifted

to 28ºC. *B. quintana* were harvested for RNA extraction and CFU enumeration on the days highlighted in green. **(B)** For each experiment, *B. quintana* growth was analyzed by enumerating CFU per plate after 3 to 11 total days of growth. CFU enumeration was done to determine the growth stage of the *B. quintana* cultures. Based on the data shown in 1B, the days highlighted in green in 1A and 1B were selected for *B. quintana* transcriptional profiling. CFU data from a single representative experiment are shown, and error bars represent the standard deviation of the mean CFU per plate from three replicates.

Figure 2. Growth stage-responsive genes comprise two large clusters and include a large proportion of the genome. The heat map depicts unsupervised clustering of data from expression arrays from two independent time courses of *B. quintana* grown at either 28°C or 37°C for 7-9 days, as outlined in Figure 1. The arrays are depicted in columns, and the rows represent the probes on the array. The dendrogram at the left describes the similarity of the gene clusters. Regardless of the temperature at which the *B. quintana* were grown, there is a clear division into two distinct transcriptional programs (genes turned on then off, and off then on over the duration of the time course). The inset legend shows the range of $\log_2$-fold changes related to the range of colors in the heatmap. Genes of interest are noted along the right–hand side of the heatmap, in their cluster position.

Figure 3. *B. quintana* were enumerated to select time points for microarray analysis of temperature-regulated genes. (A) The diagram summarizes the experimental design utilized in cultivation of *B. quintana* for *in vitro* transcriptome profiling at 37°C vs. 28°C. *B. quintana* were plated on chocolate agar and grown at 37ºC for 2 days, at which point half of the cultures were shifted to 28ºC. *B. quintana* were harvested for CFU enumeration on the days highlighted in green. **(B)** For each experiment, bacterial growth was analyzed by enumerating CFU per plate from 3 to 7 total days post plating. CFU enumeration was done to ensure that *B. quintana* cultures selected for global transcription profiling were in log phase growth at the respective temperatures. The days subsequently selected for *in vitro* transcriptional analysis of *B. quintana* are highlighted in green. CFU data from a single representative experiment are shown, and error bars indicate the standard deviation of the mean CFU per plate from three replicates.

Figure 4. RT-qPCR quantification of *B. quintana* transcription corroborates microarray data for temperature-regulated genes. Transcription of select genes up-regulated at 28°C by microarray analysis was analyzed by RT-qPCR at 37°C (black) and 28°C (gray) to validate the microarray results. Transcript level was normalized to the *B. quintana* reference gene, *purA*. Error bars indicate standard errors of the mean. *, $P \leq 0.05$; **, $P \leq 0.01$ by Student's *t* test, comparing the relative level of transcription at 37°C and 28°C for each gene.

Figure 5. *B. quintana* genes up-regulated at 28°C are overrepresented in several COG functional categories. The graph shows the COG classification of each gene that was significantly up- or down-regulated in *B. quintana* grown at 28°C, from microarray analysis (Table 1). Genes with increased transcription at 28°C are represented by black bars; genes with decreased transcription at 28°C are represented by gray bars. Of the categories with attributable function, there is an overrepresentation of genes in the transcription, signal transduction, intracellular trafficking/secretion/vesicular transport, and defense mechanisms in *B. quintana* grown at the arthropod vector temperature of 28°C.

Figure 6. MEME searching identifies an overrepresented, purine-rich motif upstream of *B. quintana* genes up-regulated at 28°C. **(A)** Sequence logo of the top scoring MEME result for the top 11 regulated genes, by SAM score; and **(B)** position and scoring of motif sites (p-value threshold <1e-3) in upstream sequences. The motif is present in upstream sequences for 8 of the top 11 genes, often with multiple instances, as shown by the blue block diagram depicting motif position within upstream sequences.

Figure 7. The number of *B. quintana* per body louse increases over time during *in vivo* infection. The number of *B. quintana* per louse was determined by qPCR analysis of DNA isolated from infected body lice. At 1 day post infection (dpi), there were approximately $1.42 \times 10^4 \pm 2.83 \times 10^3$ *B. quintana* per louse; at 5 dpi, $3.82 \times 10^4 \pm 1.02 \times 10^4$ *B. quintana* per louse; and at 9 dpi, $6.93 \times 10^4 \pm 4.00 \times 10^4$ *B. quintana* per louse. These findings corroborate the

quantification of *B. quintana* in experimentally infected body lice reported by Seki *et. al,* 2007. The average of data from three separate experiments is shown; error bars represent the standard errors of the mean.

Figure 8. Transcription of *hbpC* and BQ10280 *in vivo* corroborates transcription results *in vitro* at 28°C. *In vivo* transcription of *hbpC* and BQ10280, genes determined to be highly expressed *in vitro* at 28°C by microarray, was analyzed in *B. quintana*-infected body lice (white bars) at 1, 5, and 9 days post inoculation (dpi). The *in vitro* transcription of *hbpC* and BQ10280 in *B. quintana* grown *in vitro* on chocolate agar at 28°C (gray bars) or 37°C (black bars) also was evaluated. Transcript level was normalized to *B. quintana* 16S rRNA. The relative level of *hbpC* and BQ10280 transcript in infected body lice was similar to that observed during *in vitro* growth of *B. quintana* at 28°C. The average of data from three separate experiments is shown; error bars represent the standard errors of the mean.

Tables

Table 1. *B. quintana* genes differentially expressed at 37°C compared with 28°C

Gene ID	Description	name	28C Log2 ratio	37C Log2 ratio	Mean local fdr %	Mean Score(d)	significant oligos	Fold Change
BQ02410	Hemin binding protein c	hpbC	1.62	-3.52	0.02	-5.59	10	-35.25
BQ10280	hypothetical protein		1.51	-2.97	0	-6.41	9	-22.33
BQ00570	hypothetical protein		1.58	-2.5	0	-5.93	2	-16.96
BQ11530	hypothetical protein		1.7	-1.95	0.08	-4.04	4	-12.53
BQ06411	hypothetical protein		1.1	-1.62	0.05	-3.91	6	-6.59
BQ10530	virB secretion system component	virB2	1.05	-1.47	0	-3.7	2	-5.71
BQ10540	virB secretion system component	virB3	0.87	-1.56	0	-3.68	2	-5.38
BQ11720	hypothetical protein		0.63	-1.72	0	-2.76	2	-5.1
BQ11730	hypothetical protein		0.62	-1.59	0	-2.46	7	-4.62
BQ09200	hypothetical protein		0.87	-1.33	0	-3.07	1	-4.59
BQ10980	Sensory transduction regulatory protein	phyR	0.38	-1.63	0.78	-2.18	8	-4
BQ01390	Variable outer membrane protein	vompD	0.4	-1.34	0.36	-1.91	4	-3.34
BQ10550	virB secretion system component	virB4	0.75	-0.98	0	-2.59	8	-3.33
BQ08670	hypothetical protein		0.25	-1.19	0.72	-1.99	6	-2.7
BQ02420	Hemin binding protein a	hbpA	1.41	-0.02	5.35	-1.59	7	-2.68
BQ00240	Thioredoxin	trxA	0.68	-0.71	1.28	-1.79	2	-2.6
BQ00830	hypothetical protein		0.82	-0.53	0	-2.09	2	-2.56
BQ07681	hypothetical protein		0.7	-0.58	0	-2.11	3	-2.43
BQ11010	hypothetical protein		0.92	-0.32	0	-1.92	1	-2.36
BQ09250	Cold shock protein		0.79	-0.43	3.24	-1.71	2	-2.34
BQ08990	regulatory protein		0.6	-0.59	0	-2.59	1	-2.28
BQ06171	hypothetical protein		0.84	-0.32	1.64	-1.78	2	-2.24
BQ10570	virB secretion system component	virB6	0.63	-0.5	1.35	-2.17	6	-2.18
BQ10180	hypothetical protein		0.57	-0.54	1.08	-1.83	3	-2.16
BQ00291	hypothetical protein		0.58	-0.51	0	-3.59	1	-2.14
BQ02770	hypothetical genomic island protein		0.96	-0.11	0.39	-1.85	2	-2.09
BQ06490	transcriptional regulator		0.65	-0.39	0	-2.13	3	-2.05
BQ13370	Transmembrane protein		0.68	-0.36	5.11	-1.61	3	-2.05
BQ07300	hypothetical genomic island protein		0.6	-0.43	4.14	-1.64	1	-2.04
BQ10150	hypothetical protein		0.44	-0.58	0.32	-1.95	2	-2.02
BQ06450	hypothetical protein		0.57	-0.43	4.14	-1.65	5	-2
BQ09120	integrase recombinase		0.64	-0.33	3.58	-1.69	2	-1.96
BQ11930	hypothetical protein		0.62	-0.35	6.91	-1.54	1	-1.95
BQ06460	hypothetical protein		0.65	-0.3	2.94	-1.72	6	-1.94
BQ07302	hypothetical protein		0.49	-0.39	0.88	-1.82	1	-1.84
BQ01331	hypothetical protein		0.6	-0.26	0.07	-1.88	1	-1.82
BQ11710	Biopolymer transport exbB protein	exbB	0.44	-0.41	3.58	-1.68	6	-1.81
BQ11450	hypothetical protein		0.35	-0.5	2.09	-1.74	1	-1.8
BQ12140	Ferredoxin II	fdxA	0.27	-0.57	5.96	-1.57	1	-1.79
BQ09410	hypothetical protein		0.43	-0.37	2.6	-1.72	3	-1.74
BQ05180	hypothetical protein		0.84	0.05	4.62	-1.62	1	-1.73
BQ10790	Phosphoglucomutase	pgm	0.51	-0.19	2.65	-1.72	5	-1.63
BQ08710	DNA uvrDDNA helicase II	uvrD	0.43	-0.23	2.59	-1.75	5	-1.57
BQ01080	Heme exporter protein A	ccmA	0.41	-0.23	1.76	-1.77	4	-1.55
BQ02611	hypothetical protein		0.48	-0.13	6.66	-1.67	1	-1.52
BQ06800	Glutathione reductase	gor	0.31	-0.25	0	-1.95	1	-1.47
BQ01070	Heme exporter protein B	ccmB	0.4	-0.13	5.65	-1.74	1	-1.44
BQ03390	Transcriptional regulator ompR	ompR	0.36	-0.17	6.23	-1.56	2	-1.44
BQ05580	putative integrase dna protein		0.37	-0.1	6.88	-1.54	1	-1.39
BQ06200	Glutamate racemase	murI	0.42	0	6.66	-1.55	1	-1.34
BQ10290	Probable surface protein		0.2	-0.14	0	-3.08	1	-1.27
BQ07580	Exopolyphosphatase		0.38	0.03	6.73	-1.54	1	-1.27
BQ05030	hypothetical protein		0.25	-0.03	4.93	-1.61	5	-1.21
BQ05940	Nitrogenase cofactor synthesis protein	nifS1	0.56	0.29	6.4	-1.55	1	-1.21
BQ08870	Cell division protein	ftsW	0.03	0.15	3.48	-1.67	1	-0.92
BQ03510	Chorismate synthase	aroC	-0.14	0.06	0	-1.77	1	-0.87
BQ08760	DNA ligase	ligA	-0.33	0.26	0	2.2	1	1.5
BQ03710	Phosphatase		-0.56	0.09	0.31	1.83	1	1.57
BQ02310	ABC transporter permease protein		-0.51	0.15	1.59	1.76	1	1.58
BQ12210	transport protein transmembrane		-0.37	0.31	0.34	1.83	1	1.61
BQ00450	hypothetical protein		-0.53	0.3	0.35	1.86	5	1.78
BQ12890	SUN protein FMU protein	sun2	-0.35	0.51	0.24	1.89	6	1.81
BQ00840	Nicotinate phosphoribosyltransferase	pncB	-0.77	0.11	0	1.92	4	1.84
BQ11820	tolB protein	tolB	-0.9	0.08	0.31	1.87	3	1.97
BQ01770	ABC transporter, periplasmic binding protein		-0.71	0.31	0.3	2.09	5	2.02
BQ01780	ABC transporter, permease protein		-0.43	0.62	0.01	2.03	10	2.07
BQ12900	Heat shock protein		-0.54	0.54	0.79	1.81	2	2.11
BQ00600	Heat shock protein DnaJ	dnaJ1	-1.07	0.39	0.16	1.9	8	2.75

Table 2. Identification of homologs for unannotated, temperature-responsive *B. quintana* genes

	BlastP			pHMMER		
locus ID	Description	E-value	accession #	Description	E-value	GI/accession #
BQ00450	zinc metalloprotease	6.00E-77	NP_540932.1	zinc metalloprotease	8.50E-71	306837668
BQ00570	LysM domain/BON superfamily protein	2.00E-52	EHH06667.1	LysM domain/BON superfamily protein	3.00E-47	325291490
BQ00830	-	-	-	inner membrane protein ybaN	1.70E-23	358048936
BQ02770	XRE family transcriptional regulator	4.00E-08	YP_002004972.1	conserved hypothetical protein *Bartonella* sp. AR 15-3	3.90E-09	319405804
BQ05030	glycosyl transferase family protein	6.00E-11	YP_001877749.1	glycosyltransferase sugar-binding protein containing DXD motif	6.20E-11	299133438
BQ06171	PP-loop domain containing protein	2.00E-16	ZP_05779946.1	tRNA 2-thiocytidine biosynthesis protein TtcA	3.10E-12	81648390
BQ06411	-	-	-	similar to ankyrin 2,3/unc44, partial	2.10E-49	115950018
BQ06450	Staphylococcal nuclease homolog	2.00E-63	CBI82181.1	nuclease domain-containing protein	4.10E-29	261758074
BQ06460	uracil-DNA glycosylase	1.00E-82	EHJ97859.1	Uracil-DNA glycosylase	9.30E-79	306840362
BQ08670	-	-	-	PF11015.3 n/a Protein of unknown function (DUF2853) 2 100	9.40E-35	DUF2853
BQ09200	Transglycosylase-associated protein	3.00E-26	YP_002290717.1	transglycosylase-associated protein	3.20E-30	306843863
BQ09410	cation diffusion facilitator family transporter	1.00E-126	ZP_04680778.1	cation diffusion facilitator family transporter	8.40E-117	306844567
BQ10150	trm112p-like family protein	3.00E-17	ZP_08269540.1	Trm112p-like protein	6.20E-11	PF03966.11
BQ10180	SH3 type 3 domain-containing protein	3.00E-09	EHK80050.1	bacterial SH3 domain protein	1.00E-12	342212994
BQ10280	Inducible Bartonella autotransporter	4.00E-131	CBI80621.1	CAMP-like factor autotransporter	4.70E-267	56684460
BQ10290	inducible Bartonella autotransporter	5.00E-122	CBI80621.1	CAMP-like factor autotransporter	<1.00E-300	56684460
BQ11930	Sel1 repeat-containing protein	8.00E-53	ZP_04681125.1	Sel1 domain protein repeat-containing protein	6.70E-63	163800487
BQ11720	-	-	-	PF05957.8 Bacterial protein of unknown function (DUF883)	1.40E-18	DUF883

FIGURES

FIGURE 1

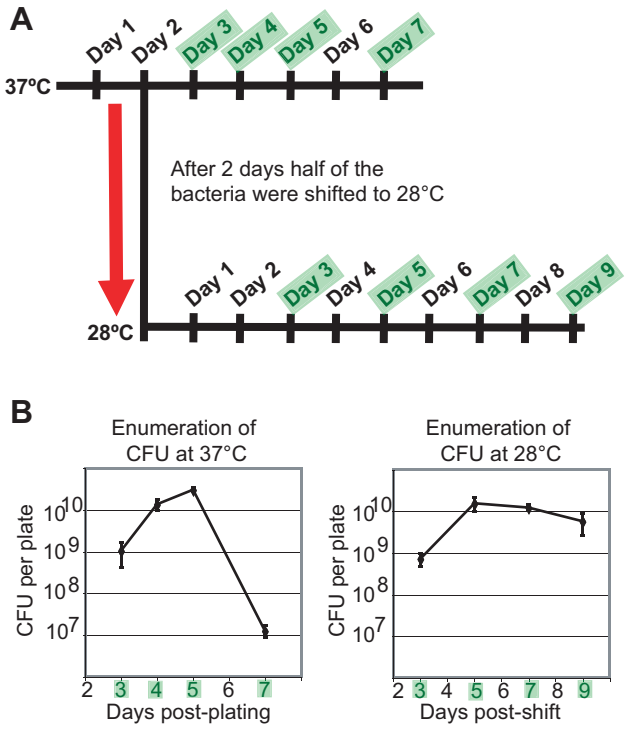


FIGURE 2

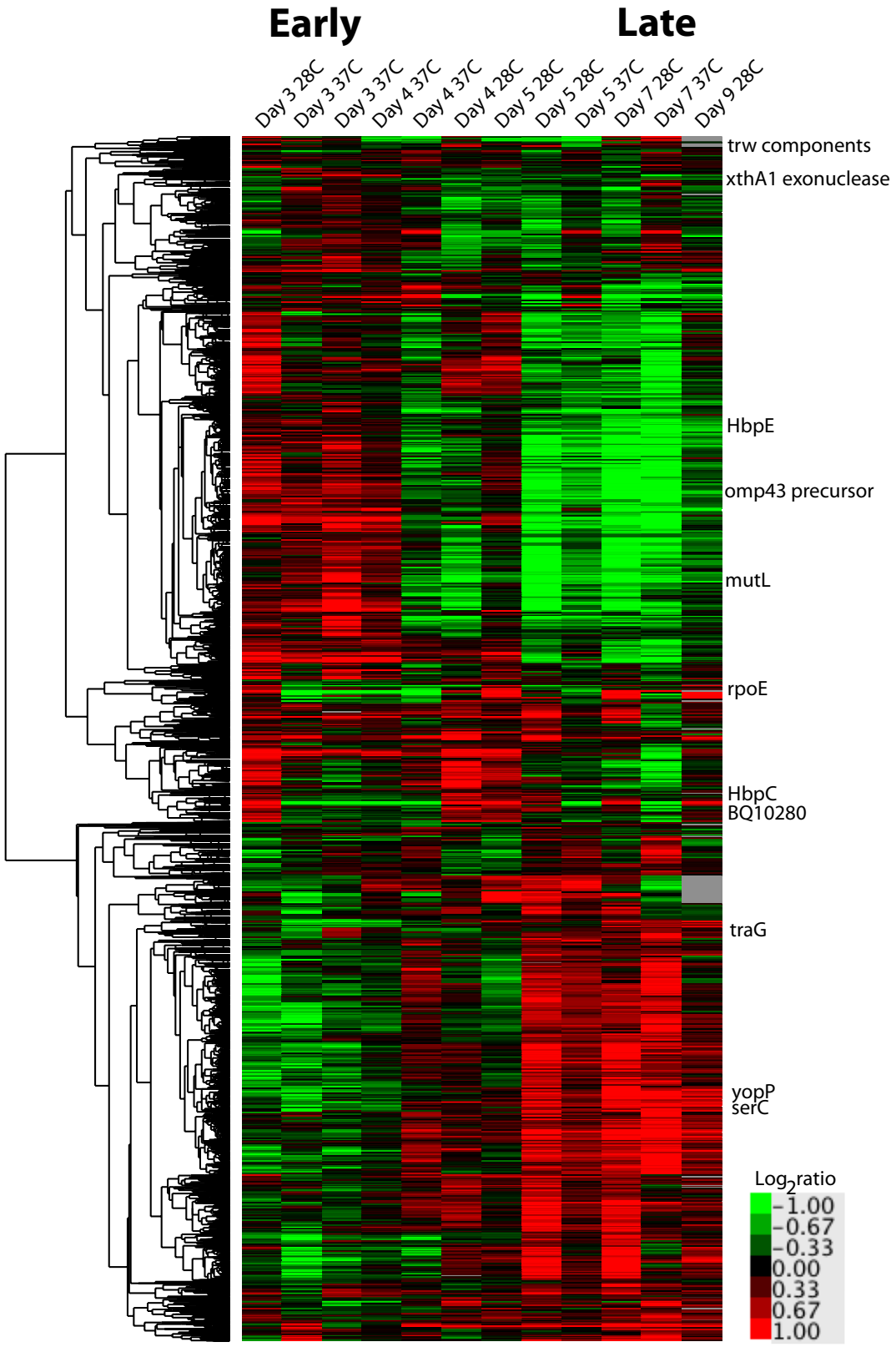

Early
Late
Day 3 28C
Day 3 37C
Day 3 37C
Day 4 37C
Day 4 37C
Day 4 28C
Day 5 28C
Day 5 28C
Day 5 37C
Day 7 28C
Day 7 37C
Day 9 28C
trw components
xthA1 exonuclease
HbpE
omp43 precursor
mutL
rpoE
HbpC
BQ10280
traG
yopP
serC
Log2 ratio
-1.00
-0.67
-0.33
0.00
0.33
0.67
1.00

FIGURE 3

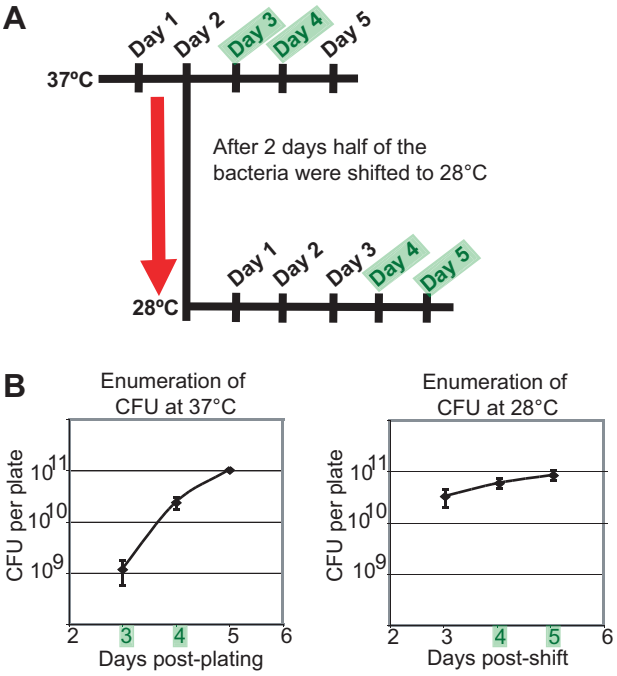


FIGURE 4

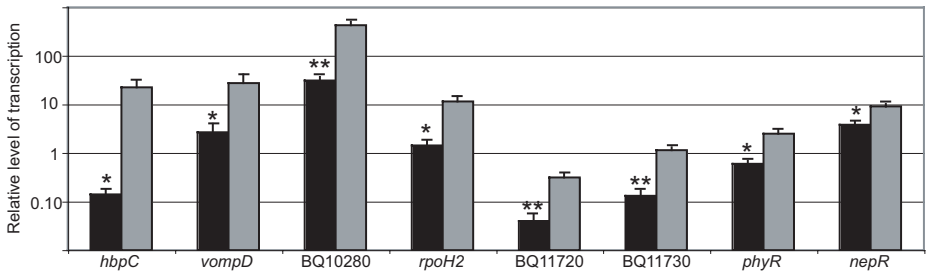
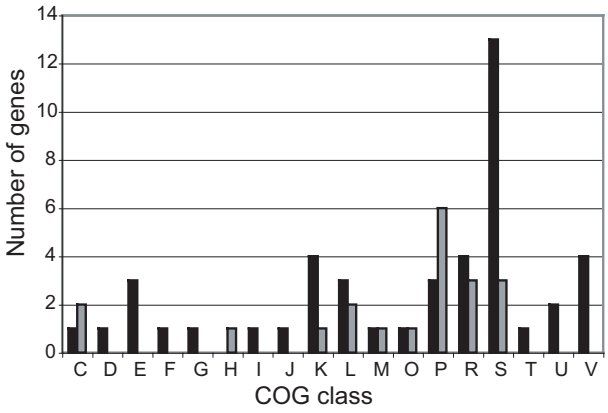


FIGURE 5



COG functional category

C: Energy production and conversion
D: Cell cycle control and mitosis
E: Amino Acid metabolism and transport
F: Nucleotide metabolism and transport
G: Carbohydrate metabolism and transport
H: Coenzyme metabolism and transport
I: Lipid metabolism and transport
J: Translation, ribosomal structure and biogenesis
K: Transcription
L: Replication, recombination, and repair
M: Cell wall/membrane/envelop biogenesis
O: Post-translational modification, protein turnover, chaperone functions
P: Inorganic ion transport and metabolism
R: General functional prediction only
S: Function unknown
T: Signal transduction mechanisms
U: Intracellular trafficking, secretion, and vesicular transport
V: Defense mechanisms

FIGURE 6

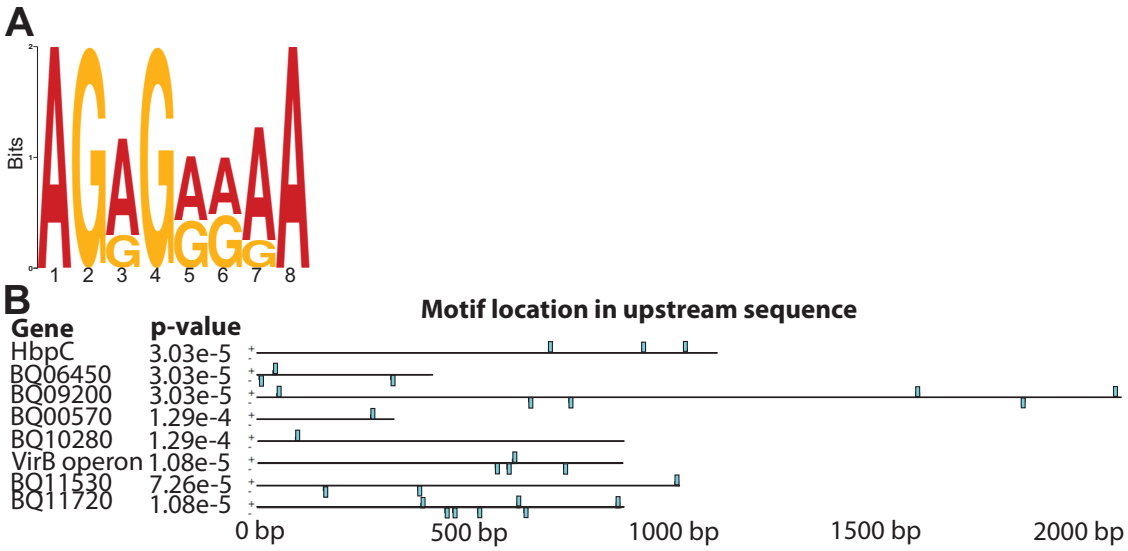


FIGURE 7

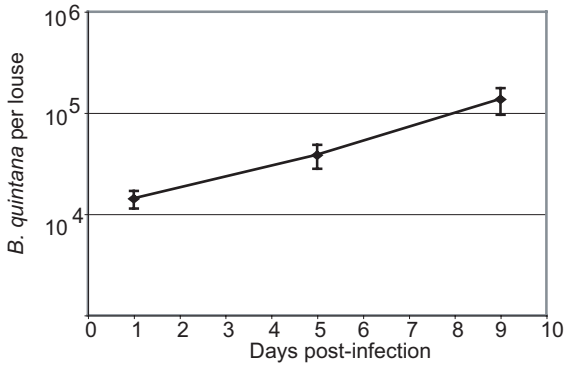


FIGURE 8

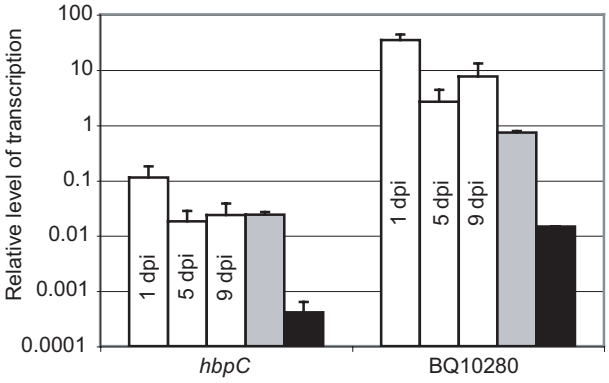


Table S1. Oligonucleotide primers used in this study

Primer	Sequence	Purpose
SA090	ggattgtacgtggcgtcttt	RT-qPCR *purA*, forward primer
SA091	aatggaccttctccaacacg	RT-qPCR *purA*, reverse primer
SA082	atgaatatgaaatggttaataacgg	RT-qPCR *hbpC*, forward primer
SA083	ccgttagcgagaatattcatctt	RT-qPCR *hbpC*, reverse primer
SA346	gcaggcaaggcgaacgt	RT-qPCR *vompD*, forward primer
SA347	tcatgtttgggccaccagta	RT-qPCR *vompD*, reverse primer
SA128	TTTTTGTTGATTGCCGTTGA	RT-qPCR *phyR*, forward primer
SA129	GGGTTCGTGCTATCCCTACA	RT-qPCR *phyR*, reverse primer
SA078	tattttacgaaattggtcaatcgtt	RT-qPCR *rpoH2*, forward primer
SA079	tatcagagctagaaaaacggaaatc	RT-qPCR *rpoH2*, reverse primer
SA224	agatgatcttctcggggtca	RT-qPCR *nepR*, forward primer
SA225	tcaaacctttctgcattgttt	RT-qPCR *nepR*, reverse primer
SA096	cagctcgtgtcgtgagatgt	RT-qPCR 16S rRNA, forward primer
SA097	cagagtgcaatccgaactga	RT-qPCR 16S rRNA, reverse primer
SA354	cgcaatgaaatttccggtat	RT-qPCR BQ11720, forward primer
SA355	gaataaacccaacccctgct	RT-qPCR BQ11720, reverse primer
SA356	acgtttatcgctccccttt	RT-qPCR BQ11730, forward primer
SA357	agaaatcggccaacacaaac	RT-qPCR BQ11730, reverse primer
SA249	aatggagacacgtcctacgg	RT-qPCR BQ10280, forward primer
SA250	ctcctgaaagttcgctttgc	RT-qPCR BQ10280, reverse primer
Seki_1	GCCGCCTTCGTTTCTCTTTC	*B. quintana* quantification, forward primer
Seki_2	AGTGTCTTCCTTAAAGTCCCAAAG	*B. quintana* quantification, reverse primer

Table S2. *B. quintana* genes that are transcriptionally responsive to the transition from logarithmic growth phase to stationary/death phase (The fold increase represents the fold difference in *B. quintana* transcription in stationary phase compared with logarithmic growth phase)

Gene ID	Gene Name	Name	Fold Change*
BQ11580	hypothetical intracellular effector	yopP	2.41
BQ12950	carbonic anhydrase protein		2.40
BQ11110	hypothetical protein		2.37
BQ09900	hypothetical protein		2.26
BQ11120	hypothetical protein		2.18
BQ02790	Phosphoserine aminotransferase	serC	2.17
BQ11350	hypothetical protein		2.13
BQ10950	hypothetical protein		2.11
BQ10620	virB11 protein homolog	virB11	2.04
BQ02400	hypothetical protein		2.04
BQ12600	trwI1 protein_10163	trwI1	-2.97
BQ04040	Aminopeptidase N_3556	pepN	-2.97
BQ06720	hypothetical protein		-2.98
BQ12590	trwJ1 protein	trwJ1	-2.99
BQ13530	Chromosome partitioning protein	parB	-2.99
BQ07520	hypothetical protein		-2.99
BQ10020	Multidrug resistance protein	vceA	-3.00
BQ08730	sco1 family protein	sco1	-3.00
BQ08430	Hemin binding protein E	hbpE	-3.00
BQ08750	Aminopeptidase P protein		-3.01
BQ10720	hypothetical protein		-3.01
BQ04920	Pyruvate dehydrogenase E1 component beta subunit	pdhB	-3.02
BQ09890	omp43 precursor	omp43	-3.02
BQ10750	Chaperonin protein groEL	mopA	-3.02
BQ11750	Cell division protein	ftsH	-3.02
BQ08460	hypothetical protein		-3.03
BQ09280	Thioredoxin reductase	trxB	-3.03
BQ13330	pH adaptation potassium efflux system E	phaE	-3.03
BQ04910	Pyruvat dehydrogenase E1 component alpha subunit	pdhA	-3.04
BQ12560	trwL8 protein	trwL8	-3.05
BQ03220	sohB Protease	sohB	-3.06
BQ08310	Amino acid permease_6924	gltJ	-3.06
BQ09810	hypothetical protein		-3.06
BQ13520	Hydroxyacylglutathione hydrolase	gloB	-3.06
BQ05970	hypothetical protein		-3.07
BQ12490	trwL1 protein	trwL1	-3.07
BQ12510	BQ12510trwL3trwL3_protein_10114	trwL3	-3.08
BQ05950	ABC transporter subunit		-3.09
BQ00540	hypothetical protein		-3.10
BQ10800	hypothetical protein		-3.13
BQ09490	hypothetical protein		-3.14
BQ09170	DNA primase	dnaG	-3.15
BQ12530	trwL5 protein	trwL5	-3.15
BQ08780	Competence lipoprotein comL precursor	comL	-3.18
BQ12840	SurF1 family protein	surF1	-3.18
BQ04430	Amidophosphoribosyltransferase precursor	purF	-3.20
BQ12040	hypothetical protein		-3.21
BQ12360	hypothetical protein		-3.22
BQ06850	Citrate synthase	gltA	-3.25
BQ00870	Cytidylate kinase	cmk	-3.26
BQ05450	hypothetical protein		-3.26

BQ08360	hypothetical protein		-3.27
BQ00310	Folylpolyglutamate synthase	folC	-3.28
BQ08320	Amino acid permease	gltK	-3.28
BQ01250	30s ribosomal protein	rpsU	-3.29
BQ03580	hypothetical protein		-3.29
BQ05310	Acyl carrier protein	acpP2	-3.32
BQ05330	hypothetical protein		-3.33
BQ07410	hypothetical protein		-3.35
BQ04600	hypothetical protein		-3.39
BQ03400	hypothetical protein		-3.40
BQ09260	Aspartate aminotransferase A	aatA	-3.41
BQ04050	DNA polymerase, bacteriophage type		-3.43
BQ01870	Preprotein translocase secA subunit	secA	-3.43
BQ02190	hypothetical protein		-3.44
BQ09040	hypothetical protein		-3.45
BQ07710	hypothetical protein		-3.46
BQ02570	DNA mismatch repair protein	mutL	-3.46
BQ09180	hypothetical protein		-3.47
BQ09140	Outer membrane protein		-3.47
BQ09380	hypothetical protein		-3.49
BQ12620	trwJ2 protein	trwJ2	-3.50
BQ02530	hypothetical protein		-3.50
BQ07600	hypothetical protein		-3.50
BQ06280	Cold shock protein		-3.52
BQ07390	Phosphoribosylglycinamide formyltransferase	purN	-3.55
BQ03780	hypothetical protein		-3.57
BQ04440	Colicin v production protein	cvpA	-3.58
BQ10070	hypothetical protein		-3.61
BQ12480	korA protein	korA	-3.61
BQ01660	Outer membrane lipoprotein precursor		-3.62
BQ12570	trwM protein	trwM	-3.73
BQ12790	DnaJ related protein		-3.78
BQ03080	hypothetical protein		-3.80
BQ03190	hypothetical protein		-3.81
BQ08500	hypothetical protein		-3.83
BQ00690	Polypeptide deformylase_695	def	-3.92
BQ04750	hypothetical protein	tatB	-3.93
BQ01860	Iron response regulator	fur2	-3.93
BQ04930	Dihydrolipoamide acetyltransferase E2	pdhC	-3.94
BQ00010	hypothetical protein		-4.15
BQ08290	Exodeoxyribonuclease III	xthA1	-4.19
BQ03790	hypothetical protein		-5.04

Chapter 4……

Dihydroartemisinin induces transcriptome arrest in artemisinin-naive and artemisinin-selected *Plasmodium falciparum*

This chapter is a summary of work done by:

Nelson C, Sorber K, Tucker M, LaCrue A, Azizan A, Kyle D, DeRisi JL.

Author contributions:

Matt Tucker derived all artemisinin resistant strains in vitro and performed all recrudescence assays. Azliyati Azizan performed the pilot drug treatment timecourse and extracted RNA from the samples. Katherine Sorber amplified the samples, and Katherine Sorber and Chris Nelson hybridized the pilot samples to microarrays. Chris Nelson, Katherine Sorber, Matt Tucker, and Alexis LaCrue performed the large drug treatment timecourse, extracted RNA from the samples, and amplified the RNA. Katherine Sorber and Chris Nelson hybridized the amplified RNA to microarrays. Chris Nelson performed all analysis of the pilot and large timecourse microarray data, and re-processed samples where necessary. Dennis Kyle and Joseph L. DeRisi conceived of and supervised the project.

Joseph L. DeRisi, Thesis Advisor

ABSTRACT

Artemisinin combination therapy (ACT) is the frontline treatment for multi-drug resistant falciparum malaria. Although artemisinin and its derivatives rapidly clear parasitemia and reduce malaria symptoms, frequent recrudescence is observed when artemisinin class compounds are administered alone. In this study, we have conducted temporal parasite morphology and transcriptome analysis experiments of sensitive and resistant *Plasmodium falciparum* parasite strains to assess the effects of dihydroartemisinin (DHA) on parasite development. Results show that following artemisinin treatment, both sensitive and resistant ring-stage parasites pause in a dormant state characterized by small rounded morphology and a transcriptional state of 8-11 hr post-invasion rings. Comparison of sensitive and resistant strains revealed constitutive and DHA-inducible differences in gene expression, as well as a distinct set of genes were strongly induced or repressed during parasite arrest. These data show that treatment with DHA results in ring-stage dormancy and transcriptional arrest, which may have implications for the study of recrudescence, and emerging artemisinin resistance in the field.

INTRODUCTION

Malaria is a disease with a large global impact, being responsible for more than 225 million new infections and over 700,000 deaths per year, most of which occur in sub-Saharan Africa (1,2). There are five species of malaria that affect humans, of which *Plasmodium falciparum* is the most pathogenic. During the

past four decades, *P. falciparum* has developed resistance to every commonly available antimalarial. As a result, many countries in Africa are now faced with high treatment failure rates when using chloroquine or sulfadoxine/pyrimethamine, the former first and second line drugs, respectively. In response to this monumental problem, the World Health Organization has supported the widespread use of artemisinin combination therapy (ACT) to treat multi-drug resistant malaria in endemic areas.

Artemisinin drugs are rapidly acting antimalarials, which contain an endoperoxide bridge that is essential for antimalarial activity (3). These potent compounds produce faster parasite and fever clearance times than any other antimalarials (4) and decrease gametocyte carriage (5,6), thereby effectively reducing transmission of malaria. Although they are effective, the short half life (~45 min) of artemisinins may contribute to the frequent recrudescence observed in patients after treatment.

Given the importance of artemisinin drugs for malaria treatment and control, considerable research has been devoted to understanding the mechanism of action of these compounds on *Plasmodium* spp. Several proposed hypotheses include non-specific alkylation or oxidation of parasite proteins due to free radical bi-products of endoperoxide breakdown (7), specific inhibition of a SERCA-like Ca^{2+}-dependent ATPase (8), inhibition of the parasite's mitochondrial electron transport chain (9), or oxidation of $FADH_2$ and dihydroflavin (10). These results though generally accepted, remain

controversial. However, evidence for emerging resistance to artemisinins in the field has added some urgency to the need for a better understanding of parasite drug interactions.

Recent studies in Southeast Asia using longitudinal and geographic comparisons demonstrate longer parasite clearance times in response to both ACTs and artesunate administered as a monotherapy (11-13). Unfortunately, these field observations have been difficult to translate into tangible phenotypes such as an increase in 50% inhibitory concentrations (IC50s), a traditional measure of *in vitro* drug susceptibility. A few recent studies observed a trend of slightly elevated IC50s in isolates from west to east Asia (14,15); however, studies attempting to directly correlate increased parasite clearance time with an increase in IC50 *ex vivo* or in culture-adapted strains remain inconclusive (12,13). These results suggest that standard short-term growth assays do not sufficiently measure artemisinin efficacy.

Recently, a novel hypothesis has emerged that may explain frequent recrudescence following artemisinin treatment (16). Studies have shown that ring-stage parasites become dormant *in vitro* (17-19) and *in vivo* (A.N. LaCrue, M. Scheel, K. Kennedy, N. Kumar, and D.E. Kyle, submitted for publication) following exposure to artemisinin and its derivatives. In this study, we expanded on these observations and characterized both a morphological and transcriptome arrest of artemisinin-sensitive and resistant lines following treatment with DHA. Specifically, we conducted transcriptome analysis of synchronous and

asynchronous *P. falciparum* after a 6 hr exposure to physiologically relevant levels of DHA. Transcriptional analysis identified genes that are differentially expressed during DHA-induced transcriptome arrest, and between the artemisinin sensitive and resistant clones. These data provide the first evidence for long-lasting transcriptome arrest in response to an antimalarial drug.

RESULTS

Stepwise artemisinin pressure yields resistant strains of the D6 line.

Recent studies reported the derivation of multiple *P. falciparum* lines resistant to artelinic acid (AL) and artemisinin (QHS) (19-21). In brief, strains were discontinuously exposed to increasing levels of AL followed by QHS to produce resistant progeny. This method produced D6 parasites that are resistant to 80 ng/mL of QHS (D6.QHS80). Further work generated a highly resistant derivative of D6 that could tolerate 2400 ng/mL QHS (D6.QHS2400x5) (19). Clones of parental D6 and D6.QHS2400x5 were obtained and used for the morphology and transcriptome studies described below.

DHA treatment of malaria parasites induces a distinct morphological state preceding the resumption of parasite growth.

A preliminary experiment was conducted to examine the kinetics of recrudescence and morphological changes induced after D6 and D6.QHS2400x5 were treated with a dose of DHA. Synchronous ring-stage parasites were exposed to DHA (200 ng/mL) for 6 hr and parasitemia were monitored for over

eight days by blood smear (Figure 1A). Matched negative control cultures treated with DMSO grew normally. Following treatment with DHA there was a predominant shift in morphology from normal rings to “morphologically dormant” forms that have a small circle of blue-staining cytoplasm along with red/purple stained chromatin (Figure 1B).

For both strains, by 48 hr post-treatment (PT) the majority of parasites were classified as morphologically dormant and by 72 hr PT, morphologically normal parasites were observed in the resistant D6.QHS2400x5 strain, whereas none were noted in the sensitive D6 strain. At 96 hr PT, normal parasites were observed for D6, but the percentage was less than D6.QHS2400x5. Overall, dormant forms persisted from 6-144 hr for each strain, but after 24 hr a greater proportion of dormant parasites were present in the resistant strain (Figure 1). Based on these results, further studies were conducted to determine whether the transcriptome of DHA-treated parasites exhibited a transcriptional perturbation concomitant with observed morphological changes, and if significant differences could be determined between parent and resistant parasites.

DHA induces transcriptome arrest resembling a ring-like state in sensitive W2 parasites.

The published descriptions of smooth progression through unique transcriptional states during the normal intraerythrocytic developmental cycle (IDC) can be used as a diagnostic for determining the hour of parasite development after invasion of an erythrocyte (22-26). In a pilot microarray

experiment, synchronized W2 ring stages were exposed to drug for 6 hr, and RNA was isolated at 6 (T6) and 27 (T27) hr PT. For untreated parasites, a 21 hr separation in the IDC would constitute a large change in the transcriptome and be expected to have a negative correlation of -0.69 (calculated from the HB3 dataset). However, these studies revealed that following DHA treatment, samples taken 21 hr apart had surprisingly similar transcriptomes with a Pearson correlation of 0.63, and correlated best with 12 and 13 hr post-invasion (hpi) in transcriptome data from the normal IDC of HB3 (Figure S1). These data suggested a drug-induced arrest in ring-stage development; therefore, a study consisting of a longer time course (Figure 2) was conducted using the D6 and D6.QHS2400x5 parasite clones to generate a specific developmental profile after treatment with DHA.

DHA induces transcriptome arrest resembling a ring-like state in QHS resistant and sensitive parasites.

Highly synchronous ring-stage cultures (approximately 8 hpi) and asynchronous mixed stage cultures of each strain were treated with DHA (200 ng/mL), while matched controls split from the same starting culture were simultaneously treated with DMSO. Samples were collected before drug was added (T0), and then 6 hr post-drug addition (T6), followed by additional time points collected at lengthening intervals. Mixed stage cultures were discontinued after 48 hr, while synchronous drug treated cultures were sampled until T104, a time where recrudescence had occurred in both strains. RNA was isolated from

time course samples and linearly amplified before hybridization to 89 *P. falciparum* expression microarrays (23).

We first analyzed the array results with respect to the normal temporal progression of the transcriptome, comparing results from each array to published array results from every hr of the IDC of normally growing HB3 parasites (24). In Figure 3, the Pearson correlation between each of our time point profiles and the published normally growing profiles is depicted. Each time point's Pearson correlation function (depicted as columns in Figure 3) has a clear peak of positive correlation that places it in developmental time, as described previously (25,26). Consecutive time points of untreated parasites' transcriptomes have a gradually shifting peak correlation (0.73±0.05 and 0.64±0.07 for D6 and for D6.QHS2400x5, respectively) that is consistent with normal progression through the IDC (Figure 3A and C).

In contrast, peak correlations (0.63±0.02 and 0.55±0.02) of the drug-treated transcriptomes were arrested at 8-11 hpi and 9-11 hpi for the parental and the resistant clone, respectively. This period of transcriptome arrest lasted for 86 hr post-drug addition and 62 hr post-drug addition for D6 and D6.QHS2400x5 respectively (Fig 3B and 3D). The exit from dormancy in each strain was verified by morphological analysis (below). Unlike in the case of an absolute arrest of transcription induced by alpha-amanitin or actinomycin D, the correlation between the DHA-treated and normal IDC ring transcriptome did not appreciably diminish across this prolonged period (25). In summary, D6 and

D6.QHS2400x5 parasites treated with DHA exhibited arrested progress of the normal transcriptome's correlation pattern at a ring-like state for 3 days.

Recovery following exposure to DHA: exit from transcriptome arrest and observance of recrudescence.

As samples were taken for RNA isolation during the time course, blood smears were also monitored to detect recrudescence after dormancy. Microscopic analysis included time points beyond those used for transcriptional studies (treated synchronized cultures were terminated after 188 hr, treated asynchronous cultures were terminated after 152 hr). The untreated cultures reached higher overall parasitemias compared to the treated groups and treated asynchronous cultures recrudesced before synchronized cultures (Figure S2A and B). In the first 24 hr PT, all strains expressed either dormant ring morphology described above or dead morphology(Figure S2D). Dormant forms were observed from 12-141 hr PT in both synchronized treated parent and resistant cultures. This period of dormancy preceded eventual recrudescence of morphologically normal parasites.

Asynchronous cultures exposed to DHA converge to a ring-like transcriptome.

An alternative to the ring stage dormancy hypothesis is the observed transcriptional arrest in rings might reflect a pause relative to their stage of development at the time of drug treatment or that signal captured on the microarrays could represent residual RNA from dead or dying rings in the

synchronized culture. Therefore, we simultaneously exposed asynchronous cultures of the D6 and D6.QHS2400x5 clones to DHA and harvested samples for microarray analysis. As expected, DMSO treated controls produced low peak correlations throughout the time course (0.18±0.07 at 21 hpi and 0.16±0.02 at 9 hpi D6 and D6.QHS2400x5, respectively) when compared to the synchronous HB3 IDC dataset, reflecting mixed parasite stages within these cultures (Figure 5A and C). If the apparent developmental point of peak correlation of the DHA-induced transcriptome were dependent on parasite stage of the initial population, then we would expect that asynchrony of DHA-treated cultures also would persist. In contrast, we found that DHA-treated mixed stage cultures converged on an arrested and semi-synchronous transcriptome that most closely resembles that of normal rings, with a peak correlation of r=0.40±0.10 at 14-16 hpi, and r=0.37±0.07 at 9-13 hpi for D6 and D6.QHS2400x5, respectively (Figure 5B and D). As with the synchronous cultures, both mixed stage artemisinin sensitive D6 and artemisinin-resistant D6.QHS2400x5 exhibited a transcriptome arrest phenotype most closely resembling ring stage parasites.

DHA treatment induces differential regulation of specific genes in both artemisinin sensitive and resistant strains.

Although the overall transcriptome of DHA-treated parasites most closely matches ring stage parasites, the data provided an opportunity to identify those genes that were reproducibly up-regulated or down-regulated during transcriptome arrest. Using two class significance analysis of microarrays (SAM, (27)), we compared the class of all time zero synchronized results to all arrested

samples (T6 hr to T48 hr post-DHA addition). The total cycle time of the two strains was reasonably similar with a slight IDC offset of ~2 hr between these two samples, which we reason is within sampling noise and not enough to flood our results with developmentally regulated genes. Antigenic gene families (Var, rifin, stevor, RESA MSP, FIKK, var-like rif pseudogenes and reticulocyte binding protein genes) were not considered for this analysis to avoid strain dependent effects. With parameters designed to yield less than one expected false positive, 21 genes that met our criteria were significantly up-regulated in the arrested condition relative to the untreated condition at time zero, and 119 genes that met these criteria were down-regulated. A representative list of these genes is displayed in Table 1 and the full list is available in the supplementary material Table 1. A subset of interesting genes within this list include: PFI1170c, PF14_0017, PF10_0327 and PF13_0088, and PFD0740w.

In the first 48 hr after drug treatment PFI1170c, which encodes thioredoxin reductase, was up-regulated 9 fold in the QHS-resistant parasite, and 23-fold in the sensitive parasite. DHA is known to deplete the native glutathione antioxidant proteins within parasitized RBCs (28). Thioredoxin reductases are antioxidant proteins complementary to the glutathioine system that are involved in detoxifying different type of peroxides (29,30). PF14_0017, upregulated 8-fold in both strains at 48 hr after DHA treatment, encodes a lysophosopholipase that is part of a nine-member family of genes (which putatively catabolize phosphoatidyl choline) encoded near the telomeres. Conversely, both PF10_0327 and PF13_0088 (which encode Myb domain transcription factors) were down-

regulated 3-fold. PF13_0088 has a reported peak of protein expression in the nucleus at the trophozoite stage (31). PFD0740w, normally expressed in the mid-ring state, encodes a putative cyclin dependent kinase.

Constitutive transcription differences between QHS-resistant and sensitive lines.

The genetic determinant of dormancy or resistance might induce constitutive up-regulation or down-regulation of some transcripts in resistant vs. susceptible lines. Therefore, we queried the data by comparing time-matched data produced from untreated parasites. For each time point from 0-48 hr, $\log_2$ transformed parental ratios were subtracted from QHS-resistant ratios. The results of these calculations were grouped and subjected to one class of SAM with parameters chosen to produce 0.73 expected false positives. From this analysis, 75 genes were found to be up-regulated in D6.QHS2400x5 when compared to D6, and 11 genes were found to be down-regulated (partial list in Table 2, full list in supplemental Table 2). Strain-specific differences may be responsible for expression of some of the observed genes, but interestingly, several genes clustered together on chromosome 10 and appeared to be up-regulated in D6.QHS2400x5 when compared to D6. Oligos from this approximately 20-gene locus are overrepresented in our list with a p-value of 2E-11 by the binomial distribution. An overlapping locus exhibited clustering of copy number variations between field isolates from Kilifi Kenya (32). Additionally, up-regulation of a longevity-assurance (LAG1) domain gene (PFE0405c), which has

been associated with the lifespan of yeast (33), was observed in the QHS resistant clone.

DHA induces differential expression of genes in QHS-resistant and sensitive parasites.

Genes differentially expressed between QHS-resistant and sensitive parasites during transcriptome arrest could potentially be involved in DHA resistance. To explore this possibility, the microarray data were analyzed for genes that showed strain-specific larger regulation change from baseline after drug treatment. The difference between the time zero-transformed D6.QHS2400x5 transcriptome inductions and time zero-transformed parental D6 transcriptome inductions was analyzed for significant outliers via SAM. Choosing a delta threshold for less than one expected false positive, we arrived at a list of 12 genes that were more up-regulated in arrested QHS-resistant parasites and 30 genes that were more down-regulated in arrested QHS-resistant parasites (delta=1.442, FDR 1.17%) (Table 3). Interesting genes in this list include PF11_0245 (encodes a translation elongation factor subunit) and PFL1330c (encodes a Pfcyc-2 cyclin-related protein), both which exhibited relatively low induction.

DISCUSSION

We have described in molecular detail a long-lasting arrest of transcriptome development that is induced when ring-stage *P. falciparum* parasites are treated with artemisinin. To the best of our knowledge, other than RNA polymerase II inhibitors such as alpha-amanitin (26,34-37), treatment of

malaria parasites with other antimalarial drugs has not resulted in long-lasting transcriptome arrest. Studies by Hu and colleagues noted instances of drug-induced developmental slowing of the transcriptome over a shorter timescale (26). Natalang et al. (38) noted that expression analysis was complicated due to "drug-induced massive slowing-down of parasite development." We believe that this observation directly corroborates our own analysis. The timing of transcriptional pause in our studies coincides with those observed previously (Kyle et al., unpublished data; (17, 18,19)) and changes in the morphology and transcriptome of arrested parasites were qualitatively similar for QHS sensitive and resistant clones. Importantly, these studies were conducted with physiologically relevant concentrations of DHA that were high enough to induce dormancy in parent and resistant parasites, but low enough to allow recovery of both sensitive and resistant parasites in less than 10 days. Although, studies by Witkowski et al. (18) reported the presence of arrested forms only in their resistant strain, based on our results we believe that dormancy is not a feature of only resistant parasites in response to drug exposure but a common response to DHA shared by all *P. falciparum* parasites.

Although it is possible that the phenomenon of transcriptome arrest could be due to the persistence of ring mRNA from dead parasites, we believe this interpretation to be implausible based on previous studies of transcriptome dynamics. Shock et al. demonstrated that *P. falciparum* mRNA decay rates are relatively rapid, with the average half-life of ring-stage transcripts being 9.5 minutes (25). Correlations scores as high as those shown in our data, quickly

decay to noise in the absence of active transcription due to intrinsic RNA degradation machinery (25). Furthermore, if the transcriptome arrest observed only reflected the mRNA of dead parasites, then it would be difficult to explain the observed DHA-induced convergence upon a ring-like state in originally asynchronous cultures. Additionally, previous studies with cidal drugs have found little to no disruption of the normal progression through the developmental cycle as assayed by expression microarrays, with studies intentionally looking for such effects (26,34). It therefore seems unlikely that residual transcripts from dead parasites could maintain such high correlation scores for days. Thus far, the data suggests that DHA induces something akin to a cell cycle arrest at the ring stage, characterized by dormant ring forms and a ring-like transcriptome.

The possibility that artemisinins induce cell cycle arrest has precedence in studies with human carcinoma cell lines. Willoughby et al. (39) concluded that artemisinin disrupts transcription at the promoter of CDK4 in prostate cancer. In pancreatic cancer, T cells, and hepatoma cells artemisinins have been shown to disrupt cyclin levels and induce G1 arrest (40-42). Additionally Efferth and colleagues suggested that in yeast, artemisinins may induce the DNA repair checkpoint based on yeast mutants with increased sensitivity (43). Though *Plasmodium* spp. are known to encode cell cycle regulatory genes, there has been doubt over the existence of any inducible checkpoint or arrest (34).

These studies have given us a few insights into the release from dormancy. In these studies, as parasites exited from arrest, there was an apparent loss of synchrony, resulting in reduced peak correlations; a finding that

is consistent with that found by Teuscher et al (17). Furthermore, QHS-resistant parasites released from transcriptome arrest >24 hr earlier than sensitive parasites. This observation stimulates speculation about the nature of the putative arrested state as it relates to artemisinin resistance. It is possible the mechanism of resistance involves an increased ability to recover from an arrested state following exposure to drug. Alternatively, a greater proportion of parasites might survive drug pressure long enough to enter an arrested state, suggesting that growth arrest is a natural phenomenon which enables parasites to cope with environmental stressors. These stressors might include oxidative or innate immune inducible oxidative stress mimicked by artemisinin class drugs.

An interesting subset of genes are induced or suppressed during DHA-induced dormancy characterized by an arrested transcriptome. PfTrxR and thioredoxin (PfTrx) are components of an essential antioxidant system for *P. falciparum* while in RBCs. This system is involved in a variety of cellular functions including DNA synthesis, regulation of transcription by interacting with transcription factors, and reducing hydrogen peroxide (44). The strong induction of thioredoxin reductase during DHA-induced dormancy is consistent with the protein's role in parasite defense against artemisinin induced oxidative stress. It is interesting that PfTrxR was up-regulated much more in D6 compared to D6.QHS2400x5. Perhaps over multiple rounds of artemisinin drug pressure, the resistant parasite developed other tolerance mechanisms so that its reliance on PfTrxR decreased over time. Because functional antioxidant and redox systems are necessary for parasite survival (normally and maybe in response to drugs),

inhibitors that target this enzyme would be effective antimalarials. Andricopulo et al. (45) identified three nitrophenyl derivatives that inhibit PfTrxR, and were active against *P. falciparum* strain K1 (CQ resistant). An inhibitor selective for the plasmodial thioredoxin reductase might complement artemisinin therapy, which may explain the synergistic effect of methylene blue and artemisinin drugs *in vitro* (46).

Down-regulated genes could indicate ring-stage metabolic pathways that are normally important for progression through the life cycle but have been shut down after drug treatment. Notable genes that were down-regulated in our transcriptional analysis included PF10_0327, PF13_0088, and a cyclin dependent kinase. PF10_0327 and PF13_0088 are genes that encode proteins containing PfMyb domains, which are linked to DNA binding and regulation of transcription (47). (In subsequent MITOMI investigations of these proteins there was never any strong DNA binding over background. Manuel Llinás and his lab came to the same conclusion using protein binding microarrays. In light of these results, I seriously doubt that these proteins are DNA binding proteins.) It may be that artemisinin treatment caused inhibition of these proteins, leading to a halt in progression of the life cycle. It is also possible that significant down-regulation of Mybs and a cyclin dependent kinase could play a role in the mechanism of arrest through an as-yet-unknown cell cycle checkpoint.

We also identified up-regulation of a LAG1-containing protein (encoded by PFE0405c) in D6 and D6.QHS2400x5 (greater magnitude in resistant parasite;

data not shown). LAG1 domain proteins are part of the LASS (longevity assurance) family of ceramide synthases (33). Deletion of LAG1 in yeast resulted in increased replication (49), whereas over-expression of LAG1 in yeast had a bimodal effect on longevity, with moderate expression resulting in increased longevity and with higher expression curtailing life span (50). Ceramide synthases are important for sphingolipid metabolism (proteins that have roles in different eukaryotic cell functions). Ceramide accumulates in response to cytotoxic agents or stress responses in eukaryotic cells (51). An increase in the intracellular ceramide content and activation of parasite sphingomyelinase(s) in *P. falciparum* were associated with the parasite death induced by artemisinin and mefloquine (52). Perhaps the observed increased ceramide content and expression associated with artemisinin treatment is a factor of the parasite creating a protective stress response. Ceramide metabolism may be important for parasite survival as it was found that certain ceramide analogs can inhibit *P. falciparum in vitro* (51).

In addition to the description of the arrested transcriptome state, we describe novel expression differences in an *in vitro* selected artemisinin resistant clone in comparison with a clone of the parental D6. As expected, many of the parasite line–specific genes are members of large gene families; however, among the list are other genes that may prove interesting given independent validation by other means. We find the co-up-regulated chromosome 10 locus near 1.2 Mb to be of interest as a potential amplification associated with artemisinin selection. Notably missing among our baseline transcriptional

differences are other candidate genes of interest postulated as being associated with resistance (e.g., SERCA ATPase), suggesting that if changes in these genes are indeed present in our selected line, they do not exert their effect at the transcript abundance level.

If our *in vitro* observations reflect those found in the field, the results could be used to create kinetic data models of the acquisition and spread of resistance. Additionally, the differentially expressed genes in arrested parasites and in QHS–selected parasites may serve as novel molecular markers for monitoring drug efficacy and resistance in the field. Given the recent emergence of clinically relevant resistance to artemisinins, confirmation of the phenomenon of transcriptome arrest *in vivo* may provide new avenues for understanding artemisinin resistance and recrudescence, which is an emerging problem for malaria control.

MATERIALS AND METHODS

Cell culture.

All strains of *P. falciparum* were maintained using previously described methods (53). Parasites were cultured at 2-4% hematocrit in RPMI 1640 medium (Invitrogen, Carlsbad, CA.) supplemented with 10% heat-inactivated A+ human plasma (complete media). All cultures were maintained at 37°C in a mixed gas incubator containing 5% O_2, 5%CO_2, and 90% N_2. Cultures were synchronized using 5% w/v D-sorbitol following the method of Lambros and Vanderberg (1979) (54).

***In vitro* drug selection and parasite cloning.**

P. falciparum laboratory clones were previously adapted to various levels of artemisinin QHS and AL (19,20). The selection of D6 initially began by applying small increments of AL to produce parasites that were resistant to 80 ng/mL of AL (D6.AL80) . D6.AL80 was then exposed to stepwise increments of QHS to produce strains of D6 that tolerated up to 2400 ng/mL of QHS (D6.QHS2400) (19). D6.QHS2400 was exposed to three subsequent treatments at 2400 ng/mL QHS and parasites were cloned by limiting dilution (55). Clones were selected and treated with one more dose of 2400 ng/mL QHS (D6.QHS2400x5). The parental D6 was also cloned via serial dilution and a single clone of D6 and D6.QHS2400x5 were selected for experiments described in this work.

Preliminary recovery assay with D6 parasites.

D6 and D6.QHS2400x5 were synchronized to ring stage and cultures were split to ~2% parasitemia. DHA (200 ng/mL) was added to each culture and an equivalent volume of DMSO was added to control cultures. At 6 hours post-drug treatment (PT), cultures were washed three times with RPMI and returned to flasks. Thick and thin smears were made before drug treatment (T0), after drug was washed out (T6), and at every 24 hours after drug was added. Parasites were monitored until parasitemia of normal parasites in drug-treated cultures exceeded 4% (192 hours post-drug). For each time point, parasitemia was determined as the total number of parasites counted/total number of erythrocytes counted (at least 800). Parasites from thin smears were sorted into

classifications of dead, dormant, ring, trophozoite, or schizont. The ratio of normal/total parasites and dormant/total parasites was also calculated at each time point.

W2 transcriptome pilot study.

Synchronized W2 ring-stage parasites at ~2% parasitemia were exposed to 100 nM dihydroartemisinin (DHA) for six hours, after which drug was washed out with RPMI and parasites were returned to culture. Time points were taken before drug addition (T0), six hr after treatment (T6), and 27 hr after treatment (T27). RNA extraction and amplification, as well as microarray analysis of these samples was carried out as described below.

Cell culture for transcriptional studies.

Plasmodium falciparum clones of D6 and D6.QHS2400x5 were divided into four groups (Figure 2). When parasites were mostly in the ring stage, one half of each culture was sorbitol synchronized and the other half of each culture was maintained as mixed stages for the duration of the experiment. Synchronous cultures were synchronized twice in the subsequent cell cycle 8 hr apart, then 10 hr apart in the next cycle, and finally 12 hr apart in the cycle after that, for a total of 7 synchronizations over a 192-hr period. Invasion in the synchronous cultures was monitored by smear every hr starting approximately 26 hr after the last synchronization. Maximum invasion (number of rings = number of schizonts) was recorded and the cultures were synchronized one last time 2 hr after maximum invasion. Both mixed stage and synchronous cultures were split to approximately 2% parasitemia.

Drug treatment.

Eight hours post-invasion, 6 mL of each culture was harvested as the T0 sample before DHA was added. Pre-warmed phosphate buffered saline (PBS) was added to the cells and the entire volume was spun down at 500xg for 5 min. The cell pellet was washed with an additional 25 mL warm PBS, then spun down again. After removal of the supernatant, pellets were flash frozen in liquid nitrogen and transferred to –80˚C. DHA (200 ng/mL final concentration) was added to the remaining experimental culture of each strain, while an equal volume of DMSO was added to untreated control culture. After 6 hr of drug or DMSO treatment (T6), another 6 mL aliquot of culture was harvested from each condition. The remaining culture was centrifuged and DHA/DMSO was washed off with 37˚C RPMI. Parasites were returned to culture at 4% hematocrit in complete media. Subsequent time points were taken at 12, 18, 24, 32, 40, and 48 hours after drug/DMSO exposure, except for mixed stage DMSO control cultures, which were sampled only at 0, 24, and 48 hr after addition. DHA-treated cultures were carried out past 48 hr in order to observe recrudescence and time points were taken at 56 hours, and then every 12 hr after that (synchronous to T188, asynchronous to T152) until cultures reached at least 3% normal parasitemia. All samples were snap frozen in liquid nitrogen for later RNA extraction and subsequent transcriptional analysis.

RNA extraction and amplification.

Total RNA was harvested from the frozen pellets using 10 mL Trizol (Invitrogen) and 2 mL chloroform for every 1-2 mL of cell pellet. Aqueous phases

were re-extracted with 1 volume of acid phenol (pH 4.3) and 1 volume of chloroform. Aqueous phases were extracted again with 1x volume of chloroform, then precipitated with 0.1x volume of 3M sodium acetate (pH 5.2) and 1x volume of isopropanol. Where possible, 110 ng of total RNA was amplified and amino allyl labeled using the Amino Allyl MessageAmp II aRNA Amplification kit (one round, 14 hr IVT, Ambion, Austin, TX). For samples with < 110 ng total RNA yield, as much total RNA as possible was used for amplification. An amplified RNA pool representing transcripts expressed throughout the IDC was compiled as a reference.

Cy dye labeling and microarray hybridization.

Aliquots of amplified pool RNA (2 µg) were coupled to Cy3, and 2 µg of each amplified sample RNA was coupled to Cy5 (GE Healthcare, Piscataway, NJ). Cy3 pool and Cy5 sample were competitively hybridized at 65 ˚C on a printed microarray containing 8,159 70-mer oligos that map to 5,338 ORFs annotated in PlasmoDB release 6.3. Prior to hybridization, microarrays were printed and post-processed as described in Bozdech et al. (23). After a minimum of 18 hr hybridizing, microarrays were washed in 65˚C 0.6x SSC, 0.03% SDS, and then in room temperature 0.06x SSC. Spun dry arrays were then scanned on an Axon 4000B scanner using Axon Genepix v 6.0 and 6.1 software (Molecular Devices, Union City, CA).

Array analysis.

Microarrays were manually gridded and Cy3 and Cy5 intensity of each spot was extracted using Genepix software. Arrays were uploaded to Nomad

v2.0 (http://ucsf-nomad.sourceforge.net/) where the data was normalized in bins of pixel intensity R^2, and then filtered to remove spots with "bad" or "missing" manual flags added during gridding and spots with sum of median intensities less than 500. The resulting ratio Cy5/Cy3 intensity tables were $\log_2$ transformed and re-centered about 0. Several arrays have technical replicates. For subsequent calculations, the results of replicate arrays without large artifacts were averaged together for the following samples: PTS0, PTS6, PTS12, PTS24, PTS32, PTS48, RTM40, RTM48, RTS0, RUS6. Sample abbreviations are P or R (parental D6 or resistant D6.QHS2400x5), T or U (DHA treated or untreated), S or M (synchronous or mixed stage), and the number of hr post-drug exposure.

Var, rifin, surfin, RESA MSP, FIKK, var-like rif pseudogenes and reticulocyte binding protein genes were removed from the data in order to eliminate confounding strain-dependent antigenic variation effects. Re-centered arrays were compared with all time points of the HB3 intra-erythrocytic developmental cycle (23) by Pearson's correlation. The lists of genes up-regulated and down-regulated in the 48 hr after drug treatment in treated sensitive and resistant strains were generated by two class significance analysis of microarrays (SAM version 3.0, http://www-stat.stanford.edu/~tibs/SAM/ (27)) after filtering for oligos with 70% of data available. We compared all synchronized time zero arrays to all arrested samples (T6 to T48, with the exception of T18 hr which did not meet the oligo data available filter). The number of arrays from D6 was balanced by an equal number of arrays from D6.QHS2400x5 for each group in order to avoid strain-specific results. Seeking <1 expected false positive; we

chose a delta score threshold of 0.665 with an overall false discovery rate (FDR) of 0.335% and 0.56 expected false positives. Although no direct stipulation was made to ensure that resulting genes were DHA-responsive in the same direction in both D6 and D6.QHS2400x5, our delta-score threshold was conservative enough that genes in Table 1 were either up-regulated or down-regulated in both strains. ORFs with multiple oligos are represented by the best scoring oligo. In Tables 1-3, fold change is expressed as the geometric mean of the fold change or fold induction from the time-point(s) in question.

To look for constitutive differences between the resistant line and clone of the parental line, we compared time-matched data produced from untreated parasites after filtering for oligos with 90% of data available (Table 2). For each time point from 0-48 hr, we subtracted $\log_2$ transformed parental D6 ratios from D6.QHS2400x5 ratios. The results of these calculations were grouped and subjected to one class SAM. A delta score threshold of 0.765 was chosen to yield an overall FDR of 0.7% and 0.73 expected false positives.

We also were interested in DHA-induced expression differences between the two strains after filtering for oligos with 70% of data available (Table 3). Array $\log_2$ ratio data from time points 6, 12, 24, 32 and 48 hr post drug exposure were zero transformed by subtracting the time zero log2 ratios of the same strain. T18 hr and T40 hr did not meet the oligo data available filter and were excluded from the analysis. The difference between zero transformed data for each strain was analyzed for significant outliers via SAM. Seeking <1 expected false positive, we

chose at delta score threshold of 1.442, with an overall FDR of 1.17% and 0.52 expected false positives.

Microscopy and smear counts.

Thin smears for each time point were independently read in a blinded fashion by two (T48-T188) to three (T0-T48) people. Each person counted between 300-900 total red blood cells per slide. Parasitemia was calculated as the number of parasite-infected cells per total number of erythrocytes counted, and stage of observed parasites (ring, trophozoite, schizont, or dormant form) was recorded.

ACKNOWLEDGEMENTS

We thank Roxana Ordoñez, Charlie Kim, Flor Caro, and Emily Wilson for their help with conducting experiments and analyzing data. The work was supported by the National Institutes of Health grant R01 AI058973 from the National Institute for Allergy and Infectious Disease. CN was supported by an NIH training grant administered through the Biomedical sciences graduate program.

AUTHOR CONTRIBUTIONS

Conceived and designed the experiments: MT, AL, CN, KS, AA, JD, DK

Performed the experiments: MT, AL, CN, KS, AA,

Analyzed the data: MT, AL, CN, KS, JD, DK

Wrote the manuscript: MT, AL, CN, KS, AA, JD, DK

REFERENCES

1. Hay, S. I., C. A. Guerra, P. W. Gething, A. P. Patil, A. J. Tatem, A. M. Noor, C. W. Kabaria, B. H. Manh, I. R. Elyazar, S. Brooker, D. L. Smith, R. A. Moyeed, and R. W. Snow. 2009. A world malaria map: *Plasmodium falciparum* endemicity in 2007. PLoS. Med. 6:e1000048. 10.1371/journal.pmed.1000048.

2. World Health Organization. 2010. World Malaria Report 2010. http://www.who.int/malaria/world_malaria_report_2010/en/index.htmll [Accessed May 7, 2011].

3. Golenser, J., J. H. Waknine, M. Krugliak, N. H. Hunt, and G. E. Grau. 2006. Current perspectives on the mechanism of action of artemisinins. Int. J. Parasitol. 36:1427-1441. 10.1016/j.ijpara.2006.07.011.

4. Nosten, F. and N. J. White. 2007. Artemisinin-based combination treatment of falciparum malaria. Am. J. Trop. Med. Hyg. 77:181-192.

5. Chen, P. Q., G. Q. Li, X. B. Guo, K. R. He, Y. X. Fu, L. C. Fu, and Y. Z. Song. 1994. The infectivity of gametocytes of *Plasmodium falciparum* from patients treated with artemisinin. Chin Med. J. (Engl.) 107:709-711.

6. Price, R. N., F. Nosten, C. Luxemburger, F. O. ter Kuile, L. Paiphun, T. Chongsuphajaisiddhi, and N. J. White. 1996. Effects of artemisinin derivatives on malaria transmissibility. Lancet 347:1654-1658.

7. Olliaro, P. L., R. K. Haynes, B. Meunier, and Y. Yuthavong. 2001. Possible modes of action of the artemisinin-type compounds. Trends Parasitol. 17:122-126.

8. Eckstein-Ludwig, U., R. J. Webb, G. Van, I, J. M. East, A. G. Lee, M. Kimura, P. M. O'Neill, P. G. Bray, S. A. Ward, and S. Krishna. 2003. Artemisinins target the SERCA of *Plasmodium falciparum*. Nature 424:957-961. 10.1038/nature01813.

9. Li, W., W. Mo, D. Shen, L. Sun, J. Wang, S. Lu, J. M. Gitschier, and B. Zhou. 2005. Yeast model uncovers dual roles of mitochondria in action of artemisinin. PLoS. Genet. 1:e36. 10.1371/journal.pgen.0010036.

10. Haynes, R. K., W. C. Chan, H. N. Wong, K. Y. Li, W. K. Wu, K. M. Fan, H. H. Sung, I. D. Williams, D. Prosperi, S. Melato, P. Coghi, and D. Monti. 2010. Facile oxidation of leucomethylene blue and dihydroflavins by artemisinins: relationship with flavoenzyme function and antimalarial mechanism of action. ChemMedChem. 5:1282-1299. 10.1002/cmdc.201000225.

11. Noedl, H., Y. Se, K. Schaecher, B. L. Smith, D. Socheat, and M. M. Fukuda. 2008. Evidence of artemisinin-resistant malaria in western Cambodia. N. Engl. J. Med. 359:2619-2620.

12. Carrara, V. I., J. Zwang, E. A. Ashley, R. N. Price, K. Stepniewska, M. Barends, A. Brockman, T. Anderson, R. McGready, L. Phaiphun, S. Proux, V. M. van, R. Hutagalung, K. M. Lwin, A. P. Phyo, P. Preechapornkul, M. Imwong, S. Pukrittayakamee, P. Singhasivanon, N. J. White, and F. Nosten. 2009. Changes in the treatment responses to artesunate-mefloquine on the northwestern border of Thailand during 13 years of continuous deployment. PLoS. ONE. 4:e4551. doi:10.1371/journal.pone.0004551.

13. Dondorp, A. M., F. Nosten, P. Yi, D. Das, A. P. Phyo, J. Tarning, K. M. Lwin, F. Ariey, W. Hanpithakpong, S. J. Lee, P. Ringwald, K. Silamut, M. Imwong, K. Chotivanich, P. Lim, T. Herdman, S. S. An, S. Yeung, P. Singhasivanon, N. P. Day, N. Lindegardh, D. Socheat, and N. J. White. 2009. Artemisinin resistance in *Plasmodium falciparum* malaria. N. Engl. J. Med. 361:455-467. 10.1056/NEJMoa0808859.

14. Noedl, H., D. Socheat, and W. Satimai. 2009. Artemisinin-resistant malaria in Asia. N. Engl. J. Med. 361:540-541. 10.1056/NEJMc0900231.

15. Lim, P., C. Wongsrichanalai, P. Chim, N. Khim, S. Kim, S. Chy, R. Sem, S. Nhem, P. Yi, S. Duong, D. M. Bouth, B. Genton, H. P. Beck, J. G. Gobert, W. O. Rogers, J. Y. Coppee, T. Fandeur, O. Mercereau-Puijalon, P. Ringwald, B. J. Le, and F. Ariey. 2010. Decreased *in vitro* susceptibility of *Plasmodium falciparum* isolates to artesunate, mefloquine, chloroquine, and quinine in Cambodia from 2001 to 2007. Antimicrob. Agents Chemother. 54:2135-2142. 10.1128/AAC.01304-09.

16. Codd, A., F. Teuscher, D. E. Kyle, Q. Cheng, and M. L. Gatton. 2011. Artemisinin-induced parasite dormancy: a plausible mechanism for treatment failure. Malar. J. 10:56. doi:10.1186/1475-2875-10-56.

17. Teuscher, F., M. X. Gatton, N. Chen, J. Peters, D. X. Kyle, and Q. Cheng. 2010. Artemisinin Induced Dormancy in *Plasmodium falciparum*: Duration, Recovery Rates, and Implications in Treatment Failure. J. Infect. Dis. 202:1362-1368. 10.1086/656476.

18. Witkowski, B., J. Lelievre, M. J. Barragan, V. Laurent, X. Z. Su, A. Berry, and F. oit-Vical. 2010. Increased tolerance to artemisinin in *Plasmodium falciparum* is mediated by a quiescence mechanism. Antimicrob. Agents Chemother. 54:1872-1877. 1128/AAC.01636-09.

19. Tucker, M. 2010. Ph.D. Dissertation. Phenotypic and Genotypic Analysis of *In Vitro* Selected Artemisinin Resistant *Plasmodium falciparum*. University of South Florida, Tampa, FL.

20. Chavchich, M., L. Gerena, J. Peters, N. Chen, Q. Cheng, and D. E. Kyle. 2010. Role of *pfmdr1* amplification and expression in induction of resistance to artemisinin derivatives in *Plasmodium falciparum*. Antimicrob. Agents Chemother. 54:2455-2464. 10.1128/AAC.00947-09.

21. Chen, N., M. Chavchich, J. M. Peters, D. E. Kyle, M. L. Gatton, and Q. Cheng. 2010. Deamplification of *pfmdr1*-containing amplicon on chromosome 5 in *Plasmodium falciparum* is associated with reduced resistance to artelinic acid *in vitro*. Antimicrob. Agents Chemother. 54:3395-3401. doi:10.1128/AAC.01421-09.

22. Le Roch, K. G., Y. Zhou, P. L. Blair, M. Grainger, J. K. Moch, J. D. Haynes, l. de, V, A. A. Holder, S. Batalov, D. J. Carucci, and E. A. Winzeler. 2003. Discovery of gene function by expression profiling of the malaria parasite life cycle. Science 301:1503-1508.

23. Bozdech, Z., M. Llinás, B. L. Pulliam, E. D. Wong, J. Zhu, and J. L. DeRisi. 2003. The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. PLoS. Biol. 1:E5. 10.1371/journal.pbio.0000005.

24. Llinás, M., Z. Bozdech, E. D. Wong, A. T. Adai, and J. L. DeRisi. 2006. Comparative whole genome transcriptome analysis of three *Plasmodium falciparum* strains. Nucleic Acids Res. 34:1166-1173. 10.1093/nar/gkj517.

25. Shock, J. L., K. F. Fischer, and J. L. DeRisi. 2007. Whole-genome analysis of mRNA decay in *Plasmodium falciparum* reveals a global lengthening of mRNA half-life during the intra-erythrocytic development cycle. Genome Biol. 8:R134. 10.1186/gb-2007-8-7-r134.

26. Hu, G., A. Cabrera, M. Kono, S. Mok, B. K. Chaal, S. Haase, K. Engelberg, S. Cheemadan, T. Spielmann, P. R. Preiser, T. W. Gilberger, and Z. Bozdech. 2010. Transcriptional profiling of growth perturbations of the human malaria parasite *Plasmodiumfalciparum*. Nat. Biotechnol. 28:91-98.

27. Tusher, V. G., R. Tibshirani, and G. Chu. 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. U. S. A 98:5116-5121. 10.1073/pnas.091062498.

28. Ittarat, W., A. L. Pickard, P. Rattanasinganchan, P. Wilairatana, S. Looareesuwan, K. Emery, J. Low, R. Udomsangpetch, and S. R. Meshnick. 2003. Recrudescence in artesunate-treated patients with falciparum malaria is dependent on parasite burden not on parasite factors. Am. J. Trop. Med. Hyg. 68:147-152.

29. Björnstedt, M., M. Hamberg, S. Kumar, J. Xue, and A. Holmgren. 1995. Human thioredoxin reductase directly reduces lipid hydroperoxides by NADPH and selenocystine strongly stimulates the reaction via catalytically generated selenols. J. Biol. Chem. 270:11761-11764.

30. Nickel, C., S. Rahlfs, M. Deponte, S. Koncarevic, and K. Becker. 2006. Thioredoxin networks in the malarial parasite *Plasmodium falciparum*. Antioxid. Redox. Signal. 8:1227-1239. 10.1089/ars.2006.8.1227.

31. Boschet, C., M. Gissot, S. Briquet, Z. Hamid, C. Claudel-Renard, and C. Vaquero. 2004. Characterization of PfMyb1 transcription factor during erythrocytic development of 3D7 and F12 *Plasmodium falciparum* clones. Mol. Biochem. Parasitol. 138:159-163. 10.1016/j.molbiopara.2004.07.011.

32. Mackinnon, M. J., J. Li, S. Mok, M. M. Kortok, K. Marsh, P. R. Preiser, and Z. Bozdech. 2009. Comparative transcriptional and genomic analysis of *Plasmodium falciparum* field isolates. PLoS. Pathog. 5:e1000644. doi:10.1371/journal.ppat.1000644.

33. Teufel, A., T. Maass, P. R. Galle, and N. Malik. 2009. The longevity assurance homologue of yeast lag1 (Lass) gene family (review). Int. J. Mol. Med. 23:135-140.

34. Ganesan, K., N. Ponmee, L. Jiang, J. W. Fowble, J. White, S. Kamchonwongpaisan, Y. Yuthavong, P. Wilairat, and P. K. Rathod. 2008. A genetically hard-wired metabolic transcriptome in *Plasmodium falciparum* fails to mount protective responses to lethal antifolates. PLoS. Pathog. 4:e1000214. 10.1371/journal.ppat.1000214.

35. Gunasekera, A. M., A. Myrick, R. K. Le, E. Winzeler, and D. F. Wirth. 2007. *Plasmodium falciparum*: genome wide perturbations in transcript profiles among mixed stage cultures after chloroquine treatment. Exp. Parasitol. 117:87-92. 10.1016/j.exppara.2007.03.001.

36. Vaidya, A., J. Morrisey, H. Painter, and M. Mather. 2006. Adaptive changes in transgenic *P. falciparum* expressing type 1A dihydroorotate dehydrogenase continuously exposed to mitochondrial electron transport inhibitors. Molecular Parasitology Meeting, Woods Hole, MA.

37. Painter, H., J. Morrisey, J. Stumhofer, and A. Vaidya. 2004. Early response to antimalarial drug treatment examined by whole genome transcriptional profiling. Molecular Parasitology Meeting, Woods Hole, MA.

38. Natalang, O., E. Bischoff, G. Deplaine, C. Proux, M. A. Dillies, O. Sismeiro, G. Guigon, S. Bonnefoy, J. Patarapotikul, O. Mercereau-Puijalon, J. Y. Coppee, and P. H. David. 2008. Dynamic RNA profiling in *Plasmodium falciparum* synchronized blood stages exposed to lethal doses of artesunate. BMC. Genomics 9:388. doi:10.1186/1471-2164-9-388.

39. Willoughby, J. A., Sr., S. N. Sundar, M. Cheung, A. S. Tin, J. Modiano, and G. L. Firestone. 2009. Artemisinin blocks prostate cancer growth and cell cycle progression by disrupting Sp1 interactions with the cyclin-dependent kinase-4 (CDK4) promoter and inhibiting CDK4 gene expression. J. Biol. Chem. 284:2203-2213. 10.1074/jbc.M804491200.

40. Chen, H., B. Sun, S. Wang, S. Pan, Y. Gao, X. Bai, and D. Xue. 2010. Growth inhibitory effects of dihydroartemisinin on pancreatic cancer cells: involvement of cell cycle arrest and inactivation of nuclear factor-kappaB. J. Cancer Res. Clin. Oncol. 136:897-903. 10.1007/s00432-009-0731-0.

41. Wang, J. X., W. Tang, L. P. Shi, J. Wan, R. Zhou, J. Ni, Y. F. Fu, Y. F. Yang, Y. Li, and J. P. Zuo. 2007. Investigation of the immunosuppressive activity of artemether on T-cell activation and proliferation. Br. J. Pharmacol. 150:652-661. 10.1038/sj.bjp.0707137.

42. Hou, J., D. Wang, R. Zhang, and H. Wang. 2008. Experimental therapy of hepatoma with artemisinin and its derivatives: *in vitro* and *in vivo* activity, chemosensitization, and mechanisms of action. Clin. Cancer Res. 14:5519-5530. 10.1158/1078-0432.CCR-08-0197.

43. Efferth, T., H. Dunstan, A. Sauerbrey, H. Miyachi, and C. R. Chitambar. 2001. The anti-malarial artesunate is also active against cancer. Int. J. Oncol. 18:767-773.

44. Krnajski, Z., T. W. Gilberger, R. D. Walter, A. F. Cowman, and S. Muller. 2002. Thioredoxin reductase is essential for the survival of *Plasmodium falciparum* erythrocytic stages. J. Biol. Chem. 277:25970-25975.

45. Andricopulo, A. D., M. B. Akoachere, R. Krogh, C. Nickel, M. J. McLeish, G. L. Kenyon, L. D. Arscott, C. H. Williams, Jr., E. vioud-Charvet, and K. Becker. 2006. Specific inhibitors of *Plasmodium falciparum* thioredoxin reductase as potential antimalarial agents. Bioorg. Med. Chem. Lett. 16:2283-2292. doi:10.1016/j.bmcl.2006.01.027.

46. Akoachere, M., K. Buchholz, E. Fischer, J. Burhenne, W. E. Haefeli, R. H. Schirmer, and K. Becker. 2005. *In vitro* assessment of methylene blue on chloroquine-sensitive and -resistant *Plasmodium falciparum* strains reveals synergistic action with artemisinins. Antimicrob. Agents Chemother. 49:4592-4597. doi:10.1128/AAC.49.11.4592-4597.2005.

47. Bischoff, E. and C. Vaquero. 2010. *In silico* and biological survey of transcription-associated proteins implicated in the transcriptional machinery during the erythrocytic development of *Plasmodium falciparum*. BMC. Genomics 11:34. doi: 10.1186/1471-2164-11-34.

49. D'mello, N. P., A. M. Childress, D. S. Franklin, S. P. Kale, C. Pinswasdi, and S. M. Jazwinski. 1994. Cloning and characterization of LAG1, a longevity-assurance gene in yeast. J. Biol. Chem. 269:15451-15459.

50. Jiang, J. C., P. A. Kirchman, M. Allen, and S. M. Jazwinski. 2004. Suppressor analysis points to the subtle role of the LAG1 ceramide synthase gene in determining yeast longevity. Exp. Gerontol. 39:999-1009.

51. Labaied, M., A. Dagan, M. Dellinger, M. Geze, S. Egee, S. L. Thomas, C. Wang, S. Gatt, and P. Grellier. 2004. Anti-*Plasmodium* activity of ceramide analogs. Malar. J. 3:49. doi:10.1186/1475-2875-3-49.

52. Pankova-Kholmyansky, I., A. Dagan, D. Gold, Z. Zaslavsky, E. Skutelsky, S. Gatt, and E. Flescher. 2003. Ceramide mediates growth inhibition of the *Plasmodium falciparum* parasite. Cell Mol. Life Sci. 60:577-587.

53. Trager, W. and J. B. Jensen. 1976. Human malaria parasites in continuous culture. Science 193:673-675.

54. Lambros, C. and J. P. Vanderberg. 1979. Synchronization of *Plasmodium falciparum* erythrocytic stages in culture. J. Parasitol. 65:418-420.

55. Rosario, V. 1981. Cloning of naturally occurring mixed infections of malaria parasites. Science 212:1037-1038.

FIGURE LEGENDS

Figure 1. Recrudescence and morphological dormancy are observed in D6 parasites. A) The proportion of normal ring, trophozoite and schizont stage parasites in D6 parrasite culture over time following 6 hr of treatment (200 ng/mL DHA). Following treatment with DHA, there was a reduction in the number of normal parasites. Normal forms reappeared and recrudesced more than 3 days post treatment. Dormant parasites were observed from 6-144 hours PT. B) At six hours PT, small dormant forms were observed, which are characterized by condensed nuclei and cytoplasmic staining.

Figure 2. Microarray time course experimental design. Clones of the parental line D6 and the resistant line D6.QHS2400x5 were expanded into both highly synchronous ring stage cultures and asynchronous mixed stage cultures. Approximately 8 hr post-invasion, cultures were split into untreated (treated with DMSO) and treated with DHA (200 ng/mL) (purple arrows). Six hr post-treatment (PT), treated cultures (orange arrows), as well as, untreated cultures (gray arrows), were washed with RPMI to remove residual DHA. The tick marks for each culture indicate samples hybridized against pool RNA on the microarray. The transcriptome time course was conducted until 104 hr PT, after which time the parasites exhibited a normal asynchronous growth by smear, reaching ≥ 3% parasitemia.

Figure 3. Dihydroartemisinin induces transcriptome arrest in synchronous *P. falciparum* parasites. Transcriptomes of the parental (D6) and resistant (D6.QHS2400x5) strains were compared against a published transcriptome for HB3 (23) to determine the gross developmental state of each strain at different time points post-treatment with DHA (200 ng/mL). A and C) As expected, the untreated synchronous parasites exhibit a transcriptome that is diagnostic of their developmental stage. B and D) In both strains, DHA treatment induces a developmental arrest in the transcriptome for approximately 68 hr in the resistant strain and 92 hr in the parent strain. After recrudescence, time points exhibit lower peak correlations and shifts away from the ring-like state. Strong positive Pearson's correlations are depicted in yellow and negative correlations in blue. The horizontal axis represents experimental time, and the width of the bars reflects the spacing of the time points. The vertical axis represents the normal timeline of transcriptome development.

Figure 4. Ratio of normal parasite forms from the microarray time course with synchronized D6 parent and QHS-selected strains. Ratio of normal/total parasites for D6 (blue diamonds) and D6.QHS2400x5 (red squares) synchronized treated cultures. The drug had the expected effect of reducing the number of normal forms in both strains as shown in Figure 1.

Figure 5. Dihydroartemisinin induces transcriptome arrest and synchrony in asynchronous *P. falciparum* parasites. A) Asynchronous, untreated

cultures of D6 do not exhibit a strong correlation with the synchronous transcriptome progression. B) However, the transcriptome of DHA-treated asynchronous D6 parasites converges on a ring-like state. C) Similarly, asynchronous D6.QHS2400x5 exhibited the expected transcriptome asynchrony, but when treated with DHA, the transcriptome converges on a ring-like state (D). The yellow color indicates a positive Pearson correlation while a negative Pearson correlation is in blue.

Figure S1. Pilot experiment of dihydroartemisinin treatment of synchronized W2 parasites shows induction of a transcriptome arrest in a ring-like state. In a pilot microarray experiment to determine the effect of DHA on the transcriptome of *P. falciparum* parasites, synchronized rings of strain W2 were exposed to 100 nM DHA for 6 hr. Pellets were collected at 6 and 27 hr post-drug treatment, RNA was isolated, and samples were analyzed by microarray. Array data from T6 and T27 were independently correlated with transcriptome data from every hour of the normal intraerythrocytic developmental cycle (IDC) of HB3 (23). The data correlated best with HB3 12-13 hr post-invasion, suggesting there is an arrest in development at a state most similar to the ring-stage following treatment with DHA. The two columns represent the two collection time points of 6 and 27 hr after drug. The yellow color indicates a positive Pearson correlation to a given published normal growth time point, while blue indicates a negative Pearson correlation.

Figure S2. *P. falciparum* parasitemia and morphology during the microarray time course assays using D6 strains. A) Following treatment with DHA (200 ng/mL), there was a decrease in parasitemia (*i.e.* normal, dead, and dormant forms) for both the parental and resistant strains was observed. As expected, the untreated parasites showed a marked increase in parasitemia around 48 hr when schizont rupture is expected to occur. B) In the asynchronous cultures, treatment with DHA results in a decrease in parasitemia for up to 50 hr post-treatment (PT). C) Assessment of the parasitemia of normal parasites, excluding dead or dormant forms, shows that overall the asynchronous cultures recrudesced faster that the synchronized parasites. A=Asynchronous; S=Synchronous. D) Representative images from Giemsa stained blood smears show parasite stages present during the time course.

Table 1 – Genes significantly up- or down-regulated in both D6 and D6.QHS2400x5 during transcriptome arrest as compared to T0 hr after treatment with DHA (Only genes mentioned specifically in the text).

PlasmoDB ID	Description	Score(d)	Fold Change	local fdr(%)
PF14_0017	lysophospholipase putative	2.55	8.44	0
PFI1170c	thioredoxin reductase	2.23	10.63	0.19
PFD0740w	cyclin-dependent kinase putative	-1.57	-2.98	0
PF13_0088	Myb1 protein	-1.74	-3.2	0
PF10_0327	Myb2 protein	-1.75	-3.31	0

Score (d) refers to delta score from SAM analysis. Fold change refers to the geometric mean fold difference between T0 and later time points. Local FDR refers to the false discovery rate for data with the corresponding delta score. The full table of significant oligos is provided in the supplemental information.

Table 2 – Genes significantly up- or down-regulated in D6.QHS2400x5 relative to D6 during the normal IDC (no DHA treatment). (Only genes mentioned in the text). Score (d) refers to delta score from SAM analysis. Fold change refers to the geometric mean fold difference in expression between D6.QHS2400x5 and D6. Local FDR refers to the false discovery rate for data with the corresponding delta score. The full table of significant oligos is provided in the supplemental information.

PlasmoDB ID	Description	Score(d)	Fold Change	local fdr(%)
PFE0405c	Longevity-assurance (LAG1) domain protein putative	1.94	1.99	1.87
PF10_0300	RNA methyltransferase putative	1.81	1.87	3.91
PF10_0298	26S proteasome subunit putative	2.08	1.68	0.26
PF10_0296	conserved *Plasmodium* protein	2.34	2.34	0
PF10_0294	RNA helicase putative	2.53	1.81	0
PF10_0291	RAP protein putative	2.5	2.62	0
PF10_0287	conserved *Plasmodium* protein	2	1.98	1.02
PF10_0283a	product unspecified	2.85	2.6	0
PF10_0282	conserved *Plasmodium* protein	2.47	3.06	0
PF10_0281	Merezoite TRAP	2.16	2.27	0

Table 3 – Genes significantly up- or down-regulated in D6.QHS2400x5 relative to D6 after treatment with DHA as compared to T0. (Only genes mentioned specifically in the text)

PlasmoDB_ID	Description	Score(d)	Induction Fold Change	local fdr(%)
PFL1330c	cyclin-related protein Pfcyc-2	-3.59	-3.18	0.1
PF11_0245	translation elongation factor EF-1 subunit alpha putative	-5.64	-4.77	0.17

Score (d) refers to delta score from SAM analysis. Induction fold change refers to the geometric mean fold difference in induction of expression between D6.QHS2400x5 and D6 after DHA exposure. Local FDR refers to the false discovery rate for data with the corresponding delta score. The full table of significant oligos is provided in the supplemental information.

FIGURES

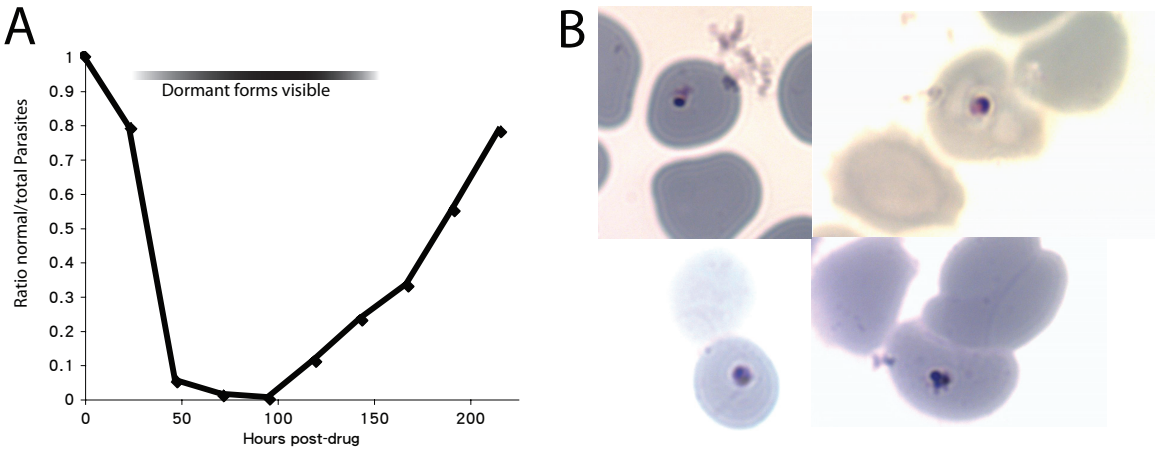


Figure 1. Recrudescence and morphological dormancy are observed in D6 parasites. A) The proportion of normal ring, trophozoite and schizont stage parasites in D6 parrasite culture over time following 6 hr of treatment (200 ng/mL DHA). Following treatment with DHA, there was a reduction in the number of normal parasites. Normal forms reappeared and recrudesced more than 3 days post treatment. Dormant parasites were observed from 6-144 hours PT. B) At six hours PT, small dormant forms were observed, which are characterized by condensed nuclei and cytoplasmic staining.

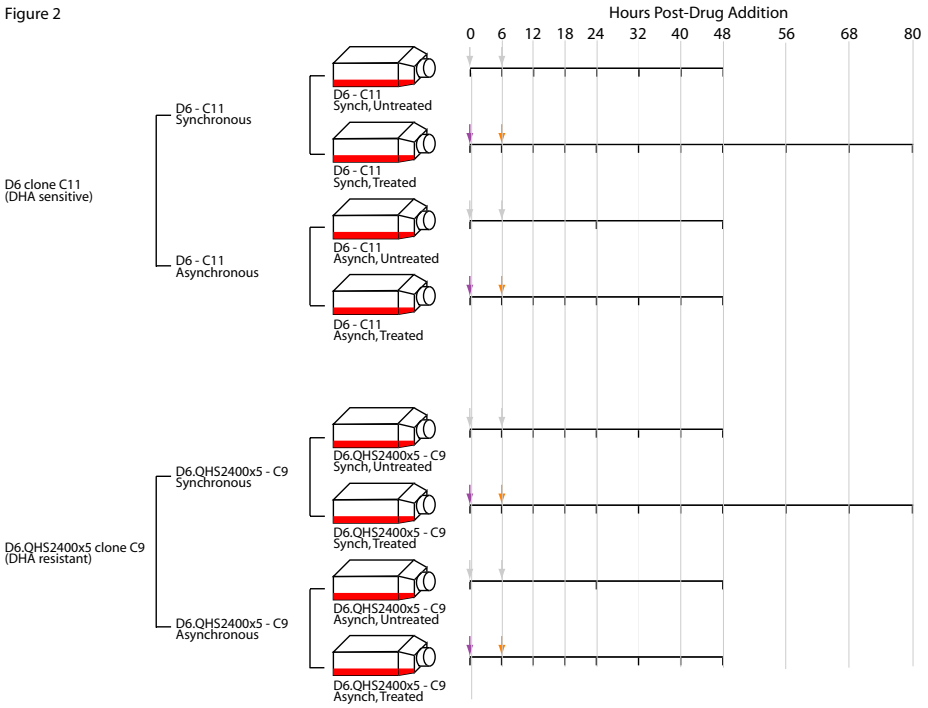


Figure 2. Microarray time course experimental design. Clones of the parental line D6 and the resistant line D6.QHS2400x5 were expanded into both highly synchronous ring stage cultures and asynchronous mixed stage cultures. Approximately 8 hr post-invasion, cultures were split into untreated (treated with DMSO) and DHA treated (200 ng/mL) (purple arrows). Six hr post-treatment (PT), treated cultures (orange arrows), as well as, untreated cultures (gray arrows) were washed with RPMI to remove residual DHA. The tick marks for each culture indicate samples hybridized against pool RNA on the microarray. The time course is only shown up to 104 hr PT, but additional samples were taken every 12 hr until treated synchronous cultures reached ≥ 3% parasitemia.

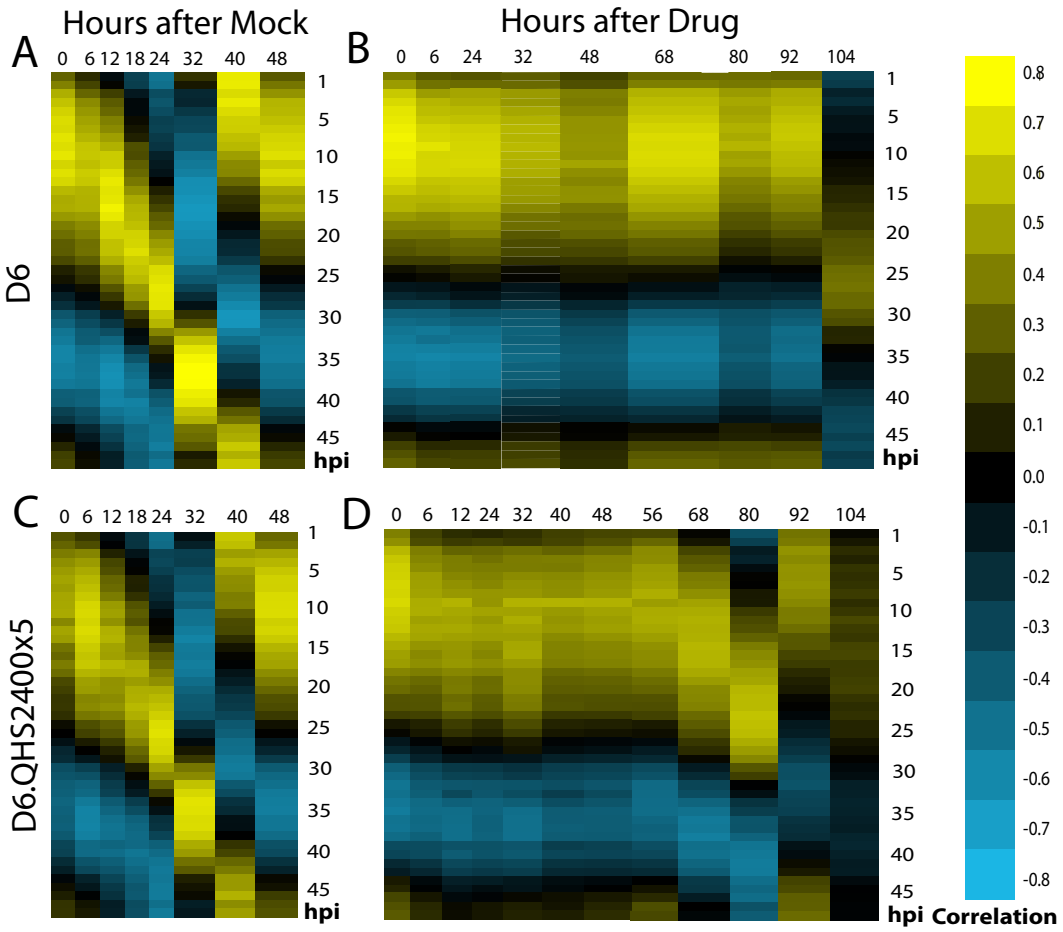


Figure 3. Dihydroartemisinin induces transcriptome arrest in synchronous *P. falciparum* parasites. Transcriptomes of the parental (D6) and resistant (D6.QHS2400x5) strains were compared against a published transcriptome for HB3 (23) to determine the gross developmental state of each strain at different time points post-treatment with DHA (200 ng/mL). A and C) As expected, the untreated synchronous parasites exhibit a transcriptome that is diagnostic of their developmental stage. B and D) In both strains, DHA treatment induces a developmental arrest in the transcriptome for approximately 68 hours in the resistant strain and 92 hours in the parent strain. After recrudescence, time points exhibit lower peak correlations and shifts away from the ring-like state. Strong positive Pearson's correlations are depicted in yellow and negative correlations in blue. The horizontal axis represents experimental time, and the width of the bars reflects the spacing of the time points. The vertical axis represents the normal timeline of transcriptome development.

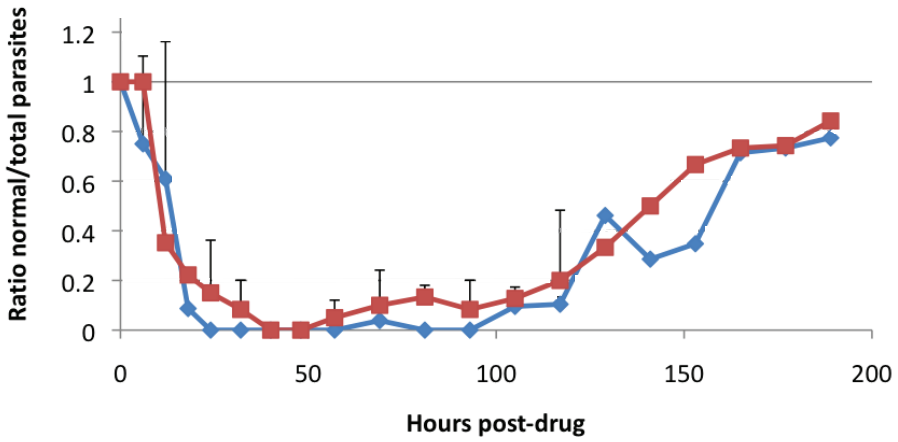


Figure 4. Recrudescence post 6 hr DHA treatment. Ratio of normal/ total parasites for D6 (blue diamonds) and D6.QHS2400x5 (red squares) synchronized treated cultures. The drug had the expected effect of reducing the number of normal forms in both strains as shown in figure 1.

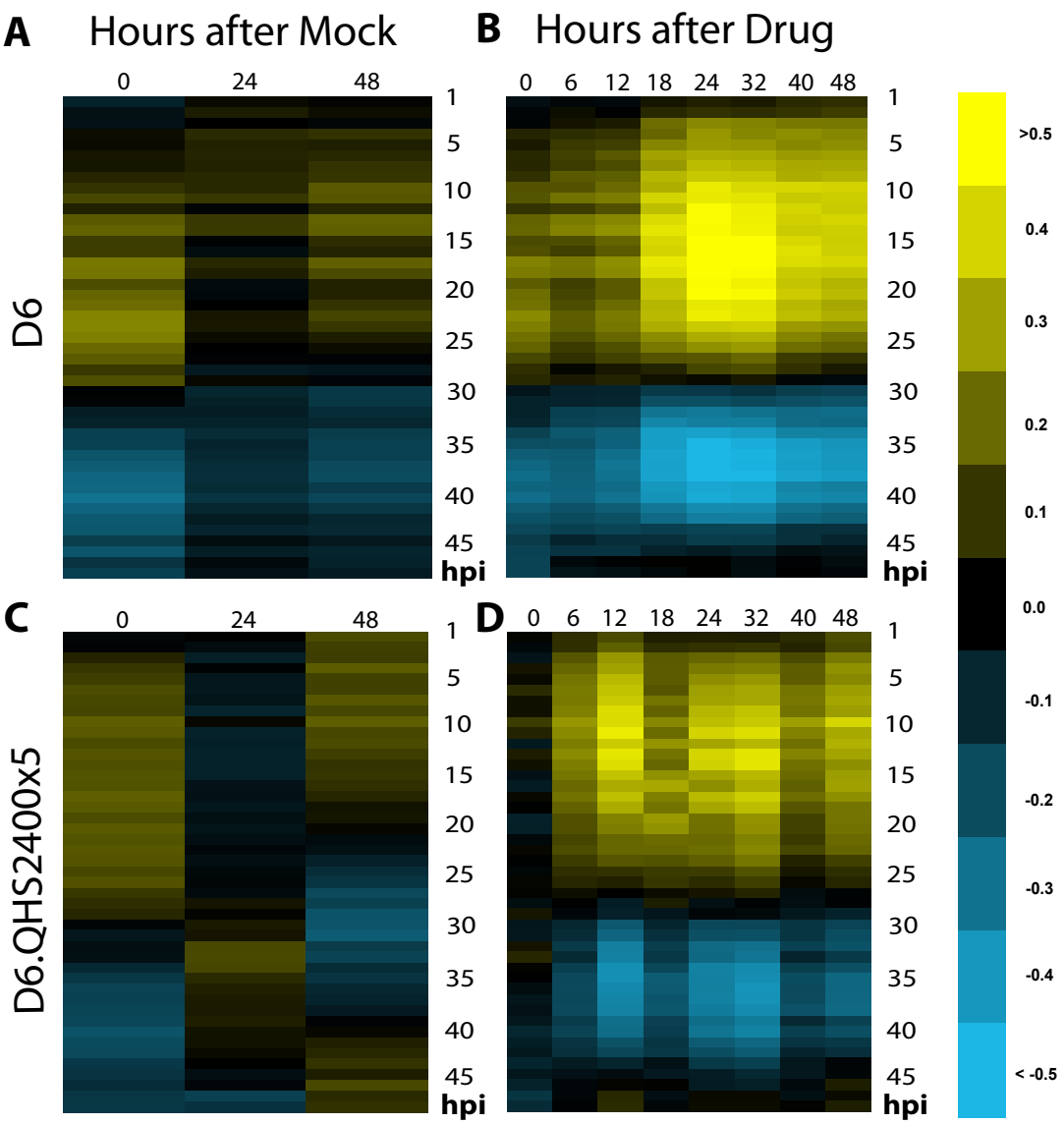


Figure 5. Dihydroartemisinin induces transcriptome arrest and synchrony in asynchronous *P. falciparum* parasites. A) Asynchronous, untreated cultures of D6 do not exhibit a strong correlation with the synchronous transcriptome progression. B) However, the transcriptome of DHA-treated asynchronous D6 parasites converges on a ring-like state. C) Similarly, asynchronous D6.QHS2400x5 exhibited the expected transcriptome asynchrony, but when treated with DHA, the transcriptome converges on a ring-like state (D). The yellow color indicates a positive Pearson correlation while a negative Pearson correlation is in blue.

SUPPORTING INFORMATION

Table S1 full list of genes associated with arrest passing significance filters

oligo ID	PlasmoDB ID	Description	Score(d)	Fold Change	local fdr(%)
N145_33	PF14 0014	Plasmodium exported protein unknown function	2.57	4.52	0
N145_22	PF14 0017	lysophospholipase putative	2.55	8.44	0
oPFI17632	PFI1170c	thioredoxin reductase	2.23	10.63	0.19
oPF08_0001_261	PF08 0001	Plasmodium exported protein unknown function	2.22	3.87	0.2
D56470_2	PFI1520w	asparagine-rich antigen putative	2.2	8.43	0.23
F62396_2	MAL7P1.144	Serine/Threonine protein kinase FIKK family	2.16	3.59	0.28
N143_54	PF14 0183	signal recognition particle RNP putative	2.13	7.01	0.31
oPFN0249	PF14 0010	glycophorin binding protein family Gbph	2.12	8.82	0.32
B587	PFB0930w	Plasmodium exported protein (hyp9) unknown function	2	3.6	0.55
oMAL6P1.106_83	PFF0510w	histone H3	1.99	7.46	0.57
oPFA0395c_496	PFA0395w	conserved Plasmodium protein unknown function	1.93	3.02	0.69
J33_27	PF10 0013	Plasmodium exported protein (hyp12) unknown function	1.92	3.59	0.71
L1_39	PFL0055c	RESA-like protein with PHIST and DnaJ domains	1.9	4.61	0.76
J33_15	PF10 0020	alpha/beta hydrolase putative	1.87	3.08	0.83
oPFG0019	MAL7P1.58	Pfmc-2TM Maurer's cleft two transmembrane protein	1.84	3.44	0.91
N143_57	PF14 0180	conserved Plasmodium protein unknown function	1.78	5.81	1.04
A26463_4	PFA0130c	Serine/Threonine protein kinase FIKK family putative	1.74	3.34	1.18
oPFL0108	PFL2175w	ubiquitin conjugating enzyme E2 putative	1.71	3.32	1.27
oPFrRNA0004	28S rRNA Chr7		1.7	5.65	1.29
oPFC0360w_45	PFC0360w	Activator of Hsp90 ATPase homolog 1-like protein putative	1.69	6.92	1.31
M20186_2	MAL13P1.470	Plasmodium exported protein (PHISTa) unknown function	1.67	3.25	1.38

Table S2 full list of genes passing significance filters for constitutive strain specific expression differences

Gene ID	PlasmoDB_ID	Description	Score(d)	Fold change	local fdr(%)
Ks488_10	Unannotated trans	700bp from the 3' end of PF11 0298	1.86	1.92	3.02
Kn8928_4	PFL2415w	Hbeta58/Vps26 protein homolog putative	2.27	2.02	0
oPFL1785c_762	PFL1785c	conserved Plasmodium protein	1.88	1.99	2.71
oPFL1750c_2362	PFL1750c	conserved Plasmodium protein	2.1	2.69	0.01
Km765_1	PFL1540c	phenylalanyl-tRNA synthetase alpha chain putative	2.26	1.64	0
L2_280	PFL0795c	male development gene 1	2.53	3.29	0
L2_212	PFL0660w	dynein light chain 1 putative	1.96	1.84	1.54
L1_32	PFL0045c	Plasmodium exported protein (PHISTc)	-2.67	-2.42	0
oPFI17673	PFI1110w	glutamine synthetase putative	2.15	1.78	0
oPFI17676	PFI0230c	bacterial histone-like protein	1.89	1.84	2.53
F38025_1	PFF1260c	conserved Plasmodium protein	-2.15	-1.93	0
F14111_3	PFF1100c	transcription factor with AP2 domain(s) putative	1.95	1.74	1.73
oMAL6P1.209_5'	PFF1055c	conserved Plasmodium protein	2.18	2.04	0
F16755_1	PFF0930w	conserved Plasmodium protein	2.02	1.91	0.88
oMAL6P1.280_5'	PFF0705c	conserved Plasmodium protein	2.19	2.49	0
oMAL6P1.106_8'	PFF0510w	histone H3	2.41	2.47	0
D57574_1	PFE1615c	Plasmodium exported protein	2.63	10.48	0
F13845_1	PFE1605w	Plasmodium exported protein (PHISTb)	4.31	20.01	0.06
oPFE1600w_785	PFE1600w	Plasmodium exported protein (PHISTb)	6.01	51.8	0
F13309_2	PFE1270c	WD domain G-beta repeat-containing protein	-2.42	-1.99	0
D53895_3	PFE0730c	ribose 5-phosphate epimerase, putative	2.07	1.92	0.31
E19231_1	PFE0570w	RNA pseudouridylate synthase putative	2.28	2.18	0
F71176_2	PFE0405c	Longevity-assurance (LAG1) domain protein putative	1.94	1.99	1.87
E6820_1	PFE0250w	conserved Plasmodium protein	-2.68	-2.34	0
E2283_4	PFE0130c	conserved Plasmodium protein	-2.53	-2.29	0
F32316_2	PFE0120c	Merozoite Surface Protein 8 MSP8	1.97	2.16	1.38
oPFD1200c_505	PFD1200c	Plasmodium exported protein (hyp6)	2.48	2.95	0
D10455_2	PFD1185w	Plasmodium exported protein (PHISTa)	3.85	3.2	0
D27953_2	PFD1180w	Plasmodium exported protein (PHISTb)	1.84	1.76	3.35
D49942_2	PFD0095c	Plasmodium exported protein (PHISTb)	2.51	2.57	0
C442	PFC0675c	mitochondrial ribosomal protein L29/L47 precursor putative	1.99	2.14	1.12
C240	PFC0370w	conserved Plasmodium protein	2.06	2.04	0.44
B603	PFB0953w	Plasmodium exported protein (hyp15)	4.62	4.84	0.05
B306	PFB0435c	transporter putative	-2.41	-2.09	0
oPFB0161c_228	PFB0161c	conserved Plasmodium protein	2.01	2.24	0.99
A11546_1	PFB0100c	knob-associated histidine-rich protein	7.17	41.98	0
B52	PFB0090c	RESA-like protein with PHIST and DnaJ domains	6.7	22.08	0
B50	PFB0085c	DNAJ protein putative	4.93	27.18	0.03
B47	PFB0080c	Plasmodium exported protein (PHISTb)	5.14	9.35	0.02
B45	PFB0075c	Plasmodium exported protein (hyp9)	3.62	4.12	0
N129_1	PF14_0745	probable protein	2.64	3.14	0
N133_39	PF14_0691	conserved Plasmodium membrane protein	1.95	1.69	1.66
N133_46	PF14_0686	conserved Plasmodium protein	3.11	2.15	0
N187_1	PF14_0678	exported protein 2	-2.3	-2.28	0
N137_29	PF14_0667	conserved Plasmodium protein	1.86	1.99	3.03
M35431_1	PF14_0631	conserved Plasmodium protein	1.87	2.23	2.85
N134_113	PF14_0586	conserved Plasmodium protein	1.97	2.02	1.47
oPFN0248	PF14_0102	rhoptry-associated protein 1 RAP1	1.94	2.18	1.82
N171_3	PF14_0073	conserved Plasmodium protein	-2.54	-1.8	0
M951_1	PF13_0222	phosphatase putative	1.83	1.76	3.63
M46928_1	PF13_0190	conserved Plasmodium protein	1.84	1.58	3.49
oPFK12891	PF11_0512	RESA-like protein with PHIST and DnaJ domains	5.93	6.29	0
Ks76_9	PF11_0423	conserved Plasmodium protein	2.08	1.92	0.18
Ks115_1	PF11_0359	coatomer delta subunit putative	3.19	2.74	0
Ks89_8	PF11_0356	conserved Plasmodium protein	1.92	1.79	2.08
Ks97_2	PF11_0192	histone acetyltransferase putative	2.01	1.68	0.94
Ks26_16	PF11_0114a	conserved Plasmodium protein	2.44	2.47	0
J116_13	PF10_0344	glutamate-rich protein	2.27	2.58	0
J461_7	PF10_0300	RNA methyltransferase putative	1.81	1.87	3.91
J64_2	PF10_0298	26S proteasome subunit putative	2.08	1.68	0.26
J564_2	PF10_0296	conserved Plasmodium protein	2.34	2.34	0
J125_3	PF10_0294	RNA helicase putative	2.53	1.81	0
J383_1	PF10_0291	RAP protein putative	2.5	2.62	0
J151_8	PF10_0287	conserved Plasmodium protein	2	1.98	1.02
J63_1	PF10_0283a		2.85	2.6	0
J63_5	PF10_0282	conserved Plasmodium protein	2.47	3.06	0
J62_1	PF10_0281	merozoite TRAP-like protein MTRAP	2.16	2.27	0
J232_8	PF10_0258	conserved Plasmodium protein	2.42	2.76	0
F42768_1	PF10_0232	Chromodomain-helicase-DNA-binding protein 1 homolog putative	1.98	2.16	1.29
J106_11	PF10_0214	RNA binding protein putative	1.86	1.77	3.01
E19346_1	PF10_0188	conserved Plasmodium membrane protein	2.06	2.03	0.46
J73_4	PF10_0084	tubulin beta chain putative	1.94	2.15	1.8
F23846_3	PF08_0034	histone acetyltransferase GCN5 putative	1.87	1.53	2.81
F67443_1	PF07_0101	conserved Plasmodium protein	1.86	1.91	3.06
F44837_1	No_ORFs	500bp from the 5' end of PFE0680w	2.13	1.94	0
D33539_1	No_ORFs	700bp from the 5' end of PFD0695w	2.06	1.68	0.45
N145_38	No_ORFs	Chr 14 1kb from 3' end of PF14 0013	-2.86	-2.92	0
F21560_1	MAL8P1.88	conserved Plasmodium protein	-2.41	-2.2	0
oMAL8P1.27_22	MAL8P1.27	translation initiation factor IF-3 putative	1.96	2.25	1.49
F18577_5	MAL8P1.134	ferlin like protein putative	-2.34	-2.14	0
F17165_1	MAL8P1.104	CAF1 family ribonuclease putative	2.58	2.31	0
E17521_1	MAL7P1.171	Plasmodium exported protein	2.45	2.83	0
Ks222_1	MAL13P1.420	hypothetical protein	1.83	1.83	3.53
M3696_2	MAL13P1.333	conserved Plasmodium protein	1.88	1.7	2.72
M35930_9	MAL13P1.228	conserved Plasmodium protein	1.97	2.15	1.41
M32813_2	MAL13P1.184	endopeptidase putative	3.23	2.83	0

Table S3. Full list of genes passing significance filters for differential induction between the two D6 lines

oligo ID	PlasmoDB_ID	Description	Score(d)	Fold Change	local fdr(%)
oPFE1020w_182	PFE1020w	U6 snRNA-associated sm-like protein Lsm2 putative	5.18	2.52	0.02
N164_5	PF14_0229	conserved Plasmodium protein unknown function	4.71	2.99	0.04
F27786_1	PF07_0086	conserved Plasmodium membrane protein unknown function	4.57	2.91	0.04
C677	PFC1016w	conserved Plasmodium protein unknown function	4.52	16.59	0.04
I11857_2	PFI0865w	XPA binding protein 1 putative	4.45	2.82	0.04
L2_104	PFL0405w	conserved Plasmodium protein unknown function	4.28	2.38	0.04
F7915_1	PFF0305c	ubiquitin conjugating enzyme E2 putative	4.14	2.33	0.03
oPF11_0152_17	PF11_0152	GTPase activator putative	4.08	2.16	0.02
oPFL0023	PFL1745c	clustered-asparagine-rich protein	3.91	8.19	0
N135_22	Gene absent in PlasmoDB v6.3	transcript between PF14_0633 and PF14_0634	3.91	2.43	0
M38913_1	PF13_0352	conserved Plasmodium protein unknown function	3.82	2.17	0
oPFM60522	MAL13P1.47	mitochondrial ATP synthase delta subunit putative	3.76	1.79	0
C237	PFC0355c	conserved Plasmodium protein unknown function	-3.43	-5.23	0.07
Ks54_4	PF11_0036	flavoprotein putative	-3.44	-2.16	0.07
A8109_9	PFA0295c	conserved Plasmodium protein unknown function	-3.45	-4.65	0.07
A31914_4	PF10_0258	conserved Plasmodium protein unknown function	-3.48	-4.41	0.08
oPFL0022	PFL2460w	coronin	-3.5	-3.69	0.09
B541	PFB0845w	conserved Plasmodium membrane protein unknown function	-3.52	-2.5	0.09
Kn5186_3	PFL1180w	chromatin assembly protein (ASF1) putative	-3.58	-2.92	0.1
J4379_1	PFL1330c	cyclin-related protein Pfcyc-2	-3.59	-3.18	0.1
F30848_1	MAL8P1.101	RNA binding protein putative	-3.61	-3.64	0.11
oPFF72487	PFF0645c	integral membrane protein putative	-3.64	-8.84	0.11
N141_27	PF14_0224	serine/threonine protein phosphatase	-3.71	-4.96	0.13
B60	PFB0105c	Plasmodium exported protein (PHISTc) unknown function	-3.72	-3.31	0.13
F23699_1	PFI0265c	RhopH3	-3.74	-3.29	0.13
Ks510_10	PF11_0381	subtilisin-like protease 2	-3.77	-3.47	0.13
F53897_2	MAL7P1.119	conserved Plasmodium protein unknown function	-3.78	-3.76	0.14
D49176_7	PFD0230c	protease putative	-3.91	-4.81	0.15
Ks370_2	PF11_0268	kelch motif containing protein putative	-3.95	-4.63	0.15
J269_10	PF10_0231	conserved Plasmodium protein unknown function	-4	-2.63	0.15
A8408_2	PF07_0101	conserved Plasmodium protein unknown function	-4.14	-3.36	0.15
oPFN0262	PF14_0589	valine-tRNA ligase putative	-4.26	-7.35	0.15
oPFrRNA0002	MAL14_5S_1	5S rRNA	-4.42	-33.03	0.15
Ks1072_1	PF11_0168	moving junction protein	-4.58	-3.55	0.15
I4719_6	PFI0175w-a	conserved Plasmodium protein unknown function	-4.63	-19.39	0.15
B50	PFB0085c	DNAJ protein putative	-4.91	-23.52	0.14
N143_57	PF14_0180	conserved Plasmodium protein unknown function	-4.93	-10.04	0.14
Ks26_12	PF11_0116	conserved Plasmodium protein unknown function	-4.96	-3.29	0.15
N145_12	PF14_0020	choline kinase	-5.17	-3.75	0.15
F27351_1	PFI0690c	conserved Plasmodium protein unknown function	-5.2	-6.67	0.15
N155_25	PF14_0031a	conserved Plasmodium protein unknown function	-5.29	-4.09	0.16
Ks259_3	PF11_0245	translation elongation factor EF-1 subunit alpha putative	-5.64	-4.77	0.17

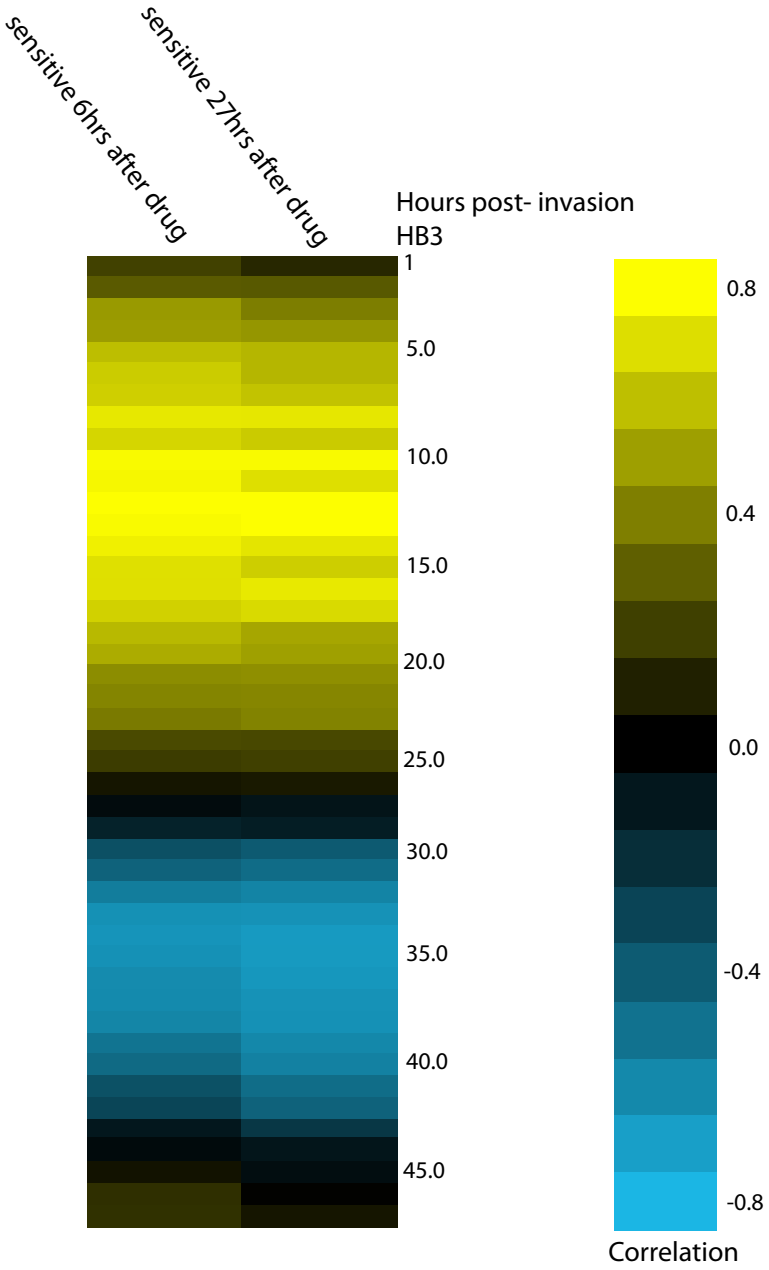


Figure S1. Pilot experiment of dihydroartemisinin treatment of synchronized W2 parasites shows induction of a transcriptome arrest in a ring-like state

In a pilot microarray experiment to determine the effect of DHA on the transcriptome of *P. falciparum* parasites, synchronized rings of strain W2 were exposed to 100 nM DHA for 6 hr. Pellets were collected at 6 and 27 hr post-drug treatment, RNA was isolated, and samples were analyzed by microarray. Array data from T=6 and T=27 were independently correlated with transcriptome data from every hour of the normal intraerythrocytic developmental cycle (IDC) of HB3 (23). The data correlated best with HB3 12-

13 hr post-invasion, suggesting there is an arrest in development at a state most similar to the ring-stage following treatment with DHA. The two columns represent the two collection time points of 6 and 27 hours after drug. The yellow color indicates a positive Pearson correlation to a given published normal growth time point, while blue indicates a negative Pearson correlation.

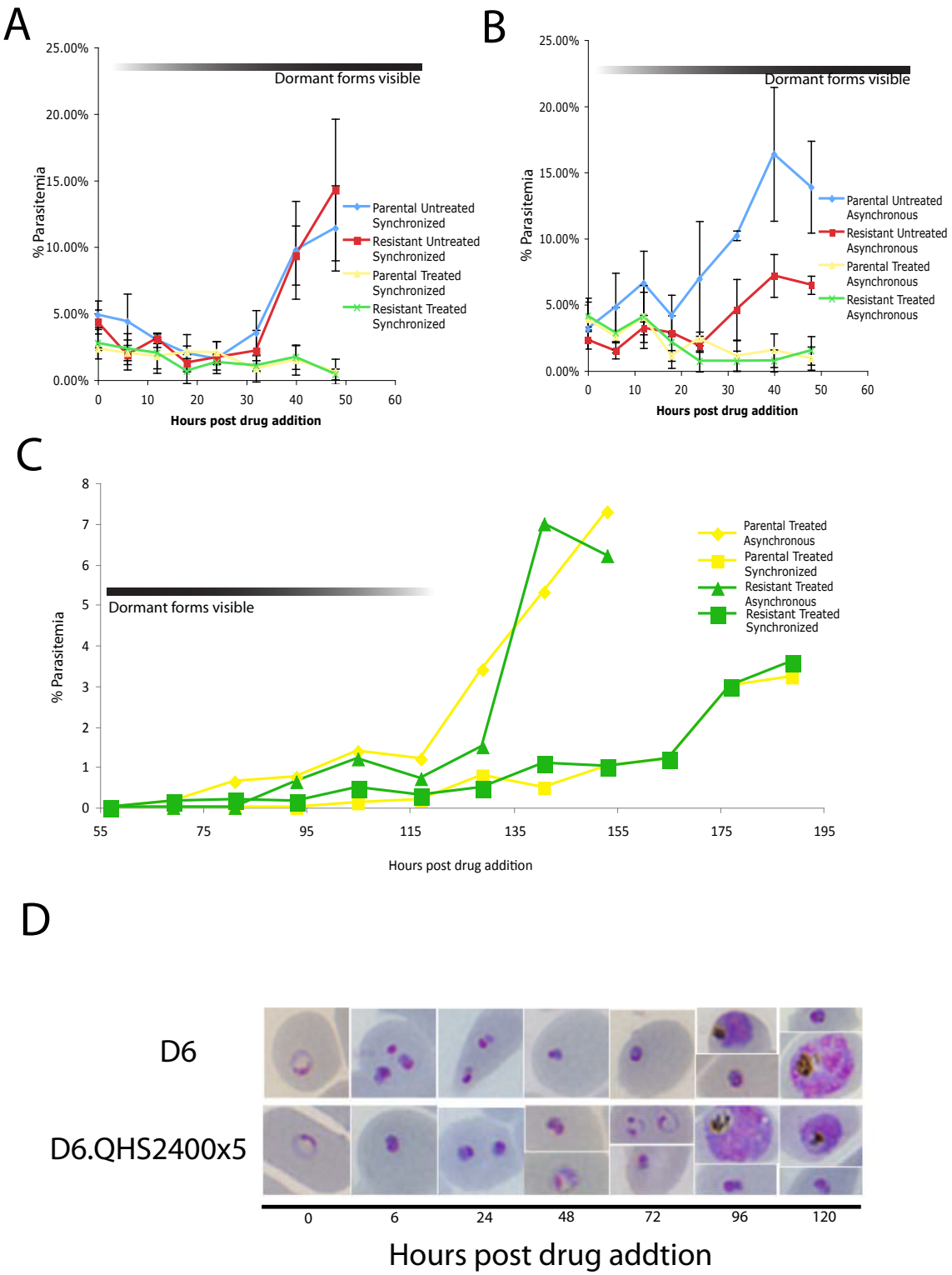


Figure S2. P. falciparum parasitemia and morphology during the microarray time course assays using D6 strains. A) Following treatment with 200 ng/ml DHA, there was a decrease in parasitemia (i.e. normal, dead, and dormant forms) for both the parental and resistant strains was

observed. As expected, the untreated parasites showed a marked increase in parasitemia around 48 hr when schizonts rupture is expected to occur. B) In the asynchronous cultures, treatment with DHA results in a decrease in parasitemia for up to 50 hr post-treatment (PT). C) Assessment of the parasitemia of normal parasites, excluding dead or dormant forms, shows that overall the asynchronous cultures recrudesced faster that the synchronized parasites. Synchronous cultures of D6.QHS2400x5 rose in parasitemia compared to D6, reflecting the earlier emergence from dormancy. A=Asynchronous; S=Synchronous. D) Example giemsa stain micrographs of both strains during the DHA challenge and outgrowth.

Chapter 5…….

MITOMI 2.0 device development

Christopher Nelson and Polly Fordyce

Author contributions:

CN and PF both participated in the device design improvements discussed here, and CN wrote the chapter.

Joseph L. DeRisi, Thesis Advisor

Abstract

There is still a considerable gap in our knowledge of how the cell reads its own genetic code. Sequence-specific DNA binding proteins comprise many families with different binding modes and binding site structures. For many DNA-binding proteins we do not know the precise preferred binding site or when we do know the binging site we do not understand the range of DNA sequences that they bind. MITOMI and MITOMI 2.0 comprise a suite of techniques to measure DNA binding affinities to libraries of DNA, in order to understand binding sit preference and structure. These techniques are still relatively new and evolving, so here we review recent design changes that we have made, along with general constraints and principles for MITOMI device design.

Introduction

Mechanically Induced Trapping of Mechanical Interactions (MITOMI) was initially developed by Sebastian Maerkl in the Quake laboratory to study the energetic effects of binding of a transcription factor against a known binding site (Maerkl and Quake 2007). These microfluidic devices measure the binding of a given transcription factor to libraries of DNA sequences (Fig 1 and 2). The devices are and are controlled by pressurized valves. Subsequently, Polly Fordyce extended the platform to enable motif discovery for transcription factors without known binding sites, calling the extended technique MITOMI 2.0 (Fordyce et al 2007). The difference between the MITOMI and MITOMI 2.0 modes of experiment is illustrated in Figure 3.

The devices, like many microfluidic devices are fabricated out of polydimethylsiloxane (*PDMS*) from molds (Fig 4). (Our molds are produced by mask photolithography with SU-8 and AZ50XT photoresists, as described in Fordyce et al 2010.) Our MITOMI devices are overlayed on top of a DNA microarray patterned using a DeRisi-style printer (Figure 5). MITOMI devices are two layer, valved devices (as opposed to single layer valve-less devices in the lineage of the Whitesides Laboratory). In push-down devices like MITOMI devices, the top control layer is comprised of dead-end channels that are loaded with water and pressurized or depressurized to allow fluid flow in the underlying flow layer. Tight seals can be made where a large footprint control valve intersects with a thin membrane on the top of the flow layer, just like stepping down on a garden hose.

In general defects in microfluidic devices are often caused by small pieces of dust. Dust can effect either the fabrication of the mold, which can result in misshapen PDMS features, or dust can enter during fabrication of the PDMS device itself. The most frequent failure modes we have noticed in MITOMI involve short circuits between the control valve lines. Figure 6 illustrates various potential means to observe such defects. To guard against such defects we moved all of our fabrication into class 1000-10,000 clean rooms.

These new devices were more than 6 times larger than the original devices to incorporate a larger random DNA library. The expansion of the MITOMI platform resulted in robustness problems. The original Quake Lab MITOMI microfluidic device was already at the leading edge of density and complexity for a silicone labs-on-a-chip (Geertz 2012). We had periods where MITOMI 2.0 devices only performed properly and generated data 30% of the time. The valves on the device would often fail to close fully, rendering them inoperable, or mixing reagents when inappropriate. This was especially true of the sandwich valve, which is constrained to snake through the device over a very long circuit, causing dissipation of the control pressure over a larger area and more opportunities for fabrication defects. An additional complication was that the expansion to larger devices increased the resistance to flow through the device, necessitating long running times and higher pressures (approaching 8 PSI for the flow pressure and 35 PSI for the control pressure)(Mortensen 2005).

In this chapter we review the developments we have made in the MITOMI platform to improve device robustness and the interpretation of the data. We also provide validation of the novel design with yeast and *Plasmodium falciparum* DNA-binding proteins.

Results

At first we redesigned the unit cell itself. Through trial and error we tried different valve design concepts, aiming at increased closure tightness and overall device robustness (Figure 7). We improved valve performance by building larger footprint valves. Additionally we spaced out the high-pressure control lines from each other to avoid the chance that a piece of dust or missing feature could short-circuit the device. At a larger scale than the individual unit cell, we wired the sandwich valves in parallel to improve sandwich valve robustness to single defects along its long circuit length, which could cause outages like a string of lights wired in series. This parallel circuit had the added benefit of improving closure speed and seal performance by halving the effective circuit length (the distance between the most distant sandwich valve and the pressure source). Before changing the format of the overall array of unit cells we revised the design of our MITOMI 2.0 DNA library.

Our initial DNA library that was useful for motif discovery incorporating enough sequence space to cover all possible 8mers. However the library had a periodicity such that neighboring 8mer motifs were closely related. This

complicated the interpretation of motif finding results. The periodicity resulted from the simple algorithm that we used to build of the random sequence library. The algorithm worked by building up the sequence, adding only 8mer strings that it hadn't seen before; however it made new motifs by changing a single base on the end of the previous motif, resulting in neighboring motifs that were related. Using programs from the Eisen lab we designed an improved library that was more thoroughly random and half of the size of the initial library (see FOXP2 chapter methods section). This halving of the size of our DNA library freed us to change the format of our MITOMI devices to avoid the robustness issues discussed above.

The format change enabled reduction in the device footprint and complexity of the devices, and resulted in much more robust devices. The format of arrays of unit cells has to conform to a few constraints. The geometries available with our printer are not infinitely adjustable. When printing with multiple pins you're restricted to choosing a number of columns in the array that is a multiple of the number of print pins, and the spacing between the columns must be a harmonic of the 4,500 μm pin spacing in the printer head. Furthermore one should consider the number of devices that can fit on a silicon mold. With our larger devices we could only fit one or two devices per mold, making our fabrication throughput relatively low. Large devices also can overlap the edge of the wafer, where the molds can be defective due to non-uniform illumination at the edge of the photolithography image. In contrast, the small DTPA-D devices, as described in

Maerkl and Quake 2007, can fit 6 devices per 4-inch wafer and avoid the edge of the wafer. Potential design layouts that would accommodate 740 oligonucleotide DNA library are described in Table1.

The last parameter that we considered was the expected flow pressure required to move reagents through the device, based on the equations of (Mortensen 2005). Slow reagent flow through the device slows the entire experimental day, and forcing the flow at higher pressures increases the risk for short circuits between layers and increases the required closing pressure. Therefore we calculated relative flow resistance to compare formats. These calculations are described below.

Single channel resistance=[viscosity* length*(cross-sectional perimeter)^2] / (cross-sectional area^3)

Most of these factors (i.e. viscosity of the flow fluid and the cross section of the channel) are going to remain pretty constant across different designs, so we can say that…

Single channel resistance = channel length times some constant.
That constant is (viscosity* perimeter^2/cross-sectional area^3)

Our goal is to compare different designs of devices with different numbers of parallel channels of varying lengths. Flow resistors add just as electrical resistors.

Therefore, to compare different designs we can use:

R_{tot}/constant=channel length/channels=1/[channels*(1/channel length)]

There are some caveats to the interpretation of these R_{tot} results. Empirically, a small DTPA-D chip tends to have less resistance than a large PF4K chip, yet with these Mortensen calculations we derive that the DTPA-D flow resistance should be twice that of the PF4K devices (Table 1). However, the general principle of expecting lower resistance with shorter channels and more parallel channels is one we used in our format selection.

We selected the formats named PC1K and CP1K for further development, because they have small footprints that would allow two replicates of the DNA library with moderate to low expected flow pressure. We made devices and aligned them to the arrays of the new DNA library. We validated the new PC1K devices and new library with pho4 protein from yeast (Fig 9A). The derived motif matched motif reports and results from prior designs. We also reconfigured the library print format to fit within a DTPA-D device. Once we had arrived at devices that performed properly more than 90% of the time we were ready to return from engineering to biology.

Our original primary biological target for MITOMI was characterizing the mostly unknown DNA sequence-specificities of *Plasmodium falciparum* transcription factors. Around the time we were revising our device designs a string of papers from Manuel Llinás', lab described the DNA binding specificity of 26 *Plasmodium* stage-specific DNA binding proteins. We profiled two AP2 proteins with our devices and confirmed their binding sites (Fig 9B). Interestingly the full-length proteins gave no DNA binding activity, suggesting that these factors may require post-translational cleavage to become active. We attempted to profile the DNA-binding affinities of other putative *Plasmodium* transcription factors (almost all of which have no documented binding motif, expect for the AP2 proteins, Table 2). However, none of the constructs that we assayed gave any activity in our MITOMI assay. Manuel Llinás, through personal communication, confirmed that his lab has attempted to profile many of the same factors, most notably the Myb family, and come up with no DNA-binding activity. Frankly, at this point we had to abandon *P. falciparum* MITOMI and turn to systems more amenable to study. With our improved MITOMI2.0 devices we first turned to a deceptively simple factor Hac1, described in the next chapter.

Acknowledgement

We would like to acknowledge the design input and facility support of Rafael Sjoberg-Gomez.

References

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

Fordyce, P. M. *et al.* De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. *Nat. Biotechnol* 28, 970–975 (2010).

Mortensen NA, Okkels F H. Bruus Reexamination of Hagen-Poiseuille flow: shape dependence of the hydraulic resistance in microchannels. Phys Rev E *Stat Nonlin Soft Matter Phys*. 2005 May;71(5 Pt 2):057301.

Geertz M, Shore D, Maerkl SJ. Massively parallel measurements of molecular interaction kinetics on a microfluidic platform. *Proc Natl Acad Sci* U S A. 2012 Oct 9;109(41):16540-5.

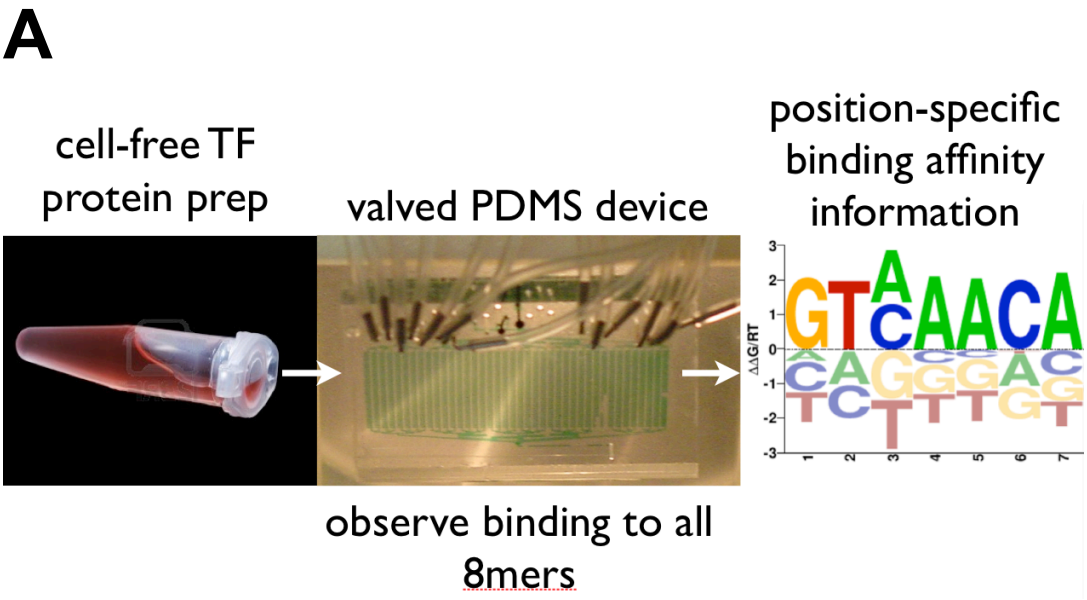
Table 1. Design parameters of various layouts of unit cells that could accommodate our new pseudorandom library, and be printed on our printer. The number of print head pins used for printing the design is noted in the pins column. “Cols” denotes the number of columns of unit cells and DNA spots. The following measurements are in microns. Row space and col space note the distance between adjacent DNA spots and thus the spacing between the unit cells. Area is the footprint area of the array of unit cells, which is useful in planning the number of chips one can pack onto a single mold. R_{tot} is the microfluidic flow resistance calculated according to Mortensen 2005, assuming an equal cross-sectional area of the flow channels resistance in arbitrary units.

pins	cols	rows	unit cells	cols/ pin	col space	row space	width	length	area	Name	Rtot
12	96	16	1536	8	562.5	325	5.E+03	5.E+04	3.E+08	-	54
12	72	21	1512	6	750	325	7.E+03	5.E+04	4.E+08	-	95
12	60	25	1500	5	900	325	8.E+03	5.E+04	4.E+08	-	135
12	48	31	1488	4	1125	325	1.E+04	5.E+04	5.E+08	-	210
12	36	42	1512	3	1500	325	1.E+04	5.E+04	7.E+08	-	379
1	16	40	640	na	680	320	1.E+04	1.E+04	1.E+08	DTPAD	800
12	60	65	3900	5	900	325	2.E+04	5.E+04	1.E+09	PF4K	352
12	60	25	1500	5	900	325	8.E+03	5.E+04	4.E+08	PF4K-1/2	135
8	40	38	1520	5	900	325	1.E+04	4.E+04	4.E+08	CP1K	309
8	48	32	1536	6	750	325	1.E+04	4.E+04	4.E+08	-	217
4x4	28	56	1568	14 or 7	321	643	4.E+04	9.E+03	3.E+08	PC1K	1286

Table2. MITOMI Plasmodium falciparum constructs attempted.

GeneID	Domain	Constructs
PFL0815w	myb	1
PFI1665c	ap2	1
PFF0200c	ap2	2
PFC0690c	c2h2 Zn finger	3
PF14_0633	ap2	2
PF14_0471	ap2	2
PF14_0271	ap2	1
PF11_0293	HTH multibridging factor	1
PF10_0327	myb	5
PF10_0091	c2h2 Zn finger	3
MAL13P1.395	krox	2

Figures



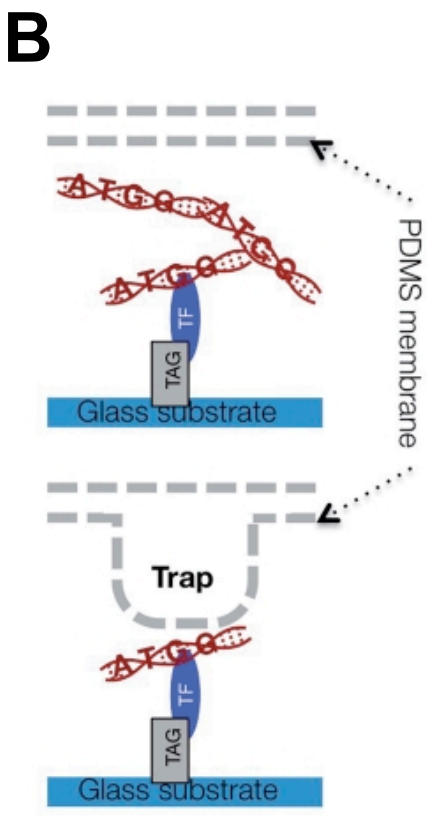


Figure 1. Overview of MITOMI2.0. A) The MITOMI2.0 devices are intended to take a protein sample, frequently expressed in a cell-free extract , and run the

sample through a microfluidic device containing a library of DNA sequences the derive the binding preferences of the protein. B) The microfluidic valve at the heart of the device is a button valve that traps the interaction between tag-immobilized protein and a DNA sequence that it has bound (from Geertz 2010).

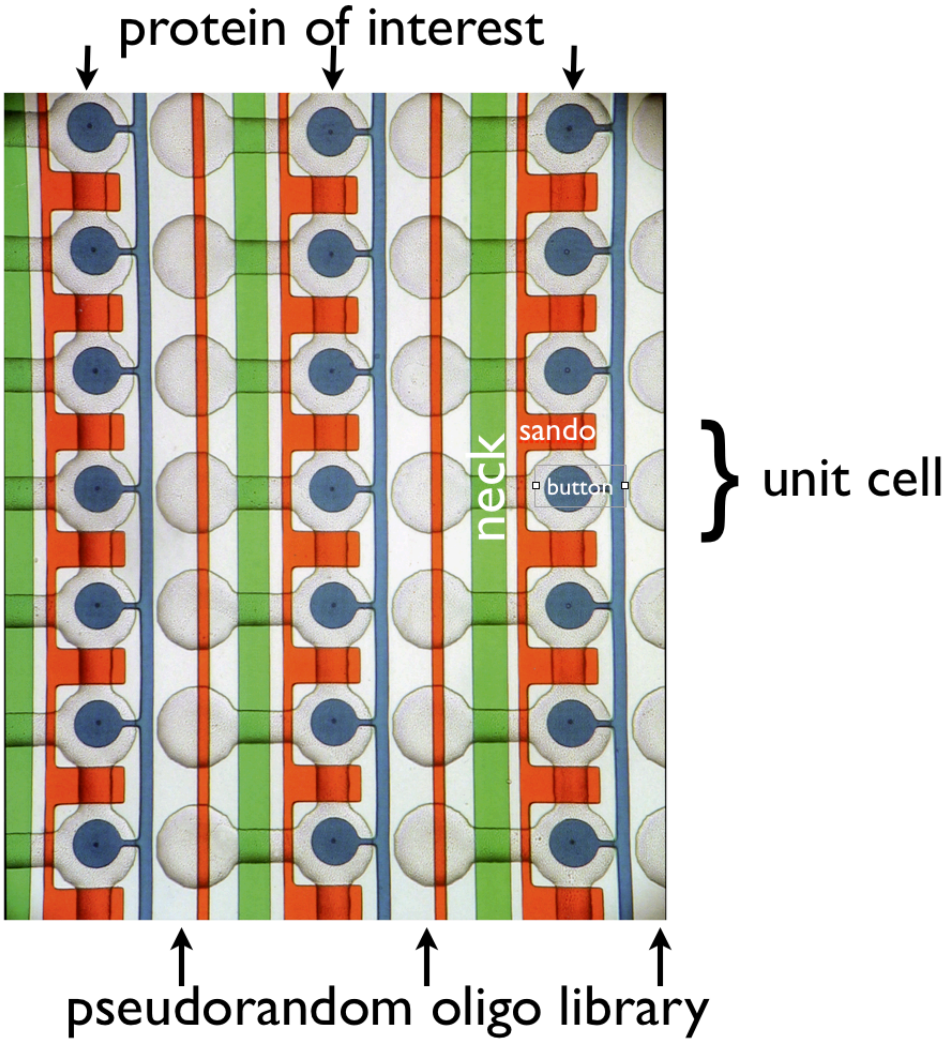


Figure 2. Schematic of the MITOMI devices. Micrograph of a portion of a MITOMI device with the control lines filled with food coloring. Blue is the button valve line, orange is the sandwich valve line and green is the neck valve line. For scale, the distance between the clear DNA chambers is 600 microns. A unit cell

of the device is defined by a pair of a DNA chamber and a protein chamber, with flow controlled by the button, sandwich and neck valves. The neck valve limits the flow between adjacent unit cells so that each reaction can be observed independently. The neck valve closes of the DNA chamber from the outside world so that the protein chamber can be derivitized and loaded with protein, without disturbing the DNA chamber. After incubation of the protein of interest with the DNA in each unit cell, the button valve is used to trap the bound protein and DNA and extrude the unbound DNA. The device is then scanned in a microarray scanner to observe the ratio of the fluorescently labeled protein and DNA at each unit cell in the device. Our devices range from 640 to 3900 unit cells.

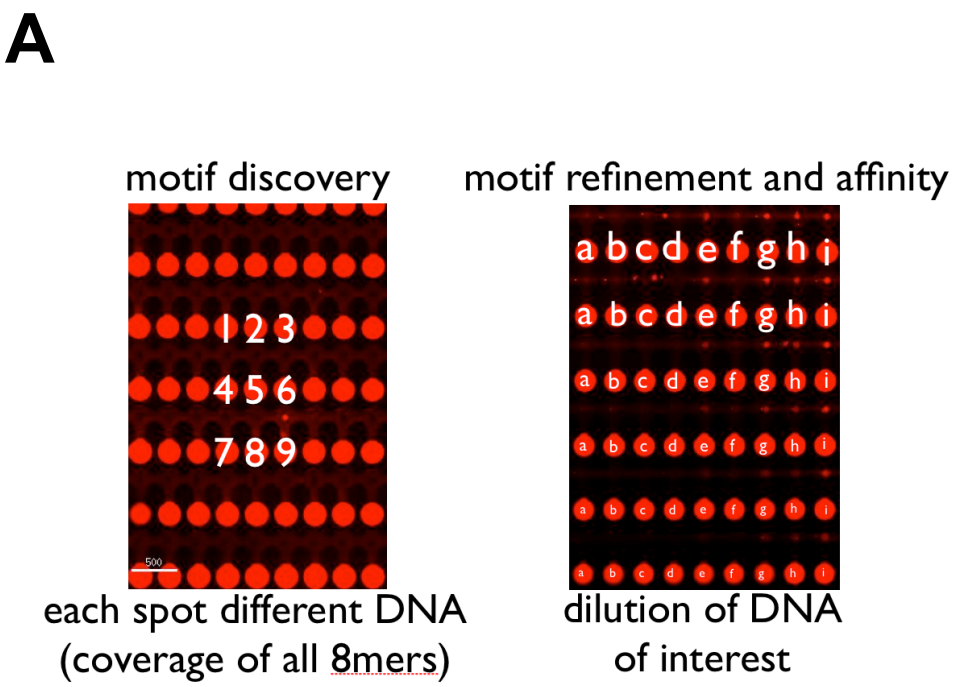


Figure 3. MITOMI2.0 and MITOMI modes. MITOMI devices can be built and programmed from two complementary purposes: motif discovery (MITOMI2.0 finding the preferred binding site of a factor with no prior knowledge) and motif refinement and affinity (MITOMI measuring the affinity changes to permutations

of a known binding site of interest). For example in MITOMI2.0 mode each of the spots contains a different sequence, and the factor only likes to bind B) Through measuring the fluorescent signal of bound DNA for each mutant DNA at the button valve after trapping we can generate affinity curves . The data in the plotted in black is from the cin5 protein. The blue line depicts a hypothetical shift in affinity. After fitting to simple hyperbolic binding curves, we can calculate the shift in the affinity, K_a.

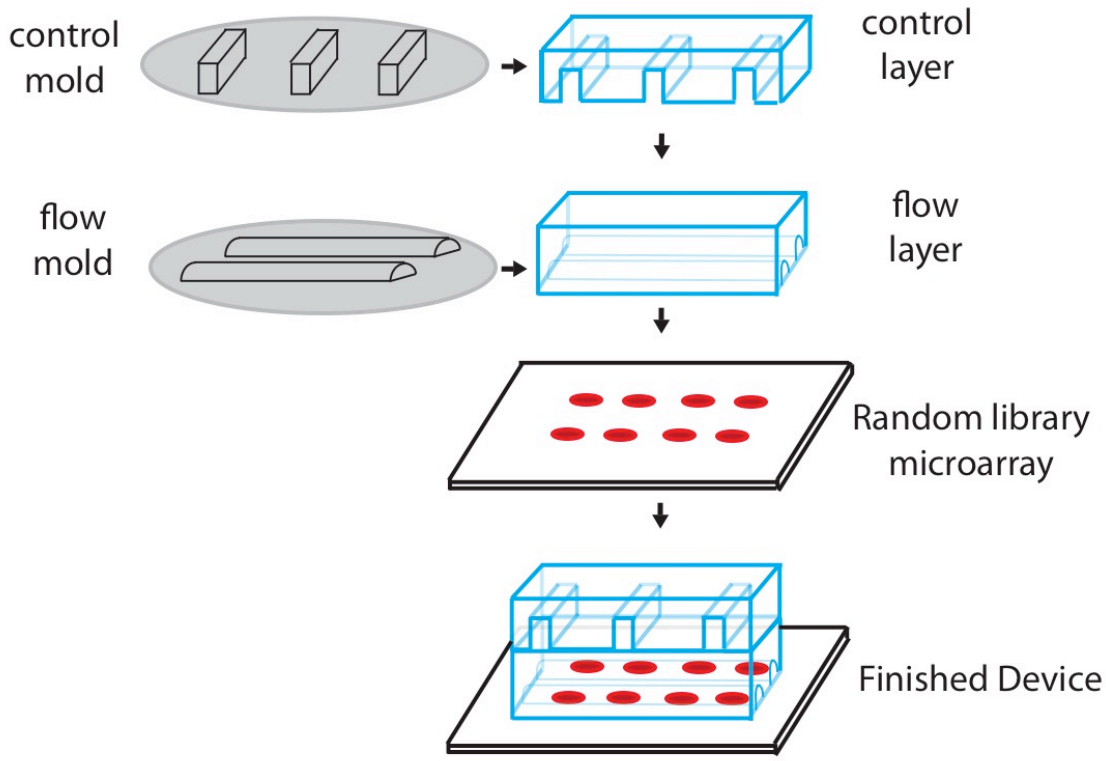


Figure 4. Schematic overview of the PDMS device fabrication process. Two layers of the device are made from independent molds on silicon wafers, These two layers are aligned to each other and then aligned to the array of DNA depicted in red.

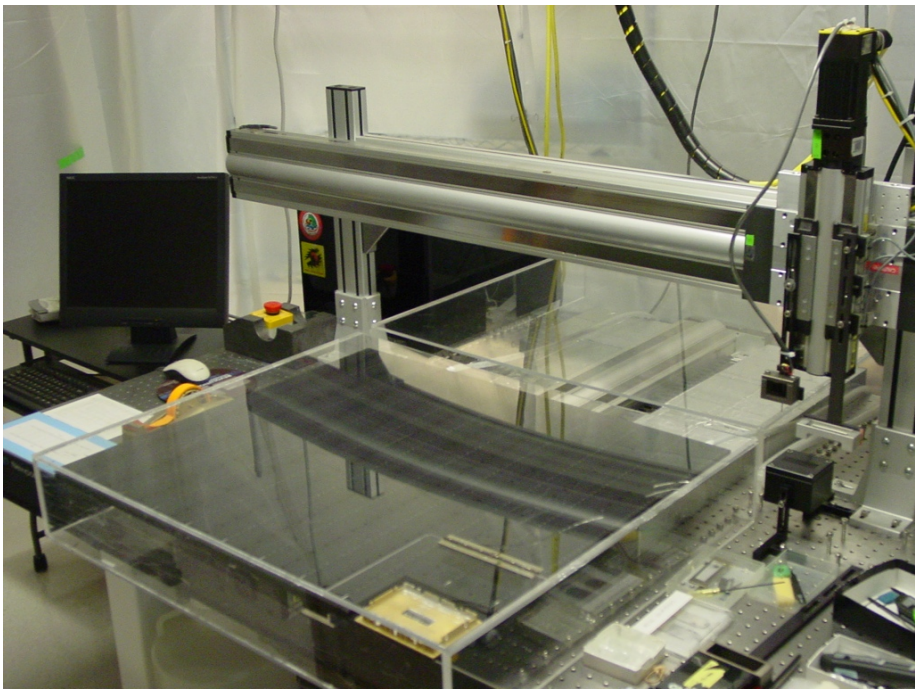


Figure 5. DeRisi-style contact microarray printer used for arraying DNA on glass substrates for MITOMI devices.

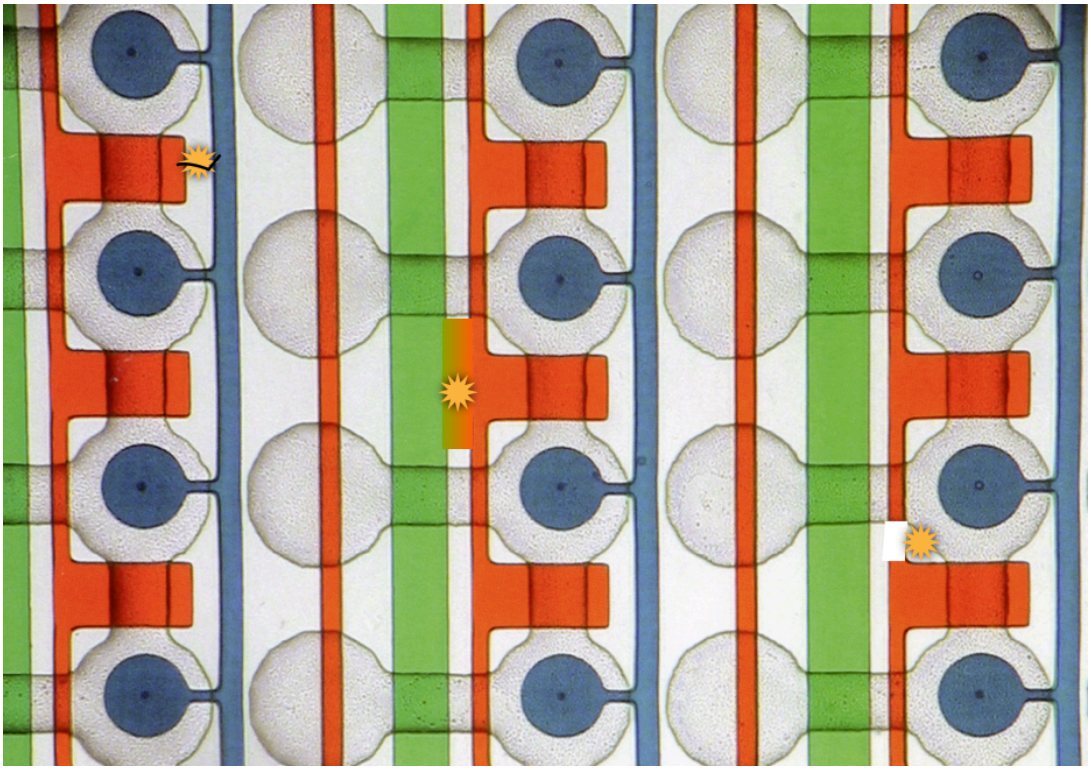


Figure 6. Common device failure modes in MITOMI devices. Most failure modes with MITOMI devices involve cross-talk between pressurized valves due to dust or missing features in the mold. To the left side is drawn a small fiber connecting. Any one of these small flaws could render the entire chip inoperable.

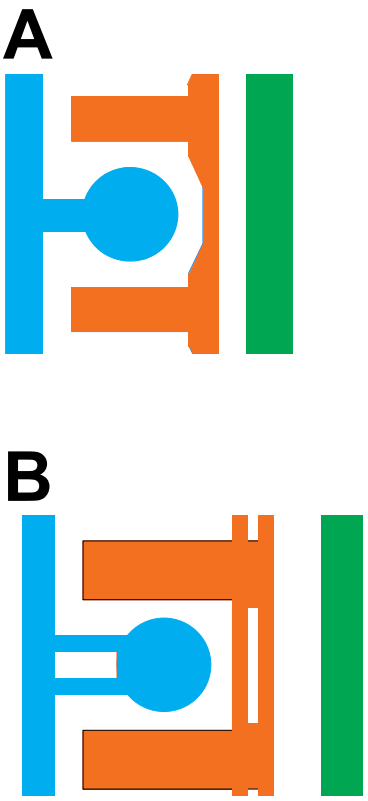
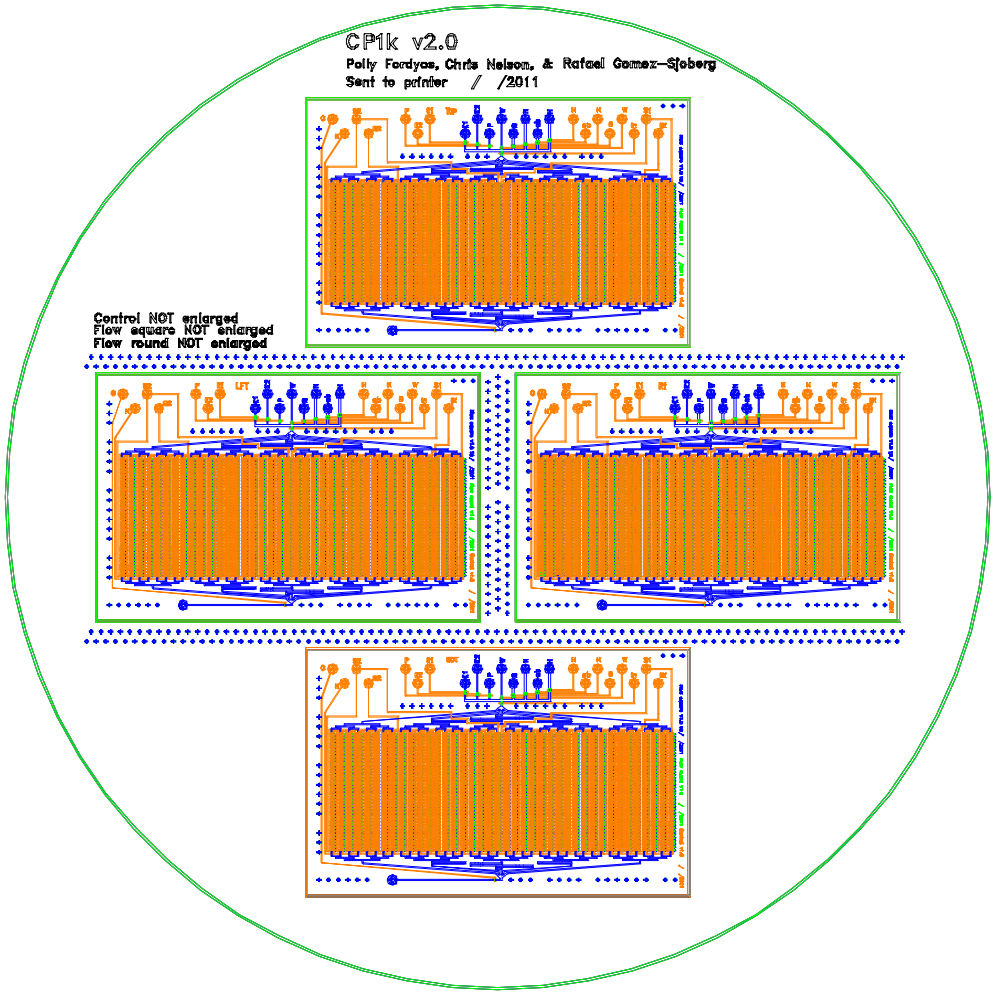


Figure 7. Unit cell schematic used between different versions of the control layers of the device (same scale). The colors are consistent from the prior schematics. Blue is the button valve, orange is the sandwich valve and green is the neck valve. Shown are different versions of the PF4K device unit cells, with evolving geometries and spacing. The spacing between the elements is meant to prevent the risk of a short circuit between the pressurized lines. The button valves grew larger to depress a larger area and more effectively. The extra cutout windows were meant to improve button and sandwich valve performance by reducing the flow pressure to fluid in these lines, while avoiding depressing non-valve areas or causing leaks of the control fluid into the flow layer. The sandwich valves grew longer and wider to improve closure performance by pressurizing a larger area and aiding alignment of the sandwich valves over the entire underlying channel.

A



B

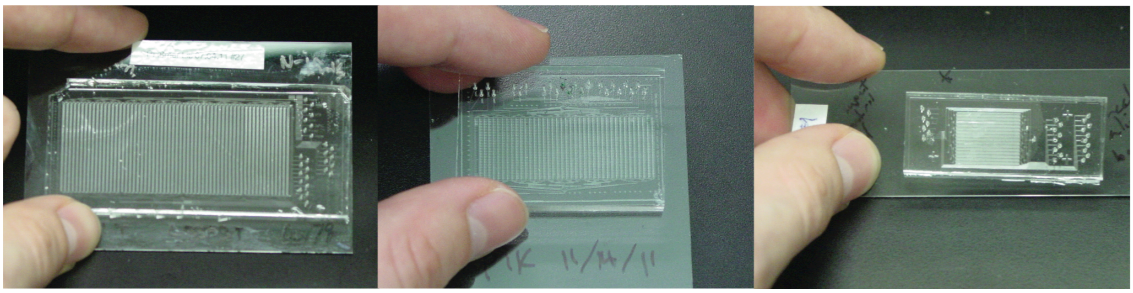


Figure 8. Example of device CAD design and finished devices of different layouts. A) Example CAD design for PC1K molds. The blue elements are the

flow channels with the input manifold on top and the output to the bottom left. The orange elements are all of the control valves. The input manifold is controlled by the cluster of 9 nearby control lines, with the button and sandwich valves split up into two circuits for the left and right half of the chip. Four devices PC1K devices will fit on the footprint of one 4-inch silicon wafer. B) Finished PDMS devices bonded to glass substrates. To the left is a PF4K 3900 chamber device, in the middle is a CP1K with 1520 unit cells and to the right is a DTPA-D 640 chamber device.

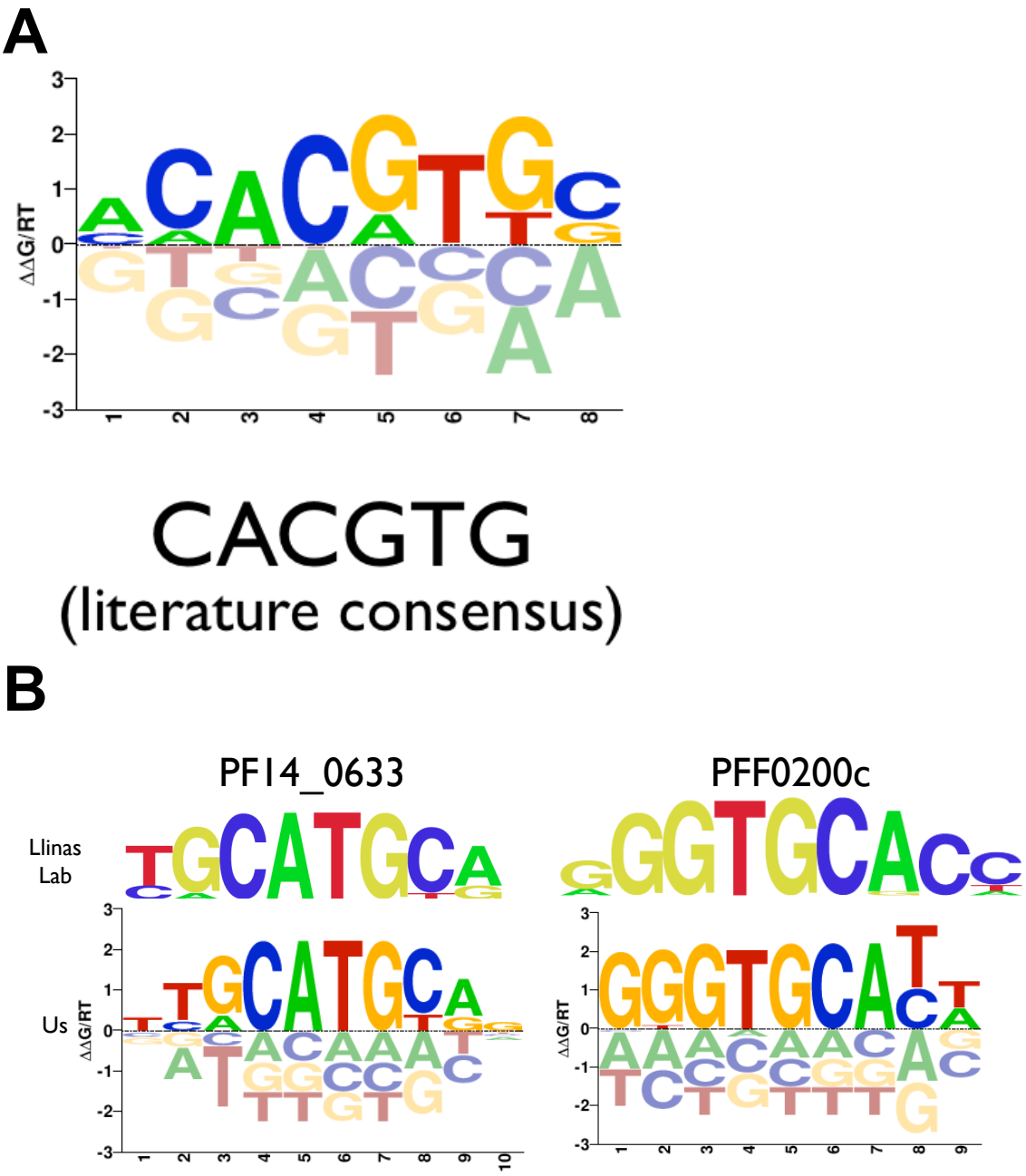


Figure 9. New library validation using proteins with known binding motifs. Positive control Pho4 protein gives the expected result with our new pseudorandom DNA library. Yeast transcription factor Pho4 gives the expected result with our new deBruijn-sequence-derived pseudorandom library. The motif logo for the Pho4 PSAM depicts the energetic effects of each mutation away from a reference sequence, with both positive and negative contributions at each

position in the motif sequence. Position in the motif is noted along the horizontal axis. B) Our MITOMI investigations of plasmodium AP2 domain proteins give strong signals and evidence motifs in agreement with published motifs.

Chapter 6……

The Basic Leucine Zipper Transcription Factor Hac1 Binds DNA In Two Distinct Modes as Revealed by Microfluidic Analyses

Polly M. Fordyce, David Pincus, Philipp Kimmig, Christopher S. Nelson, Hana El-Samad, Peter Walter, and Joseph L. DeRisi

Author contributions:

P.M.F., D.P., P.K, C.S.N., H. E.-S., P.W., and J.L.D designed research. P.M.F., D.P., P.K., and C.S.N. performed research, and P.M.F. analyzed data. P.M.F., D.P., P.K., C.S.N., H. E.-S., P.W., and J.L.D. wrote the paper.

Joseph L. DeRisi, Thesis Advisor

The following chapter is a reprint of the reference:

Fordyce PM, Pincus D, Kimmig P, Nelson CS, El-Samad H, Walter P, DeRisi JL. Basic leucine zipper transcription factor Hac1 binds DNA in two distinct modes as revealed by microfluidic analyses. Proc Natl Acad Sci U S A. 2012 Nov 6;109(45):E3084-93.

Abstract

A quantitative understanding of how transcription factors interact with genomic target sites is crucial for reconstructing transcriptional networks *in vivo*. Here, we use Hac1, a well-characterized basic leucine zipper (bZIP) transcription factor

involved in the Unfolded Protein Response (UPR), as a model to investigate interactions between bZIP transcription factors and their target sites. During the UPR, the accumulation of unfolded proteins leads to unconventional splicing and subsequent translation of *HAC1* mRNA, followed by transcription of UPR target genes. Initial candidate-based approaches identified a canonical *cis*-acting Unfolded Protein Response Element (UPRE-1) within target gene promoters; however, subsequent studies identified a large set of Hac1 target genes lacking this UPRE-1 and containing a different motif (UPRE-2). Using a combination of unbiased and directed microfluidic DNA binding assays, we established that Hac1 binds in two distinct modes: i) to short (6-7 bp) UPRE-2-like motifs, and ii) to significantly longer (11-13 bp) extended UPRE-1-like motifs. Using a library of Hac1 mutants, we demonstrate that a region of extended homology N-terminal to the basic DNA binding domain is required for this dual site recognition. These results establish Hac1 as the first bZIP transcription factor known to adopt more than one binding mode and unify previously conflicting and discrepant observations of Hac1 function into a cohesive model of UPR target gene activation. Our results also suggest that even structurally simple transcription factors can recognize multiple divergent target sites of very different lengths, potentially enriching their downstream target repertoire.

Introduction

The basic leucine zipper (bZIP) proteins form one of the largest families of eukaryotic transcription factors and play roles in a wide variety of biological phenomena, from responding to endoplasmic reticulum (ER) dysfunction to

regulating immune responses and oncogenesis[1]. Members of this superfamily contain a positively-charged DNA binding region composed of basic residues linked to a leucine zipper sequence, and homo- or hetero-dimerize via this leucine zipper. Invariant arginine and asparagine residues within the basic DNA binding region (NXAAXXCR) make direct contact with DNA bases within the major groove and drive binding specificity to palindromic or semi-palindromic target sites[2,3]. Although considered to be the simplest known protein-DNA recognition motif, crystal structures of bZIP domains bound to DNA have revealed functional variability in how these conserved residues contact DNA[2,4], and no universal code linking basic region sequence with target DNA preferences has been developed .

Here, we investigate the mechanisms that drive bZIP target site recognition using Hac1, a *Saccharomyces cerevisiae* transcription factor involved in the highly conserved Unfolded Protein Response (UPR). During the UPR, cells sense an accumulation of unfolded proteins within the endoplasmic reticulum (ER) and trigger a transcriptional upregulation of genes encoding ER-resident chaperones and protein modifying enzymes, components of ER-associated protein degradation (ERAD), and enzymes for phospholipid biosynthesis[5]. In *S. cerevisiae*, two main proteins are responsible for enacting the UPR: Ire1, a transmembrane kinase/endonuclease, and Hac1. Unfolded proteins bind to the Ire1 domain facing the ER lumen, triggering its oligomerization and activation of its cytoplasmic endonuclease domain. Once activated, Ire1 cleaves *HAC1* mRNA at two sites and tRNA ligase rejoins the severed exons via an

unconventional spliceosome-independent mechanism[5]. This splicing removes an intron to produce a new transcript (denoted *HAC1*^{i} mRNA; "i" for "induced"), thereby relieving translational inhibition exerted by the intron. Following translation of the spliced mRNA, Hac1^{i} is translocated to the nucleus, where it regulates a large set of UPR-responsive genes[6]. Despite the central role played by Hac1^{i} in activating the UPR, the rules by which Hac1^{i} recognizes UPR target genes remain unclear.

Initial studies took a candidate-based approach to identify potential Hac1^{i} binding sites within the promoters of known UPR target genes. Analysis of the promoter of *KAR2*, encoding the major Hsp70-type ER-resident chaperone Kar2 (or BiP), revealed a 22-bp *cis*-acting Unfolded Protein Response Element required for induction of UPR-dependent *KAR2* transcription (here referred to as "UPRE-1")[6]. Subsequent transcriptional activity assays identified a core 7-bp consensus (5'-*CAGNGTG-3'*; here referred to as "cUPRE-1"), in which point mutations of palindromic half sites (6 conserved bp) or changes in the half-site spacing severely reduced activity[7,8]. Gel shift assays demonstrated direct binding of Hac1^{i} to the 22-bp UPRE-1, and reporter gene assays confirmed that this element was sufficient to confer UPR-responsive transcriptional activity in an otherwise silent promoter[9]. UPRE-1-like motifs were also found in the promoters of 4 additional known UPR target genes (*PDI1*, *EUG1*, *FKB2*, and *LHS1*), lending support to its proposed role[9,10].

This central role for UPRE-1 in upregulating target gene transcription was subsequently called into question by a study employing genome-wide microarray

expression profiling to identify all candidate UPR target genes[13]. This work identified 381 candidate target genes, representing nearly 5% of all open reading frames in the *S. cerevisiae* genome and encoding numerous proteins required in the ER, the Golgi apparatus, and throughout the secretory pathway. Bioinformatic analysis of the promoter regions of these genes revealed that although most lacked the canonical UPRE-1, many contained one or more of two alternate motifs ("UPRE-2", 5'-*TACGTG*-3'; "UPRE-3", 5'-*AGGACAAC*-3') capable of driving Hac1^{i}-mediated transcription in reporter assays. Surprisingly, this analysis failed to recover the known UPRE-1 site[14]. To account for the target site variety, it was proposed that Hac1^{i} bound to these alternate sites via heterodimerization with Gcn4. Further complicating the picture, a study using protein binding microarrays (PBMs) to probe Hac1^{i} binding preferences among all possible 8-bp nucleotide sequences revealed binding only to UPRE-2[15].

In vivo studies of Hac1^{i} are complicated by both the very short half-life of the Hac1^{i} isoform derived from the spliced mRNA and the tendency of bZIP transcription factors to homo- and heterodimerize. Therefore by necessity, *in vitro* approaches provide a particularly valuable tool for accurately defining binding preferences. Here, we probe how Hac1^{i} regulates expression of target genes using microfluidic affinity analysis (MITOMI[16] and MITOMI 2.0[17]) to identify and characterize Hac1^{i} target sites. In addition, we analyze expression of reporter genes driven by a variety of Hac1^{i} mutants to identify the protein residues required for target site recognition. Our results resolve the previous conundrum regarding Hac1^{i} binding behavior to provide an integrated model of

UPR target gene regulation, and shed new light on the basic mechanisms by which bZIP transcription factors recognize their target genes.

Results

Experimental set-up for measuring Hac1^{i} binding

To obtain an unbiased assessment of Hac1^{i} binding preferences, we used a microfluidic platform, MITOMI 2.0[18], to measure relative binding affinities (DDG) between Hac1^{i} and 70-bp double-stranded oligonucleotides containing overlapping instances of all possible 8-bp combinations (Fig. 1A). In previous work, we validated this platform using a panel of 28 *S. cerevisiae* transcription factors and demonstrated the ability to quantitatively measure relative binding affinities to each oligonucleotide and recover known binding preferences[18].

In brief, each MITOMI 2.0 device contained 4,160 chambers composed of two compartments ("DNA" and "protein") controlled by 3 valves ("neck", "sandwich", and "button") (Fig. 1B). Experiments took place in six main steps (Fig. 1C): (1) DNA compartments were programmed with specific Cy5-labeled double-stranded DNA sequences by aligning devices to a spotted DNA microarray; (2) BODIPY-FL-labeled His-tagged Hac1^{i} was flowed across the protein compartments and recruited to surfaces beneath button valves that were coated with anti-His antibodies; (3) protein solution was pushed into DNA compartments, solubilizing spotted DNA and allowing Hac1^{i} and DNA sequences to interact; (4) binding interactions were mechanically trapped at equilibrium by pressurizing button valves to squeeze out unbound material; (5) neck valves were closed to isolate the compartments and allow washing away of unbound material in the protein

compartment while preserving equilibrium concentrations of both binding partners in the DNA compartment; and (6) devices were read using a fluorescence scanner. Final Cy5 intensities in each DNA chamber were previously shown to be proportional to the soluble DNA concentration available for binding[16,17], and the ratio of Cy5 (DNA) to BODIPY-FL ($Hac1^i$) intensities beneath the button valve reports the protein fractional occupancy, allowing calculation of interaction K_d and DDG (Fig. 1C).

We measured $Hac1^i$ binding in two independent experiments. In both cases, $Hac1^i$ showed strong preferences for particular sequences (Figs. 1D arrow and 1E; Fig. S1A), with Z-scores of ~ 85 for the highest affinity sequences (Fig. 1E; Fig. S1A). Measurements were fairly reproducible both between replicates within a given experiment (Pearson r^2 = 0.73 and r^2 = 0.77, Fig. S1B) and between experiments performed on different days (Pearson r^2 = 0.51, Fig. S1C); therefore, we pooled results from both experiments for further analysis.

MITOMI 2.0 analysis predicts $Hac1^i$ binding primarily to UPRE-2

Each 70-bp oligonucleotide contained multiple potential overlapping $Hac1^i$ binding sites (Fig. 1A); consequently, additional analysis was required to deconvolve results and identify the target sites responsible for $Hac1^i$ binding. First, we used fREDUCE[18] to search for 6-, 7-, and 8-bp motifs whose appearance within oligonucleotides correlated most strongly with their measured intensity ratios. Surprisingly, all searches exclusively returned variants of UPRE-2, with strong correlations between the appearance of this motif and observed intensity values (Fig. 2A, Fig. S2, Table S1). We then assessed the effects of

single nucleotide substitutions in this consensus site on DDG by using MatrixREDUCE[19,20] to generate a position-specific affinity matrix (PSAM). Importantly, PSAMs can be used to predict binding to any sequence quantitatively, and comparisons between predicted binding profiles and measured binding profiles yield additional information: in particular, oligonucleotides bound more strongly than predicted would indicate binding to additional motifs, while oligonucleotides bound more weakly would indicate motifs that repel binding. In our data, comparisons between predicted and measured binding showed strong agreement, suggesting that Hac1^{i} bound non-promiscuously to UPRE-2 *in vitro* and displaying no evidence for binding to additional sequences present in the oligonucleotide library (Fig. 2A, Fig. S2).

Hac1^{i} binds the UPRE-2 but not the cUPRE-1

It poses a paradox that our microfluidic affinity assay data and previous PBM experiments have failed to uncover evidence of UPRE-1 binding, which was well validated in previous studies[6,8,9]. This failure could be explained because either Hac1^{i} does not bind to the cUPRE-1 but requires a longer sequence that is not represented in our library, or by an insufficient sensitivity of the MITOMI 2.0 assay to pick up low-affinity cUPRE-1 binding.

To distinguish between these possibilities, we directly measured concentration-dependent binding of Hac1^{i} to a series of oligonucleotides containing either the cUPRE-1 or the UPRE-2 embedded within random sequence (Fig. 2B). In three separate experiments, we observed high-affinity binding to the oligonucleotide containing the UPRE-2, with no measurable binding above background levels to

the oligonucleotide containing the cUPRE-1 (Fig. 2B). Fits to the UPRE-2 binding data yielded a K_d of 427 ± 37 nM, similar to values we previously obtained for other bZIP transcription factors[17].

Reporter assays have suggested that the central *C* within the cUPRE-1 can be replaced by alternate nucleotides with only a slight reduction in activity[8]. We therefore assessed binding to these variants to see if any of the variations restore binding. For all variants, binding remained at the level of random sequence over multiple experimental replicates (Fig. S3).

Hac1^i binding to UPRE-1 requires an extended target site

Previous work showed that the cUPRE-1 is necessary for transcriptional activity[9]. However, it was never shown to be sufficient, and phylogenetic comparisons suggest that cUPRE-1 flanking sequences are important for Hac1^i binding. UPRE-1 sites from the promoters of multiple known Hac1^i targets (*KAR2*, *EUG1*, *PDI1*, *FKB2*, and *LHS1*) show conservation of several nucleotides both up- and downstream from the 7-bp core, even as the core is imperfectly conserved[10–12] (Fig. 3A). The same pattern is also seen for UPRE-1 sites within the promoters of *KAR2* orthologs from distant species (Fig. 3B).

To test whether flanking sequences are critical for Hac1^i binding, we measured concentration-dependent binding for the cUPRE-1 embedded within either a fragment of the *KAR2* promoter or within the *ERO1* promoter, which typically contains a UPRE-2-like motif (Fig. 3C). Consistent with the notion that cUPRE-1 flanking sequences are required, addition of the *KAR2* flanking sequences to cUPRE-1 restored high-affinity binding (Fig. 3C). Insertion of the cUPRE-1 into a

heterologous flanking context (the *ERO1* promoter) did not restore Hac1^{i} binding (Fig. 3C), establishing that UPRE-1-specific flanking sequences are required.

To identify the precise boundaries of flanking sequences required for Hac1^{i} binding to the cUPRE-1, we started with the cUPRE-1 embedded within heterologous *ERO1* flanking sequences and systematically substituted these sequences with increasing portions of *KAR2* sequences in the upstream and/or downstream direction (Fig. 3D). Restoration of upstream or downstream flanking sequences alone did not significantly increase Hac1^{i} binding affinity to cUPRE-1 (Fig. 3D, top and middle rows, respectively). By contrast, simultaneous addition of both up- and downstream *KAR2* flanking sequences had a strong effect on binding (Fig. 3D, bottom row).

Addition of one nucleotide on either side of the cUPRE-1 increased affinity 5-fold, and addition of two nucleotides on either side of the core restored affinity to that measured for the cUPRE-1 in its native *KAR2* context (Fig. 3E). Inclusion of additional *KAR2* flanking sequence did not significantly alter binding affinities, suggesting that the 12-bp sequence 5'-*GGACAGCGTGTC*-3' (hereafter termed the extended core UPRE-1, or xcUPRE-1) is sufficient for Hac1^{i} binding. Further corroboration of the importance of these nucleotides comes from the observation that single point mutations 2-bp up- and downstream from the cUPRE-1 in the *KAR2* promoter previously caused a reduction in transcriptional activity[6,8].

Systematic mutation of xcUPRE-1 and UPRE-2 target motifs reveals two distinct target motifs

Understanding precisely how xcUPRE-1 and UPRE-2 differ requires a comprehensive map of individual nucleotide preferences for each binding mode. To create such a map, we measured concentration-dependent binding curves for systematic substitutions of all possible nucleotides at each position within both targets (Fig. 4A; Table S2). In each case, we performed 3-4 experimental replicates (Fig. S4-S7) and computed the average relative binding affinity for each substitution (Fig. 4A). From these relative affinities, we then computed an average PSAM (Fig. 4B; Tables S3, S4; Fig. S8) for each motif.

The relative nucleotide preferences for UPRE-2 derived from these measurements agree well with those from our MITOMI 2.0 analysis (compare Fig. 4B and Fig. 2A). Taken together, these results establish that the complete UPRE-2 is short, subtending 6-7 nucleotides, with little degeneracy tolerated at most positions. The UPRE-2 appears to be an imperfect palindrome: attempts to create a more fully symmetric site by either adding a 5' C (Fig. 4A) or by altering multiple nucleotides (Fig. S9) do not lead to statistically significant increases in binding affinity. By contrast, mutations at nearly all positions within xcUPRE-1 have significant effects on affinity, further confirming that xcUPRE-1 subtends on the order of 11-12 bp (Fig. 4A,B). In addition, the overall composition of the motif is different: xcUPRE-1 appears to be composed of two palindromic dyad repeats (*5'-G[A/C]CAC-3'*) separated by a central degenerate nucleotide.

The absolute affinity for UPRE-2 in these experiments was slightly higher than the absolute affinity for xcUPRE-1 (UPRE-2 K_d = 497 ± 60 μM; xcUPRE-1 K_d = 720 ± 80 μM) (Table S5). However, the range of affinities measured for all

UPRE-2 and xcUPRE-1 variants largely agree, with the strongest binding measured to be 220 ± 30 μM and 360 ± 40 μM, respectively (Table S5).

Prediction of potential genomic targets using xcUPRE-1 and UPRE-2 PSAMs

An advantage of PSAMs over position weight matrices (PWMs) is that they allow *de novo* prediction of protein binding affinities to arbitrary sequences. To test the performance of our xcUPRE-1 and UPRE-2 PSAMs, we compared measured and predicted Hac1^{i} binding affinities for a variety of genomic UPREs, including those present within the *ERO1*, *KAR2*, *EUG1*, *LHS1*, *FKB2*, *SEC66*, and *PDI1* promoters (Fig. S10). Measured and predicted affinities showed relatively strong agreement (r^2 = 0.64, p = 0.03), confirming our ability to accurately predict binding to novel sequences. Next, we calculated predicted binding affinities to all annotated promoters within the yeast genome, as well as to known UPR target genes[13] (files available for download as Supplementary Information). Promoters predicted to be bound with high affinity via UPRE-2-like binding were more likely to be present within the original UPR-induced data set[14](Table S6), with the top 20 hits including multiple known UPR targets (*ULI1*, *TRA1*, *SFB3*, *MCD4*, *SNF11*, *HNT1*, and *KIC1*).

To test the ability of binding affinities measured *in vitro* to predict *in vivo* transcriptional response, we compared levels of expression of green fluorescent protein (GFP) in two *S. cerevisiae* strains following addition of dithiothreitol, which impairs the formation of disulfide bonds and leads to induction of the UPR. In one strain, GFP expression was driven by a synthetic promoter containing 4

repeats of the full *KAR2* UPRE-1[14]; in the second strain, cUPRE-1 sequences were replaced by UPRE-2 sequences, resulting in higher measured *in vitro* affinities (Fig. S11). In both strains, basal GFP expression was low and identical. Following induction of the UPR, GFP levels in the strain containing the UPRE-2 substitutions were ~2-fold higher (Fig. S11), establishing that changes in affinity measured here predict target promoter activity *in vivo*.

A region of extended homology N-terminal to basic region is required for dual site recognition

Given that xcUPRE-1 and UPRE-2 differ significantly in both their overall length and relative nucleotide compositions (Fig. 4A), recognition of each motif must accommodate distinct arrangements of contacts between Hac1^{i} and target site nucleotides. If this is the case, it should be possible to create Hac1^{i} mutants that disrupt binding to one site while largely preserving binding to the other. In the Maf subfamily of bZIP transcription factors, a region of extended homology positioned N-terminal to the basic DNA binding domain is critical for recognition of extended (13-14 nucleotide) target sites[4]. Phylogenetic alignment of Hac1 orthologs across ascomycetes reveals a similar region of extended homology (Fig. S12), suggesting that these residues may be important for DNA specificity. To identify mutants with altered binding preferences, we used a genetic screen with to assess levels of binding via each mode. To do so, we used error-prone PCR to generate a library of Hac1^{i} constructs containing random mutations to the protein between the N-terminus and the first heptad repeat of the leucine zipper (Fig. 5A, Fig. S13). We transformed this library into a yeast strain containing two

synthetic promoters controlling the expression of two fluorescent proteins. The first promoter consists of 4 repeats of the *KAR2* UPRE-1 motif driving mApple expression, and the second consists of 4 repeats of the UPRE-2 driving GFP expression (Fig. 5B). To generate the 4x-UPRE-2 promoter, we mutated two nucleotides within the *KAR2* UPRE-1 to create a UPRE-2 target site (Fig. 5B). Importantly, the PSAMs derived here (Tables S3,4; Fig. 4B) predict that these mutations are sufficient to switch Hac1^{i} binding towards the UPRE-2 recognition mode (Fig. S14). To ensure that differences in fluorescence intensity were due to changes in Hac1^{i} binding and not indirect effects from other UPR components, we ectopically expressed both wild-type Hac1^{i} and this mutant library using an estradiol-inducible system[23] (Fig. S15).

Using this approach, we identified multiple Hac1^{i} mutants with altered levels of binding to either one or both target promoters relative to wild-type constructs (Fig. 5C, Fig. S16, Fig. S17). Reassuringly, constructs sharing a given mutation displayed the same fluorescence phenotype (Fig. S17; Tables S7, S8). Most constructs with altered binding appeared to retain the ability to recognize xcUPRE-1 even as UPRE-2 recognition was impaired; this tendency could reflect the fact the xcUPRE-1 recognition appears to take place via both binding modes (Fig. S14), or could simply be due to the increased length and tolerance of degeneracy within xcUPRE-1 (Fig. 4A,B). Mutations in positively-charged arginines or lysines within the extended homology region or near the N-terminus of the basic DNA binding region preferentially reduced UPRE-2 binding while maintaining xcUPRE-1 binding (Fig. 5D). Interestingly, a single arginine within

the basic region plays a crucial role in xcUPRE-1 recognition (Fig. 5D). The diversity of these binding phenotypes and their emergence from individual mutations strongly argues that Hac1^{i} binds DNA via distinct binding modes, with individual protein residues playing different roles within each interaction.

An N-terminal truncation mutant lacking extended homology regions binds UPRE-2-like sequences with reduced affinity

To further probe this idea and test the notion that residues within the Hac1^{i} extended homology region are required for UPRE-2 recognition, we created Hac1^{i} constructs with truncations at different locations within the extended homology region and mapped their xcUPRE-1 and UPRE-2 binding preferences using microfluidic affinity analysis. One truncation mutant (N25) retained 3 residues identified as being important for UPRE-2 binding, while the other truncation mutant (N35) lost these residues (Fig. 6A). Although comparisons between relative binding affinities for nearly full-length (N10) Hac1^{i} and the N25 truncation mutant showed strong agreement (r^2 = 0.90; Fig. 6B, left), similar comparisons between nearly full-length (N10) Hac1^{i} and the N35 truncation mutant showed much weaker agreement (r^2 = 0.40; Fig. 6B, middle), suggesting a change in binding preferences. Calculation of the difference in binding preference relative to average xcUPRE-1 and UPRE-2 behaviors for each construct reveals that although all constructs show similar binding preferences for oligonucleotides containing xcUPRE-1 and single-site substitutions (Fig. 6C), the N35 construct shows dramatically reduced binding for UPRE-2 and single-site substitutions (Fig. 6D). In particular, the N35 UPRE-2 PSAM shows a

decreased tolerance for nucleotide substitutions at the 5' end of the motif, suggesting a shift towards a more extended binding site (Fig. 6D; Fig. S18). The N25 and N35 truncation mutants showed 2-fold and 10-fold decreases in overall binding affinities, respectively (Fig. S19). As a result, mapping N35 binding preferences required that experiments be performed at 4-fold higher DNA concentrations to accurately measure affinities. Comparisons between relative affinities measured for the N10 construct at both concentrations showed good agreement (r^2 = 0.76; Fig. 6B, right), signifying that changes in binding preferences do not result merely from changes in experimental conditions. These results lend additional support to the idea that residues within the extended homology region are required for dual-mode binding of Hac1^i to target sites.

Discussion

Here, we show that Hac1^i binds two divergent DNA binding sites, a compact 6- or 7-bp UPRE-2 site and a significantly longer 11- or 12-bp xcUPRE-1 site. While the compact UPRE-2 appears to be a slightly asymmetric half-site, the xcUPRE- consists of two palindromic *5'-G[A/C]CAC-3'* dyad repeats separated by a central bp that is relatively degenerate (Fig. 4A), with mutations at this position having little effect on transcriptional activity[11]. These differences in both site length and nucleotide composition suggest that Hac1^i must contact each site via distinct modes of binding. In support of this hypothesis, mutational analysis reveals that a region of extended homology N-terminal to the basic DNA binding domain is

required for Hac1^{i} dual site recognition, and microfluidic affinity analysis confirms the importance of these residues for UPRE-2 recognition. Based on these conclusions, Hac1^{i} emerges as the first natural bZIP transcription factor shown to operate in at least two different modes.

The idea that the xcUPRE-1 subtends 11-12 bp is supported by multiple previous observations. Although shown to be necessary for UPR-responsive transcriptional activation, the 7-bp cUPRE-1 was not shown to be sufficient for activation, and mutations in flanking nucleotides outside of this core motif caused severe reductions in reporter assay activity[6,8]. Such a long recognition sequence may also explain the prior failure of short word-based bioinformatic analysis of promoters to recover this motif from known UPR target genes[14]. Our results are therefore consistent with previous observations and clarify our understanding of Hac1^{i} function.

Several arguments suggest that the binding observed here reflects the behavior of Hac1^{i} alone and not of Hac1^{i} heterodimers, as previously proposed[15]. In our experiments, 6x-His tagged Hac1^{i} was produced in an *in vitro* translation system that was then flowed over a surface coated with anti-His antibodies, effectively concentrating and purifying Hac1^{i} on-chip prior to affinity measurements. We consider it likely that Hac1^{i} produced in this manner exists as an equilibrium of monomeric and homodimeric species. In addition, the shapes of the concentration-dependent binding curves suggest that both xcUPRE-1 and UPRE-2 motifs are bound by Hac1^{i} complexes with identical stoichiometries: All curves in a given experiment asymptote to an identical fluorescence intensity

ratio, establishing that the number of DNA molecules bound per labeled protein molecule remains constant. The notion that Hac1^{i} binds as a homodimer is further supported by detection of UPRE-2 binding in PBM experiments employing Hac1^{i} proteins expressed in an *E. coli* system that does not contain potential orthologous binding partners[15,22]. Moreover, gel shift assays performed in yeast extracts showed indistinguishable shifts for Hac1^{i} bound to oligonucleotides containing either xcUPRE-1 or UPRE-2 motifs[14]. Finally, the palindromic structure of the xcUPRE-1 target site is consistent with expectations of homodimeric binding.

How does Hac1^{i} bind both long xcUPRE-1 and compact UPRE-2 DNA target sites? Although most bZIP transcription factors are thought to bind relatively compact binding sites (Fig. S20), Maf subfamily transcription factors recognize unusually long motifs (13-14 bp) via an unconventional conformation of the invariant arginine and asparagine residues within the basic region of all bZIP proteins[4]. Similarly, a crystal structure of Pap1, a *S. pombe* bZIP transcription factor, complexed with DNA demonstrated that Pap1 target site specificity was also due to alternate positioning of these two residues[2]. In our genetic screen, none of the constructs with altered binding affinities were found to have mutations in these invariant residues (Fig. 5D; Fig. S17; Table S8), although such mutations were found in colonies lacking fluorescence in either channel (Table S7). We therefore suggest that these invariant residues could be required for recognition of both target sites, with changes in their conformation leading to recognition of one site or the other. Notably, the ability to recognize two closely

related sites via conformational shifts has previously been proposed for a minimal bZIP construct[25]. In a manner analogous to the Maf proteins, we propose that the extended homology region could stabilize invariant bZIP residues in the conformation required for UPRE-2 recognition. With the exception of MafG, most bZIP crystal structures have been based on constructs truncated to include only 1-9 nucleotides N-terminal to the basic DNA binding region[2,24–27] (Fig. S21). It remains to be seen whether N-terminal regions of extended homology facilitate binding of alternate sites by other bZIP proteins. Several recent studies have noted plasticity in bZIP binding preferences, although to date, this plasticity has been confined to tolerance for binding multiple related sites of the same or very similar lengths. A synthetic bZIP protein composed of the Gcn4 basic region fused to the C/EBP leucine zipper was shown to bind with high affinity to both cognate and alternate sites, indicating that protein architecture beyond the basic region can affect binding preferences[28]. In addition, a recent study employing PBMs to characterize the DNA binding specificities of multiple bZIP TFs noted that several proteins (Yap1, Yap3, and Sko1) possessed the ability to bind closely related dyad repeat sites with variable length (1-2 bp) spacers or extensions at either end.[22]. This study also noted that the DNA binding domain for Hac1^{i} shares multiple residues with the basic regions of bHLH proteins, possibly explaining the similarity between UPRE-2 and bHLH E-box target sites.

Our results have broad implications for experimental approaches to characterize transcription factor binding preferences. Many current methods (*e.g.* PBMs[15,29–

[31] and HiTS-FLIPS[21]) entail measurements of binding interactions, which are interpreted to yield relative preferences for particular nMer (usually 8-mer) sequences. In these analyses, the strength of binding for a given oligonucleotide is assumed to result solely from the presence or absence of a particular nMer sequence, without consideration for sequence context. Here, we show that even transcription factors thought to be simple in their DNA binding properties can bind multiple target sites of very different lengths, complicating attempts to identify target sites via ranked nMer preferences. The MITOMI 2.0 technique presented here attempts to avoid these pitfalls by applying a statistical mechanical model to extract binding preferences. It is important to note, however, that the size of the library that can be accommodated within the devices is still too small for simultaneous detection of binding to both Hac1^{i} motifs. The experiments presented here were guided by our previous knowledge of Hac1^{i} regulatory targets, underscoring the importance of integrating both biophysical and biological data to understand transcriptional regulation.

To what purpose does Hac1^{i} recognize multiple distinct sites? For the glucocorticoid receptor, DNA sequences can act as allosteric ligands, inducing conformational changes to preferentially recruit specific cellular co-factors with functional consequences for transcriptional activation[32]. A similar scenario may apply to Hac1^{i}, and perhaps to other bZIP family members, although additional studies will be required to determine whether changes in protein conformation within the DNA binding domain can propagate elsewhere within the protein. Alternatively, dual site recognition could represent a snapshot in evolutionary

time of a transcriptional network rewiring event in progress. According to this notion, it may have been advantageous to place an additional set of target genes under Hac1^{i} control, perhaps as a hand-off of some other transcriptional program. In this light, it is interesting to note that the Hac1^{i}-driven transcription program in *S. cerevisiae* has been split into multiple transcriptional branches in metazoans, indicating evolutionary network plasticity.

Materials and Methods

DNA library and Hac1^{i} production

DNA libraries for MITOMI 2.0 experiments were synthesized as described previously[17]. Briefly, all possible 65,536 8-bp DNA sequences were assembled into a compact DNA library spread over 1457 oligonucleotides, each of which contained an identical 3 nt 5' *'CGC'* clamp and an identical 14 nt 3' universal sequence allowing hybridization of a single Cy5-labeled primer (Fig. 1A). Following hybridization, all sequences were extended using Klenow exo- (New England Biolabs). Prior to printing, libraries were dried down and resuspended to a final concentration of 1.25 μM in 3X SSC containing polyethylene glycol (PEG) (Fluka) and D-(+)-trehalose dihydrate (Fluka) to improve spot visibility and solubilization. Libraries were printed on custom 2"x3" Scientific SuperChip Epoxysilane (Thermofisher) slides.

Linear expression templates for Hac1 proteins were created via a 2-step PCR amplification reaction, as described previously[18]. In the first PCR reaction, gene-specific primers were used to amplify gene templates and add an upstream Kozak sequence and a C-terminal 6xHis tag. In the second PCR reaction,

universal primers were used to add an upstream T7 promoter, 3' poly-A tail, and downstream T7 terminator. PCR-generated templates were added to TNT T7 Quick Coupled *in vitro* transcription translation kits (rabbit reticulocyte lysates, Promega) according to the manufacturer's instructions in the presence of BODIPY-labeled lysine tRNA (Fluorotect Green, Promega) to produce fluorescently labeled protein.

Microfluidic affinity assays

Photolithography molds and microfluidic devices were produced as described previously[16,17]. Microfluidic affinity assays and data analysis were also performed largely as described previously[17,18], with one modification. Cy5 intensities of printed slides can decrease rapidly with time, rendering calibration curves linking intensities with DNA concentration inaccurate. To sidestep this issue, we measured a single calibration curve within one day of an experiment assessing concentration-dependent binding behavior for Hac1^{i} interacting with UPRE-2 variants. The K_d values for these interactions were then used to determine the appropriate conversion between intensity and DNA concentration in the experiments shown in Figures 2 and 3. All data from Figures 4-6 were measured and calibrated independently. Average xcUPRE-1 and UPRE-2 PSAMs[21] were calculated by: (1) determining binding preferences (K_a) for wild-type and all single-substitution oligonucleotides relative to the most strongly bound xcUPRE-1-like or UPRE-2-like oligonucleotide, and (2) computing the average relative affinity over all replicates. In each case, differences in binding preferences for Hac1^{i} truncation mutants are calculated by subtracting relative

binding affinities (calculated relative to the most strongly bound oligonucleotides) from the average behavior.

Error-prone PCR

The Hac1^{i} mutant library was created using error-prone PCR[35], using a total of 48 cycles and 11 serial dilution steps (performed every 4 cycles).

Yeast strains and plasmids

Standard cloning and yeast techniques were used for construction, transformation and integration of the plasmid within strain W303. The transcription reporters used here controlled expression of the fluorescent proteins mApple and GFP via a crippled *cyc1* promoter containing 4 repeats of a 22-bp UPR-responsive *cis* element (xcUPRE-1 for mApple or mUPRE-1 for GFP). The UPRE-2 reporter was generated by site direct mutagenesis of 4xUPRE-1 (Fig. 5). The xcUPRE-1-mApple was cloned into a single integration, *HIS3*-marked vector (pNH603), while the mUPRE-1-GFP was cloned into a single integration, *LEU2*-marked vector (pNH605).

Flow cytometry

A dual reporter strain containing xcUPRE-1-mApple (integrated in *his3*) and mUPRE-1-GFP (integrated in *leu2*) also expressed a chimeric estradiol-responsive transcriptional activator with an N-terminal activation domain derived from Msn2 and a C-terminal DNA binding domain from Gal4. This parent strain was then transformed with either wild type *HAC1^{i}* or the mutant *hac1^{i}* library (cloned into a single integration, *TRP1*-marked vector under the control of the *GAL1* promoter). Cells were cultured in 2x SDC at 30°C in 96 well plates (2 ml) in

an Innova plate shaker at 900 rpm. After induction with estradiol (100 nM), cells were sampled after 4 hours using a BD LSR-II equipped with a high throughput sampler, a 488 nm 150 mW laser, 532 nM 150 mW laser, FITC and PE-Texas red emission filters, and FACS DIVA software. Flow cytometry data were analyzed using custom software written in Python. Reported mean fluorescence intensities for mutant strains were calculated via Gaussian fits to binned intensity distributions for individual cells.

Conflict of Interest Statement

The authors declare no conflict of interest.

Acknowledgements

We thank Doron Gerber, Danh Tran, and Stephen Quake for assistance with fabrication of microfluidic devices and early microfluidic assays, Marshall Burke for photographs of microfluidic devices, and Matthew Larson and Florencia Caro for careful reading of the manuscript. P.M.F was supported by a Howard Hughes Medical Institute/Helen Hay Whitney Foundation Postdoctoral Fellowship. J.D.R. and P.W. are Investigators of the Howard Hughes Medical Institute.

References

1. Asada, R., Kanemoto, S., Kondo, S., Saito, A. & Imaizumi, K. The signalling from endoplasmic reticulum-resident bZIP transcription factors involved in diverse cellular physiology. *Journal of Biochemistry* **149**, 507 –518 (2011).

2. Fujii, Y., Shimizu, T., Toda, T., Yanagida, M. & Hakoshima, T. Structural basis for the diversity of DNA recognition by bZIP transcription factors. *Nat Struct Mol Biol* **7**, 889–893 (2000).

3. Miller, M. The importance of being flexible: the case of basic region leucine zipper transcriptional regulators. *Curr. Protein Pept. Sci.* **10**, 244–269 (2009).

4. Kurokawa, H. *et al.* Structural Basis of Alternative DNA Recognition by Maf Transcription Factors. *Mol. Cell. Biol.* **29**, 6232–6244 (2009).

5. Chapman, R., Sidrauski, C. & Walter, P. INTRACELLULAR SIGNALING FROM THE ENDOPLASMIC RETICULUM TO THE NUCLEUS. *Annual Review of Cell and Developmental Biology* **14**, 459–485 (1998).

6. Mori, K. *et al.* A 22 bp cis-acting element is necessary and sufficient for the induction of the yeast KAR2 (BiP) gene by unfolded proteins. *EMBO J* **11**, 2583–2593 (1992).

7. Kohno, K., Normington, K., Sambrook, J., Gething, M. J. & Mori, K. The promoter region of the yeast KAR2 (BiP) gene contains a regulatory domain that responds to the presence of unfolded proteins in the endoplasmic reticulum. *Mol. Cell. Biol* **13**, 877–890 (1993).

8. Mori, K., Kawahara, T., Yoshida, H., Yanagi, H. & Yura, T. Signalling from endoplasmic reticulum to nucleus: transcription factor with a basic-leucine zipper motif is required for the unfolded protein-response pathway. *Genes to Cells* **1**, 803 (1996).

9. Cox, J. S. & Walter, P. A Novel Mechanism for Regulating Activity of a Transcription Factor That Controls the Unfolded Protein Response. *Cell* **87**, 391–404 (1996).

10. Mori, K., Ogawa, N., Kawahara, T., Yanagi, H. & Yura, T. Palindrome with Spacer of One Nucleotide Is Characteristic of thecis-Acting Unfolded Protein Response Element inSaccharomyces cerevisiae. *Journal of Biological Chemistry* **273**, 9912 –9920 (1998).

11. Partaledis, J. A. & Berlin, V. The FKB2 gene of Saccharomyces cerevisiae, encoding the immunosuppressant-binding protein FKBP-13, is regulated in response to accumulation of unfolded proteins in the endoplasmic reticulum. *Proceedings of the National Academy of Sciences* **90**, 5450 –5454 (1993).

12. Shamu, C. E., Cox, J. S. & Walter, P. The unfolded-protein-response pathway in yeast. *Trends in Cell Biology* **4**, 56–60 (1994).

13. Travers, K. J. *et al.* Functional and Genomic Analyses Reveal an Essential Coordination between the Unfolded Protein Response and ER-Associated Degradation. *Cell* **101**, 249–258 (2000).

14. Patil, C. K., Li, H. & Walter, P. Gcn4p and Novel Upstream Activating Sequences Regulate Targets of the Unfolded Protein Response. *PLoS Biol* **2**, e246 (2004).

15. Badis, G. *et al.* A Library of Yeast Transcription Factor Motifs Reveals a Widespread Function for Rsc3 in Targeting Nucleosome Exclusion at Promoters. *Molecular Cell* **32**, 878–887 (2008).

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

17. Fordyce, P. M. *et al.* De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. *Nat. Biotechnol* **28**, 970–975 (2010).

18. Wu, R., Chaivorapol, C., Zheng, J., Li, H. & Liang, S. fREDUCE: Detection of degenerate regulatory elements using correlation with expression. *BMC Bioinformatics* **8**, 399 (2007).

19. Foat, B. C., Houshmandi, S. S., Olivas, W. M. & Bussemaker, H. J. Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. *Proceedings of the National Academy of Sciences* **102**, 17675–17680 (2005).

20. Foat, B. C., Morozov, A. V. & Bussemaker, H. J. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* **22**, (2006).

21. Nutiu, R. *et al.* Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nat Biotech* **29**, 659–664 (2011).

22. Gordân, R. *et al.* Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights. *Genome Biol* **12**, R125 (2011).

23. Chan, I.-S., Shahravan, S. H., Fedorova, A. V. & Shin, J. A. The bZIP Targets Overlapping DNA Subsites within a Half-Site, Resulting in Increased Binding Affinities†. *Biochemistry* **47**, 9646–9652 (2008).

24. Miller, M., Shuman, J. D., Sebastian, T., Dauter, Z. & Johnson, P. F. Structural Basis for DNA Recognition by the Basic Region Leucine Zipper Transcription Factor CCAAT/Enhancer-binding Protein α. *Journal of Biological Chemistry* **278**, 15178 –15184 (2003).

25. Schumacher, M. A., Goodman, R. H. & Brennan, R. G. The Structure of a CREB bZIP·Somatostatin CRE Complex Reveals the Basis for Selective Dimerization and Divalent Cation-enhanced DNA Binding. *Journal of Biological Chemistry* **275**, 35242 –35247 (2000).

26. Keller, W., König, P. & Richmond, T. J. Crystal structure of a bZIP/DNA complex at 2.2 \AA: determinants of DNA specific recognition. *J. Mol. Biol* **254**, 657–667 (1995).

27. Ellenberger, T. E., Brandl, C. J., Struhl, K. & Harrison, S. C. The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted α Helices: Crystal structure of the protein-DNA complex. *Cell* **71**, 1223–1237 (1992).

28. Fedorova, A. V., Chan, I.-S. & Shin, J. A. The GCN4 bZIP can bind to noncognate gene regulatory sequences. *Biochimica et Biophysica Acta (BBA) - Proteins & Proteomics* **1764**, 1252–1259 (2006).

29. Berger, M. F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature biotechnology* **24**, 1429–1435 (2006).

30. Badis, G. *et al.* Diversity and Complexity in DNA Recognition by Transcription Factors. *Science* 1162327 (2009).doi:10.1126/science.1162327

31. Zhu, C. *et al.* High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res.* **19**, 556–566 (2009).

32. Meijsing, S. H. *et al.* DNA binding site sequence directs glucocorticoid receptor structure and activity. *Science* **324**, 407 (2009).

Figure Legends

Figure 1 MITOMI 2.0 experimental geometry. **(A)** Three example oligonucleotide library sequences illustrating sequence structure. All sequences contain a '*CGC'* clamp (gray text), a central variable region composed of overlapping 8-nt candidate binding sites (black text), a *'C'* spacer (gray text), and an identical 14 bp sequence (red text) for hybridization and extension of a universal Cy5-labeled oligonucleotide to create dsDNA. Transparent blue boxes show 4 potential 8-mer binding sites. **(B)** Photograph of 4,000 unit cell device with a penny for scale (left); close-up view of 5 individual unit cells (right) showing protein and DNA chambers (yellow), "neck" valve (green), "sandwich" valve (orange), and "button" valve (blue). **(C)** Schematic showing top and side views of experimental chambers at different points during the experiment. **(D)** Fluorescence scans showing final Cy5 (DNA, red) and BODIPY-FL (protein, green) intensities in DNA and protein chambers; white arrow highlights DNA recruited by surface-immobilized Hac1^{i} beneath the button valve. **(E)** Histogram of measured fluorescence intensity ratios (Cy5/BODIPY-FL) on a log-linear scale to highlight outliers; the thick black vertical bar near the *y*-axis denotes four standard deviations above the background mean. Inset: zoomed view of background events on a linear scale with a Gaussian fit (black) to the background distribution (χ^2 = 1.19, p = 1.0).

Figure 2 Hac1^{i} target binding sites revealed by MITOMI 2.0 microfluidic affinity analysis using an 8mer oligonucleotide library. **(A)** Top row: Previously published motifs determined via a candidate-based approach (UPRE-1)[9]; bioinformatic

analysis of promoters associated with genes upregulated during the UPR (UPRE-2 and UPRE-3)[14]; and *in vitro* protein binding microarray experiments (UPRE-2)[15]. Bottom row, left: 7-bp sequence whose appearance within oligonucleotide sequences correlates most strongly with measured intensity ratios, as determined using fREDUCE[18]. Bottom row, middle: PSAM for this sequence determined using MatrixREDUCE[20]. Bottom row, right: Comparison between measured and predicted binding. **(B)** Measured fluorescence intensity ratios (grey circles) as a function of soluble DNA concentration for UPRE-2 (blue box) and cUPRE-1 (orange box) embedded within standard MITOMI library sequence (grey text) in three separate experiments.

Figure 3. Efficient binding of Hac1^{i} to cUPRE-1 requires an additional 2-3 nucleotides both up- and downstream from the 7-bp core. **(A)** Alignments of known Hac1^{i} target promoters containing UPRE-1 variants. Conserved 5' and 3' flanking nucleotides are indicated by orange text and shading; imperfectly conserved nucleotides within the cUPRE-1 are indicated by light grey text. **(B)** Alignments of *KAR2* ortholog promoters. Conserved 5' and 3 flanking nucleotides are indicated by orange text and shading. **(C)** Fluorescence intensity ratios as a function of DNA concentration for Hac1^{i} binding to the cUPRE-1 (orange box) in the context of either cUPRE-1-associated *KAR2* promoter sequence (orange text) or orthologous UPRE-2-associated *ERO1* (blue text) promoter sequence. **(D)** Fluorescence intensity ratios as a function of DNA concentration for DNA constructs including increasing portions of 5' *KAR2* promoter flanking sequence (top row), 3' *KAR2* promoter flanking sequence

(middle row), and both 5' and 3' *KAR2* promoter flanking sequence (bottom row). **(E)** Bar chart showing relative binding affinities for different constructs; error bars represent errors from global fits to a single-site binding model.

Figure 4. Maps of nucleotide binding preferences at each position within xcUPRE-1 and UPRE-2. **(A)** Measured relative binding affinities for all possible single nucleotide substitutions at each position within both UPRE-2 (top, blue bars) and xcUPRE-1 (bottom, orange bars) sites. Values represent the average affinity for each substitution relative to the canonical sequence (shown at top) measured over multiple replicates; errors reported are the standard error on the mean. **(B)** AffinityLogos[21] for UPRE-2 (left) and xcUPRE-1 (right) PSAMs derived from relative affinities.

Figure 5. Mutations within the Hac1^{i} DNA binding domain and an N-terminal region of extended homology can disrupt two-mode binding. **(A)** N-terminal region of Hac1^{i} protein sequence, including proposed extended homology region (orange), basic DNA binding region (pink), beginning of leucine zipper region (blue), and region mutated via error-prone PCR (gray bar). **(B)** Top: Details of mutation of *KAR2* UPRE-1 site to generate a UPRE-2 target site. Bottom: Yeast strains used in flow cytometry assays contained: (1) a reporter with mApple expression and driven by 4 repeats of the *KAR2* UPRE-1, (2) a reporter with GFP expression driven by 4 repeats of UPRE-2 within *KAR2* flanking sequences, and (3) either wild-type *HAC1^{i}* or mutant *hac1^{i}* under the control of the Gal1 promoter within an estradiol-inducible system. **(C)** Measured mean UPRE-1-driven (mApple) and UPRE-2-driven (GFP) intensities for wild-type Hac1^{i} (left) and

mutant Hac1^{i} (right) 4 hours after addition of estradiol. Each individual point represents the mean fluorescence intensity in each channel for a population of yeast cells grown from a single clone. **(D)** UPRE-1-driven (mApple) and UPRE-2 driven (GFP) intensities as a fraction of wild-type intensity for sequenced Hac1^{i} populations sharing a given mutation; bars show average values for all populations with a specific mutation. Error bars display standard error on the mean; measurements without error bars were derived from a single population.

Figure 6. Microfluidic affinity analysis of Hac1^{i} N-terminal truncation mutants. **(A)** Schematic showing truncation points for "wild-type" Hac1^{i} (N10) and mutants either retaining (N25) or lacking (N35) three residues identified as being important for UPRE-2 binding in the genetic screen (red stars). **(B)** Comparisons between relative affinity profiles measured for Hac1^{i} constructs. Left: N10 vs N25 constructs, both printed at the standard concentration; middle: N10 vs N35 constructs, with the N35 experiment conducted at 4-fold higher concentration; right: N10 vs N10 constructs at standard and 4-fold concentrations. **(C)** Differences in relative binding preferences for N10 (red line), N25 (pink line), and N35 (blue line) Hac1^{i} constructs as compared with average xcUPRE-1 binding preferences (Fig. 4A, bottom) for wild-type xcUPRE-1 and all single-nucleotide substitutions. Shaded grey area shows a single standard deviation from the mean for each oligonucleotide. **(D)** Differences in relative binding preferences for N10 (red line), N25 (pink line), and N35 (blue line) Hac1^{i} constructs as compared with average UPRE-2 binding preferences (Fig. 4A, top).

Figure 1

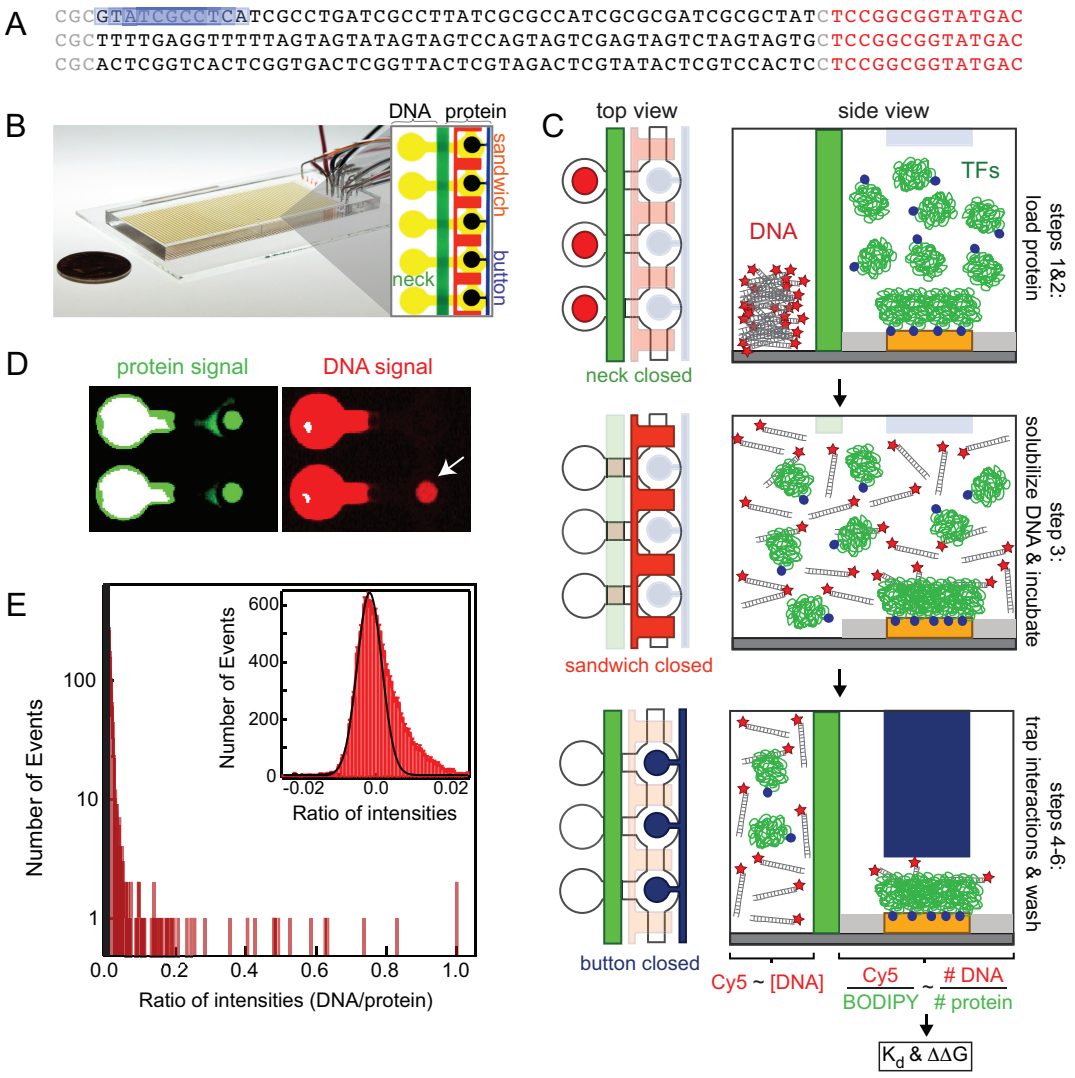

A
CGCGTATCGCCTCATCGCCTGATCGCCTTATCGCGCCATCGCGCGATCGCGCTATCTCCGGCGGTATGAC
CGCTTTTGAGGTTTTTAGTAGTATAGTAGTCCAGTAGTCGAGTAGTCTAGTAGTGCTCCGGCGGTATGAC
CGCACTCGGTCACTCGGTGACTCGGTTACTCGTAGACTCGTATACTCGTCCACTCCTCCGGCGGTATGAC
B
DNA protein
sandwich
button
neck
C
top view
side view
TFs
DNA
steps 1&2:
load protein
neck closed
step 3:
solubilize DNA & incubate
sandwich closed
steps 4-6:
trap interactions & wash
button closed
Cy5 ~ [DNA]
Cy5 / BODIPY ~ # DNA / # protein
Kd & ΔΔG
D
protein signal
DNA signal
E
Number of Events
100
10
1
0.0
0.2
0.4
0.6
0.8
1.0
Ratio of intensities (DNA/protein)
600
400
200
0
-0.02
0.0
0.02
Ratio of intensities

Figure 2

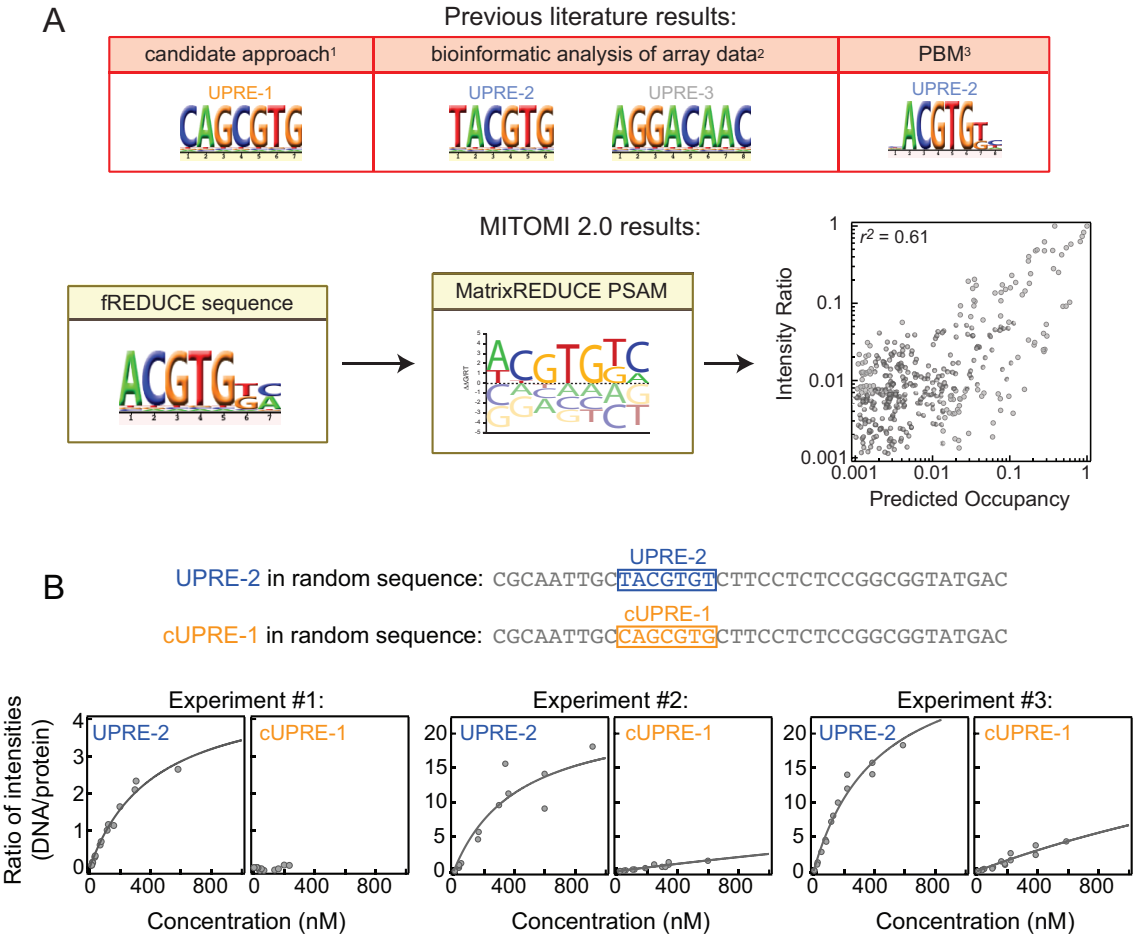

A
Previous literature results:
candidate approach[1]
bioinformatic analysis of array data[2]
PBM[3]
UPRE-1
UPRE-2
UPRE-3
UPRE-2
MITOMI 2.0 results:
fREDUCE sequence
MatrixREDUCE PSAM
r^2 = 0.61
Intensity Ratio
Predicted Occupancy
1
0.1
0.01
0.001
0.001
0.01
0.1
1
B
UPRE-2 in random sequence: CGCAATTGCTACGTGTCTTCCTCTCCGGCGGTATGAC
cUPRE-1 in random sequence: CGCAATTGCCAGCGTGCTTCCTCTCCGGCGGTATGAC
Experiment #1:
Experiment #2:
Experiment #3:
Ratio of intensities (DNA/protein)
UPRE-2
cUPRE-1
Concentration (nM)
0
400
800

Figure 3

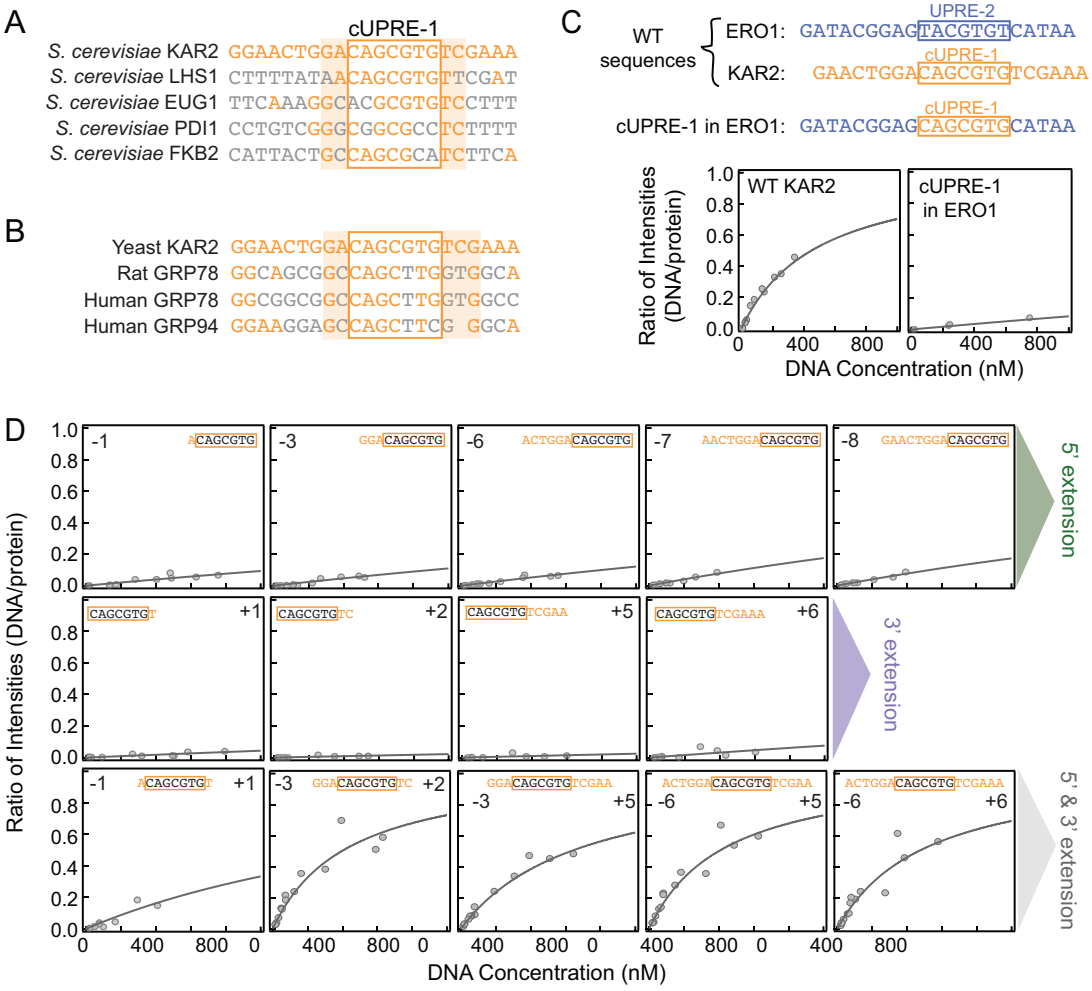

A
cUPRE-1
S. cerevisiae KAR2 GGAACTGGACAGCGTGTCGAAA
S. cerevisiae LHS1 CTTTTATAACAGCGTGTTCGAT
S. cerevisiae EUG1 TTCAAAGGCACGCGTGTCCTTT
S. cerevisiae PDI1 CCTGTCGGGCGGCGCCTCTTTT
S. cerevisiae FKB2 CATTACTGCCAGCGCATCTTCA
B
Yeast KAR2 GGAACTGGACAGCGTGTCGAAA
Rat GRP78 GGCAGCGGCCAGCTTGGTGGCA
Human GRP78 GGCGGCGGCCAGCTTGGTGGCC
Human GRP94 GGAAGGAGCCAGCTTCG GGCA
C
WT sequences
ERO1: GATACGGAGTACGTGTCATAA
UPRE-2
KAR2: GAACTGGACAGCGTGTCGAAA
cUPRE-1
cUPRE-1 in ERO1: GATACGGAGCAGCGTGCATAA
cUPRE-1
Ratio of Intensities (DNA/protein)
WT KAR2
cUPRE-1 in ERO1
DNA Concentration (nM)
D
Ratio of Intensities (DNA/protein)
5' extension
3' extension
5' & 3' extension
DNA Concentration (nM)
E
5' ext
cUPRE-1
3' ext
Relative Affinity
cUPRE-1 in KAR2 GAACTGGACAGCGTGTCGAAA
cUPRE-1 in ERO1 GATACGGAGCAGCGTGCATAA
cUPRE-1 -1 GATACGGAACAGCGTGCATAA
cUPRE-1 -3 GATACGGGACAGCGTGCATAA
cUPRE-1 -6 GATACTGGACAGCGTGCATAA
cUPRE-1 -7 GAAACTGGACAGCGTGCATAA
cUPRE-1 +1 GATACGGAGCAGCGTGTATAA
cUPRE-1 +2 GATACGGAGCAGCGTGTCTAA
cUPRE-1 +5 GATACGGAGCAGCGTGTCGAA
cUPRE-1 +6 GATACGGAGCAGCGTGTCGAAA
cUPRE-1 -1 +1 GATACGGAACAGCGTGTATAA
cUPRE-1 -3 +2 GATACGGGACAGCGTGTCTAA
cUPRE-1 -3 +5 GATACGGGACAGCGTGTCGAA
cUPRE-1 -6 +5 GATACTGGACAGCGTGTCGAA
cUPRE-1 -6 +6 GATACTGGACAGCGTGTCGAAA

Figure 4

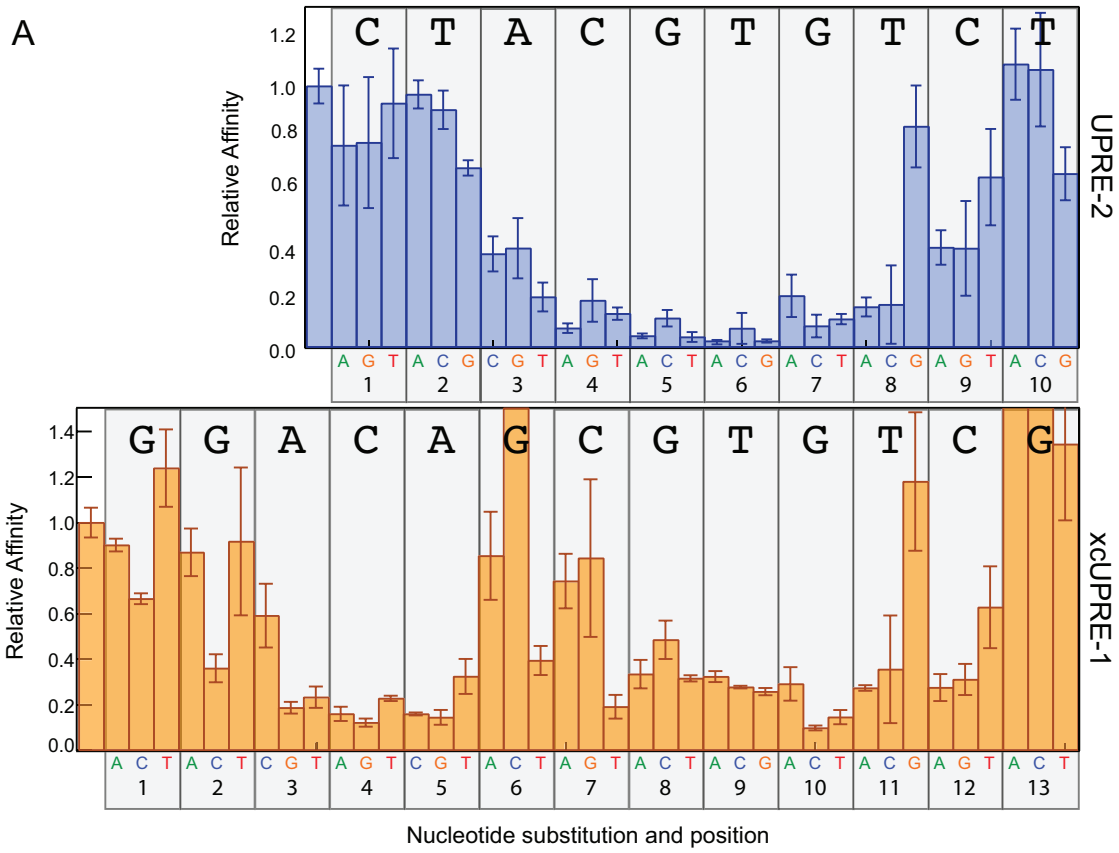

A
Relative Affinity
UPRE-2
xcUPRE-1
Nucleotide substitution and position
B
UPRE-2
xcUPRE-1
ΔΔG/RT
Position

Figure 5

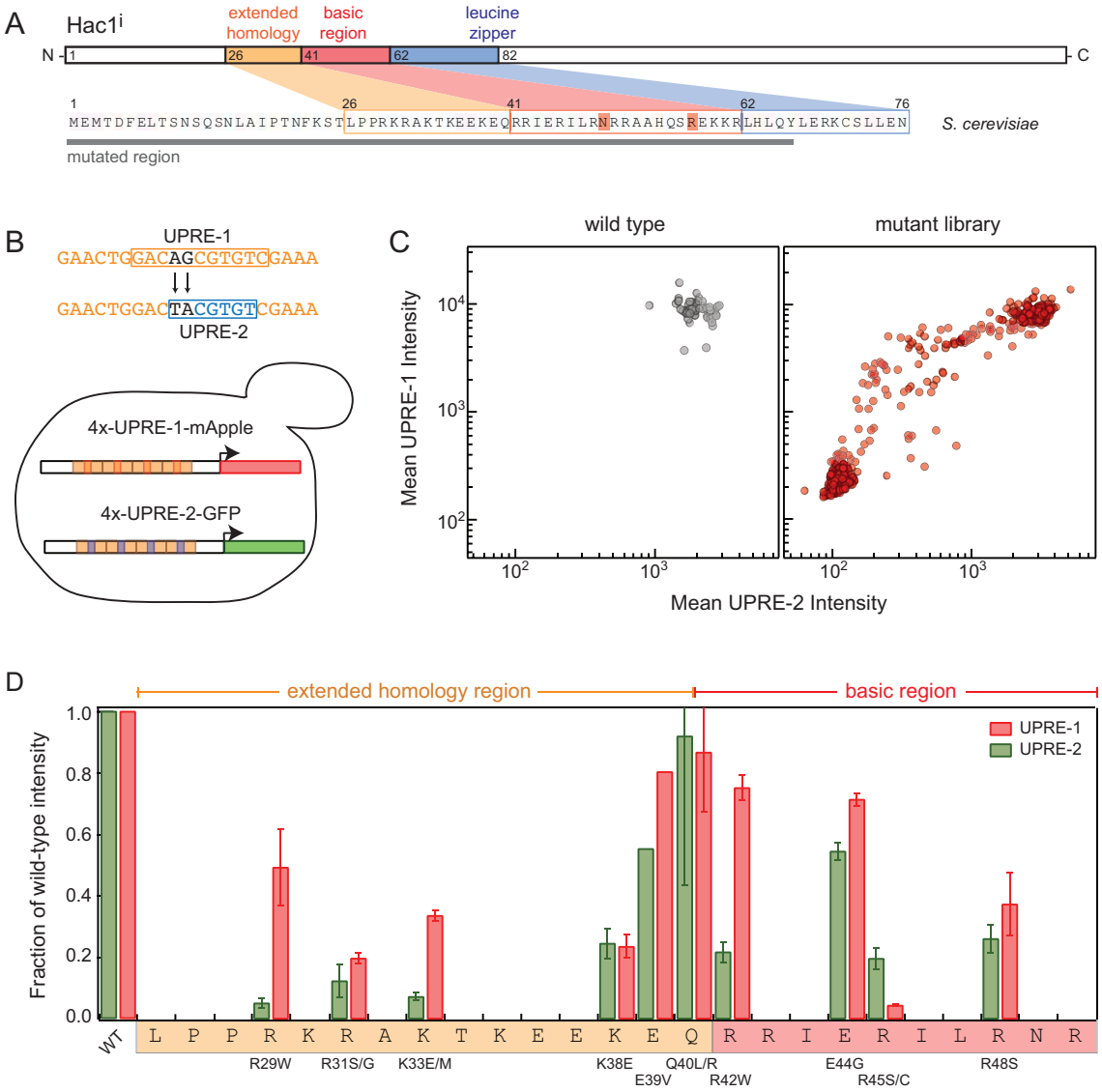
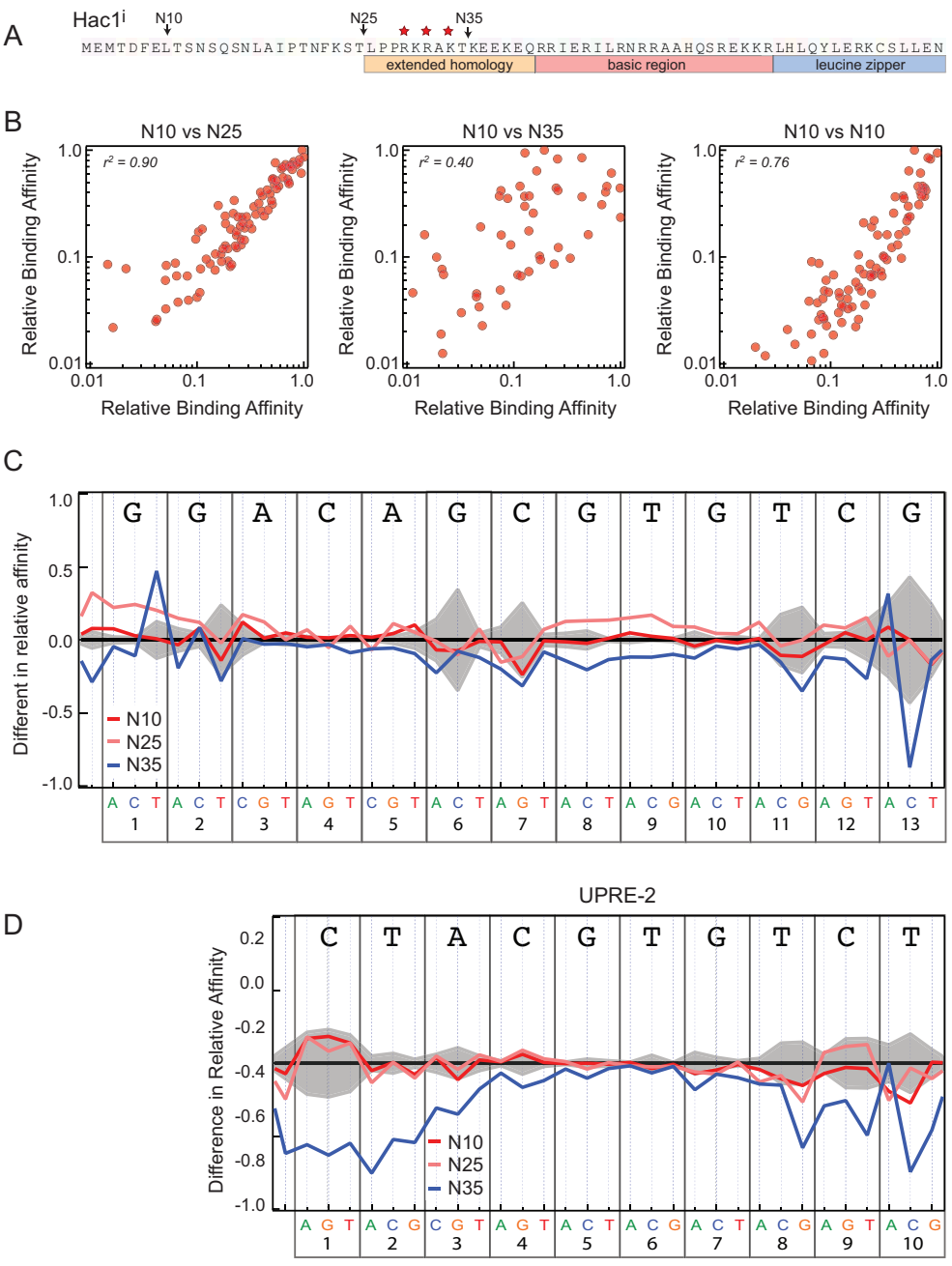

A
Hac1i
extended homology
basic region
leucine zipper
N
C
1
26
41
62
82
76
MEMTDFELTSNSQSNLAIPTNFKSTLPPRKRAKTKEEKEQRRIERILRNRRAAHQSREKKRLHLQYLERKCSLLEN
S. cerevisiae
mutated region
B
UPRE-1
GAACTGGACAGCGTGTCGAAA
GAACTGGACTACGTGTCGAAA
UPRE-2
4x-UPRE-1-mApple
4x-UPRE-2-GFP
C
wild type
mutant library
Mean UPRE-1 Intensity
Mean UPRE-2 Intensity
D
extended homology region
basic region
UPRE-1
UPRE-2
Fraction of wild-type intensity
WT
L P P R K R A K T K E E K E Q R R I E R I L R N R
R29W
R31S/G
K33E/M
K38E
E39V
Q40L/R
R42W
E44G
R45S/C
R48S

Figure 6


A
Hac1i
N10
N25
N35
MEMTDFELTSNSQSNLAIPTNFKSTLPPRKRAKTKEEKEQRRIERILRNRRAAHQSREKKRLHLQYLERKCSLLEN
extended homology
basic region
leucine zipper
B
N10 vs N25
N10 vs N35
N10 vs N10
r² = 0.90
r² = 0.40
r² = 0.76
Relative Binding Affinity
1.0
0.1
0.01
C
Different in relative affinity
1.0
0.5
0.0
-0.5
-1.0
G G A C A G C G T G T C G
N10
N25
N35
A C T | A C T | C G T | A G T | C G T | A C T | A G T | A C T | A C G | A C T | A C G | A G T | A C T
1 2 3 4 5 6 7 8 9 10 11 12 13
D
UPRE-2
Difference in Relative Affinity
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
C T A C G T G T C T
N10
N25
N35
A G T | A C G | C G T | A G T | A C T | A C G | A C T | A C G | A G T | A C G
1 2 3 4 5 6 7 8 9 10

SUPPLEMENTARY TABLES

Table S1. fREDUCE results for 6-, 7-, and 8-basepair motifs including 0-3 degenerate bases. Correlation refers to the Pearson correlation coefficient between the number of occurrences of the motif and measured intensity ratios; p-Value refers to the $-\log_{10}$ value of the probability of this correlation. All motifs with a p-Value > 100 are listed.

6Mer Motifs

Motif	*Correlation*	*p-Value*
MCACGT	0.69	1441
ACACGT	0.60	898
ACGTGG	0.27	122

7Mer Motifs

Motif	*Correlation*	*p-Value*
KMCACGT	0.80	2896
ACGTGTC	0.70	1539
ACGTGGC	0.35	226
ACACGWD	0.19	120

8Mer Motifs

Motif	*Correlation*	*p-Value*
ACGTGKMC	0.93	11178
BGACACGT	0.82	3271
ACGTGTCC	0.71	1663
ACGTGTG	0.40	314
ACGTGGCC	0.39	281
GCCACGTR	0.38	515
RMCACGTR	0.35	556
VACACGTR	0.33	462
ACACGTR	0.29	300
ATCGTGTC	0.28	129
RACACGTG	0.28	313
TGMCACGA	0.21	191
GMCACGAT	0.21	167

Table S2. Oligonucleotide sequences used for measurements of concentration-dependent binding to map nucleotide preferences within xcUPRE-1 and UPRE-2 target sites for genomic UPRE variants.

Oligo #	Name	Sequence
1	full UPRE-1	CGCAATTGCGATACGGGACAGCGTGTCGTAACTTCCTCTCCGGCGGTATGAC
2	full UPRE-1 sub 0A	CGCAATTGCGATACG**A**GACAGCGTGTCGTAACTTCCTCTCCGGCGGTATGAC
3	full UPRE-1 sub 0C	CGCAATTGCGATACG**C**GACAGCGTGTCGTAACTTCCTCTCCGGCGGTATGAC
4	full UPRE-1 sub 0T	CGCAATTGCGATACG**T**GACAGCGTGTCGTAACTTCCTCTCCGGCGGTATGAC
5	full UPRE-1 sub 1A	CGCAATTGCGATACGG**A**ACAGCGTGTCGTAACTTCCTCTCCGGCGGTATGAC
6	full UPRE-1 sub 1C	CGCAATTGCGATACGG**C**ACAGCGTGTCGTAACTTCCTCTCCGGCGGTATGAC
7	full UPRE-1 sub 1T	CGCAATTGCGATACGG**T**ACAGCGTGTCGTAACTTCCTCTCCGGCGGTATGAC
8	full UPRE-1 sub 2C	CGCAATTGCGATACGGG**C**CAGCGTGTCGTAACTTCCTCTCCGGCGGTATGAC
9	full UPRE-1 sub 2G	CGCAATTGCGATACGGG**G**CAGCGTGTCGTAACTTCCTCTCCGGCGGTATGAC
10	full UPRE-1 sub 2T	CGCAATTGCGATACGGG**T**CAGCGTGTCGTAACTTCCTCTCCGGCGGTATGAC
11	full UPRE-1 sub 3A	CGCAATTGCGATACGGGA**A**AGCGTGTCGTAACTTCCTCTCCGGCGGTATGAC
12	full UPRE-1 sub 3G	CGCAATTGCGATACGGGA**G**AGCGTGTCGTAACTTCCTCTCCGGCGGTATGAC
13	full UPRE-1 sub 3T	CGCAATTGCGATACGGGA**T**AGCGTGTCGTAACTTCCTCTCCGGCGGTATGAC
14	full UPRE-1 sub 4C	CGCAATTGCGATACGGGAC**C**GCGTGTCGTAACTTCCTCTCCGGCGGTATGAC
15	full UPRE-1 sub 4G	CGCAATTGCGATACGGGAC**G**GCGTGTCGTAACTTCCTCTCCGGCGGTATGAC
16	full UPRE-1 sub 4T	CGCAATTGCGATACGGGAC**T**GCGTGTCGTAACTTCCTCTCCGGCGGTATGAC
17	full UPRE-1 sub 5A	CGCAATTGCGATACGGGACA**A**CGTGTCGTAACTTCCTCTCCGGCGGTATGAC
18	full UPRE-1 sub 5C	CGCAATTGCGATACGGGACA**C**CGTGTCGTAACTTCCTCTCCGGCGGTATGAC
19	full UPRE-1 sub 5T	CGCAATTGCGATACGGGACA**T**CGTGTCGTAACTTCCTCTCCGGCGGTATGAC
20	full UPRE-1 sub 6A	CGCAATTGCGATACGGGACAG**A**GTGTCGTAACTTCCTCTCCGGCGGTATGAC
21	full UPRE-1 sub 6G	CGCAATTGCGATACGGGACAG**G**GTGTCGTAACTTCCTCTCCGGCGGTATGAC
22	full UPRE-1 sub 6T	CGCAATTGCGATACGGGACAG**T**GTGTCGTAACTTCCTCTCCGGCGGTATGAC
23	full UPRE-1 sub 7A	CGCAATTGCGATACGGGACAGC**A**TGTCGTAACTTCCTCTCCGGCGGTATGAC
24	full UPRE-1 sub 7C	CGCAATTGCGATACGGGACAGC**C**TGTCGTAACTTCCTCTCCGGCGGTATGAC
25	full UPRE-1 sub 7T	CGCAATTGCGATACGGGACAGC**T**TGTCGTAACTTCCTCTCCGGCGGTATGAC
26	full UPRE-1 sub 8A	CGCAATTGCGATACGGGACAGCG**A**GTCGTAACTTCCTCTCCGGCGGTATGAC
27	full UPRE-1 sub 8C	CGCAATTGCGATACGGGACAGCG**C**GTCGTAACTTCCTCTCCGGCGGTATGAC
28	full UPRE-1 sub 8G	CGCAATTGCGATACGGGACAGCG**G**GTCGTAACTTCCTCTCCGGCGGTATGAC
29	full UPRE-1 sub 9A	CGCAATTGCGATACGGGACAGCGT**A**TCGTAACTTCCTCTCCGGCGGTATGAC
30	full UPRE-1 sub 9C	CGCAATTGCGATACGGGACAGCGT**C**TCGTAACTTCCTCTCCGGCGGTATGAC
31	full UPRE-1 sub 9T	CGCAATTGCGATACGGGACAGCGT**T**TCGTAACTTCCTCTCCGGCGGTATGAC
32	full UPRE-1 sub 10A	CGCAATTGCGATACGGGACAGCGTG**A**CGTAACTTCCTCTCCGGCGGTATGAC
33	full UPRE-1 sub 10C	CGCAATTGCGATACGGGACAGCGTG**C**CGTAACTTCCTCTCCGGCGGTATGAC
34	full UPRE-1 sub 10G	CGCAATTGCGATACGGGACAGCGTG**G**CGTAACTTCCTCTCCGGCGGTATGAC
35	full UPRE-1 sub 11A	CGCAATTGCGATACGGGACAGCGTGT**A**GTAACTTCCTCTCCGGCGGTATGAC
36	full UPRE-1 sub 11G	CGCAATTGCGATACGGGACAGCGTGT**G**GTAACTTCCTCTCCGGCGGTATGAC
37	full UPRE-1 sub 11T	CGCAATTGCGATACGGGACAGCGTGT**T**GTAACTTCCTCTCCGGCGGTATGAC
38	full UPRE-1 sub 12A	CGCAATTGCGATACGGGACAGCGTGTC**A**TAACTTCCTCTCCGGCGGTATGAC
39	full UPRE-1 sub 12C	CGCAATTGCGATACGGGACAGCGTGTC**C**TAACTTCCTCTCCGGCGGTATGAC
40	full UPRE-1 sub 12T	CGCAATTGCGATACGGGACAGCGTGTC**T**TAACTTCCTCTCCGGCGGTATGAC
41	full UPRE-1 sub 1A and 11A	CGCAATTGCGATACGG**A**ACAGCGTGT**A**GTAACTTCCTCTCCGGCGGTATGAC
42	full UPRE-1 sub 2C and 10C	CGCAATTGCGATACGGG**C**CAGCGTG**C**CGTAACTTCCTCTCCGGCGGTATGAC
43	full UPRE-1 sub 3A and 9A	CGCAATTGCGATACGGGA**A**AGCGT**A**TCGTAACTTCCTCTCCGGCGGTATGAC
44	full UPRE-1 sub 4C and 8C	CGCAATTGCGATACGGGAC**C**GCG**C**GTCGTAACTTCCTCTCCGGCGGTATGAC
45	full UPRE-1 sub 5A and 7A	CGCAATTGCGATACGGGACA**A**C**A**TGTCGTAACTTCCTCTCCGGCGGTATGAC
46	full UPRE-1 sub 3A and 4C	CGCAATTGCGATACGGGA**AC**GCGTGTCGTAACTTCCTCTCCGGCGGTATGAC
47	random sequence v1	CGCAATTGCGATACG**ATCGAATGCAGT**GTAACTTCCTCTCCGGCGGTATGAC
48	full UPRE-2	CGCAATTGCGAACTGGACTACGTGTCTGAAACTTCCTCTCCGGCGGTATGAC
49	full UPRE-2 sub 0A	CGCAATTGCGAACTGGA**A**TACGTGTCTGAAACTTCCTCTCCGGCGGTATGAC
50	full UPRE-2 sub 0G	CGCAATTGCGAACTGGA**G**TACGTGTCTGAAACTTCCTCTCCGGCGGTATGAC

51	full UPRE-2 sub 0T	CGCAATTGCGAACTGGA**T**TACGTGTCTGAAACTTCCTCTCCGGCGGTATGAC
52	full UPRE-2 sub 1A	CGCAATTGCGAACTGGAC**A**ACGTGTCTGAAACTTCCTCTCCGGCGGTATGAC
53	full UPRE-2 sub 1C	CGCAATTGCGAACTGGAC**C**ACGTGTCTGAAACTTCCTCTCCGGCGGTATGAC
54	full UPRE-2 sub 1G	CGCAATTGCGAACTGGAC**G**ACGTGTCTGAAACTTCCTCTCCGGCGGTATGAC
55	full UPRE-2 sub 2C	CGCAATTGCGAACTGGACT**C**CGTGTCTGAAACTTCCTCTCCGGCGGTATGAC
56	full UPRE-2 sub 2G	CGCAATTGCGAACTGGACT**G**CGTGTCTGAAACTTCCTCTCCGGCGGTATGAC
57	full UPRE-2 sub 2T	CGCAATTGCGAACTGGACT**T**CGTGTCTGAAACTTCCTCTCCGGCGGTATGAC
58	full UPRE-2 sub 3A	CGCAATTGCGAACTGGACTA**A**GTGTCTGAAACTTCCTCTCCGGCGGTATGAC
59	full UPRE-2 sub 3G	CGCAATTGCGAACTGGACTA**G**GTGTCTGAAACTTCCTCTCCGGCGGTATGAC
60	full UPRE-2 sub 3T	CGCAATTGCGAACTGGACTA**T**GTGTCTGAAACTTCCTCTCCGGCGGTATGAC
61	full UPRE-2 sub 4A	CGCAATTGCGAACTGGACTAC**A**TGTCTGAAACTTCCTCTCCGGCGGTATGAC
62	full UPRE-2 sub 4C	CGCAATTGCGAACTGGACTAC**C**TGTCTGAAACTTCCTCTCCGGCGGTATGAC
63	full UPRE-2 sub 4T	CGCAATTGCGAACTGGACTAC**T**TGTCTGAAACTTCCTCTCCGGCGGTATGAC
64	full UPRE-2 sub 5A	CGCAATTGCGAACTGGACTACG**A**GTCTGAAACTTCCTCTCCGGCGGTATGAC
65	full UPRE-2 sub 5C	CGCAATTGCGAACTGGACTACG**C**GTCTGAAACTTCCTCTCCGGCGGTATGAC
66	full UPRE-2 sub 5G	CGCAATTGCGAACTGGACTACG**G**GTCTGAAACTTCCTCTCCGGCGGTATGAC
67	full UPRE-2 sub 6A	CGCAATTGCGAACTGGACTACGT**A**TCTGAAACTTCCTCTCCGGCGGTATGAC
68	full UPRE-2 sub 6C	CGCAATTGCGAACTGGACTACGT**C**TCTGAAACTTCCTCTCCGGCGGTATGAC
69	full UPRE-2 sub 6T	CGCAATTGCGAACTGGACTACGT**T**TCTGAAACTTCCTCTCCGGCGGTATGAC
70	full UPRE-2 sub 7A	CGCAATTGCGAACTGGACTACGTG**A**CTGAAACTTCCTCTCCGGCGGTATGAC
71	full UPRE-2 sub 7C	CGCAATTGCGAACTGGACTACGTG**C**CTGAAACTTCCTCTCCGGCGGTATGAC
72	full UPRE-2 sub 7G	CGCAATTGCGAACTGGACTACGTG**G**CTGAAACTTCCTCTCCGGCGGTATGAC
73	full UPRE-2 sub 8A	CGCAATTGCGAACTGGACTACGTGT**A**TGAAACTTCCTCTCCGGCGGTATGAC
74	full UPRE-2 sub 8G	CGCAATTGCGAACTGGACTACGTGT**G**TGAAACTTCCTCTCCGGCGGTATGAC
75	full UPRE-2 sub 8T	CGCAATTGCGAACTGGACTACGTGT**T**TGAAACTTCCTCTCCGGCGGTATGAC
76	full UPRE-2 sub 9A	CGCAATTGCGAACTGGACTACGTGTC**A**GAAACTTCCTCTCCGGCGGTATGAC
77	full UPRE-2 sub 9C	CGCAATTGCGAACTGGACTACGTGTC**C**GAAACTTCCTCTCCGGCGGTATGAC
78	full UPRE-2 sub 9G	CGCAATTGCGAACTGGACTACGTGTC**G**GAAACTTCCTCTCCGGCGGTATGAC
79	full UPRE-2 sub 1C and 8A	CGCAATTGCGAACTGGAC**C**ACGTGT**A**TGAAACTTCCTCTCCGGCGGTATGAC
80	full UPRE-2 sub 2C and 7C	CGCAATTGCGAACTGGACT**C**CGTG**C**CTGAAACTTCCTCTCCGGCGGTATGAC
81	full UPRE-2 sub 3A and 6A	CGCAATTGCGAACTGGACTA**A**GT**A**TCTGAAACTTCCTCTCCGGCGGTATGAC
82	full UPRE-2 sub 4A and 5G	CGCAATTGCGAACTGGACTAC**AG**GTCTGAAACTTCCTCTCCGGCGGTATGAC
83	random sequence v2	CGCAATTGCGAACTGGAC**CGATCTAGA**GAAACTTCCTCTCCGGCGGTATGAC
84	supersymmetric full UPRE-1	CGCAATTGCGATACG**C**GACA**C**CGTGTCGTAACTTCCTCTCCGGCGGTATGAC
85	supersymmetric full UPRE-1	CGCAATTGCGATACGGGACAGC**C**TGTC**C**TAACTTCCTCTCCGGCGGTATGAC
86	supersymmetric full UPRE-2	CGCAATTGCGAACTGGA**AC**ACGTGTCTGAAACTTCCTCTCCGGCGGTATGAC
87	supersymmetric full UPRE-2	CGCAATTGCGAACTGG**GAC**ACGTGTCTGAAACTTCCTCTCCGGCGGTATGAC
88	supersymmetric full UPRE-2	CGCAATTGCGAACTGGACTACGT**AG**CTGAAACTTCCTCTCCGGCGGTATGAC
89	KAR2	CGCAATTGCGCAACTGGACAGCGTGTCGAAACTTCCTCTCCGGCGGTATGAC
90	LHS1	CGCAATTGCCTTTTATAACAGCGTGTTCGATCTTCCTCTCCGGCGGTATGAC
91	EUG1	CGCAATTGCTTCAAAGGCACGCGTGTCCTTTCTTCCTCTCCGGCGGTATGAC
92	PDI1	CGCAATTGCCCTGTCGGGCGGCGCCTCTTTTCTTCCTCTCCGGCGGTATGAC
93	FKB2	CGCAATTGCCATTACTGCCAGCGCATCTTCACTTCCTCTCCGGCGGTATGAC
94	ERO1	CGCAATTGCGATACGGAGTACGTGTCATAAACTTCCTCTCCGGCGGTATGAC
95	SEC66	CGCAATTGCTTTTAGGAACACGTCTAAAAGTCTTCCTCTCCGGCGGTATGAC
96	random sequence v3	CGCAATTGCGAGTGTATTACCGTGACGGCCGCTTCCTCTCCGGCGGTATGAC

Table S3. xcUPRE-1 PSAM (calculated by averaging relative binding preferences over 3 independent experiments).

Position	A	C	G	T
1	0.73	0.54	0.81	1
2	0.87	0.36	1	0.92
3	1	0.59	0.19	0.23
4	0.16	1	0.12	0.23
5	1	0.16	0.14	0.32
6	0.49	1	0.57	0.23
7	0.74	1	0.84	0.19
8	0.33	0.49	1	0.32
9	0.32	0.28	0.26	1
10	0.29	0.1	1	0.15
11	0.23	0.3	1	0.85
12	0.28	1	0.31	0.63
13	0.67	1	0.43	0.57

Table S4. UPRE-2 PSAM (calculated by averaging relative binding preferences over 4 independent experiments).

Position	A	C	G	T
1	0.77	1	0.78	0.93
2	0.97	0.91	0.69	1
3	1	0.36	0.38	0.19
4	0.07	1	0.18	0.13
5	0.04	0.11	1	0.04
6	0.02	0.07	0.02	1
7	0.2	0.08	1	0.11
8	0.15	0.16	0.85	1
9	0.38	1	0.38	0.65
10	1	0.98	0.61	0.92

Table S5. Measured affinities for all 96 oligonucleotide sequences including single- and double-nucleotide substitutions within xcUPRE-1, UPRE-2, and natural variants.

Sequence	Kd	Error
CGCAATTGCGATACGGGACAGCGTGTCGTAACTTCCTCTCCGGCGGTATGAC	0.71	0.08
CGCAATTGCGATACG**A**GACAGCGTGTCGTAACTTCCTCTCCGGCGGTATGAC	0.78	0.10
CGCAATTGCGATACG**C**GACAGCGTGTCGTAACTTCCTCTCCGGCGGTATGAC	1.15	0.14
CGCAATTGCGATACG**T**GACAGCGTGTCGTAACTTCCTCTCCGGCGGTATGAC	0.67	0.08
CGCAATTGCGATACGG**A**ACAGCGTGTCGTAACTTCCTCTCCGGCGGTATGAC	1.07	0.12
CGCAATTGCGATACGG**C**ACAGCGTGTCGTAACTTCCTCTCCGGCGGTATGAC	1.53	0.25
CGCAATTGCGATACGG**T**ACAGCGTGTCGTAACTTCCTCTCCGGCGGTATGAC	1.43	0.19
CGCAATTGCGATACGGG**C**CAGCGTGTCGTAACTTCCTCTCCGGCGGTATGAC	0.96	0.12
CGCAATTGCGATACGGG**G**CAGCGTGTCGTAACTTCCTCTCCGGCGGTATGAC	3.91	1.01
CGCAATTGCGATACGGG**T**CAGCGTGTCGTAACTTCCTCTCCGGCGGTATGAC	2.44	0.58
CGCAATTGCGATACGGGA**A**AGCGTGTCGTAACTTCCTCTCCGGCGGTATGAC	4.15	1.13
CGCAATTGCGATACGGGA**G**AGCGTGTCGTAACTTCCTCTCCGGCGGTATGAC	5.42	3.26
CGCAATTGCGATACGGGA**T**AGCGTGTCGTAACTTCCTCTCCGGCGGTATGAC	2.82	0.48
CGCAATTGCGATACGGGAC**C**GCGTGTCGTAACTTCCTCTCCGGCGGTATGAC	4.15	1.35
CGCAATTGCGATACGGGAC**G**GCGTGTCGTAACTTCCTCTCCGGCGGTATGAC	3.39	0.68
CGCAATTGCGATACGGGAC**T**GCGTGTCGTAACTTCCTCTCCGGCGGTATGAC	1.49	0.23
CGCAATTGCGATACGGGACA**A**CGTGTCGTAACTTCCTCTCCGGCGGTATGAC	1.21	0.14
CGCAATTGCGATACGGGACA**C**CGTGTCGTAACTTCCTCTCCGGCGGTATGAC	0.53	0.06
CGCAATTGCGATACGGGACA**T**CGTGTCGTAACTTCCTCTCCGGCGGTATGAC	2.26	0.36
CGCAATTGCGATACGGGACAG**A**GTGTCGTAACTTCCTCTCCGGCGGTATGAC	1.18	0.16
CGCAATTGCGATACGGGACAG**G**GTGTCGTAACTTCCTCTCCGGCGGTATGAC	2.94	0.44
CGCAATTGCGATACGGGACAG**T**GTGTCGTAACTTCCTCTCCGGCGGTATGAC	4.76	1.28
CGCAATTGCGATACGGGACAGC**A**TGTCGTAACTTCCTCTCCGGCGGTATGAC	2.76	0.41
CGCAATTGCGATACGGGACAGC**C**TGTCGTAACTTCCTCTCCGGCGGTATGAC	1.95	0.30
CGCAATTGCGATACGGGACAGC**T**TGTCGTAACTTCCTCTCCGGCGGTATGAC	2.46	0.41
CGCAATTGCGATACGGGACAGCG**A**GTCGTAACTTCCTCTCCGGCGGTATGAC	1.92	0.36
CGCAATTGCGATACGGGACAGCG**C**GTCGTAACTTCCTCTCCGGCGGTATGAC	2.49	0.43
CGCAATTGCGATACGGGACAGCG**G**GTCGTAACTTCCTCTCCGGCGGTATGAC	2.99	0.76
CGCAATTGCGATACGGGACAGCGT**A**TCGTAACTTCCTCTCCGGCGGTATGAC	4.40	1.20
CGCAATTGCGATACGGGACAGCGT**C**TCGTAACTTCCTCTCCGGCGGTATGAC	9.10	6.38
CGCAATTGCGATACGGGACAGCGT**T**TCGTAACTTCCTCTCCGGCGGTATGAC	8.56	4.18
CGCAATTGCGATACGGGACAGCGTG**A**CGTAACTTCCTCTCCGGCGGTATGAC	2.83	0.72
CGCAATTGCGATACGGGACAGCGTG**C**CGTAACTTCCTCTCCGGCGGTATGAC	7.73	3.33
CGCAATTGCGATACGGGACAGCGTG**G**CGTAACTTCCTCTCCGGCGGTATGAC	0.92	0.12
CGCAATTGCGATACGGGACAGCGTGT**A**GTAACTTCCTCTCCGGCGGTATGAC	4.23	1.10
CGCAATTGCGATACGGGACAGCGTGT**G**GTAACTTCCTCTCCGGCGGTATGAC	1.97	0.42
CGCAATTGCGATACGGGACAGCGTGT**T**GTAACTTCCTCTCCGGCGGTATGAC	1.34	0.17
CGCAATTGCGATACGGGACAGCGTGTC**A**TAACTTCCTCTCCGGCGGTATGAC	0.47	0.06
CGCAATTGCGATACGGGACAGCGTGTC**C**TAACTTCCTCTCCGGCGGTATGAC	0.36	0.04
CGCAATTGCGATACGGGACAGCGTGTC**T**TAACTTCCTCTCCGGCGGTATGAC	0.89	0.11
CGCAATTGCGATACG**A**AACAGCGTGT**A**GTAACTTCCTCTCCGGCGGTATGAC	3.02	0.57
CGCAATTGCGATACGGG**C**CAGCGTG**C**CGTAACTTCCTCTCCGGCGGTATGAC	5.71	2.37
CGCAATTGCGATACGGGA**A**AGCGT**A**TCGTAACTTCCTCTCCGGCGGTATGAC	16.42	14.17
CGCAATTGCGATACGGGAC**C**GCG**C**GTCGTAACTTCCTCTCCGGCGGTATGAC	14.90	15.12
CGCAATTGCGATACGGGACA**A**C**A**TGTCGTAACTTCCTCTCCGGCGGTATGAC	5.39	1.59
CGCAATTGCGATACGGGA**AC**GCGTGTCGTAACTTCCTCTCCGGCGGTATGAC	2.09	0.39
CGCAATTGCGAACTGGACTACGTGTCTGAAACTTCCTCTCCGGCGGTATGAC	0.49	0.05
CGCAATTGCGAACTGGA**A**TACGTGTCTGAAACTTCCTCTCCGGCGGTATGAC	0.47	0.05
CGCAATTGCGAACTGGA**G**TACGTGTCTGAAACTTCCTCTCCGGCGGTATGAC	0.46	0.05
CGCAATTGCGAACTGGA**T**TACGTGTCTGAAACTTCCTCTCCGGCGGTATGAC	0.42	0.05
CGCAATTGCGAACTGGAC**A**ACGTGTCTGAAACTTCCTCTCCGGCGGTATGAC	0.49	0.05
CGCAATTGCGAACTGGAC**C**ACGTGTCTGAAACTTCCTCTCCGGCGGTATGAC	0.49	0.05
CGCAATTGCGAACTGGAC**G**ACGTGTCTGAAACTTCCTCTCCGGCGGTATGAC	0.75	0.08
CGCAATTGCGAACTGGACT**C**CGTGTCTGAAACTTCCTCTCCGGCGGTATGAC	1.13	0.13
CGCAATTGCGAACTGGACT**G**CGTGTCTGAAACTTCCTCTCCGGCGGTATGAC	1.76	0.30

CGCAATTGCGAACTGGACTTCGTGTCTGAAACTTCCTCTCCGGCGGTATGAC	2.08	0.31
CGCAATTGCGAACTGGACTAAGTGTCTGAAACTTCCTCTCCGGCGGTATGAC	5.40	2.18
CGCAATTGCGAACTGGACTAGGTGTCTGAAACTTCCTCTCCGGCGGTATGAC	1.84	0.27
CGCAATTGCGAACTGGACTATGTGTCTGAAACTTCCTCTCCGGCGGTATGAC	3.36	0.96
CGCAATTGCGAACTGGACTACATGTCTGAAACTTCCTCTCCGGCGGTATGAC	10.98	6.52
CGCAATTGCGAACTGGACTACCTGTCTGAAACTTCCTCTCCGGCGGTATGAC	4.09	1.44
CGCAATTGCGAACTGGACTACTTGTCTGAAACTTCCTCTCCGGCGGTATGAC	14.37	9.56
CGCAATTGCGAACTGGACTACGAGTCTGAAACTTCCTCTCCGGCGGTATGAC	18.15	26.24
CGCAATTGCGAACTGGACTACGCGTCTGAAACTTCCTCTCCGGCGGTATGAC	13.75	11.08
CGCAATTGCGAACTGGACTACGGGTCTGAAACTTCCTCTCCGGCGGTATGAC	28.55	65.14
CGCAATTGCGAACTGGACTACGTATCTGAAACTTCCTCTCCGGCGGTATGAC	3.76	0.99
CGCAATTGCGAACTGGACTACGTCTCTGAAACTTCCTCTCCGGCGGTATGAC	14.67	17.94
CGCAATTGCGAACTGGACTACGTTTCTGAAACTTCCTCTCCGGCGGTATGAC	4.63	1.54
CGCAATTGCGAACTGGACTACGTGACTGAAACTTCCTCTCCGGCGGTATGAC	4.21	1.72
CGCAATTGCGAACTGGACTACGTGCCTGAAACTTCCTCTCCGGCGGTATGAC	9.49	6.34
CGCAATTGCGAACTGGACTACGTGGCTGAAACTTCCTCTCCGGCGGTATGAC	0.66	0.07
CGCAATTGCGAACTGGACTACGTGTATGAAACTTCCTCTCCGGCGGTATGAC	1.48	0.17
CGCAATTGCGAACTGGACTACGTGTGTGAAACTTCCTCTCCGGCGGTATGAC	1.30	0.16
CGCAATTGCGAACTGGACTACGTGTTTGAAACTTCCTCTCCGGCGGTATGAC	0.74	0.08
CGCAATTGCGAACTGGACTACGTGTCAGAAACTTCCTCTCCGGCGGTATGAC	0.52	0.05
CGCAATTGCGAACTGGACTACGTGTCCGAAACTTCCTCTCCGGCGGTATGAC	0.59	0.07
CGCAATTGCGAACTGGACTACGTGTCGGAAACTTCCTCTCCGGCGGTATGAC	0.67	0.07
CGCAATTGCGAACTGGACCACGTGTATGAAACTTCCTCTCCGGCGGTATGAC	0.85	0.09
CGCAATTGCGAACTGGACTCCGTGCCTGAAACTTCCTCTCCGGCGGTATGAC	4.61	1.34
CGCAATTGCGAACTGGACTAAGTATCTGAAACTTCCTCTCCGGCGGTATGAC	5.93	1.57
CGCAATTGCGAACTGGACTACAGGTCTGAAACTTCCTCTCCGGCGGTATGAC	3.95	1.05
CGCAATTGCGATACGCGACACCGTGTCGTAACTTCCTCTCCGGCGGTATGAC	0.82	0.09
CGCAATTGCGATACGGGACAGCCTGTCCTAACTTCCTCTCCGGCGGTATGAC	0.65	0.07
CGCAATTGCGAACTGGAACACGTGTCTGAAACTTCCTCTCCGGCGGTATGAC	0.52	0.05
CGCAATTGCGAACTGGGACACGTGTCTGAAACTTCCTCTCCGGCGGTATGAC	0.44	0.04
CGCAATTGCGAACTGGACTACGTAGCTGAAACTTCCTCTCCGGCGGTATGAC	1.06	0.14
CGCAATTGCGCAACTGGACAGCGTGTCGAAACTTCCTCTCCGGCGGTATGAC	0.51	0.05
CGCAATTGCCTTTTATAACAGCGTGTTCGATCTTCCTCTCCGGCGGTATGAC	1.41	0.23
CGCAATTGCTTCAAAGGCACGCGTGTCCTTTCTTCCTCTCCGGCGGTATGAC	0.69	0.07
CGCAATTGCCCTGTCGGGCGGCGCCTCTTTTCTTCCTCTCCGGCGGTATGAC	1.75	0.34
CGCAATTGCCATTACTGCCAGCGCATCTTCACTTCCTCTCCGGCGGTATGAC	1.50	0.21
CGCAATTGCGATACGGAGTACGTGTCATAAACTTCCTCTCCGGCGGTATGAC	0.54	0.06
CGCAATTGCTTTTAGGAACACGTCTAAAAGTCTTCCTCTCCGGCGGTATGAC	1.71	0.26

Table S6. Top predicted genomic targets for UPRE-2 binding. 'UPRE-1 score' is the predicted affinity for xcUPRE-1 binding; 'UPRE-2 score' is the predicted affinity for UPRE-2 binding; 'UPR target' indicates whether or not the gene is considered to be a known UPR target[1].

Systematic	Standard	UPRE-1	UPRE-2	UPR
YFR026C	ULI1	0.54	3.09	Y
YJL158C	CIS3	0.63	2.85	N
YMR194C-A	NA	0.56	2.85	N
YMR195W	ICY1	0.59	2.79	N
YHR099W	TRA1	0.44	2.73	Y
YHR098C	SFB3	0.44	2.72	Y
YDR281C	PHM6	0.50	2.66	N
YGR146C-A	NA	0.58	2.48	N
YJR145C	RPS4A	0.45	2.44	N
YKL165C	MCD4	0.58	2.42	Y
YJR146W	NA	0.44	2.41	N
YKL015W	PUT3	0.24	2.32	N
YKL016C	ATP7	0.24	2.32	N
YGL045W	RIM8	0.49	2.31	N
YDR073W	SNF11	0.48	2.28	Y
YPL019C	VTC3	0.44	2.27	N
YPL018W	CTF19	0.44	2.27	N
YDL124W	NA	0.46	2.27	N
YDL125C	HNT1	0.46	2.27	Y
YHR102W	KIC1	0.38	2.26	Y
YHR101C	BIG1	0.38	2.26	N
YJR009C	TDH2	0.45	2.16	Y
YMR011W	HXT2	0.38	2.16	N
YMR250W	GAD1	0.51	2.16	N
YJL101C	GSH1	0.65	2.16	N
YPL240C	HSP82	0.68	2.15	N
YNL097C	PHO23	0.36	2.15	N
YLR300W	EXG1	0.54	2.14	N
YMR083W	ADH3	0.43	2.11	N
YDR072C	IPT1	0.43	2.10	N
YNR060W	FRE4	0.46	2.10	N
YDR233C	RTN1	0.41	2.08	N
YDR234W	LYS4	0.41	2.08	N
YGR161W-C	NA	0.65	2.07	N
YGR161C	RTS3	0.65	2.07	N
YMR291W	NA	0.47	2.06	N
YER001W	MNN1	0.47	2.06	Y
YML083C	NA	0.53	2.06	N
YJR115W	NA	0.75	2.06	N
YGR041W	BUD9	0.47	2.05	N
YOR099W	KTR1	0.51	2.04	Y
YOR098C	NUP1	0.51	2.04	N
YNR050C	LYS9	0.50	2.04	N
YML088W	UFO1	0.50	2.03	N
YDR208W	MSS4	0.45	2.03	N
YDR207C	UME6	0.45	2.03	N
YOR344C	TYE7	1.09	2.02	N
YCL054W	SPB1	0.38	2.01	N
YML082W	NA	0.52	2.01	Y

Table S7. Sequenced clones from fully induced and uninduced populations.

Phenotype	Colony #	Mutation(s)	Notes
Fully induced	1	S10C	Mutation N-terminal to EH region
Fully induced	2	None	
Fully induced	3	None	
Fully induced	4	None	
Fully induced	5	None	
Fully induced	6	None	
Fully induced	7	None	
Fully induced	8	None	
Fully induced	9	None	
Fully induced	10	None	
Fully induced	11	None	
Uninduced	1	I18N, L47S, K59I	L47 is conserved across multiple bZIP subfamilies (Fig. S17)
Uninduced	2	K33N, E39V, R51G, S46X	R51 is an invariant bZIP residue
Uninduced	3	Premature stop	
Uninduced	4	Premature stop	
Uninduced	5	Q13L, R51G, frameshift	
Uninduced	6	T25S, A53V	A53 is an invariant bZIP residue
Uninduced	7	K33M, E39V, R51G, S66C	R51 is an invariant bZIP residue
Uninduced	8	R29S, N49S	R49 is an invariant bZIP residue
Uninduced	9	K30E, A53T	A53 is an invariant bZIP residue
Uninduced	10	frameshift	

Table S8. Sequenced clones with altered promoter activities. Mutants with multiple mutations were categorized either by a single mutation (if they possessed only one mutation likely to alter DNA binding affinity) or by multiple mutations (if it was difficult to determine the residue responsible for altered binding).

# Wells	Mutation(s)	Classification
10	K33E	K33E/M
10	E44G	E44G
8	R42W	R42W
5	R31S	R31S/G
5	S10G, S14P, T20S, E44G	E44G
4	R45C	R45S/C
3	R29W, E39V	R29W+E39V
2	S12P, K38E	K38E
2	R48S	R48S
2	N21I, R48S	R48S
2	S10N, Q13L, L16V, N21D, Q40R	Q40L/R
2	F22Y, K33M, K35I	K33E/M
1	N11Y, N21D, R48S	R48S
1	E44G, H63L	E44G
1	S10E, Q13L, L16V, N21D, Q40R	Q40L/R
1	Q40L, L47S	Q40L/R
1	R29G, S56N	R29G+S56N
1	R31G	R31S/G
1	E39V	E39V

SUPPLEMENTARY FIGURES

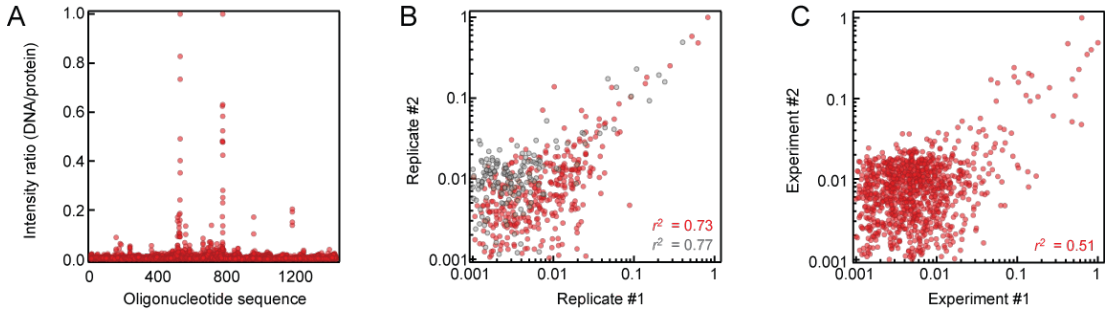


Figure S1. Hac1^{i} MITOMI 2.0 binding data. **(A)** Measured fluorescence intensity ratios as a function of oligonucleotide sequence. **(B)** Scatter plot comparing replicate measurements within experiments (experiment #1 (red) r^2 = 0.73; experiment #2 (gray) r^2 = 0.77). **(C)** Scatter plot comparing replicate measurements between experiments performed several days apart (r^2 = 0.51).

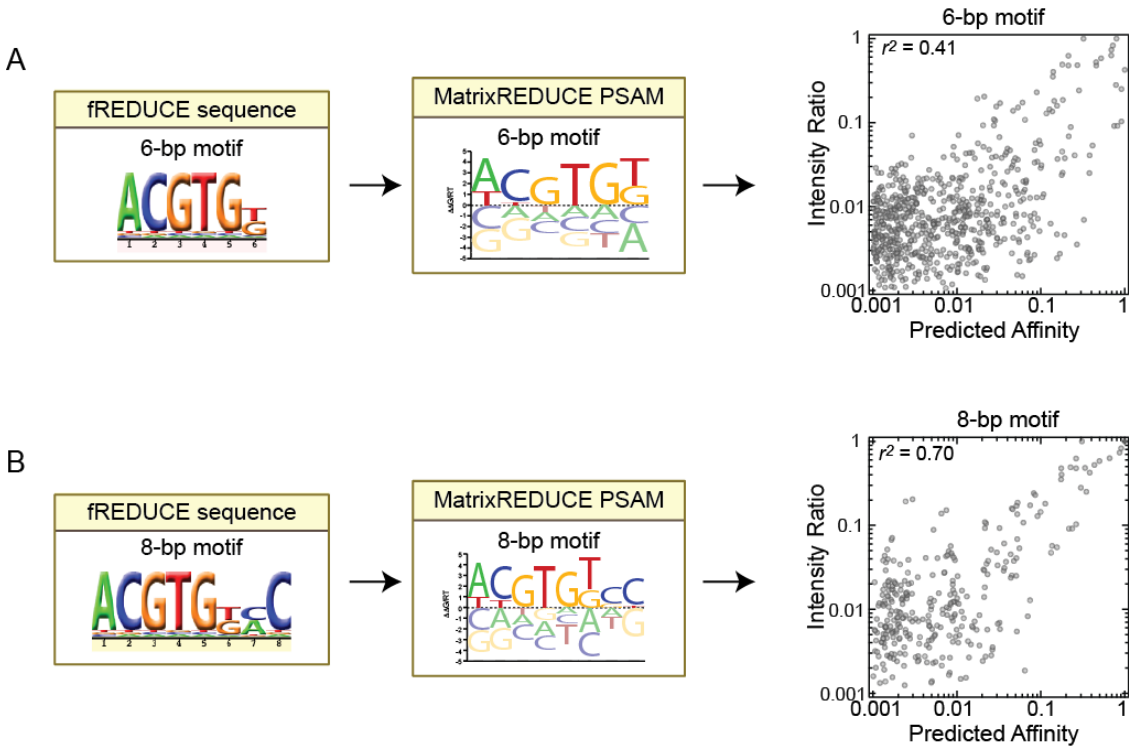


Figure S2. MITOMI 2.0 results for 6- and 8-bp motifs. **(A)** 6-bp motif results, including returned fREDUCE sequence (left), MatrixREDUCE PSAM (middle), and agreement between measured intensity ratios and predicted binding based on the 6-bp PSAM (right). **(B)** 8-bp motif results.

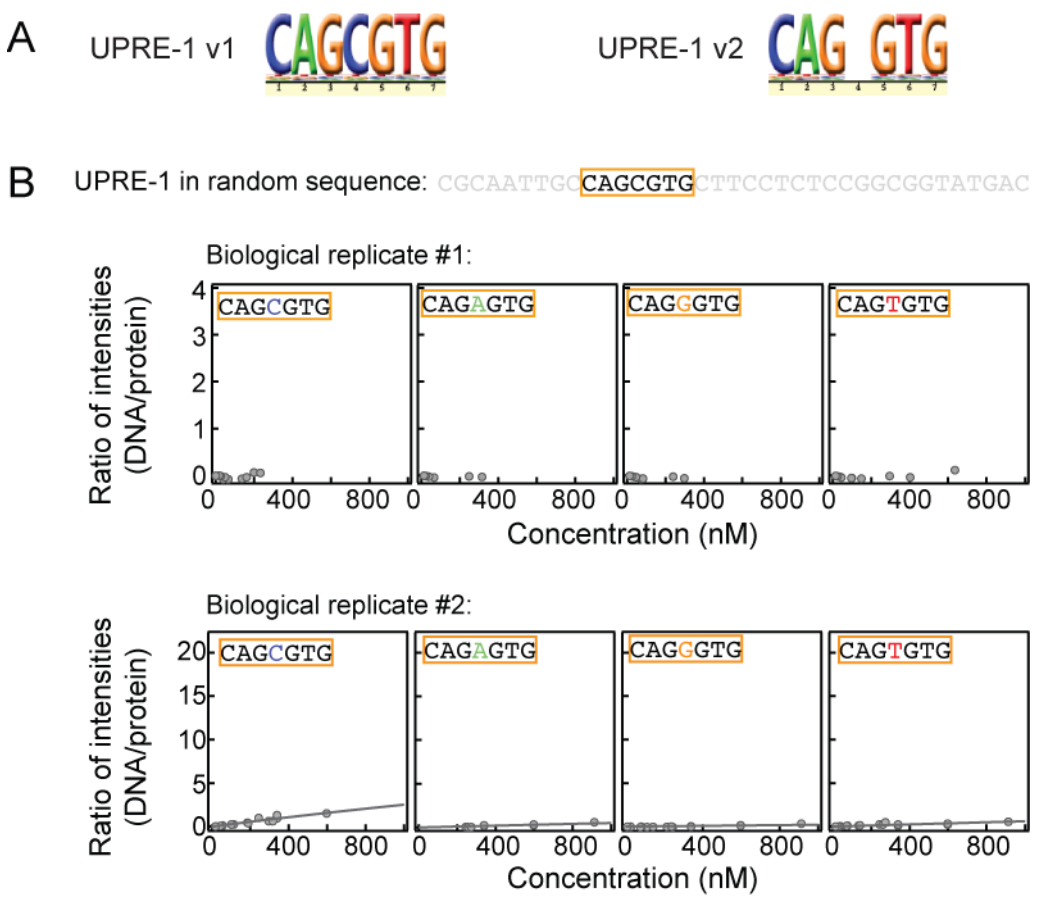
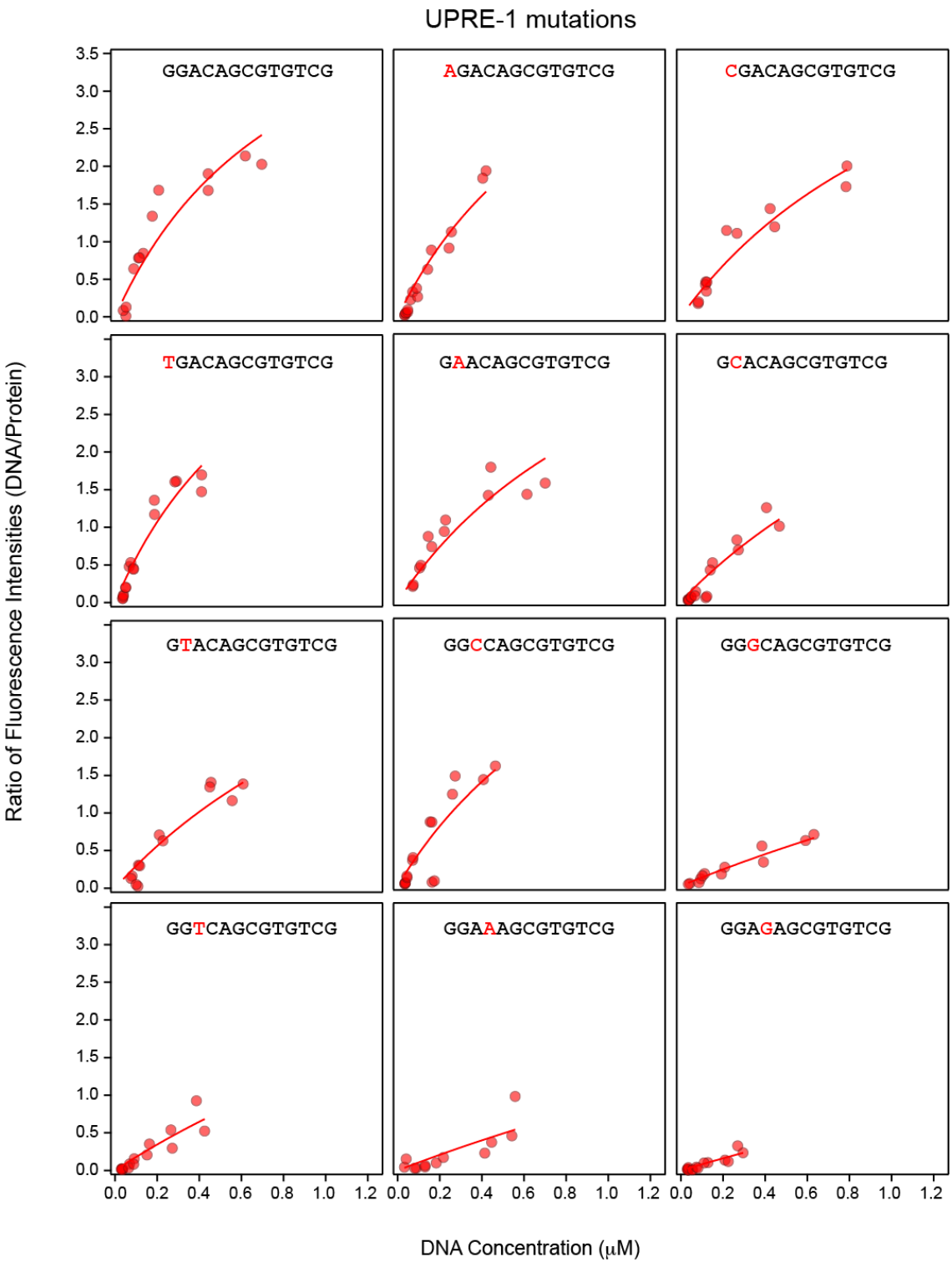
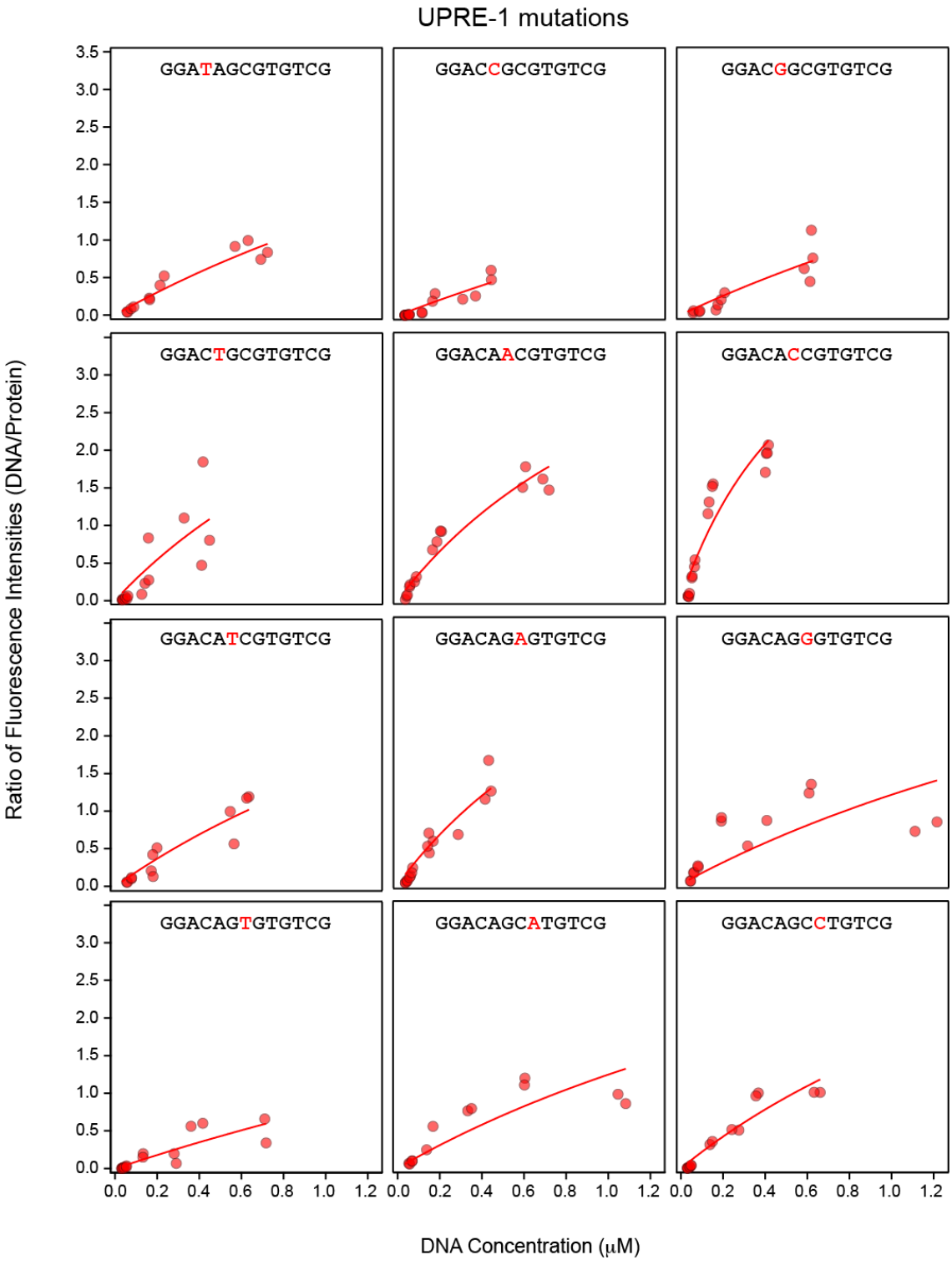
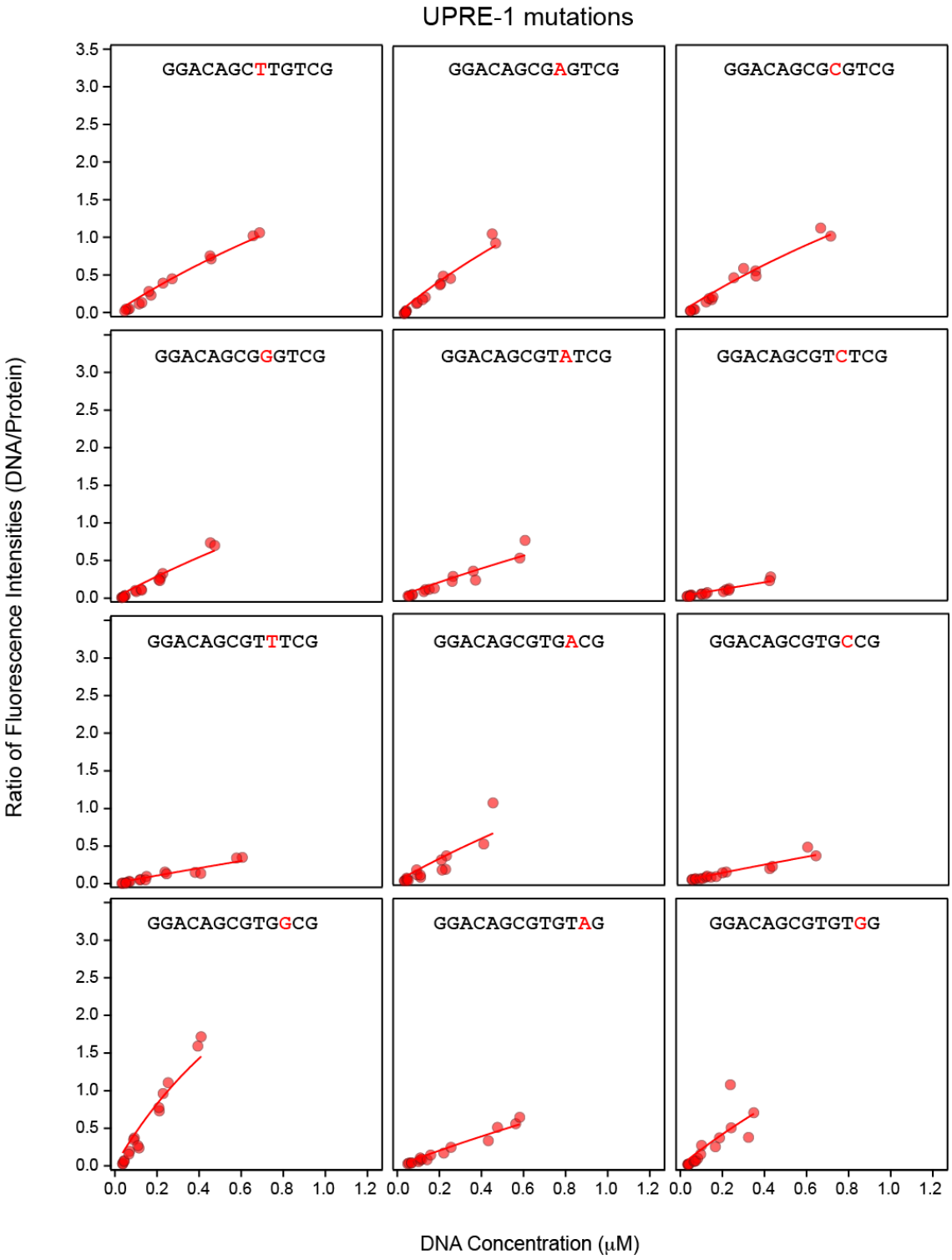


Figure S3. Experimental replicates confirm a lack of binding to any cUPRE-1 variant embedded within non-genomic MITOMI library sequence. **(A)** Alternate versions of the cUPRE-1 motif previously reported in the literature (UPRE-1 v1 and UPRE-1 v2). **(B)** Two experimental replicates showing measured fluorescence intensity ratios (grey circles) as a function of soluble DNA concentration for 4 cUPRE-1 variants (orange box) embedded within random sequence (grey).


UPRE-1 mutations
GGACAGCGTGTCG
AGACAGCGTGTCG
CGACAGCGTGTCG
TGACAGCGTGTCG
GAACAGCGTGTCG
GCACAGCGTGTCG
GTACAGCGTGTCG
GGCCAGCGTGTCG
GGGCAGCGTGTCG
GGTCAGCGTGTCG
GGAAAGCGTGTCG
GGAGAGCGTGTCG
Ratio of Fluorescence Intensities (DNA/Protein)
DNA Concentration (μM)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.2
0.4
0.6
0.8
1.2


UPRE-1 mutations
GGATAGCGTGTCG
GGACCGCGTGTCG
GGACGGCGTGTCG
GGACTGCGTGTCG
GGACAACGTGTCG
GGACACCGTGTCG
GGACATCGTGTCG
GGACAGAGTGTCG
GGACAGGGTGTCG
GGACAGTGTGTCG
GGACAGCATGTCG
GGACAGCCTGTCG
Ratio of Fluorescence Intensities (DNA/Protein)
DNA Concentration (μM)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.2
0.4
0.6
0.8
1.2


UPRE-1 mutations
GGACAGCTTGTCG
GGACAGCGAGTCG
GGACAGCGCGTCG
GGACAGCGGGTCG
GGACAGCGTATCG
GGACAGCGTCTCG
GGACAGCGTTTCG
GGACAGCGTGACG
GGACAGCGTGCCG
GGACAGCGTGGCG
GGACAGCGTGTAG
GGACAGCGTGTGG
Ratio of Fluorescence Intensities (DNA/Protein)
DNA Concentration (μM)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.2
0.4
0.6
0.8
1.2

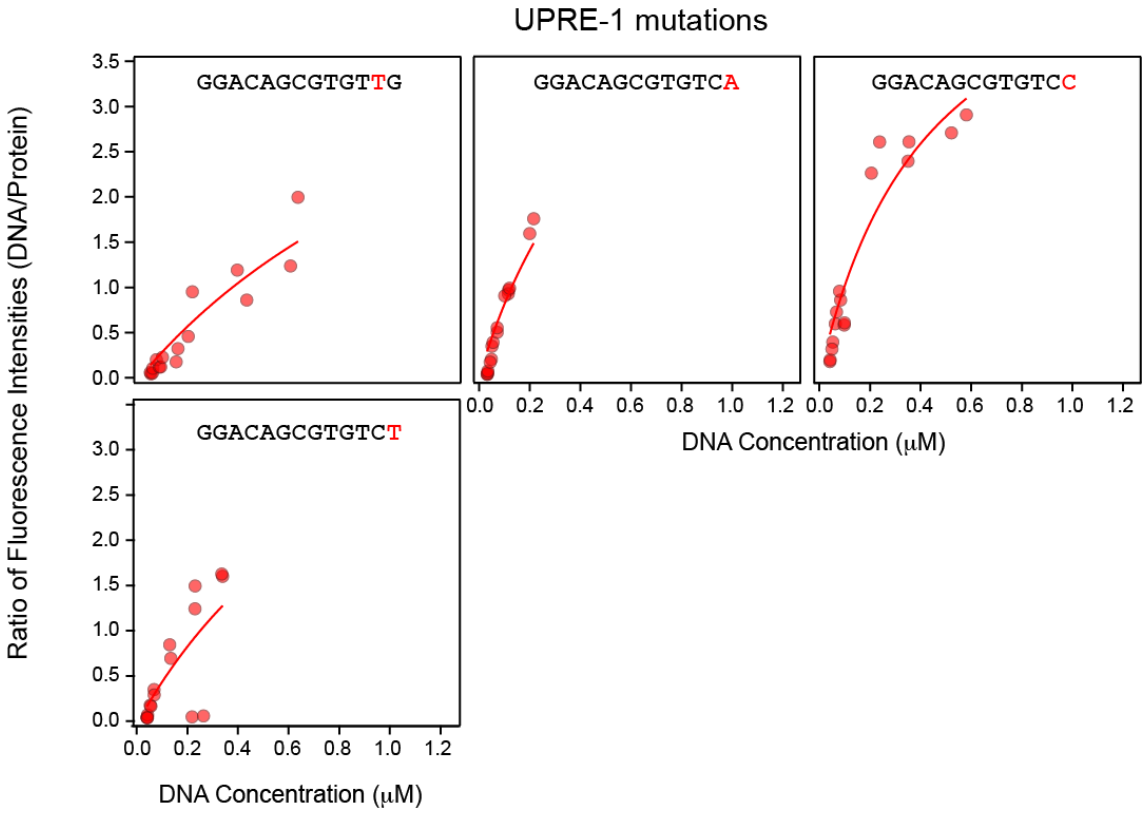


Figure S4. Concentration-dependent binding behavior for systematic mutations within xcUPRE-1 motif. For each oligonucleotide, measured fluorescence intensity ratios (DNA/Protein, red circles) are plotted as a function of soluble DNA concentration. Solid red lines show global fits to a single-site binding model.

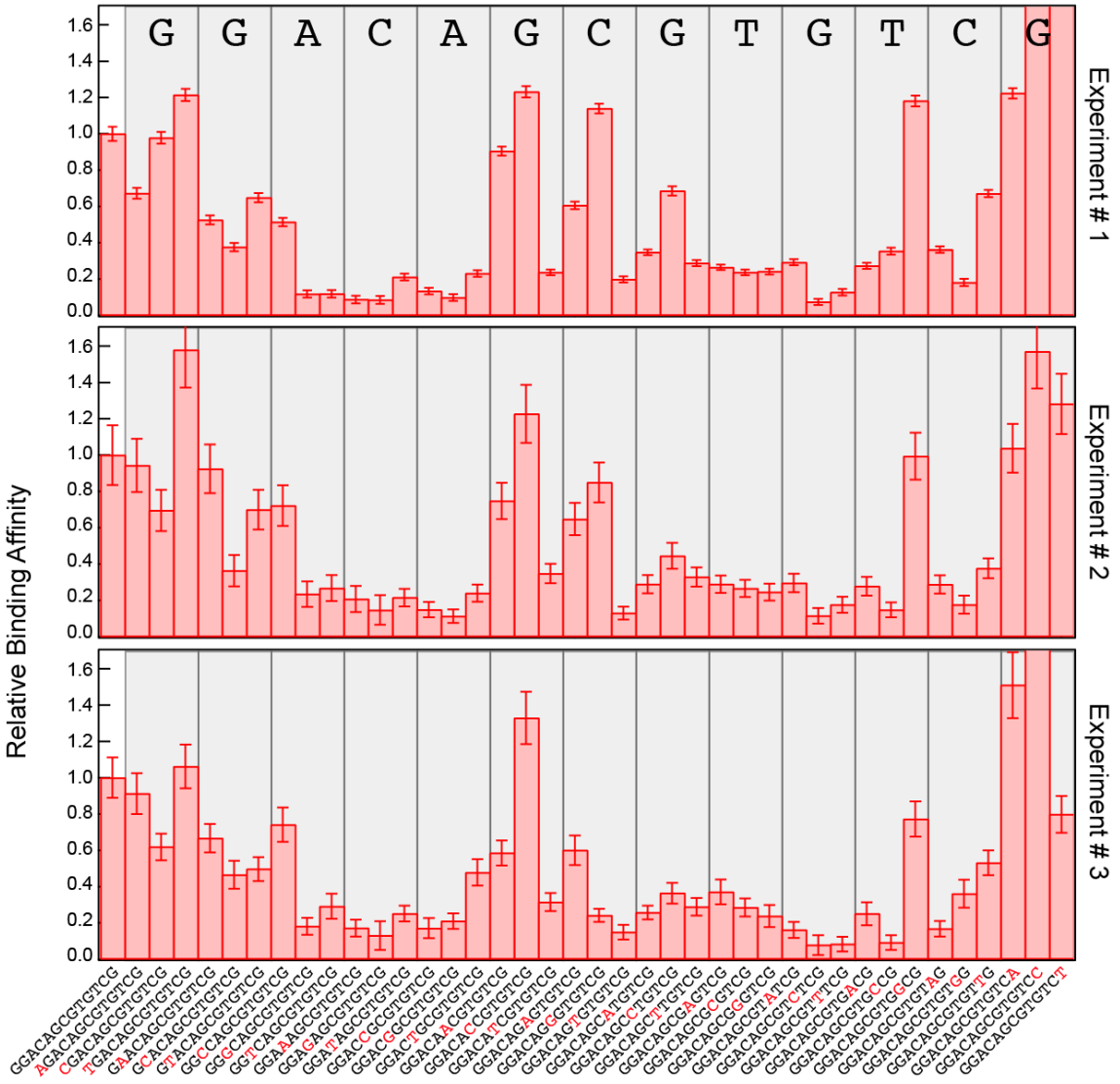
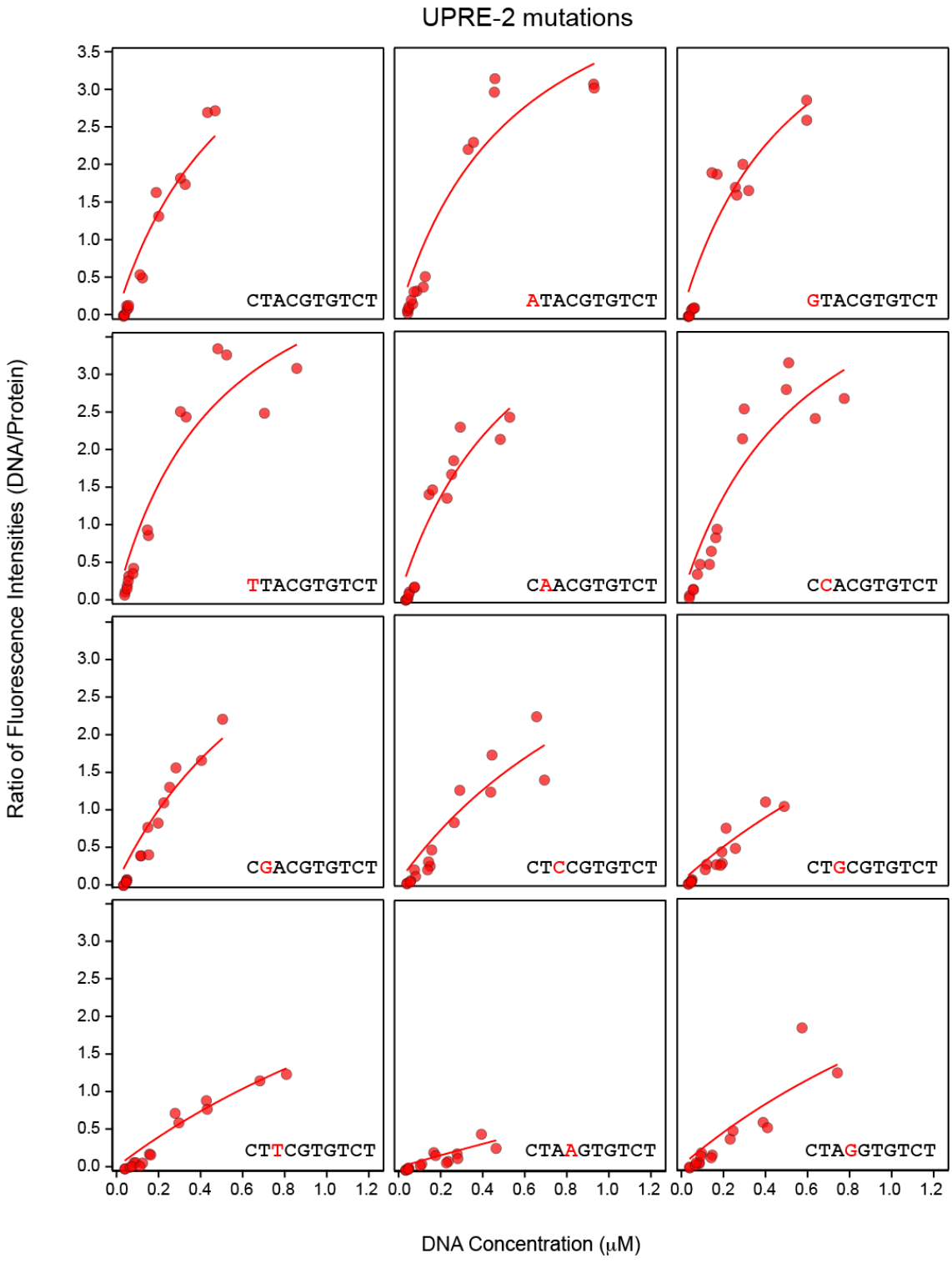
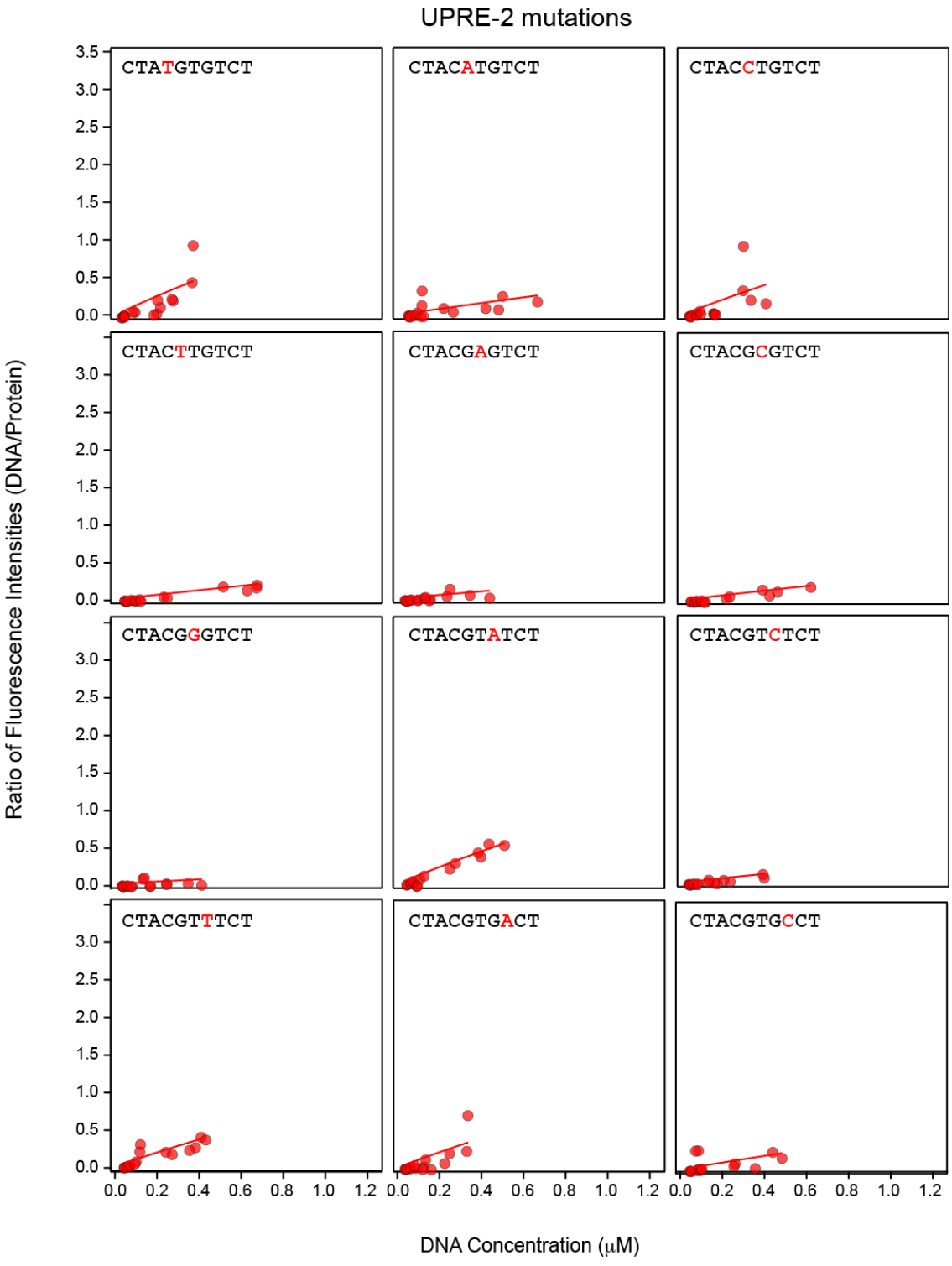


Figure S5. Relative nucleotide preferences derived from measurements of concentration-dependent binding assessing the effects of single-nucleotide substitutions at each position within the xcUPRE-1. All values are normalized relative to the measured binding affinity for the wild-type xcUPRE-1 sequence; three experimental replicates using libraries synthesized on different days and printed on different days are shown. In each case, error bars represent the error returned by from global fits to a single-site binding model.


UPRE-2 mutations
Ratio of Fluorescence Intensities (DNA/Protein)
DNA Concentration (μM)
CTACGTGTCT
ATACGTGTCT
GTACGTGTCT
TTACGTGTCT
CAACGTGTCT
CCACGTGTCT
CGACGTGTCT
CTCCGTGTCT
CTGCGTGTCT
CTTCGTGTCT
CTAAGTGTCT
CTAGGTGTCT


UPRE-2 mutations
CTATGTGTCT
CTACATGTCT
CTACCTGTCT
CTACTTGTCT
CTACGAGTCT
CTACGCGTCT
CTACGGGTCT
CTACGTATCT
CTACGTCTCT
CTACGTTTCT
CTACGTGACT
CTACGTGCCT
Ratio of Fluorescence Intensities (DNA/Protein)
DNA Concentration (μM)

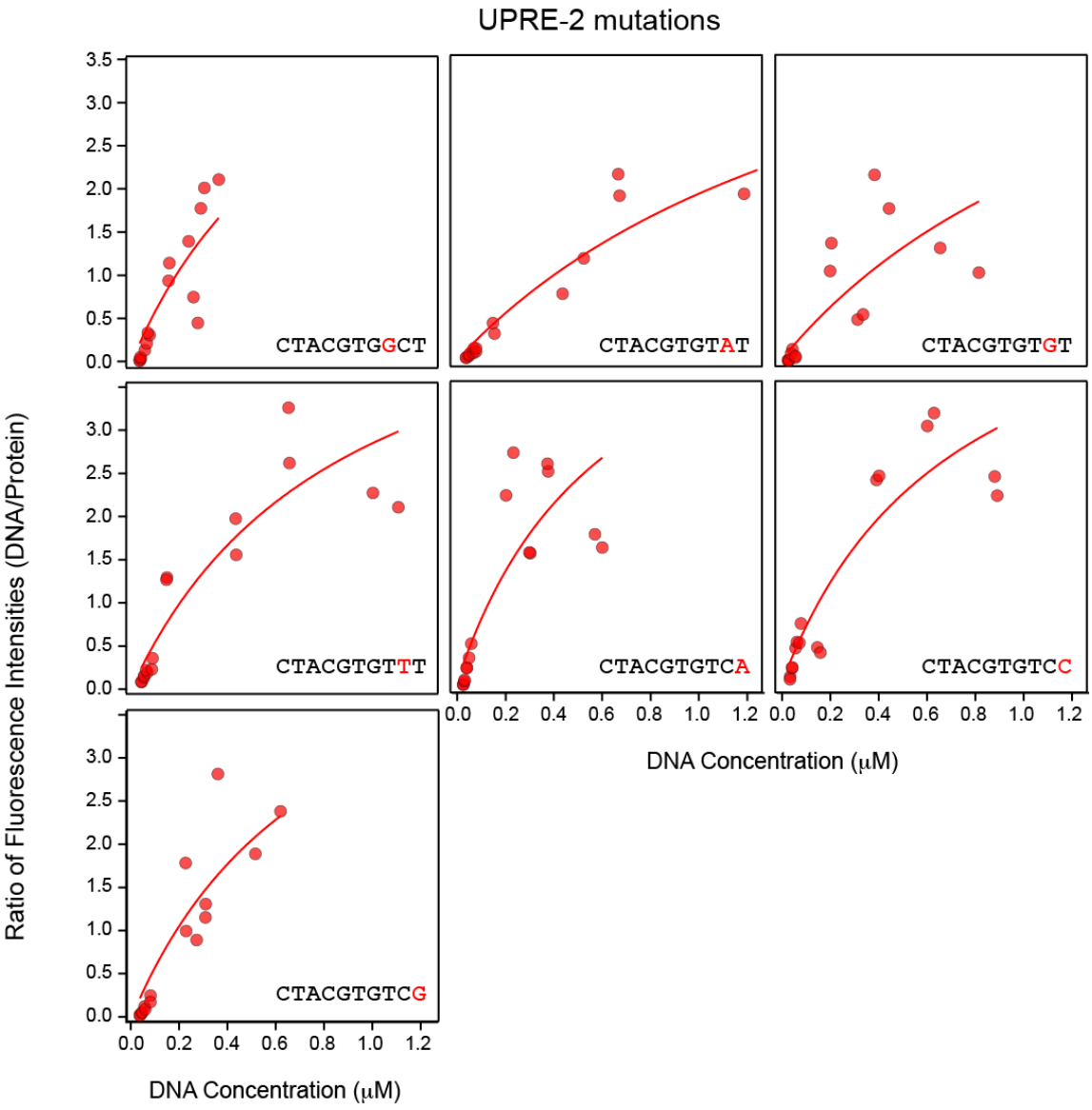


Figure S6. Concentration-dependent binding behavior for systematic mutations within the UPRE-2 target site. For each oligonucleotide, measured fluorescence intensity ratios (DNA/Protein, red circles) are plotted as a function of soluble DNA concentration. Solid red lines show global fits to a single-site binding model.

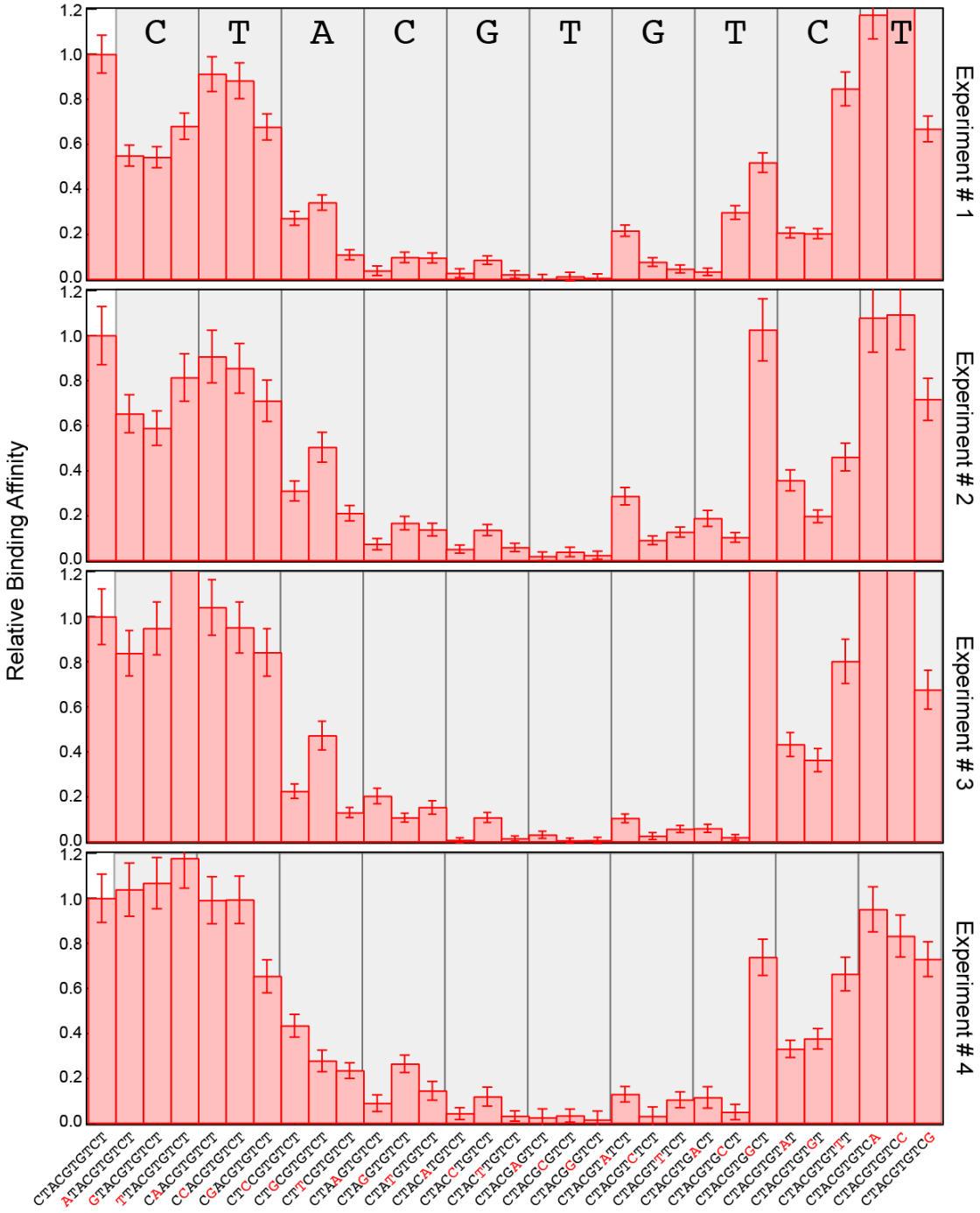


Figure S7. Relative nucleotide preferences derived from measurements of concentration-dependent binding assessing the effects of single-nucleotide substitutions at each position within the UPRE-2. All values are normalized relative to the measured binding affinity for the wild-type UPRE-2 sequence; experiments #1 and #2 were performed using the same oligonucleotide library and print. Error bars represent the error returned from global fits to a single-site binding model.

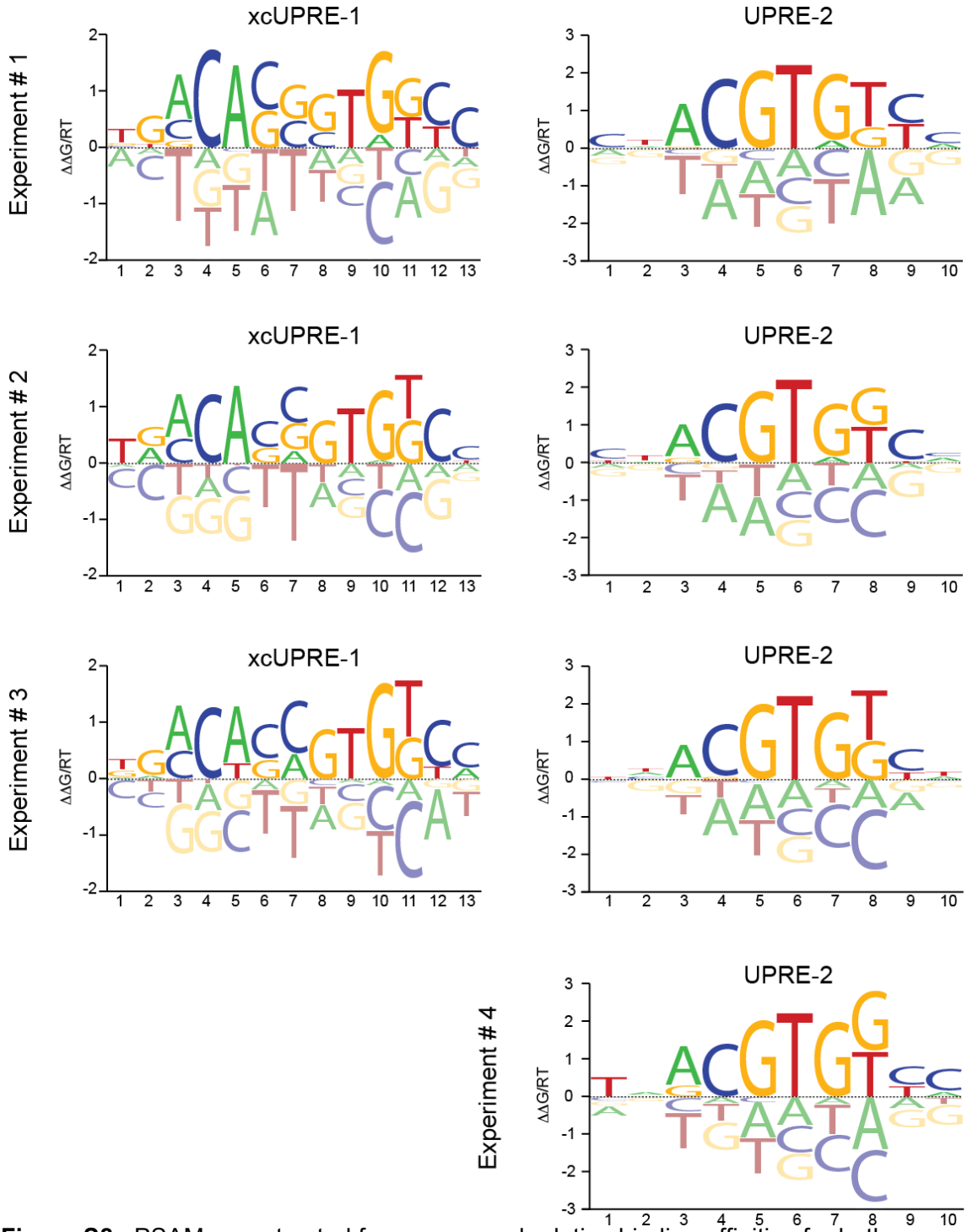


Figure S8. PSAMs constructed from measured relative binding affinities for both xcUPRE-1 (left) and UPRE-2 (right) motifs.

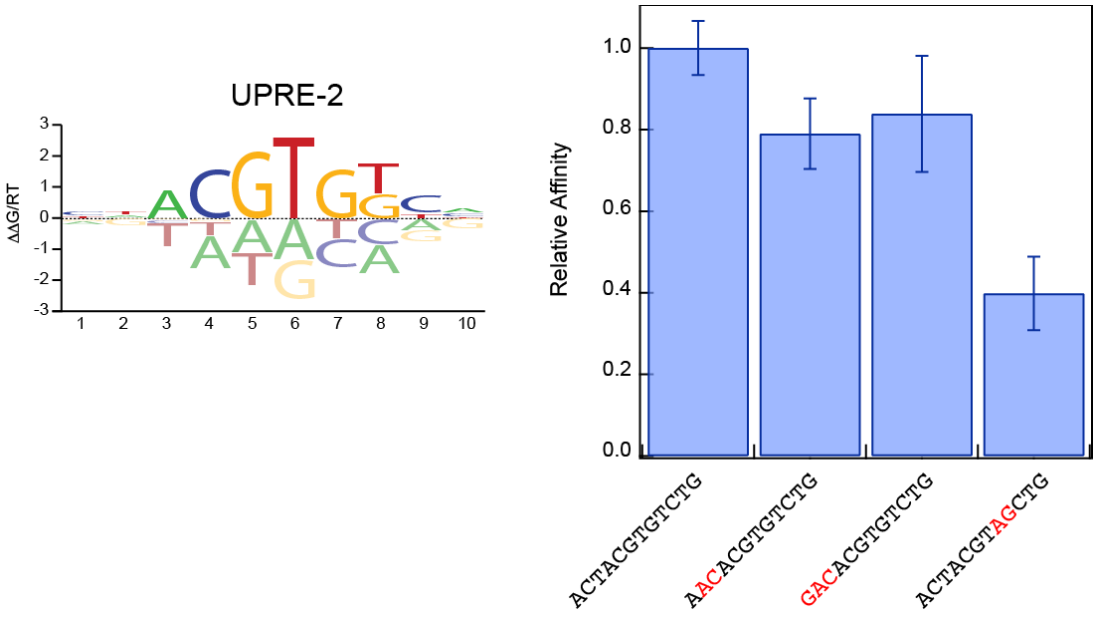


Figure S9. Creating a "supersymmetric" UPRE-2 binding site has little effect on measured UPRE-2 binding affinities. Attempts to increase binding site symmetry by mutating either 2 or 3 residues at the 5' end of the motif do not produce statistically significant changes in binding affinity (middle two columns, *5'-AACACGTGTCTG-3'* and *5'-GACACGTGTGTG-3'*). Attempts to increase binding site symmetry by mutating residues at the 3' end of the motif disrupt binding and reduce affinity ~ 2.5-fold, similar to the additive effects from each mutation individually (rightmost column, *5'-ACTACGTAGCTG-3'*).

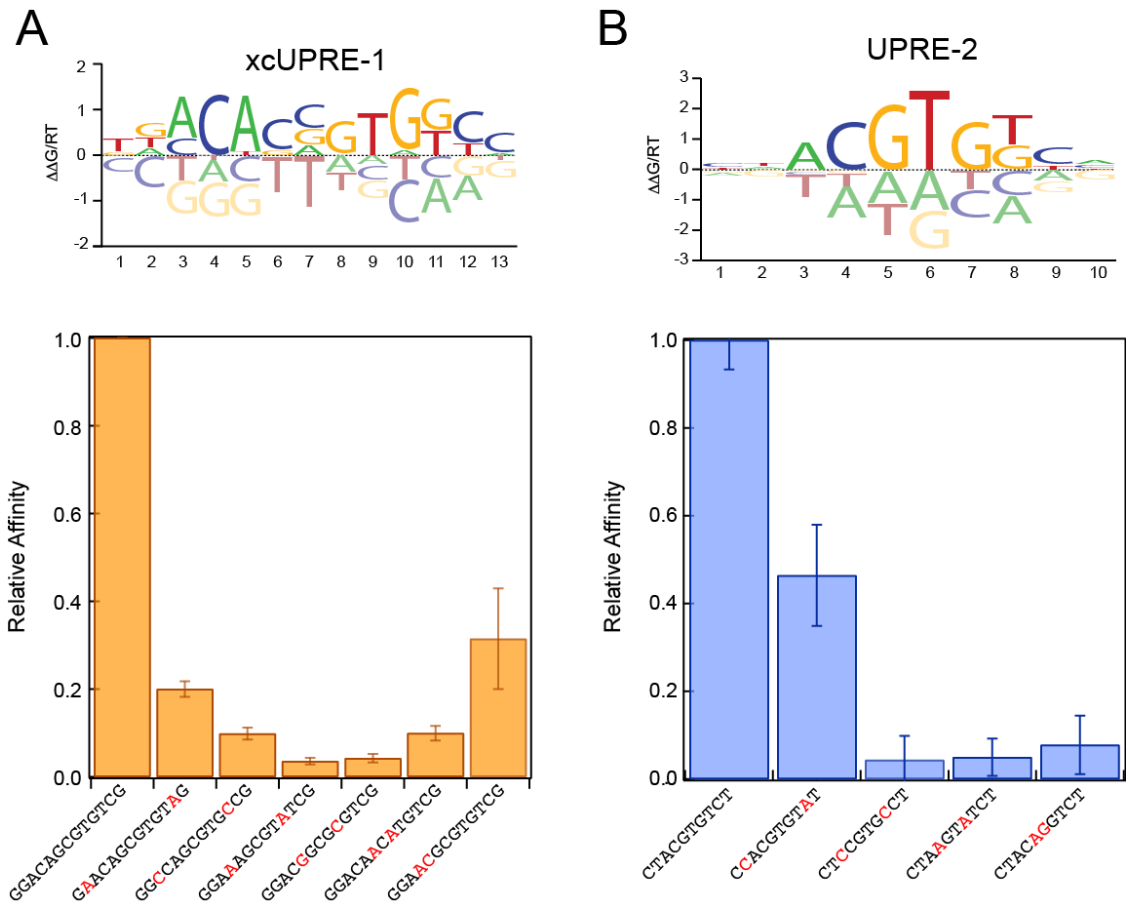


Figure S10. Effects of double nucleotide substitutions in xcUPRE-1 and UPRE-2 target sites. **(A)** Relative binding affinity measurements for the wild-type xcUPRE-1 (left-most column) and six constructs containing 2 mutations each (columns 2-7). Relative affinities were averaged over 4 independent experiments; error bars report the standard error on the mean. **(B)** Relative binding affinity measurements for the wild-type UPRE-2 (left-most column) and four constructs containing 2 mutations each (columns 2-5). As for the xcUPRE-1, relative affinities represent an average value over 4 experiments.

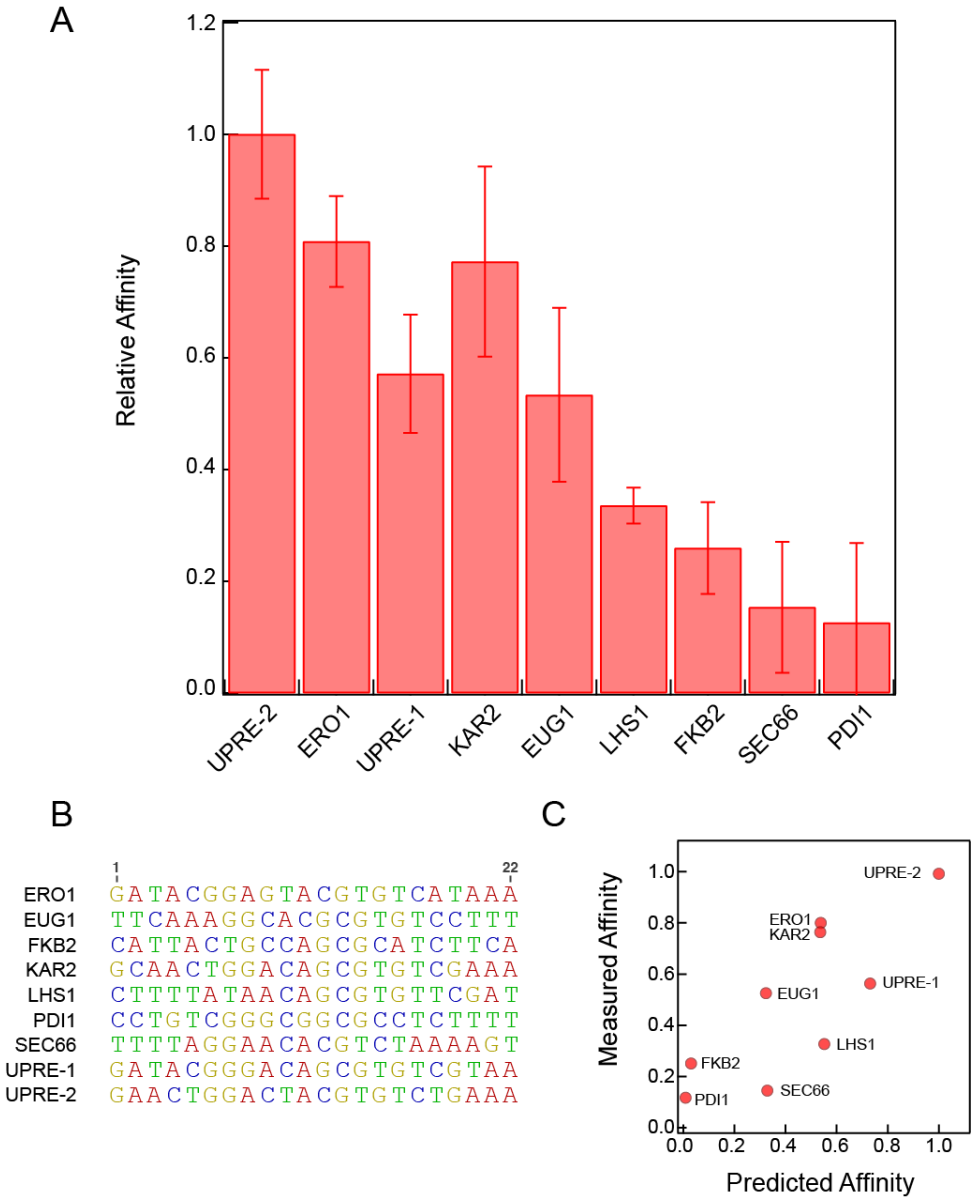


Figure S11. Measured and predicted relative affinities for xcUPRE-1, UPRE-2, and 7 known genomic UPREs (*ERO1, KAR2, EUG1*, *LHS1*, *FKB2*, *SEC66*, and *PDI1*). **(A)** Relative binding affinities, as averaged from 4 independent experiments. In each case, all affinities are normalized relative to the UPRE-2 binding affinity; error bars represent the standard error on the mean. **(B)** Genomic sequences for known UPREs within different *S. cerevisiae* promoters, shown alongside xcUPRE-1 and UPRE-2 embedded within random flanking sequences. **(C)** Comparison between relative measured affinities and PSAM-predicted affinities for each genomic UPRE. Predicted affinities are the sum of the predicted affinities for both xcUPRE-1 and UPRE-2 binding.

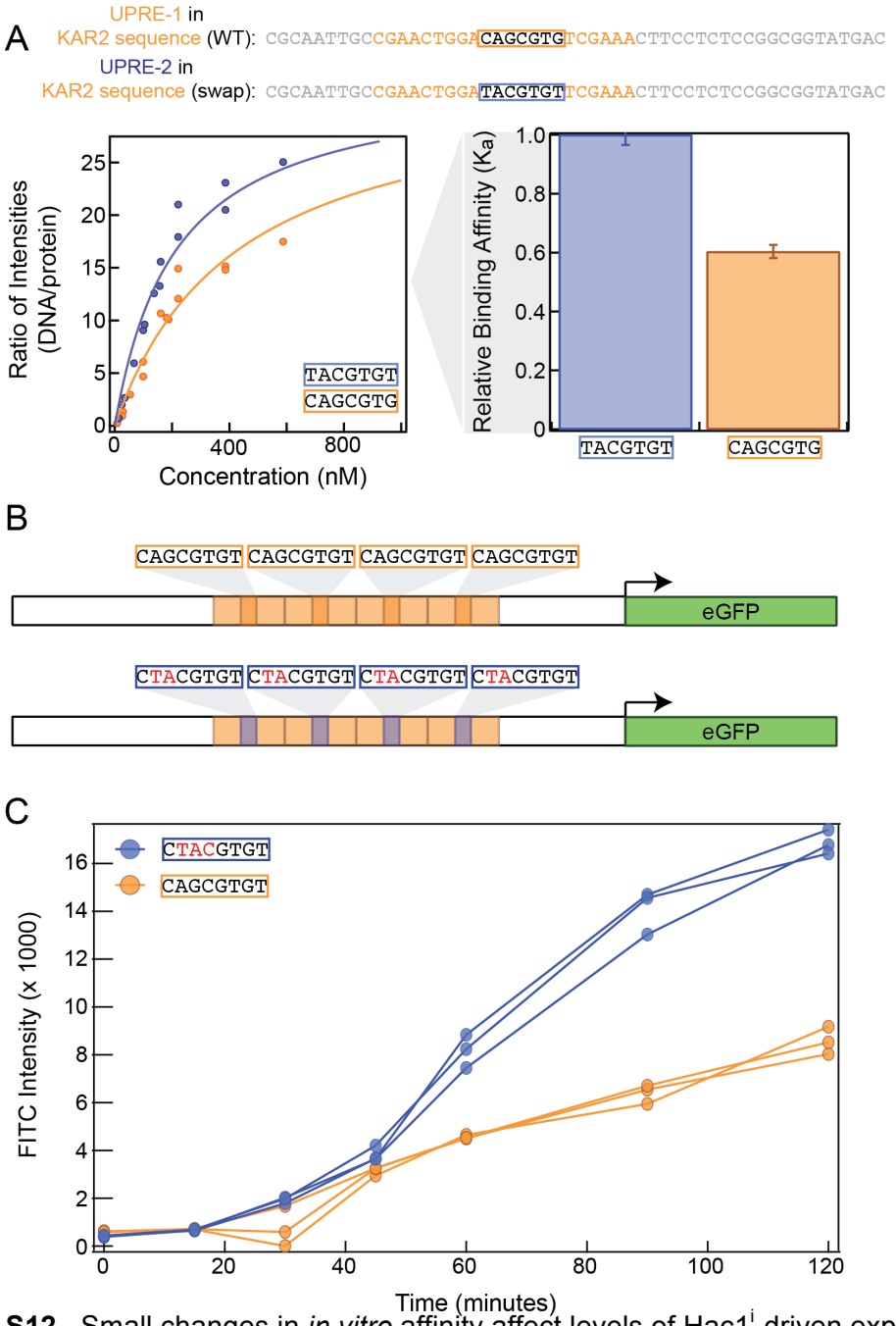


Figure S12. Small changes in *in vitro* affinity affect levels of Hac1^{i}-driven expression *in vivo*. **(A)** Fluorescence intensity ratios as a function of soluble DNA concentration for both cUPRE-1 (orange box) and UPRE-2 (blue box) motifs embedded within *KAR2* promoter flanking sequence. **(B)** Schematic representation of constructs used in reporter activity assays containing 4x repeats of either the canonical *KAR2* UPRE-1 (orange) or a mutated version of the motif designed to match the UPRE-2 (blue). **(C)** FITC intensity as a function of time for three experimental replicates showing reporter gene expression driven by the 4x UPRE-2 promoter (blue) and the 4x UPRE-1 promoter (orange).

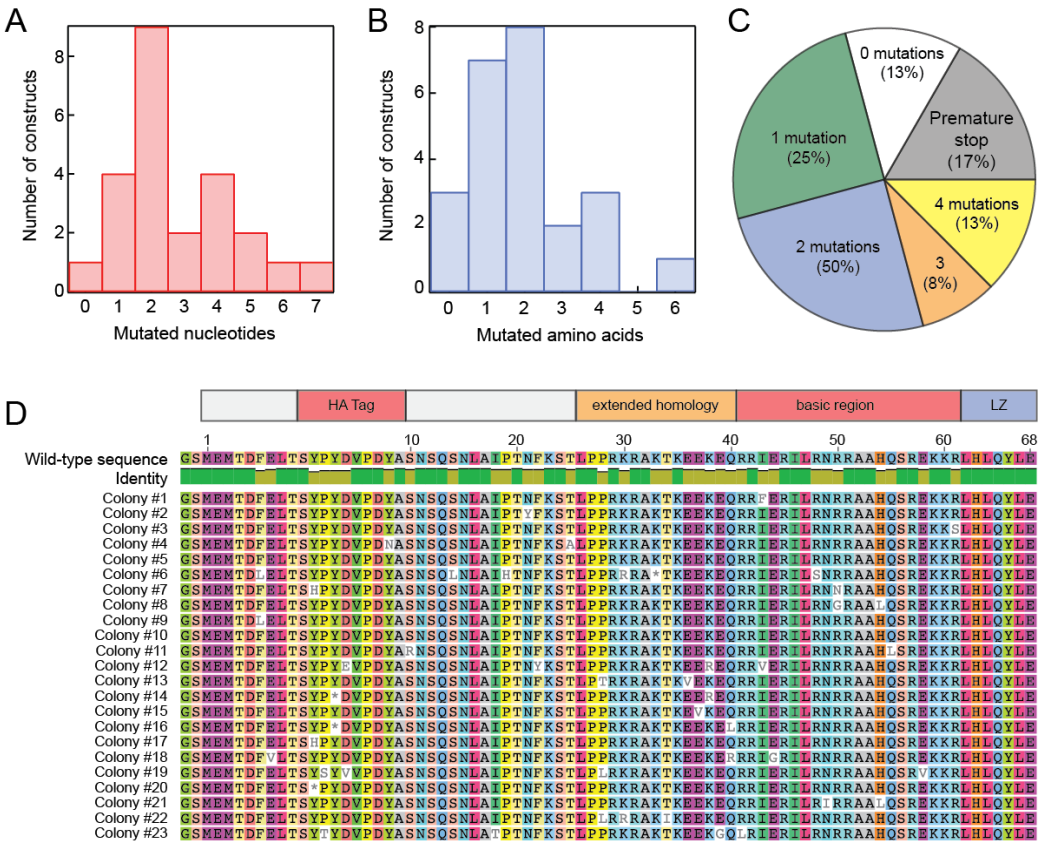


Figure S13. Sequencing results for 23 individual clones from the Hac1^{i} mutant library used within dual reporter screen. **(A)** Histogram showing the distribution of constructs containing between 0 and 7 mutated nucleotides; no constructs had greater than 7 nucleotide mutations. **(B)** Histogram showing the distribution of constructs containing between 0 and 6 amino acid substitutions; no constructs had greater than 6 amino acid substitutions. **(C)** Pie chart showing the distribution of constructs containing 0-4 mutations and premature stop codons. **(D)** Protein sequence alignment showing details of the mutated region and the distribution of sequenced mutations within this region.

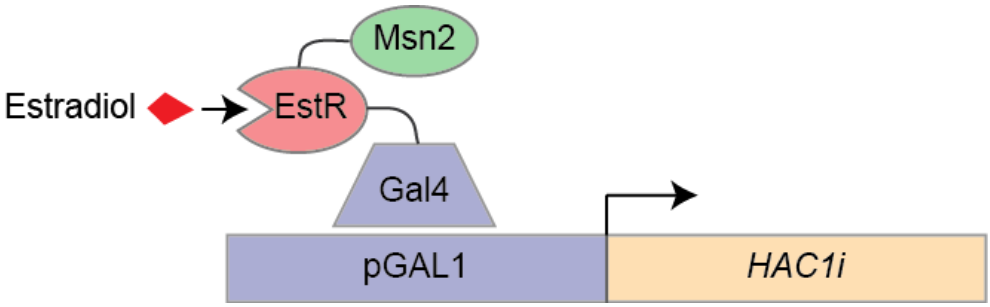


Figure S14. Estradiol-inducible ectopic expression system used for Hac1^{i} expression within dual reporter assay. Spliced Hac1 (Hac1^{i}) constructs are placed under the control of the Gal1 promoter. Transfection with an additional plasmid encoding a chimeric protein composed of the ligand binding domain of the estradiol receptor (red) linked to both the Gal4 DNA binding domain (blue) and the Msn2 activation domain (green) permits dose-dependent induction of Hac1^{i} expression upon treatment with estradiol (red diamond).

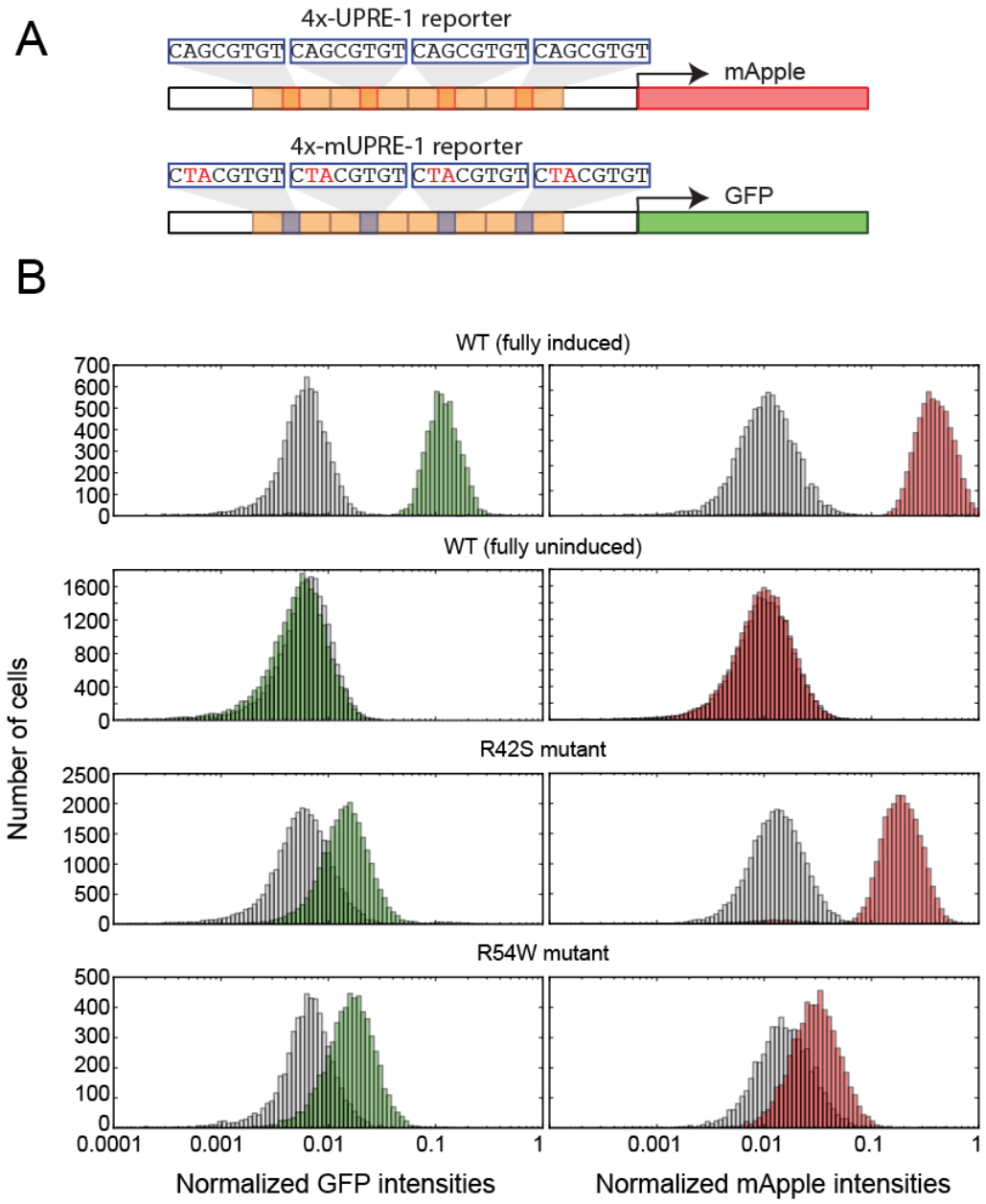


Figure S15. Flow cytometry data from the dual reporter screen designed to distinguish between xcUPRE-1 and UPRE-2 binding modes. **(A)** Dual reporter constructs. mApple expression is driven by 4 repeats of the 22-bp *KAR2* UPRE-1; GFP expression is driven by 4 repeats of a mutated version of this motif (mUPRE-1) designed to be UPRE-2-like. **(B)** Example data from wells measured via flow cytometry. Histograms of GFP intensities are shown on the left; histograms of mApple intensities are shown on the right. In each cases, grey histograms show fluorescence counts in the absence of estradiol (no $Hac1^{i}$ expression), while red and green histrograms show fluorescence counts 4 hours after addition of estradiol to induce expression of $Hac1^{i}$. Top row: fully-induced wild-type construct; second row: uninduced wild-type construct; third row: R42S construct with preferential reduction in GFP (UPRE-2 binding); bottom row: R54W construct with low affinities for both motifs, but a slight preference for UPRE-2 binding.

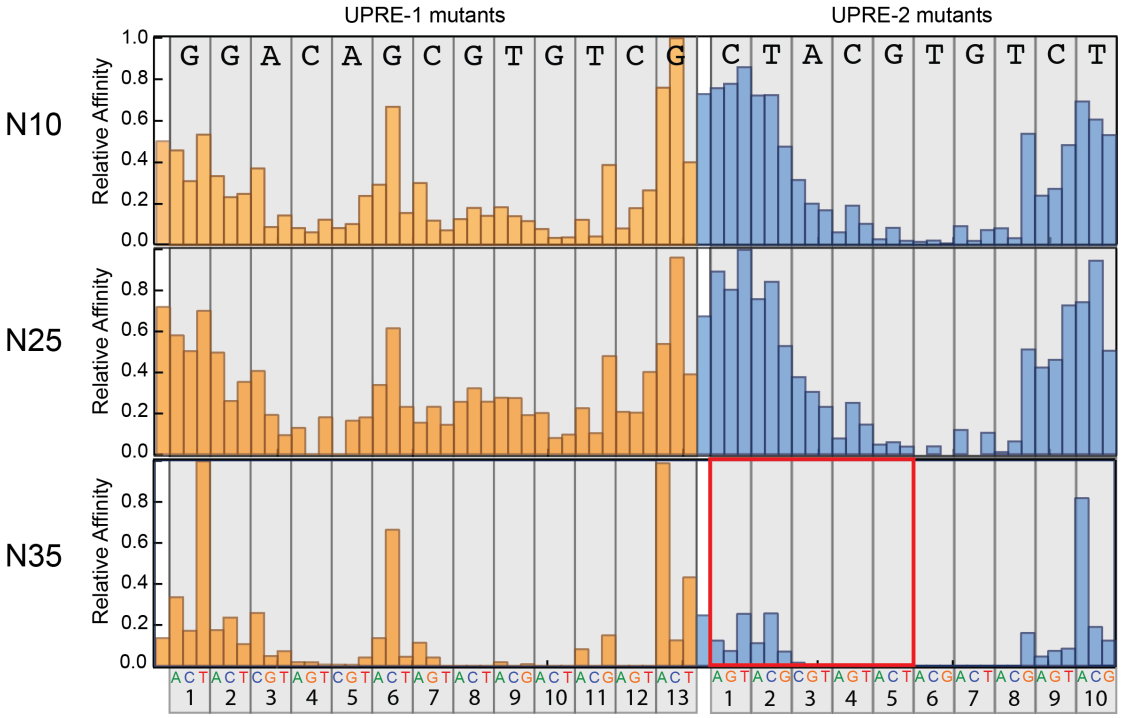


Figure S16. Relative binding affinities measured for xcUPRE-1 site and all single nucleotide substitutions (orange bars) and UPRE-2 site and all single nucleotide substitutions (blue bars) for N10 (top), N25 (middle), and N35 (bottom) constructs. The red box highlights oligonucleotides bound only weakly in the N35 mutant, indicating a decreased tolerance for substitutions at the 5' end of the motif.

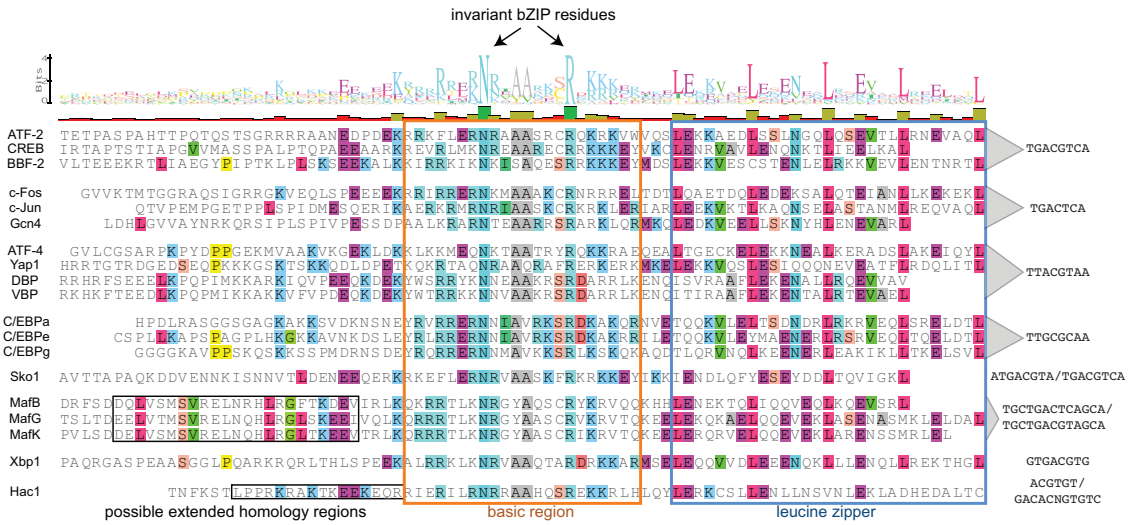


Figure S17. Alignment of bZIP family transcription factors for multiple subfamilies showing basic region (orange box), leucine zipper region (blue box), and any potential regions of extended homology N-terminal to the basic DNA binding domain (black boxes). Known DNA target sites are shown at right.

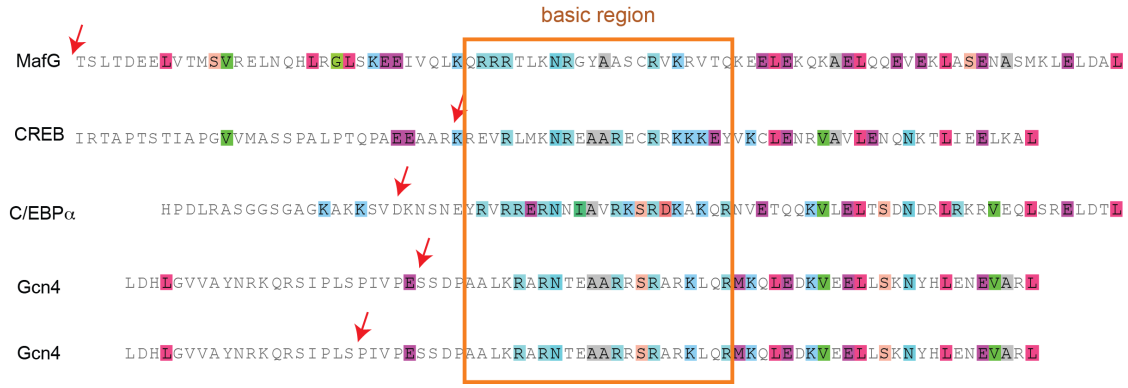


Figure S18. Protein alignment showing N-terminal truncation points for constructs used in the crystal structures for MafG[2], CREB[3], C/EBPα[4], and Gcn4[5,6], respectively.

SUPPLEMENTARY REFERENCES

1. Travers, K. J. *et al.* Functional and genomic analyses reveal an essential coordination between the unfolded protein response and ER-associated degradation. *Cell* **101**, 249–258 (2000).
2. Kurokawa, H. *et al.* Structural Basis of Alternative DNA Recognition by Maf Transcription Factors. *Mol. Cell. Biol.* **29**, 6232–6244 (2009).
3. Schumacher, M. A., Goodman, R. H. & Brennan, R. G. The Structure of a CREB bZIP·Somatostatin CRE Complex Reveals the Basis for Selective Dimerization and Divalent Cation-enhanced DNA Binding. *Journal of Biological Chemistry* **275**, 35242 –35247 (2000).
4. Miller, M., Shuman, J. D., Sebastian, T., Dauter, Z. & Johnson, P. F. Structural Basis for DNA Recognition by the Basic Region Leucine Zipper Transcription Factor CCAAT/Enhancer-binding Protein α. *Journal of Biological Chemistry* **278**, 15178 –15184 (2003).
5. Ellenberger, T. E., Brandl, C. J., Struhl, K. & Harrison, S. C. The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted α Helices: Crystal structure of the protein-DNA complex. *Cell* **71**, 1223–1237 (1992).
6. Keller, W., König, P. & Richmond, T. J. Crystal structure of a bZIP/DNA complex at 2.2 \AA: determinants of DNA specific recognition. *J. Mol. Biol* **254**, 657–667 (1995).

Chapter 7…….

Microfluidic affinity and ChIP-seq analyses converge on a deeply conserved FOXP2 binding motif in chimp and human, which enables the detection of evolutionarily novel regulatory targets

Christopher S. Nelson, Chris K. Fuller, Polly Fordyce, Alex Greninger, Hao Li and Joseph DeRisi

Author contributions:

Christopher Nelson conceived, designed, and performed MITOMI experiments, analyzed ChIP-seq and conservation data and wrote the paper. Chris K. Fuller conceived, designed, and performed ChIP-seq and ChIP-chip data analysis, and helped write the paper. Polly Fordyce designed the new pseudorandom MITOMI DNA library. Alex Greninger performed initial cDNA cloning. Joe DeRisi and Hao Li conceived and designed experiments, and helped write the paper.

Joseph L. DeRisi, Thesis Advisor

ABSTRACT

The transcription factor forkhead box P2 (FOXP2) is believed to be important in the evolution of human speech. A mutation in its DNA binding domain causes severe speech impairment. Humans have acquired two coding changes relative to the highly conserved mammalian sequence. Despite intense interest, it has remained an open question whether the human protein's DNA binding specificity and chromatin localization are conserved. Previous *in vitro* and ChIP-chip studies

have provided conflicting consensus sequences for the FOXP2 binding site. Using MITOMI2.0 microfluidic affinity assays, we describe the binding site of FOXP2 and its affinity profile in base-specific detail for all substitutions of the optimal binding site. We find that human and chimp FOXP2 have similar binding sites that are distinct from previously suggested consensus binding sites. Additionally, through analysis of FOXP2 ChIP-seq data from cultured brain cells, we find strong overrepresentation of a motif that matches our *in vitro* results and identify a set of genes likely to be direct FOXP2 targets. The FOXP2 binding sites tend to be deeply conserved, yet we identified 38 instances of evolutionarily novel sites in humans. Combined, these data present a comprehensive portrait of FOXP2 genomic targets and its binding properties. [200 words/ limit 200 words]

INTRODUCTION

FOXP2 is a transcription factor of interest in the development and evolution of language in humans (1 Scharff and Petri, 2011). Broad interest in FOXP2 began with the discovery of its linkage to autosomal dominant transmission of developmental verbal dyspraxia, a deficit of speech articulation, in the large KE family pedigree (2 Lai et al 2001). The trait was linked to a locus on chromosome 7 and eventually to a single nucleotide (residue 553) residing in the DNA binding domain of *FOXP2*, a member of the forkhead box family of sequence-specific DNA binding proteins (3 Fisher et al. 1998 and Lai et al 2001, 4 Stroud 2006, 5 Wu 2006, 6 Shu 2001). Several unrelated cases having similar phenotypes were also identified, and typically involved truncation events of the 3' end of the

FOXP2 ORF (2 Lai 2001, 7 Macdermot 2005). Affected individuals have normal intelligence and hearing but have jerky, dysfluent, and disordered speech (8 Hurst et al 1990). FOXP2 therefore offers an entry point into understanding the molecular underpinnings of the human evolution of patterned syntactic speech.

Shortly after the KE phenotype was mapped to *FOXP2*, analysis of the gene's sequence conservation revealed an interesting evolutionary history, adding another dimension to its importance in human speech. The mammalian sequence is well conserved except for two mutations in the human lineage (T303N and N325S), both N-terminal to the Zn finger domain (Figure1). Conservation analysis revealed an enhanced non-synonymous substitution rate in the hominid lineage, consistent with recent selection (9 Zhang 2002). In support of this idea, *FOXP2* locus sequences from a diverse panel of human individuals contain an excess of high-frequency derived alleles and rare intronic alleles indicative of a selective sweep in human ancestors (10 Enard et al. 2002, 11 Yu et al 2009). Animal models expressing either mutant FOXP2 or lower levels of wild type protein have borne out the involvement of FOXP2 in vocalization in mice and in zebra finches (12 Shu 2005, 13 Haesler 2007, 14 Enard 2009). These results suggest that in addition to its *developmental* role in speech, FOXP2 may have had a role in the *evolution* of speech and language.

While there exist several possible paths for the molecular evolution of FOXP2 function between ancestral primates and humans, here we investigate the simple possibilities that the selected protein mutations in the human lineage could have altered FOXP2's binding activity, driving novel targeting and functions; and/or the

genomic binding sites in humans could have changed, causing gain and loss of FOXP2 targets and modulation of targeting strength.

Evaluation of these possibilities would be aided by a thorough understanding of the FOXP2 affinity profile, yet there is surprisingly poor agreement over the identity of the FOXP2 DNA binding motif (Table1). This poor agreement may be due either to the use of different experimental techniques or reliance on prior candidate motifs identified through studies of related proteins (*e.g.* FOXP1 and FOXP3)(15,16,17 Schubert 2001, Wang 2003,Vernes et al 2007). The lack of a consistent binding site model makes it difficult to predict targets by sequence analysis, which in turn complicates the task of defining evolutionarily novel target repertoires.

Here, we clarify FOXP2's target motif using recently developed microfluidic methods that measure binding affinity of proteins to a library of different DNA sequences (18 Maerkl and Quake, 19 Fordyce et al). The resulting detailed binding site model reveals essentially identical affinity profiles for both chimp and human FOXP2 orthologs, suggesting that evolutionary differences between lineages are not due to distinct binding preferences. The derived FOXP2 motif is corroborated by an unbiased search for overrepresented motifs within FOXP2-bound ChIP-seq peaks. We find that most motif sites are deeply conserved, and they tend to be upstream of other transcription factors. However, we also find instances of evolutionarily novel FOXP2 target binding sites, including genes involved in synaptic plasticity and development, suggesting that changes in *cis* regulation may underlie FOXP2's novel functions in language.

MATERIALS AND METHODS

Cloning, mutagenesis and expression

Full length FOXP2 was initially amplified from HeLa cDNA (primers designed to isoform 1 Ensembl record CCDS5760, included in supplemental information) and placed into a PCR2.1-topo vector. Point mutations in the clone derived were corrected by site-directed mutagenesis (20 Edelheit 2009). The sequence was confirmed by Sanger sequencing and assembly with phred/phrap. Chimp and human mutant R553H FOXP2 coding versions were constructed by site-directed mutagenesis on this wildtype human plasmid. We removed the first 213 codons by PCR and added flanking promoter, polyA, and His tag sequences necessary for in vitro transcription/translation and MITOMI. The truncation removed the long polyglutamine stretch at the beginning of the protein for improved expression and solubility (Figure1). A similar truncation was previously used for EMSA studies (21 Vernes 2006), since polyglutamine stretches of over 40 residues are associated with misfolding and aggregation (22 Khare et al 2005, 23 Ross 2002). The PCR products were purified by Promega Wizard gel purification and concentrated by speedvac to ~140 ng/µl. TnT® T7 coupled reticulocyte lysate kit from Promega with the addition of 10 µM $ZnCl_2$ and was used to produce the protein of interest. We included 3 µl Fluorotect Green BODIPY charged lysine tRNA in each 75 µl translation reaction for detection of the protein by fluorescence.

MITOMI mold and device fabrication

MITOMI devices were made as described in (18, 19) Maerkl and Quake 2007 and Fordyce et al 2010. Briefly, molds for devices were fabricated on 4-inch silicon wafers by mask photolithography. Masks were based on the designs from (18 Maerkl and Quake 2007). The two layers of the device were made from RTV615 PDMS casts from the silicon molds. After partial curing the two layers were aligned and baked. The two-layer device was then aligned and bonded to a glass substrate with a printed array of the DNA library. Finished devices were run as described previously (18 Maerkl and Quake 2007, 19, 24 Fordyce et al Nat biotech 2010 and PNAS 2012).

DNA library design, synthesis and printing

The full 740 oligonucleotide pseudorandom library was designed with software from Eisen and Mintseris (25) to include all possible 65,536 8-bp DNA sequences in a relatively compact sequence space. The minimal string was then divided into 52mer oligonucleotides. We ordered these single-stranded oligonucleotides with a 3' 14-base adapter sequence to enable synthesis of the complementary strand (IDT Coralville, Iowa). A common labeled primer complementary to the common adapter (Alexa647-GTCATACCGCCGGA) was also ordered from IDT (Coralville, Iowa). The second strand was synthesized with Klenow exo- enzyme. For the targeted systematic mutation libraries, the double-stranded oligos were synthesized by the same process, and then serially diluted for final working concentrations of 0.001-2 μM DNA. Printing was carried out with silicon tips on a contact printer. Libraries were resuspended to a final concentration of 3X saline-

sodium citrate buffer, with 0.125% polyethylene glycol-6000 (Fluka) and 12.5mg/ml D-(+)-trehalose dihydrate (Fluka).

MITOMI Data Analysis

In general, we follow the analysis protocol described previously (19 Fordyce et al 2010). After running the devices, the fluorescence intensities were scanned using an arrayWoRx scanner with arrayWorx 3.0.3 software suite release 1. Fluorescence data for bound DNA and protein at the button valve and free DNA in the DNA chamber were extracted from the scanned devices with Genepix 6.1.

For initial IUPAC motifs we used fREDUCE software, which screens all degenerate Nmers in a sequence library for their Pearson correlation to associated binding scores (26 Wu et al.2007). Using ratios of protein to DNA signal at the button valve, fREDUCE was run for 6mers through 9mers with up to 3 degenerate positions. The data are not normalized for comparison of binding strengths between wildtype and mutant constructs, but normalized for all other analysis. The top scoring IUPAC sequences by correlation and p-value with respect to the whole dataset were then used as input “seeds” for MatrixREDUCE. Given a seed sequence, MartixREDUCE searches for a local optimum position specific affinity matrix (PSAM) that best fits the measured binding data (27 Foat et al. 2006). These matrices were then scored against the whole dataset by Pearson’s correlation between their observed and expected occupancies. PSAM motif logos were made with AffinityLogo software (27 Foat et al. 2006).

Binding curves were fit to a hyperbolic saturation curve with global nonlinear regression in Graphpad Prizm 4.00. A dilution series of the labeled primer flowed onto the DNA chambers was used as a standard curve to calibrate the relationship of Alexa-fluor signal to free DNA concentration in the DNA storage chamber on the devices.

Chromatin IP data

Processed ChIP-seq data from the Myers lab at Hudson Alpha was downloaded from the ENCDOE portal of the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg18&g=wgEncodeHaibTfbs). The data were derived from anti-FOXP2 (C-terminus) antibody pull-down sequence libraries from PFSK-1 and SK-N-MC cells. Samples were cross-linked and sheared chromatin was compared to libraries prepared without any antibody pulldown. The antibody epitope was the C-terminal 127 amino acids of FOXP2 (28 Marticke thesis). We used the peaks called by the Myers lab using QuEST, which collapses ChIP-seq signal from both strands of DNA and then calculates a fold enrichment of the peaks over the no IP control (29 Valouev et al). There were two biological replicates in each cell line. We used the function findOverlappingPeaks in the R Bioconductor ChIPpeakAnno package (http://www.bioconductor.org/packages/release/bioc/html/ChIPpeakAnno.html) to first merge the replicate peaks within data from each cell line, then to merge these to form a set of 71 high-confidence peak sequences across all samples from both cell lines. The peaks' positions relative to the nearest genes, regardless of gene orientation, were annotated using the annotatePeakInBatch

function of ChIPpeakAnno for genome build NCBI36. We determined GO term enrichment using the getEnrichedGO function of ChIPpeakAnno with maximum p-value of 0.05 after adjusting for multiple testing (30 Benjamini and Hochberg 1995).

Motif searching in ChIP-seq peaks

From the set of 71 high-confidence peaks, we extracted the genome coordinates and added either 50 or 200 extra nucleotides on each end. These sequences were passed to MEME version 4.3.0, which output PWMs ranked by their E-values for representation in the set of positive sequences (31 Bailey and Elkan 1994). The input parameters were a minimum motif width of 8 bp, a maximum motif width of 50 bp, a minimum of two sites, a maximum of 71 sites, and an E-value cutoff of 1E-50.

We also used the MITOMI-derived 7mer PSAM to score motifs within the 71 high-confidence peaks. For this analysis, we calculated the predicted occupancy ratio over the optimal sequence for 7mer windows across the entire oligonucleotide sequence (27 Foat 2006), and then compared the score for the highest scoring window with the distribution of scores for all 7mers. We identified candidate target sites of interest by using a score threshold of 0.06, which returns the top 0.1% of 7mer scores.

Conservation analysis

We hypothesized that FOXP2 motifs inside the high confidence peaks would exhibit elevated conservation relative to the surrounding sequence. Using the best PWM from our MEME analysis, we searched the peak regions for FOXP2 binding sites using TAMO v1.0 (32 Gordon et al. 2005) to identify predicted binding sites. We selected a threshold of 90% of the maximum bit score to yield approximately one FOXP2 motif per ChIP-seq peak. We then determined conservation scores for windows extending 100 bp upstream and downstream of each peak using the UCSC phastCons44WayPrimate track (http://hgdownload.cse.ucsc.edu/goldenpath/hg18/phastCons44way/primates/). We used these to compute both an ensemble average of conservation and the principal components of conservation (using the R prcomp package) in the region centered on each predicted TFBS.

To find novel FOXP2 targets among the human ChIP-seq peaks, we searched the merged peaks from the two biological replicates available for each cell line. From a total of 1554 peaks, we identified 472 that contain at least one instance of the core 'TGTTTAC' FOXP2 motif identified by both our MITOMI and ChIP-seq analyses. Of these, 56 contained sites with a substantial reduction in predicted FOXP2 affinity (50% or less of maximum bit score) between human and chimp sequences. By analyzing the multiz44way alignment of these sites across human, chimp, gorilla, rhesus, marmoset, tarsier, mouse lemur, and bushbaby, we identified sites for which the changes are unique to the human lineage.

RESULTS

Human R553H mutant shows no binding activity

Previous electromobility shift assay (EMSA) studies did not detect binding of the R553H mutant to an SV40 sequence (21 Vernes 2006). These results are consistent with two possibilities: the mutant could lack DNA binding activity, or the mutant could have altered target site specificity. To distinguish between these possibilities, we used a microfluidic binding assay (MITOMI 2.0) to search for binding interactions between the mutant protein and a DNA library containing all possible 8 bp sequences. In brief, MITOMI 2.0 experiments measure affinities between a single BODIPY-labeled transcription factor and many Alexa-647 or Cy5-labeled DNA sequences in parallel; the measured DNA signal intensity normalized by the protein signal intensity provides a measure of the fractional protein occupancy at a given DNA concentration. Truncated human R553H protein gives essentially zero protein occupancy signal for all assayed sequences (Figure 2a). Our data therefore suggest that R553H has lost all DNA binding activity, not just the ability to bind its normal motif.

Chimp and human proteins produce similar patterns of binding

In contrast to the R553H mutant, truncated chimp and human FOXP2 proteins give a distribution with a tail of protein occupancy signal indicative of strong binding to a subset of DNA sequences (Figure 2a). Comparing the binding pattern of chimp and human truncated FOXP2 protein, it is clear that some probes are repeatedly bound, e.g. oligonucleotide #175, while most

oligonucleotides exhibit very low binding (Figure 2a, b). The binding patterns for the two proteins are very similar (Pearson's r^2 of 0.85, Figure 2b).

Chimp and human orthologs bind similar motifs

Identifying the preferred short subsequences that correlate with binding to the library of 52mer DNA sequences requires analysis (19 Fordyce 2010). To identify these target sites, we first used fREDUCE, which identifies preferred motifs based on the correlation between measured binding intensity and the presence of subsequences within each oligonucleotide(26 Wu 2007), and searched for candidate motifs between 6 and 9 nucleotides in length. fREDUCE returns lists of degenerate consensus sequences ranked by their correlation to the observed pattern of binding to the DNA library. To derive the effects of nucleotide substitutions at each position within these target sites, we subsequently used MatrixREDUCE, which fits a local optimum PSAM to the observed pattern of binding(27 Foat et al 2006). Table S1 lists preferred sequences obtained from analysis of 4 aggregated experiments for each protein (Table S2 lists predictions from individual experiments). As expected, the similar binding patterns observed for the chimp and human proteins produce similar enriched motifs (Table S1, Figure 2C and D). In both cases, the top motifs of different lengths are essentially nested versions of each other, each containing a core TGTTKAC IUPAC sequence. In summary, the chimp and human FOXP2 bind similar DNA oligonucleotides in our library and appear to prefer very similar motifs.

Systematic mutation of the binding motif provides base-specific affinity information

To experimentally confirm our prediction that chimp and human FOXP2 binding preferences are the same at the single nucleotide level, and to explore the effects of flanking nucleotides on affinity, we measured affinities for FOXP2 constructs interacting with a series of oligonucleotides containing single-nucleotide substitutions. For this targeted binding curve library, we chose to use 13 bp containing a candidate high-affinity binding site present within a strongly bound oligonucleotide (#175) as a reference sequence. We then designed 39 DNA sequences with all possible point mutations of this 13 bp sequence within the context of the larger unchanged oligo (full DNA sequences in Table S3). We programmed the MITOMI device with a dilution series of each oligonucleotide and measured binding over the series. These experiments allowed us to calculate an apparent K_a by nonlinear regression of the binding curve for each oligonucleotide (18 Maerkl and Quake 2007).

Figure 3a plots the fold-change in the K_as for each motif variant for both truncated chimp and human versions of FOXP2 derived from analysis of individual binding curves (Supplementary Figures 4,5). The bulk of the sequence specificity lies in a 7-base pair core motif, with relatively minor contributions outside of that core. Although a number of point variants (*e.g.* T*A*TTTAC and TGTTTA*T*) are permissive for binding, with K_as only 3-fold lower than the optimal sequence, other point variants (*e.g.* TGTT*A*AC) are clearly disfavored, with K_as over 100-fold lower than the optimal sequence. Taking these measurements

together, we constructed an improved position-specific affinity matrix (PSAM) that accurately reflects the experimentally observed effects of each point mutation at each position (Figure 3b and 3c).

In agreement with the pseudorandom library measurements, we find that the pattern of motif affinity is very similar across the two proteins (Student's t-test on the mean and standard deviation of the separate experiments, $p > 0.05$), confirming that there is essentially no difference in binding preferences between species. In addition, these derived motifs are in agreement with the motifs obtained via pseudorandom library measurements, except for slightly less G/T degeneracy at the 5^{th} position in the 7mer (8^{th} position in the logo in Figure 3). These motifs represent the most detailed *in vitro* description of the specificities of the FOXP2 binding sites to date, and are distinct from all of the previously reported FOXP2 motifs (table 1).

MITOMI results match an independently-derived FOXP2 motif from ChIP-seq data

In parallel with our MITOMI efforts, we analyzed *in vivo* FOXP2 DNA binding data from human neuronal cell lines from the Myers lab released to the public as part of the ENCODE consortium (28 Marticke thesis, 33 Schroeder and Myers 2008). To study only the most reproducible ChIP-seq signal enrichment peaks, we first identified overlapping ChIP-seq signal peaks within the biological replicates for each cell line, yielding 1238 overlapping peaks (out of 5111 total peaks) for the PFSK1 cells, and 316 overlapping peaks (out of 615 total peaks) for the SK-N-

MC cells. Next, we narrowed this set to consider only those peaks that were shared between cell lines, yielding 71 high-confidence peaks.

To these 71 ChIP-seq peaks, we added 50 bp of the flanking genomic sequence and searched for enriched sequence motifs using MEME (31 Bailey and Elkan 1994). The top position weight matrix returned (E-value = 4.5E-82) is very similar to that found using our MITOMI device (motif matrix logo Figure 4A, compare with Figure 3B). When including a wider sequence window of 200 bp around each ChIP-seq peak, MEME returns 73 instances of a similar motif (E-value = 2.6E-51, Figure 4D), but also identifies 55 instances of a long putative homopolymer G/C stretch (E-value = 2.4E-56, Supplementary Figure 1). Among all high confidence peaks 47/71 (63%) contain at least one instance of our optimum predicted motif string, and 65/71 (92%) peaks contain a local PSAM motif score within the top 0.1% of all possible 7mers (Table 2 and Table S4).

High-confidence FOXP2 peaks have a stereotyped position and flanking sequence bias

To better understand the regulatory relevance of these candidate FOXP2 sites, we investigated their location relative to nearby genes. The FOXP2 sites tend to cluster near the start of the closest gene model, with over half of the ChIP peaks occurring within 1kb of a TSS (Figure 4B). Additionally, nucleotide bias calculations across the regions flanking the FOXP2 binding site revealed a G/C bias on both sides of the FOXP2 motif instance (Supplementary Figure 2). This might explain the low information content G/C biased sequences identified in our

MEME searches over the region surrounding ChIP-seq peaks (Supplementary Figure 1).

Characterization of predicted FOXP2 target genes

To functionally characterize this high confidence set of sites we mapped nearby genes. 59 genes are within 5 kb of the 71 consistent ChIP sites (Table 3). Bioconductor GO term ontology of these nearby genes returned several terms relating to other transcriptional regulators with strong p-values, suggesting that FOXP2 tends to target other transcription factors (Table 4). Targets in the list that fit this description are *ZBTB16, NFIA, TBL1X, ZNF395, CITED2, JUNB, CBX7, FOXP1*, *FOXK1*, *NR3C1* (glucocorticoid receptor), and an alternative transcript of *FOXP*2 itself. This last target site is near the start of the non-coding *FOXP2* transcript NR_033766.1, annotated as a candidate for nonsense mediated decay. The enrichment of transcription factors in FOXP2's putative target repertoire suggests that FOXP2 could act as a master regulator during development, perhaps with feedback on the expression on FOXP2 itself.

Beyond the unbiased set of high confidence sites, we searched for FOXP2 ChIP-seq peaks upstream of candidate targets, including previously suggested binding partners that we did not detect in our higher stringency list. Potential interacting forkhead box proteins *FOXP4* and *FOXJ2* are the closest annotated genes to two intergenic peaks 13-14 kb upstream with deeply conserved optimal motif sequences(37 Chokas et al. 2010). In addition, *HDAC2*, (encoding a histone deacetylase that interacts with FOXP2 (37 Chokas et al. 2010)), has two

upstream peaks. We find two strong FOXP2 localization peaks within intronic sequence of the gene that encodes CTBP1, which complexes with FOXP2 in yeast 2-hybrid and CoIP assays (38 Li et al. 2004). However, the genes for *NFATC2*, *GATAD2B*, *SFTPC*, *CC10*, and *IL6* (37 Chokas 2010, 39 Yang 2010) encoding other annotated targets or binding partners have no strong FOXP2 ChIP-seq peaks. Overall, these data suggest that FOXP2 may engage in feedback regulation of several of its annotated binding partners.

Sequence conservation at FOXP2 localization peaks

We expected a high degree of conservation for functionally important elements within ChIP-seq peak regions. Using the NCBI36 UCSC phyloP scores for site-specific conservation of multiple aligned sequences (40 Pollard et al. 2010), we observe that 51 of our 71 high-confidence peaks overlap well-conserved loci. Likewise, the average of aligned phastCons scores (41 Siepel et al. 2005) reveals an increase in conservation centered on the predicted FOXP2 binding sites, as does the first principle component of the phastCons scores (Figure 5). Such elevated conservation further confirms that we have identified the relevant DNA binding motif for FOXP2, and suggests that there is a set of well-conserved target sites for FOXP2 throughout vertebrates.

Human-specific exceptions to this generally strong conservation at the binding

site could provide insight into FOXP2's function in human-specific phenotypes. Therefore, we sought to identify particular FOXP2 targets with poorly conserved binding sites among the 71 high-confidence target loci. The sites near *HSF2BP* and *PCMTD2* localization peaks contain instances of our optimal binding sequence that are unique to the human lineage. HSF2BP (heat shock factor 2 binding protein) is known to bind the developmental transcription factor HSF2, which is required for normal brain development (42 Yoshima et al 1998, 43 Kallio 2002). A single base change in humans was responsible for creating a new optimal binding site in the 7th intron of HSF2BP. PCMTD2 (protein-L-isoaspartate (D-aspartate) O-methyltransferase domain containing 2) is an aspartate and asparganine repair enzyme, and mice lacking this enzyme have increased brain size, abnormal arborization of pyramidal neuron dendrites, and die early of progressive epilepsy (44,45 Kim 1997, Yamamoto 1998). In the second intron of *PCMTD2*, an 18 bp deletion created a new optimal FOXP2 binding site.

To conduct a broader survey of ChIP-seq peaks throughout the genome, we collected all ChIP-seq peaks that were consistently identified within either the PFSK-1 or SK-N-MC cell lines and had perfect matches to the optimal binding sequence. Out of the 1483 replicate ChIP-seq peaks, 472 have the highest-scoring TGTTTAC FOXP2 core binding site. Among these binding sites we found 38 instances of changes in sequence between chimp and human. Of these, we discarded 16 sites in which the chimp sequence alone appeared to have acquired mutations relative to the mammalian consensus, leaving 22 sites

of interest (Table 4). Roughly half of these events involve an insertion or deletion and the rest involve one or more point mutations. 63% (10/16) of the nearby genes have brain-specific functions (annotated in gray in Table 4) and several may have direct roles in neuronal function. For example, we find sites near the genes encoding gap junction protein delta 2 (GJD2), consortin (C1orf71), and neuronal calcium sensor 1 (NCS1), both of which are involved in neuronal signal transduction. GJD2 forms a class of electrical synapses that modulate the firing pattern of neurons during development (46, 47 Bennet and Zukin 2004, Blankenship 2011), and gap junction assembly requires consortin (48 del Castillo 2010). At chemical synapses NCS1 modifies synaptic activity in response to calcium current, with broader roles in plasticity and spatial memory tasks in mice (59,50 Saab 2009, Yip 2010). These candidate novel target genes are potentially important to the evolution of the FOXP2 regulon in humans.

DISCUSSION

An accurate and precise binding site model provides a useful tool to study FOXP2's evolution and molecular involvement in the development of language. Despite intense interest, the true binding preferences of FOXP2 have remained a mystery, with different experimental techniques yielding different candidate consensus sites (Table 1). To clarify FOXP2's binding site preference, we produced detailed models of the binding site from independent microfluidic affinity cell free assays, summarized in Figure 3, and neuronal cell-based ChIP-

seq datasets (Figure 4). We find that the human and chimp FOXP2 *in vitro* binding profiles are virtually identical, featuring the same degeneracies at the same positions. The *in vitro* MITOMI data provides additional information about the penalties of a given substitution, while the ChIP-seq data provides clues to genomic targets in a more physiological setting. Using our *in vitro* derived motif to identify candidate FOXP2 targets, we find 18 ChIP-seq peaks with binding sites that would have been missed by a strict 'TGTTTAC' consensus sequence search, and identify several human-specific FOXP2 binding sites that may contribute to the evolutionarily novel role of FOXP2 in language.

In addition to the strong similarity between our independently-derived motifs, several other observations suggest that we have identified the optimal FOXP2 target site. First, our motif is consistent with the accepted RYMAAYA non-FOXP Forkhead box family theme (51, 52 Pierrou 1994, Nirula 1997). Second, conservation scores within ChIP-seq peak regions tend to peak at the exact location of our predicted binding sites. Taken together, these independent lines of evidence suggest that we have resolved the functional FOXP2 binding motif modeled both in terms of positional affinity effects and positional frequencies among bound promoters.

Additional analysis demonstrates that the motif derived here improves consistency with prior FOXP2 ChIP-chip data (17 Vernes et al 2007). Our core motif, modeled as a 5mer TGTKK for the sake of comparison, is overrepresented in the most significant ChIP probes, while the previously suggested ATTTG motif is enriched at the level of expectation (Supp. Figure 3). Nucleotide biases can

complicate motif search algorithms and may explain some of the prior controversy surrounding the binding site. There is a G/C bias in the most highly enriched ChIP-chip probes, perhaps due to a tendency for FOXP2 to bind sites near TSSs within CpG islands(37 Gardener Garden 1987).

Encouragingly, the genes we identify as likely direct targets of FOXP2 also have altered patterns of expression and Foxp2 ChIP-chip signal as shown in previous experiments. Vernes and colleagues profiled expression in wildtype and Foxp2 321X mutant mice, and returned a list of 19 genes that had both ChIP-chip signal and significant expression changes (54 Vernes 2011). We found that 17 out of these 19 genes have a peak within 5 kb in at least one sample in the human ENCODE ChIP-seq data; *ALCAM, CCK, CSDE1, EBF2, GNAL, GNAS, MAPK8IP3, MAST1, NEGR1, NRN1, PLAG1, PSME4, SFXN4, TCF12, TGFBI, CITED2,* and *COL24A1*. An especially interesting target candidate from this list is CITED2 (Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 2), which modulates recruitment of the p300 histone acetyltransferase to promoters and remodeling of the chromatin locus (54 Bhatacharya 1999), and is known to modify FOXO proteins (55 Barthel et al 2005). CITED2 and these other genes appear to be reproducible FOXP2 regulatory targets as observed by independent researchers, with both activating and repressive outcomes (54 Vernes 2011).

From the ENCODE ChIP-seq data, we produced a list of consistent localization targets in neuronal cell lines and found that FOXP2 targets DNA-binding proteins such as glucocorticoid receptor and other forkhead box proteins. The set of

putative targets includes an alternative transcript of *FOXP2* itself and the gene encoding its annotated binding partner *FOXP1*. The ChIP-seq association with *FOXP1* is interesting because disruptions of these genes produce phenotypes with similar characteristics (56 Bacon 2012) and can cooperatively regulate reporter constructs (12, 36 Li 2004, Shu 2007). We speculate that autoregulation of the FOXP2 circuit may prove important to FOXP2's developmental function. In support of this hypothesis, FOXP2 is thought to be part of a coexpressed network of genes having a higher degree of connectivity in humans than in chimp and macaque (57 Konopka et al 2012). These themes are consistent with the established idea of FOXP2 as a regulator of transcriptional regulators.

Regarding the question of FOXP2's functional evolution, our data suggest that some of the genomic binding sites have evolved while the DNA-binding specificity of FOXP2 has been conserved. The FOXP2 PSAM motif and binding sites show a high degree of conservation in both biochemical affinity measurements and sequence alignment at ChIP-seq peaks. Inferring from this pattern of target site conservation, there appears to be a core set of FOXP2 targets in vertebrates, with a limited but interesting set of changed targets in humans. We have observed 22 potential examples of such *cis* evolution. These may represent newly acquired regulatory targets for human FOXP2 (Table 5, e.g. *NCS1* a synaptic calcium sensor involved in synaptic plasticity). Importantly, the FOXP2 targets listed here should not be considered an authoritative list. Rather, they were mainly used as examples to analyze the binding site and its evolution in humans. However, with a comprehensive binding site model we can now

improve our lists of direct FOXP2 targets, and better understand how its regulon may have changed over evolution. Future work of interest may include investigation of the differential protein-protein interactions of the chimp and human FOXP2, and generation of chimp FOXP2 ChIP-seq data for comparison with the existing mouse and human datasets.

FUNDING

This work was supported by the Howard Hughes Medical Institute [CN, JDR, PF]; QB3 California Institute for Quantitative Biosciences[CN, JDR, PF]; and the Helen Hay Whitney Foundation [PF].

ACKNOWLEDGEMENTS

The authors would like to acknowledge Simone Marticke, the Myers Lab and ENCODE for the ChIP-seq peak data, and the Biomolecular Nanotechnology Center at UC Berkeley and the QB3 Nanofab and Center for Advanced Technology (CAT) at UCSF for equipment support.

REFERENCES

1. Scharff,C., Petri,J. (2011) Evo-devo, deep homology and FOXP2: implications for the evolution of speech and language. Philos Trans R Soc Lond B Biol Sci. Jul 27;366(1574):2124-40.

2. Lai,C.S., Fisher,S.E., Hurst,J.A., Vargha-Khadem,F., Monaco A.P. (2001) A forkhead-domain gene is mutated in a severe speech and language disorder. Nature. Oct 4;413(6855):519-23.

3. Fisher,S.E., Vargha-Khadem,F., Watkins,K.E., Monaco,A.P., Pembrey,M.E.. Nat Genet. (1998) Feb;18(2):168-70. Localisation of a gene implicated in a severe speech and language disorder.

4. Stroud,J.C., Wu,Y., Bates,D.L., Han,A., Nowick,K., Paabo,S., Tong,H., Chen,L. (2006) Structure of the forkhead domain of FOXP2 bound to DNA. Structure 14:159–166.

5. Wu,Y., Borde,M., Heissmeyer,V., Feuerer,M., Lapan,A.D., Stroud,J.C., Bates,D.L., Guo,L., Han,A., Ziegler,S.F., et al. (2006) FOXP3 controls regulatory T cell function through cooperation with NFAT. Cell. Jul 28;126(2):375-87.

6. Shu,W., Yang,H., Zhang,L., Lu,M.M., Morrisey,E.E. (2001) Characterization of a new subfamily of winged-helix/forkhead (Fox) genes that are expressed in the lung and act as transcriptional repressors. J Biol Chem. Jul 20;276(29):27488-97.

7. MacDermot,K.D., Bonora, E., Sykes,N., Coupe,A.M., Lai,C.S., Vernes,S.C., Vargha-Khadem.F., McKenzie,F., Smith,R.L., Monaco,A.P., et al. (2005) Identification of FOXP2 truncation as a novel cause of developmental speech and language deficits, The American Journal of Human Genetics 76, 1074–1080.

8. Hurst,J.A., Baraitser,M., Auger,E., Graham,F., Norell,S. (1990) An extended family with a dominantly inherited speech disorder. Dev Med Child Neurol. Apr;32(4):352-5.

9. Zhang,J., Webb,D.M., Podlaha,O. (2002) Accelerated protein evolution and origins of human-specific features: FOXP2 as an example. Genetics. Dec;162(4):1825-35.

10. Enard,W., Przeworski,M., Fisher,S.E., Lai,C.S., Wiebe,V., Kitano,T., Monaco,A.P., Pääbo,S. (2002) Molecular evolution of FOXP2, a gene involved in speech and language. Nature. Aug 22;418(6900):869-72.

11. Yu,F., Keinan,A., Chen,H., Ferland,R.J., Hill,R.S., Mignault,A.A., Walsh,C.A., Reich,D. 2009 Detecting natural selection by empirical comparison to random regions of the genome. Hum Mol Genet. Dec 15;18(24):4853-67.

12. Shu,W., Cho,J-Y., Jiang,Y., Zhang,M., Weisz,D., Elder,G.A., Schmeidler,J., De Gasperi,R., Sosa,M.A., Rabidou,D., et al. (2005) Altered ultrasonic vocalization in mice with a disruption in the FOXP2 gene. Proc Natl Acad Sci U S A. Jul 5;102(27):9643-8.

13. Haesler,S., Rochefort,C., Georgi,B., Licznerski,P., Osten,P., Scharff,C. (2007) Incomplete and inaccurate vocal imitation after knockdown of FOXP2 in songbird basal ganglia nucleus Area X. PLoS Biol. Dec;5(12):e321.

14. Enard,W., Gehre,S., Hammerschmidt,K., Hölter,S.M., Blass,T., Somel,M., Brückner,M.K., Schreiweis,C., Winter,C., Sohr,R., et al. (2009) A humanized version of FOXP2 affects cortico-basal ganglia circuits in mice. Cell. May 29;137(5):961-71.

15. Schubert,L.A., Jeffery,E., Zhang,Y., Ramsdell,F., Ziegler,S.F. (2001) Scurfin (FOXP3) acts as a repressor of transcription and regulates T cell activation. J Biol Chem. Oct 5;276(40):37672-9.

16. Wang,B., Lin,D., Li,C., Tucker,P. (2003) Multiple domains define the expression and regulatory properties of Foxp1 forkhead transcriptional repressors. J Biol Chem. Jul 4;278(27):24259-68.

17. Vernes,S.C., Spiteri,E., Nicod,J., Groszer,M., Taylor,J.M., Davies,K.E., Geschwind,D.H., Fisher,S.E. (2007) High-throughput analysis of promoter occupancy reveals direct neural targets of FOXP2, a gene mutated in speech and language disorders. Am J Hum Genet. Dec;81(6):1232-50.

18. Maerkl,S.J., Quake,S.R. (2007) A systems approach to measuring the binding energy landscapes of transcription factors. Science. Jan 12;315(5809):233-7.

19. Fordyce,P.M., Gerber,D., Tran,D., Zheng,J., Li,H., DeRisi,J.L., Quake,S.R. (2010) De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. Nat Biotechnol. Sep;28(9):970-5.

20. Edelheit,O., Hanukoglu,A., Hanukoglu,I. (2009) Simple and efficient site-directed mutagenesis using two single-primer reactions in parallel to generate mutants for protein structure-function studies. BMC Biotechnol. Jun 30;9:61.

21. Vernes,S.C., Nicod,J., Elahi,F.M., Coventry,J.A., Kenny,N., Coupe,A.M., Bird,L.E., Davies,K.E., Fisher,S.E. (2006) Functional genetic analysis of mutations implicated in a human speech and language disorder. Hum Mol Genet. Nov 1;15(21):3154-67.

22. Khare,S.D., Ding,F., Gwanmesia,K.N., Dokholyan,N.V. (2005) Molecular Origin of Polyglutamine Aggregation in Neurodegenerative Diseases . PLoS Comput Biol 1(3): e30.

23. Ross,C.A. (2002) Polyglutamine pathogenesis: Emergence of unifying mechanisms for Huntington's disease and related disorders. Neuron 35: 819–822

24. Fordyce,P.M., Pincus,D., Kimmig,P., Nelson,C.S., El-Samad,H., Walter,P., Derisi,J.L. (2012) Basic leucine zipper transcription factor Hac1 binds DNA in two distinct modes as revealed by microfluidic analyses. Proc Natl Acad Sci U S A. Oct 10.

25. Mintseris,J. and Eisen,M.B. (2006) Design of a combinatorial DNA microarray for protein-DNA interaction studies. BMC Bioinformatics.; 7: 429.

26. Wu,R.Z., Chaivorapol,C., Zheng,J., Li,H., Liang,S. (2007) fREDUCE: detection of degenerate regulatory elements using correlation with expression. BMC Bioinformatics. Oct 17;8:399.

27. Foat,B.C., Morozov,A.V., Bussemaker,H.J. (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. Bioinformatics. Jul 15;22(14):e141-9.

28. Marticke,S.S. (2008) Ultra-high Throughput Sequencing Analysis of FOXP2 Occupancy in the Human Genome Doctorate Thesis, Stanford University. 127p.

29. Valouev,A., Johnson,D.S., Sundquist,A., Medina,C., Anton,E., Batzoglou,S., Myers,R.M., Sidow,A. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods. Sep;5(9):829-34.

30. Benjamini,Y., Hochberg,Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Roy Stat Soc B.;75:289–300.

31. Bailey,T.L. and Elkan,C. (1994) "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, Menlo Park, California.

32. Gordon,D.B., Nekludova,L., McCallum,S., Fraenkel,E. (2005) TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. Bioinformatics. Jul 15;21(14):3164-5.

33. Schroeder,D.I., Myers R.M. (2008) Multiple transcription start sites for FOXP2 with varying cellular specificities. Gene. Volume 413, Issues 1–2, 30 April, Pages 42–48

34. Benjamini,Y., Speed,TP. (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. May;40(10):e72.

35. Chen,Y., Negre,N., Li,Q., Mieczkowska,J.O., Slattery,M., Liu,T., Zhang,Y., Kim,T-K., He,H.H., Zieb,J., et al. (2012) Systematic evaluation of factors influencing ChIP-seq fidelity. Nature Methods. 9, 609–614

36. Gardiner-Garden,M., Frommer,M. (1987) CpG islands in vertebrate genomes. J Mol Biol. Jul 20;196(2):261-82.

37. Chokas,A.L., Trivedi,C.M., Lu,M.M., Tucker,P.W., Li,S., Epstein,J.A., Morrisey,E.E. (2010) Foxp1/2/4-NuRD interactions regulate gene expression and epithelial injury response in the lung via regulation of interleukin-6. J Biol Chem. Apr 23;285(17):13304-13.

38. Li,S., Weidenfeld,J., Morrisey,E.E. (2004) Transcriptional and DNA binding activity of the Foxp1/2/4 family is modulated by heterotypic and homotypic protein interactions. Mol Cell Biol. Jan;24(2):809-22.

39. Yang,Z., Hikosaka,K., Sharkar,M.T., Tamakoshi,T., Chandra,A., Wang,B., Itakura,T., Xue,X., Uezato,T., Kimura,W., et al. (2010) The mouse forkhead gene Foxp2 modulates expression of the lung genes. Life Sci. Jul 3;87(1-2):17-25.

40. Pollard,K.S., Hubisz,M.J., Rosenbloom,K.R., Siepel,A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. Jan;20(1):110-21.

41. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S., et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. Aug;15(8):1034-50.

42. Yoshima,T., Yura,T., Yanagi,H. (1998) Novel testis-specific protein that interacts with heat shock factor 2. Gene. Jul 3;214(1-2):139-46.

43. Kallio,M., Chang,Y., Manuel,M., Alastalo,T.P., Rallu,M., Gitton,Y., Pirkkala,L., Loones,M.T., Paslaru L., Larney S, et al. (2002) Brain abnormalities, defective meiotic chromosome synapsis and female subfertility in HSF2 null mice. EMBO J. Jun 3;21(11):2591-601.

44. Kim,E., Lowenson,J.D., MacLaren D.C., Clarke,S., Young,S.G. (1997) Deficiency of a protein-repair enzyme results in the accumulation of altered proteins, retardation of growth, and fatal seizures in mice. Proc Natl Acad Sci U S A. Jun 10;94(12):6132-7.

45. Yamamoto,A., Takagi,H., Kitamura,D., Tatsuoka,H., Nakano,H., Kawano,H., Kuroyanagi,H., Yahagi,Y., Kobayashi,S., Koizumi,K., et al. (1998) Deficiency in protein L-

isoaspartyl methyltransferase results in a fatal progressive epilepsy. J Neurosci. Mar 15;18(6):2063-74.

46. Bennett,M.V., Zukin,R.S. (2004) Electrical coupling and neuronal synchronization in the Mammalian brain. Neuron. Feb 19; 41(4):495-511.

47. Blankenship,A.G., Hamby,A.M., Firl,A., Vyas,S., Maxeiner,S., Willecke,K., Feller,M.B. (2011) The role of neuronal connexins 36 and 45 in shaping spontaneous firing patterns in the developing retina. J Neurosci. Jul 6;31(27):9998-10008.

48. del Castillo,F.J., Cohen-Salmon,M., Charollais,A., Caille,D., Lampe,P.D., Chavrier,P., Meda,P., Petit,C. (2010) Consortin, a trans-Golgi network cargo receptor for the plasma membrane targeting and recycling of connexins. Hum Mol Genet. Jan 15;19(2):262-75. doi: 10.1093/hmg/ddp490.

49. Yip,P.K., Wong,L.F., Sears,T.A., Yáñez-Muñoz,R.J., McMahon,S.B. (2010) Cortical overexpression of neuronal calcium sensor-1 induces functional plasticity in spinal cord following unilateral pyramidal tract injury in rat. PLoS Biol. Jun 22;8(6):e1000399.

50. Saab,B.J., Georgiou,J., Nath,A., Lee,F.J., Wang,M., Michalon,A., Liu,F., Mansuy,I.M., Roder,J.C. (2009) NCS-1 in the dentate gyrus promotes exploration, synaptic plasticity, and rapid acquisition of spatial memory. Neuron. Sep 10;63(5):643-56.

51. Pierrou,S., Hellqvist,M., Samuelsson,L., Enerbäck,S., Carlsson,P. (1994) Cloning and characterization of seven human forkhead proteins: binding site specificity and DNA bending. EMBO J. Oct 17;13(20):5002-12.

52. Nirula,A., Moore,D.J., Gaynor,R.B. (1997) Constitutive binding of the transcription factor interleukin-2 (IL-2) enhancer binding factor to the IL-2 promoter. J Biol Chem. Mar 21;272(12):7736-45.

53. Vernes,S.C., Oliver P.L., Spiteri,E., Lockstone,H.E., Puliyadi,R., Taylor,J.M., Ho,J., Mombereau,C., Brewer,A., Lowy,E., et al. (2011) FOXP2 regulates gene networks implicated in neurite outgrowth in the developing brain. PLoS Genet. Jul;7(7):e1002145.

54. Bhattacharya,S., Michels,C.L., Leung,M.K., Arany,Z.P., Kung,A.L., Livingston,D.M. (1999) Functional role of p35srj, a novel p300/CBP binding protein, during transactivation by HIF-1. Genes Dev. Jan 1;13(1):64-75.

55. Bacon,C., Rappold,G.A. (2012) The distinct and overlapping phenotypic spectra of FOXP1 and FOXP2 in cognitive disorders. Hum Genet. 2012 Nov;131(11):1687-98.

56. Konopka,G., Friedrich,T., Davis-Turak,J., Winden,K., Oldham,M.C., Gao,F., Chen,L., Wang,G.Z., Luo,R., Preuss,T.M., et al. (2012) Human-specific transcriptional networks in the brain. Neuron. Aug 23;75(4):601-17.

TABLE AND FIGURES LEGENDS

Table 1. Previously reported models of the FOXP2 binding site.

Table 2. Consistent ChIP-seq peaks near gene models. Peaks within 5 kb of a gene model are shown along with PSAM motif scores. If the PSAM score is in the top 0.1% of score for random 7mers then it is noted in the “Top 0.1%” column. The right hand column notes whether the peak contains the consensus TGTTTAC. (Intergenic peaks are described in supplementary table 4).

Table 3. Gene ontology term analysis of consistent peaks from the ENCODE ChIP-seq data.

Table 4. FOXP2 binding sites within ChIP-seq peaks where the human sequence is novel relative to chimps and other primates. Coordinates listed are relative to Hg18 / NCBI36 draft of the human genome. Sites more than 5 kb from a gene model were not given a target gene. The scores are the bitscore of the site in question relative to the human MEME matrix. Gray shading denotes a gene with brain-specific function.

Figure 1. Schematic of FOXP2 domains and truncated construct used in MITOMI experiments. FOXP2 contains a polyglutamine (polyQ) stretch that we removed by truncation of the shaded region. The human lineage substitutions are at positions 303 and 325 after the polyQ region. As in other FOXP proteins, FOXP2 has a C2H2 zinc finger domain, a leucine zipper domain and a forkhead box DNA binding domain (shown in green). R553H mutation linked to verbal

dyspraxia lies within the DNA binding domain. We 6xHis tagged the C–terminus for recruitment and retention on chip (brown).

Figure 2. Results from FOXP2 MITOMI2.0 binding assays against pseudorandom 8mer library. A) Histograms of MITOMI data. Bound DNA signal (normalized by the protein signal) for human WT, human R553H and chimpanzee alleles. R553H shows no binding to any sequence in the library while chimp and human FOXP2 produce strong binding to a subset of oligos (seen as the right hand tail in the distributions). B) Comparison of chimp and human binding ratios showing high correlation. Oligonucleotide #175, used for later targeted analysis is labeled in red. C) Top scoring human MatrixREDUCE 7mer logo D. Top scoring chimp MatrixREDUCE 7mer logo.

Figure 3. Affinity measurements for systematic mutations of the binding site and flanking sequences confirm binding site motif and provide position specific affinity profile of the motifs. A) Fold change in affinity (mutated K_a/ unmutated K_a) shown in log scale; error bars represent the standard error of the mean. Chimp data is displayed in red and human in blue. The two profiles are largely similar and contain the same valley of impermissive mutations that reflects the core binding sequence. B) PSAM affinity logo based on the affinities displayed in part A for the human allele C) PSAM affinity logo based on the affinities displayed in part A for the chimp allele. D) Difference between human and chimp positional affinity effects are relatively minor.

Figure 4. ChIP-seq analysis reveals motif consistent with MITOMI data that has a sterotyped location A. MEME motif returned with 50 bp flanks on the ChIP peaks B. Histogram of the relative positioning of FOXP2 motifs (scoring over 75% of the maximal motif score) relative to the start of the nearest neighboring gene.

Figure 5. Sequence near FOXP2 motif instances within ChIP-seq peaks are conserved. A) Example of two FOXP2 ChIP-seq peaks aligned with elements of strong conservation. Upstream of *BACCH1* on Chr17: 79,366,750-79,370,250 (hg18 / NCBI36). Shown are alignments of two high confidence peak regions with high scoring instances of our MEME motif, and the vertebrate conservation score for the underlying sequence. B) The mean of the phastcons conservation score over the FOXP2 peak regions is higher nearer the motif and well over the genomic background average conservation score in red. The same is true for the first principal component, which is plotted in blue on the same scale, noted on the right-hand axis.

Table 1. Previously reported models of the FOXP2 binding site.

Publication	**Data type**	**System**	**Motif**
Vernes et al 2007	ChIP-chip	SH-SY5Y cells overexpressing FOXP2	TCTTCGT
Vernes et al 2008	EMSA	*in vitro* binding to *CNTAP2* sequence	AATTTG
Enard et al 2009	gene expression	humanized mice	TATTTAT
Vernes et al 2011	ChIP-chip	wildtype embryonic mice	ARKTAMYT

Table 2. Consistent ChIP-seq peaks near gene models. Peaks within 5 kb of a gene model are shown along with PSAM motif scores. If the PSAM score is in the top 0.1% of score for random 7mers then it is noted in the “Top 0.1%” column. The right hand column notes whether the peak contains the consensus TGTTTAC. (Intergenic peaks are described in supplementary table 4).

Peak #	max PSAM score	Top 0.1%	"TGTTTAC"	Nearby gene	Description	RefSeq #
1	0.45	yes	no	NFIA	nuclear factor I/A	NM_001145512
2	1.00	yes	yes	TPRG1L	tumor protein p63-regulated gene 1-like protein	NM_182752
3	1.00	yes	yes	BROX	BRO1 domain and CAAX motif containing	NM_144695
4	1.00	yes	yes	RBM17	RNA binding motif protein 17	NM_001145547
6	1.00	yes	yes	PSMA1	proteasome subunit alpha type-1	NM_148976
7	0.05	no	no	ZBTB16	zinc finger and BTB domain containing 16	NM_006006
9	0.10	yes	no	NAB2	NGFI-A binding protein 2 (EGR1 binding protein 2)	NM_005967
10	0.22	yes	no	TPCN1	two pore segment channel 1	NM_001143819
11	1.00	yes	yes	BTG1	B-cell translocation gene 1, anti-proliferative	NM_001731
13	1.00	yes	yes	KLHDC2	kelch domain containing 2	NM_014315
14	0.14	yes	no	KIAA0586	Uncharacterized protein	NM_001244189
15	1.00	yes	yes	BAHCC1	Bromo adjacent homology domain and coiled-coil containing 1	NM_001080519
16	1.00	yes	yes	DHX8	DEAH (Asp-Glu-Ala-His) box polypeptide 8	NM_004941
17	1.00	yes	yes	DHX40	DEAH (Asp-Glu-Ala-His) box polypeptide 40	NM_024612
18	0.04	no	no	SPOP	speckle-type POZ protein (SPOP)	NM_001007226
19	1.00	yes	yes	PHLPP1	PH domain leucine-rich repeat-containing protein phosphatase 1	NM_194449
20	1.00	yes	yes	LTBP4	latent-transforming growth factor beta-binding protein 4	NM_001042544
21	1.00	yes	yes	JUNB	jun B proto-oncogene	NM_002229
22	0.14	yes	no	FBXO46	F-box protein 46	NM_001080469
23	1.00	yes	yes	BBC3	BCL2 binding component 3	NM_001127240
24	1.00	yes	yes	FUZ	fuzzy homolog (Drosophila)	NM_025129
25	0.14	yes	no	SPAST	spastin	NM_014946
27	0.07	yes	no	ARHGAP25	Rho GTPase activating protein 25	NM_001007231
29	1.00	yes	yes	PCMTD2	protein-L-isoaspartate O-methyltransferase domain-containing protein	NM_018257
32	1.00	yes	yes	HSF2BP	heat shock transcription factor 2 binding protein	NM_007031
33	1.00	yes	yes	PIGP	phosphatidylinositol N-acetylglucosaminyltransferase subunit P	NM_153682
34	1.00	yes	yes	C21orf77	C21orf77	NM_144659
35	1.00	yes	yes	CBX7	chromobox protein homolog 7	NM_175709
36	1.00	yes	yes	CECR3	cat eye syndrome chromosome region, candidate 3 (non-protein coding)	NR_038398
38	1.00	yes	yes	FOXP1	forkhead box P1	NM_032682
39	1.00	yes	yes	MAML3	mastermind-like protein 3	NM_018717
40	1.00	yes	yes	YTHDC1	YTH domain-containing protein 1	NM_001031732
41	0.14	yes	no	UBE2B	ubiquitin-conjugating enzyme E2B	NM_003337
42	1.00	yes	yes	POLK	DNA-directed DNA polymerase kappa	NM_016218
43	1.00	yes	yes	NR3C1	nuclear receptor subfamily 3, group C, member 1 (glucocorticoid receptor)	NM_000176
44	0.01	no	no	GPANK1	G patch domain and ankyrin repeats 1	NM_001199237
46	1.00	yes	yes	CCDC28A	coiled-coil domain containing 28A	NM_015439
47	0.14	yes	no	FAM8A1	family with sequence similarity 8, member A1	NM_016255
48	0.45	yes	no	DTNBP1	dystrobrevin binding protein 1	NM_032122
49	1.00	yes	yes	RUNX2	runt-related transcription factor 2	NM_004348
50	1.00	yes	yes	CITED2	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 2	NM_006079
51	0.00	no	no	PRKRIP1	PRKR interacting protein 1 (IL11 inducible)	NM_024653
52	1.00	yes	yes	ELN	elastin	NM_000501
53	1.00	yes	yes	CBLL1	Cas-Br-M (murine) ecotropic retroviral transforming sequence-like 1	NM_024814
54 & 55	1.00	yes	yes	FOXP2	forkhead box P2	NR_033766.1
56	1.00	yes	yes	FOXK1	forkhead box K1	NM_001037165
57	1.00	yes	yes	HIBADH	3-hydroxyisobutyrate dehydrogenase	NM_152740
58	1.00	yes	yes	THSD7A	thrombospondin type-1 domain-containing protein 7A	NM_015204
59	1.00	yes	yes	TNRC18	trinucleotide repeat-containing gene 18 protein	NM_001080495
61	1.00	yes	yes	PVT1	Pvt1 oncogene (non-protein coding)	NR_003367
62	1.00	yes	yes	ZNF395	zinc finger protein 395	NM_018660
63	0.05	no	no	FNTA	farnesyltransferase, CAAX box, alpha	NM_002027
64	1.00	yes	yes	OSR2	protein odd-skipped-related 2	NM_001142462
65	0.22	yes	no	TNFRSF10B	tumor necrosis factor (ligand) superfamily, member 10	NM_003810
66	0.45	yes	no	FBXO32	F-box protein 32	NM_058229
67	0.14	yes	no	ASAP1	ArfGAP with SH3 domain, ankyrin repeat and PH domain 1	NM_018482
69	1.00	yes	yes	BRD3	bromodomain containing 3	NM_007371
70	0.50	yes	no	TBL1X	transducin (beta)-like 1X-linked	NM_005647

Table 3. Gene ontology term analysis of consistent peaks from the ENCODE ChIP-seq data.

cell line	GO #	GO term	p-value
PFSK-1	0008134	transcription factor binding	0.0016
PFSK-1	0030528	transcription regulator activity	0.0016
SK-N-MC	0003690	double-stranded DNA binding	0.0558
SK-N-MC	0003700	sequence-specific DNA binding transcription factor activity	0.0470
SK-N-MC	0016563	transcription activator activity	0.0558
SK-N-MC	0016564	transcription repressor activity	0.0189

Table 4. FOXP2 binding sites within ChIP-seq peaks where the human sequence is novel relative to chimps and other primates. Coordinates listed are relative to Hg18 / NCBI36 draft of the human genome. Sites more than 5 kb from a gene model were not given a target gene. The scores are the bitscore of the site in question relative to the human MEME matrix. Gray shading denotes a gene with brain-specific function.

Peak Location	Human TFBS	Human Score	Chimp Sequence	Chimp Score	Change Type	Cell Type	Relative to Nearest Gene	Nearest Gene	Description
chr1:232878698-232878854	GTAAACA	13.63	CGTGTAC	3.86	SNP or SNPs	Pfsk1	-	-	-
chr1:244869749-244869890	GTAAACA	13.63		0.00	Ins / Del > 50 bp	Pfsk1	6th intron	C1orf71	consortin, connexin sorting protein
chr10:12150643-12150783	GTAAACA	13.63	GTGAACA	5.79	SNP or SNPs	Pfsk1	182 bp upstream	DHTKD1	dehydrogenase E1 and transketolase domain containing 1
chr10:1272357-1272503	TGTTTAC	13.63		0.00	Ins / Del > 50 bp	Pfsk1	5th intron	ADARB2	adenosine deaminase, RNA-specific, B2
chr10:42453548-42453718	GTAAACA	13.63	GGCAACA	3.14	Partial Ins / Del	Pfsk1	1st intron	ZNF33B	zinc finger protein 33B
chr12:68922880-68923112	GTAAACA	13.63	GTAAATA	9.25	SNP or SNPs	Pfsk1	94 bp upstream of the start	CNOT2	CCR4-NOT transcription complex, subunit 2
chr15:32832867-32833011	TGTTTAC	13.63	TGTTTAG	5.64	SNP or SNPs	Pfsk1	in the first intron	GJD2	gap junction protein, delta 2, 36kDa
chr17:31916541-31916693	GTAAACA	13.63	GTTAACA	5.79	SNP or SNPs	Pfsk1	46 bp from the start	ZNHIT3	zinc finger, HIT-type containing 3
chr18:72193002-72193148	TGTTTAC	13.63	TATTTAG	1.67	SNP or SNPs	Pfsk1	-	-	-
chr18:9063758-9063905	TGTTTAC	13.63	TATTTAC	9.25	SNP or SNPs	Sknmc	-	-	-
chr19:18263776-18263918	GTAAACA	13.63		0.00	Ins / Del > 50 bp	Pfsk1	-	-	-
chr2:179910111-179910256	TGTTTAC	13.63	TGTTTTC	9.86	SNP or SNPs	Sknmc	-	-	-
chr2:197174993-197175140	TGTTTAC	13.63	TGTCTAC	5.68	SNP or SNPs	Pfsk1	-	-	-
chr2:203161760-203161908	TGTTTAC	13.63		0.00	Ins / Del > 50 bp	Sknmc	-	-	-
chr2:236168933-236169074	GTAAACA	13.63	ATAAACA	8.85	SNP or SNPs	Sknmc	1st intron	AGAP1	centaurin, gamma 2 isoform 1
chr20:62362996-62363143	GTAAACA	13.63	CTAAACA	5.64	Partial Ins / Del	Sknmc	2nd intron	PCMTD2	protein-L-isoaspartate (D-aspartate) O-methyltransferase domain containing 2
chr21:43854041-43854198	GTAAACA	13.63	CTAAACA	5.64	SNP or SNPs	Sknmc	7th intron	HSF2BP	heat shock transcription factor 2 binding
chr21:46942689-46942846	GTAAACA	13.63		0.00	Ins / Del > 50 bp	Pfsk1	-	-	-
chr5:4755263-4755405	TGTTTAC	13.63	TATTTAC	9.25	Partial Ins / Del	Pfsk1	-	-	-
chr6:164362950-164363091	TGTTTAC	13.63	CGTTTAC	7.51	SNP or SNPs	Sknmc	-	-	-
chr7:882304-882459	TGTTTAC	13.63		0.00	Ins / Del > 50 bp	Pfsk1	1kb downstream	UNC84A	unc-84 homolog A
chr9:132039295-132039457	TGTTTAC	13.63	TGTTTCC	5.72	SNP or SNPs	Pfsk1	last exon	NCS1	neuronal calcium sensor 1

FIGURES

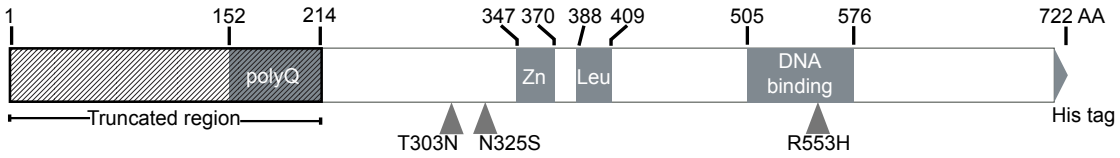


Figure 1. Schematic of FOXP2 domains and truncated construct used in MITOMI experiments. FOXP2 contains a polyglutamine (polyQ) stretch that we removed by truncation of the shaded region. The human lineage substitutions are at positions 303 and 325 after the polyQ region. As in other FOXP proteins, FOXP2 has a C2H2 zinc finger domain, a leucine zipper domain and a forkhead box DNA binding domain (shown in green). R553H mutation linked to verbal dyspraxia lies within the DNA binding domain. We 6xHis tagged the C –terminus for recruitment and retention on chip (brown).

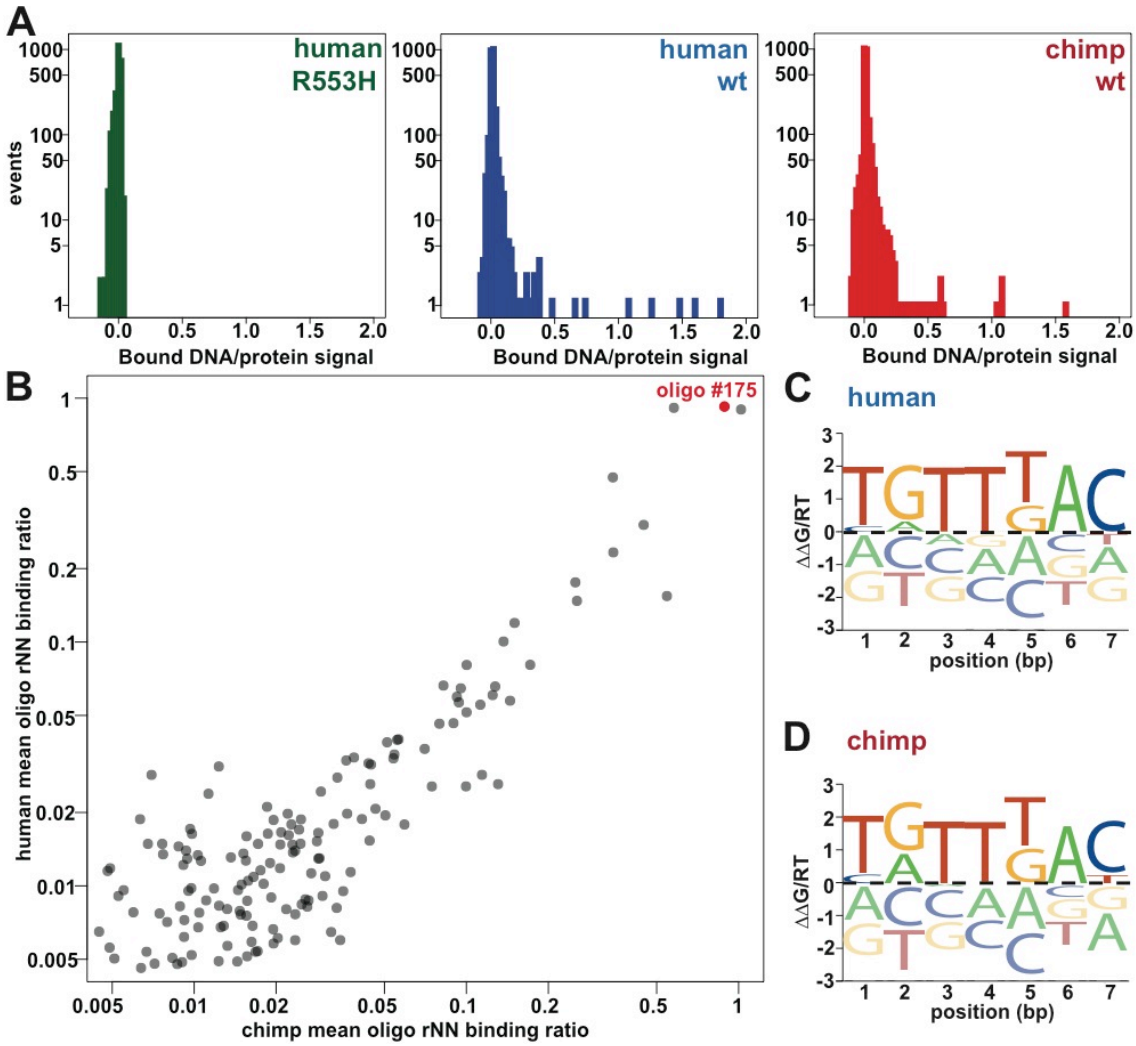


Figure 2. Results from FOXP2 MITOMI2.0 binding assays against pseudorandom 8mer library. A. Histograms of MITOMI bound DNA signal (normalized by protein signal) for human WT, human R553H and chimpanzee alleles. R553H shows no binding to any sequence in the library while chimp and human FOXP2 produce strong binding to a subset of oligos (seen as the right hand tail in the distributions). B. Comparison of chimp and human binding ratios showing high correlation. Oligonucleotide #175, used for later targeted analysis is labeled in red. C. Top scoring human MatrixREDUCE 7mer logo D. Top scoring chimp MatrixREDUCE 7mer logo.

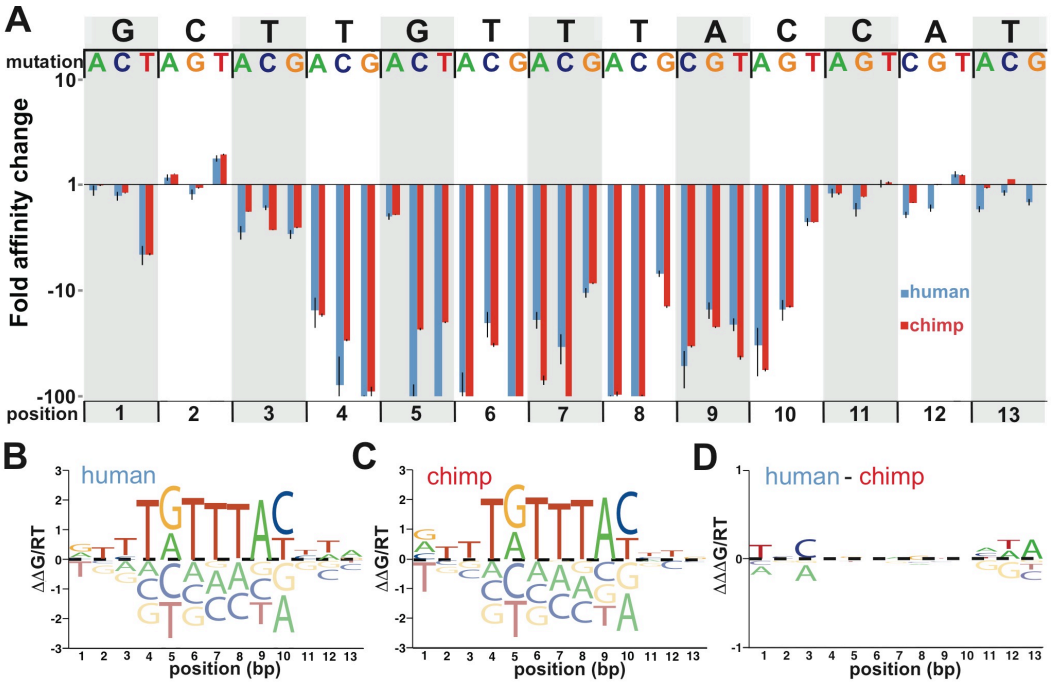


Figure 3. Affinity measurements for systematic mutations of the binding site and flanking sequences confirm binding site motif and provide position specific affinity profile of the motifs. A. Fold change in affinity (mutated K_a/ unmutated K_a) shown in log scale; error bars represent the standard error of the mean. Chimp data is displayed in red and human in blue. The two profiles are largely similar and contain the same valley of impermissive mutations that reflects the core binding sequence. B. PSAM affinity logo based on the affinities displayed in part A for the human allele C. PSAM affinity logo based on the affinities displayed in part A for the chimp allele. D. Difference between human and chimp positional affinity effects are relatively minor.

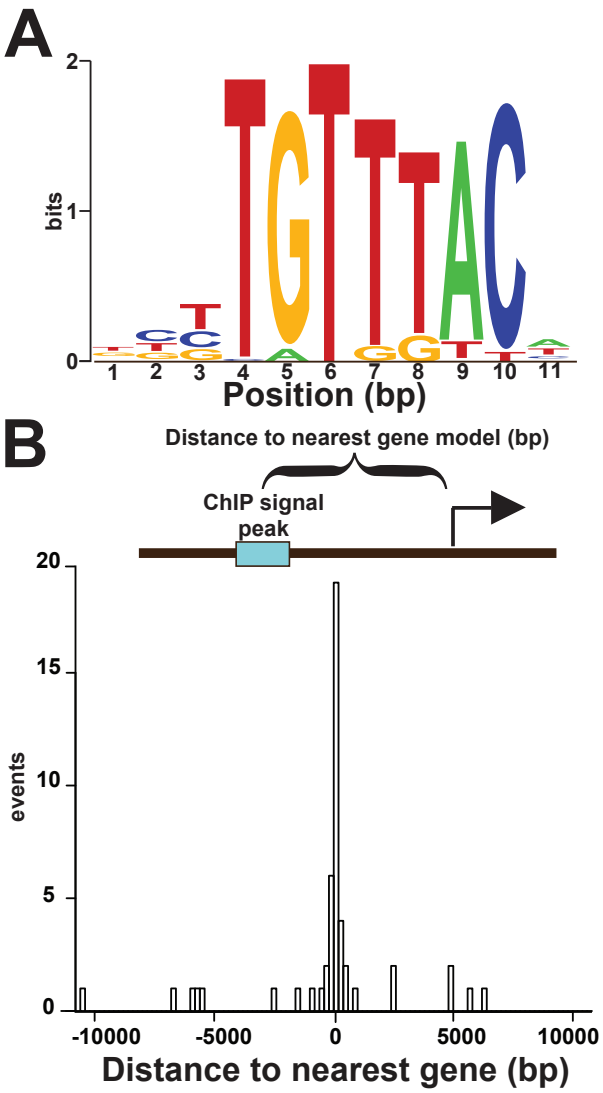


Figure 4. ChIP-seq analysis reveals motif consistent with MITOMI data that has a sterotyped location. A. MEME motif returned with 50 bp flanks on the ChIP peaks B. Histogram of the relative positioning of FOXP2 motifs (scoring over 75% of the maximal motif score) relative to the start of the nearest neighboring gene.

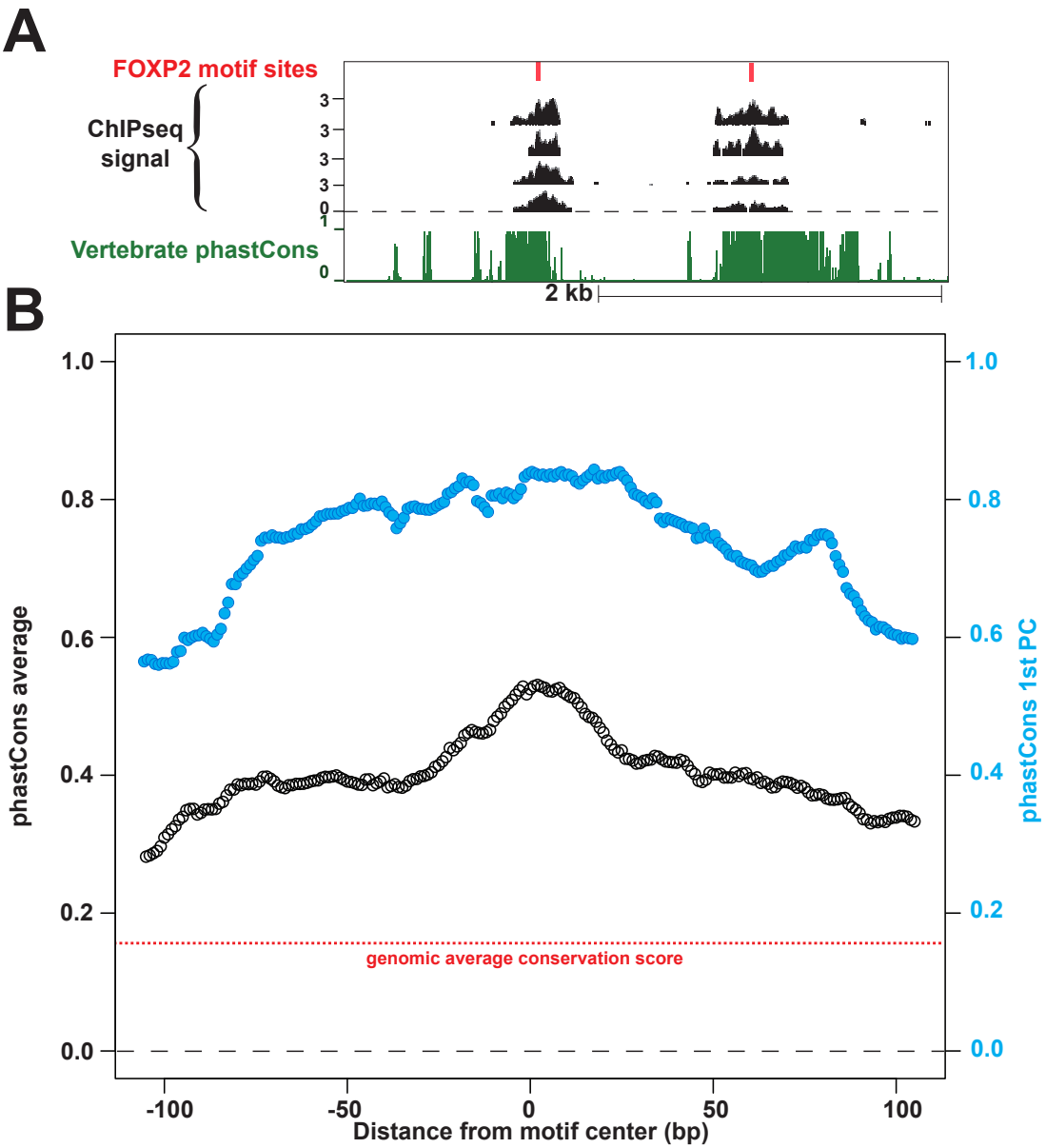


Figure 5. Sequence near FOXP2 motif instances within ChIP-seq peaks are conserved. A) Example of two FOXP2 ChIP-seq peaks aligned with elements of strong conservation. Upstream of *BACCH1* on Chr17: 79,366,750-79,370,250 (hg18 / NCBI36). Shown are alignments of two high confidence peak regions with high scoring instances of our MEME motif, and the vertebrate conservation score for the underlying sequence. B) The mean of the phastcons conservation score

over the FOXP2 peak regions is higher nearer the motif and well over the genomic background average conservation score in red. The same is true for the first principal component, which is plotted in blue on the same scale, noted on the right-hand axis.

SUPPLEMENTAL INFORMATION

Primers used for cDNA cloning from HeLa RNA:

Forward 5'-ATGATGCAGGAATCTGCGACAGAG-3'

Reverse 5'-TCATTCCAGATCTTCAGATAAAGGCTCTTC-3'

Forward flanking T7 promoter sequence added by 2 step PCR:
5'-
GATCTTAAGGCTAGAGTACTAATACGACTCACTATAGGgaatacaagctacttgttctttttcgactcgagaattc
GCCACC ATGATGCAGGAATCTGCGACAGAG-3'

Reverse flanking His tag, terminator, and PolyA sequence added by 2 step PCR:
5'-
tattacgccagccggatccAAAAACCCCTCAAGACCCGTTTAGAGGCCCCAAGGGGTTATGCTAGttttttt
ttttttttttttttttttttttttGTAGCAGCCTGAGTCGttattaATGATGATGATGATGATGTTCCAGATCTTCAGAT
AAAGGCTCTTC-3'

In vitro linear template sequence (hFOXP2):
GATCTTAAGGCTAGAGTACTAATACGACTCACTATAGGgaatacaagctacttgttctttttcgactcgagaatt
cGCCACCATGATGCAGGAATCTGCGACAGAGACAATAAGCAACAGTTCAATGAATCAAAAT
GGAATGAGCACTCTAAGCAGCCAATTAGATGCTGGCAGCAGAGATGGAAGATCAAGTGGT
GACACCAGCTCTGAAGTAAGCACAGTAGAACTGCTGCATCTGCAACAACAGCAGGCTCTCC
AGGCAGCAAGACAACTTCTTTTACAGCAGCAAACAAGTGGATTGAAATCTCCTAAGAGCAG
TGATAAACAGAGACCACTGCAGGTGCCTGTGTCAGTGGCCATGATGACTCCCCAGGTGATC
ACCCCTCAGCAAATGCAGCAGATCCTTCAGCAACAAGTCCTGTCTCCTCAGCAGCTACAAG
CCCTTCTCCAACAACAGCAGGCTGTCATGCTGCAGCAGCAACAACTACAAGAGTTTTACAA
GAAACAGCAAGAGCAGTTACATCTTCAGCTTTTGCAGCAGCAGCAGCAACAGCAGCAGCAG
CAACAACAGCAGCAACAACAGCAGCAGCAACAACAACAACAACAGCAGCAACAACAGCAG
CAGCAGCAGCAACAGCAGCAGCAGCAGCAACAGCATCCTGGAAAGCAAGCGAAAGAGCAG
CAGCAGCAGCAGCAGCAGCAACAGCAATTGGCAGCCCAGCAGCTTGTCTTCCAGCAGCAG
CTTCTCCAGATGCAACAACTCCAGCAGCAGCAGCATCTGCTCAGCCTTCAGCGTCAGGGAC
TCATCTCCATTCCACCTGGCCAGGCAGCACTTCCTGTCCAATCGCTGCCTCAAGCTGGCTT
AAGTCCTGCTGAGATTCAGCAGTTATGGAAAGAAGTGACTGGAGTTCACAGTATGGAAGAC
AATGGCATTAAACATGGAGGGCTAGACCTCACTACTAACAATTCCTCCTCGACTACCTCCTC
CaacACTTCCAAAGCATCACCACCAATAACTCATCATTCCATAGTGAATGGACAGTCTTCAGT
TCTAAGTGCAAGACGAGACAGCTCGTCACATGAGGAGACTGGGGCCTCTCACACTCTCTAT
GGCCATGGAGTTTGCAAATGGCCAGGCTGTGAAAGCATTTGTGAAGATTTTGGACAGTTTT
TAAAGCACCTTAACAATGAACACGCATTGGATGACCGAAGCACTGCTCAGTGTCGAGTGCA
AATGCAGGTGGTGCAACAGTTAGAAATACAGCTTTCTAAAGAACGCGAACGTCTTCAAGCA
ATGATGACCCACTTGCACATGCGACCCTCAGAGCCCAAACCATCTCCCAAACCTCTAAATC
TGGTGTCTAGTGTCACCATGTCGAAGAATATGTTGGAGACATCCCCACAGAGCTTACCTCA
AACCCCTACCACACCAACGGCCCCAGTCACCCCGATTACCCAGGGACCCTCAGTAATCAC
CCCAGCCAGTGTGCCCAATGTGGGAGCCATACGAAGGCGACATTCAGACAAATACAACATT

CCCATGTCATCAGAAATTGCCCCAAACTATGAATTTTATAAAAATGCAGATGTCAGACCTCC
ATTTACTTATGCAACTCTCATAAGGCAGGCTATCATGGAGTCATCTGACAGGCAGTTAACAC
TTAATGAAATTTACAGCTGGTTTACACGGACATTTGCTTACTTCAGGCGTAATGCAGCAACT
TGGAAGAATGCAGTACGTCATAATCTTAGCCTGCACAAGTGTTTTGTTCGAGTAGAAAATGT
TAAAGGAGCAGTATGGACTGTGGATGAAGTAGAATACCAGAAGCGAAGGTCACAAAAGATA
ACAGGAAGTCCAACCTTAGTAAAAAATATACCTACCAGTTTAGGCTATGGAGCAGCTCTTAA
TGCCAGTTTGCAGGCTGCCTTGGCAGAGAGCAGTTTACCTTTGCTAAGTAATCCTGGACTG
ATAAATAATGCATCCAGTGGCCTACTGCAGGCCGTCCACGAAGACCTCAATGGTTCTCTGG
ATCACATTGACAGCAATGGAAACAGTAGTCCGGGCTGCTCACCTCAGCCGCACATACATTC
AATCCACGTCAAGGAAGAGCCAGTGATTGCAGAGGATGAAGACTGCCCAATGTCCTTAGTG
ACAACAGCTAATCACAGTCCAGAATTAGAAGACGACAGAGAGATTGAAGAAGAGCCTTTAT
CTGAAGATCTGGAACATCATCATCATCATCATtaataaCGACTCAGGCTGCTACaaaaaaaaaaaa
aaaaaaaaaaaaaaaaaaCTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTg
gatccggctggcgtaat

Table S1. IUPAC motifs and MatrixREDUCE refinements from random library binding data.

Human	IUPAC seed sequence	Length	MatrixREDUCE Correlation	MatrixREDUCE TScore	MatrixREDUCE PVal
	GTTTAC	6	0.41	29.13	2.21E-157
	TGTTTAC	7	0.73	49.49	0
	NTGTTTAC	8	0.79	83.58	0
	TRTTGACSN	9	0.83	98.24	0

Chimp	IUPAC seed sequence	Length	MatrixREDUCE Correlation	MatrixREDUCE TScore	MatrixREDUCE PVal
	TRTTKA	6	0.38	25.8	5.88E-130
	TRTTKAN	7	0.59	46.4	0
	NTRTTKAN	8	0.65	59.62	0
	NTGTTKACN	9	0.72	77.14	0

Table S2. MatrixREDUCE motif results from individual experiments.

experiment	Motif	Length	Correlation v	PVal	TScore	PVal
0418chimp	TRTTKAC	7	0.6889247	3.92E-89	24.164	1.60E-91
0418chimp	TGTTKAC	7	0.6889247	3.92E-89	24.164	1.60E-91
0418chimp	TTAWTTAT	8	0.81064421	6.06E-147	42.6826	4.10E-187
0418chimp	NTRTTKAC	8	0.81064398	6.06E-147	42.6825	4.11E-187
0418chimp	NTGTTKAC	8	0.81064398	6.06E-147	42.6825	4.11E-187
0418chimp	NTGTTKACC	9	0.82267477	6.19E-155	45.3253	2.06E-199
0418chimp	TTADTTATB	9	0.82267477	6.19E-155	45.3253	2.06E-199
0418chimp	KGTAAACAN	9	0.82267477	6.19E-155	45.3253	2.06E-199
0418chimp	TTAWTTATB	9	0.82267477	6.19E-155	45.3253	2.06E-199
0418chimp	NTGTTKACN	9	0.82267477	6.19E-155	45.3253	2.06E-199
0418chimp	TTGTTKACC	9	0.82267477	6.19E-155	45.3253	2.06E-199
0418chimp	WTGTTKACM	9	0.82267477	6.19E-155	45.3253	2.06E-199
0418chimp	NTRTTKACN	9	0.82267477	6.19E-155	45.3253	2.06E-199
0418chimp	NTGTTKACM	9	0.82267477	6.19E-155	45.3253	2.06E-199
0418chimp	GGTMAACAA	9	0.82267477	6.19E-155	45.3253	2.06E-199
0418chimp	TTGTTDACC	9	0.82267477	6.19E-155	45.3253	2.06E-199
0418chimp	TTAWTTATN	9	0.82267477	6.19E-155	45.3253	2.06E-199
0418chimp	NTTGTTDACC	10	0.83634173	8.81E-165	52.8818	1.59E-232
0418chimp	NTTAWTTATBN	11	0.85421957	3.96E-179	58.5277	2.25E-255
0418chimp	NWTGTTKACMN	11	0.8542028	4.09E-179	58.5275	2.25E-255
0418chimp	NTTGTTDACCN	11	0.8542028	4.09E-179	58.5275	2.25E-255
0503chimp	TRTTKAC	7	0.57783826	3.76E-109	33.547	4.27E-175
0503chimp	TGTTKAC	7	0.57781296	3.86E-109	33.5433	4.56E-175
0503chimp	TACWGTAA	8	0.63975607	7.81E-141	41.3788	4.11E-234
0503chimp	TTAWTTAT	8	0.6397542	7.83E-141	41.3787	4.12E-234
0503chimp	NTRTTKAC	8	0.6397542	7.83E-141	41.3787	4.12E-234
0503chimp	NTGTTKAC	8	0.6397542	7.83E-141	41.3787	4.12E-234
0503chimp	CGTAAAHGN	9	0.68838404	2.31E-171	49.0455	3.67E-290
0503chimp	NTGTTKACV	9	0.68838404	2.31E-171	49.0455	3.67E-290
0503chimp	NTGTTKACN	9	0.68838404	2.31E-171	49.0455	3.67E-290
0503chimp	GGTHAACAA	9	0.68836609	2.38E-171	49.044	3.77E-290
0503chimp	NTRTTKACN	9	0.68836609	2.38E-171	49.044	3.77E-290
0503chimp	NTRTTKACV	9	0.68836608	2.38E-171	49.044	3.77E-290
0503chimp	TTAWTTATN	9	0.68836608	2.38E-171	49.044	3.77E-290
0503chimp	NTTAWTTATN	10	0.70747991	4.54E-185	57.3044	0
0503chimp	GGTHAACAAN	10	0.70747991	4.54E-185	57.3044	0
0503chimp	NGGTHAACAA	10	0.70737568	5.43E-185	54.2855	0
0503chimp	NTGTTKACVN	10	0.70737568	5.43E-185	54.2855	0
0503chimp	NTRTTKACVN	10	0.70737568	5.43E-185	54.2855	0
0503chimp	NAATRWCAAN	10	0.70531482	1.85E-183	56.0434	0
0503chimp	AATGTHWACN	10	0.70459639	6.27E-183	56.6491	0
0503chimp	NHBGTMAACAN	11	0.7251513	8.81E-199	61.555	0
0503chimp	NAATGTHWACN	11	0.72020367	7.78E-195	59.8877	0
0503chimp	NGGTHAACAAN	11	0.71550818	3.61E-191	59.7591	0
0601chimp	TGTTKAC	7	0.79826089	1.85E-132	35.572	4.09E-149
0601chimp	GTMAACA	7	0.79824767	1.88E-132	35.5723	4.08E-149
0601chimp	TGTTKACN	8	0.95720449	5.657051644	94.0688	0
0601chimp	NGTMAACA	8	0.95720449	5.657051644	94.0688	0
0601chimp	NTGTTKACN	9	0.95871807	0	96.3811	0
0601chimp	NGTMAACAN	9	0.95871807	0	96.3811	0
0601chimp	CGTAAWYAMY	10	0.95963363	0	93.0205	0
0601chimp	CGTAAWYACHK	11	0.77402129	1.32E-119	30.0823	2.35E-121
0624chimp	GTAAAYA	7	0.69853774	6.28E-94	22.4458	1.69E-82
0624chimp	WTGTTTAC	8	0.85228857	5.03E-180	45.9401	1.35E-203
0624chimp	GTAAAYAN	8	0.85228857	5.03E-180	45.9401	1.35E-203
0624chimp	NWTGTTTAC	9	0.87580866	4.16E-202	54.7134	6.42E-242
0624chimp	NWTGTTTACN	10	0.879051	1.68E-205	55.5876	1.60E-245
0624chimp	NAAACMADMAN	11	0.86419929	1.09E-190	51.5498	1.47E-228
0416human	AACCAAA	7	0.94095761	3.16E-150	49.8235	1.99E-151
0416human	AACCAAAN	8	0.94830809	4.58E-159	53.1844	2.06E-159
0416human	NAACCAAAN	9	0.96621854	1.52E-187	68.1572	1.28E-190
0422human	GTAAACA	7	0.80904703	2.59E-147	26.8667	1.74E-106
0422human	TGTTKAC	7	0.80904703	2.59E-147	26.8667	1.74E-106
0422human	NTGTTKAC	8	0.91505035	3.22E-250	56.8442	5.31E-250
0422human	NAACAAAVA	9	0.95909898	0	86.5514	0
0422human	AAACAAVCA	9	0.95909898	0	86.5514	0
0422human	AAACATTCN	9	0.95909898	0	86.5514	0
0422human	AAACAADCA	9	0.95909898	0	86.5514	0
0422human	TKTTTGTTN	9	0.95909898	0	86.5514	0
0422human	NGTAAACAN	9	0.91845631	1.45E-255	57.9729	1.64E-254
0422human	NTGTTTACM	9	0.91845631	1.45E-255	57.9729	1.64E-254
0422human	WTGTTTACN	9	0.91845631	1.45E-255	57.9729	1.64E-254
0422human	GGTHAACAAN	10	0.97113825	0	103.469	0
0422human	NGGTHAACAAN	11	0.97239025	0	114.759	0
0503 human	ACGTAAA	7	0.75618098	2.18E-110	25.9946	9.21E-100
0503 human	CGKYAAYA	8	0.95483939	4.198602222	82.3981	0
0503 human	CGTAAMAGN	9	0.95952641	0	89.2336	0
0503 human	GYTTWSGTTN	10	0.96035499	0	88.7851	0
0503 human	NGYTTWSGTTN	11	0.96011421	0	88.7938	0
0504human	TGTTKAC	7	0.79709424	5.39E-134	27.5622	8.73E-109
0504human	GTMAACA	7	0.79709424	5.39E-134	27.5622	8.73E-109
0504human	NTGTTKAC	8	0.93530528	8.39E-274	64.6359	4.98E-273
0504human	GTMAACAN	8	0.93530528	8.39E-274	64.6359	4.98E-273
0504human	KCGTAAATN	9	0.92930208	1.32E-262	73.1191	1.26E-301
0504human	NGTMAACAN	9	0.92876808	1.17E-261	72.9558	4.21E-301
0504human	TSGTAAAYRM	10	0.92546733	5.85E-256	70.8182	3.88E-294
0504human	NKTVTKGTTTN	11	0.93109711	7.57E-266	68.0945	5.09E-285

Table S3. Systematic mutations of the binding site from oligo175

sequence name	sequence	mutated position
oligo175	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTGTTTACCAATTGTtccggcggtATgac	NA
oligo175 G-3t	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTaCTTGTTTACCAATTGTtccggcggtATgac	-3
oligo175 G-3t	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTtCTTGTTTACCAATTGTtccggcggtATgac	-3
oligo175 G-3c	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTcCTTGTTTACCAATTGTtccggcggtATgac	-3
oligo175 C-2g	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGgTTGTTTACCAATTGTtccggcggtATgac	-2
oligo175 C-2a	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGaTTGTTTACCAATTGTtccggcggtATgac	-2
oligo175 C-2t	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGtTTGTTTACCAATTGTtccggcggtATgac	-2
oligo175 T-1a	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCaTGTTTACCAATTGTtccggcggtATgac	-1
oligo175 T-1g	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCgTGTTTACCAATTGTtccggcggtATgac	-1
oligo175 T-1c	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCcTGTTTACCAATTGTtccggcggtATgac	-1
oligo175 T1a	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTaGTTTACCAATTGTtccggcggtATgac	1
oligo175 T1g	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTgGTTTACCAATTGTtccggcggtATgac	1
oligo175 T1c	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTcGTTTACCAATTGTtccggcggtATgac	1
oligo175 G2c	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTcTTTACCAATTGTtccggcggtATgac	2
oligo175 G2a	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTaTTTACCAATTGTtccggcggtATgac	2
oligo175 G2t	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTtTTTACCAATTGTtccggcggtATgac	2
oligo175 T3a	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTGaTTACCAATTGTtccggcggtATgac	3
oligo175 T3g	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTGgTTACCAATTGTtccggcggtATgac	3
oligo175 T3c	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTGcTTACCAATTGTtccggcggtATgac	3
oligo175 T4c	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTGTcTACCAATTGTtccggcggtATgac	4
oligo175 T4a	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTGTaTACCAATTGTtccggcggtATgac	4
oligo175 T4g	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTGTgTACCAATTGTtccggcggtATgac	4
oligo175 T5a	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTGTTaACCAATTGTtccggcggtATgac	5
oligo175 T5g	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTGTTgACCAATTGTtccggcggtATgac	5
oligo175 T5c	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTGTTcACCAATTGTtccggcggtATgac	5
oligo175 A6c	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTGTTTcCCAATTGTtccggcggtATgac	6
oligo175 A6g	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTGTTTgCCAATTGTtccggcggtATgac	6
oligo175 A6t	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTGTTTtCCAATTGTtccggcggtATgac	6
oligo175 C7a	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTGTTTAaCAATTGTtccggcggtATgac	7
oligo175 C7g	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTGTTTAgCAATTGTtccggcggtATgac	7
oligo175 C7t	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTGTTTAtCAATTGTtccggcggtATgac	7
oligo175 C+1a	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTGTTTACaAATTGTtccggcggtATgac	8
oligo175 C+1t	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTGTTTACtAATTGTtccggcggtATgac	8
oligo175 C+1g	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTGTTTACgAATTGTtccggcggtATgac	8
oligo175 A+2g	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTGTTTACCgATTGTtccggcggtATgac	9
oligo175 A+2c	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTGTTTACCcATTGTtccggcggtATgac	9
oligo175 A+2t	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTGTTTACCtATTGTtccggcggtATgac	9
oligo175 A+3c	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTGTTTACCAcTTGTtccggcggtATgac	10
oligo175 A+3t	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTGTTTACCAtTTGTtccggcggtATgac	10
oligo175 A+3g	cgcCTGTTACGGCATCAGGGCTTTGGTTTGGGATGGCTGCTTGTTTACCAgTTGTtccggcggtATgac	10

Table S4. Intergenic high-confidence ChIP-seq peaks, motif scoring and motif conservation.

peak #	bp to nearest gene	blast hit, closest feature	PSAM score	max local score	hg19 perfect motif instance	unique in humans
5	6269	6269 bp at 3' side: elongation factor 1-gamma	1.06	1.00	chr11:62,320,767-62,320,774	no
8	13316	13316 bp at 3' side: forkhead box protein J2	1.02	1.00	chr12:8,178,994-8,179,000	no
12	26583	34847 bp at 3' side: galectin-3 isoform 2	0.15	0.14	-	-
26	46956	uncharacterized protein LOC100128905	1.06	1.00	chr2:174,890,236-174,890,242	no
28	632370	632370 bp at 5' side: uncharacterized protein LOC100996394	0.02	0.01	-	-
30	12349	12349 bp at 5' side: transmembrane protein 189 isoform 1	0.58	0.50	-	-
31	155694	155694 bp at 5' side: zinc finger protein 217	1.07	1.00	chr20:52,355,178-52,355,184	
37	162353	294928 bp at 5' side: F-box-like/WD repeat-containing protein TBL1XR1	0.12	0.07	-	-
45	14239	33574 bp at 3' side: forkhead box protein P4 isoform 1	1.52	1.00	chr6:41,499,790-41,499,796	no
60	23438	23449 bp at 5' side: vacuolar protein sorting-associated protein 41 homolog	1.19	1.00	chr7:38,972,282-38,972,288	yes
68	150163	150163 bp at 5' side: cyclin-dependent kinase 4 inhibitor B isoform 2	1.09	1.00	chr9:22,159,164-22,159,170	no
71	11593	11593 bp at 5' side: mid1-interacting protein 1	1.01	1.00	chrX:38,676,449-38,676,455	no

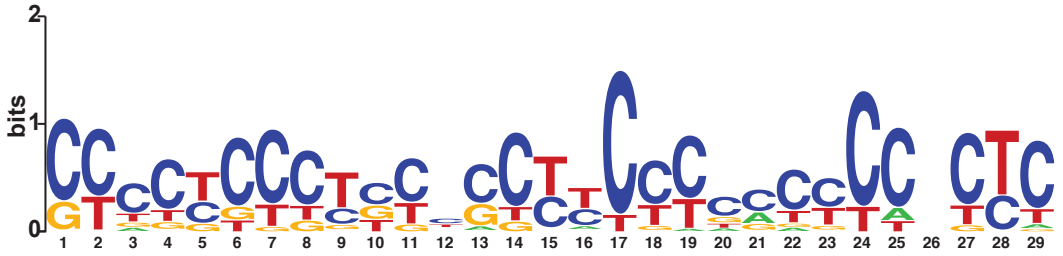


Figure S1. Low information content, C-rich motif from MEME search of ChIP-seq data.

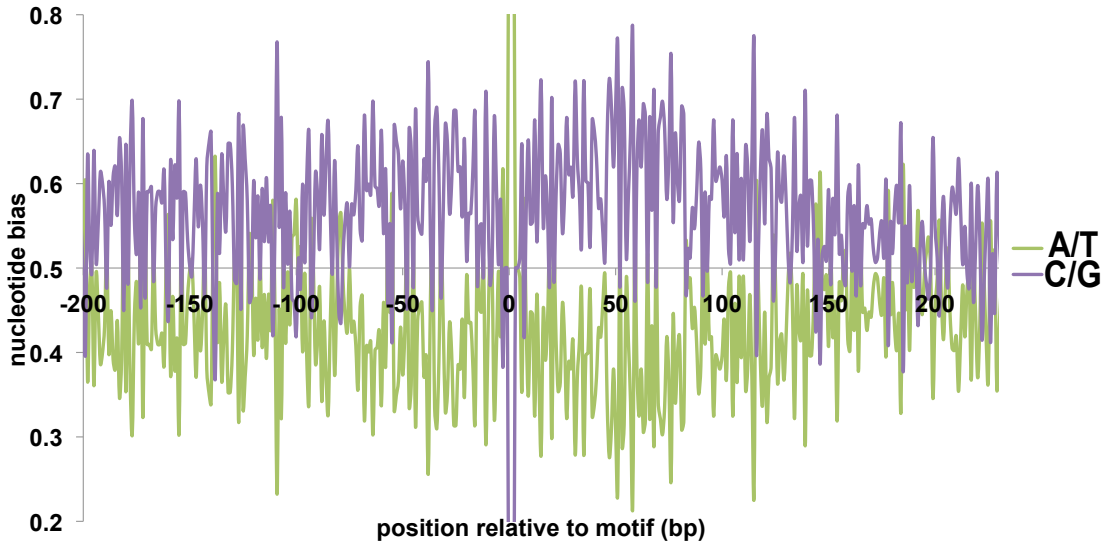


Figure S2. Nucleotide bias surrounding motifs within the 71 consistent ChIP-seq peaks. There is a region of G/C bias surrounding instances of our motif for about 100 bp on both sides.

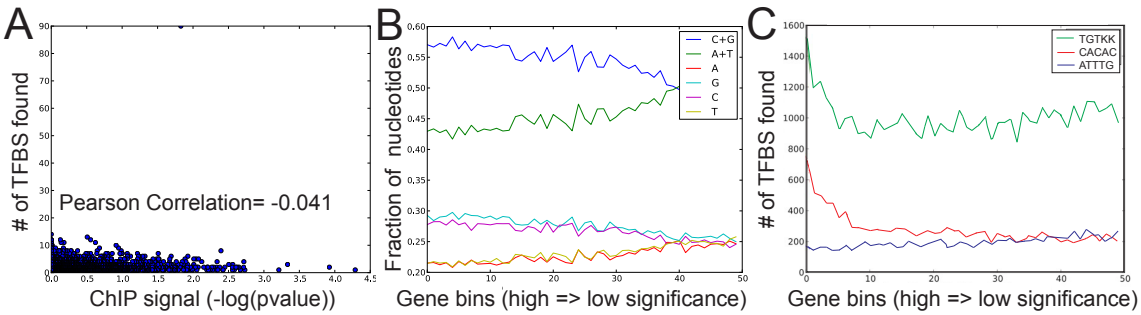
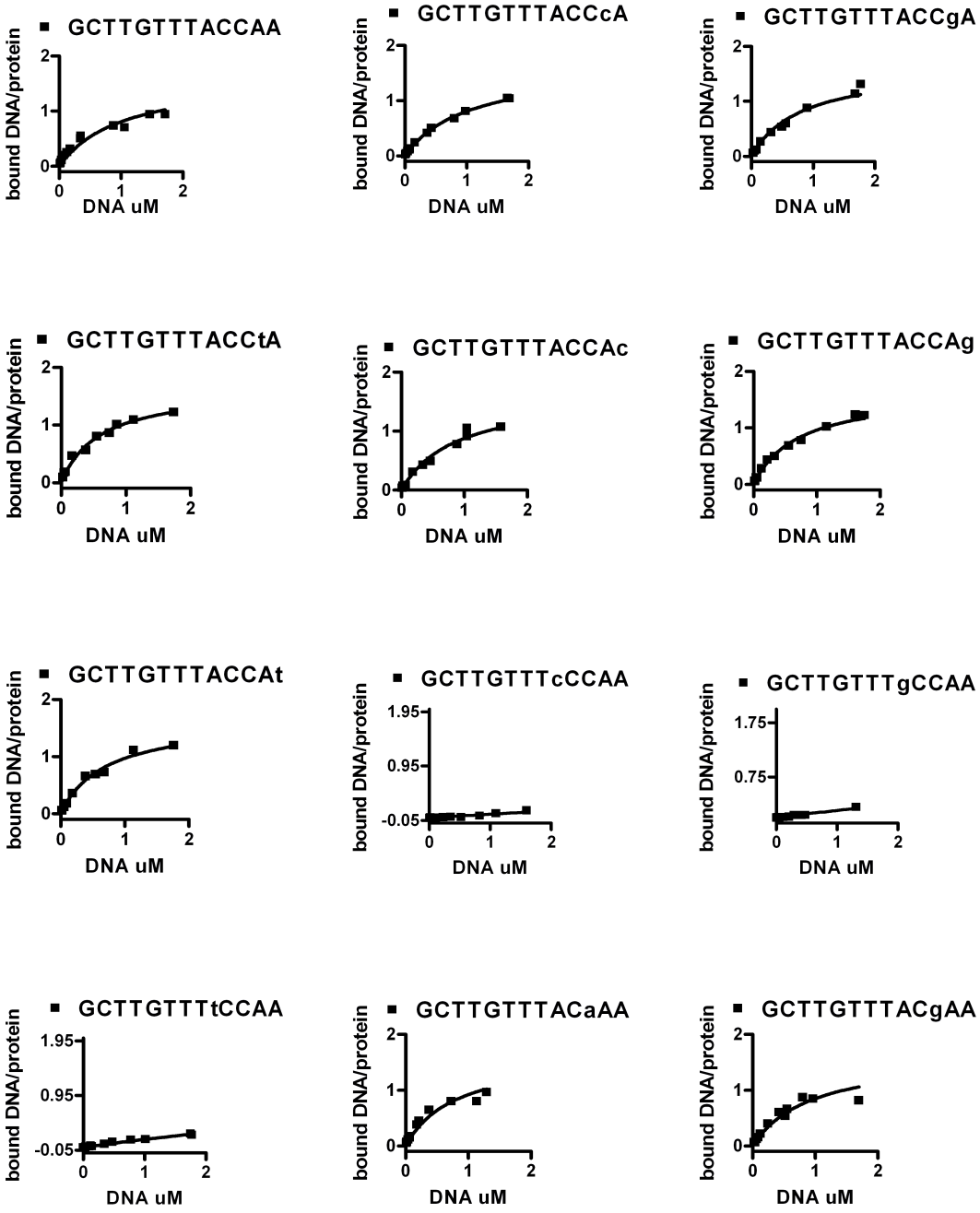
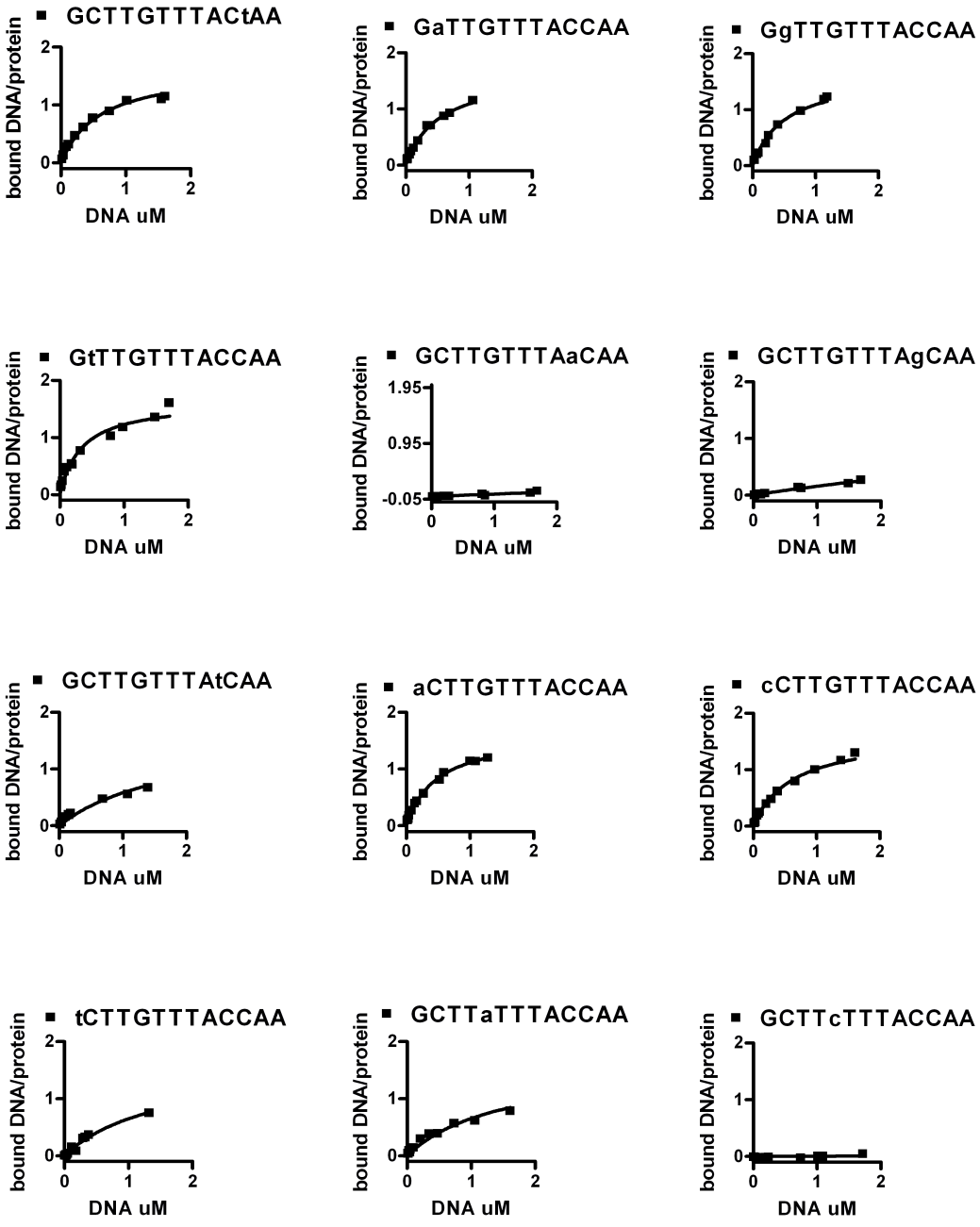


Figure S3. Reanalysis of Vernes et al 2006's ChIP-chip data showing nucleotide biases in significant genes and improved overrepresentation of our reported motifs within significant probes in their data. A. Plot of Vernes et al's model FOXP2 sites found versus ChIP-chip score. There is not a strong correlation between the prediction of the motif model they used and the array t-test statistic plotted along the horizontal axis. B. Nucleotide bias across ChIP probes ranked in order of the signal given by FOXP2 pulldown, with clear G/C bias in the more significant probes. C. Observed minus expected number of FOXP2 sites in the Vernes et al 2006 data, for three different models of the FOXP2 binding site. A 5mer version of our motif "TGTKK" plotted in green and "CACAC" plotted in red show spikes in their representation among the more significant probes on the left-hand side of the chart. In contrast Vernes et al's reported motif of "ATTTG" plotted in blue is at the level of expectation throughout the spectrum of gene scores.

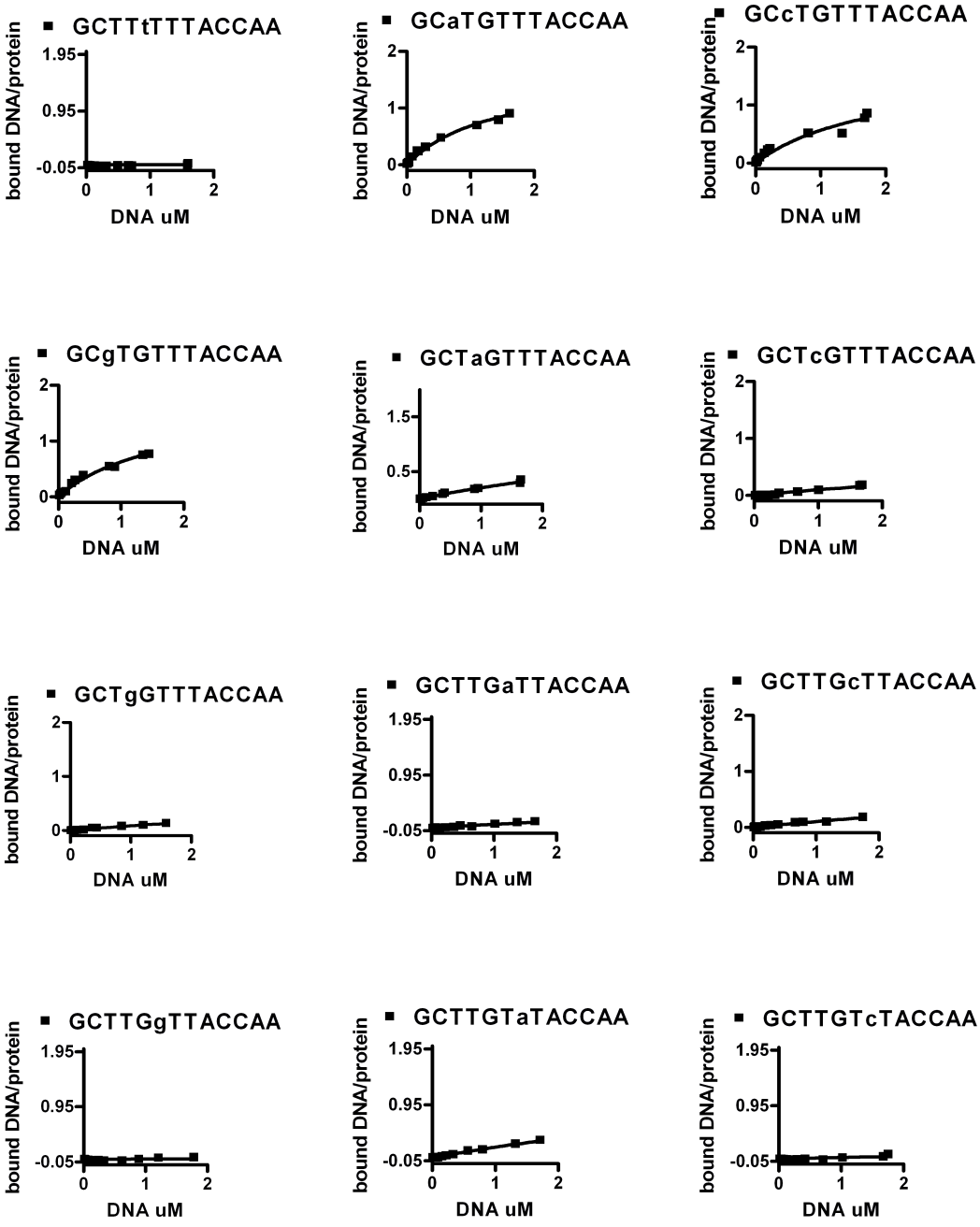
human FOXP2 binding curves



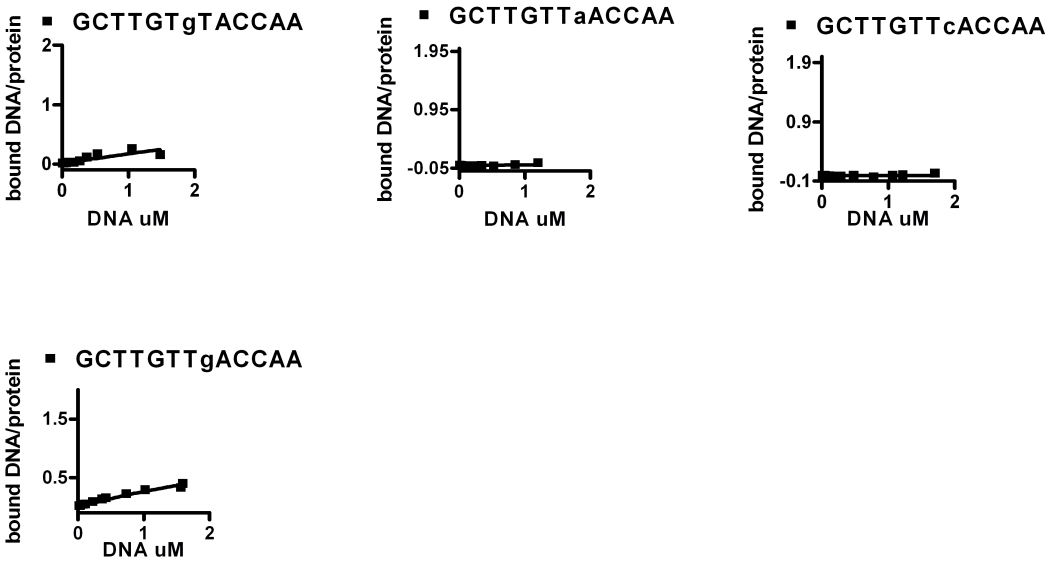
human FOXP2 binding curves



human FOXP2 binding curves

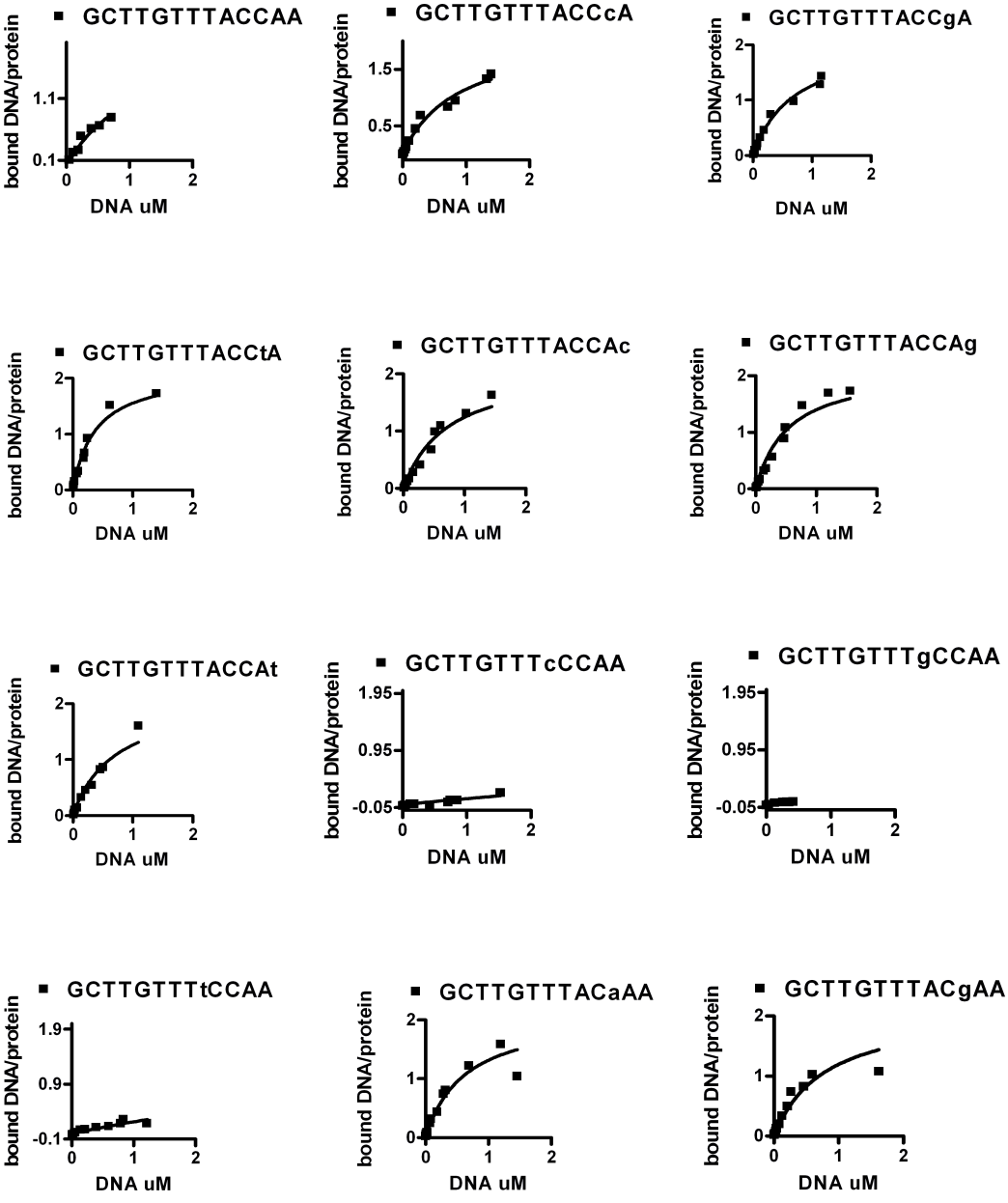


human FOXP2 binding curves

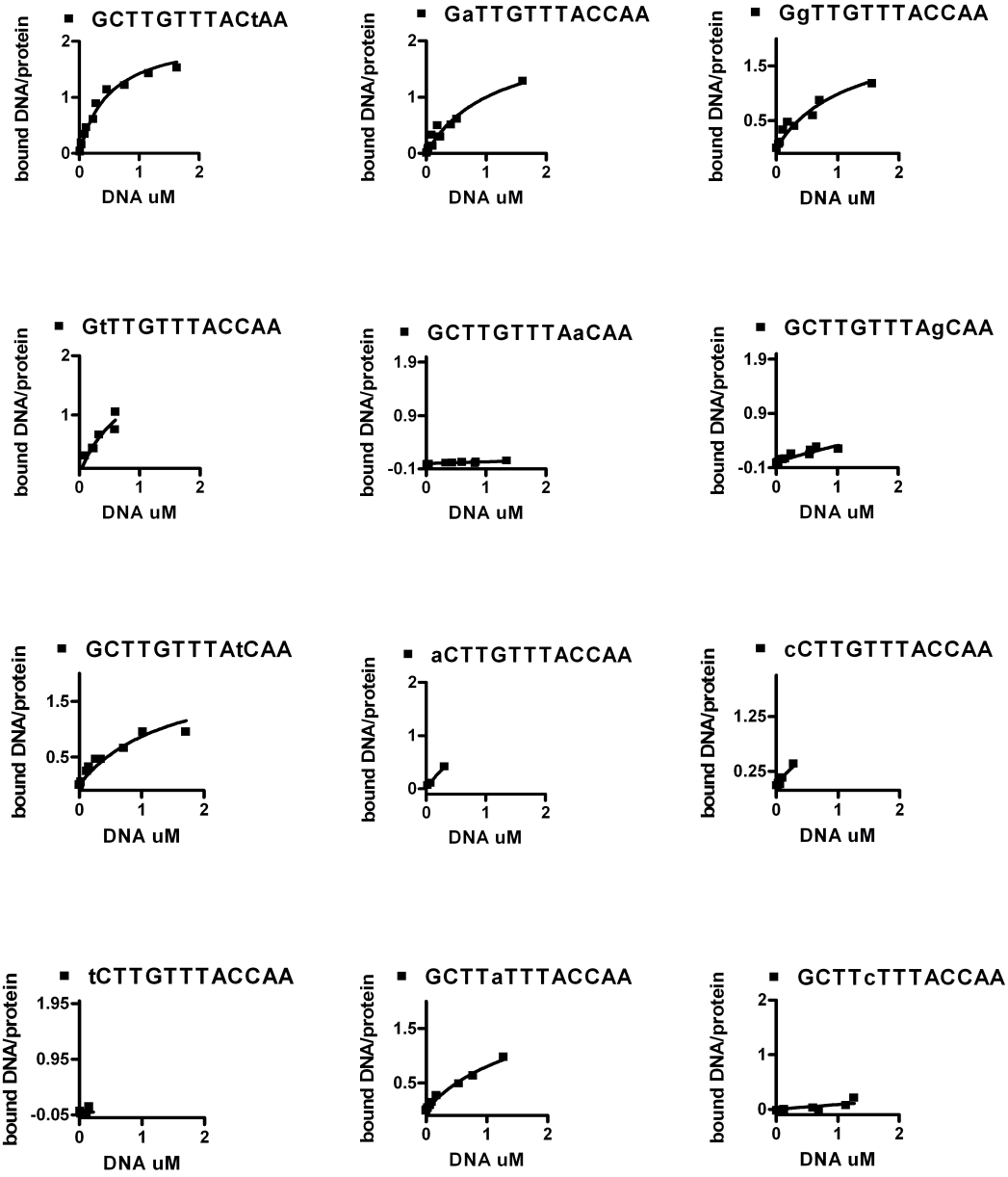


Supplementary Figure 4 Binding curves for human FOXP2 against mutations of the binding site. These curves were used to measure the Ka constants and build our final *in vitro* binding motif.**(page 4 of 4)**

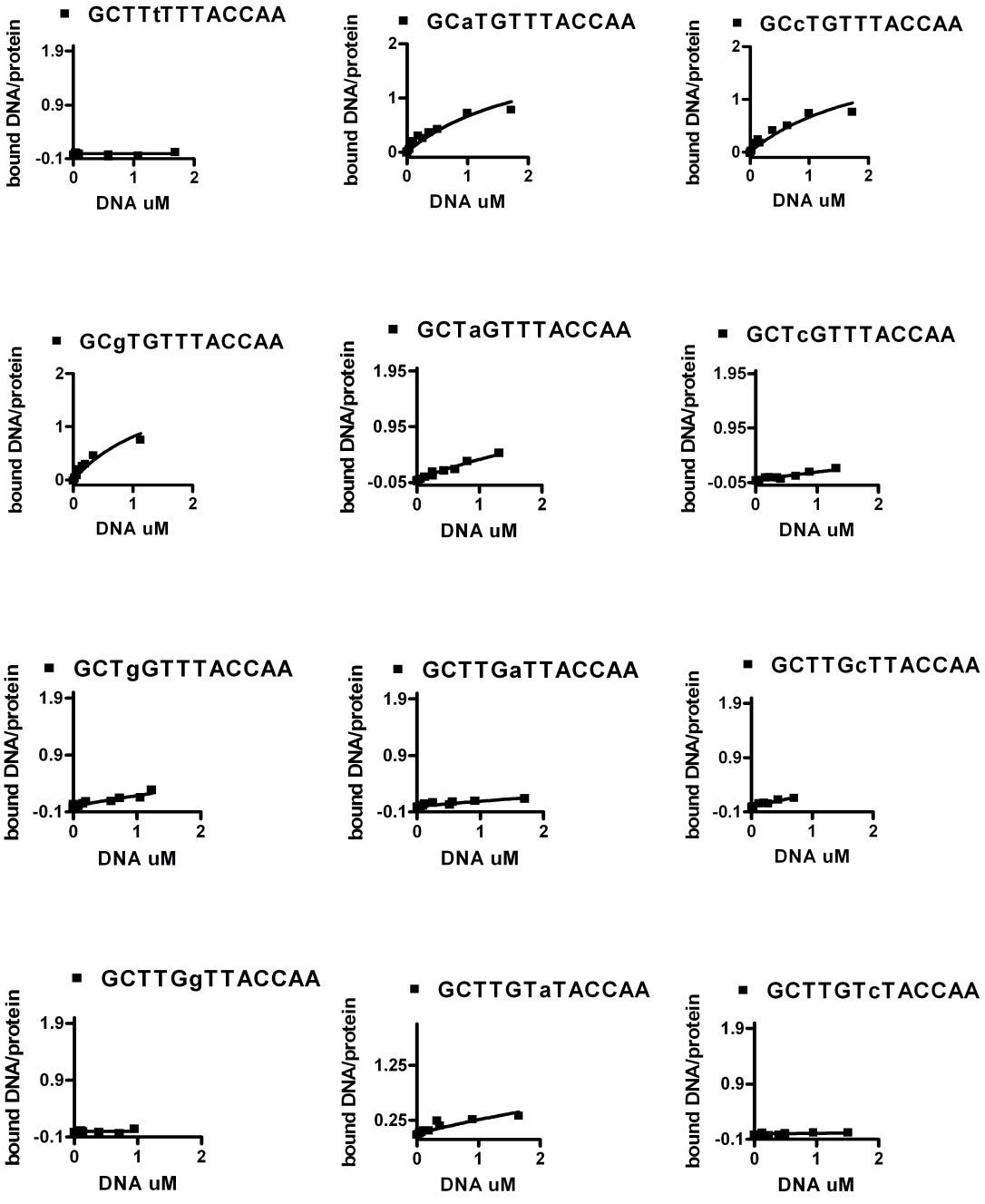
chimp Foxp2 binding curves

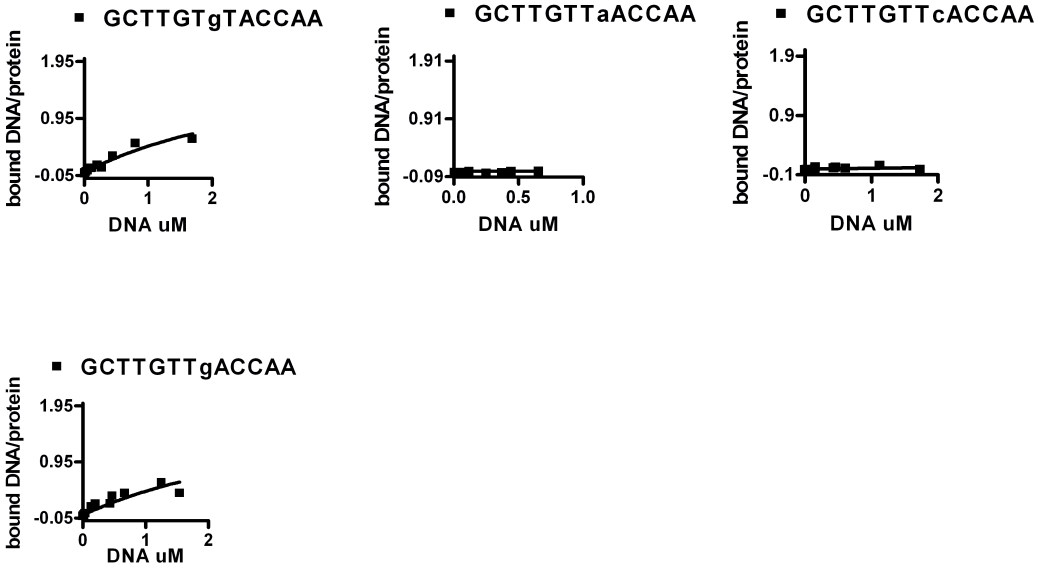


chimp Foxp2 binding curves



chimp Foxp2 binding curves





Supplementary Figure 4 Binding curves for chimp Foxp2 against mutations of the binding site. These curves were used to measure the Ka constants and build our final *in vitro* binding motif.**(page 4 of 4)**

Chapter 8…….

Graded and Co-linear Regulation from the Stress Responsive Factor Msn2

Jacob Stewart-Ornstein, Christopher Nelson, Joe DeRisi, Jonathan S. Weissman, and Hana El-Samad

Author contributions:

Jacob Stewart-Ornstein conceived and designed experiments, and constructed the yeast strains, conducted the flow cytometry reporter assays, and wrote the paper. Christopher Nelson conceived, designed, and performed MITOMI experiments, and helped write the paper. Joe DeRisi, Jonathan S. Weissman, and Hana El-Samad conceived and designed experiments, and helped write the paper.

Joseph L. DeRisi, Thesis Advisor

ABSTRACT

To thrive in challenging and rapidly changing circumstances cells tightly regulate the production of cytoprotective proteins. In the budding yeast *S. cerevisiae* a wide range of stresses evoke the Environmental Stress Response (ESR), which results in the association of the homologous transcription factors Msn2/4 to stress responsive genes. In this work, we show that Msn2 activates gene expression in a graded and uniform manner across transcriptional targets. The stress response system generates a linear relationship between Msn2 activity

and target gene expression through low affinity binding of Msn2 to target genes and an excess of binding sites relative to the quantity of Msn2 protein. These features provide a simple and general mechanism for co-linear activation of target genes, allowing proportionate response to different magnitudes of stress and maintaining stoichiometry within the Msn2/4-responsive program across a wide range of conditions.

INTRODUCTION

In order to maximize their proliferation cells need to respond to environmental cues and insults by activating stress responsive cellular pathways. Occasionally, such adjustments necessitate drastic transitions into protective physiological states characterized by massive remodeling of the cellular transcriptome and proteome. For model organisms such as *S. cerevisiae,* these emergency programs are well documented in response to carbon source switches, starvation, or large temperature changes (Gasch, et al 2000).

In addition to large environmental swings, and arguably more frequently, cells have to contend with small potentially transient perturbations. These modest modulations require proportional adjustments to cellular metabolism and physiology; an adequately responsive system should therefore operate in a graded regime. Unfortunately, our understanding of such homeostatic responses does not match our extensive knowledge of cellular emergency responses, in part because our experimental approaches have routinely relied on large

perturbations. As a result, the principles by which cellular responses can robustly achieve graded operation over a broad dynamic range remain obscure.

To adapt to environmental perturbations cells engage a complex many-gene protective transcriptional program. Although the details of these programs vary, the basic strategy is highly conserved from yeast to mammals—stressful conditions impact the activity of a core set of kinases such as protein kinase A (PKA), TOR, or AMPK. These kinases modulate the activity of a range of transcription factors that, in turn, regulate the expression of protective genes. These networks of kinases, transcription factors, and their gene targets are crucial cellular homeostats that continuously adapt the organism to its fluctuating environment.

In the budding yeast *S. cerevisiae*, PKA and two if its target transcription factors, the paralogous proteins Msn2 and Msn4 play important roles in cellular homeostatic adaptation to carbon source modulation and other stresses, with Msn2 playing a dominant role under most conditions (Broach, 2012;). Activation of Msn2/4 promotes the direct expression of at least 200 genes, including chaperones, the trehalose and glycogen synthase machinery, oxidative stress response, mitochondrial components, and alternate glycolytic enzymes (Boy-Marcotte et al., 1999; Zhu et al., 2009; Huebert et al., 2012). The expression of such a rich transcriptional program, known as the Environmental Stress Response (ESR) raises questions about how a single factor coordinates a large number of processes, and whether target genes exhibit a spectrum of responses to the same Msn2 signal. Previous data documenting the behavior of

the Msn2 transcriptional targets have been measured under strong stress conditions and show little obvious differentiation in timing or magnitude of different target genes (Gasch et al., 2000; Capaldi et al., 2008). Recently, however, more quantitative studies have uncovered subtle differences in target genes response to the duration of Msn2 activity (Hao and O'Shae., 2011; Huebert et la., 2012).

Here, we use a combination of quantitative synthetic tools, in vitro measurements, and computational modeling to systematically dissect the response of target genes to Msn2 activity. We find that Msn2 exhibits non-cooperative binding to its targets, including those whose promoters contain a large number of Msn2 binding sites. This pattern of binding is the synergistic result of low-affinity interactions between Msn2 and its cognate binding site in gene promoters and competition over a large number of Msn2 binding sites in the genome relative to the number of Msn2 molecules. These effects result in a linear relationship between the concentration of active Msn2 protein the expression of its target genes, leading to their co-linear and stoichiometric expression. Linearity is a robust feature of the system and extends over the full dynamic range of its operation, suggesting that Msn2 provides a proportional homeostatic response to stressful conditions. In addition to its properties as an 'emergency' stress response, these results position the ESR as a homeostatic system capable of providing a precisely balanced response across a range of environmental conditions. The strategy that positions Msn2 in this linear regime is simple and general, and may provide a fingerprint to identify these features in

cellular systems. Moreover, the Msn2 strategy constitutes a simple framework that may aid the design of synthetic homeostatic systems.

Results

To map the relationship between the responses of different target genes to the same Msn2 signal, we sought to measure co-expression in the same cell of prTPS1 and prPGM2, two stress responsive promoters that contain 6 and 5 Stress Response Elements (STREs), respectively. To precisely control Msn2 activity, we developed synthetic tools that can tune the concentration of active Msn2 in the nucleus while avoiding the pleiotropic effects of Msn2 activation by stress. To this end, we took advantage of a constitutively active allele of MSN2, Msn2-5A, in which every PKA phosphosite is mutated to alanine (Goner et al., 1998). We achieved copy number control of this Msn2 mutant allele by placing it under a GAL1 promoter in a *Δmsn2/4* strain expressing an estradiol-regulated Gal4 fusion protein (Stewart-Ornstein et al., 2012). This construct allows for graded regulation of the abundance of Msn2-5A by addition of the small molecule estradiol. Since this estradiol-inducible synthetic circuit provides a general strategy for controlling the expression of any protein over a wide dynamic range, we further used it to drive the production of two negative regulators of the PKA pathway, the phosphodiesterase PDE2 that degrades cAMP and a dominant negative allele of RAS2 (S24N). These two constructs was integrated into a wild type (MSN2/4) strains (Figure 1A), providing two additional means for tuning Msn2 activity.

Using these tools, we were able to titrate the activity of Msn2 and simultaneously measure the level of activity of fluorescent proteins expressed from prTPS1 and prPGM2. Since these two promoters contain several STREs, we expected their expression to be a sigmoidal function of active Msn2, resulting in a nonlinear relationship between their respective expression profiles (Figure 1B). Contrary to this expectation, we found that the prTps1-prPgm2 relationship was linear over the whole range of Msn2 activity (Figure 1B, C). This result is particularly surprising given that both the direct titration of Msn2-5A and its activation in the Ras2(S24N) strain accessed the full dynamic range of the ESR system. In fact, estradiol mediated activation of Msn2 in these strains resulted in the induction of prPGM2 or prHSP12 to levels exceeding(~4 fold higher) those observed in heat shock or mid-stationary phase (Figure 1D).

Targets of Msn2 show co-linear activation

To test whether this linear relationship was a general feature of Msn2 transcriptional regulation of its targets, we measured the expression of a large number of other Msn2 target genes. We identified 40 such genes from microarray studies (Hao and O'Shae, 2011; Capaldi et al., 2008; Gasch et al., 2000), and monitored their expression using a fluorescent reporter of their promoter activity in a strain harboring the estradiol responsive synthetic circuit driving either MSN2-5A, RAS2dn, or PDE2. Of these promoters, 32 showed measureable basal expression and greater than four-fold induction in one of the Msn2 perturbations and lost at least 50% of this induction in a *Δmsn2/4* background. These results were reproducible and specific to PKA perturbations

as overexpression of a second dominant negative allele of RAS2 (G22A) resulted in nearly identical induction as S24N (Fig. S1a). Overexpression of a constitutively active allele of Msn4 (4A -- allele) also resulted in similar patterns to those generated by induction to Msn2(5A) (S1a-d). This set was characterized further. For the bulk of these promoters all perturbations gave similar results, therefore we focused on Msn2(5A) as it is the most direct approach to regulating MSN2 activity.

As with prTPS1, most of the 32 characterized promoters (27/32), titration of Msn2-5A produced superimposable and co-linear relationships with prPgm2 (F1C, S1a). Co-linear relationships were dependent on direct Msn2 binding, as removal of the Msn2 binding consensus STREs from the promoters of two such genes (SSA1 (2 sites), and SSA4 (3 sites)) ablated induction upon Msn2(5A) overexpression (F1E, panels 1 and 2, mutants are in red).

In contrast to the 27 promoters that exhibited co-linearity, the remaining five promoters (last panels of F1E) showed a thresholded behavior: expression from these promoters was insensitive to low amounts of Msn2(5A), while an increase of Msn2 beyond a certain threshold elicited a co-linear expression with that of prPGM2. Of the five promoters that showed the thresholded relationship with prPGM2, two (HSP26, SIP18) had been suggested in a previous study to be potentially regulated by chromatin structure (Hao and O'Shae, 2012). To explore chromatin structure as the root of the thresholded behavior, we sought to alter this structure by insertion of poly-T sequences, which have been shown to disrupt nucleosome positioning (Raveh-Sadka et al., 2012). We inserted either

one or two poly-T sequences (12xdT) into the prHSP26 construct and measured its co-expression with prPgm2. Consistent with chromatin playing a role in the thresholded behavior of prHSP26, the insertion of these sequences increased the expression of prHSP26 and rendered it strongly co-linear with prPGM2 (Figure 1F). By contrast, insertion of poly-dT sequences into the promoters of three strongly co-linear genes (PGM2, TPS1, or PNC1) resulted in no substantive change in their expression (data not shown). Taken together, these results indicate that Msn2 predominantly activates of its downstream genes co-linearly. The exceptions to this linear regulation exhibit a thresholded behavior due to other contributions such as chromatin architecture—as we observe for prHSP26—or possibly due to regulation by other transcription factors. Irrespective, these promoters exhibit co-linear activation with other Msn2 targets once the threshold set by other regulatory factors is passed, and recover their co-linear activation for the full range once the thresholding influences are ablated.

Msn2 activates promoters proportional to its concentration

Traditionally, the presence of multiple binding sites for transcriptional regulators in gene promoters is thought to produce gene regulatory functions of increasing steepness (Hill coefficient) as a result of cooperative binding. Our results don't conform to this notion. Instead, the co-linearity in the expression of these promoters argues that gene expression is instead linearly related to the concentration of nuclear Msn2, even in cases where these promoters contain a large number of STREs.

The relationship between Msn2 and a promoter it regulates such as prPGM2 depends on at least two distinct kinetic steps: the translocation of Msn2 into the nucleus and binding of nuclear Msn2 to prPGM2, resulting in the production of RFP. The most parsimonious model accounting for these two steps represents the translocation of Msn2 in and out of the nucleus as first order processes proceeding at a rate K_{in} and K_{out} respectively, and the rate of production of proteins (RFP in this example) as a simple linear function of nuclear Msn2 (Supplementary materials). In addition, in this model, Msn2 is assumed to be produced in the cytoplasm at a constant rate and degraded both in the cytoplasm and nucleus with first order kinetics. The degradation rate in the nucleus is denoted g_n. At steady-state, this model predicts that steady state RFP, driven by the PGM2 promoter, as a function of total Msn2 is given by the simple expression:

$$[RFP] \propto \frac{k_{in}}{k_{in} + k_{out} + \gamma_n}[Msn2_{total}]$$

With these simplifications, this model predicts that if the residence of Msn2 into the nucleus were to be increased by either increasing k_{in} or decreasing k_{out}, then the relationship between prPGM2 expression and total Msn2 would still be linear, but with an increased slope (Figure 2A). This is a non-trivial prediction resulting from the assumptions that translocation of Msn2 into and out of the nucleus is a first order and that its binding and transcription of target genes is a linear function. This prediction can be validated by taking advantage of different Msn2 alleles with an increasing number of PKA phosphosite substitutions (Msn2-Xa, where X=0,1,2,3,4,5). These alleles span a range of nuclear localizations, from

limited (Msn2-0A) to constitutive nuclear localization (Msn2-5A) (Figure 2C, S2A).

To directly test this model we simultaneously monitored the expression of Msn2(Xa)-YFP whose concentration can be titrated using the estradiol synthetic circuit and RFP driven by the Pgm2 promoter (Fig 2B) expressed in the same cell. Induction by estradiol resulted in increasing concentrations of Msn2, and revealed that a strong linear relationship indeed exists between prPGM2 and total Msn2 concentration (Figure 2D). Consistent with the model predictions, the expression of prPGM2-RFP was linearly related to the concentration of the Msn2 alleles in all cases. The slope of the line relating prPGM2 and any one Msn2 allele was an increasing function of the allele's nuclear localization. These patterns were not unique to prPGM2, as the linear and ordered relationship also existed between the different Msn2 alleles and prHSP12 and prTPS1 (Figure S2). This uniformity of behavior across the spectrum of Msn2 alleles demonstrate that neither its nuclear import nor export dynamics contribute to the linear activation behavior.

Further, we note that total Msn2 consistently decreased in abundance for the same estradiol induction in the more active alleles, confirming previous observations of an increased rate of degradation of Msn2 in the nucleus (Chi et al., 2004; Durchschlag et al, 2004).

Msn2 binds non-cooperatively to its target promoters

We next sought to provide further evidence for the non-cooperative nature of Msn2 binding to its target promoters, focusing on the Pgm2 promoter, which has

five consensus STREs in the 500bp preceding the start codon. In the absence of cooperativity, the binding of Msn2 to any one STRE is independent from its binding to other STREs present in the same promoter. In this case, of prPGM2 expression in any two single STRE mutants should be quantitatively predictive of expression of the double mutant. To test this prediction, we focused on the Pgm2 promoter, which has five consensus STREs. We mutated each individual STRE in series—we chose to use a minimal single base pair substitution (AGGGG->AGaGG) that ablates in vivo activity of the binding site--creating five single mutant alleles. We then constructed all ten possible double mutant alleles and measured the expression of each of the wild type and 15 mutant prPGM2 promoters. The expression of all double mutant promoters was quantitatively well approximated as a product of the activity of the constituent single mutants (figure 2e). These data are in strong agreement with simple non-cooperative binding of Msn2 to STREs.

Msn2 binds STREs with low affinity

Since Msn2 does not exhibit any cooperativity in its binding to STREs, so we approximated the amount of Msn2 bound to a given promoter with a widely used model of transcription factor interactions (Michaelis-Menten kinetics):

$$\frac{[Msn2.DNA]}{[DNAtotal]} = \frac{[Msn2]}{K_d + [Msn2]}$$

We note that this model assumes that the transcription factor is largely unbound and in excess of the DNA binding sites, an assumption we will revisit later. The number of Msn2 molecules in the cell has been estimated to be 125 (Ghaemmaghami et al., 2003). Even overexpressed from a very strong Gal1

promoter we estimate there are not more than 1000-2000 active Msn2 molecules in the nucleus, likely due to the rapid degradation of active Msn2 (S2). Given a nuclear volume of 4-10pl (Jorgensen et al., 2007) this results in a concentration of Msn2 of ~0.1uM to low micromolar. We would therefore expect a linear relationship between [$Msn2_{total}$] and $Msn2_{dna}$ if the Kd is large relative to the [$Msn2_{total}$] (fig 3a).

To test this model, we sought to measure directly the binding affinity of Msn2 to STREs. We took advantage of MITOMI 2.0, an *in vitro* microfluidic technique that determines the absolute binding affinity of a given transcription factor to its cognate binding sites (Maerkl and Quake, 2007, Fordyce et al., 2010). The MITOMI 2.0 procedure works as follows: Individual cells of the microfluidic device are programmed with fluorescently labeled double-stranded oligonucleotides, arrayed at pre-determined dilutions. Using a separate fluorophore, labeled His-tagged protein (Msn2) is introduced into the device, solubilizing the deposited DNA sequences. Within each cell, anti-His antibodies, deposited below a “button” valve, recruit Msn2-DNA complexes, and these binding interactions may then be mechanically trapped by the activation of the button valve. Unbound material may then be washed away, leaving complexes at their equilibrium concentrations. By imaging, the ratio between bound protein and bound DNA may be determined beneath the button. The binding occupancy curves and the free DNA concentration are then fit and K_ds of each interaction measured (F3b). Using this approach, we measured the *in vitro* affinity of Msn2 for 27 oligonucleotide sequences each consisting of 40 bp centered at a single

genomic consensus STRE binding motif ('AGGGG'), in the event two STREs fell into this range we mutated one of the sequences to a non-consensus site. These sequences were those taken from endogenous promoters of genes that are strongly responsive to Msn2 activity (PGM2, HSP12, TPS1, HSP26, RTC3, or CYC7, supplementary table 4, S3). Measured K_d values ranged from 0.2μM to 4uM (Fig3C).

To further explore the contribution of flanking nucleotides, we selected a single binding site from the PGM2 promoter and measured Msn2 binding affinity for different bases in the first position of the 'NAGGGG' sequence (F3D). Consistent with previous data, Msn2 had a significantly lower K_d for the 'AAGGGG' motif than for motifs that have C/G/T in the first position. "TAGGGG" showed the highest K_d, while "GAGGGG" showed the second highest (Fig 3D). To validate these measurements, we determined expression from a promoter containing four copies of each of these motifs fused to a crippled CYC1-YFP promoter. In agreement with the *in vitro* data, the 'AAGGGG' motif, which has the lowest K_d, showed substantially higher activity, followed by "CAGGGG", "GAGGGG" and "TAGGGG" (F3E). Further, examining the genome as a whole we find almost two-fifths (38%) of NAGGGG sites have 'A' in the first position and this fraction increases slightly as one examines only promoters with four or more binding sites. Similarly, we see a significantly reduced frequency of the weakly binding 'T' in the first position consistent with a model where more strongly Msn2 sensitive promoters show increased tendency towards strong binding sites (S3A).

Low affinity binding is not caused by non-canonical Zinc-Finger linker arrangements

The Msn2 DNA binding domain consists of two tandem zinc fingers connected by a short linker. For many zinc finger transcription factors, this linker region is strongly conserved and has been shown to exert strong influence on DNA binding affinity (Wuttke et al., 1997). Intriguingly, the linker region of Msn2 is divergent from consensus sequences, having both increased spacing between invariant histidine residues immediately before the linker and also a strongly diverged sequence within the linker itself (Fig 3C). To investigate whether the Msn2 divergent linker might explain the protein's relatively low binding affinity to DNA, we engineered Msn2 alleles with linkers conforming to the consensus sequences. Previous work that has explored the function of the linker residues which suggests that perturbations should not affect DNA binding specificity, but only affinity (Bernstein et al., 1994; Wuttke et al., 1997; Laity et al., 2000, Kochoyan et al., 1991), drawing on this extensive literature we constructed three MSN2 mutants with altered linker arrangements, one construct (H allele) which converted the HX_4H spacing to a more conventional HX_3H spacing by removing the final valine (V669), the second construct (T allele) converted the linker sequence to the consensus (S671T, N672G,R674K), and finally an allele that combined these two mutations (HT).

Based on previous data we suspected that the H and T alleles would by themselves reduce affinity, but in combination move the protein towards the consensus and increase the affinity without altering the binding properties (F3F,

see supplement for details). Consistent with these our predictions, we found that the H allele had strongly decreased and the T allele had slightly decreased ability to activate the prHSP12 (Figure 3H). The HT allele has increased ability of activate prHSP12, although the increase was modest. Affinity measurements by MITOMI 2.0 show that as expected the H allele reduces affinity by ~2-fold. Somewhat surprisingly, the HT allele showed a relatively small binding site dependent change in affinity but on average appeared to show similar affinity to wild type. Further, as one would predict given that the linker residues are not expected to contact DNA these alleles appear not to have altered preference for DNA binding (F 3G, S3). Overexpression of these alleles retained the linear inductions of prHSP12 (fig. 3H), although as expected the HT and H allele showed increased and reduced activity respectively, suggesting that at least within the affinity regimes we could experimentally obtain linear induction was not strongly sensitive to affinity or linker arrangements.

A competitive model of Msn2 interactions with promoters may contribute to the linear interactions

A simple binding model taking our estimation of high nano-low micromolar concentration of Msn2 and the MITOMI 2.0 measured binding affinities into account suggests that whether the fraction of bound STREs follows a linear relationship with the abundance of Msn2 depends on how many available binding sites are in the genome. When many fewer Msn2 molecules than STREs are available, competition between STREs for binding to Msn2 results in operation that is far from saturation and in a linear binding regime (supplement, F4A).

There are 8450 consensus STREs (AGGGG) in the genome, roughly half of which (4122) are present in promoter regions (defined here as DNA sequences 700bp upstream of a start codon). Consistent with these sequence based estimates, recent *in vivo* measurements of Msn2 binding identified 1290 associated loci—many of which contain multiple STRE sites—that associate with Msn2 during stress conditions (Huebert et al., 2012).

In general the linearity of the relationship between the concentration of the transcription factors and the concentration of the DNA bound species depends on the affinity and the number of binding sites. With both weak affinity for STREs and large numbers of potential binding sites relative to its molecular numbers, Msn2 appears to be positioned robustly in a linear regime (F4A). This model assumes that promoter can be described as the sum of their individual binding sites as supported by our data (F2).

Testing the prediction that this excess of binding sites contributes to the linear relationship we observe would require reducing the number of MSN2 binding sites in the genome, which is not experimentally tractable. An alternative experimental approach would be to engineer Msn2 itself to recognize DNA sequences that are less abundant in the genome.

Mutations to the Msn2 DNA binding domain can abrogate its linear activation of promoters

To define a low occurrence sequence motif, we scanned a database of the *S. cerevisiae* promoter sequences for the number of occurrences of GNNGNN, which is a sequence motif known to be bound by canonical zinc fingers.

Interestingly we note that the single least common site ‘GTCGGG’ (416 occurrences) matches the consensus of the proteosomal regulator Rpn4 and that the fifth least common site ‘GCGGGG’ matches the Mig1 consensus (Harbison et al., 2004; Zhu et al., 2009). These results hint that either random transcription factor binding is deleterious so there is selection pressure to lose non-specific sites or that transcription factors binding sites selection is influenced by the frequency of related sites in the genome. Of the least common such sequences, the 9th on this list (‘GGGGGG’, 540 occurrences) is not known to be a target of any know transcription factor in yeast. To switch Msn2 recognition from AAGGGG to GGGGGG, we made two mutations to its second zinc finger (Q693R, and N690H or N690K). These mutations are predicted to alter Msn2 specificity from AAG to GGG, with high affinity for the N690H (6G(H)) and low affinity for the N690K (6G(K)) variants (fig 4B; Segal et al., 1999).

The Msn2-6G(H) and Msn2-6G(K) alleles, tested in a *msn2/4Δ* strain, specifically bound and activated a CYC1 promoter whose UAS was modified to contain three 6xG sequences (pr6GCYC-YFP). At the same time, these alleles did not induce any activity in a CYC1 promoter containing the consensus WT STRE (AGGGG, Figure 4C). However, the wild type Msn2 allele was able to bind and activate the 6xG promoter, albeit to a lesser degree than our engineered alleles, further emphasizing Msn2’s binding promiscuity. These data suggest that the mutant alleles can be used to investigate whether the Msn2 binding linearity can compromised by a combination of increased binding affinity and decreased opportunity for binding (Figure 4A, arrow).

To do so, we titrated Msn2-6G(H) and Msn2-6G(K) using the estradiol synthetic circuit and measured the expression of a pr6GCYC-YFP. The expression we observed was a nonlinear and saturating function of both the Msn2-6G(H) and Msn2-6G(K) alleles. This relationship was well fit by a simple Michaelis-Menten binding model ($R^2>0.95$). A more complex model taking into account the presence of low affinity sites achieved slightly more exact fits without changing the qualitative results (Figure 4C, supplement). Additionally, the Msn2-6G(K) allele showed weaker binding with less maximum expression and a two-fold lower apparent KD for the promoter (F4C). This data illustrates that Msn2 can be switched from its low affinity graded regime to a saturating regime with two amino acid changes. These results suggest that the graded nature of Msn2 binding to its promoters is a feature of the system, not a consequence of biophysical constraints.

Discussion

Our experiments show that transcription factor binding to promoters can be surprisingly straightforward and linear and that first principal models taking affinities and molecular numbers into account can explain these relationships. The use of synthetic biology tools to precisely set the concentrations of the active transcription factor Msn2 rather than much more pleiotropic—albeit more physiological—stress conditions that are typically used allowed us to quantitatively measure the dose response relationship between a transcription factor and many different target promoters.

Our measurements show a strictly graded activation of gene expression by the Msn2 system, which combined with other work in mammalian and yeast systems suggests that far from being switch-like many stress responsive systems are graded and capable of precise dose dependant activation (Giorgetti et al, 2010; Sadeh et al., 2012). This should perhaps come as no surprise as many stress responsive systems are tightly embedded in negative feedback loops that insure regulated and limited expression (Wang et al., 2010; Lahav et al., 2004), in these circumstances highly cooperative activation of downstream genes would be counter-productive.

We also note that this linear coupling between the transcription factor and promoter needs to be considered in the context of the PKA regulatory system, which produces dynamic changes in Msn2 activity in response to stress. Low affinity interactions with DNA allow for rapid (sub-second) binding and un-binding of Msn2 to its response elements, enabling rapid dynamic control of Msn2 activity. Our experiments were performed at relatively slow timescales, purposely ignoring dynamic features to focus on the steady state regulatory relationships, if however the rate limiting step for promoter activation is transcription factor binding then the low affinity interactions of Msn2 argue that the dynamic response of the promoters will mirror the steady state. In many cases other steps may be rate limiting such as chromatin remodeling and as this would predict different genes seem to show different dynamics at the minute timescales in response to Msn2 activity as has been demonstrated (Hao et al., 2011).

Further, obtaining this linear relationship does not require complex dynamic mechanisms but simply requires low affinity and low transcription factor abundances. The model we derive for competition between binding sites for limiting transcription factors requires only that binding sites substantially outnumber transcription factors, a situation that is common in eukaryotic systems where relatively small and frequently degenerate binding motifs are combined with large genomes. Certainly there are ways of avoiding this competitive binding, by occluding unused sites with nucleosomes, or using co-factors or cooperative binding to increase specificity, but for monomeric transcription factors with moderate affinities these results may prove general.

There are two benefits for maintaining a linear response to transcription factor concentration, first co-linear activation of target genes allows for stoichiometric expression of large groups of genes without extensive promoter tuning, and second that low affinity linear interactions allow for precise tuning of promoter activity by addition of binding sites. Although speculative, the benefits of this approach are likely substantial as a handful of mutations are sufficient to alter Msn2 pattern of binding from a linear to rapidly saturating, suggesting evolutionary pressure has maintained Msn2 in this linear regime. These features also potentially make the Msn2 regulon very 'evolvable' with binding site changes modifying the quantitative relationship between different genes without changing quantitative features, consistent with this stress responsive networks in fungi show rapid rewiring across evolutionary timescales (Rhind et al., 2011; Wapinski et al., 2007; Tirosh et al., 2011).

Experimental Procedures

Yeast Strains

All yeast strains used for these experiments are derived from W303A-1 in which the *ade2* marker was reverted to ADE2+ to reduce the autoflorescence. Promoter constructs were integrated at the TRP1 locus of a HIS3+ MATa strain. Overexpression constructs were integrated into the TRP1 locus of a LEU2+ MATalpha strain, which contained the estradiol inducible construct. These strains were then mated and diploids selected in SD-leu/his media. All strains were constructed using standard yeast protocols and LioAc/PEG transformation, for a complete list of strains and plasmids see supplementary table 1/2.

Growth and fluorescence measurements by flow cytometry

For all measurements cells were grown to saturation in 96-shallow well plates (Costar) and then diluted into fresh media, grown at 30C on orbital shakers (Elim) for 12hrs to an OD of ~0.5. These cells were subsequently diluted and estradiol added as necessary and grown for 6hrs to an OD of ~0.05 before measurement. Expression of the estradiol-regulated system was activated by addition of 0-200nm estradiol (stock of 1.6mM in 90/10 mixture of EtOH and DMSO (sigma)), typically applied in a log1.6 titration series.

All cytometry measurements were made on a Becton Dickinson LSRII flow cytometer, along with an autosampler device (HTS) to collect data over a sampling time of 6-12 seconds, typically corresponding to 2000-10000 cells. GFP and YFP were excited at 488nm, and fluorescence was collected through a HQ530/30 bandpass filters (Chroma), mCherry and mKate2 were excited at 561 nm and fluorescence collected through 610/20 bandpass filter (Chroma) .

Microscopy and image analysis

Cells expressing Msn2-YFP or related constructs were plated in SD complete onto ConcanavalinA coated 96 well glass bottom plates, allowed to settle and then washed twice with fresh media. Samples were imaged on a Nikon Ti inverted scope with arc-lamp illumination using RFP(560/40nm excitation, 630/75nm emission, Chroma) and YFP (510/10nm excitation, 542/27nm emission, Semrock) filters. Images were processed and analyzed with ImageJ and Matlab. Nuclear enrichment was computed by dividing the average intensity of the brightest 10pixels in the cell by the median intensity of the cell.

Flow cytometry data analysis

All data was analyzed with custom Matlab software. Raw cytometry data were filtered to remove errors due to uneven sampling and remove outliers using an MCD method (Rousseau and Van Driessen, 1999). Variability in cell size was corrected using a linear transformation from the side scatter parameter (see supplemental materials for detail).

Sequence Analysis

The latest S288C genome was downloaded from SGD (yeastgenome.org) and the DNA sequence was processed and examined with custom matlab scripts. Promoters were operationally defined as 700bp upstream of the start (ATG) codon of the gene.

References

Bernstein BE, Hoffman RC, Horvath S, Herriott JR, Klevit RE (1994) Structure of a histidine-X4-histidine zinc finger domain: insights into ADR1-UAS1 protein-DNA recognition. Biochemistry **33**(15):4460-70.

Berry DB, Gasch AP (2008). Stress-activated genomic expression changes serve a preparative role for impending stress in yeast. Mol Biol Cell. 19(11):4580-7.

Boy-Marcotte E, Lagniel G, Perrot M, Bussereau F, Boudsocq A, Jacquet M, Labarre J (1999). The heat shock response in yeast: differential regulations and contributions of the Msn2p/Msn4p and Hsf1p regulons. Mol Microbiol. 33(2):274-83.

Boy-Marcotte E, Garmendia C, Garreau H, Lallet S, Mallet L, Jacquet M (2006). The transcriptional activation region of Msn2p, in Saccharomyces cerevisiae, is regulated by stress but is insensitive to the cAMP signalling pathway. Mol Genet Genomics. **275**(3):277-87.

Camus C, Hermann-Le Denmat S, Jacquet M (1995). Identification of guanine exchange factor key residues involved in exchange activity and Ras interaction. Oncogene, **11**(5):951-9.

Cai L, Dalal CK, Elowitz MB (2008). Frequency-modulated nuclear localization bursts coordinate gene regulation. Nature. 455(7212):485-90.

Capaldi AP, Kaplan T, Liu Y, Habib N, Regev A, Friedman N, O'Shea EK (2008). Structure and function of a transcriptional network activated by the MAPK Hog1. Nat Genet. **40**(11):1300-6.

Chi Y, Huddleston MJ, Zhang X, Young RA, Annan RS, Carr SA, Deshaies RJ (2001). Negative regulation of Gcn4 and Msn2 transcription factors by Srb10 cyclin-dependent kinase. Genes Dev. **15**(9):1078-92.

Dreier B, Beerli RR, Segal DJ, Flippin JD, Barbas CF 3rd. (2001). Development of zinc finger domains for recognition of the 5'-ANN-3' family of DNA sequences and their use in the construction of artificial transcription factors. J Biol Chem. **276**(31):29466-78.

Durchschlag E, Reiter W, Ammerer G, Schüller C (2004). Nuclear localization destabilizes the stress-regulated transcription factor Msn2. J Biol Chem. 279(53):55425-32.

Ferguson SB, Anderson ES, Harshaw RB, Thate T, Craig NL, Nelson HC (2005). Protein kinase A regulates constitutive expression of small heat-shock genes in an Msn2/4p-independent and Hsf1p-dependent manner in Saccharomyces cerevisiae. Genetics. 169(3):1203-14.

Finn RD, Clements J, Eddy SR (2011). HMMER web server: interactive sequence similarity searching.Nucleic Acids Res. **39**(Web Server issue):W29-37.

Gasch AP, Spellman PT, Kao CM,Carmel-Harel O, Eisen MB,Storz G, Botstein D, Brown PO (2000). Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. Molecular Biology of the Cell **11**: 4241–4257.

Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS (2003). Global analysis of protein expression in yeast. Nature. 425(6959):737-41.

Giorgetti L, Siggers T, Tiana G, Caprara G, Notarbartolo S, Corona T, Pasparakis M, Milani P, Bulyk ML, Natoli G (2010). Noncooperative interactions between transcription factors and clustered DNA binding sites enable graded transcriptional responses to environmental inputs. Mol Cell. 37(3):418-28.

Görner W, Durchschlag E, Martinez-Pastor MT, Estruch F, Ammerer G, Hamilton B, Ruis H, Schüller C (1998). Nuclear localization of the C2H2 zinc finger protein Msn2p is regulated by stress and protein kinase A activity. Genes Dev. **12**, 586-597.

Görner W, Durchschlag E, Wolf J, Brown EL, Ammerer G, Ruis H, Schüller C (2002). Acute glucose starvation activates the nuclear localization signal of a stress-specific yeast transcription factor. EMBO J. **21**(1-2):135-44.

Hahn S, Young ET (2011). Transcriptional regulation in Saccharomyces cerevisiae: transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. Genetics. 189(3):705-36.

Hao N, O'Shea EK (2011). Signal-dependent dynamics of transcription factor translocation controls gene expression. Nat Struct Mol Biol. **19**(1):31-9.

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KS, Lander ES, Gifford DK, Fraenkel E, and Young RA (2004). transcriptional regulatory code of a eukaryotic genome. Nature **431**:99-104.

Huebert DJ, Kuan PF, Keles, S, Gasch AP (2012) Dynamic changes in nucleosome occupancy are not predictive of gene expression dynamics but are linked to transcription and chromatin regulators. Mol Cell Biol. 32(9):1645-53.

Jorgensen P, Edgington NP, Schneider BL, Rupes I, Tyers M, Futcher B (2007). The size of the nucleus increases as yeast cells grow. Mol Biol Cell. 9:3523-32.

Kochoyan M, Keutmann HT, Weiss MA (1991). Alternating zinc fingers in the human male-associated protein ZFY: HX3H and HX4H motifs encode a local structural switch. Biochemistry. **30**(39):9396-402.

Lahav G, Rosenfeld N, Sigal A, Geva-Zatorsky N, Levine AJ, Elowitz MB, Alon U (2004). Dynamics of the p53-Mdm2 feedback loop in individual cells. Nat Genet. 2004 Feb;36(2):147-50.

Laity JH, Dyson HJ, Wright PE (2000). DNA-induced alpha-helix capping in conserved linker sequences is a determinant of binding affinity in Cys(2)-His(2) zinc fingers. J Mol Biol, **295**(4):719-27.

Mumberg D, Müller R, Funk M (1994). Regulatable promoters of Saccharomyces cerevisiae: comparison of transcriptional activity and their use for heterologous expression. Nucleic Acids Res. 22(25):5767-8.

Pomerantz JL, Sharp PA, Pabo CO (1995). Structure-based design of transcription factors. Science. 1995 **267**(5194):93-6.

Poullet P, Créchet JB, Bernardi A, Parmeggiani A (1995). Properties of the catalytic domain of sdc25p, a yeast GDP/GTP exchange factor of Ras proteins. Complexation with wild-type Ras2p, [S24N]Ras2p and [R80D, N81D]Ras2p. Eur J Biochem, **227**(1-2):537-44.

Rhind N, Chen Z, Yassour M, Thompson DA, Haas BJ, Habib N, Wapinski I, Roy S, Lin MF, Heiman DI, Young SK, Furuya K, Guo Y, Pidoux A, Chen HM, Robbertse B, Goldberg JM, Aoki K, Bayne EH, Berlin AM, Desjardins CA, Dobbs E, Dukaj L, Fan L, FitzGerald MG, French C, Gujja S, Hansen K, Keifenheim D, Levin JZ, Mosher RA, Müller CA, Pfiffner J, Priest M, Russ C, Smialowska A, Swoboda P, Sykes SM, Vaughn M, Vengrova S, Yoder R, Zeng Q, Allshire R, Baulcombe D, Birren BW, Brown W, Ekwall K, Kellis M, Leatherwood J, Levin H, Margalit H, Martienssen R, Nieduszynski CA, Spatafora JW, Friedman N, Dalgaard JZ, Baumann P, Niki H, Regev A, Nusbaum C (2008). Comparative functional genomics of the fission yeasts. Science, **332**(6032):930-6.

Sadeh A, Movshovich N, Volokh M, Gheber L, Aharoni A (2012). Fine-tuning of the Msn2/4-mediated yeast stress responses as revealed by systematic deletion of Msn2/4 partners. Mol Biol Cell. 22(17):3127-38.

Segal D.J., Dreier B, Beerli RR, Barbas III CF (1999). Toward controlling gene expression at will, selection and design of zinc finger domains recognizing each of the 5'-GNN-3' DNA target sequences. Proc. Natl Acad. Sci. USA, **96**;2758–2763.

Spitz F and Furlong EEM (2012). Transcription factors: from enhancer binding to developmental control. Nature Reviews Genetics 13, 613-626

Slattery MG, Liko D, Heideman W (2007). Protein kinase A, TOR, and glucose transport control the response to nutrient repletion in Saccharomyces cerevisiae. Eukaryot Cell. 7(2):358-67.

Stewart-Ornstein J, Weissman JS, El-Samad H (2012). Cellular noise regulons underlie fluctuations in Saccharomyces cerevisiae. Mol Cell. **45**(4):483-93.

Tirosh I, Wong KH, Barkai N, Struhl K (2011). Extensive divergence of yeast stress responses through transitions between induced and constitutive activation. Proc Natl Acad Sci U S A. 108(40):16693-8.

Wang X, Xu H, Ha SW, Ju D, Xie Y (2010). Proteasomal degradation of Rpn4 in Saccharomyces cerevisiae is critical for cell viability under stressed conditions. Genetics. 184(2):335-42.

Wapinski I, Pfeffer A, Friedman N, Regev A (2007). Natural history and evolutionary principles of gene duplication in fungi. Nature, 449(7158):54-61.

Wharton RP, Brown EL, Ptashne M (1984). Substituting an alpha-helix switches the sequence-specific DNA interactions of a repressor. Cell. 38(2):361-9.

Wuttke DS, Foster MP, Case DA, Gottesfeld JM, Wright PE (1997). Solution structure of the first three zinc fingers of TFIIIA bound to the cognate DNA sequence: determinants of affinity and sequence specificity. J Mol Biol. **273**(1):183-206.

Zhu C, Byers KJ, McCord RP, Shi Z, Berger MF, Newburger DE, Saulrieta K, Smith Z, Shah MV, Radhakrishnan M, Philippakis AA, Hu Y, De Masi F, Pacek M, Rolfs A, Murthy T, Labaer J, Bulyk ML (2009). High-resolution DNA-binding specificity analysis of yeast transcription factors. Genome Res. **19**(4):556-66.

Figure Legends

Figure 1—Msn2 Regulated Genes are Induced Activated Co-Linearly (a)Overexpression of a constitutively active allele of MSN2 or PKA inhibitors (PDE2 or RAS2dn) result in activation of Msn2 target genets. (b) A model of gene expression where target genes are induced linearly by MSN2, linear induction by MSN2 leads to co-linear relationships across genes. (c) The promoters of PGM2 and TPS1 show co-linear activation over a wide range of expression levels when induced by either Msn2(5A) or PKA inhibitors. Each dot represents the average expression each promoter at a given concentration of the inducer estradiol. (d) the HSP26 promoter and PGM2 show a non-monotonic relationship which differs between PKA inhibition and Msn2(5A) overexpression. (e) Each plot shows the expression of one promoter-YFP fusion under a range of induction levels of Msn2(5A), plotted against the expression of prPGM2-YFP on the X-Axis. The red dots in SSA1 and SSA4 show the expression of those

promoters when mutated to removed the Msn2 binding (STRE) elements. (f) the promoter of HSP26 shows a thresholded response to Msn2(5A) activation, this response can be converted to a monotonic linear relationship by insertion of one or two chromatin disrupting polyT elements.

Figure2—Gene expression is linear with respect to MSN2, can be described by a very simple model

(a)The expression of Msn2-YFP is induced by estradiol and the activity of a downstream PGM2 promoter monitored by RFP production. (b) A model of the response of a Msn2 sensitive promoter to overexpression of increasingly active Msn2 alleles. (c) Nuclear localization (d) Measurement of prPGM2 expression by RFP florescence ploted against Msn2-YFP abundance. (e) Mutational analysis of prPGM2 promoter, mutations were made to each of the five consensus Msn2 binding sites, all ten possible double mutants were also constructed, expression of each of these promoters was measured. A simple multiplicative model was used to predict the expression of the double mutants from the single mutant data (x-axis), this model was compared to the actual double mutant values and a best fit line of linear model plotted. Errorbars represent standard error (n=3).

Figure3—Msn2 affinity measurements and binding

(a) A simple model of transcription factor binding to DNA shows linear relationships when the affinity is low. (b) MITOMI measurements of MSN2 affinity for DNA. (c)Histogram of the affinity of Msn2 for a range of binding sites from the promoters of in vivo targets of Msn2. (d) MITOMI measurements of

Msn2 preference for the first nucleotide in the binding site. (e) *in vivo* measurements of expression changes in a 4xSTRE promoter with a single nucleotide difference, expression is measured by YFP florescence (AU). (f)Design of high and low affinity alleles of Msn2, the conservation between MSN2/MSN4 is shown along with the amino acid changes in each allele. The crucial zinc coordinating residues are highlighted in red, the mutational regions in green or blue. (g) MITOMI data on H/HT alleles showing similar binding preferences, but altered affinity. (h) The H and T alleles show reduced expression of HSP12 and the HT allele somewhat increased expression.

Figure4—Testing a model of co-linear induction by plurality of target gene sequestration.

(a)Model describing the relationship between number of binding sites and binding affinity that govern the linearity of the relationship between transcription factor concentration and DNA binding. (b)Two point mutations to MSN2 convert its specificity from AAGGGG to GGGGGG. (c) Msn2(6G) specifically activates a construct with 3x6G sites, and shows no cross activation with the wild type Msn2 Binding sites. (d) Msn2(6H), in green, shows saturating activation of the transcriptional reporter, with the K allele (in blue) showing ~2 fold reduced affinity. Dashed lines represent fits to the [TF]/([TF]+kd).

FIGURES

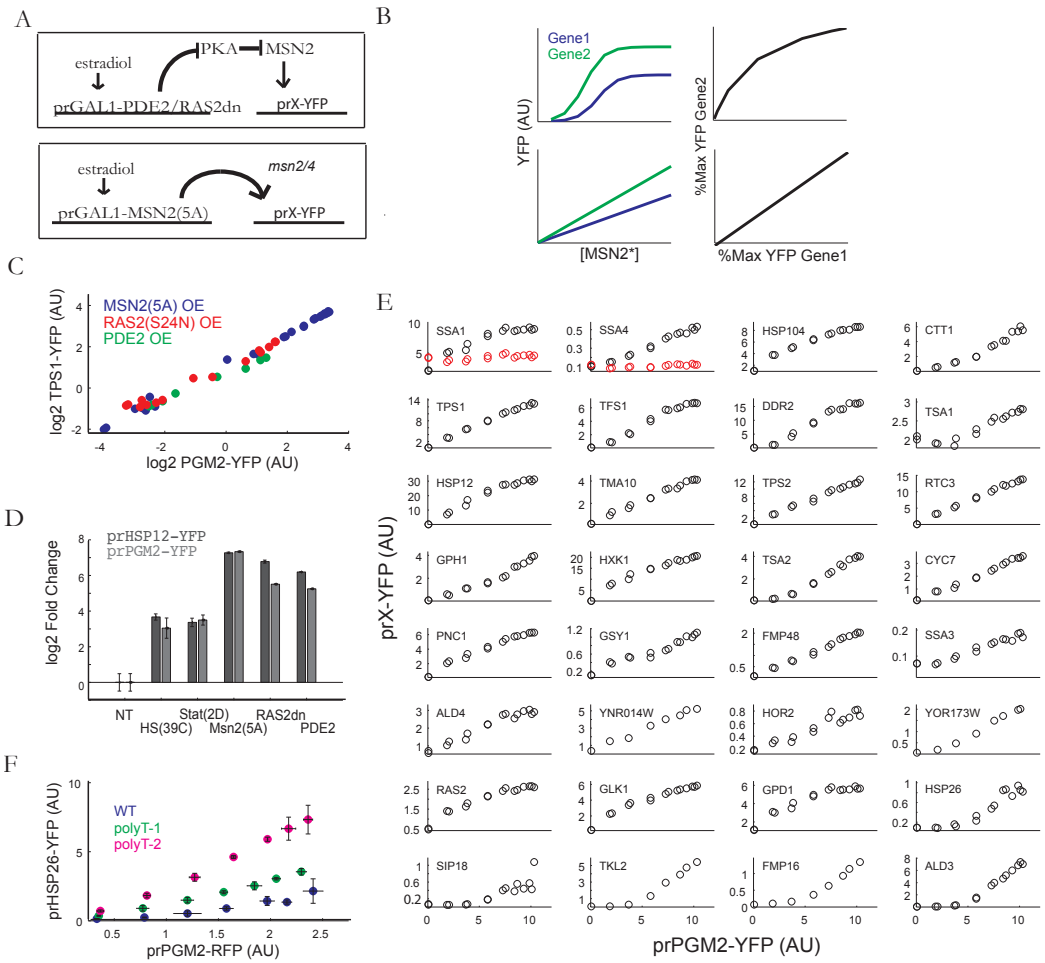


Figure 1

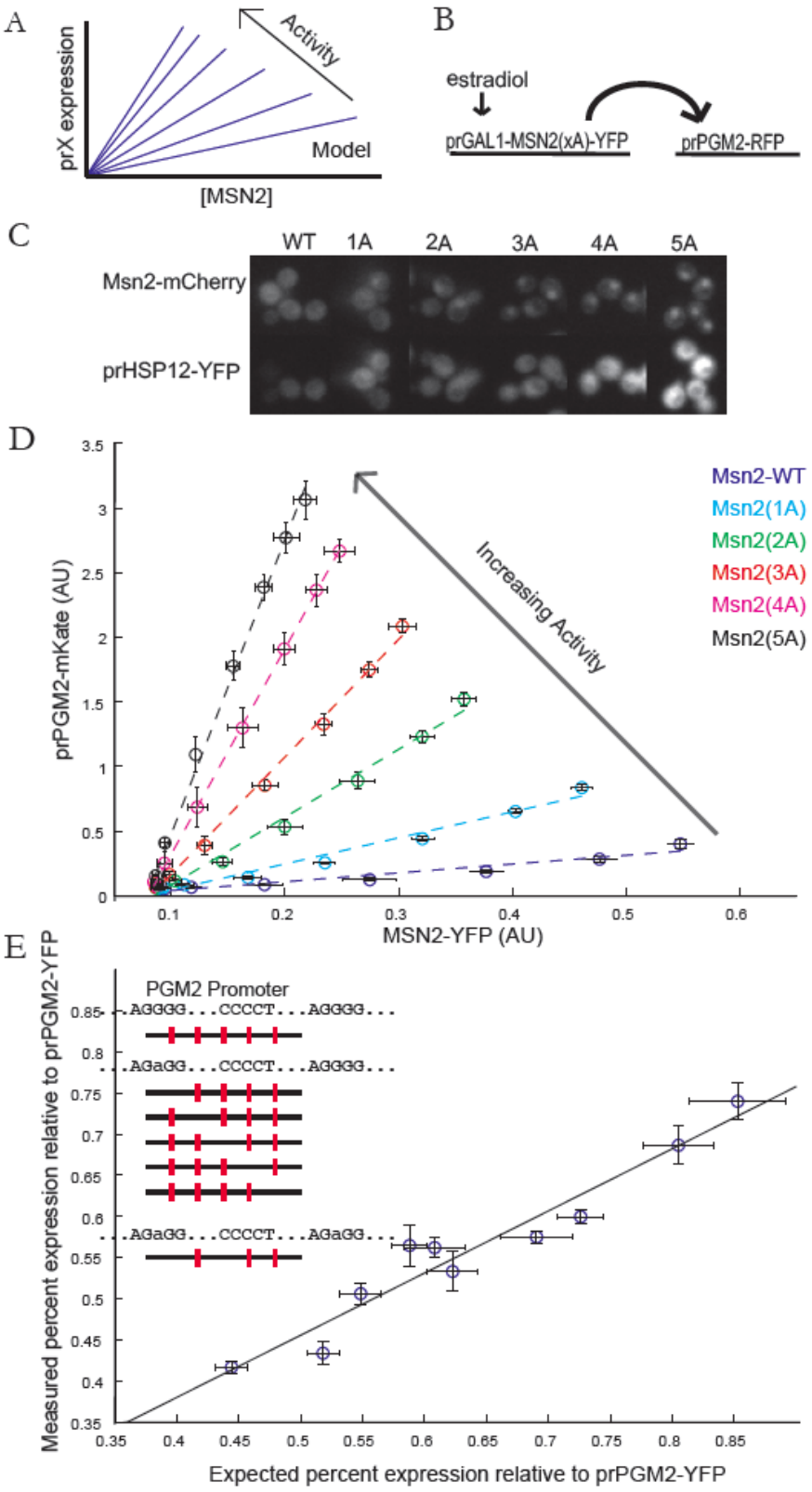

A
prX expression
Activity
Model
[MSN2]
B
estradiol
prGAL1-MSN2(xA)-YFP
prPGM2-RFP
C
WT
1A
2A
3A
4A
5A
Msn2-mCherry
prHSP12-YFP
D
prPGM2-mKate (AU)
MSN2-YFP (AU)
Msn2-WT
Msn2(1A)
Msn2(2A)
Msn2(3A)
Msn2(4A)
Msn2(5A)
Increasing Activity
E
PGM2 Promoter
AGGGG...CCCCT...AGGGG...
AGaGG...CCCCT...AGGGG...
AGaGG...CCCCT...AGaGG...
Measured percent expression relative to prPGM2-YFP
Expected percent expression relative to prPGM2-YFP

Figure 2

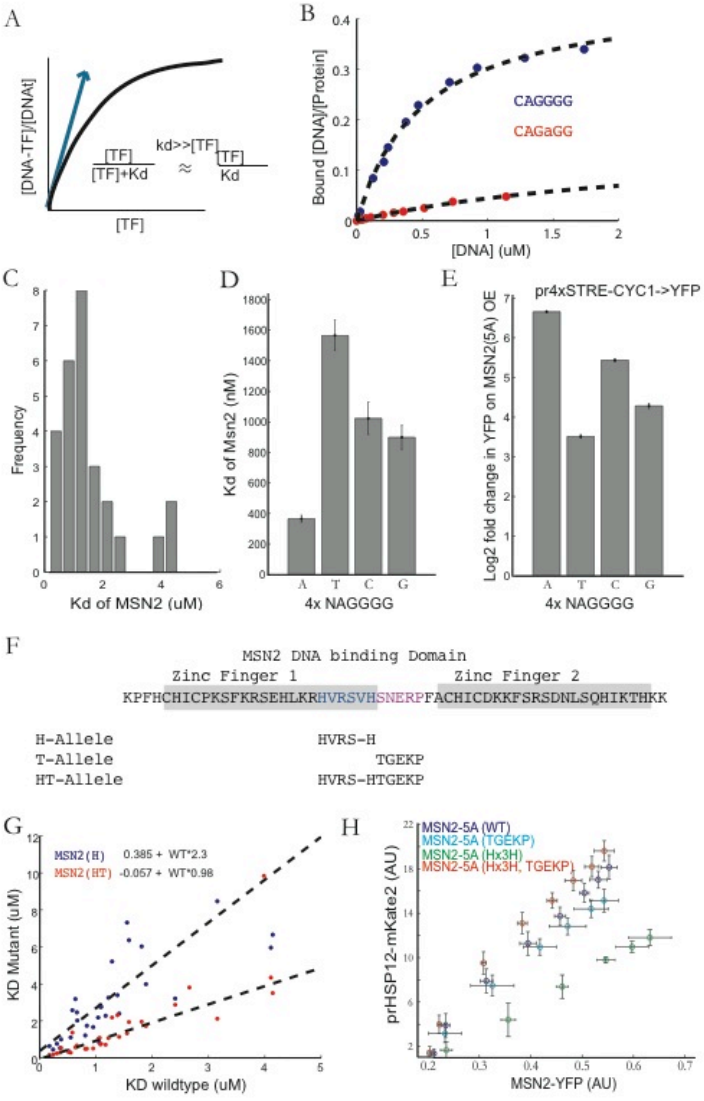

A
[DNA-TF]/[DNAt]
[TF]
[TF]/[TF]+Kd
kd>>[TF]
≈
[TF]/Kd
B
Bound [DNA]/[Protein]
[DNA] (uM)
CAGGGG
CAGaGG
C
Frequency
Kd of MSN2 (uM)
D
Kd of Msn2 (nM)
4x NAGGGG
E
pr4xSTRE-CYC1->YFP
Log2 fold change in YFP on MSN2(5A) OE
4x NAGGGG
F
MSN2 DNA binding Domain
Zinc Finger 1
Zinc Finger 2
KPFHCHICPKSFKRSEHLKRHVRSVHSNERPFACHICDKKFSRSDNLSQHIKTHKK
H-Allele HVRS-H
T-Allele TGEKP
HT-Allele HVRS-HTGEKP
G
MSN2(H) 0.385 + WT*2.3
MSN2(HT) -0.057 + WT*0.98
KD Mutant (uM)
KD wildtype (uM)
H
MSN2-5A (WT)
MSN2-5A (TGEKP)
MSN2-5A (Hx3H)
MSN2-5A (Hx3H, TGEKP)
prHSP12-mKate2 (AU)
MSN2-YFP (AU)

Figure 3

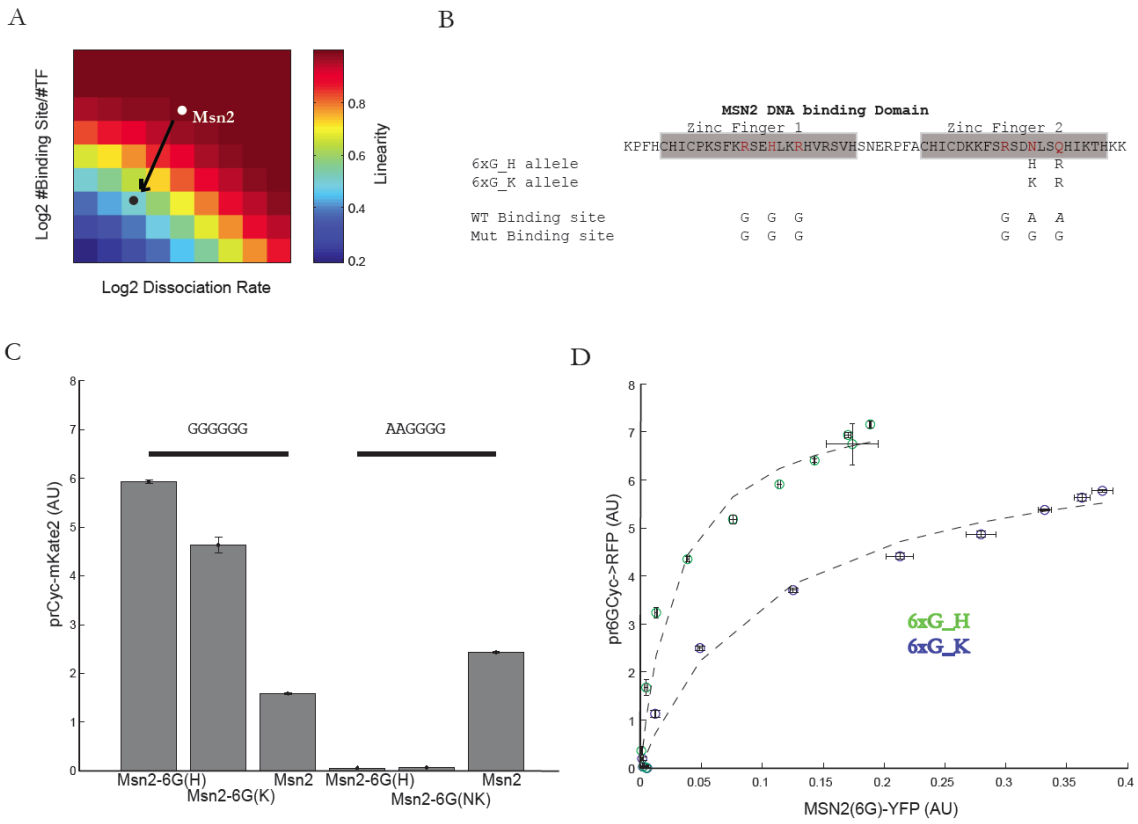


Figure 4

Supplemental Information

Supplemental Figures

S1: Multiple approaches to modulating Msn2 activity result in qualitatively similar results with minor quantitative difference, data from these assays can be predicted with reasonable accuracy by a simple model of Msn2 activity based on binding site number of each promoter. (a-d) Expression of YFP driven by 90 promoters enriched for stress response was measured in exponential growth and in response to overexpression of Msn2(5A), Msn4(4A), Ras2(S24N),

Ras2(G22A), and PDE2. Log2 fold change in expression over exponentially growing cells for each perturbation is plotted. (e) Expression of YFP from each promoter in a WT or *msn2/4Δ* strain with overexperssion of Ras2(S24N). The size of each circle represents the number of Msn2 binding sites in that promoter.

S2: Sequential removal of repressive PKA phosphorylation sites results is graded increase in Msn2 nuclear localization and transcriptional activity but does not change the essentially linear character of the transcriptional regulation. (a) We constructed Msn2 alleles with serial removal of each of the five PKA phosphorylation sites resulting in six alleles (WT, 1-5A). Each of these alleles tagged with YFP was expressed in a strain also expressing a TPS1 promoter driving the expression of the RFP mKate2, the degree of nuclear localization was measured by microscopy as well as the expression of RFP. Quantification of the nuclear localization and abundance of each Msn2 allele, increased nuclear localization results in lower protein likely due to degradation. Errorbars are standard deviation across single cells (N=20-50) (b) Experimental measurement of the expression of the Msn2 reporter prTPS1-mKate2 in response to variable expression of each allele of Msn2 obtained using the titratable estradiol system.

S3: To explore the role of flanking nucleotides in vivo we scanned the genome for all potential Msn2 binding sequences ('NAGGGG') and compared the frequency of nucleotides in the 1st position for promoters with varying numbers of binding sites. Promoters with many Msn2 sites tend to be strongly regulated and among these promoters we see a mild increase in the number of strong 'AAGGGG' sites and a decline in the number of weak 'TAGGGG' sites.

Errorbars are standard error computed from bootstrapping of the genomic dataset. (b) kD of wild type and mutant alleles of Msn2 from MITOMI 2.0 measurements for five DNA oligos lacking any consensus Msn2 binding sites show no correlation. (c) Kd measurements of 34 oligos containing at least one 'AGGGG' sequence for the three Msn2 alleles, and for five negative controls oligos lacking such as sequence. (d) The location of binding sites measured by MITOMI 2.0 are indicated on each promoter by red dashes where the height shows the Kd.

Supplemental Experimental Procedures

Construction of Promoter Library

One kilobase upstream of each gene of interest was PCRed from genomic DNA with primers containing pspOMI (or Not1) and Xho1 (or Sal1) sites, digested, and cloned into a TRP1 marked single integration vector directly upstream of a yeast optimized Venus YFP. Clones were sequence verified and integrated into yeast by digestion with Nae1, transformation using a standard PEG/LioAc protocol and selection on SD-TRP plates.

Deletion of STRE elements was accomplished by quickchange mutagenesis. The binding sites AGGGG were mutated to AGaGG, a mutation that appeared to eliminate activity *in vivo* and we observed to dramatically reduce the affinity for MSN2 *in vitro*.

Insertion of polyT(12x) elements was accomplished by quickchange mutagenesis, insertions sites were chosen based on location in the promoter

(250-400bp upstream of the ATG) and availability of suitable regions for design of efficient quickchange primers (16GC basepairs within a 40bp region). Primer design was accomplished by custom matlab scripts complemented by manual editing.

Construction of synthetic promoters

We constructed a vector with the prCYC1(1-243) sequence cloned directly upstream of Venus in a TRP1 marked vector. Additionally, to reduce background expression we included the prCYC1(1000-684) sequence. In between these sequences (which lack UASs) we inserted our STRE sequences or STRE like sequences.

4xSTRE: GGGCCCCTNCATTACCCCTNCTTTACCCCTNCAAACCCCTN

Where the N nucleotide was varied

3x6C: GGGCCCATTTACCCCCCATTTACCCCCCATTTACCCCCCA

These constructs were then integrated into yeast and assayed as above.

Construction of over expression constructs

Genes of interest were PCRed from yeast genomic DNA and cloned downstream of a prGAL1 promoter in a TRP1 marked single integration plasmid. Site directed mutagenesis was preformed as necessary to construct the constitutively active alleles of MSN2 (S288A, S582A, S620A, S625A, S633A), Msn4(S263A, S316A, S531/2A, S558A), or dominant negative alleles of RAS2 (S24N or G22A) using a standard QuickChange protocol and the pfuTurbo enzyme mix (stratagene). Constructs were sequence verified, cut with Pme1 and integrated into yeast with selection on SD-TRP plates.

Measurement of MSN2 abundance

To convert arbitrary intensity units to absolute values we expressed constructed strains where we tagged one of NUF2, ASC1, and SPC42 with YFP. These proteins localize to the kinetochore and spindle pole body structures and have a well-described abundance in these structures and have been used previously as molecular standards. To measure these abundances we imaged cells expressing these tagged proteins and defined a linear function relating their expression to the intensity measured in the microscope [R^2>0.95]. We then applied this function to measurements of Msn2-YFP to convert measured intensity to molecular numbers.

Zinc Finger Engineering H/T/HT alleles

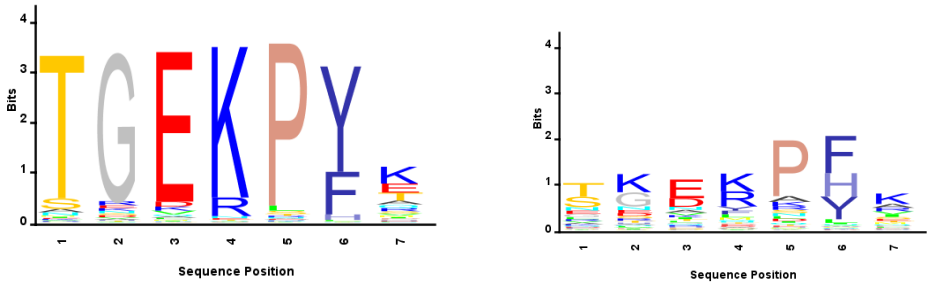
Our goal was to construct MSN2 alleles with altered affinity but identical binding preferences. We began by comparing the sequence of the MSN2 DNA binding domain to that of related cerevisiae 2xC2H2 zinc finger proteins MSN4, COM2, GIS1, RPH1, MIG1, MIG2, ADR1ADR1

```
MSN2 KPFHCHICPKSFKRSEHLKRHVRSVHSNERPFACHI--CDKKFSRSDNLSQHIKT-HKKHGDI
MSN4 KPFKCKDCEKAFRRSEHLKRHIRSVHSTERPFACMF--CEKKFSRSDNLSQHLKT-HKKHGDF
COM2 KQFACDFCDRRFKRQEHLKRHVRSLHMGEKPFDCHI--CGKKFSRSDNLNQHIKT-HTDGEVL
GIS1 KVYVCQECRRQFSSGHHLTRHKKSVHSGEKPHSCPK--CGKKFKRRDHVLQHLNK--KIPC
RPH1 KIYICKECQRKFSSGHHLTRHKKSVHSGEKPHSCPK--CGKRFKRRDHVLQHLNK--KIPC
MIG1 RPYVCPICSRAFHRLEHQTRHIRT-HTGEKPHACEFPGCGKRFSRSDELTRHTRI-HTNPAPKGKRGRKKKSKS
MIG2 RPFRCETCARGFHRLEHKKRHMRT-HTGEKPHHCAFPGCGKGFSRSDELKRHLRT-HTSTSQRRTK-KPKMR
ADR1 RSFVCEVCTRAFARQEHLKRHYRS-HTNEKPYPCGL--CNRCFTRRDLLIRHAQKIHSGNLGETI
     : : *  * : *    *  ** :: *  *:*  *    * :** * * :  *    *
```

The critical C2H2 residues which coordinate the zinc ion are marked in red, the linker between the two fingers in purple. The linker region is of particular note as the consensus sequence (TGEKP) is highly conserved across zinc fingers and

mutations to this sequence are strongly associated with reduced affinity. It is thought this linker caps the end of the recognition helix and positions the adjacent finger for effective binding to DNA. However Msn2 also has one additional feature which is the unusual spacing between the two histidines in the first finger (highlighted in green it makes up four residues, not the more common three we observed in Mig1 or Adr1 for example). To see if the increased spacing of histidines could explain the non-cannonical spacing we querried a protein structure database HMMER (http://hmmer.janelia.org/) and using the MSN2 DBD(residues 648-705) as a querry downloaded all matching 2xC2H2 proteins and compared the linker structure for fingers with a HX_3H (101864) and those with a HX_4H spacing (4626) by constructing consensus seqlogs (with matlab scripts).

HX_3H **HX_4H**



We observe substantially less conservation of the core linker sequence in the HX_4H fingers, perhaps due to the disrupted helix which has been observed in the NMR structure of the second finger of ADR1, a HX_4H finger [].

Drawing on previous work that has explored the function of the linker residues which suggests that perturbations should not affect DNA binding specificity, but only affinity [], we constructed three MSN2 mutants with altered linker arrangements, one construct (H allele) which converted the HX_4H spacing to a more conventional HX_3H spacing by removing the final valine (V669), the second construct (T allele) converted the linker sequence to the consensus (S671T, N672G,R674K), and finally an allele that combined these two mutations (HT). Based on previous data we suspected that the H and T alleles would by themselves reduce affinity, but in combination move the protein towards the consensus and increase the affinity without altering the nucleotide preference of Msn2.

Zinc Finger Engineering 6G alleles

To alter the Msn2 binding preference towards 'GGGGGG' we examined the second finger (as the first already recognizes GGG). The recognition helix in this finger is the somewhat unusual RSDNLSQ, the residues highlighted in red are largely responsible for DNA recognition in a canonical zinc finger. A scan of the literature suggests that this helix should recognize AAG (with the Q and N both recognizing A and the R, G), consistent with its in vivo binding properties. Previously a RSDHLTR helix has been shown to recognize GGG, as has a RSDKLTR with less affinity []. We therefore constructed a Q693R, N690H/K MSN2 which did indeed have strong affinity for a 6xG site.

Model of MSN2 promoter regulation (figure 2)

Consider a model of MSN2 gene regulation whereby MSN2 is produced in the cytoplasm (at rate *alpha*), and degraded (at rate *γ1*) and translocated into the nucleus (at rate *kin*). Nuclear MSN2 then is degraded (at the higher rate *γ2*) and can re-enter the cytoplasm (at rate *kout*). RFP is produced at a rate proportional to the amount of MSN2n and degraded (*γ3)*.

$$dMSN2c/dt = alpha + MSN2n * kout - MSN2c * kin - MSN2c * \gamma_1$$
$$dMSN2n.dt = MSN2c * kin - MSN2n * kout - MSN2n * \gamma_2$$
$$dRFP/dt = MSN2n * k - RFP * \gamma_3$$

Solving this system of equations for at steady state (dx/dt=0) we get the following for RFP

$$RFP = \frac{k}{\gamma_3}\alpha\frac{Kin}{(kout+\gamma_2)(kin+\gamma_1)} * \frac{1}{1-\frac{kin*kout}{(kout+\gamma_2)(kin+\gamma_1)}}$$

Showing a linear relationship between the production rate of MSN2 and RFP concentration with a slope governed by the relative rates of import/export and degradation.

For the simulations shown in figure 2 we use the following parameters

K=1;

Kin=0.1-1

Kout=0.4

α=0.1-1

$\gamma_1 = \gamma_3 = 0.0077$

$\gamma_2 = 5*0.0077$

Models which lead to linear association of transcription factors with promoters (Figures 3/4)

Where O is the occupancy of a given site/promoter, X is the concentration of the transcription factor, and k is the k_d of that transcription factor for the promoter.

$$O(x) = \frac{X}{X + k}$$

$$X << k$$

$$O(x) = \frac{X}{k}$$

$$O(x) \propto X$$

$$O(x) \propto \frac{1}{k}$$

Alternatively we can write a second model to explore the effect of the concentration of transcription factor *X* and its binding sites *Y* on the formation of the *XY* complex.

We can begin by writing a simple system of differential equations to describe the binding and unbinding of X from Y.

$$\frac{dx}{dt} = -\alpha * x * y + \gamma * xy$$

$$\frac{dxy}{dt} = \alpha * x * y - \gamma * xy$$

$$\frac{dy}{dt} = -\alpha * x * y + \gamma * xy$$

Solving this system for steady states (d/dt=0), adding the assumption that the [Ytotal]>>[Xtotal] we can simplify the final expression

$$xy = \frac{\alpha}{\gamma} * x * y$$

$$xy = \frac{\alpha}{\gamma} * (x_t - xy) * (y_t - xy)$$

$$y_t >> xy$$

$$xy = \frac{\alpha}{\gamma} * (x_t - xy) * y_t$$

$$xy = x_t * \frac{\frac{\alpha}{\gamma} y_t}{\frac{\alpha}{\gamma} y_t + 1} = x_t * \frac{y_t}{y_t + \frac{\gamma}{\alpha}}$$

To extend this analysis to a system where two species (transcription factor and DNA) may have arbitrary ratios and affinities and where the bound form results in production of a third species (mRNA or Protein) we adopted the same model as above.

$$\frac{dx}{dt} = -\alpha * x * y + \gamma * xy$$

$$\frac{dxy}{dt} = \alpha * x * y - \gamma * xy$$

$$\frac{dy}{dt} = -\alpha * x * y + \gamma * xy$$

For our simulations we set α=10^3, γ=10^2 and γ_2=10^0. X(t=0) was set at 1, and Y(T=0) was varied over 9 log2s centered on 0.1, the off rate (γ) was similarly varied over 9log2s centered on 10^2.

Linearity between input and output at each point was computed by the R^2 coefficient between the concentration of X_{total} and XY.

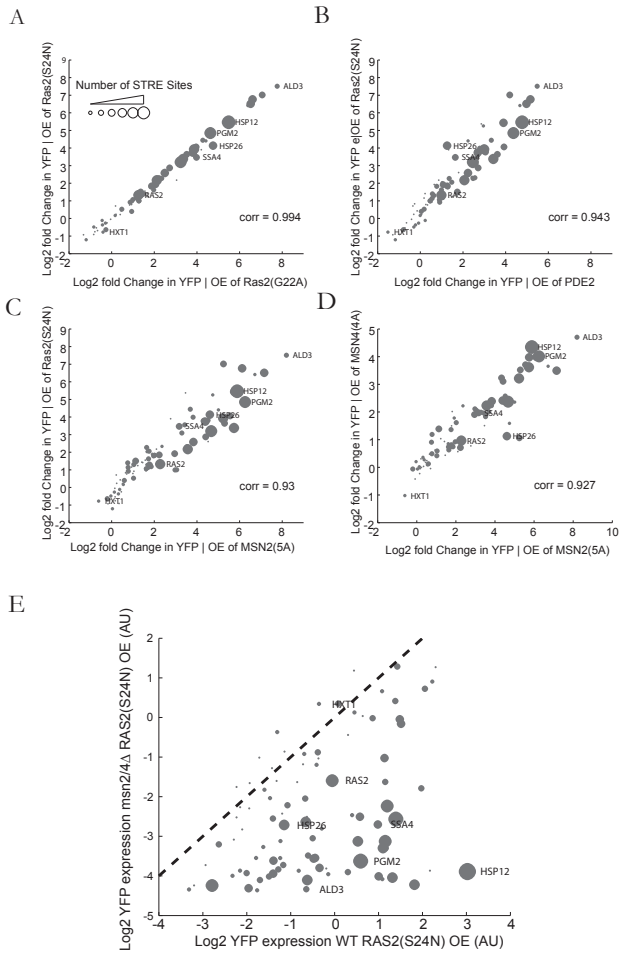

A
B
C
D
E
Number of STRE Sites
ALD3
HSP12
PGM2
HSP26
SSA4
RAS2
HXT1
corr = 0.994
corr = 0.943
corr = 0.93
corr = 0.927
Log2 fold Change in YFP | OE of Ras2(S24N)
Log2 fold Change in YFP | OE of Ras2(G22A)
Log2 fold Change in YFP | OE of PDE2
Log2 fold Change in YFP | OE of MSN2(5A)
Log2 fold Change in YFP | OE of MSN4(4A)
Log2 YFP expression msn2/4Δ RAS2(S24N) OE (AU)
Log2 YFP expression WT RAS2(S24N) OE (AU)

Figure S1

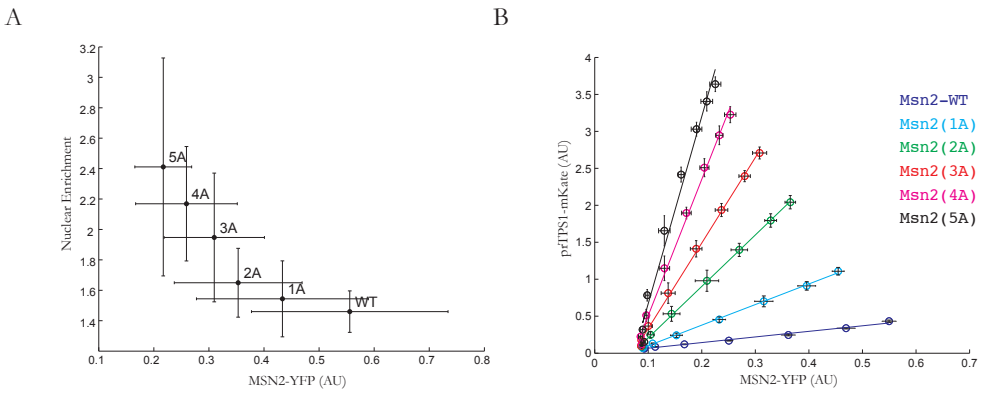

A
B
Nuclear Enrichment
MSN2-YFP (AU)
5A
4A
3A
2A
1A
WT
prTPS1-mKate (AU)
MSN2-YFP (AU)
Msn2-WT
Msn2(1A)
Msn2(2A)
Msn2(3A)
Msn2(4A)
Msn2(5A)

Figure S2

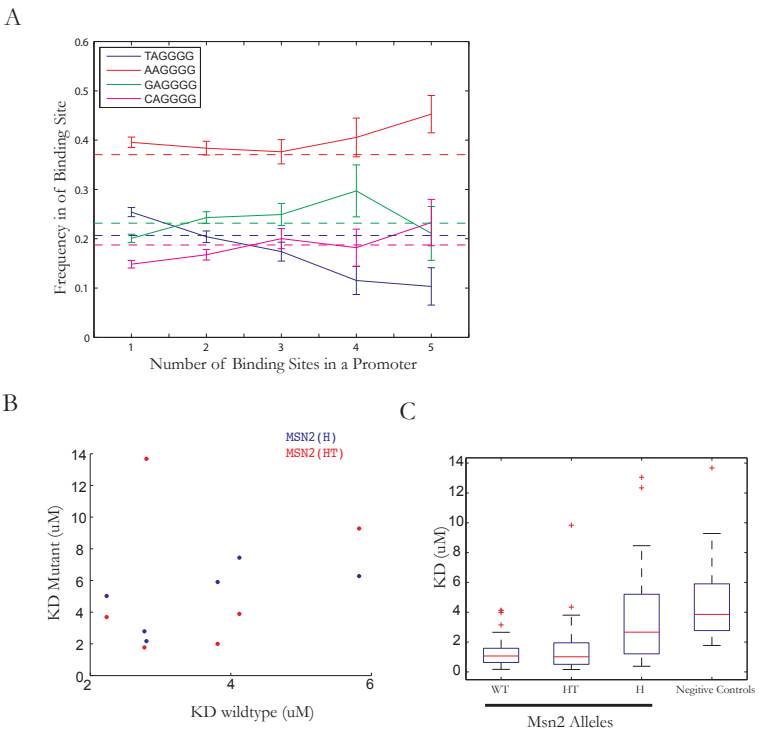

A
TAGGGG
AAGGGG
GAGGGG
CAGGGG
Frequency in of Binding Site
Number of Binding Sites in a Promoter
B
MSN2(H)
MSN2(HT)
KD Mutant (uM)
KD wildtype (uM)
C
KD (uM)
WT
HT
H
Negitive Controls
Msn2 Alleles

Figure S3

Chapter 9……...
MITOMI analysis of the tuberculosis virulence regulator EspR as a monomer and dimer

Chris Nelson

ABSTRACT

Mycobacterium tuberculosis causes millions of chronic infections and deaths every year. It invades and replicates within lung macrophages. This unique niche requires remodeling by excreted effectors. One such system is controlled by EspR, a transcriptional regulator, in order to shield the effector proteins from antigenic surveillance. It has recently been recognized that EspR adopts a unique DNA binding mode among helix-turn helix (HTH) DNA binding proteins. Furthermore, the sequence-specificty of EspR is somewhat in question, as ChIP-seq experiments suggest that it can bind wide areas of the genome in clusters. Here I present MITOMI results suggesting that EspR binding *in vitro* is indeed sequence-specific, and that similar half-site specificities can be derived from both monomeric and dimeric forms of the protein.

INTRODUCTION

13.7 million people have chronic tuberculosis, (WHO 2009). At the heart of *Mycobacterium tuberculosis*' success in prolonged infections is its ability the evade immunity and remodel it's niche inside alveolar macrophages. This ability is conferred in part by the ESX-1 secretion system that is induced upon invasion

(Stanley 2003). In fact, the benign BCG strain used to vaccinate children is traces its avirulence to a genomic region encoding most of the ESX-1 system (Mahairas 1996). The secretion system must be tightly regulated so that the secreted effectors are not exposed to immune neutralization over the decades long course of infection. The bacteria will only build its full secretion apparatus when necessary. Raghavan et al 2010 found that the system is regulated by a DNA binding factor EspR, that is itself secreted to turn off the circuit.

Two recent crystal structures have shown that EspR dimers adopt and atypical DNA-recognition conformation (Rosenberg 2011, Blasco 2012). The two halves of the dimer are splayed out in crystal structures relative to other HTH proteins, suggesting that instead of two nearby half-sites, the espR homodimer binds two distant sites, perhaps connected by a long loop. Subsequent work by Blasco et al found that EspR might act as a nucleoid-associated protein. These proteins are somewhat similar to eukaryotic histones in that they bind and organize the genome. These factors exhibit a wide range of binding site specificities. EspR is thought to bind a consensus AGCAAA, based on a few candidate ChIP-Chip sites (Rosenberg et al. 2011).

In the work described here we set out to analyze EspR's preferred binding motif and specificity, and determine how dimerization might effect binding preference.

MATERIALS AND METHODS

MITOMI 1.0 and 2.0 analyses were carried out as described earlier in this thesis. Full-length and delta107 EspR were cloned into linear translation constructs with C-terminal 6xHis tags for recruitment on the MITOMI chips. Three library replicate binding experiments were carried out for the full-length construct, and four replicate measurements were taken from the truncated (monomeric) protein. Motif finding was carried out as described previously in the chapters on FOXP2 and Hac1.

For systematic mutation of the binding site we selected oligonucleotide 332, which gave consistent signal in the random library experiments. The mutated oligonucleotides were spaced out, skipping unit cells in the MITOMI device to avoid oligonucleotide-oligonucleotide cross talk, which can greatly reduce the specificity of the measurements. At least 3 measurements of each oligonucleotide was taken and the data was fit in graphpad 4.0 to derive relative affinity constants for each mutation. These measurements were then consolidated into a position specific affinity matrix for display and future use in predicting *in vivo* sites of regulation.

RESULTS

MITOMI2.0 experiments consistently produced motifs with a core AGCAAA sequence (figure 1). This was true across the full length and monomeric forms of the protein. AGCAAA was typically the top scoring 6mer in fREDUCE analysis. For the full length protein the top cumulative matrixREDUCE 6mer motif gave a

Pearson correlation of 0.48, with a p-value of 4.7E-100. The top 8mer motif gave a Pearson correlation of 0.53, with a p-value of 3.6E-129. The logos associated with these top-scoring PSAM binding models are shown in figure 1A and 1B. Since we are interested in how the abnormal dimeric structure affects DNA binding we employed a truncation mutant, that ablated the domain responsible for dimerization (Rosenberg et al. 2011). The 6mer motif derived from the truncated protein gives a 0.39 Pearson correlation, with a p-value of 2.3E-105. The top 8mer motif gave a 0.48 Pearson correlation, with a p-value of 1.6E-168. These values, while still significant are much lower than those derived for Hac1 and FOXP2 described in the preceding chapters, suggesting that there is more of the binding activity that is not fully accounted for by a single simple binding model.

This led us to look at the distribution of oligonucleotides with and without a perfect match to the consensus sequence AGCAAA. Figure 2 shows that the AGCAAA-containing oligos tend to bind well above background, while only a few oligonucleotides without a perfect match to AGCAAA score within the top oligonucleotides ranked by normalized bound DNA signal (RNN). Some of these binders could be explained by point variants of the consensus AGCAAA, such as GGCAAA.

To confirm further investigate permissive motif variants and explore the degree of sequence-specificity of EspR we performed binding curve measurements for all possible mutations of an 8mer binding site (Figure3). These experimental

measurements are broadly in line with the predictions of our random library analysis. The largest positional difference between the two predictions is that AGCATA appears to be a permissive variant when measured in a targeted binding curve fashion, as opposed to in the context of a random library at single concentrations. In the core 6mer, all of the A positions can be substituted without major penalties. In addition we seem to have recapitulated the predicted flanking base predictions, extending the motif to a degenerate 8mer. Overall the positional effects tend to be below 2 on the deltadeltaG/RT scale, suggesting that the specificity of EspR is somewhat moderate. For comparison see the FOXP2 chapter, Figure 3B and C.

Discussion

On the question of the sequence specificity of EspR we can say that it does exhibit clear >100 fold preferences between different oligonucleotides and sub motifs. The *in vivo* clustering effect described by Blasco et al. 2012 may be due to a binding activity other than that, and may require genomic context. In any case we do not see a great deal of nonspecific binding, and EspR has a clear preference for a subset of our library. The binding site model we have described explains the majority of the observed binding.

On the other hand in terms of site availability and information content a degenerate 6mer like EspR might be expected to be more promiscuous than factors like Hac1 or FOXP2 with longer binding sites and stiffer energetic

penalties for variants. Perhaps the consistent flanking base preferences that we found lend a bit more specificity, or added occupancy probability for sites requiring more intense EspR regulation. In conclusion it seems best to consider EspR as a factor that may function with moderate specificity as either a monomer or a dimer.

The differences that we observed between motifs derived from the monomeric and potentially dimeric forms of the protein are relatively minor, with the same top scoring AGCAAA sequence. That said it is possible that the minor variants of the motif are bound differently by different monomeric or dimeric species.

At the time of submission this complementary studies were being carried out in the Cox Laboratory at UCSF. The active question is of how EspR might loop and organize genomic DNA in the context of these binding sites. The ESX-1 locus EspA contains at least three instances of the consensus motif discussed here, and their relative impacts on reporter activity, and requirements of spacing between the half-sites is ongoing.

REFERENCES

Blasco B, Chen JM, Hartkoorn R, Sala C, Uplekar S, Rougemont J, Pojer F, Cole ST. Virulence regulator EspR of Mycobacterium tuberculosis is a nucleoid-associated protein. PLoS Pathog. 2012;8(3):e1002621

Blasco B, Stenta M, Alonso-Sarduy L, Dietler G, Peraro MD, Cole ST, Pojer F. Atypical DNA recognition mechanism used by the EspR virulence regulator of Mycobacterium tuberculosis. Mol Microbiol. 2011 Oct;82(1):251-64.

Mahairas GG, Sabo PJ, Hickey MJ, Singh DC, Stover CK Molecular analysis of genetic differences between Mycobacterium bovis BCG and virulent M. bovis. J Bacteriol. 1996 Mar; 178(5):1274-82.

Raghavan S, Manzanillo P, Chan K, Dovey C, Cox JS. Secreted transcription factor controls Mycobacterium tuberculosis virulence. Nature. 2008 Aug 7;454(7205):717-21.

Rosenberg OS, Dovey C, Tempesta M, Robbins RA, Finer-Moore JS, Stroud RM, Cox JS. EspR, a key regulator of Mycobacterium tuberculosis virulence, adopts a unique dimeric structure among helix-turn-helix proteins. Proc Natl Acad Sci U S A. 2011 Aug 16;108(33):13450-5.

Stanley SA, Raghavan S, Hwang WW, Cox JS Acute infection and macrophage subversion by Mycobacterium tuberculosis require a specialized secretion system. Proc Natl Acad Sci U S A. 2003 Oct 28; 100(22):13001-6.

World Health Organization (2009). "Epidemiology". Global tuberculosis control: epidemiology, strategy, financing. pp. 6–33. ISBN 978-92-4-156380-2.

FIGURE LEGENDS

Figure 1. MITOMI 2.0 investigations of the EspR binding site. Full length and truncated EspR was incubated with a random library of DNA in a microfluidic device and the amount of DNA bound for each DNA sequence in the library was measured. The top matrixREDUCE motifs are shown for A) truncated espR queried for a 6mer motif, B) truncated espR queried for an 8mer motif, C) full-

length espR queried for an 6mer motif, and D) full-length espR queried for an 6mer motif.

Figure 2. Distribution of RNN binding signal for Oligonucleotides with and without AGCAAA perfect matches. The normalized DNA/protein signal bound at the MITOMI button valves is plotted along the vertical axis, and the percent rank is plotted along the horizontal axis(tighter binders are to the left, weaker binders to the right).

Figure 3. Systematic position specific affinity measurements confirm the optimal binding motif. To confirm the motifs derived by matrix reduce calculations we measured relative affinities of EspR for all variants of an 8bp motif candidate. The resulting matrix of experimentally measured affinity changes is displayed here as a motif logo.

FIGURES

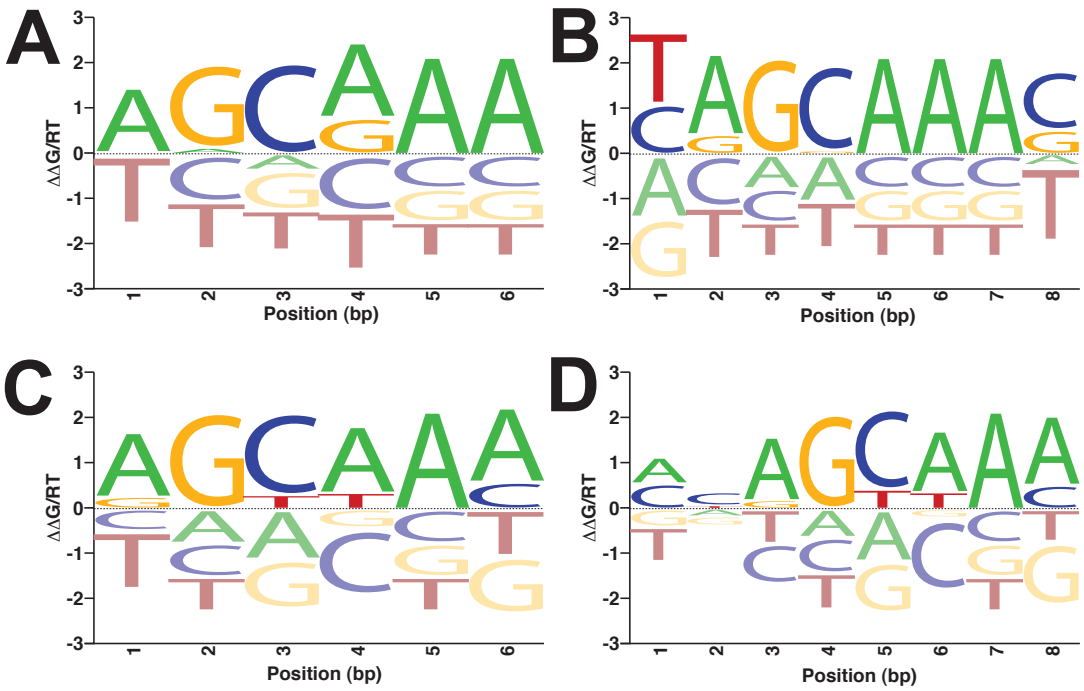


Figure 1.

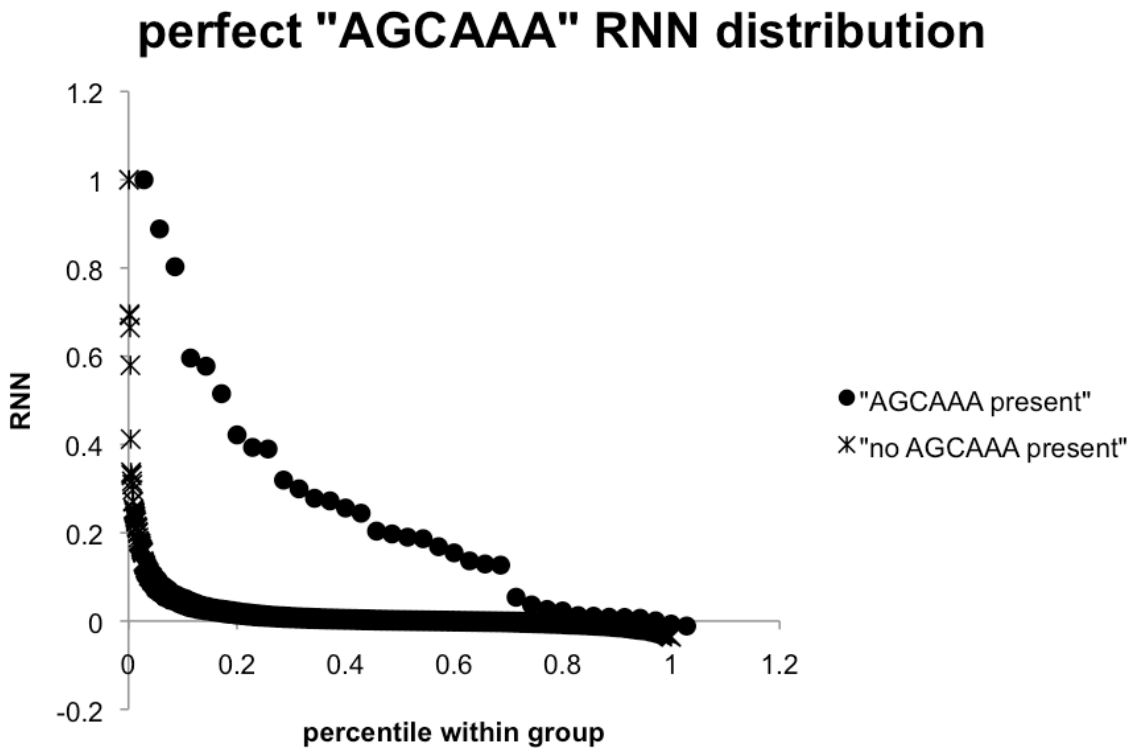


Figure 2.

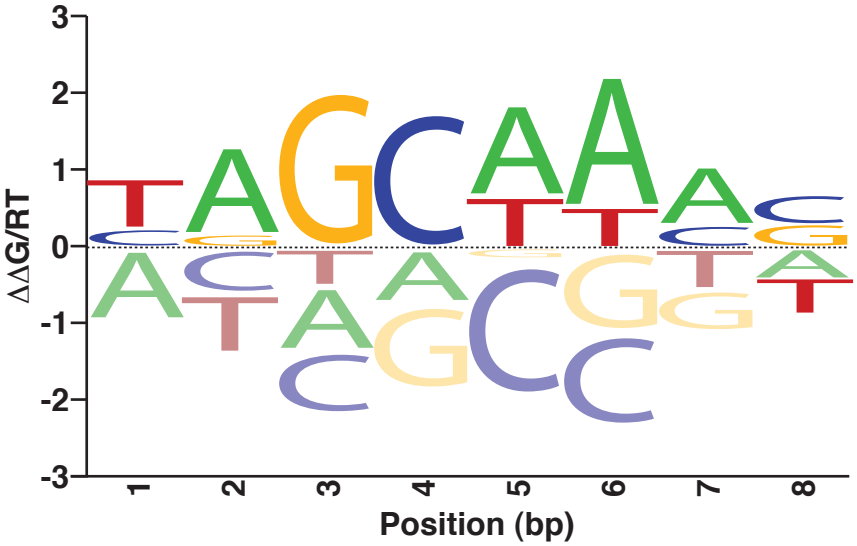


Figure 3.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San

Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

January 9th 2011

Author Signature Date