Summer 9-1-2014

Genetic and Epigenetic Determinants of Transcription in the Divergent Eukaryote Leishmania major

Britta Anderson
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology & Biomedical Sciences

Molecular Cell Biology

Dissertation Examination Committee:
Stephen M. Beverley, Chairperson
Douglas Chalker
Tamara Doering
Sarah Elgin
Daniel Goldberg
Christina Stallings
Ting Wang

Genetic and Epigenetic Determinants of Transcription in the Divergent Eukaryote *Leishmania major*

by

Britta Amelia Anderson

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2014

Saint Louis, Missouri

Table of Contents

List of Figures……………………………………………………………………………….iv

List of Tables………………………………………………………………………………..ix

Acknowledgments…………………………………………………………………………..xi

Abstract…………………………………………………………………………………….xii

Chapter One: Introduction……………………………………………………………...1

Chapter One: References………………………………………………………….....…36

Chapter Two: Kinetoplastid-specific histone variant functions are conserved in *Leishmania major*………………………………………………………………………………..56

Chapter Two: References……………………………………………………………...81

Chapter Three: The chromatin landscape of the early-diverging eukaryote *Leishmania major*...105

Chapter Three: References…………………………………………………………….146

Chapter Four: Identification of *cis*-regulatory elements associated with transcription of protein-coding genes in *Leishmania major* using an integrated bidirectional reporter…………………169

Chapter Four: References………………………………………………………………195

Chapter Five: Mutation of poly(dG:dC) tracts in the dSSR core suggests that transcription directionality can be biased in an irregular manner suggestive of epigenetic control………….215

Chapter Five: References……………………………………………………………….234

Chapter Six: Elements, strategies, and tactics oriented towards the development of a system for inducible transcription in *Leishmania*…………………………………………………………..245

Chapter Six: References………………………………………………………………264

Chapter Seven: Concluding remarks and future directions…………………………………….272

Chapter Seven: References……………………………………………………………284

List of Figures

Chapter 1: Introduction

Figure 1-1. Models of active promoter elements in eukaryotes……………………………...49

Figure 1-2. Depiction of polycistronic transcription and *trans*-splicing of protein-coding genes and the spliced leader (SL) RNA…………………………………………………………...50

Figure 1-3. Description of genetic and epigenetic elements associated with divergent and convergent strand switch regions…………………………………………………………...51

Figure 1-4. Proposed model for the nature and function of *cis*-acting elements in *Leishmania*……………………………………………………………………………….52

Chapter 2: Kinetoplastid-specific histone variant functions are conserved in *Leishmania major*

Figure 2-1. *H2AZ* is essential in *L. major*……………………………………………………84

Figure 2-2. *H2BV* is essential in *L. major*……………………………………………………85

Figure 2-3. Deletion of *H3V* in *L. major* does not alter growth or metacyclogenesis…………86

Figure 2-4. Deletion of *H3V* does not increase read-through transcription as observed by SL-RNA-seq……………………………………………………………………………………87

Supplemental Figure S2-1. Demonstration of anti-H3 and anti-H3.V specificity by western blotting…………………………………………………………………………………...90

Supplemental Figure S2-2. Transcription termination is unaltered in *Δh3v* parasites………...91

Supplemental Figure S2-3. Transcription termination is unaltered in *Δh3v* parasites, regardless of whether tRNAs are present in the cSSR…………………………………………………103

Chapter 3: The chromatin landscape of the early-diverging eukaryote *Leishmania major*

Figure 3-1. Outline of PAVED bioinformatics pipeline…………………………………...149

Figure 3-2. Nuclease hypersensitive (NH) sites can be readily identified in tRNA and rRNA genes, but not in divergent SSRs……………………………………………………....150

Figure 3-3. Nucleosome positioning analysis reveals well-positioned nucleosomes upstream of tRNA genes………………………………………………………………………….151

Figure 3-4. FAIRE-seq enriches dSSR-proximal loci and NH sites identified by MNAse-seq……………………………………………………………………………………...152

Figure 3-5. MNAse overdigestion reveals qualitative differences in dSSR-proximal nucleosomes…………………………………………………………………………153

Supplemental Figure S3-1. Preparation and characterization of MNAse-treated and sheared DNA samples……………………………………………………………………..154

Supplemental Figure S3-2. Sub-mononucleosome-sized DNA fragments show similar size distributions…………………………………………………………………………155

Supplemental Figure S3-3. Variations in somy and copy number errors in the reference genome are corrected by normalization to a control dataset…………………………...156

Supplemental Figure S3-4. SL-derived sequences are absent from raw data for MNAse-treated DNA and chromatin datasets……………………………………………………..157

Supplemental Figure S3-5. Normalized fragment depth histograms for MNAse-seq datasets……………………………………………………………………………...158

Supplemental Figure S3-6. Quantitative PCR validation of standard MNAse-seq data…....159

Supplemental Figure S3-7. Generation and quantitation of randomly distributed genomic intervals……………………………………………………………………………160

Supplemental Figure S3-8. Genome-wide nucleosome positioning and spacing……………161

Supplemental Figure S3-9. Nucleosome distance analysis for regions of interest…………..162

Supplemental Figure S3-10. Restriction endonuclease sensitivity assays demonstrate similar nuclease sensitivities in PGC-internal loci and dSSRs………………………………………163

Chapter 4: Identification of *cis*-regulatory elements associated with transcription of protein-coding genes in *Leishmania major* using an integrated bidirectional reporter

Figure 4-1. Description of the dSSR and planned integration of the bidirectional dual-luciferase reporter into chromosome 1…………………………………………………………….203

Figure 4-2. Mapping of the minimal functional splice acceptors in the dSSR of chromosome 1………………………………………………………………………………………………204

Figure 4-3. Confirmation of planned integration of reporter constructs into the chromosome 1 dSSR………………………………………………………………………………………206

Figure 4-4. Reporter gene data for integrated bidirectional reporter clones containing the WT dSSR………………………………………………………………………………………208

Figure 4-5. FAIRE-qPCR indicates maintenance of the epigenetic state of the dSSR……….209

Figure 4-6. dSSR deletion and substitution constructs demonstrate that the core dSSR of chromosome 1 lacks *cis*-regulatory elements………………………………………………..210

Supplemental Figure S4-1. Plasmid map for pLUC v2……………………………………...211

Supplemental Figure S4-2. Correlation of reporter mRNA levels with FLuc, but not RLuc activity………………………………………………………………………………………….212

Supplemental Figure S4-3. Identification of the splice acceptor dinucleotides in studies mapping the minimal functional splice acceptor by RT-PCR………………………………213

Supplemental Figure S4-4. Spliced leader (SL)-primed RT-PCR reveals the utilization of the expected splice acceptor dinucleotides for the *FLuc* and *RLuc* mRNAs in the integrated reporter bearing the WT dSSR, but identifies a cryptic splice acceptor in some Δ489 + S lines………215

Supplemental Figure S4-5. Comparisons of the *T. brucei*-derived genic stuffer and the WT dSSR sequence between minimal functional splice acceptors……………………………...216

Chapter 5: Mutation of poly(dG:dC) tracts in the dSSR core suggests that transcription directionality can be biased in an irregular manner suggestive of epigenetic control

Figure 5-1. Landmarks of interest in the chromosome 1 dSSR and description of dSSR mutants used in this chapter…………………………………………………………………...241

Figure 5-2. Expected behavior of parasite lines containing mutations or deletions in the dSSR reporter……………………………………………………………………………….242

Figure 5-3. Reporter gene activity differs in some but not all Δ73 clones…………………….243

Figure 5-4. Reporter gene activity is also different in lines bearing poly(dG:dC) mutations...244

Figure 5-5. Substitution of poly(dA:dT) tracts does not protect against the unidirectional promoter phenotype………………………………………………………………………….245

Figure 5-6. Selective pressure can reverse the unidirectional phenotype without altering the genetic identity of the dSSR…………………………………………………………...246

Figure 5-7. Δ73 switchers have no genomic rearrangements or copy number variations affecting the dSSR reporter relative to parent lines……………………………………………………..247

Figure 5-8. FAIRE-qPCR analysis of “switchers” demonstrates differences in the epigenetic state of the reporter-associated dSSR………………………………………………………248

Chapter 6: Elements, strategies, and tactics oriented towards the development of a system for inducible transcription in *Leishmania*

Figure 6-1. Regulation of destabilization domain-gene of interest fusion proteins using the small molecule FK506……………………………………………………………………………...271

Figure 6-2. Schematic representation of constructs described in this chapter………………...272

Figure 6-3. Validation of pIR1 integration into the ribosomal SSU…………………………...273

Figure 6-4. Western blot confirmation of ddT7RNAPnls protein expression and regulation...274

Figure 6-5. Regulation of T7RNAP activity using ligand-mediated conditional protein stabilization……………………………………………………………………………………..275

Figure 6-6. Proposed models for "silent" nature of LAC-based reporter systems……………276

Chapter 7: Concluding remarks and future directions

Figure 7-1. Model for the definition of the *de facto* promoter activity of divergent SSRs in *L. major*………………………………………………………………………………………291

Figure 7-2. Model for *trans*-splicing acceptor sites as genetic *cis*-acting signals in defining dSSRs and polycistronic gene cluster (PGC)-internal transcription initiation regions………...292

List of Tables

Chapter 1: Introduction

Table 1-1. Histone variant genes in *Leishmania major*……………………………………...53

Table 1-2. Chromatin readers, writers, and erasers in trypanosomatid protozoa……………….54

Chapter 2: Kinetoplastid histone variant functions are conserved in *Leishmania major*

Supplemental Table S2-1. Primer sequences used for generation of deletion constructs for *H2AZ*, *H2BV*, and *H3V* and demonstration of *Δh3v* planned replacements……………………..88

Supplemental Table S2-2. Primer sequences used for amplification of histone variant ORFs for protein expression and episomal complementation vectors……………………………………..89

Chapter 3: The chromatin landscape of the early-diverging eukaryote *Leishmania major*

Supplemental Table S3-1. Alignment metrics for Illumina sequencing data………………...164

Supplemental Table S3-2. Normalized fragment depth metrics according to genomic context……………………………………………………………………………………..165

Supplemental Table S3-3. Comparison of relative fragment depth maximum thresholds for identification of NH sites……………………………………………………………………….166

Supplemental Table S3-4. Comparison of well-positioned nucleosomes within regions of interest……………………………………………………………………………………..167

Chapter 4: Identification of *cis*-regulatory elements associated with transcription of protein-coding genes in *Leishmania major* using an integrated bidirectional reporter

Table 4-1. Transfection efficiencies for episomal and linearized constructs used in this study……………………………………………………………………………………….217

Table 4-2. Primer sequences used for construction of pLUC plasmids……………………….218

Table 4-3. Primer sequences used for modification of pLUC plasmids or for amplification of dSSR fragments……………………………………………………………………………...219

Table 4-4. Accessory primers used for validation of proper integration, for splice acceptor dinucleotide mapping, or for quantitative PCR and reverse transcriptase PCR………………..220

Chapter 5: Chapter 5: Mutation of poly(dG:dC) tracts in the dSSR core suggests that transcription directionality can be biased in an irregular manner suggestive of epigenetic control

Table 5-1. Quantitation of episomal and integrated colonies from pLUC poly(dG:dC) tract mutants…………………………………………………………………………………….249

Table 5-2. Primer sequences used in this study……………………………………………….250

Acknowledgements

The projects contained in this thesis were greatly enhanced by the constructive discussions and valuable insight provided by my thesis committee, and I am grateful for their help in shaping this work and for their patience as I learned my way through new skill sets and foreign methods. I am especially thankful for Steve's mentorship throughout my graduate career; his insistence that experiments be done rigorously and his enthusiasm for devising scientific models and toppling them with well-designed experiment are qualities that I can only hope that I picked up from him. Steve was fully supportive when I chose to pursue a project that was very different from the others in the lab, and he always reminded me that I could do whatever I set my mind to, even when I wasn't convinced that was true. The members of the Beverley lab have become fantastic coworkers and terrific friends during my time in Steve's lab, and I have really appreciated their camaraderie, their scientific input, and their interest in celebrating worldwide sporting events with our own lab games. I am truly indebted to my family and friends who have supported me along the way. My parents have been a constant source of encouragement, and they've instilled in me the aspiration to be the best scientist, the best friend, and the best person I can be. My brother Alec has always exemplified what hard work can do for you, and he was a great motivator to finish graduate school, since I don't think I can handle his reminders about needing a real job for much longer. Lastly, I owe a giant 'thank you' to my companion, rest day coach, and sous-chef Drew. He is an infinite well of encouragement and support, and he finds a way to make the tough days easier, and the good days even better.

ABSTRACT OF THE DISSERTATION

Genetic and Epigenetic Determinants of Transcription in the Divergent

Eukaryote *Leishmania major*

by

Britta Amelia Anderson

Doctor of Philosophy in Biology and Biomedical Sciences

Molecular Cell Biology

Washington University in St. Louis, 2014

Dr. Stephen M. Beverley, Chairperson

Leishmania spp. and other trypanosomatid protozoa use a highly unusual mechanism to generate functional messenger RNAs (mRNAs) in which protein-coding genes are transcribed polycistronically. Here, transcription initiates primarily in divergent strand switch regions (dSSRs), where two polycistronic gene clusters are oriented head-to-head. These regions lack all known eukaryotic *cis*-regulatory elements, and it is not known how genetic and epigenetic factors cooperate to define dSSRs as regions of productive initiation. To quantitatively identify regulatory elements and to study the contribution of epigenetic factors to dSSR function, we combined genome-wide studies of chromatin structure with a focused interrogation of a single dSSR using a novel integrated bidirectional, dual-luciferase reporter. Chromatin-based studies demonstrated that *Leishmania* lack well-positioned nuclease-hypersensitive sites associated with promoters in other eukaryotes. Rather, nuclease-hypersensitive sites are positioned heterogeneously across broad regions associated with epigenetic marks indicative of active transcription, suggesting that transcription initiation events occur promiscuously within regions associated with a transcriptionally-permissive epigenetic state. Our studies using the bidirectional reporter validate these observations and strongly suggest that *Leishmania* do not require *cis*-regulatory elements for efficient bidirectional transcription initiating in dSSRs, as a large region of the dSSR can be replaced with unrelated sequences without altering bidirectional reporter gene expression. In addition to these genetic studies we also focused on epigenetic determinants of transcriptional activity in *Leishmania*, with respect to both transcription initiation and transcription termination. We showed that the histone variants H2A.Z and H2B.V, which are associated with transcriptionally permissive regions in *T. brucei*, are essential in *L. major*, while the transcription termination-associated histone variant H3.V is not. Interestingly, unlike *Leishmania* lacking the DNA modification base J, *H3.V*-null *L. major* shows no defects in transcription termination. Although the study of essential genes in *Leishmania* is challenging at this time, we present preliminary data describing elements of inducible gene expression systems which may improve our ability to study essential genes. Together, the data in this thesis show that transcription of protein-coding genes is primarily determined epigenetically, and suggest that chromatin-related processes may be an attractive target for therapeutic intervention.

Chapter 1

Introduction

Preface

The first draft of this chapter was written by BA, and comments from SMB were incorporated into the final draft presented here.

Leishmania: Relevance to Global Health

The disease leishmaniasis is caused by protozoan parasites of the genus *Leishmania*, which are transmitted between mammalian hosts by phlebotomine sand flies. In humans, *Leishmania* spp. cause three major forms of the disease which typically correlate with the infecting species: cutaneous leishmaniasis, which results in ulcers near the site of the infection and can lead to lifelong, disfiguring scars; mucocutaneous leishmaniasis, which results in visible destruction of the nasal and oral mucosa; and visceral leishmaniasis or kala-azar, which results in enlargement of the spleen and liver and is invariably fatal if left untreated (1). The World Health Organization (WHO) estimates that 1.3 million new cases of leishmaniasis occur each year, and 20,000-30,000 deaths are attributed to leishmaniasis annually (1). However, attempts to quantify asymptomatic infections, which have the potential to reactivate, suggests that the actual rates of infection are at least 10-fold higher worldwide (2–5), and ratios of asymptomatic to symptomatic infections were as high as 18:1 in Brazil (6) and 50:1 in Spain (7). Leishmaniasis is endemic through vast portions of the world, including throughout southeastern Asia, Africa, South and Central America, and the Mediterranean basin, leaving an estimated 310 million people at risk for infection. However, changes in temperature and rainfall spurred by climate change are predicted to have profound effects on the distribution of *Leishmania* vectors, suggesting that these numbers could increase significantly in the future (1,8–10).

The prevalence and widespread distribution of leishmaniasis and the potential for a drastic expansion in the range of sand fly vectors have led to the identification and implementation of mechanisms which may help control leishmaniasis and include efforts to target the phlebotomine sand fly and the mammalian host. Vector control programs that rely primarily on insecticide-based methods have been implemented in several countries, including the distribution of insecticide-laced bednets and promotion of indoor insecticide spraying. While these are effective when implemented properly, they require high levels of compliance, and resistance to various insecticides has been documented (1). Recently, the development of transgenic insect vectors that have the capacity to block or decrease transmission were reported for the tropical disease malaria (11). The feasibility of this concept was demonstrated in sand flies using antibodies that block *Leishmania* interactions with receptors in the sand fly midgut (12), and efforts to sequence two phlebotomine sand fly vector species will facilitate the development of methods that decrease or block parasite transmission at this stage (13). At present no vaccines exist to prevent leishmaniasis in humans, although a T-cell directed vaccine, LEISHDNAVAX, was recently reported and shows promise for human disease (14). Finally, therapies for leishmaniasis exist but are expensive, require lengthy treatments, are usually not administered orally, and often have significant side effects which limit their use (1). Resistance to therapeutics has been documented [reviewed in (15)], demonstrating the need for the development of additional therapies that are orally administered, inexpensive, and well-tolerated.

Importantly, *Leishmania* are highly diverged from both of their hosts, and methods that interfere with essential parasite processes may have the added benefit of leaving the host relatively untouched. It is clear that the study of biological processes that are necessary for the survival, proliferation, and differentiation of *Leishmania* in its mammalian and phlebotomine hosts is critical for controlling leishmaniasis. The work described in this thesis focuses on the mechanisms controlling expression of protein-coding genes in *Leishmania*, a process which is extremely different from that of both of its hosts, is highly dynamic throughout the parasite life cycle, and is an essential component of parasite viability. Our studies focus on the stages present in the phlebotomine host, but the tools and reagents developed in these studies are easily adapted for use in other species and life cycle stages. We believe that the work presented here enhances our understanding of the fundamentals of gene expression in *Leishmania*, and also presents an interesting perspective on the interactions between the genetic and epigenetic determinants of gene expression in eukaryotic systems.

The life cycle and differentiation of *Leishmania*

Leishmania parasites have a digenetic life cycle, alternating between sand flies and the mammalian host. In the sand fly, *Leishmania* reside primarily in the midgut as procyclic promastigotes, which are non-infective and replicative. These parasites undergo a density-dependent differentiation into non-replicating, infective metacyclic promastigotes, which are regurgitated into the bite on the mammalian host during the sand fly's next blood meal. Upon transmission, metacyclic promastigotes are phagocytosed by various immune cells, including neutrophils, dendritic cells, and macrophages. Inside the phagolysosome, they differentiate again into the amastigote stage where they are able to thrive and replicate. When a sand fly feeds from the infected mammalian host, these parasites are again transmitted to the sand fly where they differentiate back to procyclic promastigotes, thus completing their life cycle [reviewed in (16,17)].

This life cycle requires *Leishmania* to adapt to diverse environmental stresses and results in significant changes in cell morphology and metabolism. These include transitions from motile promastigotes to amotile amastigotes via drastic alterations in flagellar structure [reviewed in (18)]; alterations in surface glycoconjugate and phospholipid composition (19,20); changes in organellar biology and metabolism [(21,22); reviewed in (23)]; and changes in cell size and shape [reviewed in (24)]. In diverse eukaryotes, alterations in gene expression often correlate tightly with responses to environmental stresses [reviewed extensively in (25)]. Furthermore, coordinated changes in gene expression are associated with developmental cues in many systems [for an example, see (26)], including in protozoa (27). Much like in other eukaryotes, a wide variety of mechanisms have been implicated in the differentiation process in *Leishmania* and related kinetoplastids [reviewed in (28)]. Post-translational mechanisms such as differential phosphorylation have been observed throughout the differentiation process [reviewed in (29)], but mechanisms altering gene expression also play a role in the differentiation process. Interestingly, comparative analysis of gene expression between procyclic promastigotes, metacyclic promastigotes, and intracellular *L. major* amastigotes by DNA microarrays demonstrated that only ~1% of sampled genes are stage-regulated (30,31), and more recent work documenting the process of promastigote-to-amastigote differentiation in *L. donovani* showed significant fluctuations in many transcripts, but the correlation between transcript abundance and protein levels was generally very poor (32). However, coordinated regulation of transcript abundance based on location within a polycistronic gene cluster has been observed in *T. brucei* after heat shock (33), suggesting that transcriptional regulation occurs in the context of these other mechanisms.

In addition, alternative splicing between life cycle stages has been documented in both trypanosomes and in *Leishmania* and can result in the production of different proteins through the use of alternate start codons (26, Myler and Beverley, in preparation). More recently, Kolev and colleagues demonstrated that overexpression of a single RNA binding protein was sufficient to promote differentiation from noninfectious procyclic promastigote *T. brucei* to the infectious metacyclic form *in vitro* (35), and additional RNA binding proteins have been shown to play roles in other stages of differentiation in trypanosomes [reviewed in (36)]. In *L. major*, coordinated changes in chromatin compaction may be important in differentiation, as global alterations in chromatin condensation were observed between procyclic and metacyclic *L. major* promastigotes and intracellular amastigotes (Wong and Beverley, in preparation). In addition, alterations in histone modifications occur in promastigotes as they approach stationary phase (37), suggesting that epigenetic alterations may be important in chromatin condensation throughout the life cycle. Interestingly, a significant downregulation in transcriptional rates was observed in *L. major* metacyclic promastigotes compared to procyclic promastigotes (Akopyants and Beverley, in preparation), suggesting a functional link between these phenomena. These epigenetic phenomena will be discussed in greater detail later in this introduction. Together, these data demonstrate that kinetoplastid differentiation is highly complex and requires the orchestration of many diverse processes in response to environmental stimuli, including changes in gene expression. Interference with a number of these processes would likely alter parasite biology in a manner sufficient to inhibit parasite growth or transmission, and the development of methods that target broad-acting factors, such as epigenetic modifiers or master regulators of transcription, could have wide-reaching effects on such processes.

Principles of eukaryotic transcriptional regulation

Cis-*regulation and epigenetic regulation*

To understand the questions relating to transcriptional regulation and the potential for interfering with this process in *Leishmania*, I begin with a discussion of the mechanisms involved in transcriptional regulation in other eukaryotes that have been studied extensively (described in Figure 1-1). In the standard model of eukaryotic transcription of protein-coding genes [discussed at length in (38,39)], transcription initiation is regulated on a gene-by-gene basis by a combination of DNA-encoded sequences (*cis*-acting elements) and protein factors which recognize these elements (*trans*-acting factors). Transcription initiation begins with direct interactions between DNA binding proteins called transcription factors with DNA-encoded promoter sequences. Two classes of transcription factors combine to provide many layers of transcriptional regulation. The general (basal) transcription factors (TFs) bind to the core promoter motif, which can include the TATA box, Initiator (Inr) element, B-recognition element (BRE), and downstream positioning element (DPE), or some combination of these four elements. In contrast, sequence-specific transcriptional activators are not ubiquitously required for transcription and interact with discrete promoter motifs outside of the core promoter called upstream activating sequences (UAS). These UAS vary significantly among transcriptional activators, and individual genes often contain multiple UAS corresponding to different transcriptional activators. In contrast to the general TFs, which interact directly with the RNA polymerase complex at the majority of genes, transcriptional activators regulate a subset of genes in response to environmental stimuli or differentiation signals by facilitating recruitment of the transcriptional machinery to loci bearing the correct UAS in addition to the core promoter motifs.

Although these *cis*- and *trans*-acting factors are major determinants of gene expression, they do not act in isolation on naked DNA; the linear chromosomal DNA in eukaryotes is highly compacted, interacting with a variety of proteins to form chromatin [reviewed in (40)]. The basic unit of chromatin is a nucleosome, consisting of approximately 146 base pairs (bp) of DNA wound around a protein core composed of eight histone proteins (two each of the histones H2A, H2B, H3, and H4). These nucleosomes are organized as "beads on a string" along the DNA strand, and the location and spacing of nucleosomes are accomplished by a combination of

DNA-encoded structural properties and chromatin remodeling proteins, which slide nucleosomes along the DNA strand [reviewed in (41)]. The organization of DNA into higher-order structures requires additional proteins including linker histones, which are important for the condensation of linear chromosomes into intermediate (30 nm) fibers and for further condensation during chromosome segregation. Importantly, the nucleosome provides an additional layer of transcriptional regulation by controlling the access of regulatory sequences to the transcriptional machinery—nucleosomes present a barrier between *cis*-regulatory elements and their cognate transcription factors, effectively blocking many transcription factors from recognizing their *cis*-regulatory motif [reviewed in (42,43)]. In addition, biochemical evidence suggests that RNA polymerases move inefficiently through nucleosome-bound templates (44), and eukaryotic RNA polymerase II is especially sensitive to nucleosomal barriers *in vitro*, arresting at discrete points that correlate with well-positioned nucleosomes [(45); reviewed in (46)]. In the context of eukaryotic chromatin *in vivo*, ATP-dependent chromatin remodelers are frequently found in close proximity to RNA polymerase II, working to enhance the capacity of RNA polymerases to move through nucleosome-bound templates [reviewed in (47); (48,49)]. Importantly, many of these chromatin remodelers are often coupled to domains or proteins that confer nucleosome-destabilizing epigenetic marks, which could decrease requirement for chromatin remodelers by decreasing the affinity of histones for DNA [reviewed in (47,50)]; these properties will be discussed in detail in the next section of this introduction. However, these data suggest that modulation of nucleosome stability and placement is also a major determinant of gene expression in eukaryotes.

A variety of mechanisms exist to promote or repress transcriptional activity independently *cis*-regulatory elements and their sequence-specific DNA binding proteins; these epigenetic modifiers function by tipping the balance toward a euchromatic, transcriptionally active state or toward a heterochromatic, transcriptionally silenced state. This can be accomplished in several ways: by altering the position of nucleosomes relative to *cis*-acting elements; altering the affinity of histones for one another or for DNA by inclusion of histone variants or through post-translational modification of histone proteins; or by chemical modification of the DNA itself. The literature describing these areas is broad and deep, as there is significant variation among eukaryotes in epigenetic marks and the protein interaction networks regulating these processes. A brief review of the general properties of histone variants, post-translational modification of histones, and chemical modification of DNA will be included below, with a focus on those which are conserved across a broad range of eukaryotes including kinetoplastids.

Nucleosome positioning and occupancy

The expression of genes from chromatin-bound DNA relies on the accessibility of relevant *cis*-acting elements, as the RNAP II transcriptional apparatus is not able to interact with nucleosome-bound sequences [reviewed in (44,50)]. Therefore, gene expression is heavily influenced by the specific placement of nucleosomes, and DNA-encoded elements that alter the favorability of nucleosome formation have been maintained in most eukaryotic genomes [reviewed in (51)]. Nucleosome positioning is broadly categorized into two classes: translational positioning, which indicates the preferred position of a nucleosome relative to other sequences of a similar length; and rotational positioning, which indicates the preferred position of the nucleosome relative to the 10.5 bp helical repeat of double-stranded DNA. The functional consequences of these two concepts are distinct in the context of nucleosome positioning *in vivo*, and these will be discussed in the context of *Leishmania* gene expression in Chapter 3.

The translational positioning of a nucleosome is most significantly altered by DNA sequences that strongly disfavor nucleosome formation, as nucleosomes occupy more stable positions upstream and downstream of these sequences. Both poly(dA:dT) and poly(dG:dC) homopolymer tracts function in this manner, as they are highly inflexible and are therefore refractory to nucleosome formation (52). Functionally, these sequences promote a more open chromatin environment around the homopolymer tract, increasing the accessibility of the DNA to transcription factors and the transcriptional apparatus [reviewed in (53)]. In yeast, poly(dA:dT) are essential components of the promoters of some housekeeping genes [(54,55); reviewed in (53)]. Interestingly, these sequences can be functionally substituted by structurally different poly(dG:dC) tracts, demonstrating that these sequences function by virtue of their inflexibility rather than by recruiting a sequence-specific DNA binding protein. The phenotypic effects of poly(dA:dT) tracts on the transcriptional activity of core promoters and UAS were more recently demonstrated using careful manipulations of homopolymer length, perfectness, and downstream promoter motifs (56). Here, the strongest effects on gene expression were observed for genes containing weak, degenerate promoter motifs when poly(dA:dT) tracts were present upstream of the promoter. Importantly, the length of the poly(dA:dT) tract was a major determinant of gene expression, and mismatches within the poly(dA:dT) tract decreased the effect but did not ablate it completely.

While the translational positioning of nucleosomes has major consequences on gene expression, the rotational positioning of a nucleosome may also play a role in gene expression by altering nucleosome positioning across regions with equivalent translational positioning potential. This quality is strongly influenced by individual dinucleotide pairs: AA/AT/TA/TT dinucleotides are significantly more “bendable”, while CC/CG/GC/GG dinucleotides are relatively inflexible [reviewed in (51)]. The number and periodicity of these dinucleotides produces tremendous variation in the favorability of a sequence for nucleosome formation, and quantitative analysis of the nucleosome-forming capacity of natural and non-natural 147 bp sequences demonstrates that the affinity of DNA sequences for histones varies over three orders of magnitude (57). Interestingly, in these studies the presence of these “bendable” A/T and inflexible G/C dinucleotides spaced with 10 bp periodicity was a key determinant of the nucleosome-forming capacity, and nucleosomes preferentially form when the A/T dinucleotides are positioned to interact with the histone core while the G/C dinucleotides are solvent-exposed. When this concept is extended to *in vivo* nucleosome positioning studies, it is apparent that the tendency for eukaryotic DNA to assemble into nucleosomes has been maintained throughout evolution, as eukaryotic genomes show consistent 10 bp periodicity across a population in regions lacking strongly favorable or disfavorable nucleosome positioning sequences [(58,59); reviewed in (51)]. Moreover, *in vitro* chromatin assembly experiments demonstrated that nucleosomes were nearly 10-fold more likely to assemble on eukaryote-derived genomic DNA than on bacterial genomic DNA, giving further credence to this concept (60). This phenomenon has a direct consequence in the choice of reporter genes in quantitative studies of eukaryotic *cis*-regulatory elements, such as those described in Chapters 4 and 5: we deliberately chose two eukaryotic reporter genes, the firefly and Renilla luciferases, as these would likely be incorporated easily into the endogenous chromatin environment.

Histone variant incorporation

Eukaryotes contain a number of histone variants which replace one of the core histones in the octamer and differ from their core histone counterparts in amino acid sequence [reviewed in (50,61)]. Histone variants are expressed throughout the cell cycle and incorporated into chromatin in a replication-independent manner, rather than only during DNA synthesis like core histones. It is unclear whether this phenomenon is also true in kinetoplastids, as histone variant and core histone mRNAs are expressed using similar mechanisms and are not transcriptionally regulated [reviewed in (62)]. Histone variant genes are typically present in a single copy in the genome while core histone genes are multicopy genes in most eukaryotes, including kinetoplastids; see Table 1-1 for documentation of the core and histone variant genes in *Leishmania.* This has important consequences for genetic studies of these proteins and facilitates localization studies using tagged proteins (63). Finally, core histone mRNAs usually contain an unusual secondary structure at the 3' end in place of a poly(A) tail, while histone variants are polyadenylylated like other cellular mRNAs. This does not appear to be the case in kinetoplastids, as core histone mRNAs are polyadenylylated [reviewed in (62)].

The functions of histone variant proteins in nucleosomes are diverse, with known roles in transcriptional regulation, DNA repair, DNA replication, and chromosome segregation [reviewed in (40)]. A summary of widely conserved histone variants, as well as the identity of the core histone and histone variant genes in *Leishmania*, are presented in Table 1-1; few of these histone variants are conserved in kinetoplastids and will not be discussed beyond this table. The extensively conserved variant H2A.Z is present in all eukaryotes studied to date with the notable exception of *Drosophila melanogaster,* which contains an H2A.V variant that combines the function of H2A.Z and the DNA repair-associated histone variant H2A.X. The biochemical consequences of H2A.Z incorporation are heavily debated, and the properties of chromatin containing H2A.Z-bearing nucleosome varies among eukaryotes [reviewed in (64)]. However, in all cases this variant associates with transcriptionally-active loci such as promoter elements, and typically contributes to the destabilization of the nucleosome particle. In the kinetoplastid *Trypanosoma brucei*, H2A.Z appears to function in a similar manner but has an unusually broad distribution (65); this and the other kinetoplastid-specific histone variants will be discussed later in this introduction and are the focus of Chapter 2 in this thesis.

Post-translational modification of histones

In addition to more substantial alteration of nucleosomes via the incorporation of histone variants, the core histones can be altered by post-translational modification (PTM), both on their flexible N- and C-terminal tails or on the globular core of the histone, which forms the "bead" of the nucleosome. These modifications are often reversible and require chromatin "writer" proteins, which transfer specific chemical groups to histones and chromatin "erasers", which remove these chemical groups. Although some modifications can directly alter the stability of the nucleosome via changes in histone-histone or histone-DNA contacts, many modifications interact with effector proteins called chromatin "readers", which bind to these chemical groups and recruit additional effector proteins to modulate the local chromatin environment [reviewed in (40,66)]. The repertoire of chromatin readers, erasers, and writers is ever expanding, and the number of chemical modifications to histones has increased substantially with advances in mass spectrometry and protein sequencing [reviewed in (66)]. A correlation of trypanosomatid histone modifications relative to those identified in other eukaryotes was previously described by Figueireido and colleagues (67), and the reader is referred to their work for a visual comparison of homology between trypanosomatid and human core histones and their modifications.

However, a summary of putative chromatin readers, writers, and erasers and their associated modifications in *Leishmania* and other trypanosomatid protozoa can be found in Table 1-2.

The identification of PTMs on the histone tails in various combinations led to the model of a “histone code”, which incorporates cross-talk between histone modifications, histone variants, and DNA modifications (68). Notably, functional networks relating to combinations of epigenetic marks have been elucidated, facilitated in large part by the development of chemical systems for direct ligation of chemical modifications to histones for *in vitro* characterization and the production of modification-specific antibodies, which can be used to co-localize epigenetic marks and their interacting partners *in vivo*. Much of the work relating to these functional networks has been performed in the context of the establishment and maintenance of heterochromatic domains, which function in transcriptional repression [see (69) for a detailed description of the epigenetic signatures of heterochromatin in *D. melanogaster*; also reviewed in (66,70)]. However, a large proportion of epigenetic marks associated with constitutive or facultative heterochromatin, including trimethylation of lysines 9 and 27 of histone H3 (H3K9me3, H3K27me) and their effector proteins in the HP1 and Polycomb families are not present in kinetoplastids (63). As a result, the focus of this section will be primarily on activating histone modifications associated with euchromatic loci.

Two well-studied classes of histone marks which are conserved in kinetoplastids are the reversible acetylation and methylation of lysine residues in the N-terminal tails of histones. These modifications were first described in 1964 and were postulated to have a function in modifying the efficiency of RNA synthesis from nucleosome-bound DNA templates (71)]. Histone acetyltransferases (HATs) function by transferring acetyl moieties from acetyl-CoA to lysine residues of histones, frequently on the basic, positively charged N-terminal histone tail.

Importantly, acetyllysine is negatively charged and has profound effects on both histone-DNA contacts, which require positively-charged histones to interact with negatively-charged DNA, and nucleosome-nucleosome interactions via similar charge repulsions. As a result, histone acetylation is typically considered to be an activating epigenetic mark associated with euchromatin and actively-transcribed loci, as it destabilizes nucleosomes and effectively decompacts chromatin. However, a family of proteins containing bromodomains interacts specifically with acetyllysine and can produce additional changes beyond those facilitated by the charge of the acetyl moiety. Removal of these acetyl groups is accomplished by a number of histone deacetylases (HDACs) [reviewed in (72)]; these proteins participate in epigenetic regulation of transcription at diverse loci, with tissue-specific functions and preferences for specific genomic loci.

In contrast, the function of histone methylation requires modification-specific effector proteins, as the addition of methyl groups to lysine residues does not alter the polarity or charge of the histone. However, up to three methyl groups can be added to a single lysine residue, and the reader proteins that recognize methyllysine residues are typically able to discriminate among mono-, di-, and trimethylated lysines. Furthermore, these reader proteins recognize methyllysine relative to its surroundings and also differentiate between different lysine residues. As a result, histone methylation can have drastically different consequences depending on the residue modified and the extent to which it is methylated [reviewed in (73)]. As mentioned previously, trimethylation of residues K9 and K27 of histone H3 are associated with heterochromatin formation, while di- and trimethylation of K4 of histone H3 (H3K4me2, H3K4me3) has a role in transcriptional activation near promoter elements, and mono-, di-, and trimethylation of K79 of histone H3 (H3K79me, H3K79me2, and H3K79me3) have important roles in cell division, DNA replication, and transcriptional regulation [reviewed in (40,66,73,74)]. Proteins containing chromodomains are the typical binding partners of methyllysine and include the proteins HP1 and members of the Polycomb family, which bind trimethylated H3K9 and H3K27, respectively [reviewed in (73,75)]. The proteins responsible for histone demethylation were only recently discovered (76), although approximately 20 different histone demethylases have now been characterized [reviewed in (77)]. Notably, many chromatin writers, erasers, and readers are dysregulated in a wide variety of human diseases, including cancer [reviewed in (78)]. This has led to significant interest among members of the pharmaceutical industry in developing small molecule inhibitors of these proteins, a topic that will be revisited in Chapter 7.

Covalent modification of DNA

A third facet of epigenetic regulation of gene expression in eukaryotes arises from the ability of the DNA itself to be covalently modified by various chemical groups. Importantly, many *trans*-acting factors that bind to sequence motifs in DNA interact extensively with the DNA backbone, and modifications can downregulate gene expression by preventing TFs from binding to their cognate *cis*-acting element (79). One highly studied DNA modification is that of cytosine methylation at the 5' position, which is accomplished by DNA methyltransferases (DNMTs). This modification has been found in a variety of eukaryotes but is not ubiquitous—the presence of DNA methylation in insects and in some fungi is low, if it exists at all (80). In eukaryotes in which DNA methylation and the associated DNMTs are present, the extent of methylation and the specific motifs which can be modified vary, and the rules governing the effect of DNA methylation on gene expression are not clear. More recently, it was shown that TET family proteins oxidize 5mc into several different derivatives which can be detected *in vivo*, including 5-hydroxymethylcytosine (hmC), 5-formylcytosine (fC), and 5-carboxycytosine (caC) (81). Notably, these derivatives were shown to have distinct functions *in vivo* and interact with distinct readers (82), suggesting that they have unique roles in eukaryotic DNA metabolism. In addition to possessing a canonical DNA methylation pathway, kinetoplastids also contain an unusual DNA modification referred to as base J which will be discussed at length in this introduction and in Chapter 2.

Leishmania **are just different: transcriptional regulation in kinetoplastids**

Polycistronic transcription and trans-*splicing in kinetoplastid protozoa*

Throughout the preceding sections, the multilayered strategies of eukaryotic transcriptional regulation were reviewed, encompassing the interactions between *cis*-acting elements, *trans*-acting factors, and epigenetic regulators. While these overarching principles are also implemented in kinetoplastid protozoa, the specific mechanisms and functional consequences of transcriptional regulation are quite different. In sharp contrast to the one gene-one promoter model common in most eukaryotes, *Leishmania* and other kinetoplastid protozoa including *Trypanosoma brucei* and *Trypanosoma cruzi* employ a highly unusual mechanism to generate messenger RNAs (mRNAs), which is reflected in their genome organization. Their protein-coding genes are organized in head-to-tail arrays containing potentially hundreds of functionally unrelated genes, referred to as polycistronic gene clusters (PGCs). PGCs are transcribed as polycistronic pre-mRNAs by RNAP II, which initiates primarily in divergent strand switch regions (dSSRs) where two PGCs are oriented head-to-head, and terminates in convergent strand switch regions (cSSRs) where two PGCs meet tail-to-tail [reviewed in (83); described in Figure 1-2]. A second transcript called the spliced leader (SL) RNA is transcribed by RNAP II from a separate locus and contains the cap and 5' end for all mature mRNAs in the cell. These RNAs are co-transcriptionally *trans*-spliced to generate monocistronic mRNAs, and the transcript immediately upstream of the site of *trans*-splicing is polyadenylylated in a reaction coupled to *trans*-splicing to generate mature mRNAs (84,85). Although the SL and pre-mRNA substrates involved in *trans*-splicing differ significantly from those in other eukaryotes, the machinery involved in *trans*-splicing is similar to that required for *cis*-splicing of introns in other eukaryotes [reviewed in (86)].

The consequences of polycistronic transcription in kinetoplastids are tremendous: mRNA ends are defined by *trans*-splicing, not transcription; gene expression is constitutive (87); and transcription initiation and termination events are concentrated at relatively few loci genome-wide. This presents a plethora of interesting questions related to the mechanisms regulating RNAP II transcription of protein-coding genes throughout the life cycle and makes it apparent that these processes may be ideal targets for therapeutic intervention at all stages of the life cycle, as they likely differ significantly from their mammalian counterparts. The remaining sections in this introduction will focus on what is known to date regarding transcription initiation and termination of protein-coding genes in kinetoplastids. Because the interpretation of published data and the experiments which will be discussed in this thesis require one to not only consider the variations throughout the life cycle but also to account for unusual features which do not exist in other eukaryotes, a brief review of the experimental tools and reagents which are available will also be included in this introduction.

Cis-*regulation of RNAP II transcription in Leishmania and related kinetoplastids*

Just as RNAP II transcribes protein-coding genes in other eukaryotes, RNAP II is responsible for the transcription of both PGCs and the SL RNA genes. However, the mechanisms controlling transcription of PGCs differ significantly from those used for genes in other eukaryotes, and they are also distinct from the mechanisms regulating SL RNA gene transcription. The publication of the kinetoplastid genomes demonstrated that *Leishmania* and other kinetoplastid protozoa contain no apparent specific transcriptional activator proteins and lack a significant number of the general transcription factors, namely those which confer specificity to the site of transcription initiation (63). These differences include extremely weak similarity to components of the TFIID subunit, which interacts directly with the TATA box and other promoter motifs in other eukaryotes, and a highly divergent TFIIB complex, which facilitates the appropriate definition of the transcription start site relative to the promoter in other eukaryotes (63,88). Importantly, although kinetoplastids contain a TFIID component TRF4 that resembles TATA binding protein (TBP), this gene is an orthologue of TBP-related factors and *T. cruzi* TBP/TRF4 demonstrates a preference for G/C-rich sequences *in vitro* (89). Chromatin immunoprecipitation studies in *Leishmania* demonstrate that TRF4 and the transcription factor complex SNAP$_c$, which is required for RNAP II-mediated transcription of small nuclear RNAs (snRNAs) in other eukaryotes, bind to regions associated with transcription initiation of both PGCs and SL RNA genes (37). However, binding of these proteins requires a well-defined promoter motif in SL RNA genes which is absent in dSSRs (90,91), demonstrating significant differences in the transcriptional regulation of these two gene classes.

In agreement with the lack of specificity-conferring transcription factors in *Leishmania*, examination of the *Leishmania* genome demonstrates a paucity of canonical eukaryotic transcription factor binding sites that might be used for transcription of protein-coding genes, including TATA boxes, Inr elements, BREs, and DPEs (63). Furthermore, genome-wide mapping of TSS in *T. brucei* and 5' RACE experiments in *L. major* showed multiple TSS were associated with individual PGCs, suggesting that a more delocalized process was occurring in dSSRs and at a limited number of PGC-internal regions of transcription initiation (34,92). Interestingly, comparative genomics in *T. brucei* demonstrated an overrepresentation of poly(dG:dC) tracts in regions of RNAP II transcription initiation, and the orientation of the poly(dG:dC) tract was hypothesized to confer directionality to RNAP II transcription initiating from these regions (65). This possibility has led to the speculation that these homopolymers may be the long sought after *cis*-acting elements involved in transcription of protein-coding genes in kinetoplastids. Notably, although two poly(dG:dC) tracts are present in the 73-bp region between opposing TSS mapped to the dSSR of chromosome 1 in *L. major*, these sequences are scattered throughout the genome and anre not overrepresented in regions assocaiated with transcription initiation (37). Although several attempts were made to demonstrate that this dSSR possesses promoter activity in reporter-based assays, these experiments were performed using multicopy episomal DNAs, which do not require any *Leishmania* elements for transcription, or stable integration into the ribosomal RNA locus, which contains an extremely strong promoter element (92). Even with these caveats, these experiments showed that inclusion of the dSSR produced extremely weak effects on reporter gene activity, and the amount of clonal variation in these lines makes these results rather unconvincing. We hypothesized that poly(dG:dC) tracts and other unknown sequences could function as *cis*-regulatory elements by nucleosome exclusion, similar to poly(dA:dT) tracts present in yeast. The potential roles of poly(dG:dC) tracts and the search for other novel *cis*-regulatory elements associated with PGC transcription are a major focus of this thesis and will be addressed in detail in Chapter 3, with respect to their roles in nucleosome positioning and in Chapters 4 and 5, with respect to their roles in bidirectional transcription initiating within a dSSR.

Epigenetic regulation of transcription in kinetoplastids

Because few indications pointed to a major role of *cis*-regulatory elements in controlling transcription of protein-coding genes in kinetoplastids, significant effort has been poured into the characterization of epigenetic networks that may be important for transcriptional regulation (described in Figure 1-3; data represent marks identified in *Leishmania*). The completed genomic sequences of *Leishmania* and trypanosomes demonstrated the presence of histone variants which appeared to replace histones H2A, H2B, and H3 (see Table 1-1); a histone H4 variant has been identified in *T. brucei* and *T. cruzi*, but the high level of amino acid divergence among H4 genes in *Leishmania* has made it difficult to identify an H4 variant in this species (93). Phylogenetic comparisons suggest that the H2A variant is related to the H2A.Z histone variant which is highly conserved among eukaryotes, while the H2B and H3 variants (H2B.V and H3.V) appear to be kinetoplastid-specific. A major breakthrough in the identification of transcription-associated epigenetic networks came in 2009, when Siegel and colleagues demonstrated the localization of these histone variants to the boundaries of PGCs: H2A.Z/H2B.V-containing nucleosomes are present in broad peaks at and around dSSRs, and H3.V/H4.V-containing nucleosomes are present at cSSRs. The *H2AZ* and *H2BV* genes in *T. brucei* could not be deleted without prior inclusion of an ectopic copy of the gene, suggesting these proteins are essential for viability. However, both the *H3V* and *H4V* genes were readily deleted, and no phenotypes relating to transcription termination have been documented in these mutants to date (65,94,95). The possible roles the histone variants H2A.Z, H2B.V, and H3.V in

Leishmania major will be discussed in greater detail in Chapter 2: there, we describe genetic tests to determine the essentiality of *H2A.Z* and *H2B.V*, as well as phenotypic assessment of *H3.V*-null *L. major* with respect to transcription termination.

The completed genome sequences also revealed the presence of a variety of chromatin readers, erasers, and writers which may be involved in post-translational modification of histones (63); genes identifying putative chromatin readers, writers, and erasers are summarized in Table 1-2, but will be discussed here in greater detail. Identification of PTMs on the core histones using mass spectrometry and Edman degradation in *T. brucei* demonstrated the presence of acetylated and methylated residues (96,97). However, both the number of modifications and the number of putative chromatin modifiers are greatly reduced in kinetoplastids compared to other eukaryotes (see Tables 1-1 and 1-2). Despite this, a few notable points were made in these studies (96,97). First, all 4 core histones contain modifications on their N-terminal residue which are highly abundant, suggesting that most histones contain this modification (methylalanine in H2A, H2B, and H4; acetylserine in H3); this phenomenon has not been observed in other eukaryotes to date, and it is unclear which writers, erasers, and readers might interact with these modifications. Second, the N-terminal tails of histones H2A and H2B contain very few modifications, unlike most model eukaryotes; rather, the C-terminal tail of H2A is highly acetylated, and mass spectrometry peptide analysis suggests ubiquitination also occurs on the H2A C-terminal tail. The function of these modifications are not clear at this time, but recent work demonstrated that phosphorylation of threonine 130 in H2A functions in DNA damage signaling and may functionally mimic γ-H2A.X in other eukaryotes (98). Although no modifications of the C-terminus of H2B were detected, attempts to tag H2B at the C-terminus in *Leishmania* were unsuccessful and generated mislocalized tagged protein (Robinson and

Beverley, unpublished data). This H2B-GFP fusion protein has been utilized in other eukaryotes for analysis of DNA content with no deleterious effects (99), suggesting that the C-terminus of H2B may be important for function or localization.

Finally, the N-terminal tails of histones H3 and H4 contain a variety of modifications, some of which appear to correlate with marks in other eukaryotes; as mentioned previously, these are summarized in Table 1-2. Importantly, the extreme N-terminus of H3 was not amenable to analysis in *T. brucei* due to its N-terminal acetylation, and documentation of histone modifications in these protozoa has not been peformed exhaustively. One example of a key histone modification that was not identified by mass spectrometry was that of trimethylation of histone H3 at lysine 4 (H3K4me3), which can be detected with modification-specific antisera (100). Nucleosomes containing this modification localize to dSSRs in *T. brucei* and *T. cruzi* (101,102) and preferentially include the histone variant H2B.V (100). This strongly resembles the observation that H3K4me3 marks regions of active transcription in many eukaryotes [reviewed in (66)]. Interestingly, antisera designed to detect acetylated histone H3 at lysines 9 and 14 (H3K9/K14ac) in *Tetrahymena thermophila* cross-reacts with *Leishmania* H3, and chromatin immunoprecipitations reveals broad peaks associated with dSSRs, similar to those shown with *T. brucei* H2A.Z and H2B.V. While the absolute identity of these marks is not known, these marks appear to denote sites of active transcription initiation and decrease during transcriptional downregulation in stationary phase promastigotes (37). Moreover, bromodomain-containing protein BDF3, which is expected to recognize acetyllysine residues, localizes to dSSRs in *T. brucei* (65), again suggesting a functional role for histone acetylation at these loci. An important fact to note is that in contrast to typical eukaryotic transcription initiation-associated marks, these modifications localize to relatively broad regions in and around dSSRs, typically encompassing 5-10 kilobases of the chromosome. However, promiscuous transcription initiation events were found spanning these regions in *T. brucei*, suggesting that these likely are not associated specifically with transcription elongation. Our examination of chromatin structural features in Chapter 3 identified similar phenomena, suggesting a similar role for these marks in *Leishmania*.

In addition to these modifications, mono-, di-, and trimethylation of histone H3 on lysine 76 (H3K76me, H3K76me2, H3K76me3) were identified by mass spectrometry and were later shown to be catalyzed by the SET-domain DOT1 histone methyltransferases DOT1A and DOT1B, as expected based on the location of the modification in the N-terminal tail (103). DOT1 proteins are important in chromosome segregation in many eukaryotes, and DOT1 mutants in *T. brucei* show defects in the cell cycle and DNA replication, suggesting some conservation in histone modification function (103,104). However, DOT1 proteins are part of transcriptional regulatory networks involving H3K4me3 and other histone modifications, and DOT1B mutants show defects in antigenic variation, a process which is regulated epigenetically in *T. brucei* (105).

Additional modifications of the H3 N-terminus were also identified in *T. brucei*—acetylation of lysine 23 and trimethylation of lysine 32 in histone H3 were observed, but no known functions have been assigned to these modifications. Similarly, the histone H4 N-terminus contains acetylation and methylation of a variety of residues (97). The HATs responsible for acetylation of H4K4, H4K10, and H4K14 have been identified in *T. brucei* or *Leishmania donovani* (106–108), and the patterns of histone acetylation have also been characterized extensively in *T. cruzi* (109). Although there is regulation of histone acetylation in response to ultraviolet irradiation or throughout the cell cycle, there is no apparent localization of these marks to dSSRs or cSSRs, suggesting they may not be required for transcriptional regulation; however, a role for these marks in transcriptional elongation has not been ruled out, although genetic studies of the chromatin writers responsible for these marks could test this idea.

A third group of epigenetic regulators in kinetoplastids is that of covalent modification of DNA, either by cytosine methylation or through the kinetoplastid-specific DNA modification β-D-glucopyranosyloxymethyluracil, referred to as base J. Cytosine methylation has been identified in *T. brucei*, *T. cruzi*, and *L. major* and is likely performed by the sole annotated DNMT gene, which belongs to the DNMT6 family (110–112). The broad distribution of DNA methylation and the usage of CG, CHG, and CHH motifs for cytosine methylation in kinetoplastids does not reflect a likely role in transcriptional regulation; however, treatment of *T. cruzi* with the DNA methyltransferase inhibitor 5-azacytidine resulted in an increase in cell growth. Although the authors postulate a role for DNA methylation in cell division or DNA replication, this phenomenon could also result from a direct effect on transcription (113).

In contrast to DNA methylation, major roles in transcriptional regulation have been identified for base J in *T. cruzi* and *L. major*. This DNA modification was originally localized to telomeric DNA and the variant surface glycoprotein genes in *T. brucei* (114,115), but chromatin immunoprecipitation coupled to high-throughput sequencing (ChIP-seq) using J-specific antisera also demonstrated that this modification is also present in dSSRs and cSSRs (116–118) in *T. brucei*, *T. cruzi*, and *L. major*. Interestingly, the epigenetic networks which base J participates in appear to differ among these three organisms based on genetic studies of the JBP1 and JBP2 proteins, which are members of the TET/JBP dioxygenase family of thymidine hydroxylase proteins and catalyze the first steps in J biosynthesis. Although their functions are similar, these proteins appear to play different roles in J biosynthesis, as JBP2 contains a SWI/SNF-family chromatin remodeling domain and appears to possess some sequence-specificity in J deposition (119). In *T. brucei*, neither *JBP1* nor *JBP2* is essential, and deletion of both genes generates normal, viable parasites (120). In *T. cruzi*, *JBP1-/-* and *JBP2-/-* parasites were viable but showed significant alterations in the rates of transcription initiation in dSSRs (117,121). To date, a double J-null mutant has not been successfully generated, suggesting that the base J modification may be essential in *T. cruzi*. In further contrast yet, only *JBP2*-/- parasites have been generated in *L. major*, as *JBP1* is an essential gene (118,122). Extended cultures of these parasites *in vitro* or treatment with BrdU decreases J levels to less than 30% of WT parasites (122), and these parasites are no longer viable. Transcriptome analysis demonstrated that transcription termination in cSSRs is hugely defective, and many dSSRs show alterations in transcription initiation (118). Importantly, the glucosyltransferase responsible for catalyzing the final step in J biosynthesis was recently identified (123), and additional studies of this protein will allow the separation of the other possible functions of *JBP1* and *JBP2* in transcriptional biology from the function of base J. The potential roles for DNA base J in *Leishmania* epigenetic networks will be addressed in Chapter 2 with respect to transcription termination, and its potential for functioning as a signal to define a transcriptionally permissive epigenetic state in dSSRs will be revisited in the future directions described in Chapter 7.

***Leishmania* are easy: tools and reagents for studies of parasite biology**

The work described throughout the preceding sections in this introduction arose out of significant investments in laboratory tools and reagents for the culture of *Leishmania* and related kinetoplastids. Robust *in vitro* systems exist for the study of *Leishmania* parasites, especially during the insect stage of the life cycle. Promastigotes from many species have been cultured in the laboratory in liquid suspension using rich growth media containing nutrients for which the parasites are auxotrophic. In this *in vitro* system, logarithmically-growing cells are representative of procyclic promastigotes, which divide rapidly (*L. major* doubling time = 4-6 hours) and are actively transcribing. The procyclic-to-metacyclic differentiation process can be accomplished by simply allowing the parasites to persist in stationary phase for several days, and nonreplicative, infective metacyclic promastigotes can be isolated using a Ficoll gradient (124) or using negative agglutination with peanut agglutinin (PNA) (125). Importantly, it appears that the vast majority of biological processes are maintained in this *in vitro* culture system; however, the capacity of *Leishmania* promastigotes to undergo genetic exchange through a sexual cycle has only been observed in sand fly infections (126).

While *Leishmania* promastigotes are easily cultured in the laboratory, much of the work characterizing *Leishmania* amastigotes has been performed in mouse and hamster models, and *in vivo* infections are the gold standard for experiments which may indicate the relevance of a particular pathway in the context of human leishmaniasis. However, several species of *Leishmania* are capable of differentiating *in vitro* to axenic amastigotes using careful manipulations of pH, nutrients, and incubation temperature (127–132). These systems have provided a useful tool for characterization of amastigote gene expression (132,133), and they are an important component of high-throughput small molecule screens for potential therapeutics for use in humans or other mammals (129,134). Moreover, some of these axenic systems are also amenable to transfection (135,136), and efforts are underway to establish conditions for successful transfection and reproducible plating of *L. braziliensis* axenic amastigotes in our laboratory. Although these systems will not be explored in detail in the work described here, they provide a very provocative avenue for future work which expands on these data, which will be discussed in Chapter 7.

In addition to the development of *in vitro* systems for the growth of *Leishmania*, previous members of the Beverley lab have established highly reproducible conditions for transfection of DNA molecules using electroporation (137), and isolation of individual transfectants can be accomplished easily by plating on semisolid medium containing selective antibiotics compatible with the selectable marker used. Antibiotic-marker gene pairs which are commonly used in *Leishmania* include G418/*neo*, hygromycin B/*hyg*, blasticidin/*BSD*, phleomycin/*ble (PHLEO)*, nourseothricin/*SAT (NAT)*, and puromycin/*PAC*. As a result of these efforts, genetic manipulations of *Leishmania* have become routine, and many expression systems and techniques for manipulations of genes of interest have been developed for use in a variety of contexts. A number of these systems are used in the work described in this thesis and will be discussed throughout this section, as the subtle differences between expression systems allow many aspects of *Leishmania* transcription to be explored.

Leishmania are capable of incorporating linearized DNA fragments introduced by transfection into their genomes using homologous recombination (HR), and transfection of linearized DNA fragments containing homologous sequences at the 5' and 3' ends usually results in integration of the DNA fragment into the preferred locus (138,139). The generation of null mutants or the *in situ* tagging of an endogenous gene is accomplished using a DNA fragment bearing selectable markers and accompanying RNA processing sequences flanked by ~500 base pairs (bp) of sequence homologous to the 5' and 3' flanking sequences of the gene of interest. In cases in which the flanking sequences are difficult to ascertain based on problems with genome assemblies or in which these regions contain repetitive or low-complexity sequences, gene disruptions are often used instead; here, a cassette bearing 5' and 3' sequences sufficient for *trans*-splicing flanking the selectable marker are inserted into the middle of the open reading frame (ORF), and the ORF itself is used for homologous recombination. In addition, a widely used expression system which integrates into the small subunit (SSU) of the ribosomal RNA locus allows one to generate parasites expressing one or more genes at extremely high levels from the nucleolar-localized ribosomal RNA promoter (137).

In addition to these methods, many *Leishmania* spp. are able to stably propagate circular extrachromosomal DNA fragments called episomes (140), which are transcribed promiscuously on both strands by a "run around" mechanism (141). Interestingly, not a single nucleotide of *Leishmania*-derived sequence is required for transcription from these DNA elements, and origins of replication or centromeric elements are not required for replication or transmission of the DNA to daughter cells during cell division (142). As a result, in many situations episomal transfections are used as controls for the integrity of *trans*-splicing signals and selectable markers in constructs used in HR-based methods: if the episomal transfection fails to generate colonies, there is likely a problem with the DNA construct itself. However, if the episomal transfection is successful but the linearized transfection is not, this may represent issues with HR at this locus or could suggest that the gene or chromosomal element being replaced is essential.

The utility of episomal DNA fragments is most obvious in the validation of a gene's essentiality. Here, the chromosomal alleles of a gene of interest (GOI) are deleted in the presence of an episomal copy of the gene, which is on a plasmid that bears the green fluorescent protein (GFP) gene and a selectable marker. Removal of selection for the episome allows the DNA to be lost over multiple rounds of DNA replication and cell division, and the loss of the episome in individual cells can be quantified by GFP levels. Sorting for GFP-negative cells in this population will generate viable progeny for genes that are not essential; this has been used for genes which have been challenging to delete, as it separates the physical isolation of null mutants from the processes of transfection and allelic replacement. However, if the gene is essential, few GFP-negative cells will be isolated, and those cells which do survive are false positives which maintained low levels of the plasmid (143). This method will be discussed in greater detail in Chapter 2, in which we used it to study the histone variants H2A.Z and H2B.V in *L. major*.

Although the previously described genetic techniques have made it possible to study the functions of many genes, the process is rather time-consuming: traditional techniques typically target one allele at a time, and the asexual nature of *Leishmania* replication *in vitro* prevents the utilization of genetic crossing to generate homozygous mutants. In addition, *Leishmania* display a remarkable tolerance for variation in chromosomal copy number, both during normal growth in culture (144–146) and under situations of stress, such as during drug treatment (147,148) or when attempts are made to delete an essential gene (122,143). As a result, under standard circumstances at least two rounds of transfection and homologous recombination are required to generate null mutants, but for those unlucky individuals who are studying genes present on an aneuploid chromosome, more would be required. The recent demonstration of a functional RNA interference (RNAi) pathway in parasites of the *Leishmania (Viannia)* subgenus effectively circumvents the vast majority of these road blocks (149) and has generated much excitement in the *Leishmania* field. This process, first demonstrated in *C. elegans* by Fire and Mello, allows specific, potent targeting of genes independent of their copy number using a single genetic manipulation, and has revolutionized the study of gene functions in eukaryotic systems (150). In the *Leishmania* system, double-stranded RNA (dsRNA) complementary to the gene of interest

(GOI) is generated from a stem-loop (StL) transgene, typically transcribed from a nucleolar-localized, ribosomally-integrated construct (149). This dsRNA is processed by the RNA-induced silencing complex (RISC), which uses 21-22 nucleotide fragments of the dsRNA trigger to identify its target mRNA and degrade it using the "slicer" activity present within the complex. Despite the utility of this process in other eukaryotic systems and its ability to target endogenous mRNAs in *Leishmania* (149), the study of essential genes using this method is just out of reach in *Leishmania*, as viable transfectants are not produced (Lye, Brettmann, Fowlkes, and Beverley, unpublished data). I will return to this subject in Chapter 6, in which I describe attempts to improve upon inducible gene expression technologies in *Leishmania*.

Aims and scope of thesis: investigations in transcriptional biology in *Leishmania* at the intersection of genetics and genomics

Although many contributions have been made in recent years to our understanding of the mechanisms controlling transcription of PGCs in kinetoplastids, significant gaps remain in our knowledge of these processes, specifically relating to the nature and function of DNA-encoded elements. We believe that the essential nature of this process throughout the *Leishmania* life cycle make many aspects of this process desirable targets for therapeutic intervention, as inhibition of even a single target within this could have wide-reaching effects on the parasite. The groundwork for laboratory-based studies of *Leishmania* have been laid through many years of hard work, and recent advances in high-throughput sequencing have vaulted genome-scale experiments to the forefront of the field of transcriptional regulation. Our abilities to do thorough genetic studies of individual loci in combination with global epigenome analysis put us in a position to make significant contributions to our understanding of the mechanisms controlling transcription in *Leishmania*, which could have additional ramifications if these processes are conserved across the kinetoplastid lineage.

The primary focus of the work in this thesis is on the characterization of *cis*-acting elements and epigenetic factors in transcription of PGCs in *Leishmania* promastigotes. We began these studies with the notion that *Leishmania* divergent SSRs likely contained some kind of *cis*-regulatory element, albeit a very elusive one, as these are ubiquitous among eukaryotic protein-coding genes. Additional studies demonstrating the overrepresentation of poly(dG:dC) tracts in dSSRs in *T. brucei* suggested that perhaps a more generic, nucleosome-disfavoring sequence was the *cis*-acting element (described in Figure 1-4). To identify novel *cis*-regulatory sequences and to explore the functional consequences of poly(dG:dC) in *Leishmania*, we devised a broad set of independent but complementary experiments, described in the following chapters. In Chapter 2, I describe genetic studies of the *Leishmania* histone variants H2A.Z, H2B.V, and H3.V and attempts to place these histone variants in the epigenetic regulatory networks at divergent and convergent SSRs. In Chapter 3, I present genome-scale experiments which seek to identify chromatin-based signatures indicative of active regulatory elements, with a specific focus on poly(dG:dC) tracts and their presumed nucleosome-disfavoring characteristics. While the profound tolerance for aneuploidy in *Leishmania* initially prompted us to develop a novel computational pipeline to analyze this data relative to control datasets, we found that this pipeline also addressed a number of technical and computational artifacts that have riddled prior nucleosome mapping studies, which will be discussed in this chapter. In Chapters 4 and 5, I describe the development and application of a novel dSSR-based bidirectional reporter system for transcriptional activity, which has proven to be a versatile reagent in the investigation of both genetic determinants of transcriptional activity. In Chapter 4, we utilized this reporter to assess the genetic contributors to dSSR-mediated transcription and present strong evidence against the existence of *cis*-regulatory elements in dSSRs. In addition, I present some unexpected findings in Chapter 5 regarding the possible role for poly(dG:dC) tracts in promoting the directionality of transcription, which appears to occur via an epigenetic mechanism. Finally, in Chapter 6 I discuss the implications of the new knowledge generated by these projects in the context of regulatable gene expression in *Leishmania*. The development of an inducible system for expression of protein-coding and RNAi transgenes would lead to significant advancements in the studies of essential genes, and this has encompassed a part of my work in the Beverley laboratory. Together, the body of work presented in this thesis describes advances in our understanding of parasite gene expression and set the stage for further characterization of these processes in life cycle stages which are relevant to human disease.

Figure Legends

Figure 1-1. Models of active promoter elements in eukaryotes. Black lines indicate genomic DNA, purple and green circles represent nucleosomes, and gray ovals represent *trans*-acting factors, including sequence-specific transcription factors and general transcription factors. Red arrows indicate transcribed mRNA; black circle indicates the cap of eukaryotic mRNAs. Specific DNA-protein contacts are typically important for active transcription, and the state of the chromatin at and around the *cis*-acting elements strongly alters the activity of these promoters.

Figure 1-2. Depiction of polycistronic transcription and *trans*-splicing of protein-coding genes and the spliced leader (SL) RNA. Yellow boxes indicate the SL RNA genes; the transcript generated from this locus bears a 5’ cap, indicated by a black circle. Genes present within polycistronic gene clusters are indicated in blue and red; these pre-mRNAs are transcribed as a polycistronic pre-mRNA. The polycistronic pre-mRNA and SL RNA undergo co-transcriptional *trans*-splicing, and the upstream transcript is polyadenylylated to form “typical” mature mRNAs.

Figure 1-3. Description of genetic and epigenetic elements associated with divergent and convergent SSRs. Genes within polycistronic gene clusters are indicated as blue and red box arrows; the polycistronic transcripts arising from transcription of these loci are depicted as blue and red line arrows. Green boxes indicate the divergent SSR. Gray circles indicate the general transcription factors TRF4 and $SNAP_c$; the gray arch indicates loci associated with the acetylated H3 histone modification. Loci bearing the covalent DNA modification base J are depicted as black circles with a J inside.

Figure 1-4. Proposed model for the nature and function of *cis*-acting elements in *Leishmania*. Elements in the schematic are identical to those in Figure 1-1.

Table 1-1. Histone variant genes in *Leishmania major*. Genes were identified using BLAST comparisons to known histone variants and by comparison to core histones; data are adapted from (63).

Table 1-2. Chromatin readers, writers, and erasers in trypanosomatid protozoa. Data are adapted from (37,63,67), as well as from current InterPro annotations on TriTrypDB (www.tritrypdb.org).

References

1. Control of the Leishmaniases: Report of a Meeting of the WHO Expert Committee on the Control of Leishmaniases. World Health Organization technical report series. 2010 Jan p. 1–186.

2. Bern C, Haque R, Chowdhury R, Ali M, Kurkjian KM, Vaz L, et al. The epidemiology of visceral leishmaniasis and asymptomatic leishmanial infection in a highly endemic Bangladeshi village. Am J Trop Med Hyg. 2007 May;76(5):909–14.

3. Ostyn B, Gidwani K, Khanal B, Picado A, Chappuis F, Singh SP, et al. Incidence of symptomatic and asymptomatic Leishmania donovani infections in high-endemic foci in India and Nepal: a prospective study. PLoS Negl Trop Dis. 2011 Oct;5(10):e1284.

4. Stauch A, Sarkar RR, Picado A, Ostyn B, Sundar S, Rijal S, et al. Visceral leishmaniasis in the Indian subcontinent: modelling epidemiology and control. PLoS Negl Trop Dis. 2011 Nov;5(11):e1405.

5. Hasker E, Kansal S, Malaviya P, Gidwani K, Picado A, Singh RP, et al. Latent infection with Leishmania donovani in highly endemic villages in Bihar, India. PLoS Negl Trop Dis. 2013 Jan;7(2):e2053.

6. Evans T, Teixeira M, McAuliffe I, Vasconcelos I, Vasconcelos A, Sousa A, et al. Epidemiology of visceral leishmaniasis in northeast Brazil. J Infect Dis. 1992 Nov;166(5):1124–32.

7. Moral L, Rubio E, Moya M. A leishmanin skin test survey in the human population of l'Alacanti region (Spain): implications for the epidemiology of Leishmania infantum infection in southern Europe. Trans R Soc Trop Med Hyg. 2002;96(2):129–32.

8. Chaves L, Pascual M. Climate cycles and forecasts of cutaneous leishmaniasis, a nonstationary vector-borne disease. PLoS Med. 2006 Aug;3(8):e295.

9. González C, Wang O, Strutz S, González-Salazar C, Sánchez-Cordero V, Sarkar S. Climate change and risk of leishmaniasis in north america: predictions from ecological niche models of vector and reservoir species. PLoS Negl Trop Dis. 2010 Jan;4(1):e585.

10. Moo-Llanes D, Ibarra-Cerdeña C, Rebollar-Téllez E, Ibáñez-Bernal S, González C, Ramsey J. Current and future niche of North and Central American sand flies (Diptera: psychodidae) in climate change scenarios. PLoS Negl Trop Dis. 2013 Jan;7(9):e2421.

11. Ito J, Ghosh A, Moreira L, Wimmer E, Jacobs-Lorena M. Transgenic anopheline mosquitoes impaired in transmission of a malaria parasite. Nature. 2002 May 23;417(6887):452–5.

12. Kamhawi S, Ramalho-Ortigao M, Pham V, Kumar S, Lawyer P, Turco S, et al. A role for insect galectins in parasite survival. Cell. 2004 Oct 29;119(3):329–41.

13. McDowell M, Ramalho-Ortigao M, Dillon R. Proposal for Sequencing the Genome of the Sand Flies , Lutzomyia longipalpis and Phlebotomus papatasi.

14. Das S, Freier A, Boussoffara T, Oswald D, Losch F, Selka M, et al. Modular Multiantigen T Cell Epitope-Enriched DNA Vaccine Against Human Leishmaniasis. Sci Trans Med. 2014 Apr 30;6(234):234ra56–234ra56.

15. Croft S, Sundar S, Fairlamb A. Drug Resistance in Leishmaniasis. Clin Microbiol Rev. 2006;19(1):111–26.

16. Gossage SM, Rogers ME, Bates P a. Two separate growth phases during the development of Leishmania in sand flies: implications for understanding the life cycle. Int J Parasitol. 2003 Sep;33(10):1027–34.

17. Handman E, Bullen DVR. Interaction of Leishmania with the host macrophage. Trends Parasitol. 2002 Aug;18(8):332–4.

18. Gluenz E, Ginger M, McKean P. Flagellum assembly and function during the Leishmania life cycle. Curr Opin Micro. Elsevier Ltd; 2010 Aug;13(4):473–9.

19. De Assis R, Ibraim I, Nogueira P, Soares R, Turco S. Glycoconjugates in New World species of Leishmania: polymorphisms in lipophosphoglycan and glycoinositolphospholipids and interaction with hosts. Biochim Biophys Acta. 2012 Sep;1820(9):1354–65.

20. Zhang K, Beverley S. Phospholipid and sphingolipid metabolism in Leishmania. Mol Biochem Parasitol. 2010 Apr;170(2):55–64.

21. Besteiro S, Williams R, Morrison L, Coombs G, Mottram J. Endosome sorting and autophagy are essential for differentiation and virulence of Leishmania major. J Biol Chem. 2006 Apr 21;281(16):11384–96.

22. Silva A, Cordeiro-da-Silva A, Coombs G. Metabolic variation during development in culture of Leishmania donovani promastigotes. PLoS Negl Trop Dis. 2011 Dec;5(12):e1451.

23. Michels P, Bringaud F, Herman M, Hannaert V. Metabolic functions of glycosomes in trypanosomatids. Biochim Biophys Acta. 2006 Dec;1763(12):1463–77.

24. Wheeler R, Gluenz E, Gull K. The cell cycle of Leishmania: morphogenetic events and their implications for parasite biology. Mol Microbiol. 2011 Feb;79(3):647–62.

25. Atkinson B, Walden D. Changes in Eukaryotic Gene Expression in Response to Environmental Stress. Academic Press (Orlando); 1985.

26. Mallo M, Alonso C. The regulation of Hox gene expression during animal development. Development. 2013 Oct;140(19):3951–63.

27. Bozdech Z, Llinás M, Pulliam B, Wong E, Zhu J, DeRisi J. The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. PLoS Biol. 2003 Oct;1(1):E5.

28. Kramer S. Developmental regulation of gene expression in the absence of transcriptional control: the case of kinetoplastids. Mol Biochem Parasitol. 2012 Feb;181(2):61–72.

29. Tsigankov P, Gherardini P, Helmer-Citterich M, Zilberstein D. What has proteomics taught us about Leishmania development? Parasitology. 2012 Aug;139(9):1146–57.

30. Akopyants N, Matlib R, Bukanova E, Smeds M, Brownstein B, Stormo G, et al. Expression profiling using random genomic DNA microarrays identifies differentially expressed genes associated with three major developmental stages of the protozoan parasite Leishmania major. Mol Biochem Parasitol. 2004 Jul;136(1):71–86.

31. Almeida R, Gilmartin B, McCann S, Norrish A, Ivens A, Lawson D, et al. Expression profiling of the Leishmania life cycle: cDNA arrays identify developmentally regulated genes present but not annotated in the genome. Mol Biochem Parasitol. 2004 Jul;136(1):87–100.

32. Lahav T, Sivam D, Volpin H, Ronen M, Tsigankov P, Green A, et al. Multiple levels of gene regulation mediate differentiation of the intracellular pathogen Leishmania. FASEB J. 2011 Feb;25(2):515–25.

33. Kelly S, Kramer S, Schwede A, Maini PK, Gull K, Carrington M. Genome organization is a major component of gene expression control in response to stress and during the cell division cycle in trypanosomes. Open Biol. 2012;2(4):120033.

34. Kolev N, Franklin J, Carmi S, Shi H, Michaeli S, Tschudi C. The transcriptome of the human pathogen Trypanosoma brucei at single-nucleotide resolution. PLoS Pathog. 2010 Jan;6(9):e1001090.

35. Kolev N, Ramey-Butler K, Cross G, Ullu E, Tschudi C. Developmental progression to infectivity in Trypanosoma brucei triggered by an RNA-binding protein. Science. 2012 Dec 7;338(6112):1352–3.

36. Kolev N, Ullu E, Tschudi C. The emerging role of RNA-binding proteins in the life cycle of Trypanosoma brucei. Cell Microbiol. 2014 Apr;16(4):482–9.

37. Thomas S, Green A, Sturm N, Campbell D, Myler P. Histone acetylations mark origins of polycistronic transcription in Leishmania major. BMC Genomics. 2009 Jan;10:152.

38. Lodish H, Berk A, Zipursky S, Matsudaira P, Baltimore D, Darnell J. Molecular Cell Biology. Molecular Cell Biology. 2000.

39. Lee T, Young R. Transcription of eukaryotic protein-coding genes. Annu Rev Genet. 2000;34:77–137.

40. Bell O, Tiwari V, Thomä N, Schübeler D. Determinants and dynamics of genome accessibility. Nat Rev Genet. 2011 Aug;12(8):554–64.

41. Felsenfeld G. Chromatin as an essential part of the transcriptional mechanism. Nature. 1992;355:219–24.

42. Cairns BR. The logic of chromatin architecture and remodelling at promoters. Nature. 2009 Sep 10;461(7261):193–8.

43. Zaret KS, Carroll JS. Pioneer transcription factors: establishing competence for gene expression. Genes Dev. 2011 Nov 1;25(21):2227–41.

44. Izban M, Luse D. Factor-stimulated RNA Polymerase I1 Transcribes at Physiological Elongation Rates on Naked DNA but Very Poorly on Chromatin Templates. J Biol Chem. 1992;267(19):13647–55.

45. Kireeva ML, Walter W, Tchernajenko V, Bondarenko V, Kashlev M, Studitsky VM. Nucleosome Remodeling Induced by RNA Polymerase II. Mol Cell. 2002 Mar;9(3):541–52.

46. Studitsky VM, Walter W, Kireeva M, Kashlev M, Felsenfeld G. Chromatin remodeling by RNA polymerases. Trends Biochem Sci. 2004 Mar;29(3):127–35.

47. Workman JL. Nucleosome displacement in transcription. Genes Dev. 2006 Aug 1;20(15):2009–17.

48. Orphanides G, LeRoy G, Chang CH, Luse DS, Reinberg D. FACT, a factor that facilitates transcript elongation through nucleosomes. Cell. 1998 Jan 9;92(1):105–16.

49. Carey M, Li B, Workman JL. RSC exploits histone acetylation to abrogate the nucleosomal block to RNA polymerase II elongation. Mol Cell. 2006 Nov 3;24(3):481–7.

50. Workman J, Kingston R. Alteration of nucleosome structure as a mechanism of transcriptional regulation. Ann Rev Biochem. 1998 Jan;67:545–79.

51. Struhl K, Segal E. Determinants of nucleosome positioning. Nat Struct Mol Biol. 2013 Mar;20(3):267–73.

52. Nelson H, Finch J, Luisi B, Klug A. The structure of an oligo(dA)-oligo(dT) tract and its biological implications. Nature. 1987;330:221–6.

53. Segal E, Widom J. Poly(dA:dT) tracts: major determinants of nucleosome organization. Curr Opin Struct Biol. 2009 Feb;19(1):65–71.

54. Struhl K. Naturally occurring poly(dA-dT) sequences are upstream promoter elements for constitutive transcription in yeast. Proc Natl Acad Sci USA. 1985;82(December):8419–23.

55. Iyer V, Struhl K. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. EMBO J. 1995 Jun 1;14(11):2570–9.

56. Raveh-Sadka T, Levo M, Shabi U, Shany B, Keren L, Lotan-Pompan M, et al. Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. Nat Genet. 2012 Jul;44(7):743–50.

57. Thåström A, Lowary P, Widlund H, Cao H, Kubista M, Widom J. Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. J Mol Biol. 1999 Apr 30;288(2):213–29.

58. Gaffney D, McVicker G, Pai A, Fondufe-Mittendorf Y, Lewellen N, Michelini K, et al. Controls of nucleosome positioning in the human genome. PLoS Genet. 2012;8(11):e1003036.

59. Moyle-Heyrman G, Zaichuk T, Xi L, Zhang Q, Uhlenbeck O. Chemical map of Schizosaccharomyces pombe reveals species-specific features in nucleosome positioning. Proc Natl Acad Sci USA. 2013;110(50):20158–63.

60. Zhang Y, Moqtaderi Z, Rattner BP, Euskirchen G, Snyder M, Kadonaga JT, et al. Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. Nat Struct Mol Biol. 2009 Aug;16(8):847–52.

61. Marzluff W, Wagner E, Duronio R. Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. Nat Rev Genet. 2008;9:843–54.

62. Alsford S, Horn D. Trypanosomatid histones. Mol Microbiol. 2004;53(2):365–72.

63. Ivens A, Peacock C, Worthey E, Murphy L, Aggarwal G, Berriman M, et al. The genome of the kinetoplastid parasite, Leishmania major. Science. 2005 Jul 15;309(5733):436–42.

64. Bönisch C, Hake S. Histone H2A variants in nucleosomes and chromatin: more or less stable? Nucleic Acids Res. 2012 Nov;40(21):10719–41.

65. Siegel T, Hekstra D, Kemp L, Figueiredo L, Lowell J, Fenyo D, et al. Four histone variants mark the boundaries of polycistronic transcription units in Trypanosoma brucei. Genes Dev. 2009 May 1;23(9):1063–76.

66. Rothbart S, Strahl B. Interpreting the language of histone and DNA modifications. Biochim Biophys Acta. 2014 Mar 12;

67. Figueiredo LM, Cross G a M, Janzen CJ. Epigenetic regulation in African trypanosomes: a new kid on the block. Nat Rev Microbiol. 2009 Jul;7(7):504–13.

68. Strahl B, Allis C. The language of covalent histone modfications. Nature. 2000;403(January):41–5.

69. Riddle N, Minoda A, Kharchenko P, Alekseyenko A, Schwartz Y, Tolstorukov M, et al. Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin. Genome Res. 2011 Feb;21(2):147–63.

70. Grewal S, Jia S. Heterochromatin revisited. Nat Rev Genet. 2007 Jan;8(1):35–46.

71. Allfrey V, Faulkner R, Mirsky A. Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. Proc Natl Acad Sci USA. 1964;315(1938):786–94.

72. Delcuve G, Khan D, Davie J. Roles of histone deacetylases in epigenetic regulation: emerging paradigms from studies with inhibitors. Clin Epigenet. 2012 Jan;4(1):5.

73. Wozniak G, Strahl B. Hitting the “mark”: Interpreting lysine methylation in the context of active transcription. Biochim Biophys Acta. 2014 Mar 12;2–10.

74. Nguyen A, Zhang Y. The diverse functions of Dot1 and H3K79 methylation. Genes Dev. 2011 Jul 1;25(13):1345–58.

75. Daniel J, Pray-Grant M, Grant P. Effector proteins for methylated histones: an expanding family. Cell Cycle. 2005;4(7):919–26.

76. Shi Y, Lan F, Matson C, Mulligan P, Whetstine J, Cole P, et al. Histone Demethylation Mediated by the Nuclear Amine Oxidase Homolog LSD1. Cell. 2004;119:941–53.

77. Shi Y, Tsukada Y. The discovery of histone demethylases. Cold Spring Harb Persp Biol. 2013 Sep;5(9).

78. Blancafort P, Jin J, Frye S. Writing and rewriting the epigenetic code of cancer cells: from engineered proteins to small molecules. Mol Pharmacol. 2013 Mar;83(3):563–76.

79. Watt F, Molloy P. Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. Genes Dev. 1988 Sep 1;2(9):1136–43.

80. Capuano F, Mülleder M, Kok R, Blom H, Ralser M. Cytosine DNA Methylation Is Found in Drosophila melanogaster but Absent in Saccharomyces cerevisiae,

Schizosaccharomyces pombe, and Other Yeast Species. Anal Chem. 2014 Apr 15;86(8):3697–702.

81. Ito S, Shen L, Dai Q, Wu S, Collins L, Swenberg J, et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science. 2011 Sep 2;333(6047):1300–3.

82. Spruijt C, Gnerlich F, Smits A, Pfaffeneder T, Jansen P, Bauer C, et al. Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. Cell. 2013 Mar 28;152(5):1146–59.

83. Gunzl A, Vanhamme L, Myler P. Transcription in trypanosomes: a different means to the end. In: Barry J, Mottram J, McCulloch R, Acosta-Serrano A, editors. Trypanosomes - After the Genome. Horizon Bioscience, Wymondham, Norfolk, UK; 2007. p. 177–208.

84. Matthews K, Tschudi C, Ullu E. A common pyrimidine-rich motif governs trans-splicing and polyadenylation of tubulin polycistronic pre-mRNA in trypanosomes. Genes Dev. 1994 Feb 15;8(4):491–501.

85. LeBowitz J, Smith H, Rusche L, Beverley S. Coupling of poly(A) site selection and trans-splicing in Leishmania. Genes Dev. 1993 Jun 1;7(6):996–1007.

86. Agabian N. Trans Splicing of Nuclear Pre-mRNAs. Cell. 1990;61(6):1157–60.

87. Leifso K, Cohen-Freue G, Dogra N, Murray A, McMaster WR. Genomic and proteomic expression analysis of Leishmania promastigote and amastigote life stages: the Leishmania genome is constitutively expressed. Mol Biochem Parasitol. 2007 Mar;152(1):35–46.

88. Palenchar J, Liu W, Palenchar P, Bellofatto V. A Divergent Transcription Factor TFIIB in Trypanosomes Is Required for RNA Polymerase II-Dependent Spliced Leader RNA Transcription and Cell Viability. Euk Cell. 2006;5(2):293–300.

89. Cribb P, Esteban L, Trochine A, Girardini J, Serra E. Trypanosoma cruzi TBP shows preference for C/G-rich DNA sequences in vitro. Exp Parasitol. 2010 Mar;124(3):346–9.

90. Yu M, Sturm N, Saito R, Roberts T, Campbell D. Single nucleotide resolution of promoter activity and protein binding for the Leishmania tarentolae spliced leader RNA gene. Mol Biochem Parasitol. 1998 Aug 1;94(2):265–81.

91. Gilinger G, Bellofatto V. Trypanosome spliced leader RNA genes contain the first identified RNA polymerase II gene promoter in these organisms. Nucleic Acids Res. 2001 Apr 1;29(7):1556–64.

92. Martínez-Calvillo S, Yan S, Nguyen D, Fox M, Stuart K, Myler P. Transcription of Leishmania major Friedlin chromosome 1 initiates in both directions within a single region. Mol Cell. 2003;11(5):1291–9.

93. Dobson DE, Scholtes LD, Myler PJ, Turco SJ, Beverley SM. Genomic organization and expression of the expanded SCG/L/R gene family of Leishmania major: internal clusters and telomeric localization of SCGs mediating species-specific LPG modifications. Mol Biochem Parasitol. 2006 Apr;146(2):231–41.

94. Lowell J, Cross G. A variant histone H3 is enriched at telomeres in Trypanosoma brucei. J Cell Sci. Co Biol; 2004 Nov 15;117(Pt 24):5937–47.

95. Lowell J, Kaiser F, Janzen C, Cross G. Histone H2AZ dimerizes with a novel variant H2B and is enriched at repetitive DNA in Trypanosoma brucei. J Cell Sci. 2005 Dec 15;118(24):5721–30.

96. Janzen C, Fernandez J, Deng H, Diaz R, Hake S, Cross G. Unusual histone modifications in Trypanosoma brucei. FEBS Lett. 2006 Apr 17;580(9):2306–10.

97. Mandava V, Fernandez J, Deng H, Janzen C, Hake S, Cross G. Histone modifications in Trypanosoma brucei. Mol Biochem Parasitol. 2007 Nov;156(1):41–50.

98. Glover L, Horn D. Trypanosomal (gamma)H2A and the DNA damage response. Mol Biochem Parasitol. 2012;183(1):78–83.

99. Kanda T, Sullivan KF, Wahl GM. Histone-GFP fusion protein enables sensitive analysis of chromosome dynamics in living mammalian cells. Curr Biol. 1998 Mar 26;8(7):377–85.

100. Mandava V, Janzen C, Cross G. Trypanosome H2Bv replaces H2B in nucleosomes enriched for H3 K4 and K76 trimethylation. Biochem Biophys Res Comm. 2008 Apr 18;368(4):846–51.

101. Wright J, Siegel T, Cross G. Histone H3 trimethylated at lysine 4 is enriched at probable transcription start sites in Trypanosoma brucei. Mol Biochem Parasitol. 2010 Aug;172(2):141–4.

102. Respuela P, Ferella M, Rada-Iglesias A, Aslund L. Histone acetylation and methylation at sites initiating divergent polycistronic transcription in Trypanosoma cruzi. J Biol Chem. 2008 Jun 6;283(23):15884–92.

103. Janzen C, Hake S, Lowell J, Cross G. Selective di- or trimethylation of histone H3 lysine 76 by two DOT1 homologs is important for cell cycle regulation in Trypanosoma brucei. Mol Cell. 2006 Aug;23(4):497–507.

104. Gassen A, Brechtefeld D, Schandry N, Arteaga-Salas J, Israel L, Imhof A, et al. DOT1A-dependent H3K76 methylation is required for replication regulation in Trypanosoma brucei. Nucleic Acids Res. 2012 Nov 1;40(20):10302–11.

105. Figueiredo L, Janzen C, Cross G. A histone methyltransferase modulates antigenic variation in African trypanosomes. PLoS Biol. 2008 Jul 1;6(7):e161.

106. Siegel T, Kawahara T, Degrasse J, Janzen C, Horn D, Cross G. Acetylation of histone H4K4 is cell cycle regulated and mediated by HAT3 in Trypanosoma brucei. Mol Microbiol. 2008 Feb;67(4):762–71.

107. Kawahara T, Siegel T, Ingram A, Alsford S, Cross G, Horn D. Two essential MYST-family proteins display distinct roles in histone H4K10 acetylation and telomeric silencing in trypanosomes. Mol Microbiol. 2008 Aug;69(4):1054–68.

108. Kumar D, Rajanala K, Minocha N, Saha S. Histone H4 lysine 14 acetylation in Leishmania donovani is mediated by the MYST-family protein HAT4. Microbiology. 2012 Feb;158(2):328–37.

109. Nardelli S, da Cunha J, Motta M, Schenkman S. Distinct acetylation of Trypanosoma cruzi histone H4 during cell cycle, parasite differentiation, and after DNA damage. Chromosoma. 2009 Aug;118(4):487–99.

110. Rojas M, Galanti N. DNA methylation in Trypanosoma cruzi. FEBS Lett. 1990 Apr 9;263(1):113–6.

111. Militello K, Wang P, Jayakar S, Pietrasik R, Dupont C, Dodd K, et al. African trypanosomes contain 5-methylcytosine in nuclear DNA. Eukaryot Cell. 2008 Nov;7(11):2012–6.

112. Huff J, Zilberman D. Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes. Cell. Elsevier Inc.; 2014 Mar 13;156(6):1286–97.

113. Rojas M, Galanti N. Relationship between DNA methylation and cell proliferation in Trypanosoma cruzi. FEBS Lett. 1991 Dec 16;295(1-3):31–4.

114. Gommers-Ampt J, Leeuwen F, de Beer A, Vliegenthart J, Dizdaroglu M, Kowalak J, et al. A Novel Modified Base Present in the DNA of the Parasitic Protozoan T . brucei. Cell. 1993;75:1129–36.

115. Van Leeuwen F, Wijsman E, Kieft R, Van Der Marel G, Van Boom J, Borst P. Localization of the modified base J in telomeric VSG gene expression sites of Trypanosoma brucei. Genes Dev. 1997 Dec 1;11(23):3232–41.

116. Cliffe L, Siegel T, Marshall M, Cross G, Sabatini R. Two thymidine hydroxylases differentially regulate the formation of glucosylated DNA at regions flanking polymerase

II polycistronic transcription units throughout the genome of Trypanosoma brucei. Nucleic Acids Res. 2010 Jul;38(12):3923–35.

117. Ekanayake D, Minning T, Weatherly B, Gunasekera K, Nilsson D, Tarleton R, et al. Epigenetic regulation of transcription and virulence in Trypanosoma cruzi by O-linked thymine glucosylation of DNA. Mol Cell Biol. 2011 Apr;31(8):1690–700.

118. Van Luenen H, Farris C, Jan S, Genest P, Tripathi P, Velds A, et al. Glucosylated hydroxymethyluracil, DNA base J, prevents transcriptional readthrough in Leishmania. Cell. 2012 Aug 31;150(5):909–21.

119. DiPaolo C, Kieft R, Cross M, Sabatini R. Regulation of trypanosome DNA glycosylation by a SWI2/SNF2-like protein. Mol Cell. 2005 Feb 4;17(3):441–51.

120. Cliffe L, Kieft R, Southern T, Birkeland S, Marshall M, Sweeney K, et al. JBP1 and JBP2 are two distinct thymidine hydroxylases involved in J biosynthesis in genomic DNA of African trypanosomes. Nucleic Acids Res. 2009 Apr;37(5):1452–62.

121. Ekanayake D, Sabatini R. Epigenetic regulation of polymerase II transcription initiation in Trypanosoma cruzi: modulation of nucleosome abundance, histone modification, and polymerase occupancy by O-linked thymine DNA glucosylation. Eukaryot Cell. 2011 Nov;10(11):1465–72.

122. Genest P, ter Riet B, Dumas C, Papadopoulou B, van Luenen H, Borst P. Formation of linear inverted repeat amplicons following targeting of an essential gene in Leishmania. Nucleic Acids Res. 2005 Jan;33(5):1699–709.

123. Bullard W, Lopes da Rosa-Spiegler J, Liu S, Wang Y, Sabatini R. Identification of the glucosyltransferase that converts hydroxymethyluracil to base J in the trypanosomatid genome. J Biol Chem. 2014 Jun 2;In press.

124. Späth G, Beverley S. A lipophosphoglycan-independent method for isolation of infective Leishmania metacyclic promastigotes by density gradient centrifugation. Exp Parasitol. 2001 Oct;99(2):97–103.

125. Da Silva R, Sacks DL. Metacyclogenesis is a major determinant of Leishmania promastigote virulence and attenuation. Infect Immun. 1987;55(11):2802–6.

126. Akopyants N, Kimblin N, Secundino N, Patrick R, Peters N, Lawyer P, et al. Demonstration of genetic exchange during cyclical development of Leishmania in the sand fly vector. Science. 2009 Apr 10;324(5924):265–8.

127. Bates P, Robertson C, Tetley L, Coombs G. Axenic cultivation and characterization of Leishmania mexicana amastigote-like forms. Parasitology. 1992;105:193–202.

128. Saar Y, Ransford A, Waldman E, Mazareb S, Amin-Spector S, Plumblee J, et al. Characterization of developmentally-regulated activities in axenic amastigotes of Leishmania donovani. Mol Biochem Parasitol. 1998;95(1):9–20.

129. Sereno D, Lemesre J. Axenically cultured amastigote forms as an in vitro model for investigation of antileishmanial agents. Antimicrob Agents Chemother. 1997;41(5):972–6.

130. Teixeira M, de Jesus Santos R, Sampaio R, Pontes-de-Carvalho L, Dos-Santos W. A simple and reproducible method to obtain large numbers of axenic amastigotes of different Leishmania species. Parasitol Res. 2002;88(11):963–8.

131. Hodgkinson V, Soong L, Duboise S, McMahon-Pratt D. Leishmania amazonensis: cultivation and characterization of axenic amastigote-like organisms. Exp Parasitol. 1996;83:94–105.

132. Walker J, Vasquez J, Gomez M, Drummelsmith J, Burchmore R, Girard I, et al. Identification of developmentally-regulated proteins in Leishmania panamensis by proteome profiling of promastigotes and axenic amastigotes. Mol Biochem Parasitol. 2006;147(1):64–73.

133. Rochette A, Raymond F, Corbeil J, Ouellette M, Papadopoulou B. Whole-genome comparative RNA expression profiling of axenic and intracellular amastigote forms of Leishmania infantum. Mol Biochem Parasitol. 2009 May;165(1):32–47.

134. De Rycker M, Hallyburton I, Thomas J, Campbell L, Wyllie S, Joshi D, et al. Comparison of a high-throughput high-content intracellular Leishmania donovani assay with an axenic amastigote assay. Antimicrob Agents Chemother. 2013 Jul;57(7):2913–22.

135. Coburn CM, Otteman KM, McNeely T, Turco SJ, Beverley SM. Stable DNA transfection of a wide range of trypanosomatids. Mol Biochem Parasitol. 1991 May;46(1):169–79.

136. Roy TAN, Lemesre JL, Papadopoulou B, Sereno D, Ouellette M. DNA Transformation of Leishmania infantum Axenic Amastigotes and Their Use in Drug Screening. 2001;45(4):1168–73.

137. Robinson K, Beverley S. Improvements in transfection efficiency and tests of RNA interference (RNAi) approaches in the protozoan parasite Leishmania. Mol Biochem Parasitol. 2003 May;128(2):217–28.

138. Cruz A, Beverley S. Gene replacement in parasitic protozoa. Nature. 1990;348(6297):171–3.

139. Cruz A, Coburn C, Beverley S. Double targeted gene replacement for creating null mutants. Proc Natl Acad Sci USA. 1991 Aug 15;88(16):7170–4.

140. LeBowitz J, Coburn C, McMahon-Pratt D, Beverley S. Development of a stable Leishmania expression vector and application to the study of parasite surface antigen genes. Proc Natl Acad Sci USA. 1990 Dec;87(24):9736–40.

141. Curotto de Lafaille M, Laban A, Wirth D. Gene expression in Leishmania: analysis of essential 5’ DNA sequences. Proc Natl Acad Sci USA. 1992 Apr 1;89(7):2703–7.

142. Papadopoulou B, Roy G, Ouellette M. Autonomous replication of bacterial DNA plasmid oligomers in Leishmania. Mol Biochem Parasitol. 1994 May;65(1):39–49.

143. Murta S, Vickers T, Scott D, Beverley S. Methylene tetrahydrofolate dehydrogenase/cyclohydrolase and the synthesis of 10-CHO-THF are essential in Leishmania major. Mol Microbiol. 2009 Mar;71(6):1386–401.

144. Sterkers Y, Lachaud L, Crobu L, Bastien P, Pagès M. FISH analysis reveals aneuploidy and continual generation of chromosomal mosaicism in Leishmania major. Cell Microbiol. 2011 Feb;13(2):274–83.

145. Sterkers Y, Lachaud L, Bourgeois N, Crobu L, Bastien P, Pagès M. Novel insights into genome plasticity in Eukaryotes: mosaic aneuploidy in Leishmania. Mol Microbiol. 2012 Oct;86(1):15–23.

146. Mannaert A, Downing T, Imamura H, Dujardin J. Adaptive mechanisms in pathogens: universal aneuploidy in Leishmania. Trends Parasitol. 2012 Sep;28(9):370–6.

147. Ubeda J, Légaré D, Raymond F, Ouameur A, Boisvert S, Rigault P, et al. Modulation of gene expression in drug resistant Leishmania is associated with gene amplification, gene deletion and chromosome aneuploidy. Genome Biol. 2008 Jan;9(7):R115.

148. Downing T, Imamura H, Decuypere S, Clark T, Coombs G, Cotton J, et al. Whole genome sequencing of multiple Leishmania donovani clinical isolates provides insights into population structure and mechanisms of drug resistance. Genome Res. 2011;21:2143–56.

149. Lye L, Owens K, Shi H, Murta S, Vieira A, Turco S, et al. Retention and loss of RNA interference pathways in trypanosomatid protozoans. PLoS Pathog. 2010;6(10):e1001161.

150. Fire A, Xu S, Montgomery M, Kostas S, Driver S, Mello C. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature. 1998;391(February):806–11.

151. Alsford S, Horn D. Elongator protein 3b negatively regulates ribosomal DNA transcription in african trypanosomes. Mol Cell Biol. 2011 May;31(9):1822–32.

152. Ingram AK, Horn D. Histone deacetylases in Trypanosoma brucei: two are essential and another is required for normal cell cycle progression. Mol Microbiol. 2002 Jul;45(1):89–97.

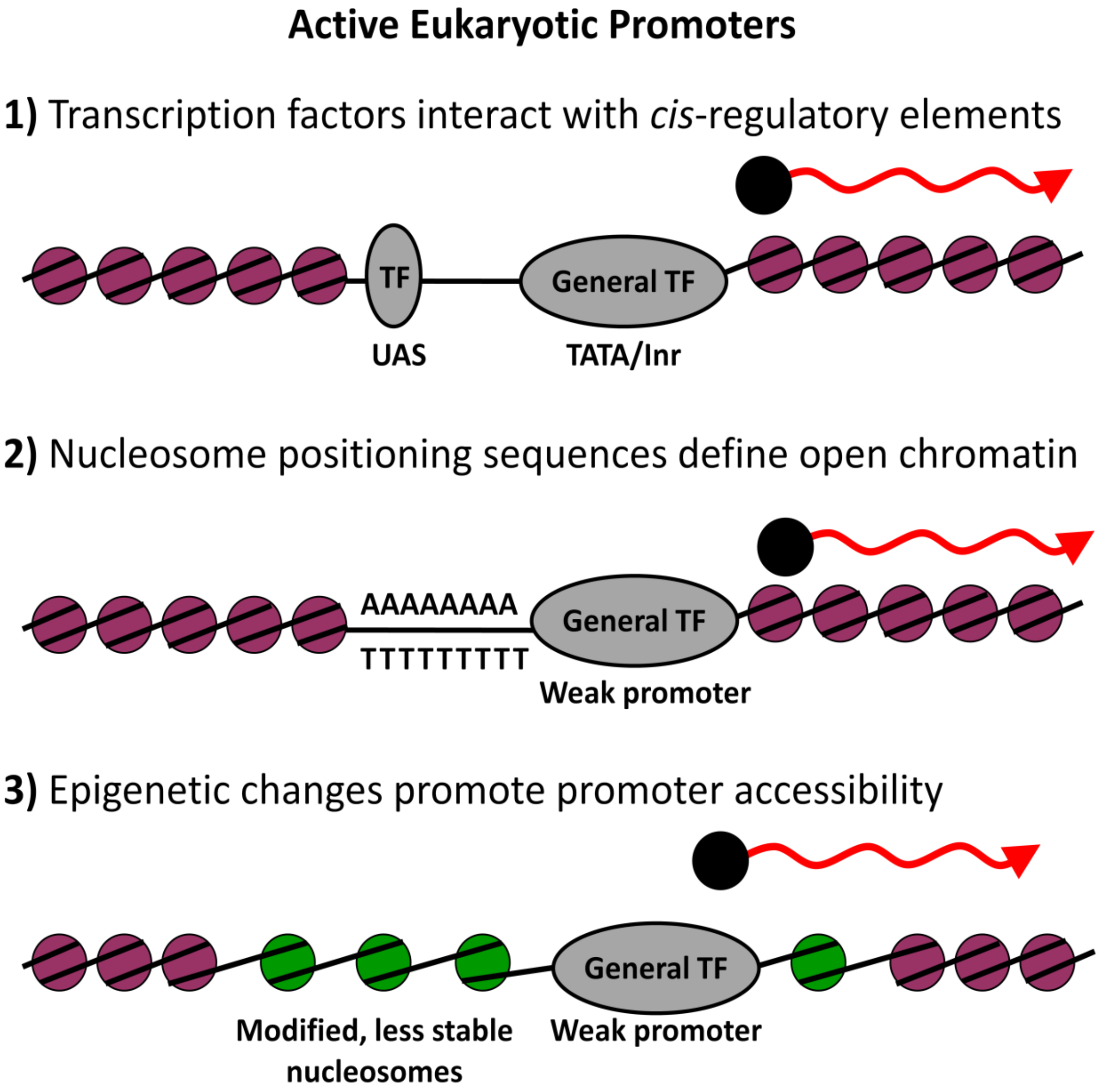
153. Alsford S, Kawahara T, Isamah C, Horn D. A sirtuin in the African trypanosome is involved in both DNA repair and telomeric gene silencing but is not required for antigenic variation. Mol Microbiol. 2007 Feb;63(3):724–36.

154. Fisk JC, Zurita-Lopez C, Sayegh J, Tomasello DL, Clarke SG, Read LK. TbPRMT6 is a type I protein arginine methyltransferase that contributes to cytokinesis in Trypanosoma brucei. Eukaryot Cell. 2010 Jun;9(6):866–77.

155. Fisk JC, Sayegh J, Zurita-Lopez C, Menon S, Presnyak V, Clarke SG, et al. A type III protein arginine methyltransferase from the protozoan parasite Trypanosoma brucei. J Biol Chem. 2009 Apr 24;284(17):11590–600.

156. Villanova GV, Nardelli SC, Cribb P, Magdaleno A, Silber AM, Motta MCM, et al. Trypanosoma cruzi bromodomain factor 2 (BDF2) binds to acetylated histones and is accumulated after UV irradiation. Int J Parasitol. 2009 May;39(6):665–73.

Figure 1-1.



Active Eukaryotic Promoters

1) Transcription factors interact with *cis*-regulatory elements

2) Nucleosome positioning sequences define open chromatin

3) Epigenetic changes promote promoter accessibility

Figure 1-2.

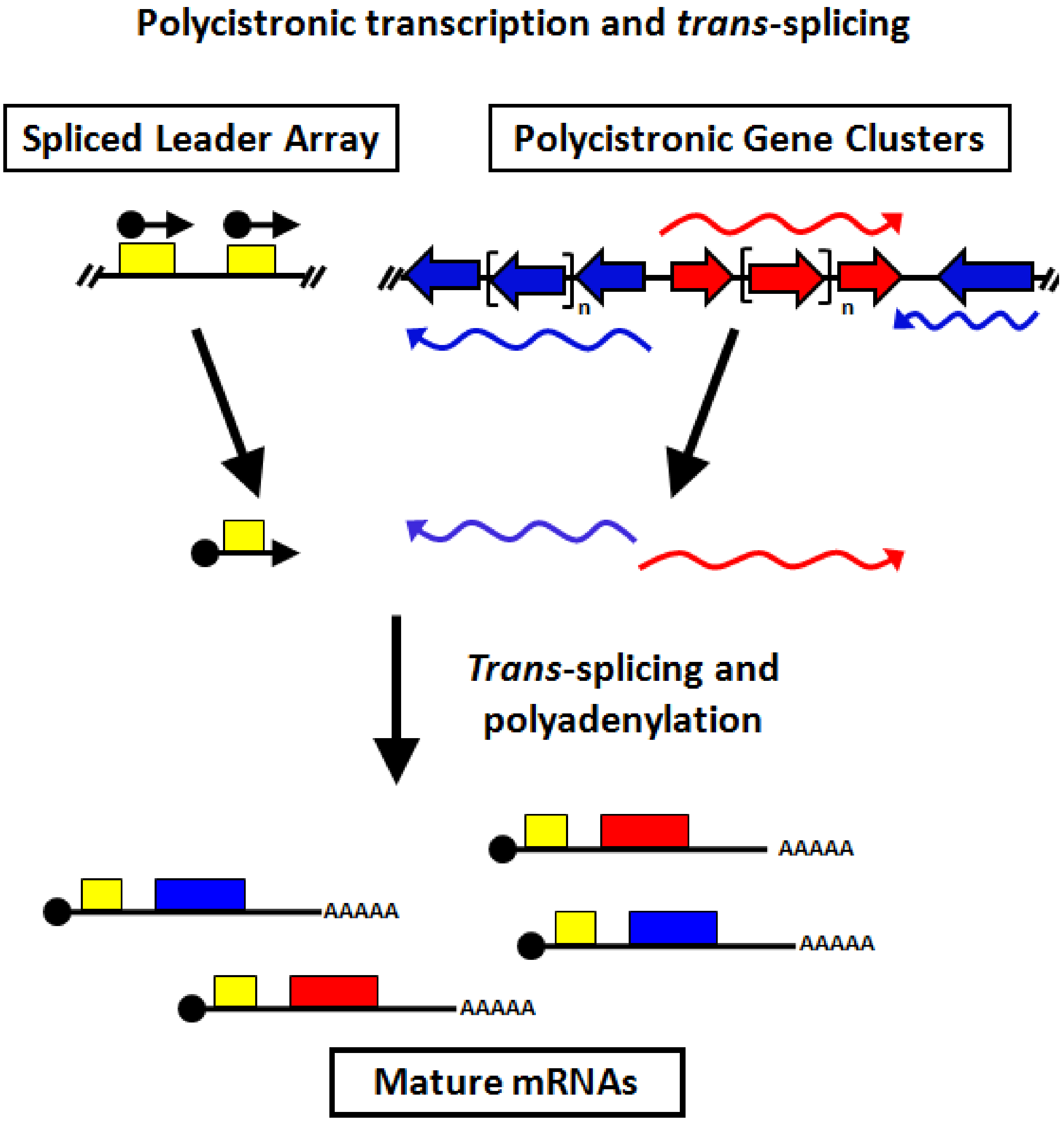
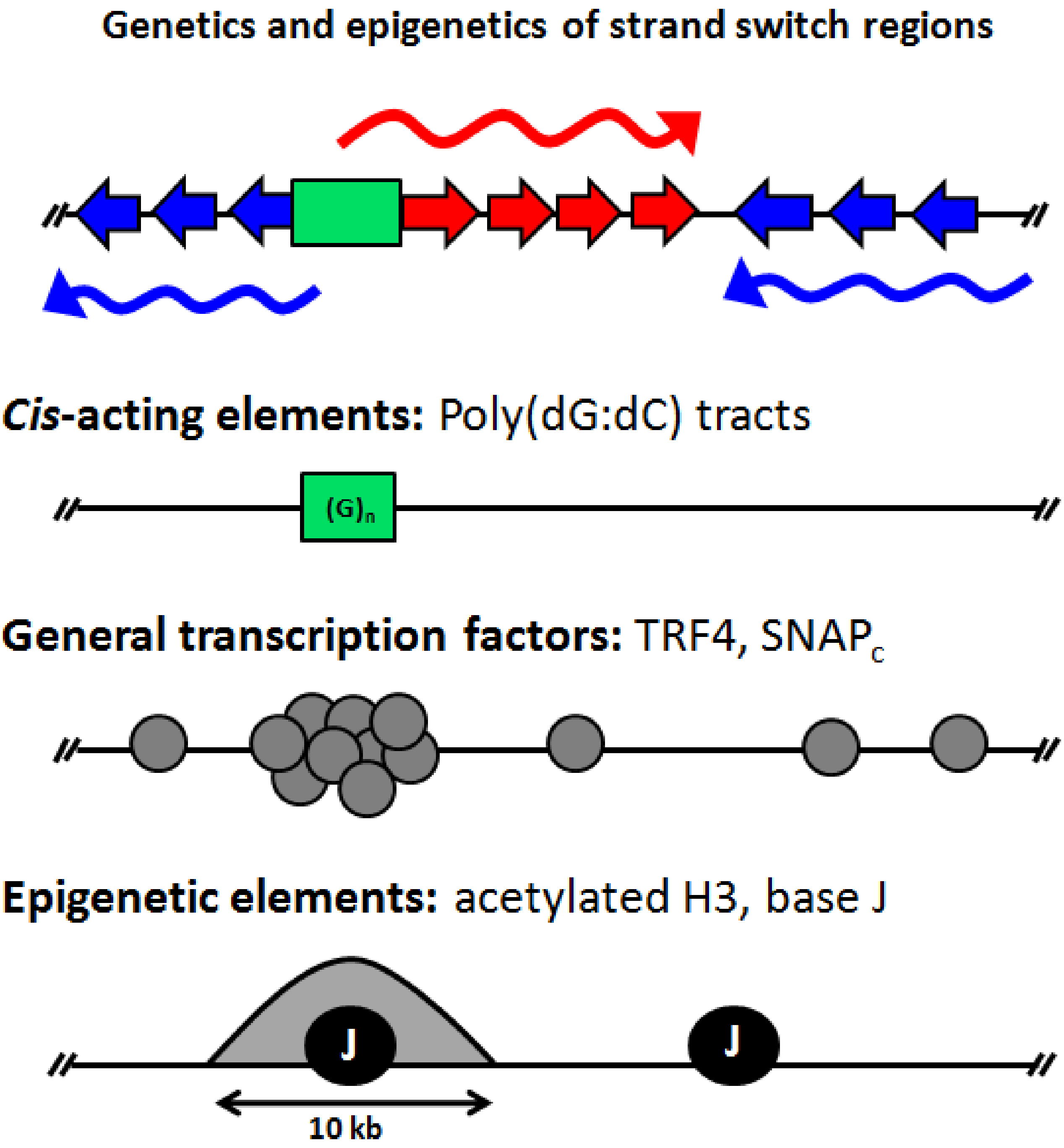

Polycistronic transcription and *trans*-splicing
Spliced Leader Array
Polycistronic Gene Clusters
n
n
Trans-splicing and polyadenylation
AAAAA
AAAAA
AAAAA
AAAAA
Mature mRNAs

Figure 1-3.



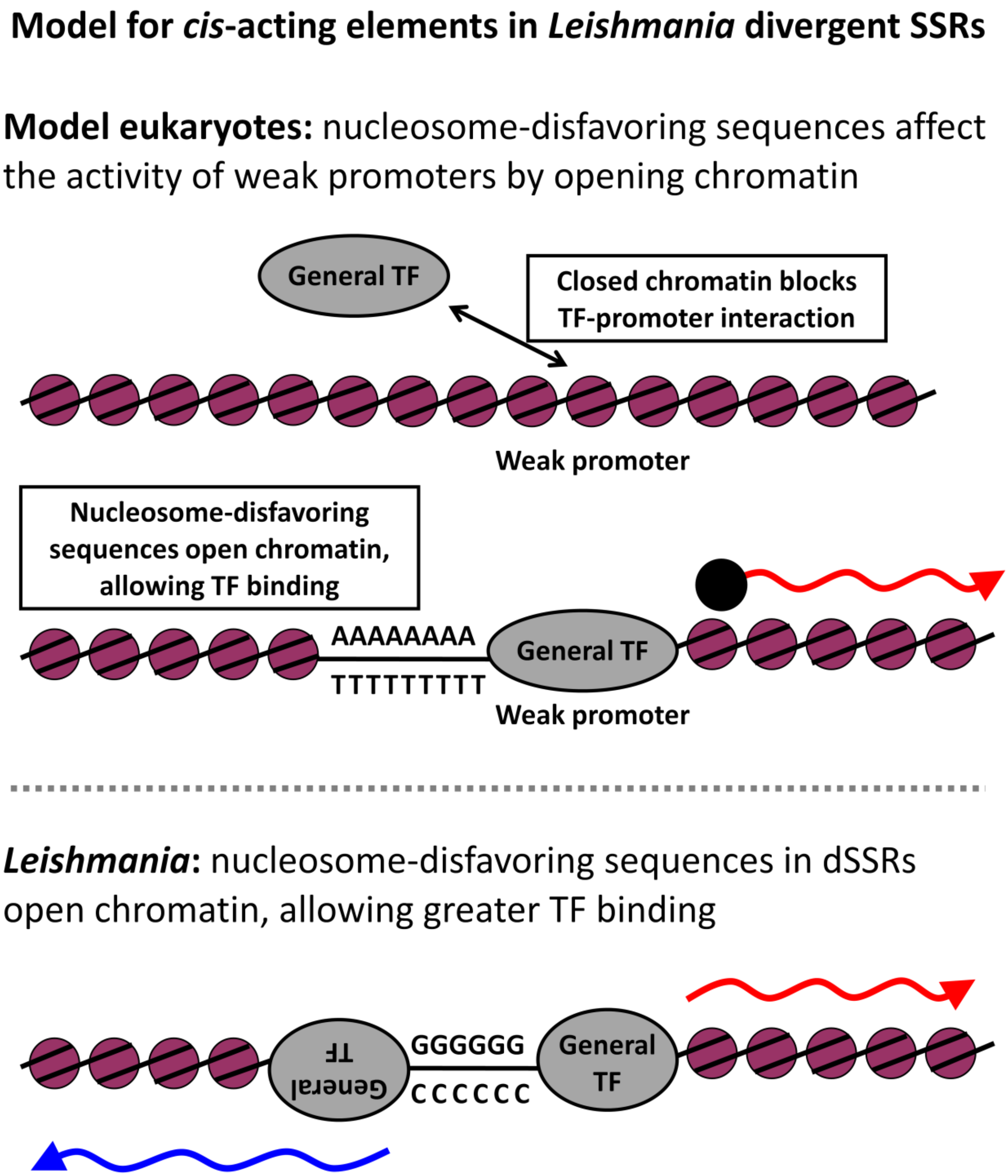
Genetics and epigenetics of strand switch regions

***Cis*-acting elements:** Poly(dG:dC) tracts

General transcription factors: TRF4, SNAP$_c$

Epigenetic elements: acetylated H3, base J

Figure 1-4.



Model for *cis*-acting elements in *Leishmania* divergent SSRs

Model eukaryotes: nucleosome-disfavoring sequences affect the activity of weak promoters by opening chromatin

General TF
Closed chromatin blocks TF-promoter interaction
Weak promoter
Nucleosome-disfavoring sequences open chromatin, allowing TF binding
AAAAAAAA
TTTTTTTTT
General TF
Weak promoter

***Leishmania*:** nucleosome-disfavoring sequences in dSSRs open chromatin, allowing greater TF binding

General TF
GGGGGG
CCCCCC
General TF

Table 1-1.

Gene Product	**Function**	**Gene ID in *L. major***
H2A	**Core histone**	LmjF21.0915 LmjF21.0920 LmjF21.0930 LmjF29.1720 LmjF29.1730 LmjF29.1740
H2A.X	Histone variant involved in DNA damage signaling; is phosphorylated and recruited to double-strand DNA breaks	**None**
H2A.Z	Histone variant involved in transcription initiation; localizes near transcription start sites	LmjF.17.0280
H2B	**Core histone**	LmjF19.0030 LmjF19.0040 LmjF19.0050 LmjF17.1220 LmjF09 1340
H2B.V	Trypanosomatid-specific histone variant, localizes near transcription start sites; present in nucleosomes containing H2A.Z	LmjF28.0210
H3	**Core histone**	LmjF10.0870 LmjF10.0990 LmjF16.0600 LmjF16.0610
H3.3	Localizes near transcription start sites in other eukaryotes; present in nucleosomes containing H2A.Z	**None**
Cenp-A (H3-like)	Centromeric histone variant	**None**
H3.V	Trypanosomatid-specific histone variant, localizes to transcription termination sites in *T. brucei*	LmjF19.0630
H4	**Core histone**	LmjF02.0020 LmjF06.0010 LmjF15.0010 LmjF25.2450 LmjF31.3180 LmjF35.1310 LmjF36.0020
H4.V	Trypanosomatid-specific histone variant, localizes to transcription termination sites in *T. brucei*	**None, but H4 genes very divergent**

Table 1-2.

Gene Product	**Function**	**Gene ID in *L. major***
HAT1	MYST-family histone acetyltransferase; associated with chromosome segregation and telomeric silencing in *T. brucei* (107)	**LmjF14.0140**
HAT2	MYST-family histone acetyltransferase; is essential and catalyzes H4K10Ac in *T. brucei* (107)	**LmjF28.2270**
HAT3	MYST-family histone acetyltransferase; catalyzes H4K4 acetylation in *T. brucei* but is not essential (106)	**LmjF36.6990**
HAT4	MYST-family histone acetyltransferase; ortholog present in *T. cruzi* but not *T. brucei* (63)	**LmjF13.0170**
ELP3.1	Elongator-type histone acetyltransferase	**LmjF16.0240**
ELP3.2	Elongator-type histone acetyltransferase; negatively regulates transcription of rRNA locus in *T. brucei* (151)	**LmjF23.1350**
HDAC1	Class I (nuclear, zinc-dependent) histone deacetylase; essential in *T. brucei* (152)	**LmjF21.0680**
HDAC2'	Class I (nuclear, zinc-dependent) histone deacetylase; specific to *Leishmania* and is orthologous to HDAC1, not HDAC2 in *T. brucei* and *T. cruzi*	**LmjF24.1370**
HDAC3	Class II (nuclear and cytoplasmic, zinc-dependent) histone deacetylase; is essential in *T. brucei* (152)	**LmjF21.1870**
HDAC4	Class II (nuclear and cytoplasmic, zinc-dependent) histone deacetylase; is not essential, but is important for normal cell cycling (152)	**LmjF08.1090**
SIR2RP1	Class III (NAD+-dependent) deacetylase; is not essential but is involved in DNA repair and RNA pol I repression near telomeres in *T. brucei* (153)	**LmjF26.0210**
SIR2RP2	Class III (NAD+-dependent) deacetylase; is mitochondrial in *T. brucei* (153)	**LmjF23.1210**
SIR2RP3	Class III (NAD+-dependent) deacetylase; is mitochondrial in *T. brucei* (153)	**LmjF34.2140**
DOT1A	SET-domain histone demethylase; di-methylates H3K76 and is essential for viability in *T. brucei* (103) and in *L. major* (Wong and Beverley, in preparation); regulates cell cycle progression (104)	**LmjF07.0025**
DOT1B	SET-domain histone demethylase; tri-methylates H3K76 but is not essential for viability in *T. brucei* (103,105) or *L. major* (Wong and Beverley, in preparation; involved in VSG silencing in *T. brucei* (105)	**LmjF20.0030**
MT3	SET-domain, DOT1-like methyltransferase	**LmjF33.1790**
SET1	Multi-SET domain methyltransferase	**LmjF35.4550**
SET2	Single SET domain methyltransferase	**LmjF21.1750**
SET3	Contains SET and post-SET domains important for zinc binding	**LmjF36.0210**
RMT1	Protein arginine methyltransferase (Type I, asymmetric dimethylation)	**LmjF12.1270**

RMT2	Protein arginine methyltransferase	**LmjF03.0600**
RMT3 (PRMT6)	Protein arginine methyltransferase (Type I, asymmetric dimethylation); associated with histones, flagellar proteins, and nuclear pore in *T. brucei* (154)	**LmjF16.0030**
RMT4 (PRMT7)	Protein arginine methyltransferase (Type III, monomethylation); is cytoplasmic in *T. brucei* (155)	**LmjF06.0870**
JHDM1	Jumonji-C domain histone demethylase	**LmjF35.2940**
JHDM2	Jumonji-C domain histone demethylase	**LmjF31.0240**
JHDM3	Jumonji-C domain histone demethylase	**LmjF30.1190**
JHDM4	Jumonji-C domain histone demethylase	**LmjF27.1320**
JHDM5	Jumonji-C domain histone demethylase	**LmjF27.1150**
JHDM6	Jumonji-C domain histone demethylase	**LmjF26.1290**
BDF1	Bromodomain-containing protein	**LmjF36.6880**
BDF2	Bromodomain-containing protein; binds to acetylated H4K10 and acetylated H2A in *T. cruzi* (156); binds at VSG expression sites and across polycistronic gene clusters in *T. brucei* (Schulz and Papavasiliou, KMCB Meeting 2013)	**LmjF36.2980**
BDF3	Bromodomain-containing protein; is essential and localizes to divergent SSRs in *T. brucei* (65)	**LmjF36.3360**
BDF4	Bromodomain-containing protein; localizes to divergent SSRs in *T. brucei* (Schulz and Papavasiliou, KMCB Meeting 2013); *L. major* ortholog is much longer and has zinc-finger motif	**LmjF14.0360**

Chapter Two

Kinetoplastid-specific histone variant functions are conserved in *Leishmania major*

Preface

BA designed and performed the majority of these experiments, analyzed RNA-seq data, generated the figures, and wrote the manuscript. ILKW generated and validated antisera. LB generated RNA-seq libraries, and GR performed the RNA-seq read alignments. SMB supervised these studies, assisted with the design and analysis of these experiments, and assisted in writing and editing the manuscript. PJM supervised RNA-seq studies and contributed to editing the manuscript.

This work has been previously published in a condensed form and is reused with permission from Elsevier (license #3442631422582):

Anderson, BA, Wong, ILK, Baugh, L, Ramasamy, G, Myler, PJ, and Beverley, SM. Kinetoplastid-specific histone variant functions are conserved in *Leishmania major*. Mol Biochem Parasitol 2013; 191(2): 53-7.

Abstract

Protein-coding genes in kinetoplastid protists are transcribed from polycistronic arrays, yielding RNA precursors that are processed to form mature transcripts bearing a 5’ spliced leader (SL) and 3’ poly(A) tract. Regions of transcription initiation and termination lack known eukaryotic promoter and terminator elements, and current data suggest that transcription is instead controlled predominantly through epigenetic mechanisms. Several epigenetic marks, including histone modifications, histone variants, and an atypical DNA modification known as base J have been localized to transcription initiation or termination regions in *Trypanosoma brucei*, *Trypanosoma cruzi*, and/or *Leishmania major*. Despite this conservation, the phenotypes of base J mutants vary significantly across trypanosomatids, suggesting that the specific epigenetic networks governing transcription initiation and termination have diverged significantly during evolution. In this light, we sought to characterize and compare the roles of the histone variants H2A.Z, H2B.V, and H3.V in *L. major.* As in *T. brucei*, the histone variants H2A.Z and H2B.V were shown to be essential in *L. major* using a powerful quantitative plasmid segregation-based test. In contrast and again similar to *T. brucei*, H3.V is not essential in *Leishmania* as *H3V*-null lines grew normally, resembled WT, and remained infectious. Using spliced leader (SL)-primed RNA-seq, we found that H3.V-null parasites have steady-state transcript levels comparable to WT parasites and display no defects in the efficiency of transcription termination at convergent strand switch regions (SSRs). Our results show a genetic conservation of histone variant phenotypes between *L. major* and *T. brucei*, in contrast to the diversity of phenotypes associated with genetic manipulation of the epigenetic DNA base J modification.

Keywords: histone variants, transcriptional read-through, chromatin

Introduction

The generation of mature messenger RNAs (mRNAs) in *Leishmania* and other kinetoplastid parasites involves a bipartite mechanism of transcription by RNA polymerase II (RNAP II), unlike the majority of eukaryotes studied to date. All protein-coding genes are transcribed as pre-mRNAs arising from long head-to-tail arrays called polycistronic gene clusters (PGCs), while the RNAs encoding the capped 5' ends of mature transcripts are transcribed separately from the spliced leader (SL) RNA array (reviewed in (1,2). Polycistronic pre-mRNAs are then processed by 5' *trans*-splicing of the SL RNA to generate the capped 5' end of the mRNA and cleavage and polyadenylylation to generate the 3' end. Notably, polyadenylylation of the upstream transcript is coupled to *trans*-splicing of the downstream transcript (3,4). In this system, individual transcription units are mostly defined by the boundaries of PGCs: transcription primarily initiates within divergent strand-switch regions (dSSRs), where two PGCs are oriented head-to-head, and terminates in convergent strand-switch regions (cSSRs), where two PGCs meet tail-to-tail. These regions lack known eukaryotic promoter and terminator elements (5–7), and trypanosomatid genomes reveal the presence of general but not sequence-specific RNAP II transcription factors (8**).**

Eukaryotic transcription is heavily regulated by chromatin-associated epigenetic factors including histone variants and reversible covalent modification of histones and DNA (reviewed in (9,10)). By organizing transcriptionally permissive or repressive chromatin environments, these heritable epigenetic factors can regulate transcription genome-wide through global alterations in epigenetic patterns or provide more complex local regulation at individual loci. A number of epigenetic marks have been identified in trypanosomatids, including histone variants, histone modifications, and the trypanosomatid-specific DNA modification β-D-glucopyranosyloxymethyluracil (base J), many of which have been mapped to dSSRs or cSSRs in one or more trypanosomatid species (5,11–15). However, transcription termination and re-initiation may also potentially occur within a PGC, since these chromatin signatures have been found within PGCs (5,11,16). In addition to marking regions of transcriptional initiation or termination, epigenetic mechanisms may also play a role in other aspects of transcription. One example may be global transcriptional regulation in *L. major* promastigotes, where histone H3 acetylation levels decline greatly in stationary phase (11), a time when total RNA levels and transcription decline [Akopyants and Beverley, unpublished results].

Despite the apparent conservation of epigenetic marks and their genomic localization amongst trypanosomatid species, recent data suggest that their functions may differ greatly. This is most clearly seen in studies of DNA base J, perturbations of which show widely varying consequences in the three lineages examined thus far. In *T. brucei*, *T. cruzi*, *L. tarentolae*, and *L. major*, base J been localized to convergent and divergent SSRs as well as telomeres, including the inactive subtelomeric variant surface glycoprotein genes in *T. brucei* (12–14,17). In *T. brucei*, deletion of the genes encoding the thymidine hydroxylases JBP1 and JBP2, which catalyze the first step in base J biosynthesis, generates viable parasites lacking J with no other observable phenotypes or changes in gene expression (18). In *T. cruzi*, the *JBP1-/JBP2-* double null mutant was not viable, while individual *JBP1-* or *JBP2* mutants showed altered transcriptional rates and polymerase occupancy near dSSRs, but normal transcription termination at the cSSRs examined (14). In contrast, in *Leishmania tarentolae*, *JBP1* is essential (19), and *JBP2*-null mutants showed massive transcriptional read-through at cSSRs in addition to increased antisense transcription and use of alternative transcription start sites in dSSRs (12).

The evolutionary diversity evident from base J perturbations prompted us to ask whether other epigenetic marks might show functional divergence as well. In *T. brucei*, chromatin immunoprecipitations studies localized H2A.Z and H2B.V to dSSRs and H3.V and H4.V to cSSRs (5), implicating these proteins in regulation of transcription initiation and termination, respectively. Genetic studies in a variety of organisms have confirmed the vital role of histone variants, and H2A.Z is conserved in most eukaryotes studied to date (reviewed in (20)). Both *H2AZ* and *H2BV* are essential in *T. brucei* , while *H3V* and *H4V* are not (5). Here we focus on the histone variants of *L. major* and explore the functional consequences of their genetic inactivation on viability and transcription. In anticipation that one or more histone variants would be essential in *Leishmania* as well, we employed a recently developed definitive test which relies on segregational loss of an episomal complementation vector (21). First, a positive/negative GFP-expressing episomal vector (pXNG) expressing the test gene is introduced into a wild type (WT) or heterozygous line, followed by generation of chromosomal-null mutants. Removal of selection for the complementation vector allows cells to lose the plasmid during subsequent rounds of cell division, should the test gene not be essential. Loss of the plasmid can readily be visualized by flow cytometry (GFP expression) and selected for by sorting GFP-negative cells or using the Herpes simplex virus thymidine kinase (TK) (22). Importantly, this technique separates the test of gene function from the relatively inefficient process of transfection and allelic replacement and allows for screening high number of events rapidly. This improves the chances of isolating null mutants whose fitness may be compromised and mutants from loci where homologous recombination is less efficient; furthermore, when null mutants are not obtained, one has a higher confidence in conclusions concerning the essentiality of the gene of interest (21,23).

By this powerful test, we show that both *H2AZ* and *H2BV* are essential in *L. major*. Thus the requirement of *Leishmania* for these histone variants closely resembles that seen in *T. brucei*. In contrast to *H2AZ* and *H2BV*, we were readily able to delete *H3V*, and these null mutants remained phenotypically normal with little alteration in transcriptional patterns. In this regard, loss of the H3.V 'termination' mark differed greatly from the loss of the DNA base J mark reported previously in *L. tarentolae*. Therefore, while the epigenetic mark base J has divergent functions in different kinetoplastids, we have shown that histone variants likely have conserved roles in these organisms.

Results

H2A.Z and H2B.V are essential

To probe the roles of H2A.Z and H2B.V in transcription in *Leishmania,* we attempted unsuccessfully to generate H2A.Z- and H2B.V-null *L. major* promastigotes using successive homologous allelic replacement methods (24). While a sign that these genes are essential, the ability of *Leishmania* to undergo aneuploidy with high frequency (25,26) and concerns about negative results arising from complex targeting protocols prompted us to employ more rigorous tests, specifically an episome segregation approach as described in the introduction (21). We generated heterozygote lines bearing an episomal complementation vector expressing either *H2AZ* or *H2BV* along with GFP (*H2AZ/HYG [pXNG-H2AZ]* and *H2BV/SAT [pXNG-H2BV]*). In the presence of the episomal gene, it was now possible to remove the second chromosomal allele, yielding the chromosomal-null lines *Δh2az[pXNG-H2AZ]* and *Δh2bv [pXNG-H2BV]*. Typically the finding of 'replaceable in the presence of complementation' has been taken as *a priori* evidence of essentiality; however, we have shown recently that for some loci this is misleading as the episome can subsequently be lost, suggesting that the failure to recover chromosomal-null parasites by the classic route arose from some other cause [23, Guo and Beverley, unpublished data].

To carry out the episome segregation tests, the episome-bearing lines were grown without selection for two culture passages (approximately 12 cell doublings) to permit loss of the episomal complementation vector. We observed that only 0.2% of the *Δh2az [pXNG-H2AZ]* were GFP-dim compared to the 33.7% of the *H2AZ/HYG [pXNG-H2AZ]* heterozygote line, potentially heralding that this gene is essential (Fig. 1A). Similar results were obtained for *Δh2bv [pXNG-H2BV]* (0.1% GFP-dim) and *H2BV/PAC [pXNG-H2BV]* (10.4% GFP-dim) (Fig. 2A). Single cells from both the chromosomal-null and heterozygote lines were sorted into multiple 96-well microtiter plates on the basis of GFP fluorescence, focusing on GFP-dim cells which had potentially lost the complementation vector, or as a control, GFP-bright cells which had retained it. These plates were then incubated with media until robust growth was seen in control wells.

Sorting of the parental heterozygous lines bearing the episome (*H2AZ/HYG [pXNG-H2AZ]* and *H2BV/SAT [pXNG-H2BV]*) yielded growth in 70-80% of wells for both the GFP-dim and GFP-bright populations (Fig. 1B and 2B, respectively); this provides a basal measure of cell survival and growth following sorting. As expected, all of the GFP-bright clones retained episomes containing the streptothricin acetyltranferase (*SAT*) or hygromycin B phosphotransferase (*HYG*) markers, while most (80-100%) cells arising from the GFP-dim populations completely lost the episome and became sensitive to the selective antibiotics (Fig. 2B, 2D). Sorting of the GFP-bright cells from the *Δh2az [pXNG-H2AZ]* and *Δh2bv [pXNG-H2BV]* populations yielded 70-85% growth, comparable to that of the heterozygous control populations. In contrast, only a small fraction (0.6-0.8%) of the GFP-dim cells showed growth following sorting; none of these cells had lost the episome bearing *H2AZ* or *H2BV* as judged by retention of the selectable marker from the episome (Fig. 2B, 2D). Previous studies showed that these cells most likely arose from imperfect sorting or recovery of cells bearing low episome copy numbers (21). From the plating efficiency and numbers of wells tested, we estimated that approximately 610 events were scored in this assay for *H2AZ* and 740 events for *H2BV*, many more than typically screened by traditional non-segregational methods. Thus, we conclude from these experiments that both *H2AZ* and *H2BV* are essential in *L. major*.

Loss of H3.V does not affect viability or differentiation

In contrast to H2A.Z and H2B.V, we were able to delete both *H3V* alleles by the standard method of two rounds of allelic replacement, yielding homozygous null mutants (*Δh3v*). Colonies were readily obtained from the second round of allelic replacement, and out of six colonies screened five had lost the *H3V* gene. This was shown by the presence of the planned replacements as revealed by PCR using primers flanking and internal to the targeting fragment (data not shown), the absence of the *H3V* ORF by PCR using primers within the *H3V* ORF (Fig. 3A) and absence of H3.V protein by Western blotting with H3.V-specific antisera (Fig. 3B). Complemented lines were generated by transfection of an *H3V*-containing episome and showed restoration of H3.V protein levels to levels comparable to WT (Fig. 3B). Since typically episomes are present in multiple copies leading to overexpression of encoded genes, these data suggest the possibility that H3.V levels are regulated at the protein level. Analysis of several clonal *Δh3v* lines showed that they were phenotypically normal, showing WT growth *in vitro* (Fig. 3C). Although we observed a significant increase in the fraction of metacyclic parasites in both *Δh3v* clones, a similar increase was observed in the complementing lines and thus is unrelated to loss of H3.V expression (Fig. 3D). To identify defects in parasite virulence we inoculated BALB/c mice in the footpad with 10^7 stationary phase parasites from six independent clones. All lines generated lesions within one month, similar to WT *L. major* (data not shown). Together these data demonstrate that deletion of *H3V* does not alter viability of *L. major* in the promastigote stage or significantly impair the infectivity of amastigote stages in murine infections.

Loss of H3.V does not affect transcription termination or steady-state transcript levels

To elucidate potential roles for H3.V in transcription in *Leishmania*, we analyzed mRNA levels in WT and *Δh3v* parasites by high-throughput sequencing of spliced leader (SL)-primed cDNA libraries. This method quantifies steady-state RNA levels in a population of cells by specifically amplifying only transcripts with an SL sequence at their 5' end (12,16,27,28). Importantly, studies in *L. tarentolae* demonstrate that this approach is also a sensitive method for detecting read-through transcription arising from defects in transcription termination, as these abnormal RNAs can give rise to stable RNAs after processing using cryptic splice acceptor sites (see Fig. 2 in reference (12) for an example of this arising from base J deficiency in *L. tarentolae*). The sensitivity of detection of both normal 'sense' and cryptic 'antisense' splice acceptors is very high, with ranges in the hundreds of reads per million reads mapped for both 'normal' and 'cryptic' splice acceptors (12).

We focus first on transcriptional read-through, a hallmark of defects in transcriptional termination. As in previous studies in *L. tarentolae* and *T. brucei* (12,16,28), in WT *L. major* the vast majority of SL-containing reads map to the coding strand, with very few mapping to antisense regions beyond cSSRs (see Fig. 4 A-B, Supplemental Figs. S2, S3). Remarkably, this pattern was unchanged in *Δh3v* parasites across the parasite genome (Supplemental Fig. S2). This included 'simple' cSSRs (Fig. 4 A, B show two representative examples), cSSRs containing RNA polymerase III-transcribed genes (which are known to suppress transcriptional read-through in the absence of base J in *L. tarentolae* (12); Supplemental Fig. S3A,B), or the single cSSR known to lack base J in *L. major* (located on chromosome 28; Supplemental Fig. S3C). Quantitative measurement of transcriptional read-through (the antisense-to-sense ratio of reads mapping within 10 kb of a cSSR) shows a very similar distribution in the WT, heterozygous, and *Δh3v* lines (Fig. 4C). These findings are in stark contrast to the results seen in *L. tarentolae* by SL-primed RNA sequencing, where perturbations of base J synthesis in *JBP2dKO* parasites led to high levels of transcriptional read-through (12).

Lastly, we compared mRNA levels by plotting the normalized number of reads mapping to the sense strand of individual *L. major* genes for WT against *Δh3v* parasites. Again, normalized sense transcript levels were remarkably similar between WT and *Δh3v* parasites, with R^2 values >0.96 for two independent *Δh3v* clonal transfectants (Fig. 4D). Examination of all genes containing at least 50 mapped reads in the WT and/or *Δh3v* datasets showed that only two genes showed greater than two-fold differences, occurring in both independent *Δh3v* clonal lines. These genes are located in the middle of PGCs and would appear unlikely candidates to be unaffected by any potential alterations in regulation of transcription initiation or termination. The P27 protein (encoded by *Lmj28.0980*), a component of the cytochrome c oxidase complex (29,30), was up-regulated 2.3-fold in both *Δh3v* lines tested relative to WT. In addition, a protein tyrosine phosphatase 1-like protein (*LmjF36.2180*) was up-regulated 2.2-2.3-fold in these lines. This protein has not been characterized to date in *Leishmania* but is an important regulator of cell differentiation in *T. brucei* (31). Given the absence of detectable phenotypes in *Δh3v* mutants, the significance of these small changes or whether they even result in changes in protein levels is uncertain. Together, these data suggest that H3.V is not required for defining transcriptional stops in *Leishmania* and likely does not play a critical role in controlling mRNA abundance.

Discussion

Epigenetic regulation by histone variant incorporation and reversible covalent modification of histones and/or DNA is a common thread in eukaryotic transcription, acting to both broadly regulate global transcription and to fine-tune transcription of specific genes. In kinetoplastid protists, which lack sequence-specific transcription factors (8), epigenetic control may be the primary source of transcriptional regulation, and a growing body of work shows that many epigenetic marks localize to sites of transcription initiation and termination (5,11–15). While it is often standard practice to translate the functional aspects of epigenetic networks from one system to another based on localization patterns, studies of the hypermodified DNA base J in kinetoplastids demonstrates that assumptions of conserved function based on conserved localization patterns may be incorrect (12–14,32). In this light, we characterized three histone variants in *Leishmania*: H2A.Z and H2B.V, which have been localized to dSSRs in *T. brucei*, and H3.V, which was localized to cSSRs (5).

In our survey of histone variants in *Leishmania*, we found that *H2AZ* and *H2BV* were essential (much like in *T. brucei*), suggesting that their functions are likely conserved. Given their genomic distribution in *T. brucei* and the high degree of H2A.Z conservation among all eukaryotes, we suspect that these proteins are integral components of the epigenetic networks controlling transcription initiation. However, the dSSR-associated epigenetic network could differ in *T. cruzi*, as base J mutants show a transcription initiation-related phenotype (14,32) that is not replicated in base J mutants in *T. brucei* (13) or *L. tarentolae* (12). In such a case, H2A.Z and/or H2B.V may play functionally different roles which could differ significantly from eukaryotes studied to date. Elucidation of the effects of H2A.Z/H2B.V incorporation on chromatin compaction and characterization of histone variant incorporation during parasite differentiation will allow us to more specifically define the roles of these proteins in kinetoplastids.

In contrast to H2A.Z and H2B.V, we found that *H3V*-null *L. major* were viable, morphologically normal, and infectious; moreover, they behaved as WT parasites with respect to transcriptional regulation (Fig. 4D) and most interestingly, transcription termination (Fig. 4A, B; Supplemental Fig. S2, S3). These data, when interpreted in the light of recent work demonstrating the deleterious effects of perturbation of transcription termination-associated epigenetic networks in *Leishmania* mediated by base J (12), suggests that H3.V is not an essential component of this epigenetic network. Transcription termination-associated phenotypes were not examined in *H3V*-null *T. brucei* (33) and no data exists regarding the essentiality of this protein in *T. cruzi*, so it remains unclear whether this protein is functioning redundantly with other components in the epigenetic network of these parasite species. Although H3.V may not be a critical component of the transcription termination-associated epigenetic networks, chromatin-based studies of H3.V mutants may further define the roles of this conserved, kinetoplastid-specific histone variant.

Materials and Methods

Parasite growth

All studies used derivatives of *Leishmania major* Friedlin V1 (MHOM/JL/81/Friedlin), grown at 26°C in M199 medium (US Biologicals) supplemented with 40 mM 4-(2-hydroxyethyl)-1-piperazineethanesuphonic acid (HEPES) pH 7.4 (Fisher Scientific), 100 uM adenine (Sigma), 1 µg mL^{-1} biotin (Sigma), 10 µg mL^{-1} hemin (Sigma), 2 µg mL^{-1} biopterin (Schircks Laboratories), 50 units/mL penicillin (Gibco), 50 µg/mL streptomycin (Gibco), and 10% (v/v) heat inactivated fetal calf serum (HyClone). Cell density was determined by using a model Z1 Coulter counter (logarithmic phase) or hemocytometer (stationary phase). Metacyclic promastigotes were purified from stationary phase day 4 cultures using density gradient centrifugation (34). Semisolid M199 medium was prepared using supplemented M199 medium with 1% (w/v) Difco noble agar (BD Diagnostic Systems).

Generation of recombinant proteins and antisera

The N-terminal 47 amino acids of *H3V* (LmjF.19.0620) and the N-terminal 33 amino acids of *H3* were amplified using primers described in Supplemental Table S1. The PCR products were digested with BamHI and NdeI and inserted into BamHI- and NdeI-digested into pET-16B to generate the protein expression vectors pET-16B-H3-N (B5994) and pET-16B-H3V-N (B5995). Constructs were confirmed by restriction digestion and sequencing.

B5994 and B5995 were transformed into BL21(DE3) pLysS cells (Invitrogen), and H3-N and H3V-N protein expression were induced using 1 mM IPTG and incubating cells at 37°C with agitation for 5 hours. Cells were lysed by sonication and centrifuged, and the cell pellet was solubilized using 8M urea, pH 8.0 (Fisher). Proteins were purified with the Ni-NTA purification system using denaturing conditions (Invitrogen). Polyclonal antisera were raised against the N-termini of H3 and H3.V using a commercial service (Proteintech). Two rabbits were injected with each antigen, and the primary injections were followed with boosts at 28, 42, 60, and 78 days; pre-immune sera were collected as well as sera after the immunization program was completed. Specificity of antisera was validated using immunogens and the acid-soluble fraction extracted from *Leishmania* chromatin (Supplemental Fig. 1).

Western blotting

Logarithmic phase promastigotes (~2x10^6 cells/mL) were collected, resuspended at a concentration of 4x10^8 cells/mL in Laemmli buffer [10% glycerol (Sigma), 2% sodium dodecyl sulfate (Sigma), 63 mM Tris-HCl pH 6.8 (Fisher Scientific), 0.1% 2-mercaptoethanol (Sigma), and 0.0005% bromophenol blue (Bio-Rad)], and boiled for 10 minutes. Total lysates from 8 x 10^6 cells were resolved by SDS-PAGE, electroblotted onto Hybond-ECL nylon membranes (Amersham Biosciences), and blocked with Odyssey blocking buffer (Li-Cor). Primary incubations were performed using 1:500 anti- H3V-N or 1:5,000 anti-H3-N in Odyssey blocking buffer. Secondary incubations were performed with 1:10000 IR680-labeled goat anti-rabbit antibody (Li-Cor) and blots were analyzed and quantified using the Odyssey imaging system (Li-Cor).

Generation of constructs for targeted deletion and episomal complementation

Cassettes for targeted deletion of *H2AZ* (LmjF.17.0280), *H2BV* (LmjF28.0210), and *H3V* (LmjF.19.0620) were generated by fusion PCR (35) using primers described in Supplemental Table S1. Briefly, 500-1000 bp 5’ of the ORF were amplified using primers containing the fusion sequence GGTAACGGTGCGGGCTGACG at the 3’ end, and 500-1000 bp 3’ of the ORF were amplified using primers containing the fusion sequence

CGAGATCCCACGTAAGGTGC at the 5’ end. Drug resistance marker sequences were amplified using primers to introduce complementary fusion sequences at the 5’ and 3’ ends and the sequence CCACC directly upstream of the marker ORF. Amplicons were purified by gel extraction (Qiagen), and deletion cassettes were assembled in a second PCR containing the 5’ and 3’ sequences and the drug resistance marker ORF. The resulting cassettes were cloned into pGEM-T Easy to generate the constructs pGEM-H2AZ-HYG (B6623), pGEM-H2AZ-BSD (B6624), pGEM-H2BV-PAC (B6569), pGEM-H2BV-SAT (B6572), pGEM-H3V-HYG (B6570), and pGEM-H3V-BSD (B6571). All constructs were confirmed by restriction enzyme digestion and sequencing. Deletion cassettes were released by restriction enzyme digestion (*H2AZ,* XmaI; *H2BV*, BglII; *H3V*, BamHI) and treated with calf intestinal phosphatase (New England Biolabs) to minimize re-ligation of transfected fragments. All deletion cassettes were gel purified before transfection.

The ORFs for *H2AZ*, *H2BV*, and *H3V* were amplified using the primers described in Supplemental Table S2. BglII-digested PCR products were cloned directly into BglII-digested pXNG4-SAT (B5840) or pXNG4-HYG (B6559) [described in (21)] to generate the episomal complementation constructs pXNG-H2AZ-SAT (B6651), pXNG-H2BV-HYG (B6657), and pXNG-H3V-SAT (B6652). Constructs were confirmed by restriction enzyme digestion and sequencing.

Generation of chromosomal-null cell lines

Linearized *H2AZ-HYG* and *H2BV-PAC* targeting fragments were transfected separately into WT *L. major* FV1 promastigotes as described (36). Heterozygous clones *H2AZ/Δh2az::HYG* (*H2AZ/HYG*) and *H2BV/Δh2bv::PAC* (*H2BV/HYG*) were isolated by plating on semisolid supplemented M199 medium containing 50 µg mL^{-1} hygromycin (Calbiochem) or 30 µg mL^{-1} puromycin (Sigma), respectively. The presence of the *HYG* or *PAC* genes were confirmed by PCR using the primers described in Supplemental Table S2, and allelic replacements were confirmed by Southern blotting. *H2AZ/HYG* clone 4 and *H2BV/ PAC* clone 5 were transfected with the respective episomal complementation constructs pXNG-H2AZ-SAT or pXNG-H2BV-HYG to generate the lines *H2AZ/Δh2az::HYG[pXNG-H2AZ]*, referred to as *H2AZ/HYG[pXNG-H2AZ]* and *H2BV/Δh2bv::PAC[pXNG-H2BV]*, referred to as *H2BV/PAC[pXNG-H2BV]*. *H2AZ/HYG [pXNG-H2AZ]* and *H2BV/PAC [pXNG-H2BV]* clones were isolated by plating on semisolid supplemented M199 medium containing 25 µg mL^{-1} hygromycin and 100 µg mL^{-1} nourseouthricin (Werner BioAgents) or 15 µg mL^{-1} puromycin and 50 µg mL^{-1} hygromycin, respectively. The presence of the episomal complementation construct was demonstrated by GFP expression. *H2AZ/HYG [pXNG-H2AZ]* clone 22 and *H2BV/PAC [pXNG-H2BV]* clone 52 were transfected with the linearized targeting fragments *H2AZ-BSD* or *H2BV-SAT*. Chromosomal-null *Δh2az::HYG/Δh2az::BSD[pXNG-H2AZ]* and *Δh2bv::PAC/Δh2bv::SAT[pXNG-H2BV]* clones, referred to as *Δh2az[pXNG-H2AZ]* and *Δh2bv [pXNG-H2BV]*, were selected by plating on semisolid supplemented M199 containing 25 µg mL^{-1} hygromycin, 50 µg mL^{-1} nourseouthricin, and 10 µg mL^{-1} blasticidin (Fisher) or 15 µg mL^{-1} puromycin, 25 µg mL^{-1} hygromycin, and 100 µg mL^{-1} nourseothricin, respectively. The presence of the expected resistance markers were confirmed by PCR using primers described in Supplemental Table S2, and integration of the replacement cassettes and loss of the chromosomal alleles were confirmed by Southern blotting. *Δh2az[pXNG-H2AZ]* clone 11 and *Δh2bv [pXNG-H2BV]* clone 521 were used for all experiments shown.

The linearized *H3V-HYG* targeting fragment was transfected into WT *L. major* FV1 as previously described and *H3V/Δh3v::HYG* (*H3V/HYG*) heterozygotes were isolated by plating on semisolid supplemented M199 containing 30 µg mL^{-1} hygromycin. Presence of the *HYG* gene and integration of the targeting fragment were confirmed by PCR using primers described in Supplemental Table S2. *H3V/HYG* clone 7 was transfected with the linearized *H3V-BSD* targeting fragment and *Δh3v::HYG/Δh3v::BSD* (*Δh3v*) clones were isolated by plating on semisolid supplemented M199 containing 15 µg mL^{-1} hygromycin and 10 µg mL^{-1} blasticidin. Integration of the targeting fragments and the loss of the *H3V* allele were confirmed by PCR using primers described in Supplemental Table S2**.** *Δh3v* clones 3 and 4 were used for all experiments shown. To generate complemented chromosomal-null lines, *Δh3v* clone 4 was transfected with pXNG-H3V-SAT as previously described. Complemented clones were isolated by plating on semisolid supplemented M199 containing 5 µg/mL blasticidin, 25 µg/mL hygromycin, and 100 µg/mL nourseothricin. The presence of the episomal complementation vector was confirmed by GFP expression and restoration of H3.V protein expression.

Single cell sorting

Δ*h2az [pXNG-H2AZ]* clone 11 and *Δh2bv [pXNG-H2BV]* clone 521 and their immediate parental lines *H2AZ/HYG [pXNG-H2AZ]* clone 22 and *H2BV/PAC [pXNG-H2BV]* clone 52 were grown for two cell passages in supplemented M199 medium in the absence of all selective drugs. Logarithmic-phase cells were collected, resuspended in phosphate-buffered saline, and filtered through CellTrics 50 µm filters (Partec) to remove clumps. A Dako MoFlo high-speed cell sorter was used to sort and recover single cells based on their GFP fluorescence. Gates for the GFP-dim populations were set using WT *L. major* FV1, and gates for the GFP-bright population were set using *H2AZ/HYG [pXNG-H2AZ]* clone 22. Single cells were recovered into individual wells of a 96-well plate containing 150 µL Schneider's medium (Sigma) supplemented with 100 uM adenine, 10 µg mL^{-1} hemin, 2 µg mL^{-1} biopterin, 50 units/mL penicillin, 50 µg/mL streptomycin, and 10% (v/v) heat inactivated fetal calf serum; supplemented Schneider's medium was used as this was found empirically to increase the recovery in control test sorts. Plates were incubated at 26°C for 2 weeks and parasite growth was scored. Positive wells were screened for presence of the *PAC*, *HYG*, *SAT*, and *BSD* drug resistance markers by growing lines in M199 containing 30 µg mL^{-1} puromycin, 50 µg mL^{-1} hygromycin, 100 µg mL^{-1} nourseothricin, or 10 µg mL^{-1} blasticidin. Lines lacking the expected drug resistance markers associated with allelic replacement (i.e. *BSD* and *HYG* for Δ*h2az [pXNG-H2AZ]*, *PAC* and *SAT* for *Δh2bv [pXNG-H2BV]*) were excluded from further analysis, as cells lacking the appropriate allelic replacement markers represent contamination from the WT cells used for gate setting or from parental lines used in previous sorts and do not represent candidate null mutants.

Spliced leader (SL) RNA-primed sequencing

Logarithmically-growing promastigotes from WT *L. major* FV1, one *H3V/HYG* transfectant (clonal line 7), and two *Δh3v* transfectants (clonal lines 3 and 4) were collected and resuspended at a concentration of 5x10^{8} cells/mL in TriZOL (Invitrogen). The aqueous phase was separated by addition of 0.2 mL chloroform (Fisher Scientific) and centrifugation at 12,000 x *g* for 15 minutes at 4°C. The aqueous phase was isolated and RNA was precipitated by adding 1 volume of 100% isopropanol (Fisher Scientific) and centrifugation at 12,000 x *g* for 10 minutes at 4°C. The RNA pellet was washed with 75% ethanol (Pharmco) and was resuspended in nuclease-free water (Ambion). Purified RNA was treated with 20 units of DNAse I (Ambion) and was precipitated using 1/10 volumes 3M sodium acetate (Sigma) and 3 volumes 100% ethanol. RNA was pelleted by centrifugation at 15,000 g^{-1} and the resulting pellet was washed with cold 75% ethanol and resuspended in RNAse-free water (Ambion). SL RNA-primed libraries for Illumina sequencing were prepared, sequenced, and analyzed as previously described (27).

Read-through transcription analysis

Transcription termination sites (TTS) within cSSRs were defined using base J localization data in combination with genome annotations. Briefly, peaks were called from base J immunoprecipitation data from WT *L. major* Friedlin (GEO Accession GSE23976, sample GSM816864) (12) using MACS using the default parameters (37). Peaks overlapping with convergent SSRs were extracted using BEDTools intersectBed (38). Transcription termination sites were defined as the midpoint of the base J peak within the cSSR, and strand-specific windows were generated encompassing 10 kb upstream of the TTS (sense) or 10 kb downstream of the TTS (antisense/read-through). Strand-specific coverage was obtained using BedTools coverageBed, and the antisense:sense ratio was calculated using custom Unix scripts. Strand-specific coverage plots normalized to the total number of reads aligned were generated using BEDTools genomeCoverageBed, specifying the –scale and –bga (bedgraph) outputs. Positive and negative strand coverage plots were merged and formatted using custom Unix scripts. Data was viewed using Integrative Genomics Viewer (39,40).

Acknowledgments

We thank the Alvin J. Siteman Cancer Center at Washington University School of Medicine and Barnes-Jewish Hospital in St. Louis, Mo., for the use of the Siteman Flow Cytometry Core, which provided the single-cell sorting service. The Siteman Cancer Center is supported in part by NCI Cancer Center Support Grant #P30 CA91842. This work was funded by the National Institutes of Health (R0129646 to SMB and ILKW, T32 GM007067 to BAA) and the Washington University Berg-Morse and Schlesinger Graduate Fellowships to BAA. We thank Andrew Haydock and members of our laboratories for discussions.

Figure Legends

Figure 2-1. *H2AZ* is essential in *L. major*. (A) Quantitation of pXNG-H2AZ levels by GFP flow cytometry following removal of nourseothricin selection. The dark gray shaded regions represent GFP fluorescence of the experimental lines, and the dotted line shows GFP fluorescence values from WT *L. major* FV1. Light gray shaded regions represent FACS gates used for recovery of 'GFP-dim' (left shaded region) and 'GFP-bright' (right shaded region) cells; parasites with a GFP fluorescence signal of 1 or less were not included in the 'GFP-dim' gate. The lines *H2AZ/HYG[pXNG-H2AZ]* (left panel) and *Δh2az[pXNG-H2AZ]* (right panel) were grown for 48 hours (~12 cell doublings) in the absence of nourseothricin to allow loss of the episome before GFP fluorescence was analyzed. Boxes show percent of parasites classified as 'GFP-bright' or 'GFP-dim'. (B) Single cells from 'GFP-dim' and 'GFP-bright' *H2AZ/HYG[pXNG-H2AZ]* and *Δh2az[pXNG-H2AZ]* were sorted into 96-well plates containing supplemented Schneiders' medium (see Materials and Methods). Boxes show the percentage of wells scored for robust growth after two weeks of incubation at 26°C; numbers in parentheses represent the total number of cells sorted (total from two independent experiments). For these, retention of pXNG-H2AZ

was tested by growth in the presence of nourseothricin, conferred by the plasmid *SAT* marker. Boxes show the percentage of cells demonstrating nourseothricin resistance; numbers in parentheses represent the total number of wells subjected to nourseothricin resistance testing.

Figure 2-2. H2BV is essential in L. major. (A) Quantitation of pXNG(HYG)-H2BV levels by GFP flow cytometry following removal of hygromycin selection. GFP fluorescence panels and boxes are defined as in Fig. 1A. *H2BV/PAC[pXNG-H2BV]* (left panel) and *Δh2bv[pXNG-H2BV]* (right panel) cells were grown for 48 hours (~12 cell doublings) in the absence of hygromycin to allow loss of the episome. (B) Single cells from ‘GFP-dim’ and ‘GFP-bright’ *H2BV/PAC[pXNG-H2BV]* and *Δh2bv[pXNG-H2BV]* were sorted and scored as described in Fig. 1. Boxes are defined as in Fig. 1B. Plasmid retention was tested using hygromycin resistance of cells from robustly growing wells and is presented as described in Fig. 1B.

Figure 2-3. Deletion of *H3V* in *L. major* does not alter growth or metacyclogenesis. (A) Deletion of *H3V* was shown by PCR analysis using *H3V* ORF primers located as depicted in the upper figure in this panel. (B) Loss of H3.V expression shown by western blotting using anti-H3.V antisera. The migration position of H3.V is shown, as is a nonspecific band evident in all samples. The nonspecific band does not arise from cross-reactivity with H3 (Fig. S1). (C) WT and *Δh3v* mutants grow comparably *in vitro.* (D) Metacyclogenesis was quantitated after 3 d in stationary phase using the density gradient method (34). Error bars represent standard deviation of three biological replicates.

Figure 2-4. Deletion of *H3V* does not increase read-through transcription as observed by SL-RNA-seq. (A-B) Integrative Genomics Viewer (39,40) screenshots demonstrating SL-RNA-seq coverage across ‘simple’ cSSRs. The Y-axes represent normalized read counts (per million reads mapped) and the X-axis represents physical location on each chromosome; a 20 kb window showing 10 kb flanking the transcription termination site (TTS) is shown (A: Chromosome 4, 118,903-138,903 bp; B: Chromosome 7, 49,636-69,636 bp). Unlike random RNA-seq reads, SL-RNA-seq results in clustering of reads on a limited number of splice acceptor sites (regardless of whether they are ‘normal’ or ‘cryptic’ (12)). (C) Quantitative analysis of transcription termination assessed by SL-RNA-seq. Following previous studies (12), TTS within cSSRs were defined using the midpoint of base J ‘peaks’ associated with TTS; reads mapping to the ‘sense’ and ‘antisense’ strand within 10 kb of the TTS were quantitated and the ratio of antisense to sense reads is shown by a box plot. The middle line represents the median, while the box represents the 25th through 75th percentiles. Whiskers represent the 10th through 90th percentiles, and dots represent individual cSSRs which lie below the 10th or above the 90th percentile. (D) Total mRNA levels quantitated by SL-RNA-seq are unchanged in *Δh3v* parasites. Read counts were normalized to the median number of reads mapped to each gene (see (27) for methods used). The X- and Y-axes shows sense strand read counts for genes from WT and *Δh3v* clone 3, respectively. The solid line shows the slope (1) expected for no changes in transcript levels; dotted lines represent 2-fold higher and lower boundaries. The correlation coefficient (R^2) comparing WT and *Δh3v* clone 3 was 0.9646. Comparable results were obtained in comparisons of WT with *H3V/HYG* (R^2 = 0.9905) or *Δh3v* clone 4 (R^2 = 0.9918).

Supplemental Data

Supplemental Table S2-1. Primer sequences used for generation of deletion constructs for *H2AZ*, *H2BV*, and *H3V* and demonstration of *Δh3v* planned replacements. Restriction sites are underlined, and fusion sequences are in boldface.

Supplemental Table S2-2. Primer sequences used for amplification of histone variant ORFs for protein expression and episomal complementation vectors. Restriction sites are underlined.

Supplemental Figure S2-1. Demonstration of anti-H3 (A, B) and anti-H3.V specificity by western blotting. (A, C) Antisera were tested using the recombinant proteins used as immunogens (A, H3-N; C, H3.V-N). (B, D) Antisera were tested against a purified acid-soluble fraction from *L. major* chromatin (B, anti-H3-N; D, anti-H3.V-N). The migration of molecular weight markers is shown; the expected MW are 14.6 kDa for H3 and 16.3 kDa for H3.V.

Supplemental Figure S2-2. Transcription termination is unaltered in *Δh3v* parasites. (A-L) IGV screenshots are shown for *L. major* chromosomes 1-3 (A), 4-6 (B), 7-9 (C), 10-12 (D), 13-15 (E), 16-18 (F), 19-21 (G), 22-24 (H), 25-27 (I), 28-30 (J), 31-33 (K), and 34-36 (L), displaying SL-RNA-seq mappings as described in Fig. 4. Y-axes represent normalized read counts (per million reads mapped) and are scaled to 1000 reads per million reads mapped, and X-axes represent physical location on the chromosome. Unlike random RNA-seq reads, SL-RNA-seq results in clustering of reads on a limited number of splice acceptor sites.

Supplemental Figure S2-3. Transcription termination is unaltered in *Δh3v* parasites, regardless of whether tRNAs are present in the cSSR. (A-B) IGV screenshots demonstrating SL-RNA-seq coverage across convergent SSRs containing one (A) or multiple (B) RNA polymerase III-transcribed genes. Y-axes represent normalized read counts (per million reads mapped) and X-axes represents physical location on the chromosome; 20 kb windows are shown as described in Fig. 4. (C) IGV screenshot demonstrating SL-RNA-seq coverage across the sole cSSR lacking base J in *L. major*, located on chromosome 28 (12). Despite the absence of both base J and H3.V, transcription termination is not altered.

References

1. Günzl A. The pre-mRNA splicing machinery of trypanosomes: complex or simplified? Eukaryot Cell 2010;9(8):1159–70.

2. Gunzl A, Vanhamme L, Myler P. Transcription in trypanosomes: a different means to the end. In: Barry J, Mottram J, McCulloch R, Acosta-Serrano A, editors. Trypanosomes - After the Genome. Horizon Bioscience, Wymondham, Norfolk, UK; 2007. p. 177–208.

3. LeBowitz J, Smith H, Rusche L, Beverley S. Coupling of poly(A) site selection and *trans*-splicing in *Leishmania*. Genes Dev 1993;7(6):996–1007.

4. Matthews K, Tschudi C, Ullu E. A common pyrimidine-rich motif governs *trans*-splicing and polyadenylation of tubulin polycistronic pre-mRNA in trypanosomes. Genes Dev 1994;8(4):491–501.

5. Siegel T, Hekstra D, Kemp L, Figueiredo L, Lowell J, Fenyo D, et al. Four histone variants mark the boundaries of polycistronic transcription units in *Trypanosoma brucei*. Genes Dev 2009;23(9):1063–76.

6. Martínez-Calvillo S, Nguyen D, Stuart K, Myler PJ. Transcription initiation and termination on *Leishmania major* chromosome 3. Eukaryot Cell 2004;3(2):506–17.

7. Martínez-Calvillo S, Yan S, Nguyen D, Fox M, Stuart K, Myler P. Transcription of *Leishmania major* Friedlin chromosome 1 initiates in both directions within a single region. Mol Cell 2003;11(5):1291–9.

8. Ivens A, Peacock C, Worthey E, Murphy L, Aggarwal G, Berriman M, et al. The genome of the kinetoplastid parasite, *Leishmania major*. Science 2005;309(5733):436–42.

9. Bell O, Tiwari V, Thomä N, Schübeler D. Determinants and dynamics of genome accessibility. Nat Rev Genet 2011;12(8):554–64.

10. Martinez-Calvillo S, Vizuet-de-Rueda J, Florencio-Martinez L, Manning-Cela R, Figueroa-Angelo E. Gene expression in trypanosomatid parasites. J Biomed Biotechnol 2010;2010:525241.

11. Thomas S, Green A, Sturm N, Campbell D, Myler P. Histone acetylations mark origins of polycistronic transcription in *Leishmania major*. BMC Genomics 2009;10:152.

12. Van Luenen H, Farris C, Jan S, Genest P, Tripathi P, Velds A, et al. Glucosylated hydroxymethyluracil, DNA base J, prevents transcriptional readthrough in *Leishmania*. Cell 2012;150(5):909–21.

13. Cliffe L, Siegel T, Marshall M, Cross G, Sabatini R. Two thymidine hydroxylases differentially regulate the formation of glucosylated DNA at regions flanking polymerase II polycistronic transcription units throughout the genome of *Trypanosoma brucei*. Nucleic Acids Res 2010;38(12):3923–35.

14. Ekanayake D, Sabatini R. Epigenetic regulation of polymerase II transcription initiation in *Trypanosoma cruzi*: modulation of nucleosome abundance, histone modification, and polymerase occupancy by O-linked thymine DNA glucosylation. Eukaryot Cell 2011;10(11):1465–72.

15. Wright J, Siegel T, Cross G. Histone H3 trimethylated at lysine 4 is enriched at probable transcription start sites in *Trypanosoma brucei*. Mol Biochem Parasitol 2010;172(2):141–4.

16. Kolev N, Franklin J, Carmi S, Shi H, Michaeli S, Tschudi C. The transcriptome of the human pathogen *Trypanosoma brucei* at single-nucleotide resolution. PLoS Pathog 2010;6(9):e1001090.

17. Van Leeuwen F, Wijsman E, Kieft R, Van Der Marel G, Van Boom J, Borst P. Localization of the modified base J in telomeric VSG gene expression sites of *Trypanosoma brucei*. Genes Dev 1997;11(23):3232–41.

18. Cliffe L, Kieft R, Southern T, Birkeland S, Marshall M, Sweeney K, et al. JBP1 and JBP2 are two distinct thymidine hydroxylases involved in J biosynthesis in genomic DNA of African trypanosomes. Nucleic Acids Res 2009;37(5):1452–62.

19. Genest P, ter Riet B, Dumas C, Papadopoulou B, van Luenen H, Borst P. Formation of linear inverted repeat amplicons following targeting of an essential gene in *Leishmania*. Nucleic Acids Res 2005;33(5):1699–709.

20. Malik H, Henikoff S. Phylogenomics of the nucleosome. Nat Struct Biol 2003;10(11):882–91.

21. Murta S, Vickers T, Scott D, Beverley S. Methylene tetrahydrofolate dehydrogenase/cyclohydrolase and the synthesis of 10-CHO-THF are essential in *Leishmania major*. Mol Microbiol 2009;71(6):1386–401.

22. LeBowitz J, Cruz A, Beverley S. Thymidine kinase as a negative selectable marker in *Leishmania major*. Mol Biochem Parasitol 1992;51:321–5.

23. Feng X, Rodriguez-Contreras D, Polley T, Lye L, Scott D, Burchmore R, et al. "Transient" genetic suppression facilitates generation of hexose transporter null mutants in *Leishmania mexicana*. Mol Microbiol 2013;87(2):412–29.

24. Cruz A, Coburn C, Beverley S. Double targeted gene replacement for creating null mutants. Proc Natl Acad Sci USA 1991;88(16):7170–4.

25. Cruz A, Titus R, Beverley S. Plasticity in chromosome number and testing of essential genes in *Leishmania* by targeting. Proc Natl Acad Sci USA 1993;90(4):1599–603.

26. Sterkers Y, Lachaud L, Crobu L, Bastien P, Pagès M. FISH analysis reveals aneuploidy and continual generation of chromosomal mosaicism in *Leishmania major*. Cell Microbiol 2011;13(2):274–83.

27. Mittra B, Cortez M, Haydock A, Ramasamy G, PJ M, Andrews N. Iron uptake controls the generation of *Leishmania* infective forms through regulation of ROS levels. J Exp Med 2013;210(2):401–16.

28. Nilsson D, Gunasekera K, Mani J, Osteras M, Farinelli L, Baerlocher L, et al. Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of *Trypanosoma brucei*. PLoS Pathog 2010;6(8):e1001037.

29. Dey R, Dagur P, Selvapandiyan A, McCoy J, Salotra P, Duncan R, et al. Live attenuated *Leishmania donovani* p27 gene knockout parasites are nonpathogenic and elicit long-term protective immunity in BALB/c mice. J Immunol 2013;190(5):2138–49.

30. Dey R, Meneses C, Salotra P, Kamhawi S, Nakhasi H, Duncan R. Characterization of a *Leishmania* stage-specific mitochondrial membrane protein that enhances the activity of cytochrome c oxidase and its role in virulence. Mol Microbiol 2010;77(2):399–414.

31. Szöőr B, Wilson J, McElhinney H, Tabernero L, Matthews K. Protein tyrosine phosphatase TbPTP1: a molecular switch controlling life cycle differentiation in trypanosomes. J Cell Biol 2006;175(2):293–303.

32. Ekanayake D, Minning T, Weatherly B, Gunasekera K, Nilsson D, Tarleton R, et al. Epigenetic regulation of transcription and virulence in *Trypanosoma cruzi* by O-linked thymine glucosylation of DNA. Mol Cell Biol 2011;31(8):1690–700.

33. Lowell J, Cross G. A variant histone H3 is enriched at telomeres in *Trypanosoma brucei*. J Cell Sci 2004;117(24):5937–47.

34. Späth G, Beverley S. A lipophosphoglycan-independent method for isolation of infective *Leishmania* metacyclic promastigotes by density gradient centrifugation. Exp Parasitol 2001;99(2):97–103.

35. Kuwayama H, Obara S, Morio T, Katoh M, Urushihara H, Tanaka Y. PCR-mediated generation of a gene disruption construct without the use of DNA ligase and plasmid vectors. Nucleic Acids Res 2002;30(2):e2.

36. Robinson K, Beverley S. Improvements in transfection efficiency and tests of RNA interference (RNAi) approaches in the protozoan parasite *Leishmania*. Mol Biochem Parasitol 2003;128(2):217–28.

37. Zhang Y, Liu T, Meyer C, Eeckhoute J, Johnson D, Bernstein B, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol 2008;9(9):R137.

38. Quinlan A, Hall I. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26(6):841–2.

39. Robinson J, Thorvaldsdottir H, Winckler W, Guttman M, Lander E, Getz G, et al. Integrative Genomics Viewer. Nat Biotechnol 2011;29(1):24–6.

40. Thorvaldsdottir H, Robinson J, Mesirov J. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 2013;14(2):178–92.

Figure 2-1. *H2AZ* is essential in *L. major.*

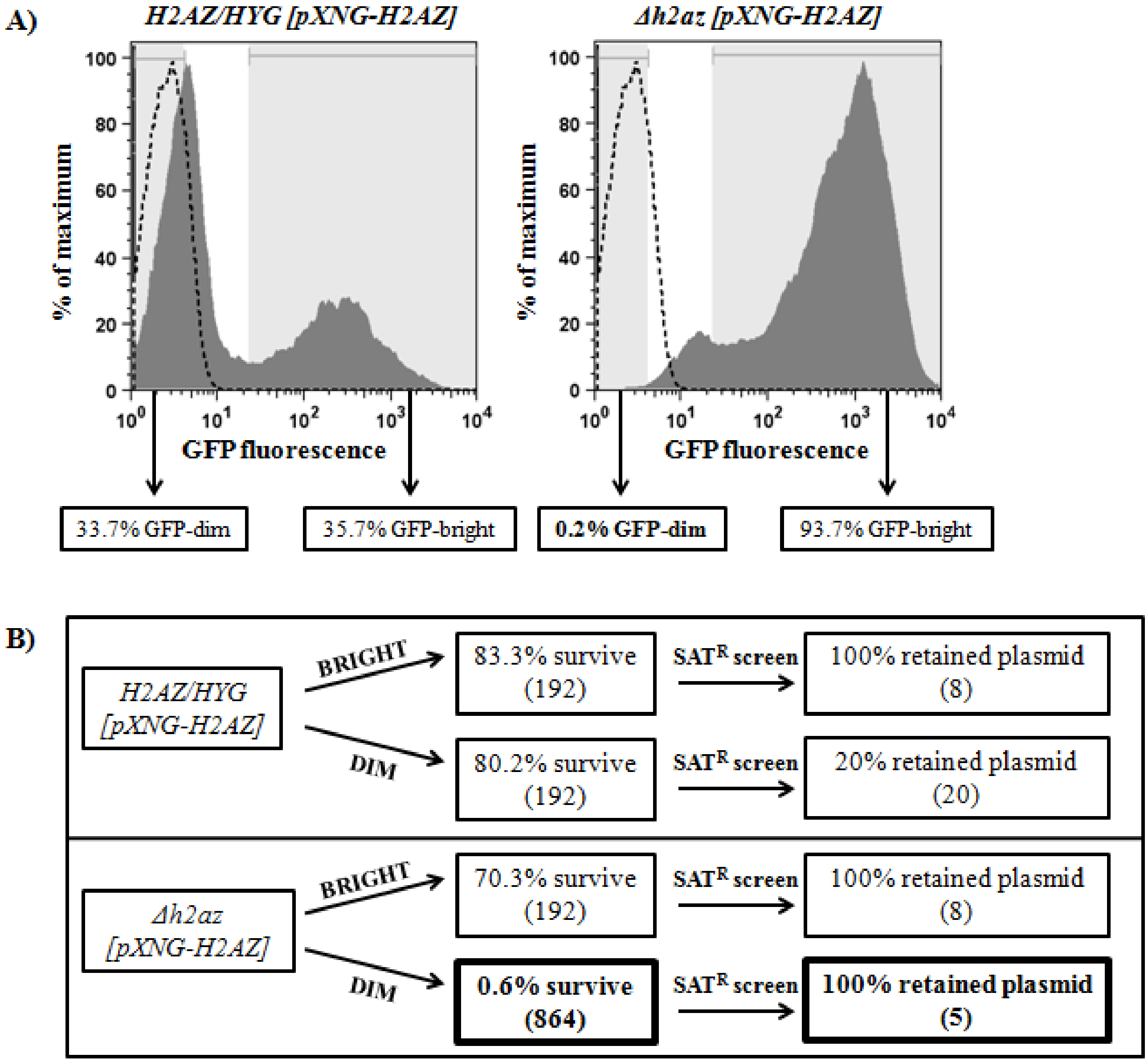


Figure 2-2. *H2BV* is essential in *L. major*.

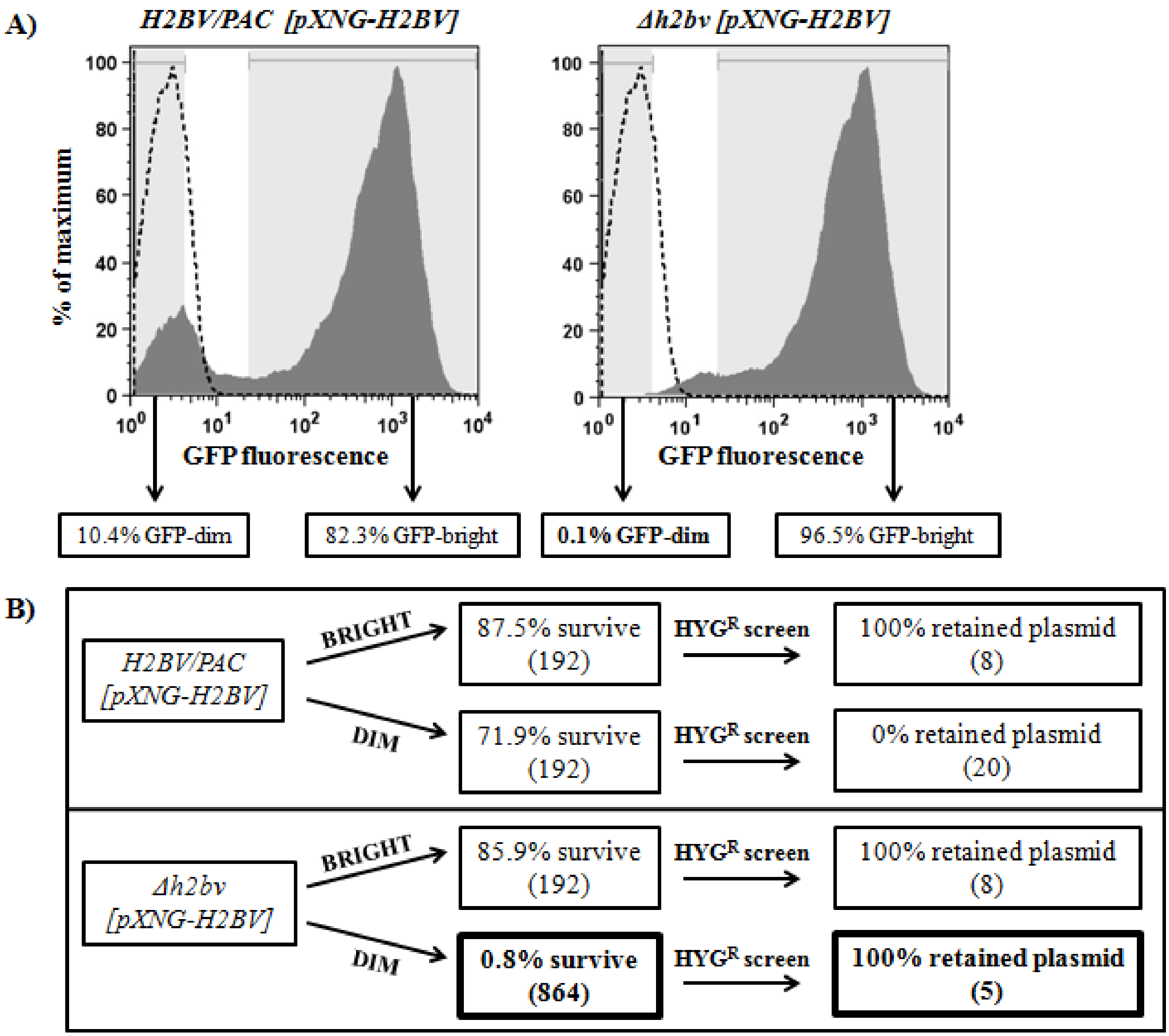


Figure 2-3. Deletion of *H3V* in *L. major* does not alter growth or metacyclogenesis.

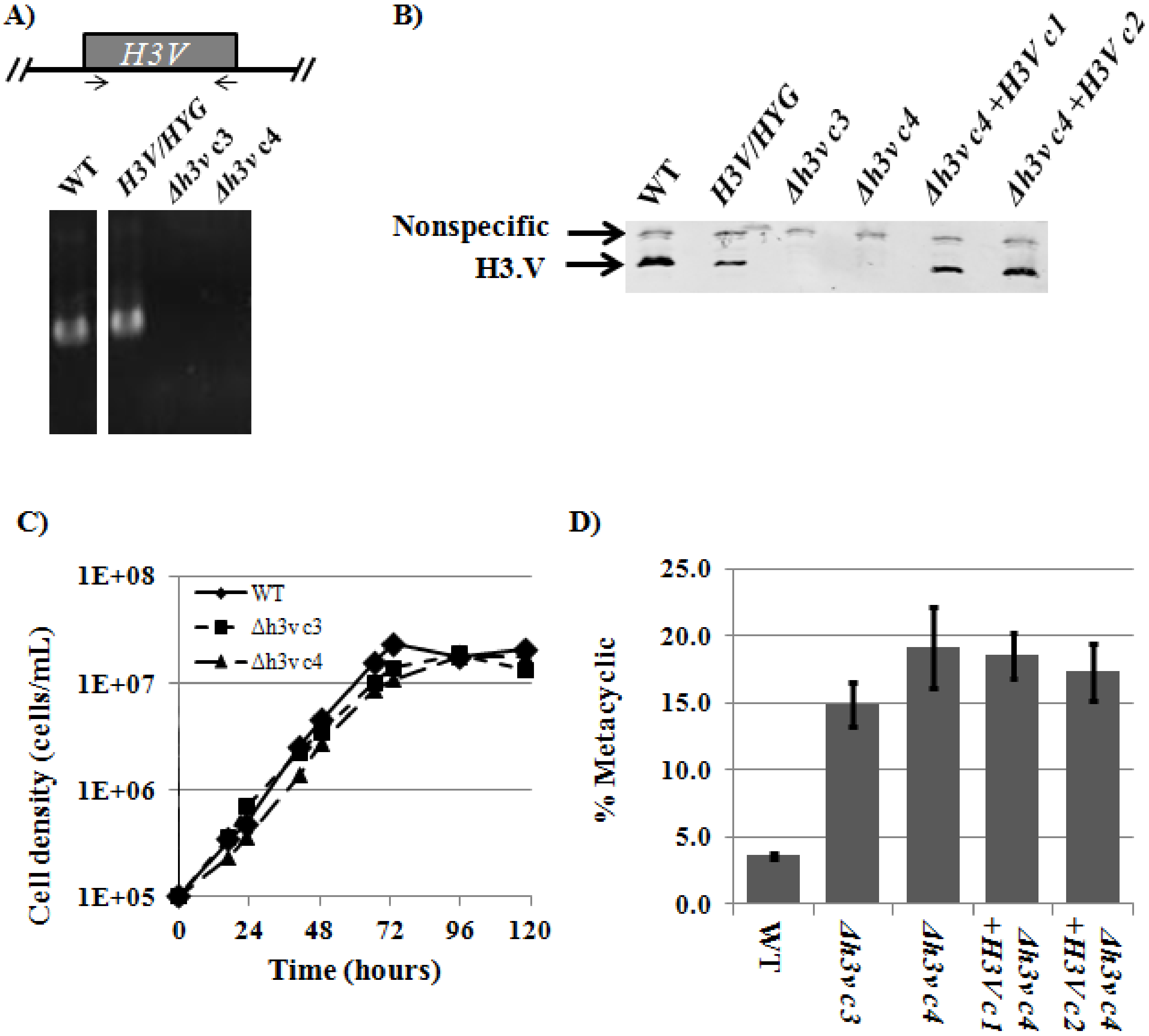
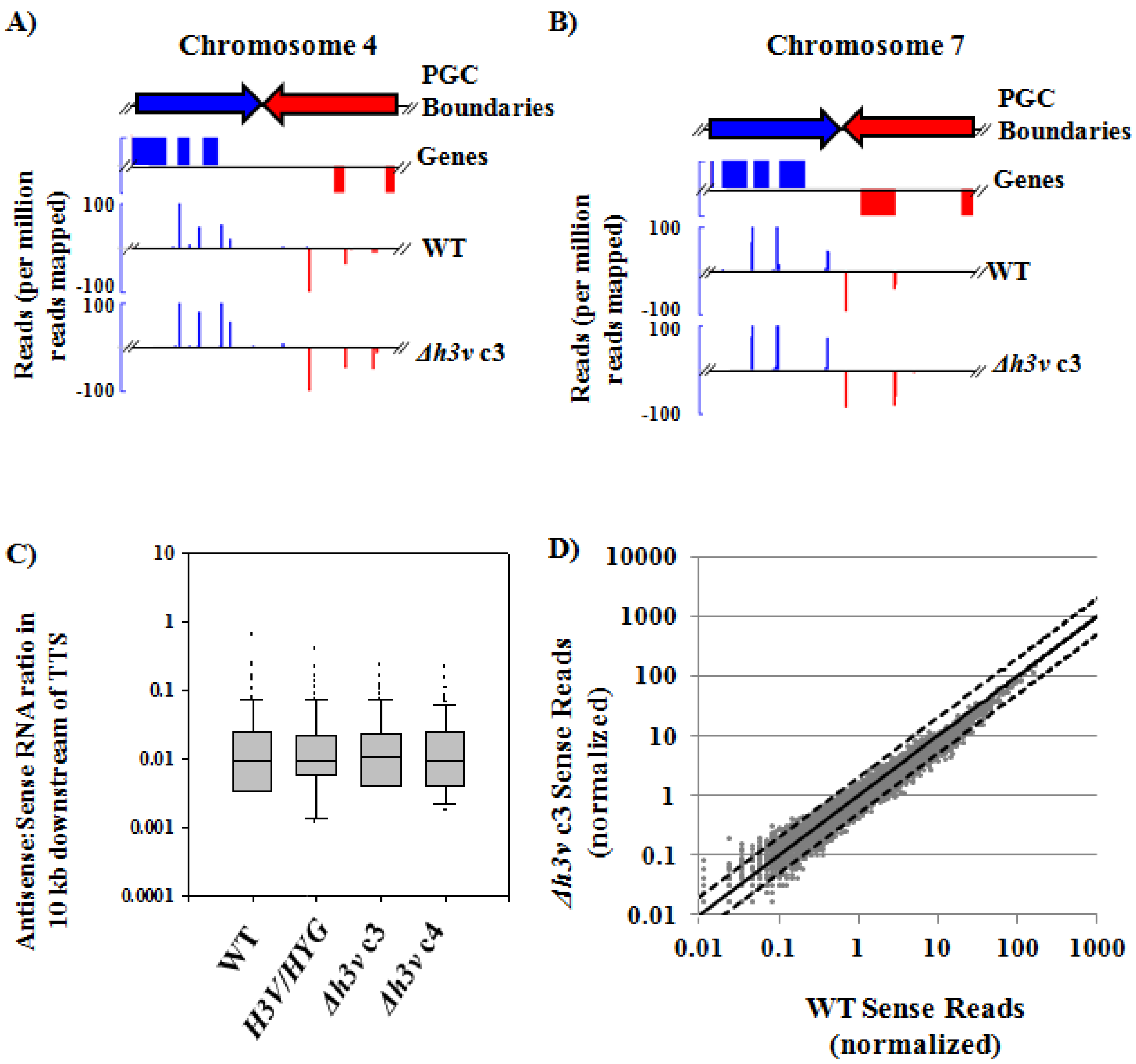
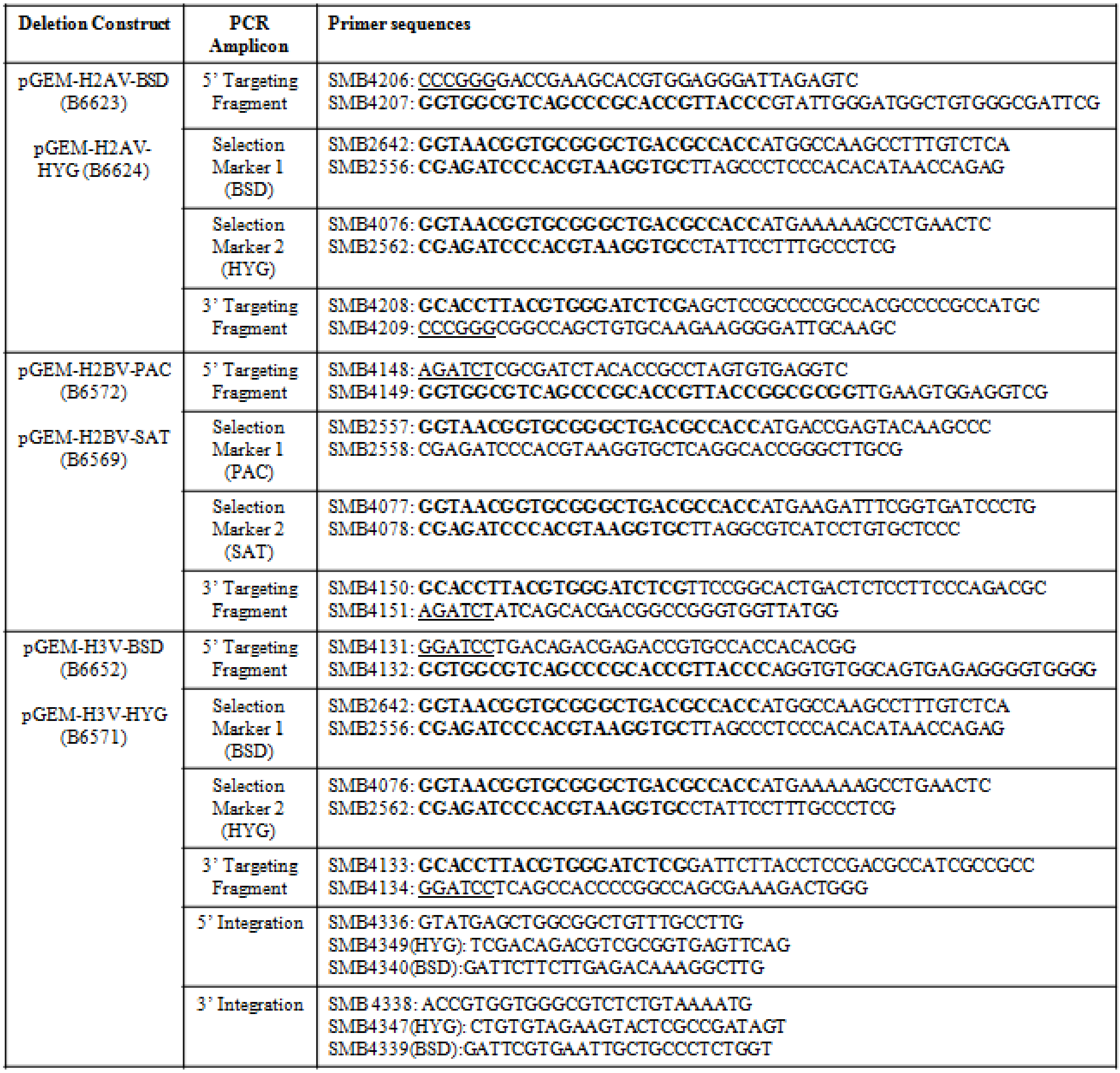


Figure 2-4. Deletion of *H3V* does not increase read-through transcription as observed by SL-RNA-seq

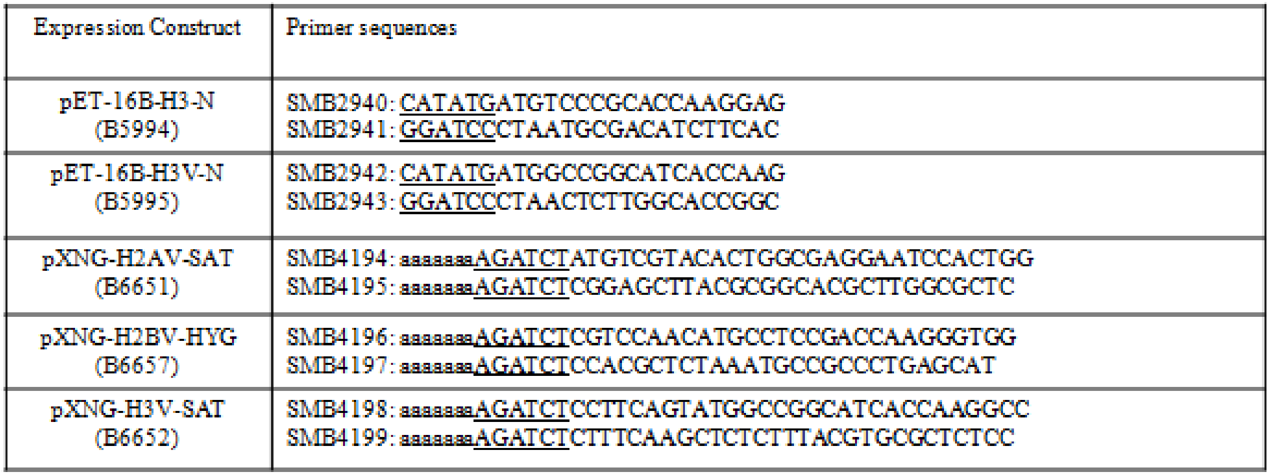


Supplemental Table S2-1. Primer sequences used for generation of deletion constructs for *H2AZ, H2BV,* and *H3V.*



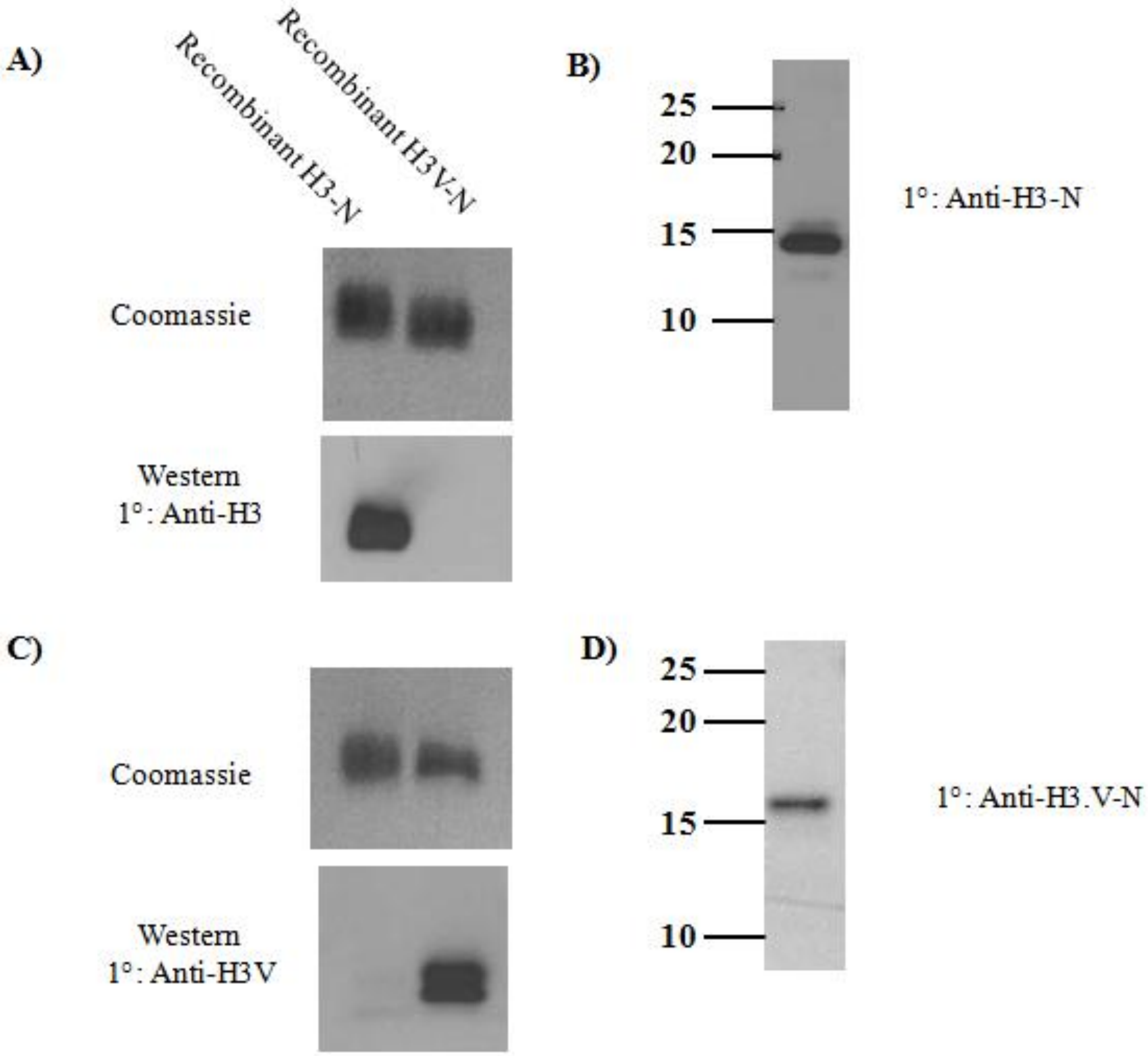
Deletion Construct	PCR Amplicon	Primer sequences
pGEM-H2AV-BSD (B6623) pGEM-H2AV-HYG (B6624)	5' Targeting Fragment	SMB4206: CCCGGGGACCGAAGCACGTGGAGGGATTAGAGTC SMB4207: **GGTGGCGTCAGCCCGCACCGTTACCC**GTATTGGGATGGCTGTGGGCGATTCG
	Selection Marker 1 (BSD)	SMB2642: **GGTAACGGTGCGGGCTGACGCCACC**ATGGCCAAGCCTTTGTCTCA SMB2556: **CGAGATCCCACGTAAGGTGC**TTAGCCCTCCCACACATAACCAGAG
	Selection Marker 2 (HYG)	SMB4076: **GGTAACGGTGCGGGCTGACGCCACC**ATGAAAAAGCCTGAACTC SMB2562: **CGAGATCCCACGTAAGGTGC**CTATTCCTTTGCCCTCG
	3' Targeting Fragment	SMB4208: **GCACCTTACGTGGGATCTCG**AGCTCCGCCCCGCCACGCCCCGCCATGC SMB4209: CCCGGGCGGCCAGCTGTGCAAGAAGGGGATTGCAAGC
pGEM-H2BV-PAC (B6572) pGEM-H2BV-SAT (B6569)	5' Targeting Fragment	SMB4148: AGATCTCGCGATCTACACCGCCTAGTGTGAGGTC SMB4149: **GGTGGCGTCAGCCCGCACCGTTACCGGCGCGG**TTGAAGTGGAGGTCG
	Selection Marker 1 (PAC)	SMB2557: **GGTAACGGTGCGGGCTGACGCCACC**ATGACCGAGTACAAGCCC SMB2558: CGAGATCCCACGTAAGGTGCTCAGGCACCGGGCTTGCG
	Selection Marker 2 (SAT)	SMB4077: **GGTAACGGTGCGGGCTGACGCCACC**ATGAAGATTTCGGTGATCCCTG SMB4078: **CGAGATCCCACGTAAGGTGC**TTAGGCGTCATCCTGTGCTCCC
	3' Targeting Fragment	SMB4150: **GCACCTTACGTGGGATCTCG**TTCCGGCACTGACTCTCCTTCCCAGACGC SMB4151: AGATCTATCAGCACGACGGCCGGGTGGTTATGG
pGEM-H3V-BSD (B6652) pGEM-H3V-HYG (B6571)	5' Targeting Fragment	SMB4131: GGATCCTGACAGACGAGACCGTGCCACCACACGG SMB4132: **GGTGGCGTCAGCCCGCACCGTTACCC**AGGTGTGGCAGTGAGAGGGGTGGGG
	Selection Marker 1 (BSD)	SMB2642: **GGTAACGGTGCGGGCTGACGCCACC**ATGGCCAAGCCTTTGTCTCA SMB2556: **CGAGATCCCACGTAAGGTGC**TTAGCCCTCCCACACATAACCAGAG
	Selection Marker 2 (HYG)	SMB4076: **GGTAACGGTGCGGGCTGACGCCACC**ATGAAAAAGCCTGAACTC SMB2562: **CGAGATCCCACGTAAGGTGC**CTATTCCTTTGCCCTCG
	3' Targeting Fragment	SMB4133: **GCACCTTACGTGGGATCTCG**GATTCTTACCTCCGACGCCATCGCCGCC SMB4134: GGATCCTCAGCCACCCCGGCCAGCGAAAGACTGGG
	5' Integration	SMB4336: GTATGAGCTGGCGGCTGTTTGCCTTG SMB4349(HYG): TCGACAGACGTCGCGGTGAGTTCAG SMB4340(BSD):GATTCTTCTTGAGACAAAGGCTTG
	3' Integration	SMB 4338: ACCGTGGTGGGCGTCTCTGTAAAATG SMB4347(HYG): CTGTGTAGAAGTACTCGCCGATAGT SMB4339(BSD):GATTCGTGAATTGCTGCCCTCTGGT

Supplemental Table S2-2. Primer sequences used for amplification of histone variant ORFs for protein expression and episomal complementation vectors.

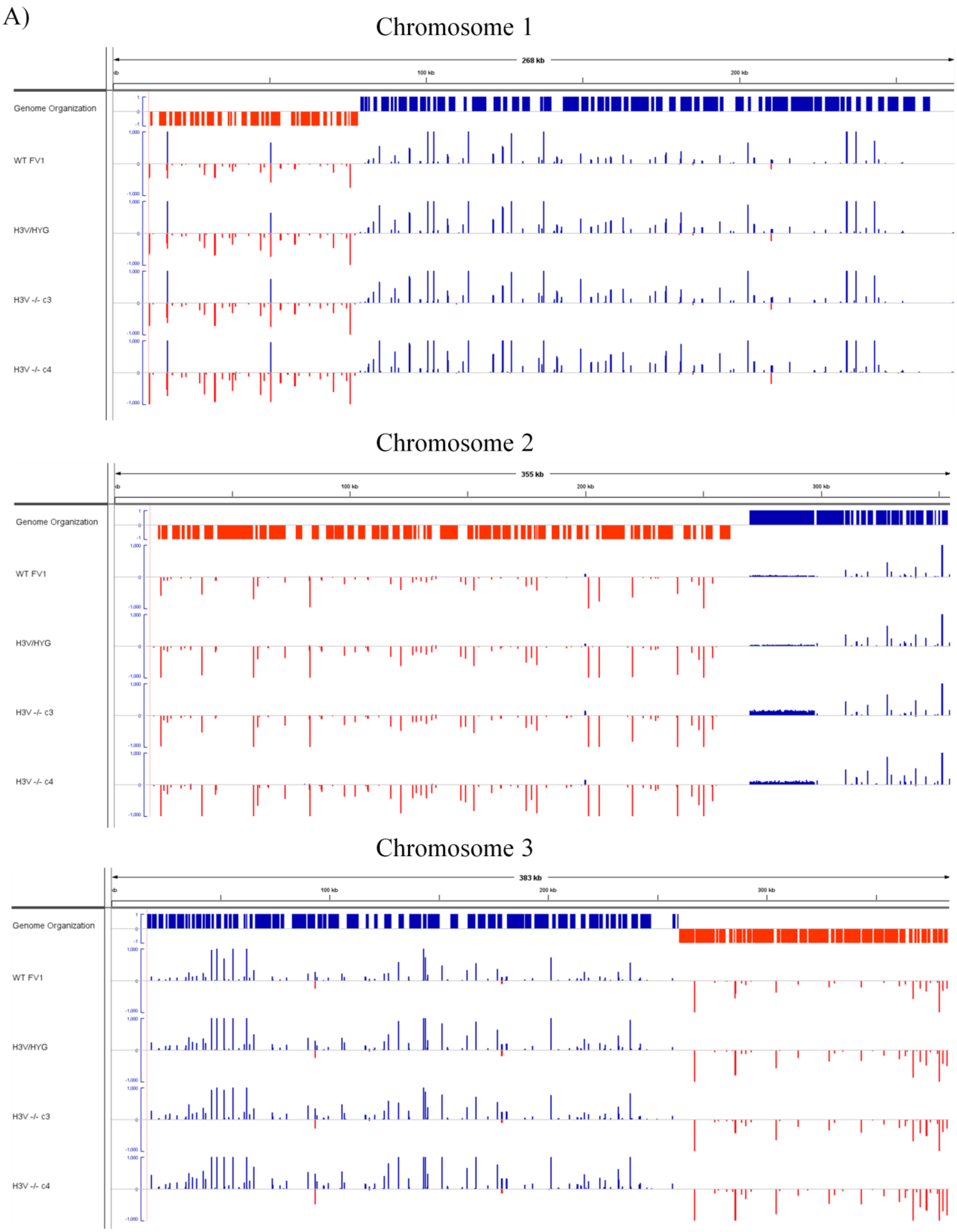


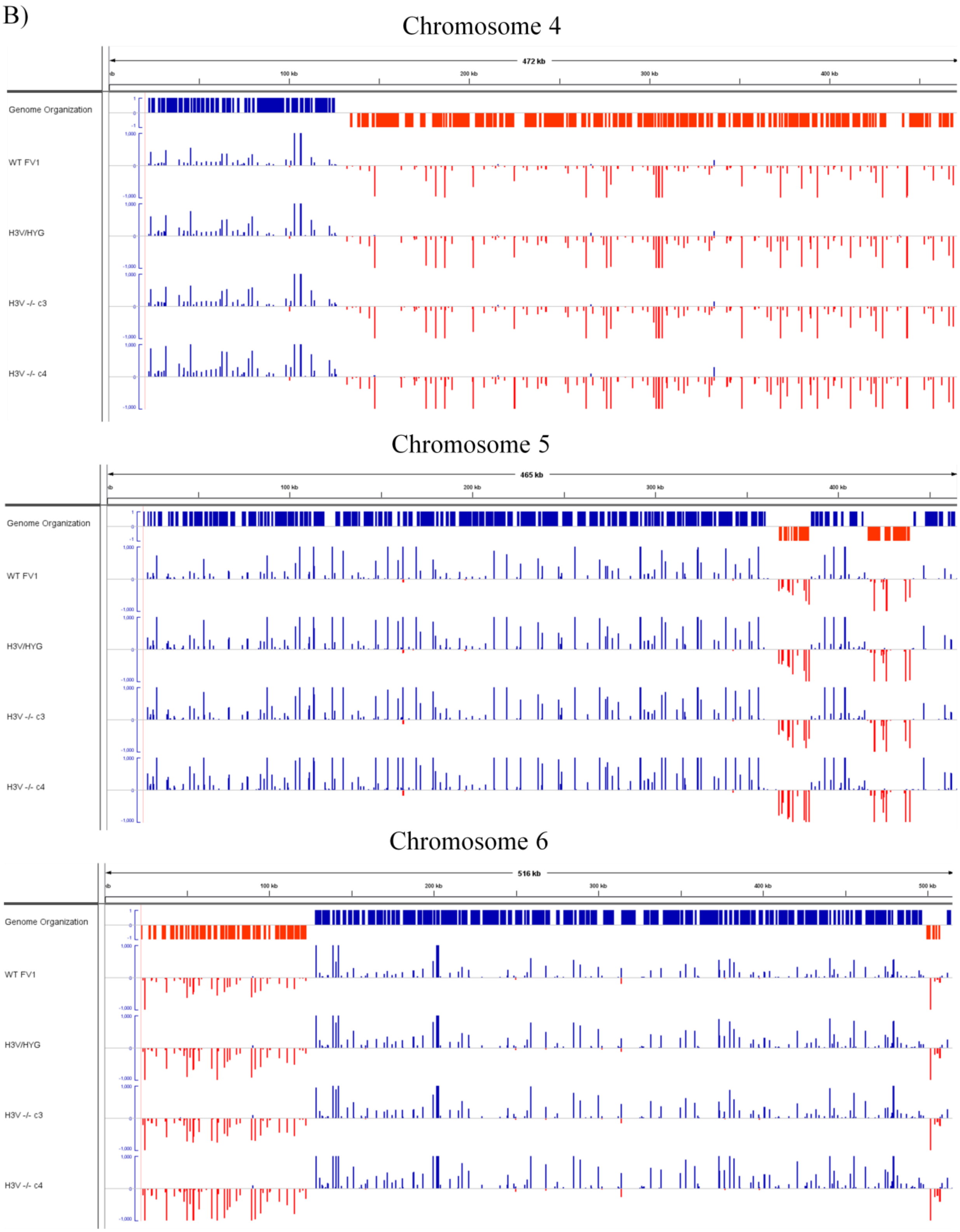
Expression Construct	Primer sequences
pET-16B-H3-N (B5994)	SMB2940: CATATGATGTCCCGCACCAAGGAG SMB2941: GGATCCCTAATGCGACATCTTCAC
pET-16B-H3V-N (B5995)	SMB2942: CATATGATGGCCGGCATCACCAAG SMB2943: GGATCCCTAACTCTTGGCACCGGC
pXNG-H2AV-SAT (B6651)	SMB4194: aaaaaaaAGATCTATGTCGTACACTGGCGAGGAATCCACTGG SMB4195: aaaaaaaAGATCTCGGAGCTTACGCGGCACGCTTGGCGCTC
pXNG-H2BV-HYG (B6657)	SMB4196: aaaaaaaAGATCTCGTCCAACATGCCTCCGACCAAGGGTGG SMB4197: aaaaaaaAGATCTCCACGCTCTAAATGCCGCCCTGAGCAT
pXNG-H3V-SAT (B6652)	SMB4198: aaaaaaaAGATCTCCTTCAGTATGGCCGGCATCACCAAGGCC SMB4199: aaaaaaaAGATCTCTTTCAAGCTCTCTTTACGTGCGCTCTCC

Supplemental Figure S2-1. Demonstration of anti-H3 and anti-H3.V specificity by western blotting.



Supplemental Figure S2-2. Transcription termination is unaltered in *Δh3v* parasites.



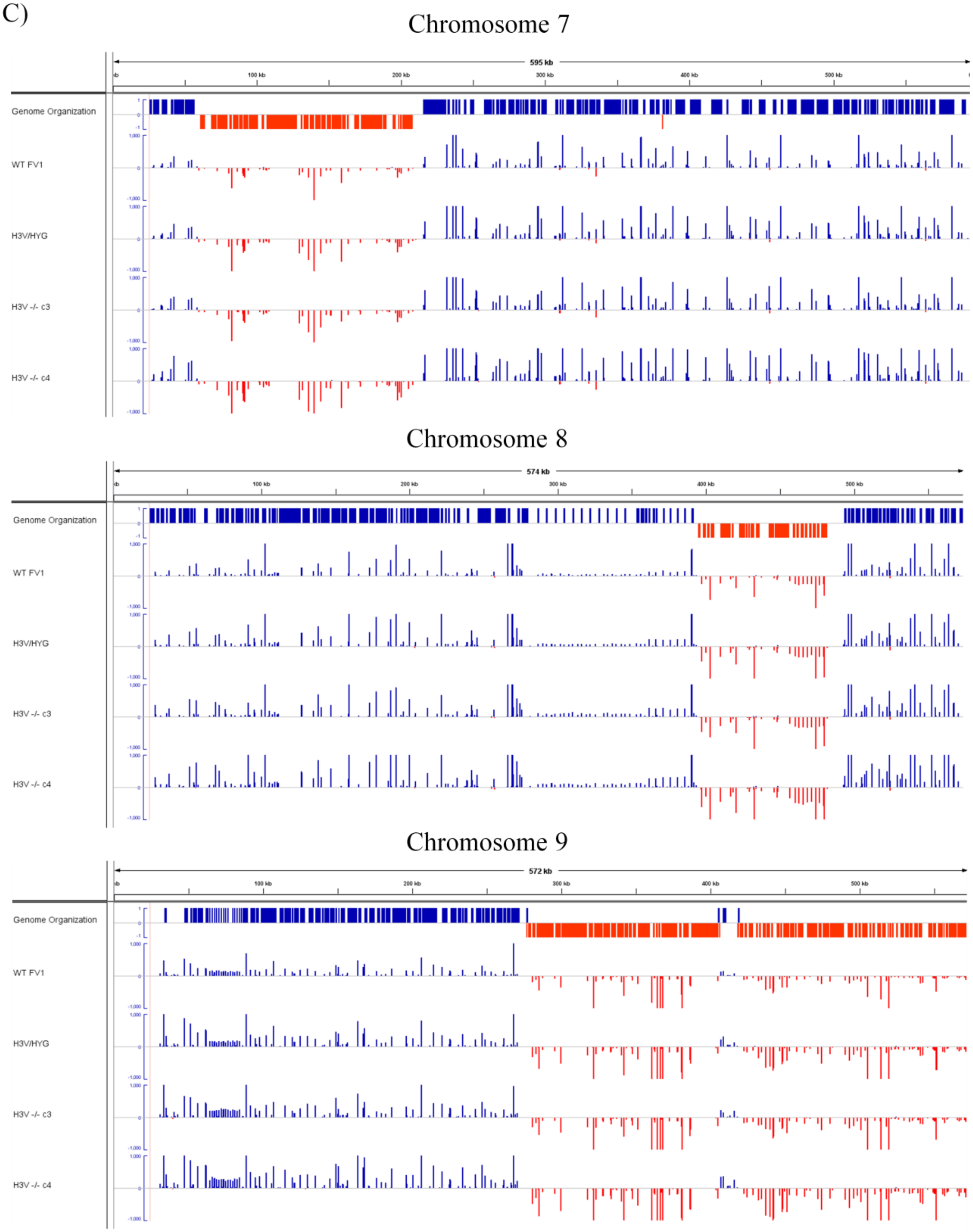

B)

Chromosome 4

Chromosome 5

Chromosome 6

C)

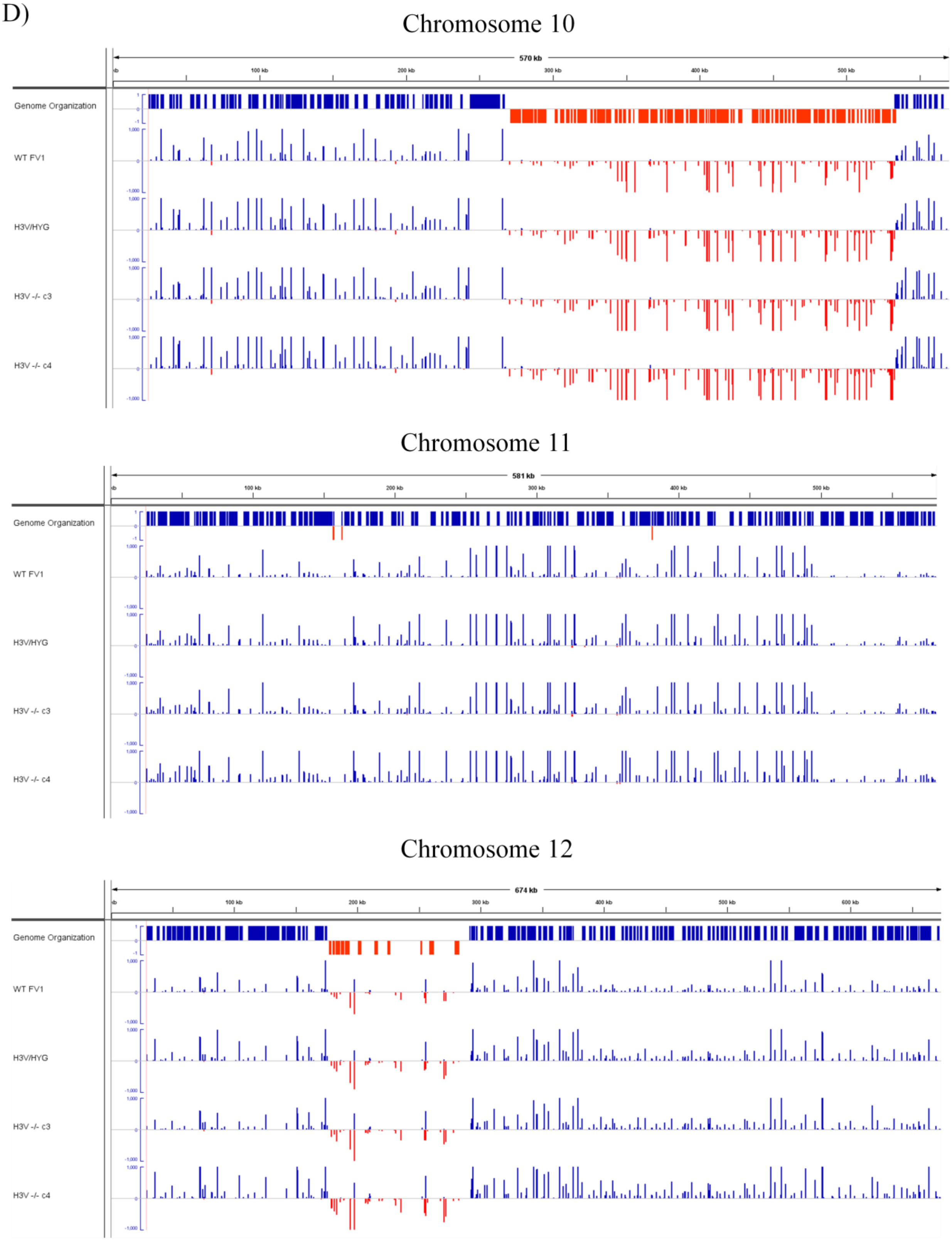


Chromosome 7

Chromosome 8

Chromosome 9

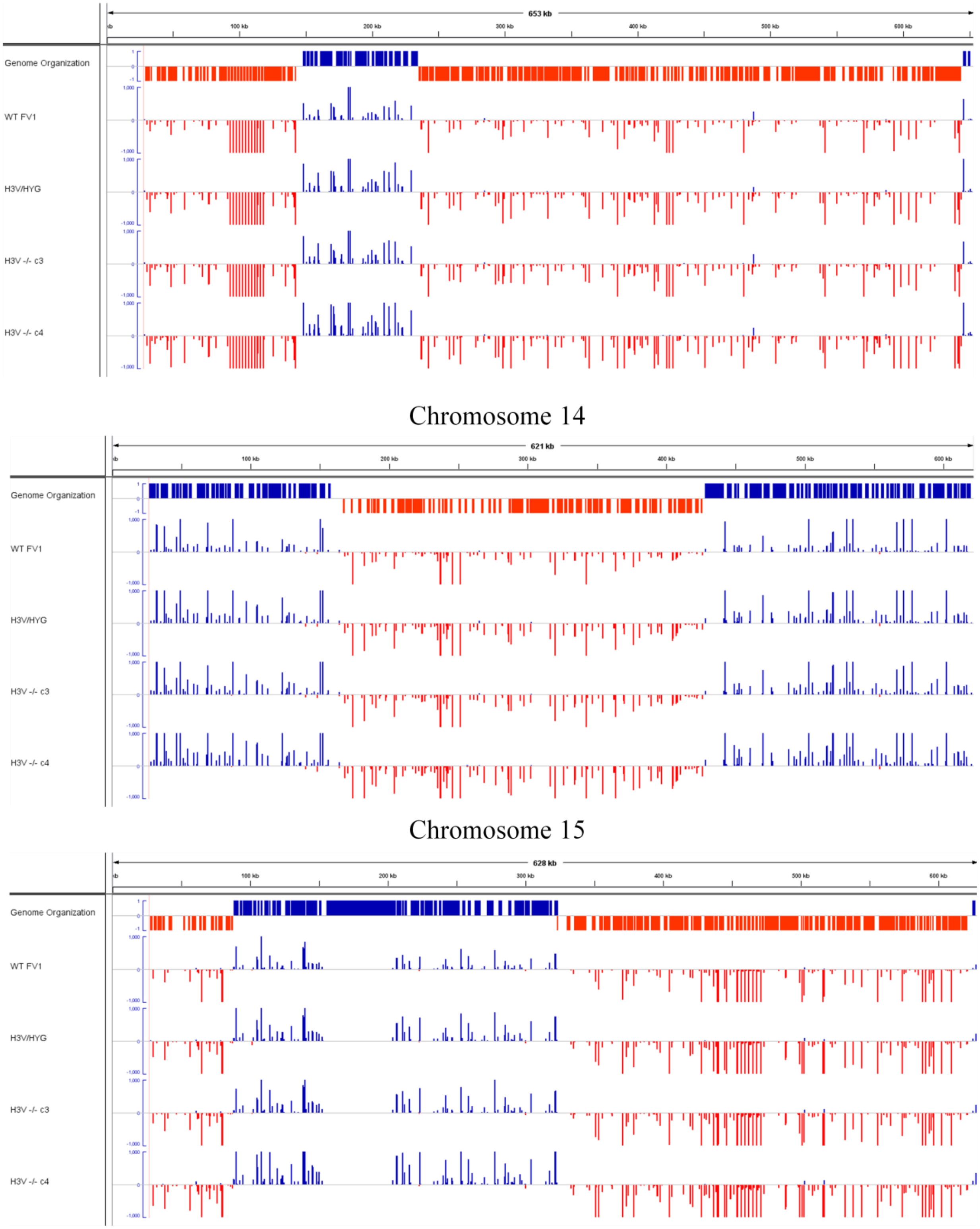
D)


Chromosome 10

Chromosome 11

Chromosome 12

E)



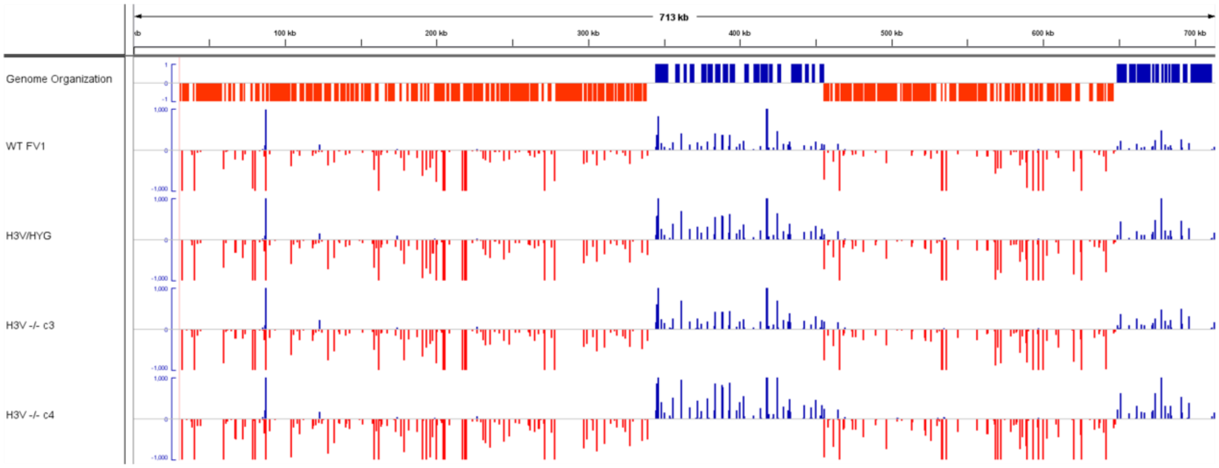
Chromosome 13

Chromosome 14

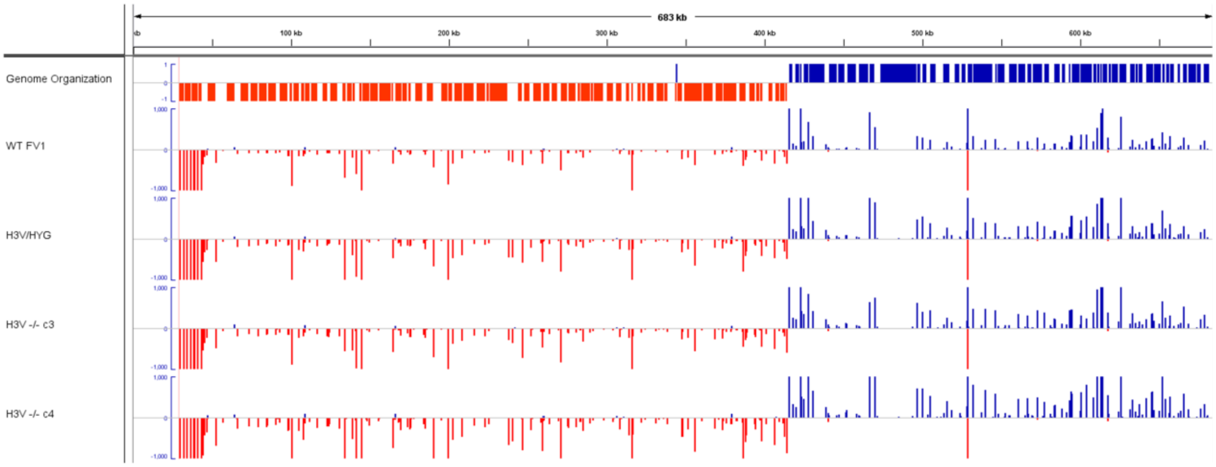
Chromosome 15

F)

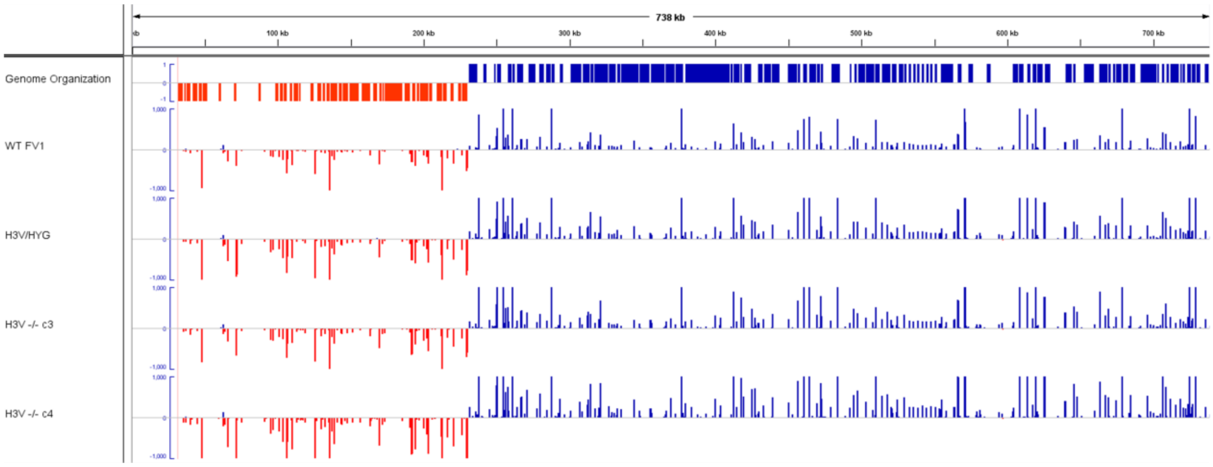
Chromosome 16



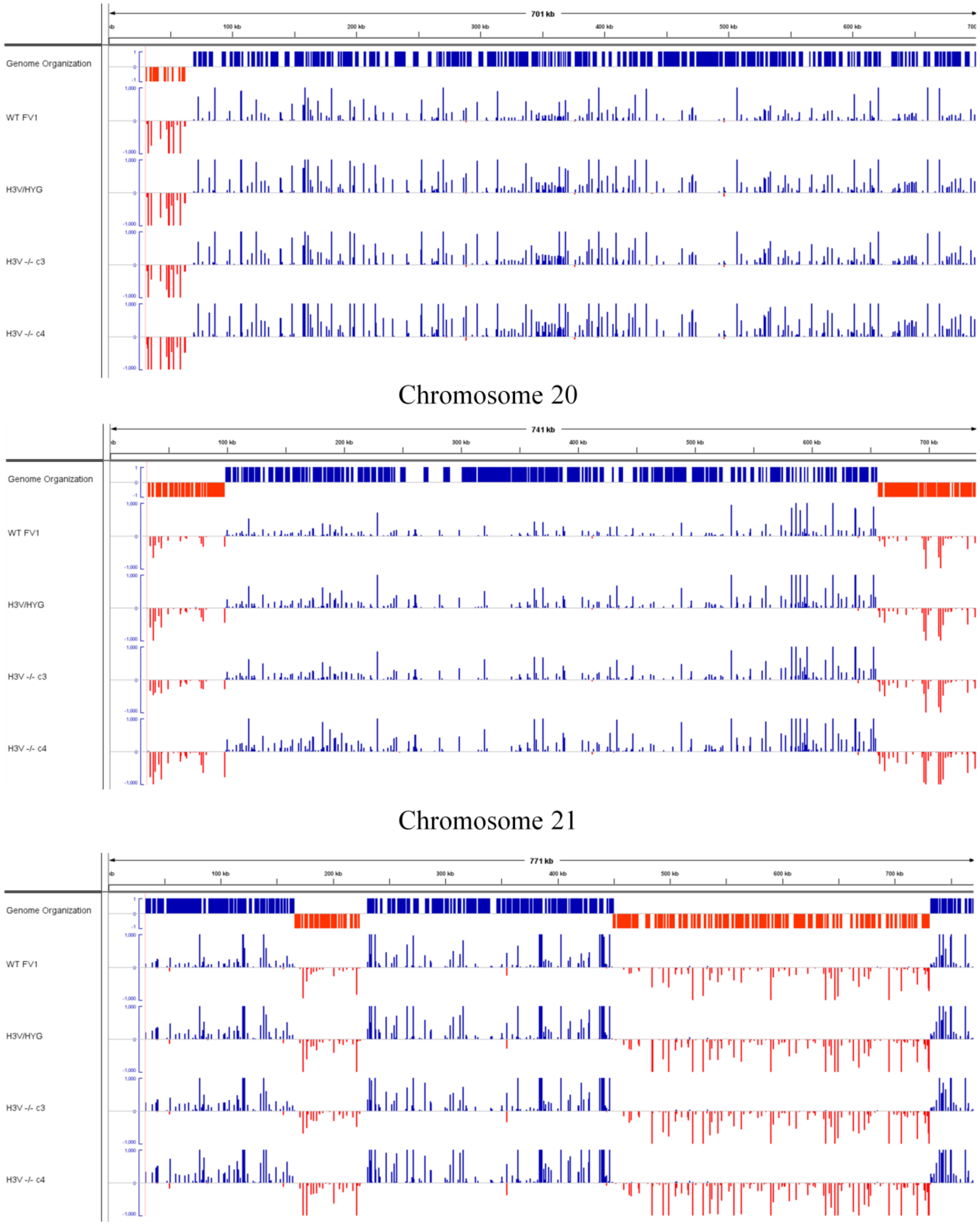
Chromosome 17



Chromosome 18



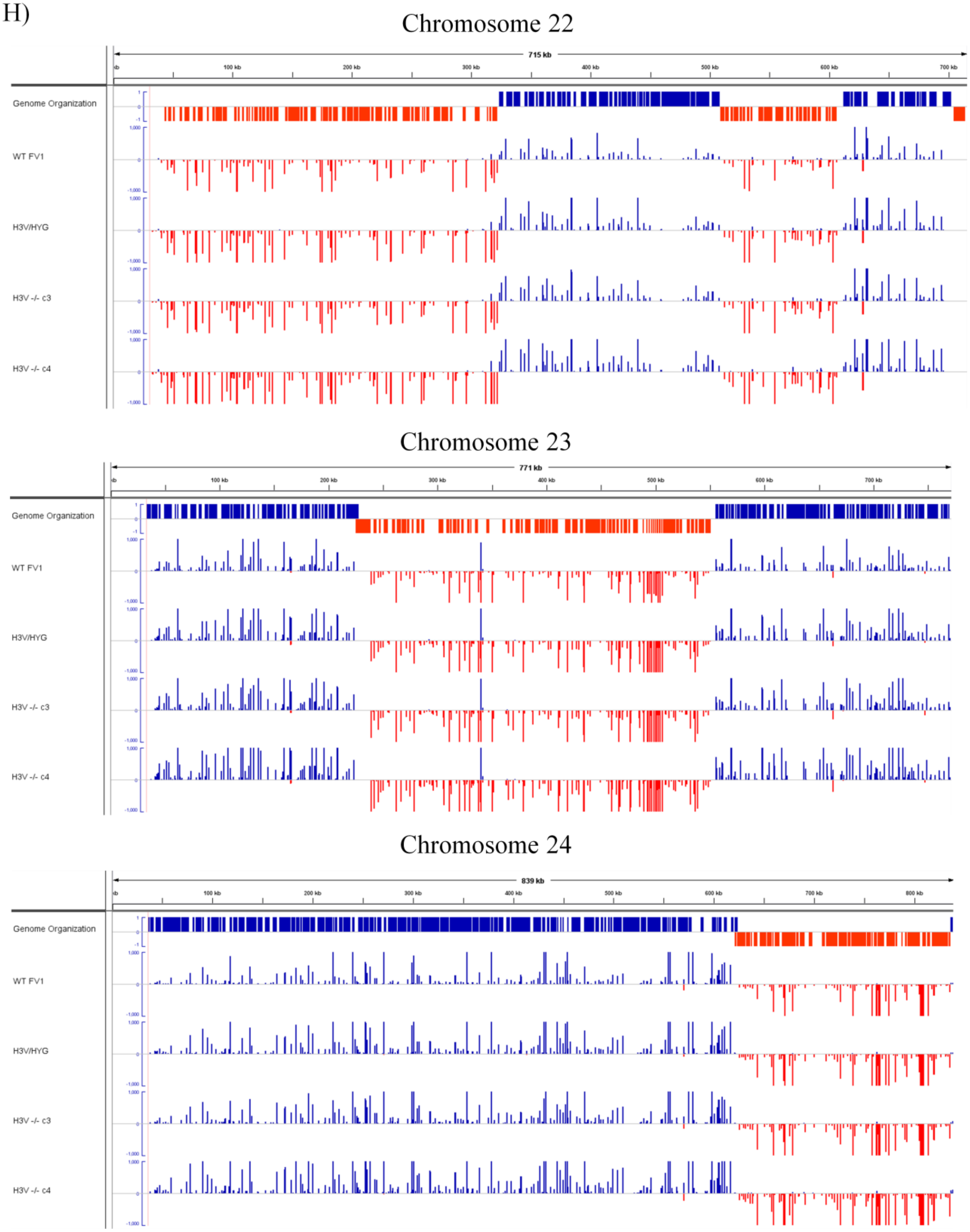
G)


Chromosome 19

Chromosome 20

Chromosome 21

H)

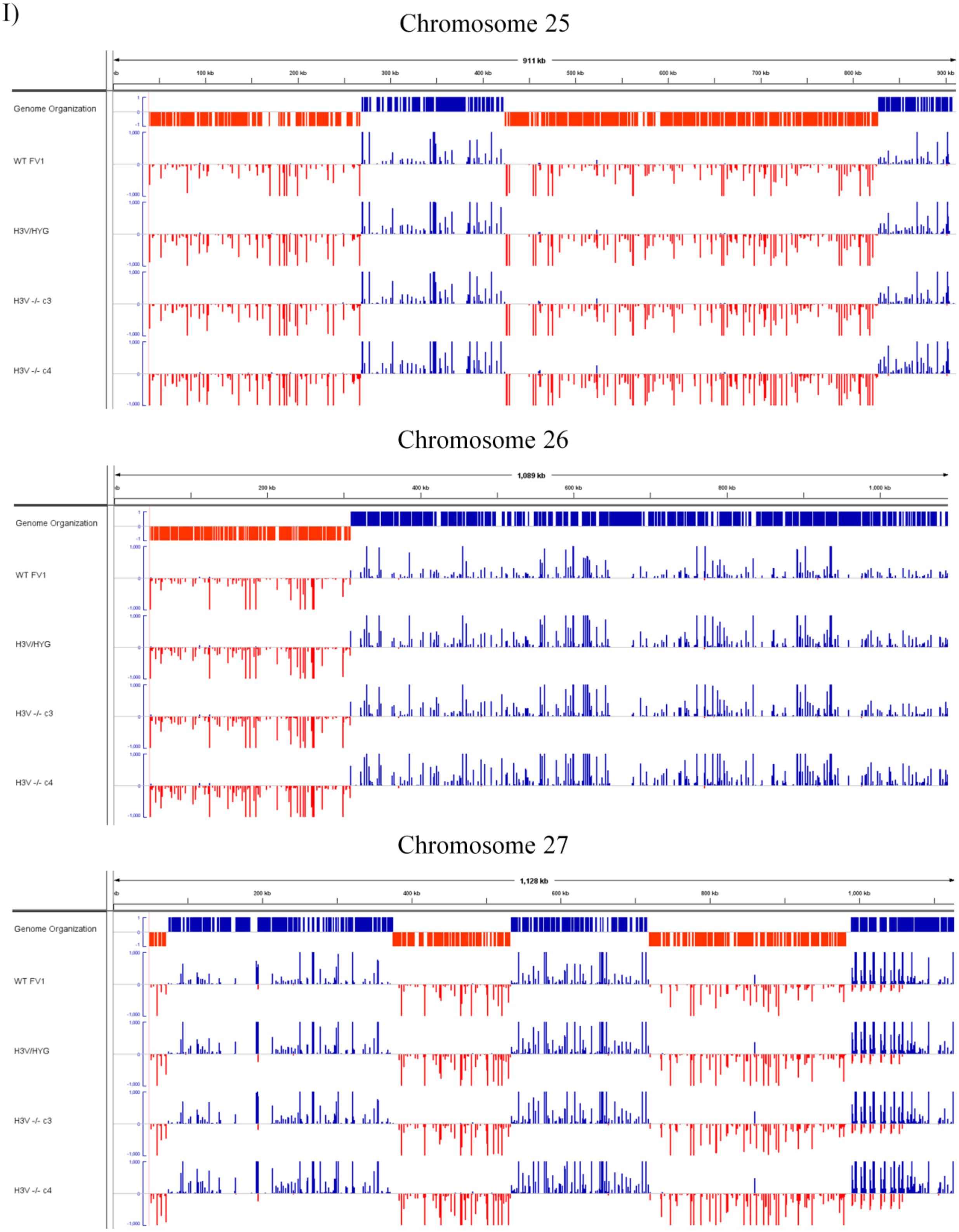


Chromosome 22

Chromosome 23

Chromosome 24

I)



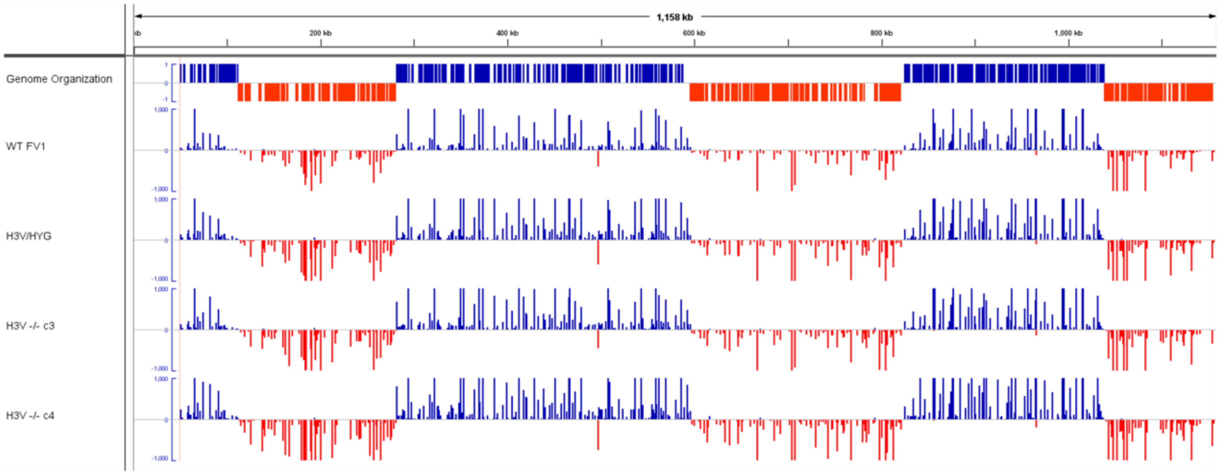
Chromosome 25

Chromosome 26

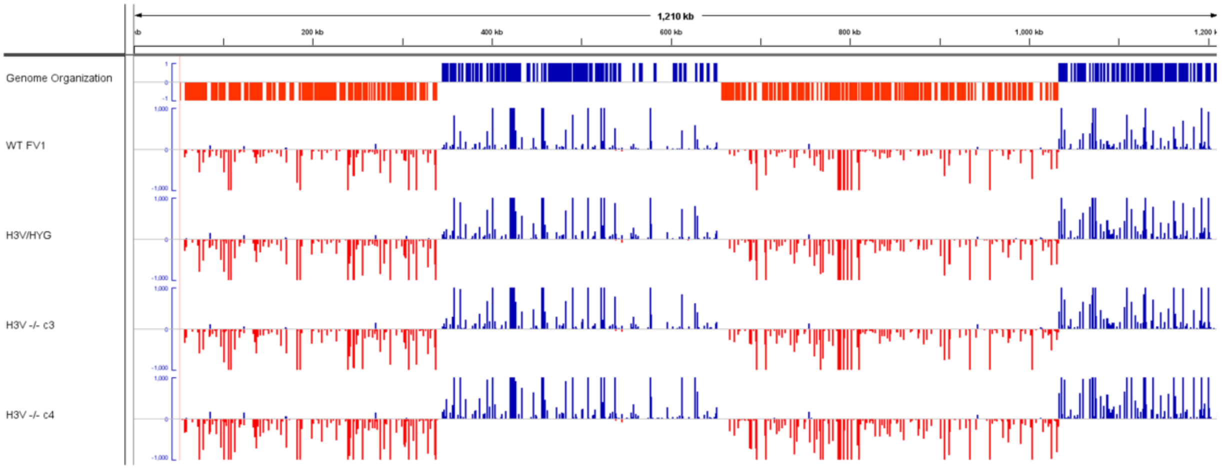
Chromosome 27

J)

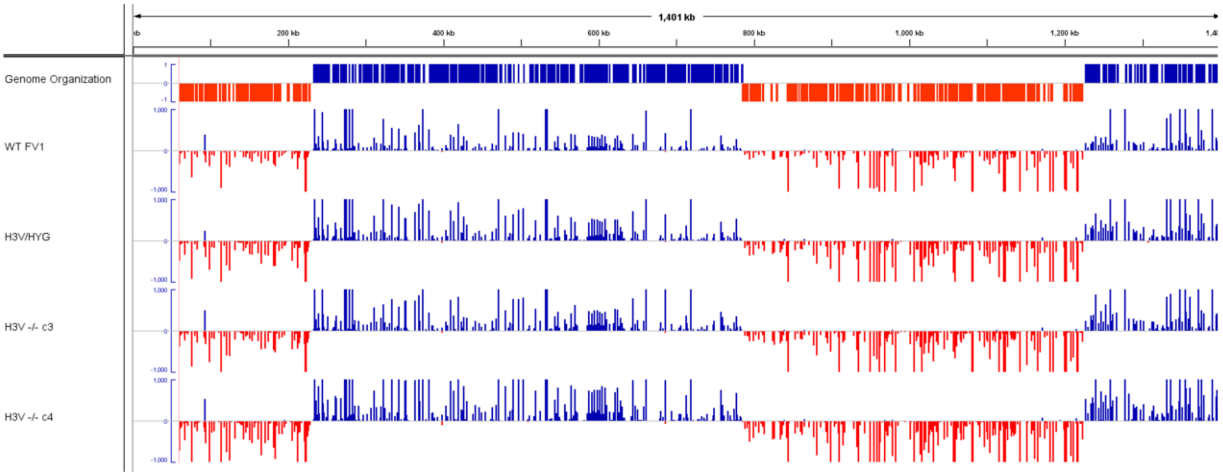
Chromosome 28



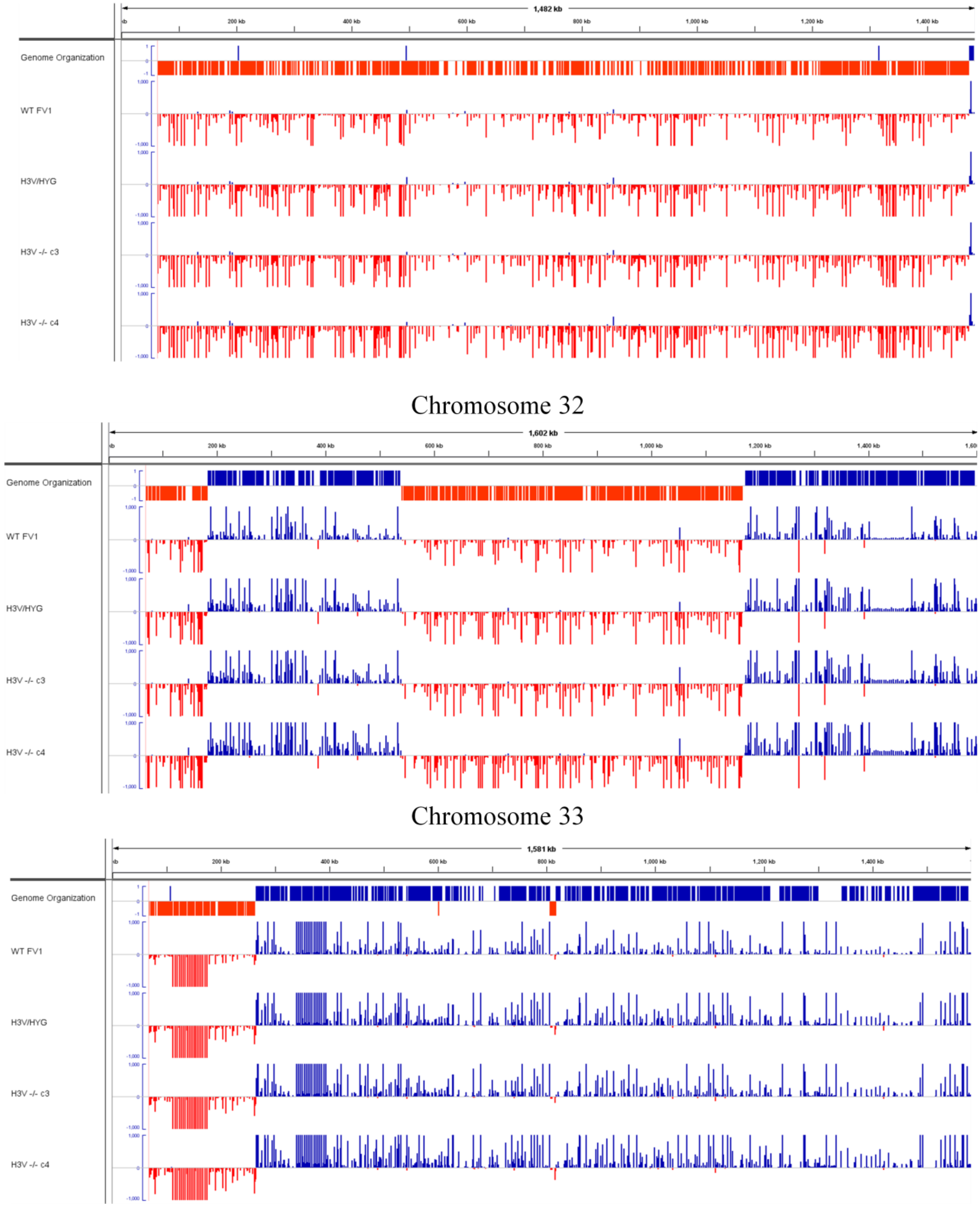
Chromosome 29



Chromosome 30



K)

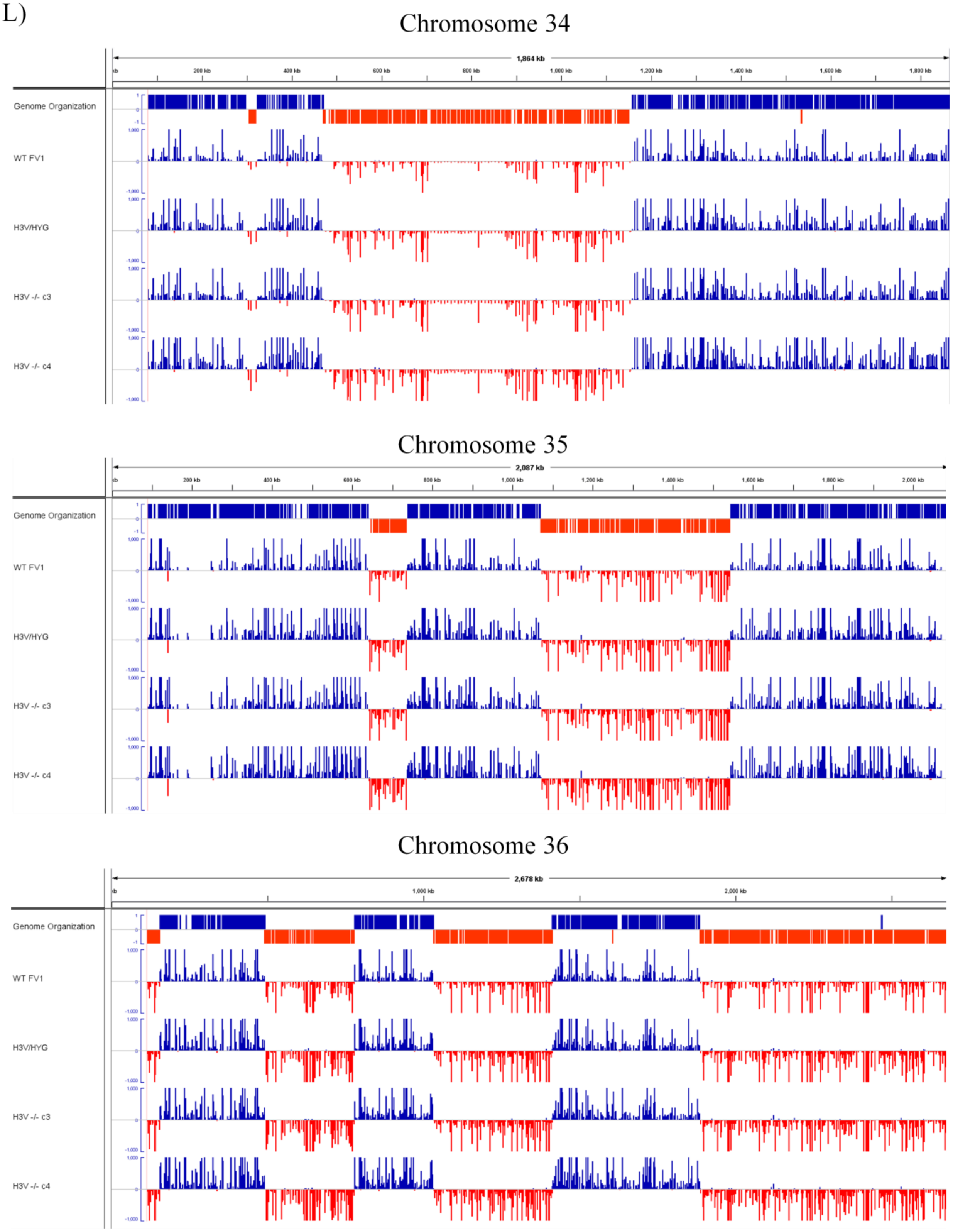


Chromosome 31

Chromosome 32

Chromosome 33

L)

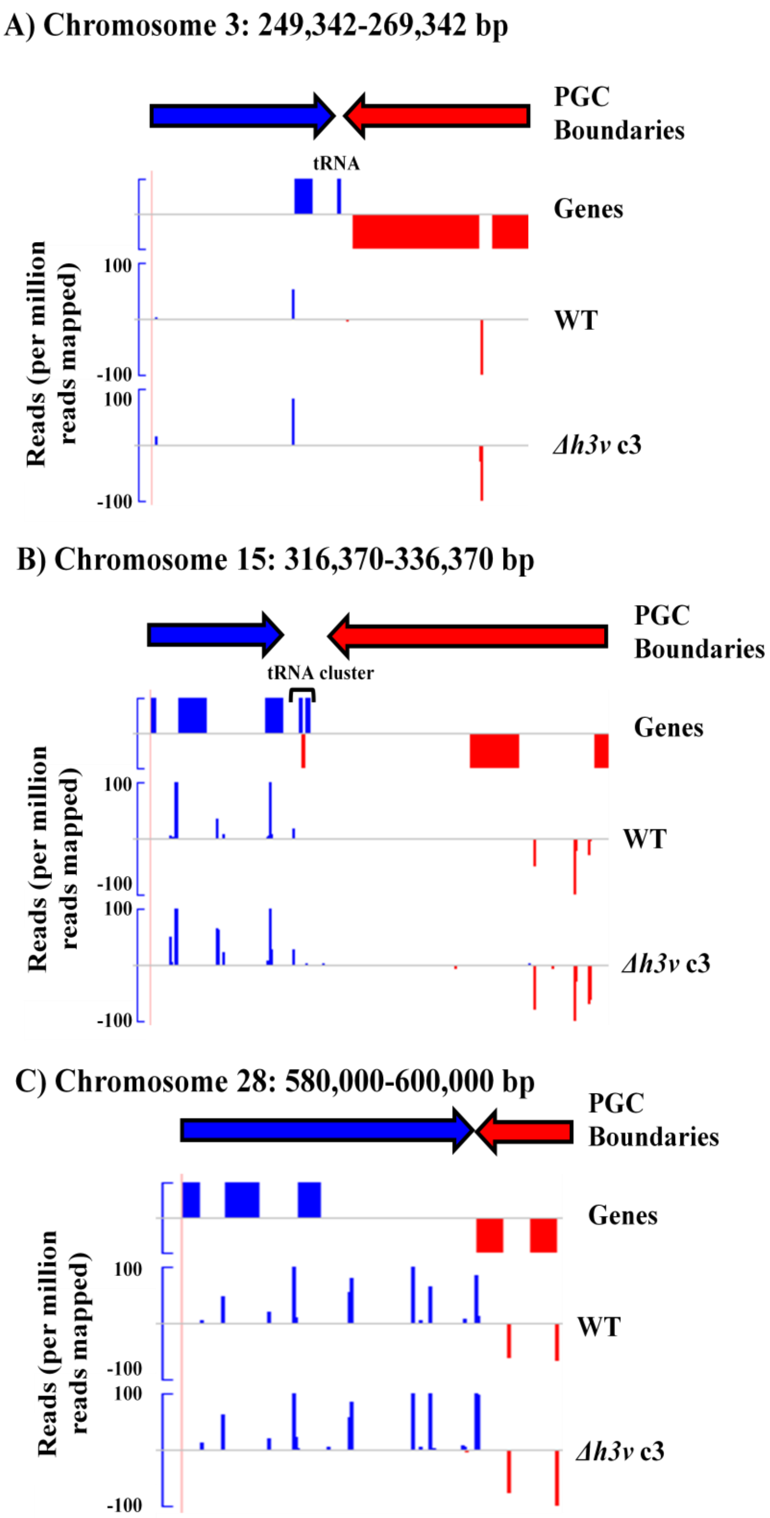


Chromosome 34

Chromosome 35

Chromosome 36

Supplemental Figure S2-3. Transcription termination is unaltered in *Δh3v* parasites, regardless of whether tRNAs are present in the cSSR.



A) Chromosome 3: 249,342-269,342 bp

B) Chromosome 15: 316,370-336,370 bp

C) Chromosome 28: 580,000-600,000 bp

Chapter Three

The chromatin landscape of the early-diverging eukaryote *Leishmania major*

Preface

BA designed and performed all experiments, assisted with NGS library preparation at the Genome Technology Access Center (GTAC), and performed data analysis. JS wrote Java-based bioinformatics platform, which was streamlined and implemented by BA. SMB supervised these studies and assisted in the design of these experiments. Profs. Sarah Elgin and Gary Stormo provided helpful comments regarding the bioinformatics pipeline and the data analysis. The first draft of this chapter was written by BA, and comments from SMB are included in the final version presented here.

We intend to publish this work in two parts:

Anderson, B.A., Shaik, J.S., and Beverley, S.M. The chromatin landscape of the early-diverging eukaryote *Leishmania major*. In preparation.

Shaik, J.S., Anderson, B.A., and Beverley, S.M. Peak and valley detection from epigenome-derived data. In preparation.

Abstract

Polycistronic transcription of protein-coding genes in *Leishmania* and other kinetoplastid protozoa initiates primarily in divergent strand switch regions (dSSRs), which lack canonical eukaryotic promoter motifs but possess activating epigenetic marks typical of functional promoters. In eukaryotic chromatin, active regulatory elements are nucleosome-depleted and hypersensitive to endonuclease digestion, qualities which distinguish them from the bulk of chromatin and facilitate the characterization of novel regulatory elements therein. Here, we describe the chromatin landscape of *Leishmania major* in an effort to similarly identify active regulatory elements genome-wide using two complementary techniques coupled to Illumina sequencing: micrococcal nuclease digestion of chromatin (MNAse-seq), which identifies nuclease-hypersensitive (NH) sites, and formaldehyde-assisted isolation of regulatory elements (FAIRE-seq), which isolates protein-depleted DNA sequences. These techniques do not require specialized reagents, and advances in next-generation sequencing technologies have made them increasingly useful for assessment and identification of novel regulatory elements. To address challenges associated with variations in copy number, aneuploidy, and sequencing bias arising during identification of nuclease-hypersensitive (NH) sites from MNAse-seq data and enriched regions from FAIRE-seq data, we developed a flexible Java-based software suite called Peak and VallEy Detector (PAVED), which facilitates the comparison of experimental datasets with appropriately designed control datasets. Using MNAse-seq, we identified NH sites spanning highly-transcribed tRNA and rRNA genes as expected, and we identified well-positioned nucleosomes at the 5' end of many tRNA genes. However, we observed very few NH sites in dSSRs and detected no consistent positioning or phasing among nucleosomes in these and most other regions of the genome. Because heterogeneous or transient NH sites within a population of

cells would be obscured by abundant, poorly phased nucleosomes within dSSRs, we turned to FAIRE-seq, calling peaks at a threshold of 5-fold or greater over input DNA. This method again detected tRNA and rRNA genes, as well as significant peaks over dSSRs, spanning broad regions overlapping those bearing known epigenetic marks. Because histones appear to be present at similar densities within dSSRs and internal regions, we tested whether nucleosomes from FAIRE-enriched regions were less stable or more transient using MNAse overdigestion. This now reveals regions correlating with FAIRE peaks, indicating a qualitative difference between dSSR-proximal nucleosomes compared to those found in the rest of the genome. Thus, transient and/or heterogeneous nucleosome-depleted regions are distributed broadly in dSSRs, rather than localized to discrete loci defined by DNA-encoded structural elements. These experiments support a model of delocalized transcription initiation occurring within permissive epigenetic environments, a mechanism compatible with the reliance in these organisms on *trans*-splicing instead of transcription initiation to define mRNA 5' ends.

Introduction

Leishmania and other kinetoplastid protozoa generate mature messenger RNAs (mRNAs) using an unusual bipartite mechanism of transcription by RNA polymerase II (RNAP II). Protein-coding genes are transcribed polycistronically from long, unidirectional arrays called polycistronic gene clusters (PGCs), which can contain hundreds of functionally unrelated genes [reviewed in (1)]. Separately, the transcripts encoding the capped 5' end of each mRNA are transcribed from the spliced leader (SL) RNA gene array. Maturation of polycistronic pre-mRNAs occurs via coupled *trans*-splicing and polyadenylylation reactions, where the capped 39-nt SL RNA is *trans*-spliced at a 5' splice acceptor site, and polyadenylylation of the upstream transcript follows (2,3). The mechanisms regulating transcription are also unusual: kinetoplastid genomes lack canonical RNAP II promoter and terminator elements (4–6), and comparison to other eukaryotic genomes reveals the presence of general but not sequence-specific RNAP II transcription factors (7). The sole motifs which have been identified in dSSRs are long, G-rich stretches of DNA (4,5), but the functional significance of these loci remains unknown.

Despite the paucity of obvious DNA-encoded elements in these regions, transcription start site (TSS) and transcription termination site (TTS) mapping by RNA-seq and characterization of epigenetic marks in *Leishmania* and the related kinetoplastids *Trypanosoma brucei* and *Trypanosoma cruzi* demonstrate that individual transcription units are primarily defined by the boundaries of PGCs, referred to as divergent and convergent strand switch regions (dSSRs and cSSRs, respectively). In dSSRs, two PGCs are oriented head-to-head and are marked by broad peaks of trimethylated histone H3 lysine 4 (H3K4me3) [*T. brucei, T. cruzi*; (8,9)], acetylation of the N-terminal tail of histone H3 [*L. major*; (10)], and the incorporation of the histone variants H2A.Z and H2B.V [*T. brucei*; (4)]. These peaks can completely encompass shorter dSSRs, while longer dSSRs contain two separate peaks; in all cases, a small number of PGC-internal peaks were observed, which coincide with *bona fide* regions of transcription initiation in *T. brucei* (11). In agreement with the lack of readily identifiable promoter elements, transcription initiation is delocalized within dSSRs, as multiple transcription start sites (TSS) were identified within the dSSR of chromosome 1 in *L. major* (5) and in all dSSRs in *Trypanosoma brucei* (11). Interestingly, chromatin immunoprecipitation (ChIP) of the RNAP II general transcription factor TBP (TRF) in *L. major* and *T. brucei* demonstrate widespread binding across the entire genome, with higher levels in dSSRs (10,12). This PGC-internal transcription factor binding likely corresponds to sites of infrequent transcription initiation, as very low levels of transcription initiation were detected genome-wide in *T. brucei* (11).

Together, the lack of known promoter elements, the identification of widespread transcription initiation events, and the broad peaks of dSSR-associated epigenetic marks have led many to hypothesize that although transcription may initiate promiscuously genome-wide, dSSRs act as *de facto* promoters through maintenance of a transcriptionally-permissive epigenetic environment. However, many questions remain regarding the epigenetic nature and function of dSSRs, including whether DNA-encoded elements might facilitate the acquisition of this permissive chromatin state. In budding yeast, homopolymeric sequences such as poly(dA:dT) tracts are inherent components of promoters for some housekeeping genes [reviewed in (13)] and function primarily by defining nucleosome-free regions [reviewed in (14)], which can drastically alter the behavior of weak or degenerate promoter sequences by their inherent nucleosome-disfavoring properties (15). Interestingly, the function of poly(dA:dT) tracts does not require perfect homopolymers (15) and can be substituted by poly(dG:dC) tracts (16), suggesting that the structural properties of DNA sequences, rather than the specific sequences themselves, confer promoter activity. The long, G-rich tracts previously identified in *T. brucei* and *Leishmania* and other nucleosome-disfavoring sequences could play a role in facilitating the transcriptionally-permissive environment found in and around dSSRs in *Leishmania*.

To study the role of poly(dG:dC) tracts and to address the potential for additional sequences which influence nucleosome placement in dSSRs, we characterized the chromatin landscape genome-wide in *L. major* using two independent but complementary methods coupled to paired-end Illumina sequencing: micrococcal nuclease digestion of intact chromatin (MNAse-seq) and formaldehyde-assisted enrichment of regulatory elements (FAIRE-seq). In MNAse-seq, nucleosome-bound DNA sequences are isolated, and the boundaries of individual nucleosomes can be accurately determined by using paired-end sequencing. Importantly, these datasets provide two forms of insight into the factors influencing the positioning and spacing of nucleosomes, specifically in TSS- and TTS-proximal regions (17–21). These studies also define nuclease-hypersensitive (NH) sites by their lack of coverage, yielding NH sites which are well-conserved across cellular populations including poly(dA:dT) tracts (16) or active promoters (20,22). In addition, knowledge of the boundaries of the nucleosome enables determination of the positioning and spacing of nucleosomes by locating the midpoint of the sequenced DNA fragments. In a number of systems, the nucleosomes downstream of many TSS are spaced at regular intervals [reviewed in (14)], suggesting that the presence of regularly spaced nucleosomes in dSSRs could also indicate the location of active regulatory elements. As a complement to these experiments, we used FAIRE-seq, which relies on isolating protein-depleted loci from crosslinked chromatin by a phenol-based extraction and frequently isolates loci which correlate well with NH sites identified by MNAse-seq (23). However, FAIRE is also capable of detecting heterogeneous or transient NH loci which are obscured by abundant, poorly-phased nucleosomes in MNAse-seq experiments by preferentially isolating these loci from a sea of nucleosome-bound fragments. These loci display a smaller degree of enrichment by FAIRE, which correlates with the proportion of loci which are nucleosome-depleted in the population (23).

The proper analysis of epigenome-derived datasets from *Leishmania* such as those generated from MNAse- and FAIRE-seq requires adjustment to account for aneuploidy, which is highly prevalent in this species [(24); reviewed in (25,26)] and can vary significantly among cells in a single culture (24). Other potential artifacts in epigenome-focused next generation sequencing experiments can arise from enzyme-induced biases (27) and errors during next-generation sequencing (28). These challenges could distort the interpretation of the results, especially in *Leishmania*, as poly(dG:dC) tracts within dSSRs are of particular interest. We developed several experimental and computational strategies to better address these issues. First, we used comparisons so similar experimental treatments of purified DNA and developed an analytical pipeline, PAVED, to filter out loci which are covered poorly in control datasets due to technical issues or errors during sequencing or alignment, and to extract loci of interest from both types of datasets: nuclease hypersensitive "valleys" from MNAse-seq data, and enriched peaks from FAIRE or ChIP-seq datasets (Shaik et al., in preparation). Although we demonstrate its utility in *Leishmania*, we believe it addresses technical challenges that arise in other eukaryotic systems as well, and we aimed to make this pipeline versatile and easy to implement.

Using this pipeline to analyze MNAse-seq datasets derived from MNAse-treated chromatin and naked DNA, we identified NH sites at tRNA and rRNA genes, consistent with the high promoter activity at these loci. In contrast, we identified few NH sites in dSSRs; both MNAse-seq and quantitative PCR analysis show that the poly(dG:dC) tracts in the dSSR of chromosome 1 are not marked by an NH site. Furthermore, analysis of nucleosome positions genome-wide demonstrates that well-positioned, regularly-spaced nucleosomes are rarely found across the genome. We observed positioned nucleosomes upstream of many tRNA genes, although the specific distance from the tRNA gene and promoters varied among genes and we could not detect these in when tRNA and other RNAP III-transcribed genes were clustered in close proximity. Interestingly, nucleosome positioning analysis suggests the possibility of periodic rotational positioning based on the helical turns of the DNA strand, but demonstrates that strongly positioned, regularly spaced nucleosomes are infrequent in dSSRs and elsewhere in the genome.

In contrast to MNAse-seq, we observed broad peaks of FAIRE enrichment across dSSRs which closely mirror the patterns of known histone modifications, suggesting that transient and/or heterogeneous NH sites which arise in transcriptionally permissive chromatin environments may be responsible for more frequent transcription initiation events. Using restriction endonuclease sensitivity assays, we show that dSSRs and PGC-internal loci have similar nucleosome densities despite showing significant differences in FAIRE-associated NH sites. To understand the origin of these transient NH sites, we tested whether dSSR-proximal nucleosomes showed qualitative differences in stability using MNAse overdigestion. In contrast to standard MNAse-seq, this now reveals broad regions around dSSRs which are depleted of nucleosomes, suggesting that dSSR-proximal nucleosomes may have altered stabilities or rates of displacement. Comparison of regions depleted by MNAse overdigestion, FAIRE peaks, and known patterns of acetylated histone H3 (10) demonstrate a high degree of correlation among these datasets. Together, these data support a model in which transcription initiates promiscuously from a permissive epigenetic environment in which destabilized nucleosomes generate transient NH sites, rather than one which is defined by transcription factor binding or by nucleosome-disfavoring sequences like many eukaryotic promoter elements.

Results

Development and implementation of Peak and Valley Detector (PAVED)

We sought to identify NH sites and putative transcription start sites genome-wide in *Leishmania* using MNAse-seq and FAIRE-seq. To accomplish these analyses and include relevant control datasets, we developed Peak and Valley Detector (PAVED), a Java-based computational pipeline which accepts read alignments in the commonly used BAM format, and generates versatile output datasets in BED format. This pipeline can be implemented using simple shell scripts on any operating system, can be used downstream of most commonly used alignment algorithms, and is capable of detecting both “valleys” and “peaks” (Fig. 3-1). For paired-end datasets, forward and reverse reads from each sample are aligned to the reference genome together, retaining the mate-pair information in a BAM output file. The DNA fragments are reconstructed *in silico* using the beginning of the forward read and the end of the reverse read, and the fragment depth at each nucleotide is calculated; this step avoids counting overlapping reads twice. This step additionally incorporates a filter that restricts the maximum fragment length to filter out pairs which have aligned nonspecifically or improperly to the reference genome. The fragment depth files are then normalized such that the sum of the fragment depth over the entire genome is identical between datasets, allowing us to make comparisons between experimental replicates for which differing numbers of reads were obtained or in which insert sizes are variable. Next, the normalized fragment depth ratios between the experimental (chromatin-derived or MNAse-treated DNA) and control (purified DNA control) datasets are calculated, excluding regions which are not covered in the control dataset from further analysis. Regions of low or high coverage can be extracted from the normalized fragment depth ratio files, using filters for the minimum or maximum threshold and the minimum length of the region. Extracted regions will be returned in the flexible BED format. With BEDTools (29), we can not only identify overlapping or nonoverlapping regions between datasets but also can categorize these regions by transcription type, gene class, or epigenetic state. Additionally, BED files are supported by the versatile genome browser IGV (30,31), allowing visual depiction of loci of interest relative to other genomic features.

Generation and sequencing of MNAse-seq datasets

We used MNAse-seq to characterize genome-wide nucleosome density and nucleosome positioning in *L. major*. In addition to nuclease-treated chromatin, we included mechanically-sheared DNA (150-350 bp fragments) to assess regions performing poorly during sequencing and/or read alignment (Supplemental Fig. S3-1A), and MNAse-digested purified DNA to assess nuclease digestion bias (Supplemental Fig. S3-1B). Finally, we prepared MNAse-treated chromatin samples by digesting purified nuclei to a mononucleosome-sized fraction with MNAse (Supplemental Fig. S3-1C). Gel electrophoresis of this DNA showed predominantly ~146 bp bands expected for mononucleosomes (Supplemental Fig. S3-1C), as well as a lower level of sub-mononucleosome-sized fragments, but further examination of the alignments of reads derived from mononucleosomal and sub-mononucleosomal fractions showed no significant differences in two biological replicates (Supplemental Fig. S3-2). These preparations were subjected to 101 bp paired-end Illumina sequencing and the resulting reads were aligned to the *L.*

major FV1 reference genome (Supplemental Table S3-1). We analyzed two replicates each of MNAse-treated purified DNA and MNAse-treated chromatin using PAVED (Shaik et al., in preparation), designating the mechanically sheared DNA dataset as the "control" dataset. We found that this corrected for aneuploidy in addition to variations in copy number relative to the reference genome (Supplemental Fig. S3-3). While the spliced leader (SL) RNA array likely contains positioned nucleosomes and nuclease-hypersensitive sites, similar to that in *L. tarentolae* (32), we observed that the 3' end of the SL RNA gene was sequenced at very low levels in both replicates of the MNAse-treated chromatin and MNAse-treated DNA datasets, but was covered normally in the sheared genomic DNA datasets (Supplemental S3-4A). This bias was seen in BLAST analysis of the raw datasets as well, and is thus independent of the alignment methodology (Supplemental Fig. S3-4B). MNAse has a known bias toward A/T rich sequences (27), but the SL RNA array does not demonstrate an obvious overrepresentation of these sequences, and it is not clear whether these arose from MNAse bias or a technical issue during sequencing. For this reason we were unable to include in this analysis the SL RNA locus, which contains the only known RNA pol II promoter in *Leishmania*.

Standard MNAse-seq identifies NH sites at tRNA and rRNA genes but not in divergent SSRs

We focus first on tRNA and rRNA genes, which can be viewed as controls in these experiments: these genes have active, defined promoters, and the rRNA array is known to be nucleosome-depleted in *T. brucei* (33). We observed a marked decrease in the normalized fragment depth at tRNA genes in two biological replicates of MNAse-treated chromatin, which showed average normalized fragment depths of 0.15 and 0.33 respectively, compared to two replicates of MNAse-treated DNA, which showed average normalize fragment depths of 1.0 and

0.57 at tRNA genes (Fig. 3-2A, Supplemental Table S2). A more widespread decrease in normalized fragment depth was observed across rRNA gene array, which is annotated as 6 cistrons in the reference genome but likely contains close to 20 (34) (Fig. 3-2B). In *Leishmania*, the mature rRNAs are polycistronic but each cistron has its own promoter, and individual cistrons are separated by a 63-nt repetitive element (34). We observe similar levels of nucleosome depletion across the entire cistron (Fig. 3-2C). Quantitative analysis (Supplemental Table S2) showed that the normalized fragment depth in rRNA genes is much lower in both replicates of MNAse-treated chromatin (normalized fragment depth = 0.11, 0.16) than in MNAse-treated DNA (normalized fragment depth = 0.81, 0.36). In contrast, the normalized fragment depth values in both datasets are relatively similar in these 63 bp repeats between rRNA cistrons, suggesting these are nucleosome-bound (normalized fragment depth = 1.14, 1.09 for MNAse-treated DNA; 1.03, 0.95 for MNAse-treated chromatin; Supplemental Table S2). Thus, MNAse-seq shows nucleosome depletion at known RNAP I and RNAP III genes as expected.

We then examined the dSSR in chromosome 1, the only one where evidence of promoter function has been presented to date (5). The value of the control datasets is evident in that it is clear that poly(dG:dC) tracts are not represented at all, consistent with reports that homopolymeric sequences are a frequent source of errors in next generation sequencing platforms (35) (Fig. 3-2D). However, most of the dSSR is covered adequately, including those regions within the proposed ‘promoter region’ containing the putative transcription start sites and the regions bearing promoter-like epigenetic marks (Fig. 3-2D). In dSSRs, the normalized fragment depth are similar for MNAse-treated DNA (normalized fragment depth = 0.79, 0.57 for two biological replicates) and MNAse-treated chromatin (normalized fragment depth = 0.71, 0.67 for two technical replicates; Supplemental Table S2) and contrasts sharply from loci containing RNAP I or RNAP III promoters, as these effects are relatively small. This suggests that these loci are predominantly bound by nucleosomes, consistent with ChIP analysis of the core histone H3 (10).

Because poly(dG:dC) tracts were not assessed by next-generation sequencing experiments, we performed quantitative PCR (qPCR) comparing MNAse-treated chromatin and MNAse-treated DNA to undigested DNA, which also confirms the results above independently. We observed a high degree of correlation between next-generation sequencing data and qPCR for 18S rRNA genes, which appear highly depleted in MNAse-treated chromatin but not MNAse-treated DNA in our analysis described previously (Supplemental Fig. 3-6D). We similarly assayed 3 loci on chromosome 1: a gene located in the middle of a PGC that is not marked with histone modifications indicative of transcription initiation (*LmjF01.0400*; Supplemental Fig. 3-6C) , a locus within the dSSR which is associated with these epigenetic marks but is nucleosome-bound in our previous analysis (Supplemental Fig. 3-6B), and a locus spanning the two poly(dG:dC) tracts in the dSSR of chromosome 1 (Supplemental Fig. S3-6B). As expected, the loci within the dSSR show similar enrichment compared to the PGC-internal locus in the MNAse-treated DNA datasets. In the MNAse-treated chromatin samples, we observe significant but variable nucleosome density in loci in the dSSR, including across the poly(dG:dC) tracts (Supplemental Fig. S3-6A). This suggests that in contrast to poly(dA:dT) and poly(dG:dC) tracts in both budding and fission yeast (15,36,37), these regions do not explicitly exclude nucleosomes in *Leishmania*.

To identify NH sites globally, we identified regions with a low normalized fragment depth by comparison of MNAse-treated DNA and MNAse-treated chromatin and functionally annotated them as RNAP I-transcribed genes, RNAP III-transcribed genes, RNAP II-transcribed noncoding RNAs (ncRNAs, including snoRNAs), RNAP II-transcribed protein coding genes, and noncoding intergenic regions. We explored several parameters for normalized fragment depth and its length and chose a cutoff of a normalized fragment depth < 0.1 and a minimum length of 10 bp, as these parameters correctly identified tRNA and rRNA genes in both MNAse-treated chromatin replicates and excluded them in the MNAse-treated DNA replicates. Similar results were obtained using less-stringent thresholds (normalized fragment depth < 0.3, length 10 bp), but tRNA and rRNA genes were infrequently identified using a threshold of 0. We used comparisons of these regions to identify high-confidence NH sites, defined as ones present in both MNAse-treated chromatin replicates but neither MNAse-treated DNA replicate (Fig. 3-2E). Similarly, we identified “false positive” sites, defined as ones present in both MNAse-treated purified DNA replicates but neither MNAse-treated chromatin replicate (Fig. 3-2E). As anticipated based on the parameters used to define regions of interest showing low normalized fragment depth, a significant number of high-confidence NH sites were annotated as RNAPI- and RNAPIII-transcribed genes. In contrast, similar numbers of high-confidence NH sites and false positive sites were annotated as RNAPII-transcribed ncRNAs and protein-coding genes, dSSRs, and noncoding intergenic regions, suggesting that these were unlikely to be *bona fide* NH sites.

To test this statistically we randomly distributed the intervals from both groups across the genome using the BEDTools shuffleBed utility and annotated them as previously described (Supplemental Fig. S3-7A). We performed 1000 iterations to identify the mean and standard deviation of the number of intervals within each annotation category; as expected, the mean number of intervals within each category reflects the total percentage of the genome annotated by that category (Supplemental Fig. S3-7B). Using this theoretical random distribution, we calculated a Z-score quantitating how many standard deviations our observed values are from the mean of the theoretical random distribution. As expected, we observe an extremely large Z-score for RNAP I- and III-transcribed genes from the high-confidence NH site intervals but not from the false positive group (Supplemental Fig. S3-7B). In contrast, we observe that both the high-confidence NH sites and false positives show large but similar Z-scores for all other transcriptional categories, including dSSRs. Interestingly, in both groups we find many fewer NH sites than expected in ORFs, suggesting that the composition of these sequences may make them less susceptible to MNAse digestion. While these experiments reliably identified NH sites in genes with well-defined promoter elements, we failed to detect NH sites across poly(dG:dC) tracts and identify no loci which likely represent other nucleosome-disfavoring sequence elements in dSSRs.

Most nucleosomes are not positioned or phased in the L. major genome

A common feature of promoters, including RNAP II promoters, is their ability to confer positioning and spacing of nucleosomes in the adjacent regions, either through DNA-encoded properties or through the networks of chromatin remodelers associated with transcription (17,19,20,22). This allows a second test to be performed on putative promoters located within dSSRs. In this method, one infers the position of the mononucleosomes from sequencing data and then associates this with features of interest. Our datasets collectively provided information on 10 million nucleosomes, well within the number needed for the 32 megabase *Leishmania* genome as judged from other studies of nucleosome positioning and phasing (17–19). In the context of eukaryotic promoters, well-positioned nucleosomes are identified by the presence of multiple nucleosome-derived fragments at a distinct position. Using the fragment midpoint to define the position of individual nucleosomes, we observe that roughly one-third of nucleosomes (32.9%) are singletons that do not share a position with another nucleosome (Supplemental Table S3-4), indicating poor phasing of these nucleosomes. We identified potentially well-positioned nucleosomes, defined as at least three nucleosomes at one position, and compared their location relative to known promoters and dSSRs. We observed that tRNA genes that are not part of a tRNA cluster frequently show a well-positioned nucleosome at their 5' ends, demonstrating that we can identify positioned nucleosomes at known promoters (Fig. 3-3A). Interestingly, this positioning clearly varies amongst different tRNAs (Fig. 3-3A). Metagene analysis of tRNA genes bearing well-positioned nucleosomes shows that the RNAP III promoter, which is intragenic in kinetoplastid protozoa (38), and the tRNA TSS would not be obstructed by these positioned nucleosomes (Fig. 3-3B). We then performed a more thorough analysis of several regions of interest: dSSRs; peri-SSR ORFs and intergenic regions, which are within 5 kilobases of a dSSR; and PGC-internal ORFs and intergenic regions, which are greater than 5 kilobases from a dSSR (Supplemental Table S3-4). We find a similar percentage of potentially well-positioned nucleosomes across all of these loci (Supplemental Table S3-4), suggesting that this phenomenon may not be an indicator of transcriptionally-coordinated events.

Although we observed little indication of well-positioned nucleosomes across most of the genome, we sought to examine the spacing between nucleosomes as a function of genomic context. We extracted well-positioned nucleosomes and calculated the distance to the middle of the next nucleosome. Genome-wide analysis of this distance shows very few nucleosome pairs which show distances between 170 and 200 bp, the expectation for phased nucleosomes (Supplemental Fig. S3-8B). However, we do observe a weak 10-bp periodicity in nucleosome spacing at short distances (Supplemental Fig. S3-8C). This phenomenon has been observed in other systems [reviewed in (14)], and likely occurs as nucleosomes occupy overlapping positions in a population of cells according to energetically preferred DNA:histone contacts which occur along the 10.5 bp DNA helical repeat. This periodicity appears genome-wide in a variety of genomic contexts, including in dSSRs, peri-SSR ORFs and intergenic regions, and PGC-internal ORFs and intergenic regions (Supplemental Figs. S3-9A-E).

FAIRE-seq enriched loci correlate with activating histone marks in dSSRs and dSSR-proximal regions

FAIRE is a powerful technique which allows the identification of active regulatory elements in the absence of information about histone modification and/or histone variant incorporation. FAIRE signal is correlated with nucleosome occupancy (23), allowing the identification of NH sites which may be heterogeneous in a population. Because MNAse-seq failed to demonstrate the presence of NH sites in dSSRs in *Leishmania*, we hypothesized that a heterogeneous or transient population of NH sites may be present in and around dSSRs instead and sought to characterize the distribution of these loci by FAIRE-seq. We compared FAIRE-isolated DNA to a FAIRE input DNA using the previously described pipeline; in contrast to MNAse-seq, here we aim to identify regions with a high relative fragment depth.

We observed a high level of enrichment (50- to 100-fold over input DNA) for NH sites in rRNA genes, likely representing the maximum enrichment in our assays due to the high degree of nucleosome depletion in these loci (Fig. 3-4A; note the scale of the Y-axis). We observed a clear but lesser degree of enrichment at short NH sites in tRNA genes (Fig. 3-4B), a finding which likely stems from sonication of chromatin to fragment sizes which are longer than tRNA genes. In contrast to our standard MNAse-seq experiments, we find a robust enrichment of FAIRE signal (5- to 10-fold over input DNA) in and around dSSRs (see Figs. 3-4C and 3-4D for examples of two dSSRs). Although we observed a high degree of nucleosome density in these regions in our previous experiments, the FAIRE signal observed at these loci is much lower than in the rRNA array. Notably, similar degrees of enrichment are observed at putative internal transcription initiation regions (Fig. 3-4D), suggesting that these loci also maintain a transcriptionally-permissive chromatin environment. Interestingly, we observe several smaller, narrow FAIRE peaks which are located within PCGs but do not correlate with known peaks of acetylated H3 (red arrows in Figs. 3-4C and 3-4D); although these loci are unique sites in the genome, the significance of these peaks is not clear. To quantify the distribution of FAIRE peaks across various classes of genomic loci, we extracted all regions which were enriched at a level of at least 5-fold over the input DNA and annotated them according to their genomic location, with an added separation of peri-dSSR regions less than 5 kb from a dSSR from PGC-internal regions which are 5 kb or greater from a dSSR. We find that FAIRE-enriched loci are primarily located at dSSR-proximal loci and rRNA genes (Fig. 3-4E), and many fewer FAIRE-enriched loci were found within PGCs. Comparison of FAIRE peaks to randomly distributed intervals of similar length and number demonstrate a specific enrichment of FAIRE peaks within dSSR-proximal regions, including both peri-SSR genes and intergenic regions (Fig. 3-4E); a similar phenomenon was observed for both RNAPI and III-transcribed loci. A significant proportion of FAIRE-associated loci overlap with known regions of histone H3 acetylation (Fig. 3-4D), suggesting a functional link between these two phenomena.

As a correlate to MNAse-seq and FAIRE-seq, we used restriction endonuclease digestion of chromatin to characterize DNA accessibility (Fig. S3-10A). This technique facilitates the rapid quantitation of the relative nucleosome density in a population of cells. In this assay, restriction endonuclease sites which are bound by stable protein-DNA complexes such as nucleosomes will be protected from nuclease-catalyzed cleavage, while sites which are in linker regions or in open chromatin will be more accessible to the nuclease. The number of intact loci remaining in the population can be assessed by qPCR using PCR amplicons which span the restriction site; sites which are 100% protected will amplify as well as uncut DNA, while sites which are 100% unprotected will amplify as poorly as digested purified DNA. We subjected purified nuclei and an equivalent amount of purified DNA to digestion by a panel of restriction enzymes. Examination of several restriction sites across chromosomes 1 and 6 shows an intermediate level of restriction endonuclease susceptibility at PGC-internal, peri-dSSR, and SSR-internal loci, with values ranging from approximately 40-80% protected (Fig. S3-10B). Importantly, there are no differences between restriction sites in dSSRs, peri-dSSR, or mid-PCG regions, while a clear deprotection can be seen in the 18S rRNA gene, which contains many nuclease HS sites by standard MNAse-seq. Although this method does not discriminate between nucleosomes and other stable protein-DNA complexes, it is in good agreement with assumed nucleosome densities assessed by MNAse-seq, and correlates with known histone H3 ChIP patterns (10).

MNAse overdigestion reveals qualitative differences in nuclease susceptibility of dSSR-proximal nucleosomes

The presence of histone modifications and histone variants in transcription start site-proximal nucleosomes can strongly influence the stability of these nucleosome particles when chromatin is subjected to a higher degree of MNAse digestion (20). *Leishmania* possess modified histones in dSSR-proximal loci (10), although MNAse-seq and restriction endonuclease susceptibility assays demonstrate that nucleosome densities are relatively similar between dSSRs and PGCs. Thus, we sought to test whether dSSR-proximal nucleosomes exhibited altered stability by examining their susceptibility to MNAse overdigestion using purified nuclei. Interestingly, we observed that overdigestion of chromatin by MNAse generates a very high population of sub-mononucleosome-sized particles which accumulate at discrete fragment sizes, with bands appearing at 125 bp and 115 bp (Fig. S3-1D). In both replicates, the proportion of sub-mononucleosome-sized fragments was higher than the proportion of mononucleosome-sized fragments. Alignment of read pairs corresponding to mononucleosome- and each sub-mononucleosome-sized population again showed no differences in the distribution of alignments (Supplemental Fig. S3-2B).

We first performed a qualitative comparison of "standard" MNAse-seq and overdigested MNAse-seq. While the general density of nucleosomes appeared similar between experiments in most PGCs and in tRNA and rRNA genes, we observed broad nuclease overdigestion hypersensitive (NOH) regions in and around dSSRs in the overdigested MNAse-seq experiments (see Fig. 3-5A and 3-5B for examples of two dSSRs). Interestingly, long dSSRs which contain two peaks of acetylated H3 similarly show two regions of NOH in the overdigested MNAse-seq replicates (Fig. 3-5B). In addition, we observe NOH sites at putative internal transcription start sites, which are also marked with acetylated H3 and are enriched by FAIRE (note the head-to-tail blue arrows in Fig. 3-5B). Quantitation of the average normalized fragment depth according to genomic context demonstrates a significantly lower normalized fragment depth in dSSRs (0.36, 0.32) peri-SSR regions (0.71 and 0.68, ORFs; 0.38 and 0.44, intergenic) compared to PGC-internal loci (1.37 and 1.24, ORFs; 0.70 and 0.82, intergenic) (Supplemental Table 3-2).

After normalization to mechanically sheared purified DNA, we extracted regions with a relative fragment depth <0.1 and a length of 10 bp from these datasets to identify NOH loci. We then annotated these loci according to their genomic categorization, including classes for dSSR-proximal ORFs and intergenic regions. It is apparent that NOH sites are much more abundant in dSSRs and dSSR-proximal loci than in PCG-internal loci, and comparison to standard MNAse-seq and MNAse-treated purified DNA datasets suggest that this effect is specific to MNAse overdigestion (Fig. 3-5C). To understand the correlations between NOH sites, FAIRE-enriched loci, and known patterns of histone modification, we divided the genome into 1 kb windows and categorized each window according to whether it was categorized as a dSSR and whether it contained FAIRE-enriched loci, NOH sites, and acetylated histone H3. We then examined the degree of overlap between each of these datasets in both directions (i.e. overlap of dSSRs with FAIRE-enriched loci and the overlap of FAIRE-enriched loci with dSSRs; see Fig. 3-5D). We find that the majority of each dSSR overlap with FAIRE-enriched loci (61.9%) and patterns of acetylated H3 enrichment (69.3%), and most regions within a dSSR contain NOH sites (95.9%). Notably, the regions of dSSRs which are resistant to nuclease overdigestion appear to be regions between acetylated histone H3 peaks, which are known to contain the DNA modification base J in *L. major* (39). We find in the converse direction more modest enrichment of dSSRs in FAIRE-enriched loci (11.4%) and peaks of histone H3 acetylation (21.3%), which reflect enrichment of these loci in peri-SSR regions in addition to dSSRs. We find a significant overlap between FAIRE-enriched and acetyl-H3 enriched loci in these studies; a smaller percentage of FAIRE peaks overlap with acetyl-H3 peaks (25.4%) than acetyl-H3 peaks overlap with FAIRE peaks (47.6%), which may reflect the high degree of enrichment of FAIRE signal in SL RNA and rRNA genes. Importantly, we find that most FAIRE-enriched loci and regions of acetylated

H3 enrichment (99.6% and 95.5%, respectively) contain NOH sites, further validating the correlation between nuclease overdigestion hypersensitivity and markers of transcription initiation-associated chromatin.

Discussion

In this work we performed a thorough characterization of chromatin structure in *Leishmania* using MNAse-seq and FAIRE-seq and developed a novel, highly versatile bioinformatics platform to rigorously assess these datasets. This platform contains utilities to detect peaks and valleys in experimental datasets relative to an experimental control dataset, and it accurately corrects for variations in chromosome copy number. More importantly, it facilitates the filtering of loci that are not covered in the experimental control dataset due to technical challenges or other experimentally-induced artifacts, allowing one to reduce the likelihood of identifying false positive nuclease hypersensitive sites in MNAse-seq data. We demonstrate the versatility and utility of this software platform in our characterization of the chromatin landscape in *Leishmania major* and validate these observations using restriction endonuclease sensitivity assays and qPCR. We demonstrated that genomic loci transcribed by RNAP I, II, and III possess distinct chromatin structural characteristics which reflect known and unknown influences on their transcriptional regulation. First, we observed that rRNA genes and tRNA genes, transcribed by RNAP I and III respectively, appear to be predominantly nucleosome-depleted in actively transcribing *Leishmania* promastigotes. Individual rRNA gene arrays and all tRNA genes show a very low relative fragment depth in standard MNAse-seq experiments, and validation using restriction endonuclease sensitivity shows that a locus within the 28S rRNA gene is cleaved at a similar frequency as naked DNA in these assays. Systematic identification of NH sites demonstrates that low relative fragment depth sites in rRNA and tRNA genes represent *bona fide* NH sites, as they are not detected in MNAse-treated naked DNA controls. Although tRNA and rRNA genes behave similarly in MNAse-seq experiments, we find a much more robust enrichment of FAIRE signal at rRNA genes than tRNA genes, and we note that many tRNA genes possess a well-positioned nucleosome at their 5' ends. Notably, similar FAIRE enrichment levels were observed for rRNA genes in *T. brucei* as we observed in *Leishmania* (33).

In contrast to loci with known promoters, we find that dSSRs display a similar relative fragment depth compared to PGCs, a finding which was validated using restriction endonuclease sensitivity assays. Systematic identification of high-confidence NH sites present in two biological replicates of MNAse-digested chromatin identified NH sites in some, but not all dSSRs. Moreover, similar numbers of nonspecific NH sites were identified in dSSRs using MNAse-digested naked DNA datasets, and comparison to a theoretical random distribution suggests that these sites are distributed largely by chance or perhaps arise by MNAse sequence biases. Finally, although it was suggested that the presence and direction of poly(dG:dC) tracts could relate to promoter activity in dSSRs (4), we demonstrate using quantitative PCR of MNAse-digested chromatin that these loci are not nucleosome-free, suggesting that the nucleosome-disfavoring properties of this sequence do not manifest in these assays in *Leishmania*. Nucleosome positioning analysis demonstrates similar densities of nucleosomes, similar numbers of well-positioned nucleosomes, and similar midpoint-to-midpoint distances in dSSRs and in PGCs, leading us to conclude that the standard model of well-positioned nucleosomes flanking discrete TSS does not hold in kinetoplastid protozoa.

Because transcriptome-based data in *T. brucei* demonstrated the presence of heterogeneous TSS within an individual dSSR, as do more limited data for the *L. major* chromosome 1 dSSR (5,11,40,41), we also sought to identify heterogeneous NH sites using FAIRE-seq. In contrast to MNAse-seq, we observed robust FAIRE peaks associated with dSSRs which overlapped known patterns of histone modifications, indicating that heterogeneous or transient NH sites occur throughout regions which are likely to be transcriptionally permissive. Because we observed heterogeneous NH sites in dSSRs with no apparent alterations in nucleosome density in these regions, we sought to understand the physical properties of dSSR-proximal nucleosomes which may facilitate the development of NH sites. Examination of chromatin which was overdigested by MNAse reveals large regions of chromatin which contain nucleosomes which are susceptible to overdigestion but not standard digestion, reflecting a qualitative difference in nucleosome stability in these regions. Loci which contain these nucleosomes are also marked by acetylated histone H3, an indication of the functional relationship between the epigenetic marks and the stability of nucleosome particles. With this work, we demonstrate that poly(dG:dC) tracts behave differently in *Leishmania* than in other model systems, and do not disfavor the placement of nucleosomes. In addition, we were unable to identify any DNA-encoded elements which disfavor nucleosomes, and instead observe heterogeneous populations of NH sites which correlate with heterogeneous transcription start sites within dSSRs. These data support a model in which nucleosome instability, rather than nucleosome-disfavoring sequences, facilitate the development of transient nucleosome-depleted loci in regions of frequent transcription initiation.

Materials and Methods

Cell culture

All studies used derivatives of *Leishmania major* Friedlin V1 (MHOM/JL/81/Friedlin), grown at 26°C in M199 medium (US Biologicals) supplemented with 40 mM 4-(2-hydroxyethyl)-1-piperazineethanesuphonic acid (HEPES) pH 7.4 (Fisher Scientific), 100 uM adenine (Sigma), 1 µg mL^{-1} biotin (Sigma), 10 µg mL^{-1} hemin (Sigma), 2 µg mL^{-1} biopterin (Schircks Laboratories), 50 units/mL penicillin (Gibco), 50 µg/mL streptomycin (Gibco), and 10% (v/v) heat inactivated fetal calf serum (HyClone). Cell density was determined by using a model Z1 Coulter counter.

Micrococcal nuclease digestion

Logarithmic-phase cells were grown to a density of 2-4 x 10^6 cells/mL. Cells were collected, washed with Dulbecco's phosphate buffered saline (DPBS), and resuspended in ice-cold cell lysis buffer composed of 10 mM Tris-HCl, pH 7.4 (Sigma), 10 mM sodium chloride (Sigma), 3 mM magnesium chloride (Sigma), 0.5% Igepal-CA630 (Sigma), 0.15 mM spermine (Sigma), and 0.5 mM spermidine (Sigma). Nuclei were pelleted at 4°C and washed with ice-cold micrococcal nuclease digestion buffer composed of 10 mM Tris-HCl, pH 7.4, 15 mM sodium chloride, 60 mM potassium chloride (Sigma), 2 mM calcium chloride (Sigma), 0.15 mM spermine, and 0.5 mM spermidine. Nuclei were pelleted and resuspended at a density of 1.3 x 10^9 nuclei/mL in digestion buffer. Micrococcal nuclease (Sigma, cat. N3755) was added to a concentration of 1 unit per 3.3 x 10^7 nuclei and incubated at room temperature for 20 minutes. 1/25 volumes of micrococcal nuclease stop buffer composed of 100 mM EDTA (Sigma) and 10 mM EGTA (Sigma), pH 7.5 was added to stop the reaction. Nuclei were lysed with 1/10 volume of 20% sodium dodecyl sulfate (Sigma) and incubated overnight at 37°C with 750 µg/mL proteinase K (Sigma). The samples were extracted with an equal volume of phenol-chloroform-isoamyl alcohol (25:24:1, Sigma), and the aqueous phase was treated with 200 µg/mL RNAse A (Sigma) for 3 hours at 37°C. DNA was precipitated with 1/10 volumes 3M sodium acetate, pH 5.2 (Sigma) and 3 volumes of 100% ethanol (Pharmco-Aaper). The pellet was washed with 70% ethanol and was resuspended in TE buffer for library preparation or quantitative PCR. Overdigested MNAse samples were prepared identically, incubating for 25 minutes with MNAse.

MNAse-digested naked DNA was prepared by performing nuclear isolation, DNA extraction, RNAse A treatment, and purification as described above. Purified DNA was resuspended in micrococcal nuclease digestion buffer at a density of 2.2×10^8 nuclear equivalents/mL. 1 unit of MNAse was added to the reaction, and the reaction was incubated at room temperature. Aliquots of DNA were removed at 0, 15, 30, 60, 90, 120, 300, and 480 seconds and mixed with 1/25 volumes stop buffer. DNA was phenol extracted and precipitated as described previously. Samples were resolved on a 3% agarose gel, and gel fragments for MNAse-treated naked DNA controls were cut out as described in Supplementary S3-1. DNA was extracted using a Qiagen Gel Extraction kit according the manufacturer's instructions.

Formaldehyde-assisted isolation of regulatory elements (FAIRE) sample preparation

Logarithmic-phase cells were grown to a density of 2-4 x 10^6 cells/mL. FAIRE was performed according to published protocols with minor modifications (42). Cells were crosslinked for 15 minutes in 1% formaldehyde (Fisher), washed three times with ice-cold phosphate-buffered saline, and resuspended in lysis buffer B containing 1X Roche cOmplete protease inhibitor cocktail. Nuclei were pelleted and resuspended in lysis buffer A containing 1X Roche cOmplete protease inhibitor cocktail at a density of 3.3×10^8 nuclear equivalents/mL.

Nuclei were lysed with 10 strokes in a Dounce homogenizer, and 300 µL aliquots were transferred to 1.5 mL TPX tubes (Diagenode). Cells were sonicated at 4°C for 60 cycles of 30 seconds on (high), 30 seconds off, using a Bioruptor (Diagenode) with a circulating water bath. FAIRE and input samples were extracted and prepared as indicated (42).

Illumina library preparation

For preparation of sheared genomic DNA, 3 µg of purified *L. major* FV1 genomic DNA was diluted in 130 µL nuclease-free water (Ambion) and was sheared using a Covaris focused ultrasonicator using the following parameters: duty cycle 10%, intensity 5, cycles per burst 200, time 240 s, set mode frequency sweeping, and temperature 4°C. Sheared DNA was purified with AMPure XP beads (Beckman Coulter) according to the manufacturer's protocol. For sheared DNA and FAIRE samples, end repair was performed using T4 DNA polymerase (Enyzmatics) and Klenow DNA polymerase (Enzymatics), and ends were phosphorylated using T4 polynucleotide kinase (Enzymatics). DNA was purified using AMPure XP beads, and fragments were A-tailed using Klenow (3'-5' exo-) (Enzymatics). Libraries for paired end sequencing were prepared from all samples using T4 DNA ligase (Enzymatics) to add adapters containing the following sequences to the ends of the purified DNA fragments: adapter 1, (5' phosphate)-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC; adapter 2, ACACTCTTTCCCTACACGACGCTCTTCCGATCT. Fragments were purified with AMPure beads to remove adapter dimers. Index sequences were added by PCR using the following primers: primer 1.0 AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT; primer 2.0 GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT; and indexing primer CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTCAGACGTGTGC, where N designates the location of the unique 7-mer index sequence from each library. Libraries were validated using an Agilent 2100 Bioanalyzer, quantified using the Invitrogen Quant-iT HS DNA kit (Life Technologies), pooled in equimolar ratios, and subjected to 2x101 paired end sequencing using an Illumina HiSeq 2000.

Data analysis

Quality control on all datasets was initially performed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Paired-end data files were aligned to the *L. major* FV1 reference genome (TriTrypDB version 4.0) by NovoalignMPI (http://novocraft.com) using the following parameters: -o SAM; -r random; -l 30; -e 100; -H; -a AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCG ATCT GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCTTAGATCTCGTAT GCCGTCTTCTGCTTG. Datasets were processed to BAM format using Samtools import, sort, and index programs. Depth of coverage parameters were calculated using the BEDTools genomeCoverageBed tool (29). Java pipeline and BEDTools analysis was implemented using custom shell scripts, as described in the results section (Shaik et al., in preparation); genome-wide relative fragment depth totals were normalized to a mean of 1 for all MNAse-seq datasets. Midpoint analysis was performed using fragment length information extracted from the alignment files using custom scripts, and all comparisons were generated using custom shell scripts which implemented the BEDTools intersectBed and annotateBed functions (29). All scripts are available upon request.

Quantitative PCR validation of MNAse-seq data

MNAse-treated chromatin and naked DNA samples were quantified using the Invitrogen Qubit BR DNA kit (Life Technologies) and were diluted to a concentration of 1 ng/µL. PCR samples were prepared using the SYBR Green PCR Master Mix (Life Technologies), using 20 µL reaction volumes containing 0.4 pmol primers and 5 ng DNA. Primers sequences are as follows: Chr. 1 PolyG, AGAATGGCTGCATGACGAAC and ACACCTTCTGTGACCGATCT; Chr. 1 SSR, GCAAAGTGAACAGCATGTAGAA and CTGAGAAGTCTGCCTGAGTTT; 28S rRNA, TTTCTGCGTGCGTCTTCA and ATCCCGTTGGTTCAGTTTACA; LmjF01.0400, GTCGCATCTCGAGGCACGCAAGGTGATGTA and AAGGCGTAGAACGAGCCGGTGCA GCTG. Quantitative PCR was performed using an ABI Prism 7000 sequence detection system (Applied Biosystems), using an initial denaturation of 10 minutes at 95°C, followed by 45 cycles of 2-step qPCR (15 seconds at 95°C, 60 seconds at 60°C) and the dissociation curve. Thresholds and baselines were set manually and fold changes were calculated using the $2^{-\Delta\Delta Ct}$ method (43), normalizing to MNAse-treated genomic DNA.

Restriction endonuclease susceptibility assays

Logarithmic-phase cells were collected and washed twice with ice-cold phosphate-buffered saline. Cells were lysed with 10 mL per 2 x 10^8 cells of ice-cold lysis buffer (10 mM Tris-HCl, pH 7.4, 50 mM NaCl, 3 mM $MgCl_2$, 0.15 mM spermine, 0.5 mM spermidine, 0.5% Igepal CA-630; all reagents were supplied by Sigma-Aldrich) supplemented with 1X cOmplete protease inhibitor cocktail (Roche). Nuclei were collected and resuspended in 1 mL per 2x10^8 cells of digestion buffer (10 mM Tris-HCl, pH 7.4, 50 mM NaCl, 10 mM $MgCl_2$, 0.15 mM spermine, 0.5 mM spermidine, 0.2 mM EDTA, 0.2 mM EGTA, 5 mM 2-mercaptoethanol; all reagents were supplied by Sigma-Aldrich). Nuclei were divided into three 300 µL aliquots; one aliquot was digested using 50 units of BglII, KpnI, and SacII (New England Biolabs) for 1 hour at 37°C. After this digestion, 5 µL of 1.5 M Tris-HCl, pH 8.8 (Sigma-Aldrich) and 5 µL 200 mM EDTA were added to all three aliquots; nuclei were lysed with 15 µL 20% sodium dodecyl sulfate (Sigma-Aldrich) and incubated with 50 µg of proteinase K (Sigma-Aldrich) for 30 minutes at 55°C. Aliquots were extracted with equal volumes of phenol-chloroform-isoamyl alcohol , 25:24:1 (Sigma-Aldrich) and precipitated with one-tenth volumes of 3M sodium acetate (Sigma-Aldrich) and 3 volumes of 100% ethanol (Pharmco-AAPER). DNA was collected, and the second aliquot was resuspended in 300 µL digestion buffer and digested identically to the first aliquot. This aliquot was purified by ethanol precipitation as previously described, and DNA samples were quantified using the Invitrogen Qubit BR DNA kit. Quantitative PCR was performed as described in section 5.6, using primers which generate amplicons spanning the restriction sites. The primer pairs, internal restriction sites, and categorization in Fig. 3-3B are as follows: LmjF01.0315 (no restriction site, uncut control) CTCTCCACACGCGCAGAAT and CAGGCAAACGAGGAGCTCAT; LmjF01.0180 (KpnI, PGC-internal) CGGACCCTGTCGAG AAGCACATGCCCAC and CTACGCCTCTGGTGGCGGCATTGCAG; LmjF01.0220 downstream (KpnI, PGC-internal) ACGGCGGGATTCCGGCACGCAAG and TCCTTTGCGT CCCTCGGCGAGCTAGCGAG; Chr. 1 dSSR (KpnI, dSSR) GATCACATGGACGCAGTCGC ATCAGTAGATC and ATGGGCGGTTCGTCATGCAGCCATTCTTGC; Chr. 1 dSSR (BglII, dSSR) GCAAAGTGAACAGCATGTAGAA and CTGAGAAGTCTGCCTGAGTTT; LmjF06.0330 (KpnI, PGC-internal), CACGTTGGGCACAAGCCGCAATCCTTG and GAGGTGCACAAACTCACCACACGGATCG; LmjF06.0370 (KpnI, peri-SSR), GGCAAGCACGAGACGTCGAAGGTATCAG and AAGACGTGAGGTCACCAAGTAG

GGTC; LmjF06.0400 (KpnI, PGC-internal), GTCTTGCGGCCTGAGCGAGCTGCAGTC and TGCGCTCGTCCTTGCCCTTCCACGTCC; and Chr. 6 dSSR (SacII, dSSR), AAGCACGGACCATCCAATC and AATAAACGCGCTGAGGCA. LmjF01.0400 (KpnI, PGC-internal) and 28S rRNA (KpnI) primers are described in section 5.6.

Acknowledgments

The authors are grateful to Drs. Sarah Elgin and Gary Stormo at Washington University for their thoughtful comments. We thank Drs. Jenny Duncan and David Rudnick at Washington University in St. Louis for the use of their Bioruptor. We thank the Genome Technology Access Center in the Department of Genetics at Washington University School of Medicine for library preparation and sequencing. The Center is partially supported by NCI Cancer Center Support Grant P30 CA91842 to the Siteman Cancer Center and by ICTS/CTSA Grant UL1 TR000448 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), and NIH Roadmap for Medical Research. Computations were performed using the facilities of the Washington University Center for High Performance Computing, which were partially provided through grant NCRR 1S10RR022984-01A1. This work was funded by the National Institutes of Health (R0129646 to SMB, JSS, and BAA; T32 GM007067 to BAA) and the Washington University Berg-Morse and Schlesinger Graduate Fellowships to BAA. This publication is solely the responsibility of the authors and does not necessarily represent the official view of NCRR or NIH.

Figure Legends

Figure 3-1. Outline of PAVED bioinformatics pipeline. Paired-end or single-end sequencing platforms and alignment algorithms are flexible and defined by the user, and processing of alignment files to BAM file input can be accomplished with programs such as SAMTools. Paired-read fragments are reconstructed *in silico* using the beginning of the forward read and the end of the reverse reads; filters can be implemented to restrict the fragment length. Datasets are normalized such that the mean fragment depth ratio in the final pipeline is 1, and loci which are not covered in the input or control datasets are filtered during the fragment depth ratio calculation. Regions of interest are extracted based on minimum or maximum fragment depth ratio and the minimum length of the region. Comparisons and annotations can be accomplished using custom scripts or BEDTools.

Figure 3-2. Nuclease hypersensitive (NH) sites can be readily identified in tRNA and rRNA genes but not in divergent SSRs. Normalized fragment depth values were calculated relative to sheared genomic DNA for two replicates of MNAse-treated purified DNA and two replicates of MNAse-treated chromatin using PAVED as described previously (Shaik et al., in preparation). (A-E) Normalized fragment depth plots of MNAse-treated purified DNA and MNAse-treated chromatin. X-axis indicated physical distance on the chromosome in kilobase pairs, and scales are designated using double-edged arrows below panel. Y-axis indicates normalized fragment depth per base pair. Genes of interest are described with black arrows above panel, and other regions of interest are designated with red bars above panel. (A) tRNA-lysine gene on chr. 3; (B) all rRNA cistrons on chr. 27; (C) rRNA cistron 1 on chr. 27; (D) dSSR on chr. 1. Red arrow in (D) indicates the location of two poly(dG:dC) tracts which are sequenced by zero reads in all

sheared DNA, MNAse-treated DNA, and MNAse-treated chromatin datasets. (E) Representation of filters applied to define high confidence NH sites and false positive regions from MNAse-treated purified DNA and MNAse-treated chromatin replicates. Loci were extracted from the respective datasets using a maximum relative fragment depth of 0.1 and a minimum length of 10 bp. Comparisons of datasets were accomplished using the BEDTools intersectBed function. (F) High-confidence NH sites and false positives were annotated by transcription type and/or genomic context using the BEDTools annotate function; genomic categories were assigned using gene annotations from the *L. major* Friedlin TriTrypDB version 4.0 gff. Y-axis represents the number of loci identified per category. The black bars represent high-confidence NH sites identified from MNAse-treated chromatin; the gray bars represent false positive sites identified from MNAse-treated purified DNA.

Figure 3-3. Nucleosome positioning analysis identifies well-positioned nucleosomes upstream of tRNA genes. (A) Midpoints of nucleosome-sized fragments were calculated by extracting the fragment length and position of the forward read; individual midpoints were used to generate a WIG file for display in IGV. Individual tRNA genes are depicted (to scale) as blue arrows below each panel, and the 300 bp upstream of each tRNA gene start is depicted below the panel using double-edged arrows. Midpoint peaks are denoted with boxes above each peak. (B) Metagene analysis was accomplished by extracting all midpoints upstream of tRNA genes containing no RNAP III transcription units directly upstream of the 5’ end of the tRNA gene and calculating the distance of each midpoint from the tRNA start. Midpoints within 300 bp of the tRNA start were considered in this analysis, and midpoints were summed and averaged. Datapoints were binned into 25 bp intervals.

Figure 3-4. FAIRE-seq enriches dSSR-proximal loci and NH sites identified by MNAse-seq. (A-D) relative fragment depth values for FAIRE were calculated using input DNA as the normalization control; graphical plots for MNAse-treated chromatin and acetylated H3 enrichment are shown as described in Fig. 3-3. Acetylated H3 data was obtained from the Gene Expression Omnibus series GSE 13415, sample GSM338433; data was converted to WIG format for display in IGV. (A) rRNA gene array on chr. 27; (B) tRNA gene on chr. 3; (C-D) Chromosome-wide views of chr. 1 (C) and chr.5 (D). Red vertical arrows indicate sites which show modest enrichment by FAIRE and no enrichment for acetylated histone H3. Red and blue horizontal arrows indicate PCGs as described in Fig. 3-3. (E) FAIRE peaks were annotated as described in Fig. 3-3; additional categories were included to separate peri-SSR (within 5 kb of a dSSR) and PGC-internal (5 kb or greater from a dSSR) protein-coding genes and intergenic regions. The distribution of annotations for randomized genomic intervals over 1000 iterations was calculated using BEDTools shuffleBed and annotateBed features, and mean, standard deviation, and range of the expected distribution were calculated using custom shell scripts. Error bars for the expected (random distribution) data represent the standard deviation of 1000 iterations.

Figure 3-5. MNAse overdigestion reveals qualitative differences in dSSR-proximal nucleosomes. (A-B) relative fragment depth values for MNAse-treated naked DNA, MNAse-treated chromatin, and MNAse-overdigested chromatin were calculated and plotted as described in Fig. 3-3; acetylated H3 data is plotted as described in Fig. 3-4. Red and blue horizontal arrows indicate PGCs as described in Fig. 3-3. Data represent chromosome-scale views of

chromosomes 1 (A) and 5 (B). (C) NOH sites identified in both replicates of MNAse-overdigested chromatin were subjected to annotation as described in Fig. 3-4. The Y-axis indicates the total number of sites identified per category. The black bars represent MNAse-overdigested chromatin, the dark gray bars represent "standard" MNAse-treated chromatin, and the light gray bars represent MNAse-treated purified DNA. (F) Quantitation of dSSR, FAIRE, acetyl-H3, and NOH correlations genome-wide. The *L. major* genome was divided into 1 kb windows starting at the beginning of each chromosome. Windows were annotated using BEDTools according to whether they contained FAIRE peaks, acetyl-H3 peaks (Thomas *et al.*), or NOH sites; dSSRs were defined as the region between dSSR-proximal open reading frames. Overlaps between peaks or dSSRs were curated manually.

Supplemental Figure S3-1. Preparation and characterization of MNAse-treated and sheared DNA samples. (A-D) DNA preparations were run on 2% agarose gels using a low molecular weight DNA ladder. Mapped fragment lengths were extracted from BAM files using custom Java scripts. All plots indicate datapoints which were retained after filtering for fragment length. The X-axis of the graphs indicates the length from the beginning of the forward read to the end of its reverse mate; the Y-axis indicates the number of fragments with a particular length. (A) Mapped fragment lengths of mechanically sheared genomic DNA. (B) Agarose gel preparation and mapped fragment lengths of MNAse digestion series of purified genomic DNA. Size selection was performed by cutting out gel slices indicated in red (replicate 1) and blue (replicate 2) boxes. (C) Agarose gel preparation and mapped fragment lengths of MNAse-treated chromatin subjected to standard digestion conditions. (D) Agarose gel preparation and mapped fragment lengths of MNAse-treated chromatin subjected to MNAse overdigestion.

Supplemental Figure S3-2. Sub-mononucleosome-sized DNA fragments show similar size distributions**.** Paired reads which generated mononucleosome- and sub-mononucleosome-sized particles were extracted and re-aligned to the reference genome. Top panel indicates read pairs from MNAse-treated chromatin replicate 1, separated into fragment sizes of 100-115 bp, 120-135 bp, and 140-155 bp. Bottom panel indicates read pairs from MNAse-overdigested chromatin replicate 1, separated similarly. X-axis indicates the physical position on chromosome 1; Y-axis indicates the read depth at each position.

Supplemental Figure S3-3. Variations in somy and copy number errors in the reference genome are corrected by normalization to a control dataset. Comparison of fragment depths and relative fragment depth for mechanically sheared purified DNA and MNAse-treated purified DNA. Known tri- and pentasomic chromosomes and locations of repetitive gene families known to be misassembled in the reference genome are indicated with red arrows. The Y-axis for the top panel indicates gene density in the chromosome; the Y-axis for the middle two panels indicates coverage depth; and the Y-axis for the bottom panel indicates relative fragment depth.

Supplemental Figure S3-4. SL-derived sequences are absent from raw data for MNAse-treated DNA and chromatin datasets. (A) Top panel indicates read pairs from MNAse-treated naked DNA (mean fragment size 171 bp); bottom panel indicates read pairs from sheared genomic DNA (mean fragment size 204 bp). Red arrows indicate forward reads, and blue indicates the mate pair of that read. Light pink and blue reads indicate that these reads do not map uniquely to the reference genome. Horizontal lines connecting paired reads indicate the distance between the reads. Blue arrows at the bottom indicate SL RNA genes. (B) SL RNA-derived sequences were used as BLAST queries for databases generated from the raw data derived from forward and reverse reads for one replicate of MNAse-treated naked DNA. The number of copies assigned to the reference genome and the number of BLAST hits per query are depicted in this table.

Supplemental Figure S3-5. Normalized fragment depth histograms for MNAse-seq datasets. Histogram-level data including the relative fragment depth and the number of loci displaying a given relative fragment depth was collected at each base pair in the genome. Data were collapsed into bins of 0.01 width. The Y axis indicates the number of base pairs in the genome with a given relative fragment depth; the X-axis indicates the relative fragment depth bin.

Supplemental Figure S3-6. Quantitative PCR validation of standard MNAse-seq data. Upper panel describes quantitative PCR analysis of MNAse-digested chromatin and MNAse-treated DNA replicates relative to undigested DNA. PCR amplicons relative to relative fragment depth ratios are described in the lower panels; in lower panels, Y-axis indicates the relative fragment depth values for each dataset relative to sheared genomic DNA, and the X-axis indicates the physical position on chromosome 1 (LmjF01.0400 and dSSR), or chromosome 27 (18S rRNA). Position on the chromosome is indicated at the top of the plot.

Supplemental Figure S3-7. Generation and quantitation of randomly distributed genomic intervals. (A) Flow chart demonstrating the experimental details of randomly distributed genomic intervals and their annotation. (B) Description of observed annotations of high-confidence and false-positive NH sites and the distribution of annotations for randomized genomic intervals over 1000 iterations. Mean, standard deviation, and range of the expected distribution were calculated using BEDTools shuffleBed and annotateBed features, implemented by custom shell scripts. Z scores were calculated using the formula [(observed-mean)/(standard deviation)].

Supplemental Figure S3-8. Genome-wide nucleosome positioning and spacing analysis. (A) Nucleosome midpoints were calculated as described in Fig. 3-3. The number of midpoints at each base pair was quantified and plotted as a histogram. X-axis represents the number of midpoints per base pair. Y-axis represents the number of times a given midpoints per base pair occurred across the entire genome. (B-C) Midpoint-to-midpoint distances were calculated for all adjacent midpoint maxima, which are defined as 3 or more midpoints at a given base pair. The distances are plotted as a histogram; X-axis represents the midpoint-to-midpoint distance, and Y-axis indicates the number of occurrences of that midpoint-to-midpoint distance across the entire genome. (B) All midpoint maxima are plotted. (C) Only midpoint-to-midpoint distances between 1 and 200 bp are shown.

Supplemental Figure S3-9. Nucleosome distance analysis for regions of interest. Midpoints were extracted as described in Fig. 3-3 and were converted into BED format. BEDTools intersectBed was used to extract midpoints falling within the designated genomic categories: (A) divergent SSRs; (B) peri-SSR open reading frames (ORFs), designated as loci within 5 kb of a dSSR; (C) peri-SSR intergenic regions, designated as interORF regions within 5 kb of a dSSR; (D) PGC-internal ORFs, which are 5 kb or greater from a dSSR; (E) PGC-internal intergenic regions, designated as inter ORF regions 5 kb or greater from a dSSR. For the last midpoint in an interval, the next midpoint maximum, which was outside of that region, was used for distance calculations.

Supplemental Figure S3-10. Restriction endonuclease sensitivity assays demonstrate similar nuclease sensitivities in PGC-internal loci and dSSRs. (A) Graphical representation of restriction endonuclease susceptibility assay. Purple circles indicate nucleosomes, and green X marks represent restriction endonuclease cleavage sites. (B) Quantitative PCR analysis of restriction endonuclease susceptibility using loci on chr. 1 and 6 and the 28S rRNA genes located on chr. 27. Loci are categorized as PGC-internal (greater than 5 kb from a dSSR), peri-dSSR (within 5 kb of a dSSR) or dSSR. Equivalent amounts of purified DNA or chromatin were subjected to restriction endonuclease overdigestion for 1 hour. Quantitative PCR amplicons span individual restriction sites. Dark gray bars represent restriction endonuclease-digested chromatin and light gray bars represent restriction endonuclease-digested purified DNA. Data are normalized to uncut purified DNA. Error bars represent the average of three biological replicates.

Supplemental Table S3-1. Alignment metrics for Illumina sequencing data. Percent alignment was calculated during alignment by Novoalign. Mean, median, and quartile depth of coverage metrics were calculated from BAM files using BEDTools genomeCoverageBed.

Supplemental Table S3-2. Normalized fragment depth metrics according to genomic context. Densities of regions of interest were calculated using the TrackNRest feature in PAVED (Shaik et al., in preparation). Regions of interest are described in Supplemental Fig. S3-9; divergent

SSRs and convergent SSRs are subdivided length or the presence of RNAP III-transcribed genes as noted in the table. H3Ac intervals are derived from Thomas, et al. (10).

Supplemental Table S3-3. Comparison of relative fragment depth maximum thresholds for identification of NH sites. MNAse-treated chromatin and MNAse-treated DNA replicates were analyzed as described in Fig. 3-2, and regions of low coverage were extracted using a maximum relative fragment depth threshold of 0, 0.1, or 0.3 and a minimum length of 10 bp. Values represent the total number of base pairs as a percentage of the total number of base pairs in the reference genome.

Supplemental Table S3-4. Comparison of well-positioned nucleosomes within regions of interest. Regions of interest are defined as in Supplemental Fig. S3-9. Well-positioned nucleosomes are calculated by using the number of positions containing at least 3 midpoints divided by the total number of midpoints in the region.

References

1. Gunzl A, Vanhamme L, Myler P. Transcription in trypanosomes: a different means to the end. In: Barry J, Mottram J, McCulloch R, Acosta-Serrano A, editors. Trypanosomes - After the Genome. Horizon Bioscience, Wymondham, Norfolk, UK; 2007. p. 177–208.

2. Matthews K, Tschudi C, Ullu E. A common pyrimidine-rich motif governs *trans*-splicing and polyadenylation of tubulin polycistronic pre-mRNA in trypanosomes. Genes Dev 1994;8(4):491–501.

3. LeBowitz J, Smith H, Rusche L, Beverley S. Coupling of poly(A) site selection and *trans*-splicing in *Leishmania*. Genes Dev 1993;7(6):996–1007.

4. Siegel T, Hekstra D, Kemp L, Figueiredo L, Lowell J, Fenyo D, et al. Four histone variants mark the boundaries of polycistronic transcription units in *Trypanosoma brucei*. Genes Dev 2009;23(9):1063–76.

5. Martínez-Calvillo S, Yan S, Nguyen D, Fox M, Stuart K, Myler P. Transcription of *Leishmania major* Friedlin chromosome 1 initiates in both directions within a single region. Mol Cell 2003;11(5):1291–9.

6. Martínez-Calvillo S, Nguyen D, Stuart K, Myler P. Transcription initiation and termination on *Leishmania major* chromosome 3. Eukaryot Cell 2004;3(2):506–17.

7. Ivens A, Peacock C, Worthey E, Murphy L, Aggarwal G, Berriman M, et al. The genome of the kinetoplastid parasite, *Leishmania major*. Science. 2005;309(5733):436–42.

8. Wright J, Siegel T, Cross G. Histone H3 trimethylated at lysine 4 is enriched at probable transcription start sites in *Trypanosoma brucei*. Mol Biochem Parasitol 2010;172(2):141–4.

9. Respuela P, Ferella M, Rada-Iglesias A, Aslund L. Histone acetylation and methylation at sites initiating divergent polycistronic transcription in *Trypanosoma cruzi*. J Biol Chem 2008;283(23):15884–92.

10. Thomas S, Green A, Sturm N, Campbell D, Myler P. Histone acetylations mark origins of polycistronic transcription in *Leishmania major*. BMC Genomics. 2009;10:152.

11. Kolev N, Franklin J, Carmi S, Shi H, Michaeli S, Tschudi C. The transcriptome of the human pathogen *Trypanosoma brucei* at single-nucleotide resolution. PLoS Pathog 2010;6(9):e1001090.

12. Ruan J, Arhin G, Ullu E, Tschudi C. Functional characterization of a *Trypanosoma brucei* TATA-binding protein-related factor points to a universal regulator of transcription in trypanosomes. Mol Cell Biol 2004;24(21):9610–8.

13. Segal E, Widom J. Poly(dA:dT) tracts: major determinants of nucleosome organization. Curr Opin Struct Biol 2009;19(1):65–71.

14. Struhl K, Segal E. Determinants of nucleosome positioning. Nat Struct Mol Biol 2013;20(3):267–73.

15. Raveh-Sadka T, Levo M, Shabi U, Shany B, Keren L, Lotan-Pompan M, et al. Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. Nat Genet 2012;44(7):743–50.

16. Iyer V, Struhl K. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. EMBO J 1995;14(11):2570–9.

17. Gaffney D, McVicker G, Pai A, Fondufe-Mittendorf Y, Lewellen N, Michelini K, et al. Controls of nucleosome positioning in the human genome. PLoS Genet 2012;8(11):e1003036.

18. Brogaard K, Xi L, Wang J, Widom J. A map of nucleosome positions in yeast at base-pair resolution. Nature 2012;486:496–501.

19. Valouev A, Johnson S, Boyd S, Smith C, Fire A, Sidow A. Determinants of nucleosome organization in primary human cells. Nature 2011;474:516–20.

20. Weiner A, Hughes A, Yassour M, Rando O, Friedman N. High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. Genome Res 2010;20(1):90–100.

21. Zaret K. Micrococcal nuclease analysis of chromatin structure. Curr Prot Mol Biol 2005;Chapter 21:Unit 21.1.

22. Lee W, Tillo D, Bray N, Morse R, Davis R, Hughes T, et al. A high-resolution atlas of nucleosome occupancy in yeast. Nat Genet 2007;39(10):1235–44.

23. Giresi P, Kim J, McDaniell R, Iyer V, Lieb J. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res 2007;17(6):877–85.

24. Sterkers Y, Lachaud L, Crobu L, Bastien P, Pagès M. FISH analysis reveals aneuploidy and continual generation of chromosomal mosaicism in *Leishmania major*. Cell Microbiol 2011;13(2):274–83.

25. Sterkers Y, Lachaud L, Bourgeois N, Crobu L, Bastien P, Pagès M. Novel insights into genome plasticity in eukaryotes: mosaic aneuploidy in *Leishmania*. Mol Microbiol 2012;86(1):15–23.

26. Mannaert A, Downing T, Imamura H, Dujardin J. Adaptive mechanisms in pathogens: universal aneuploidy in *Leishmania*. Trends Parasitol 2012;28(9):370–6.

27. Chung H, Dunkel I, Heise F, Linke C, Krobitsch S, Ehrenhofer-Murray A, et al. The effect of micrococcal nuclease digestion on nucleosome positioning data. PLoS One 2010;5(12):e15754.

28. Ross M, Russ C, Costello M, Hollinger A, Lennon N, Hegarty R, et al. Characterizing and measuring bias in sequence data. Genome Biol. 2013;14(5):R51.

29. Quinlan A, Hall I. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26(6):841–2.

30. Robinson J, Thorvaldsdottir H, Winckler W, Guttman M, Lander E, Getz G, et al. Integrative Genomics Viewer. Nat Biotechnol 2011;29(1):24–6.

31. Thorvaldsdottir H, Robinson J, Mesirov J. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 2013;14(2):178–92.

32. Hitchcock R, Thomas S, Campbell D, Sturm N. The promoter and transcribed regions of the *Leishmania tarentolae* spliced leader RNA gene array are devoid of nucleosomes. BMC Microbiol 2007;7:44.

33. Figueiredo L, Cross G. Nucleosomes are depleted at the VSG expression site transcribed by RNA polymerase I in African trypanosomes. Eukaryot Cell 2010;9(1):148–54.

34. Martínez-Calvillo S, Sunkin S, Yan S, Fox M, Stuart K, Myler P. Genomic organization and functional characterization of the *Leishmania major* Friedlin ribosomal RNA gene locus. Mol Biochem Parasitol 2001;116(2):147–57.

35. Minoche A, Dohm J, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol 2011;12(11):R112.

36. Struhl K. Naturally occurring poly(dA-dT) sequences are upstream promoter elements for constitutive transcription in yeast. Proc Natl Acad Sci USA 1985;82(24):8419–23.

37. Moyle-Heyrman G, Zaichuk T, Xi L, Zhang Q, Uhlenbeck O. Chemical map of *Schizosaccharomyces pombe* reveals species-specific features in nucleosome positioning. Proc Natl Acad Sci USA 2013;110(50):20158–63.

38. Nakaar V, Dare A, Hong D, Ullu E, Tschudi C. Upstream tRNA genes are essential for expression of small nuclear and cytoplasmic RNA genes in trypanosomes. Mol Cell Biol 1994;14(10):6736–42.

39. Van Luenen H, Farris C, Jan S, Genest P, Tripathi P, Velds A, et al. Glucosylated hydroxymethyluracil, DNA base J, prevents transcriptional readthrough in *Leishmania*. Cell 2012;150(5):909–21.

40. Siegel TN, Hekstra DR, Wang X, Dewell S, Cross G a M. Genome-wide analysis of mRNA abundance in two life-cycle stages of *Trypanosoma brucei* and identification of splicing and polyadenylation sites. Nucleic Acids Res 2010;38(15):4946–57.

41. Nilsson D, Gunasekera K, Mani J, Osteras M, Farinelli L, Baerlocher L, et al. Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of *Trypanosoma brucei*. PLoS Pathog 2010;6(8):e1001037.

42. Simon J, Giresi P, Davis I, Lieb J. Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. Nat Protoc 2012;7(2):256–67.

43. Livak K, Schmittgen T. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 2001;25(4):402–8.

Figure 3-1. Outline of PAVED bioinformatics pipeline.

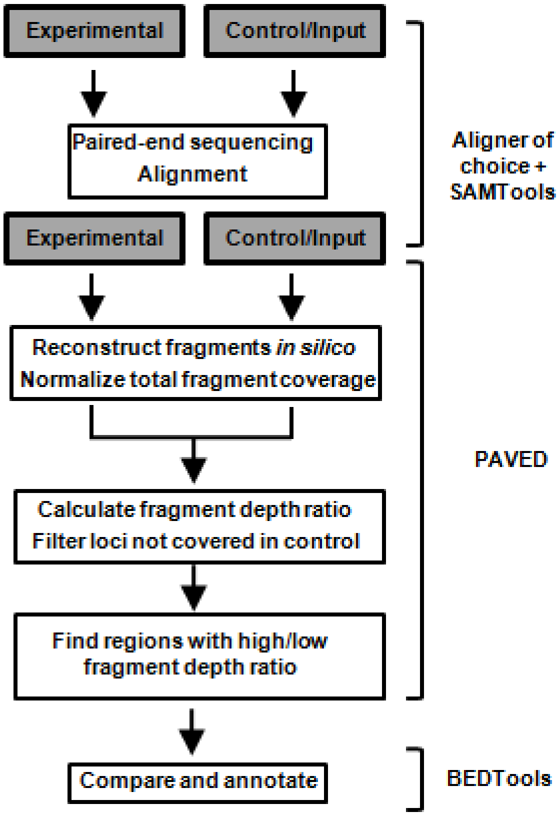

Experimental
Control/Input
Paired-end sequencing
Alignment
Experimental
Control/Input
Aligner of choice + SAMTools
Reconstruct fragments *in silico*
Normalize total fragment coverage
Calculate fragment depth ratio
Filter loci not covered in control
Find regions with high/low fragment depth ratio
PAVED
Compare and annotate
BEDTools

Figure 3-2. Nuclease hypersensitive (NH) sites can be readily identified in tRNA and rRNA genes, but not in divergent SSRs.

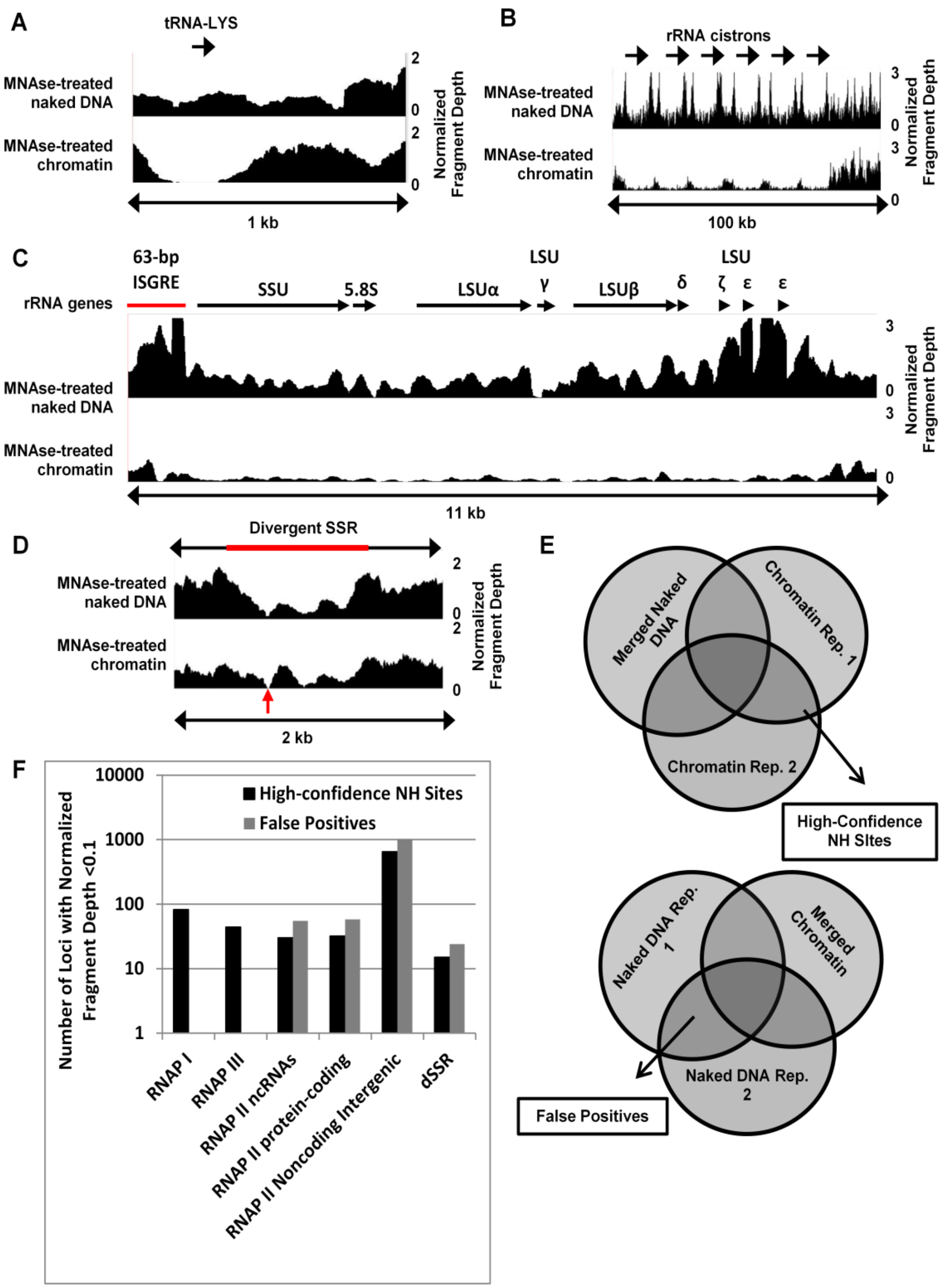


Figure 3-3. Nucleosome positioning analysis reveals well-positioned nucleosomes upstream of tRNA genes.

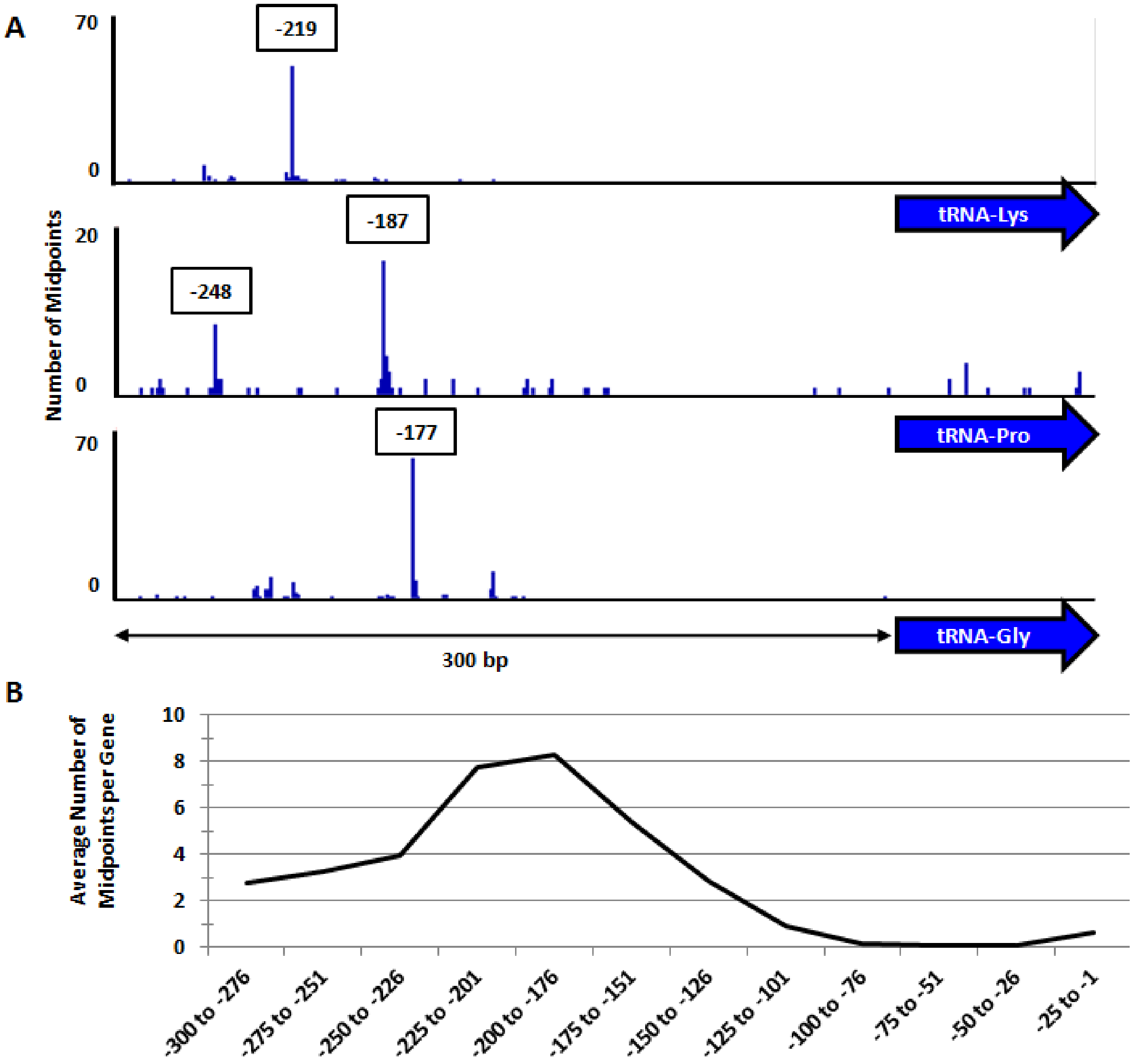


Figure 3-4. FAIRE-seq enriches dSSR-proximal loci and NH sites identified by MNAse-seq.

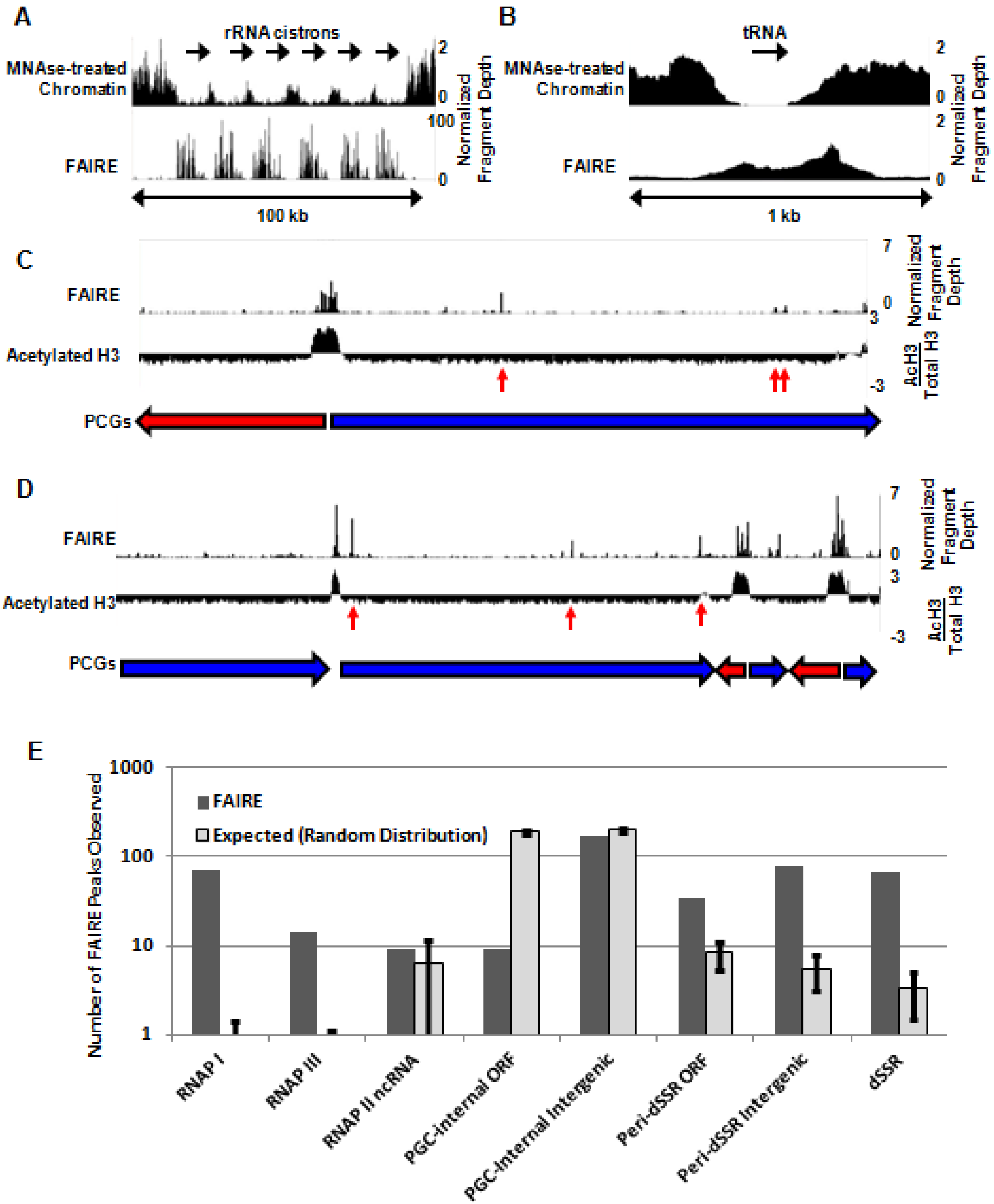
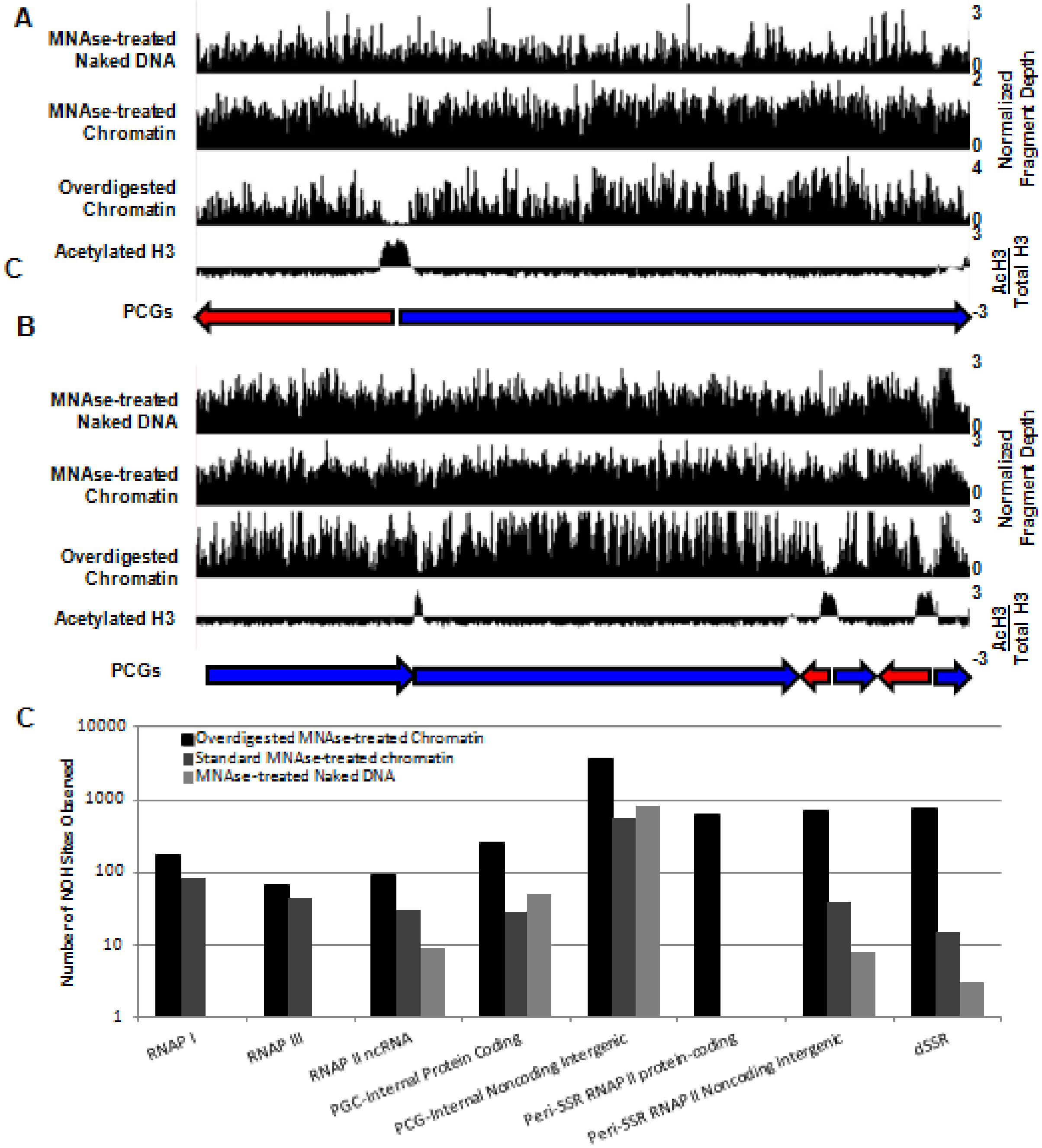


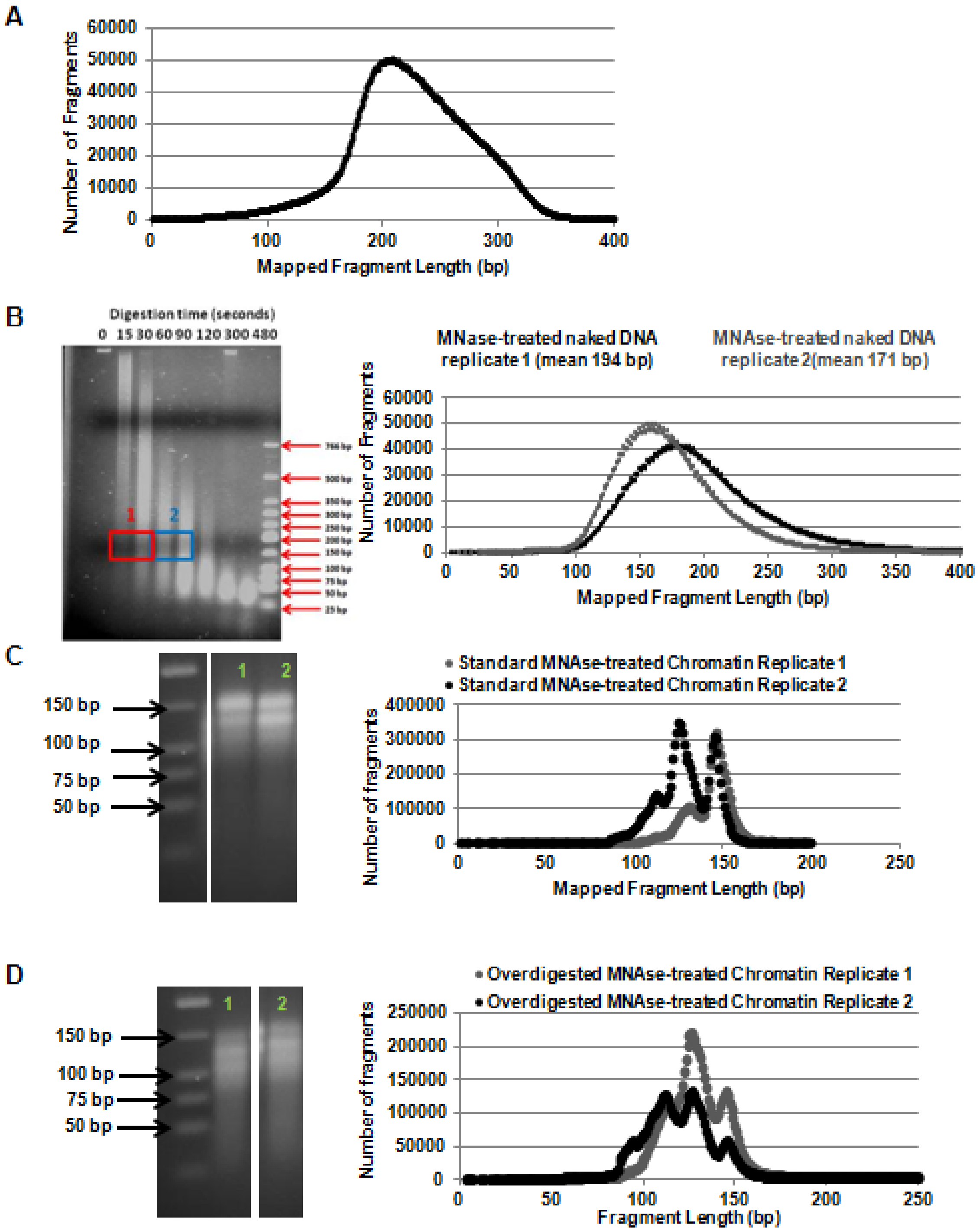
Figure 3-5. MNAse overdigestion reveals qualitative differences in dSSR-proximal nucleosomes.



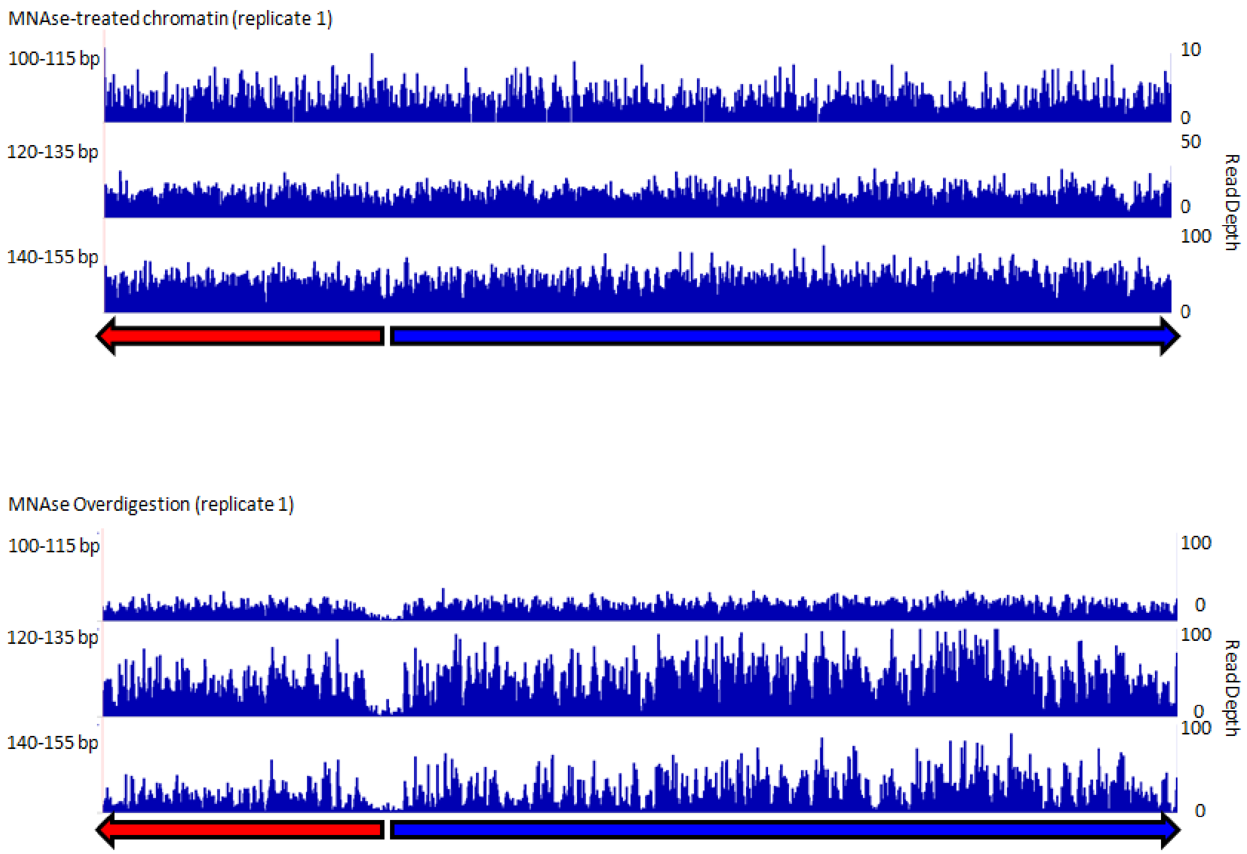
D

	Overlap with dSSR	Overlap with FAIRE	Overlap with Acetyl-H3	Overlap with NOH	Total Windows Containing Feature
dSSR		226 (61.9%)	253 (69.3%)	350 (95.9%)	365
FAIRE	226 (11.4%)		505 (25.4%)	1979 (99.6%)	1987
Acetyl-H3	253 (21.3%)	505 (47.6%)		1011 (95.5%)	1059
NOH	350 (1.13%)	1979 (9.9%)	1011 (5.1%)		19950

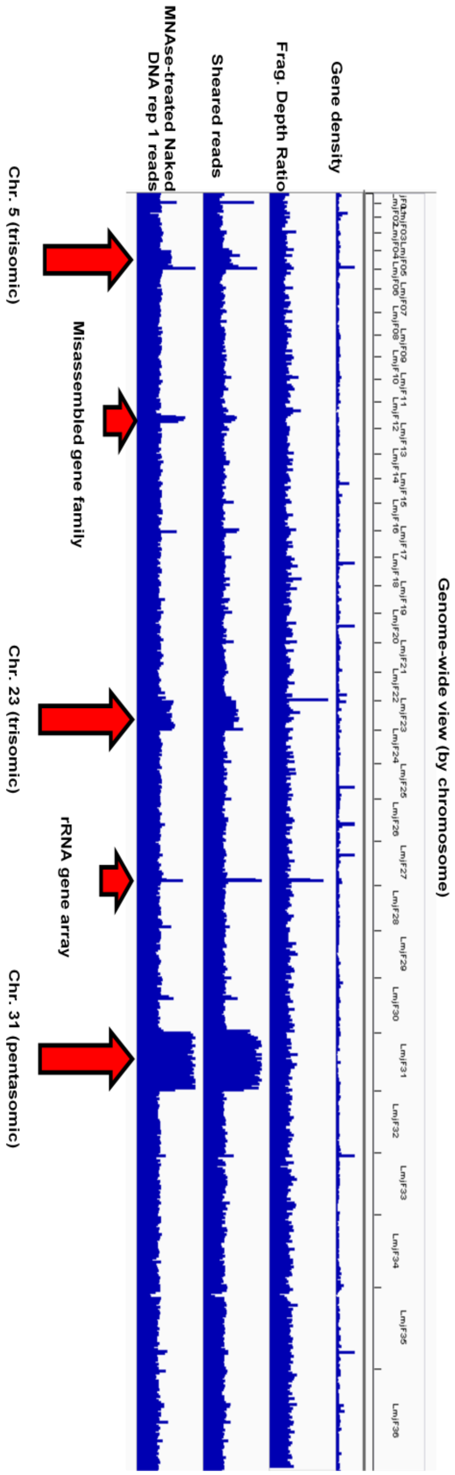
Supplemental Figure S3-1. Preparation and characterization of MNAse-treated and sheared DNA samples.



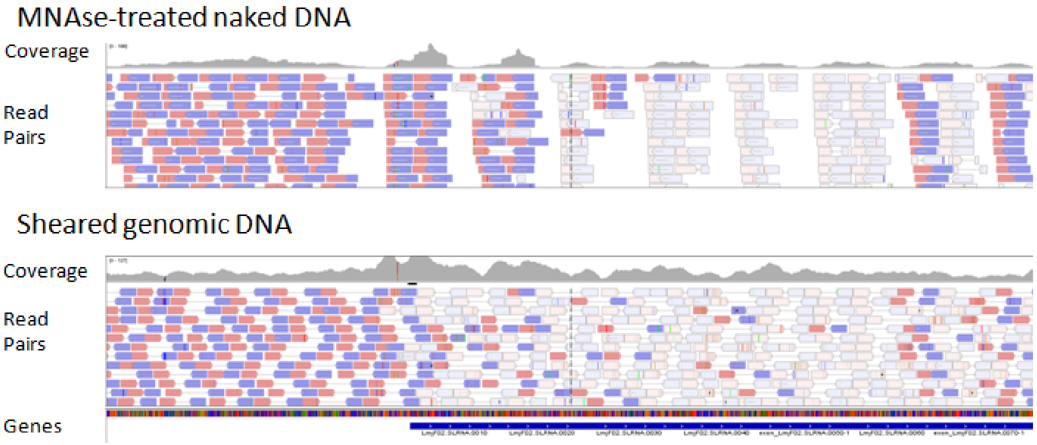
Supplemental Figure S3-2. Sub-mononucleosome-sized DNA fragments show similar size distributions.



Supplemental Figure S3-3. Variations in somy and copy number errors in the reference genome are corrected by normalization to a control dataset.



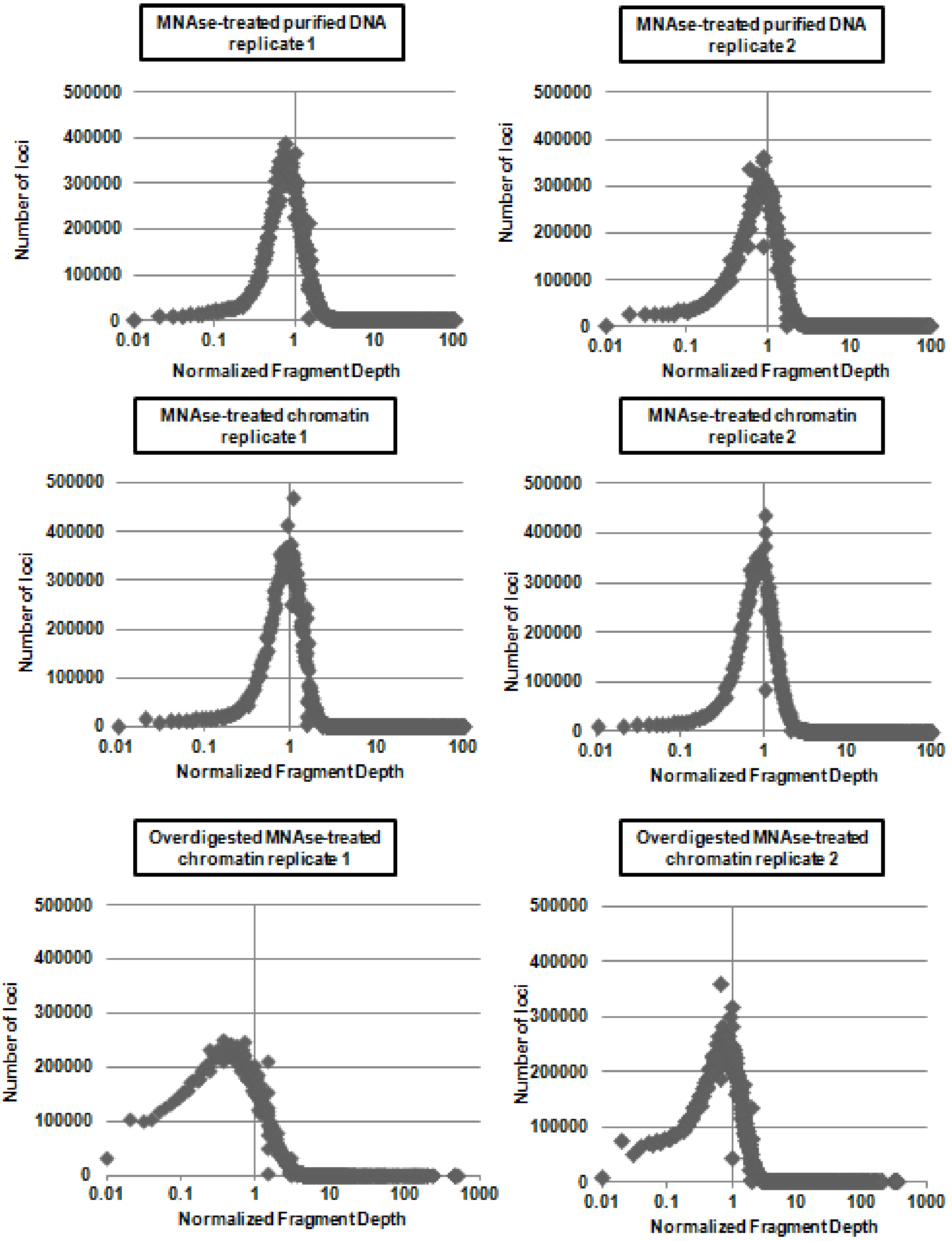
Supplemental Figure S3-4. SL-derived sequences are absent from raw data for MNAse-treated DNA and chromatin datasets.



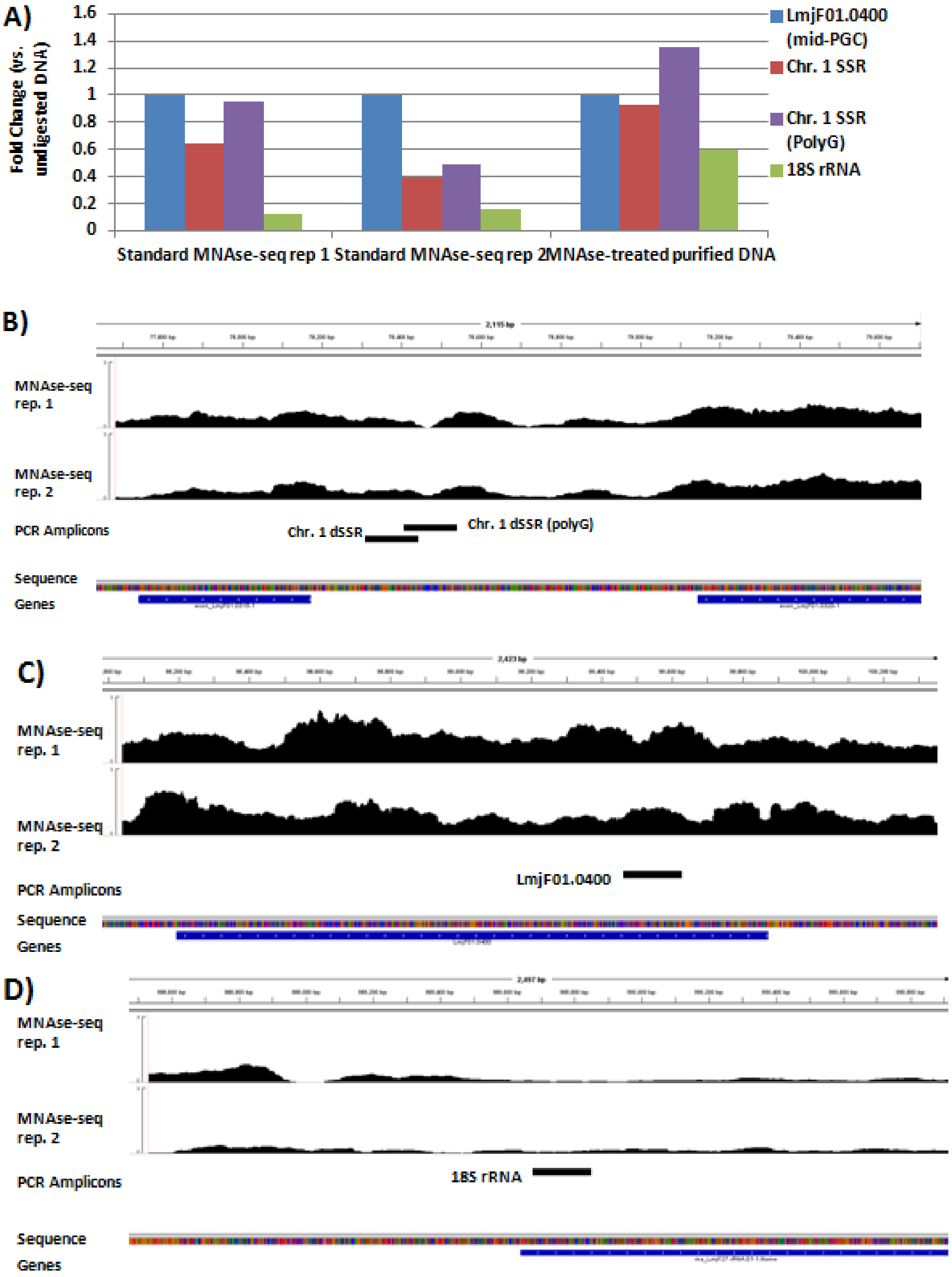
B)

BLAST Query	Copies in reference genome	BLAST Hits (MNAse-treated naked DNA, forward)	BLAST Hits (MNAse-treated naked DNA, reverse)
SL RNA-proximal (chr. 2: 269,061-269,160)	1	68	55
SL RNA 5' (chr. 2: 269,732-269,832)	62	932	1079
SL RNA Middle (chr. 2: 269,833-269,932)	62	1196	1477
SL RNA 3' (chr. 2: 270,021-270,112)	62	142	95

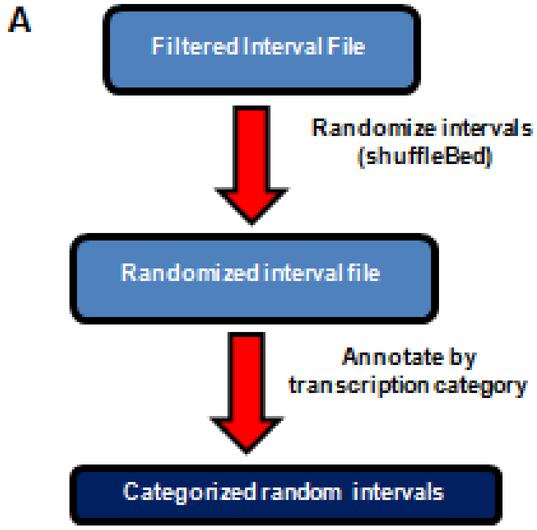
Supplemental Figure S3-5. Normalized fragment depth histograms for MNAse-seq datasets.



Supplemental Figure S3-6. Quantitative PCR validation of standard MNAse-seq data.



Supplemental Figure S3-7. Generation and quantitation of randomly distributed genomic intervals.



B

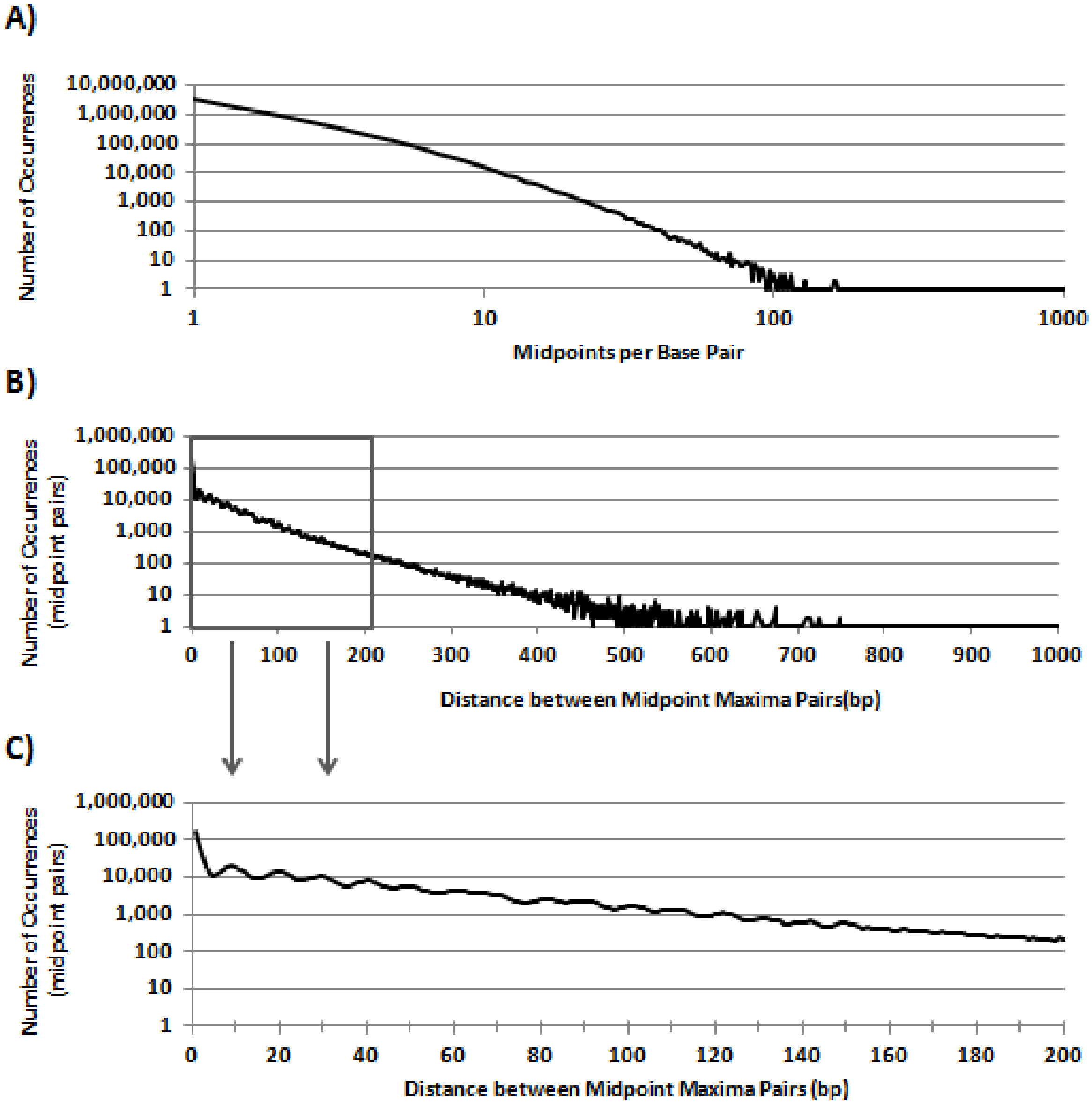
High-confidence NH sites

	Observed	Shuffled Mean	Shuffled St.Dev	Shuffled Range	Z score
Pol I	82	0.961	0.995	0-5	81.45
Pol II ncRNAs	30	5.521	3.25	0-18	7.53
ORFs	32	371.96	14.03	334-420	-24.23
Pol III	44	0.243	0.512	0-3	85.46
Divergent SSRs	15	6.545	2.55	0-16	3.32
Intergenic	648	385.223	14.27	335-426	18.41

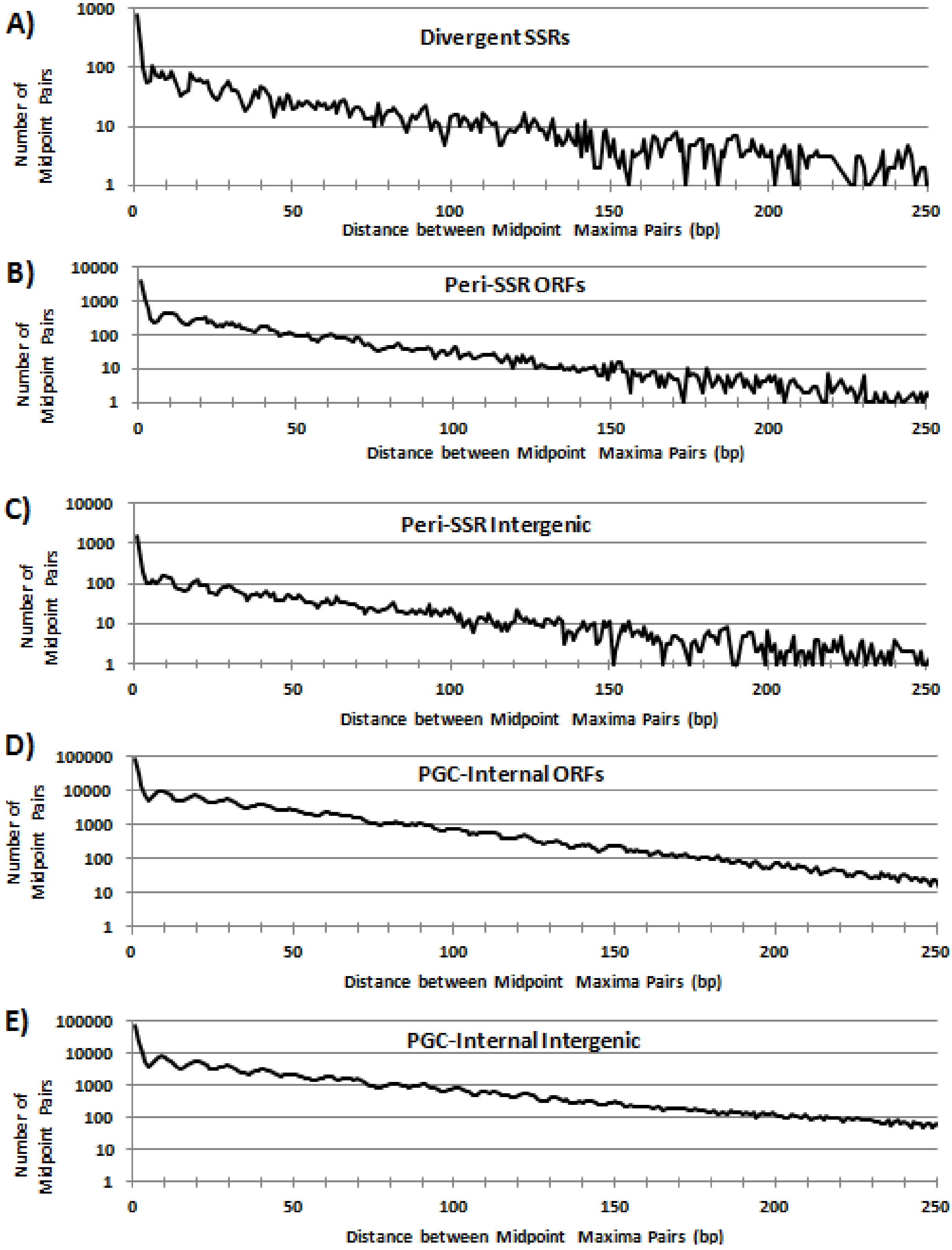
False-positive NH sites

	Observed	Shuffled Mean	Shuffled St.Dev	Shuffled Range	Z score
Pol I	0	1.357	1.2	0-3	-1.13
Pol II ncRNAs	55	8.824	4.07	0-10	11.35
ORFs	58	573.528	17.4	106-157	-29.63
Pol III	0	0.331	0.63	0-2	-0.53
Divergent SSRs	24	10.352	3.21	0-8	4.25
Intergenic	1030	594.656	17.46	105-158	24.93

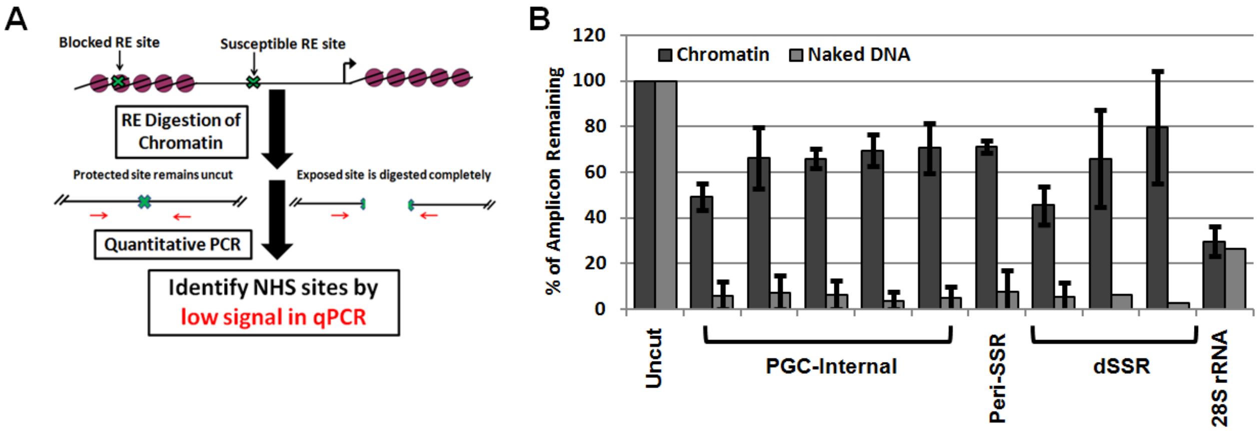
Supplemental Figure S3-8. Genome-wide nucleosome positioning and spacing.



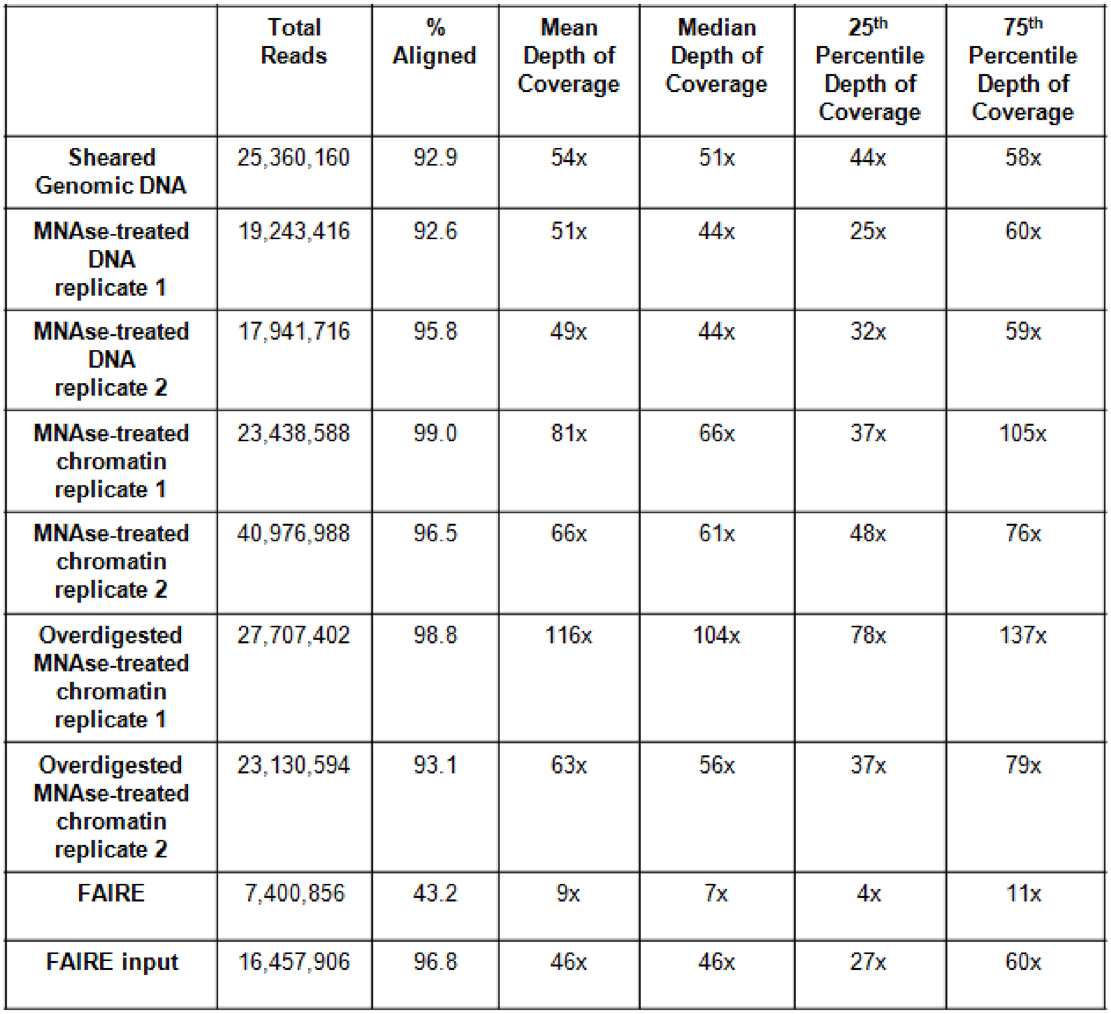
Supplemental Figure S3-9. Nucleosome distance analysis for regions of interest.



Supplemental Figure S3-10. Restriction endonuclease sensitivity assays demonstrate similar nuclease sensitivities for PGC-internal loci and dSSRs.

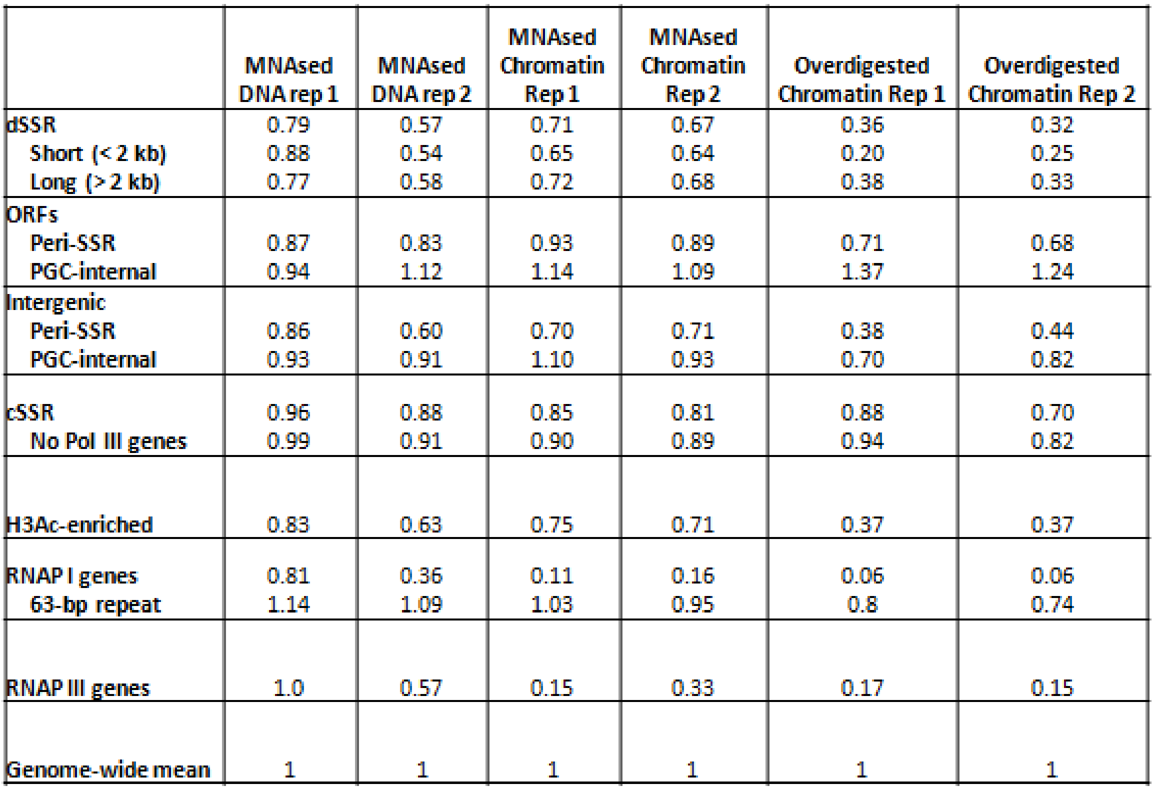


Supplemental Table S3-1. Alignment metrics for Illumina sequencing data.



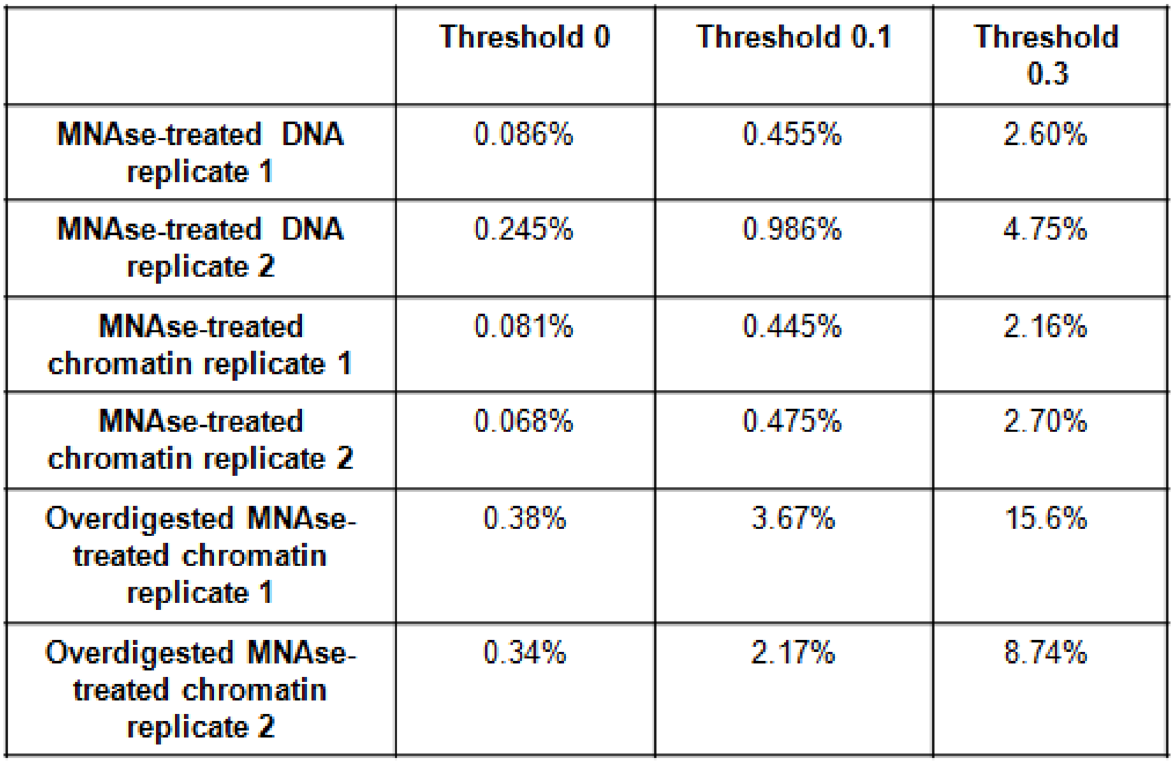
	Total Reads	% Aligned	Mean Depth of Coverage	Median Depth of Coverage	25th Percentile Depth of Coverage	75th Percentile Depth of Coverage
Sheared Genomic DNA	25,360,160	92.9	54x	51x	44x	58x
MNAse-treated DNA replicate 1	19,243,416	92.6	51x	44x	25x	60x
MNAse-treated DNA replicate 2	17,941,716	95.8	49x	44x	32x	59x
MNAse-treated chromatin replicate 1	23,438,588	99.0	81x	66x	37x	105x
MNAse-treated chromatin replicate 2	40,976,988	96.5	66x	61x	48x	76x
Overdigested MNAse-treated chromatin replicate 1	27,707,402	98.8	116x	104x	78x	137x
Overdigested MNAse-treated chromatin replicate 2	23,130,594	93.1	63x	56x	37x	79x
FAIRE	7,400,856	43.2	9x	7x	4x	11x
FAIRE input	16,457,906	96.8	46x	46x	27x	60x

Supplemental Table S3-2. Normalized fragment depth metrics according to genomic context.



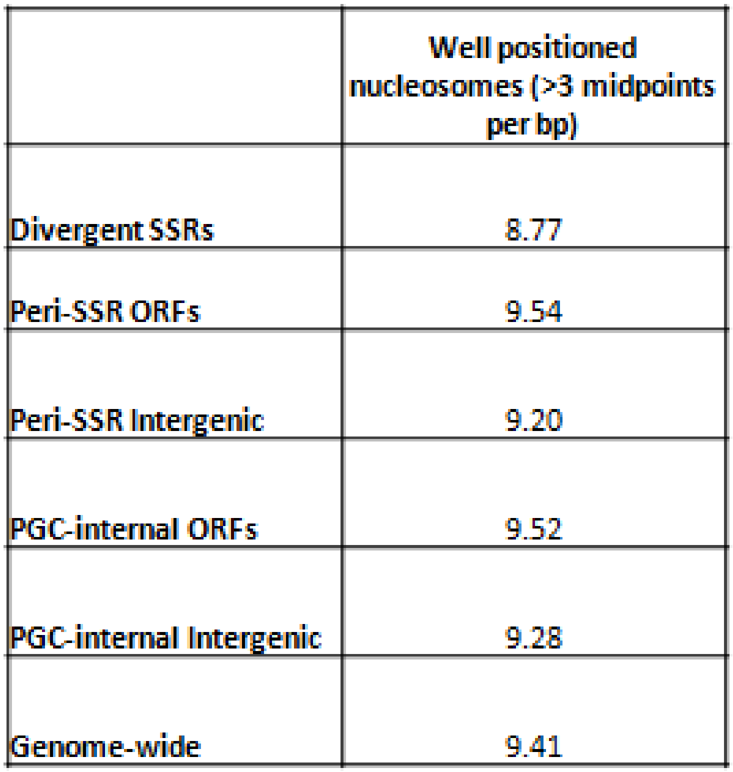
	MNAsed DNA rep 1	MNAsed DNA rep 2	MNAsed Chromatin Rep 1	MNAsed Chromatin Rep 2	Overdigested Chromatin Rep 1	Overdigested Chromatin Rep 2
dSSR	0.79	0.57	0.71	0.67	0.36	0.32
Short (< 2 kb)	0.88	0.54	0.65	0.64	0.20	0.25
Long (> 2 kb)	0.77	0.58	0.72	0.68	0.38	0.33
ORFs						
Peri-SSR	0.87	0.83	0.93	0.89	0.71	0.68
PGC-internal	0.94	1.12	1.14	1.09	1.37	1.24
Intergenic						
Peri-SSR	0.86	0.60	0.70	0.71	0.38	0.44
PGC-internal	0.93	0.91	1.10	0.93	0.70	0.82
cSSR	0.96	0.88	0.85	0.81	0.88	0.70
No Pol III genes	0.99	0.91	0.90	0.89	0.94	0.82
H3Ac-enriched	0.83	0.63	0.75	0.71	0.37	0.37
RNAP I genes	0.81	0.36	0.11	0.16	0.06	0.06
63-bp repeat	1.14	1.09	1.03	0.95	0.8	0.74
RNAP III genes	1.0	0.57	0.15	0.33	0.17	0.15
Genome-wide mean	1	1	1	1	1	1

Supplemental Table S3-3. Comparison of relative fragment depth maximum thresholds for identification of NH sites.



	Threshold 0	Threshold 0.1	Threshold 0.3
MNAse-treated DNA replicate 1	0.086%	0.455%	2.60%
MNAse-treated DNA replicate 2	0.245%	0.986%	4.75%
MNAse-treated chromatin replicate 1	0.081%	0.445%	2.16%
MNAse-treated chromatin replicate 2	0.068%	0.475%	2.70%
Overdigested MNAse-treated chromatin replicate 1	0.38%	3.67%	15.6%
Overdigested MNAse-treated chromatin replicate 2	0.34%	2.17%	8.74%

Supplemental Table S3-4. Comparison of well-positioned nucleosomes within regions of interest.



	Well positioned nucleosomes (>3 midpoints per bp)
Divergent SSRs	8.77
Peri-SSR ORFs	9.54
Peri-SSR Intergenic	9.20
PGC-internal ORFs	9.52
PGC-internal Intergenic	9.28
Genome-wide	9.41

Chapter 4

Identification of *cis*-regulatory elements associated with transcription of protein-coding genes in *Leishmania major* using an integrated bidirectional reporter

Preface

Ideas, tactics, and strategies arose from discussions between BA and SMB. BA designed and performed all experiments, analyzed data, and wrote the first draft of this chapter. SMB supervised these studies and provided comments that were incorporated into the final version presented here. This chapter represents a draft stage intended for publication, pending additional experiments.

Abstract

The early-diverging eukaryote *Leishmania* and other related trypanosomatid protozoa differ from other eukaryotes and transcribe protein-coding genes polycistronically from long, head-to-tail gene arrays called polycistronic gene clusters (PGCs). Currently, it is thought that transcription initiates primarily within divergent strand switch regions (dSSRs), where two PGCs are oriented head-to-head; these regions lack canonical eukaryotic *cis*-regulatory elements, although they are associated with broad regions of epigenetic marks associated with active transcription. Some dSSRs contain G-rich tracts, and potentially there are other conserved cryptic elements that have escaped detection; whether such signals function as *cis*-acting elements directly or function as organizers of epigenetic marks remains an open question. To enable functional tests of dSSRs, we developed a dual-luciferase platform that integrates into an endogenous dSSR and utilized it to interrogate the genetic factors contributing to transcription, using the dSSR of chromosome 1 as a model. We began by delineating the "core dSSR", which includes known sites of transcription initiation and contains two G-rich motifs, from the minimal functional splice acceptor sequences, which cannot be manipulated without altering *trans*-splicing of the dSSR-proximal reporter genes. We demonstrated that this integrated reporter system replicates features of endogenous dSSRs, including bidirectional transcription and acquisition of a transcriptionally-permissive environment. As expected, complete deletion of the core dSSR ablated the functionality of the dSSR reporter completely; unexpectedly, a complete swap of the core dSSR with unrelated DNA sequences showed little effect on functionality or bidirectional reporter gene expression. This demonstrates clearly that G-rich motifs and other sequences within the core dSSR are not required for bidirectional transcription from these loci, although *cis*-acting elements could remain within the endogenous splice acceptor sequences. Potentially, as-yet cryptic elements within the minimal splice acceptors provide signals mediating bidirectional expression. Alternatively, potentially the splice acceptors themselves are key elements of this signal, in addition to their known roles in simultaneously demarcating sites of both splicing and polyadenylylation.

Introduction

Leishmania and other trypanosomatid protozoa are responsible for several neglected tropical diseases, including leishmaniasis, African sleeping sickness, and Chagas' disease. In contrast to most eukaryotes, in which *trans*-acting factors interact with specific *cis*-regulatory elements to facilitate transcription of a single gene, trypanosomatid protein-coding genes are transcribed polycistronically by RNA polymerase II (RNAP II) from long, head-to-tail arrays called polycistronic gene clusters (PGCs), decoupling transcription from the regulation of individual gene products [reviewed in (1,2)]. Co-transcriptional *trans*-splicing and polyadenylylation reactions process these polycistronic pre-mRNAs into monocistronic mRNAs using the RNAP II-transcribed spliced leader (SL) RNA, which contributes the 5' cap and 5' end to all mature mRNAs in the cell (3,4). Transcription start site (TSS) mapping studies in *L. major* (5) and *T. brucei* (6,7) and chromatin immunoprecipitation (ChIP) studies of the basal transcription factor TRF4 (8,9) have localized transcription initiation events primarily, but not exclusively to divergent strand switch regions (dSSRs), where two PGCs are oriented head-to-head.

Several lines of evidence suggest that the regulation of RNAP II-mediated transcription of protein-coding genes differs from standard models in other eukaryotes. First, although the ubiquitous distribution of *cis*-regulatory elements in eukaryotic genomes suggests that similar mechanisms would control transcription of trypanosomatid protein-coding genes, trypanosomatids lack canonical eukaryotic RNAP II *cis*-regulatory elements including the TATA box, Initiator (Inr) elements, the B recognition element (BRE), and the downstream positioning element (DPE) (10), which in various combinations constitute a core promoter that functions in the transcription of many eukaryotic genes. Moreover, sequence-based comparisons of all *L. major* dSSRs failed to identify well-conserved motifs in these regions, although a weakly conserved *trans*-splicing acceptor consensus sequence was detected (Anderson and Beverley, unpublished data). While some structural features such as GC skew and DNA bending are different in dSSRs compared to the remainder of the genome, functional characterization has not been performed and it is unclear what, if any role these properties play in facilitating transcription from these loci (11,12). Interestingly, comparative genomics studies in *T. brucei* demonstrated an overrepresentation of poly(dG:dC) tracts in dSSRs (13), and the ability of these sequences to function as promoters in yeast (14,15) makes this motif particularly intriguing. Although these tracts are present between the identified TSS in chromosome 1 in *L. major* (10), these motifs are not overrepresented in regions associated with transcription initiation in *L. major*, as poly(dG:dC) motifs are scattered throughout the *L. major* genome and approximately half of putative sites of transcription initiation lack them altogether (9).

In light of the large gaps in knowledge regarding the genetic and epigenetic factors controlling transcription initiating at dSSRs, we set out to develop a reporter gene-based platform for assaying dSSR-mediated transcriptional activity in *Leishmania major,* which could be used to characterize determinants of the putative promoter activity of a dSSR. Similar assays have been used in diverse eukaryotes to characterize *cis*-regulatory elements for many years [reviewed in (16)]; briefly, the region of interest containing the putative *cis*-regulatory element is placed upstream of an easily visualized reporter gene, such as green fluorescent protein (GFP) or firefly luciferase (FLuc). There, comparisons to promoterless vectors and mutagenized promoters allow the robust identification of required *cis*-regulatory elements, and highly sensitive luciferase-based assays allow quantitative assessment of promoter activity. In *Leishmania*, promoter-trapping studies have been attempted using the dSSR of chromosome 1 to drive expression of a luciferase reporter. Although experiments using stable, episomal transfectants or integrated vectors targeted to the ribosomal RNA locus demonstrated a small increase in luciferase activity when the dSSR was present, these effects were weak and extremely variable (5). In addition, the interpretation of these experiments is complicated in the context of “run around” transcription from episomes, which does not require a promoter element (17), by the highly active ribosomal RNA promoter, and by the presence of an unannotated gene in the dSSR characterized (18). Together, these technical issues make it difficult to conclude anything about the putative promoter activity of dSSRs, and additional studies to determine the influence of dSSRs on bidirectional transcription are warranted.

In this work, we probed the expression arising from a single dSSR using a bidirectional, dual luciferase-based reporter that is integrated into an endogenous dSSR. Here, we chose the dSSR of chromosome 1, which has been well-characterized previously (5,18) and contains two poly(dG:dC) tracts, which are the sole motif identified in comparative genomic studies in any trypanosomatid protozoa (13). We sought to replace the dSSR in its entirety with the bidirectional reporter, using the dSSR-flanking ORFs to target the reporter to the correct locus for integration using homologous recombination. In this reporter, the Renilla (RLuc) and firefly (FLuc) luciferases are positioned immediately downstream of the dSSR, allowing the assessment of transcription originating from this locus. In addition, selectable markers are present downstream of the reporter genes, enabling selection of transfectants and assessment of bidirectional transcription, which should be independent of which marker was selected for. Importantly, we utilize episomal transfections to further understand the requirements for dSSRs in bidirectional transcription. These extrachromosomal DNA elements are promiscusously transcribed on both strands, and appropriate signals for *trans*-splicing are the only requirement for expression of episome-derived genes. In these studies, deletion of an element required for transcription would alter the ability to obtain transfectants from the linearized DNA construct without altering the efficiency of transfection of episomal DNA.

In order to make deletions within the dSSR, we first dissected the dSSR to separate the "core dSSR" from the minimal functional splice acceptor sequences, which are essential for proper *trans*-splicing of the dSSR-flanking reporter genes. The identification of these minimal functional splice acceptors was amenable to study on episomal vectors, and usage of the major endogenous splice acceptor dinucleotide was confirmed in all subsequent studies using the integrated bidirectional reporter. We show that after correct integration and replacement of the endogenous chromosome 1 dSSR on one allele, the bidirectional reporter was able to facilitate bidirectional reporter gene transcription independent of selective pressure, and that the integrated reporter maintains the epigenetic signatures of dSSRs. Deletion of the "core dSSR", which includes the entire region between minimal functional splice acceptor sequences, resulted in the inability to generate viable transfectants from linearized DNA. Control transfections using episomal DNA readily generated normal, viable transfectants, demonstrating that the construct itself is fully functional, but a *cis*-acting element required for transcription from a chromosomal locus may be present within the deleted region. However, we found that replacement of the core dSSR with completely unrelated DNA sequences of a similar length was able to sustain bidirectional reporter gene activity in the chromosomally integrated reporter, suggesting that in fact the core dSSR does not contain a *cis*-acting element and that poly(dG:dC) tracts are not required for transcription from dSSRs. These experiments suggest instead that splice acceptor sequences may contain an as-yet cryptic element required for bidirectional transcription. Alternatively, potentially the splice acceptor itself could be *cis*-acting elements, on top of its roles in both splicing and polyadenylylation. These data provide further evidence of a potential link of the fundamental eukaryotic processes of transcription, splicing, and polyadenylation within this deep-branching eukaryote.

Results

Design of an integrated, bidirectional reporter for characterization of dSSR function

We focused our studies on the dSSR of chromosome 1, which has been studied in detail previously and exhibits features representative of most dSSRs. In all studies described here, we define this dSSR as the inter-ORF regions between *LmjF01.0315* and *LmjF01.0320*; relevant features within this region, which are described below, are depicted graphically in Figure 4-1A. Several transcription start sites within this dSSR have been mapped by 5' RACE (5), and this dSSR contains sequences which others have hypothesized to be important: two poly(dG:dC) tracts are present between the innermost transcription start sites, and this dSSR possesses DNA-encoded structural properties that may distinguish dSSRs from the rest of the genome, as defined by *in silico* models of DNA bending (12). This dSSR is bound by the basal RNAP II transcription factor TRF4 and is marked by activating histone modifications (9), and analysis of nuclease-hypersensitive sites, which typically correlate with active regulatory elements and sites of transcription initiation in other eukaryotes, showed broadly distributed, heterogeneous nuclease-hypersensitive sites that correlate with activating histone modifications (Anderson, Shaik, and Beverley, in preparation). In addition, genetic studies of this dSSR demonstrate the tractability of this locus, although an unannotated gene of unknown function (now *LmjF01.0315*) complicates the interpretation of some of the genetic manipulations published previously (18).

We developed a bidirectional reporter gene construct that integrates directly into the dSSR of chromosome 1 using homologous recombination (Fig. 4-1B; Supplemental Fig. S4-1). This bidirectional reporter contains the Renilla (RLuc) and firefly (FLuc) luciferase genes immediately proximal to the endogenous dSSR sequence, allowing the quantiation of transcription events stemming from the dSSR. While the activity assays of these luciferases are similar, these proteins use very different substrates and can be assayed independently of one another (19). In experiments comparing episomal and integrated lines, we observed that FLuc activity assayed by bioluminescence was proportional to mRNA levels, demonstrating that this bioluminescence assay is a suitable proxy for FLuc transcript levels (Supplemental Fig. S4-2A). We observed robust RLuc activity in vectors designed in our first iteration of these vectors, and a unidirectional reporter called the RLuc ½ construct (Fig.4-2A) was used for studies mapping the minimal functional splice acceptor. However, RLuc expression from the integrated construct was relatively low (~20-fold over background; data not shown), and the Kozak sequence of the RLuc gene was modified in an effort to improve RLuc expression in the bidirectional pLUC v2 construct (Supplemental Fig. S4-1). Unexpectedly, when this modification was made in the context of lines containing the endogenous dSSR, the alteration now generates an upstream ORF that overlaps the start codon in the 5' UTR of the RLuc mRNA which reduces RLuc protein expression to the point that it was unusable (Supplemental Fig. S4-1B); as a result, quantitative RT-PCR was used to measure RLuc levels in subsequent experiments using the pLUC v2 vector backbone.

Downstream of the reporter genes, the *SAT* and *HYG* antibiotic resistance genes are preceded by "strong" SA sequences derived from the α-tubulin locus or from previously characterized expression vectors (26,27; Lye and Beverley, unpublished data) (Fig. 4-1A; Supplemental Fig. S4-1; Table 4-3). This enables the selection of parasites using one or both "halves" of the reporter construct: in the absence of bidirectional promoter activity, one would expect that selection with nourseothricin (SAT) would generate parasites expressing only FLuc, while selection with hygromycin B (HYG) would generate parasites expressing only RLuc. However, a *bona fide* bidirectional reporter would function independently of which marker was selected for. Finally, the reporter construct is flanked by targeting sequences derived from the dSSR-proximal genes *LmjF01.0315* and *LmjF01.0320*, allowing targeted integration into the dSSR of chromosome 1 in a manner which should leave their sequence and expression levels unaffected (Fig. 4-1B).

Defining the core dSSR, separate from known elements required for trans-*splicing*

To delineate the sequences required for *trans*-splicing from the rest of the dSSR, we used an episome-based approach to identify the minimal functional SA sequences present within this dSSR. Because episomes are present in many copies and are transcribed promiscuously by a "run around" mechanism (17), differences in reporter gene activity are likely to reflect differences in *trans*-splicing or in the 5' UTR of the mRNA, rather than from differences in transcription of the reporter gene. We generated lines containing an episomal copy of the RLuc "half" of the bidirectional reporter preceded by a fragment of the dSSR, oriented according to the directionality of the SA site (Fig. 4-2A). We generated a panel of constructs bearing fragments of the dSSR starting from the dSSR-ORF junctions on the 5' or 3' side of the dSSR, described in Figures 4-2B and 4-2C, and quantified RLuc activity from lines containing these episomes (Figs. 4-2D, E). Comparison of three transfectants from each construct demonstrated relatively little variation among lines bearing the same construct (Figs. 4-2 D, E). We found that 84 bp of dSSR-derived sequence from the 5' (*LmjF01.0315*) side was sufficient for full RLuc activity (Fig. 4-2E); similarly, 301 bp of sequence on the 3' side of the dSSR was also sufficient for both full RLuc activity (Fig. 4-2D). To verify that these constructs were utilizing the correct, major splice acceptor dinucleotides within the dSSR, we utilized a spliced leader (SL)-based reverse transcription and PCR (RT-PCR) assay to map the splice acceptor dinucleotides in the *RLuc* mRNA (Supplemental Fig. S4-3A). Here, a reverse primer specific for the gene of interest, here the *RLuc* mRNA, is paired with a SL-specific forward primer to amplify across the SL-mRNA junction. Splice acceptor utilization can be roughly quantified by resolving these amplicons on an agarose gel, and the splice acceptor site can be identified by the product size and/or sequencing. We utilized this approach to map the major splice acceptor dinucleotides in one clone from each of the constructs described in Fig. 4-3 (Supplemental Fig. S4-3B). Using the known major splice acceptor dinucleotides in the dSSR to predict the length of the SL-RLuc amplicon, we expected amplicons of 386 bp on the "right" side of the dSSR and 248 bp on the "left". We found that constructs that conferred high RLuc activity levels (R301, R401, and all "left" side constructs; Fig. 4-2D,E) showed SL-RLuc amplicons of the expected length, indicating use of the expected splice acceptor dinucleotide. While sequencing validated that the R301 and L182 constructs utilized the expected splice acceptor dinucleotide (Supplemental Figs. S4D, E), we observe that the L84 construct instead utilizes a minor splice acceptor dinucleotide, and does not match the predicted amplicon (Supplemental Fig. S4C). Using these data, we have roughly mapped the minimal functional splice acceptor sequences, here defined as 182 bp and 301 bp of sequence from the left and right ends of the dSSR. Now that these sequences have been defined, we can interrogate the "core dSSR" between these sequences using deletions and substitutions to identify *cis*-acting elements without altering RNA processing. Importantly, to definitively determine that reporter gene expression is not altered by differential splice acceptor usage, we also confirmed the use of the correct, major splice acceptor dinucleotide in all lines described in this chapter using spliced leader (SL)-primed RT-PCR and Sanger sequencing; SL-based RT-PCR data from several representative integrated and episomal lines described in this study are shown in Supplemental Figure S4-4.

The integrated bidirectional reporter system functions as a bona fide *dSSR*

The dSSR of chromosome 1 contains no homology to other loci in the *Leishmania* genome, and specific integration of this reporter construct into the correct locus can be easily confirmed by allele-specific PCR (Fig. 4-3A). Moreover, allelic replacement by homologous recombination typically modifies only one allele with each transfection. The isolate used in this work contains only two copies of chromosome 1 (Anderson, Shaik, and Beverley, in preparation); curiously, this same line obtained from other sources can show 3 or more copies of this chromosome (18), a phenomenon associated with the extreme plasticity of the *Leishmania* genome. The presence of a wile type provides a valuable internal control for epigenetic characterization of the integrated reporter.

We first quantified our ability to target the bidirectional reporter to the appropriate locus, comparing transfections using linearized DNA, which should integrate into the expected locus using homologous recombination, and episomal DNA, which are stably propagated extrachromosomally. Episomal transfections are frequently used as controls for the integrity of constructs used for allelic replacement, as these do not require elements beyond a selectable marker and elements required for *trans*-splicing of the marker gene (17). Thus, comparisons between episomal and integrated bidirectional reporters allow us to discern effects specifically associated with integration and "bidirectional" expression in the proper chromosomal context. In transfections of the bidirectional reporter containing the full-length dSSR, we observed that both linearized and episomal DNA readily generated transfectants (Table 4-1). However, we observe roughly 100-fold lower transfection efficiencies for the linearized construct than for the episomal DNA, possibly a consequence of the size of the bidirectional reporter (Table 4-1).

We next surveyed several clones transfected with linearized DNA to determine whether the bidirectional reporter integrates as expected. We used primer pairs spanning the reporter-chromosome junction to determine whether the reporter had integrated in the proper locus; here, the forward primer to assess 5' integration and the reverse primer to assess 3' integration are located outside of the targeting fragments used for homologous recombination and cannot amplify episomal DNA (Fig. 4-3A). To confirm that both halves of the reporter were integrated into the same allele, we utilized a third primer set spanning the FLuc-dSSR-RLuc junction (Fig. 4-3A). In all cases, correct integration is verified when a PCR amplicon of the correct size and/or sequence is obtained. Using these primer pairs to examine several clones containing the bidirectional reporter, we observe that the majority of these lines integrate as expected, showing bands of the expected size at the 5', 3', and middle PCRs (Fig. 4-3B,C,D). A small number of clones integrated only one side the construct using homologous recombination events at the dSSR and at the 5' or 3' ends (Fig. 4-3B, D; see lines marked with a yellow X), which was confirmed by PCR (data not shown).

Using lines that had successfully integrated the entire bidirectional reporter, we next assayed FLuc and RLuc reporter gene levels using bioluminescence assays and qRT-PCR. As predicted for a bidirectional reporter, we observe similar FLuc activity and RLuc mRNA levels in all lines bearing the integrated reporter, independent of the selective pressures applied during transfection and maintenance in culture (Fig. 4-4A-B). This is a clear indication that our dSSR reporter strategy allows replacement of the endogenous dSSR and is able to reveal bidirectional transcription from the dSSR.

The integrated bidirectional reporter recapitulates the epigenetic state of dSSRs

We then asked whether the integrated bidirectional reporter maintained the epigenetic characteristics that define dSSRs. To do so, we used formaldehyde-assisted isolation of regulatory elements (FAIRE) coupled to quantitative PCR (qPCR); this method isolates regions of nucleosome-depleted DNA from bulk chromatin by crosslinking histones to DNA and removing protein-linked DNA fragments using phenol extraction (22,23). This method has been used extensively in other eukaryotes to identify novel regulatory elements *in vivo* and integrates relevant epigenetic properties associated with regulatory elements and sites of active transcription (24). We previously showed that this method detects broad regions of open chromatin associated with dSSRs (Anderson, Shaik, and Beverley, in preparation), including the dSSR of chromosome 1 (Figure 4-5A). We designed qPCR amplicons that specifically detect genomic regions derived from the wild type (WT) allele, from various points within the integrated bidirectional reporter, and from a locus far from the dSSR that is not marked by FAIRE enrichment (Fig. 4-5B). Comparison of the FAIRE signal relative to an input control at these loci demonstrates a similar level of enrichment in the bidirectional reporter allele compared to the WT allele, which extends out through the ends of the reporter in a pattern similar to that observed in genome-wide FAIRE experiments (Fig.4-4C). In combination with reporter gene transcription data, this demonstrates that the integrated, bidirectional dual-luciferase reporter effectively replicates the behavior of an endogenous dSSR, allowing us to interrogate the genetic and epigenetic determinants of dSSR behavior in more detail *in vivo*.

The putative promoter activity of a dSSR is not dependent on DNA sequence or structural properties

Given the uncertainties about the nature and properties of potential *cis*-acting elements within dSSRs, we began our search for *cis*-regulatory elements in dSSRs with a large-scale deletion of the chromosome 1 dSSR. We generated a construct bearing a 489-bp deletion encompassing the entire core dSSR between the minimal functional SA sequences, including both poly(dG:dC) tracts (Δ489; Figure 4-6A). This was transfected into WT *L. major* FV1 as a linearized construct which if functional should integrate into the desired locus by homologous recombination; episomal transfections acted as a control for the integrity of the plasmid, markers, and splice acceptors. Interestingly, in 5 independent transfections, we were unable to obtain any colonies from the linearized Δ489 construct, despite a substantial number of colonies from the linearized WT dSSR construct and comparable efficiencies with both the WT and Δ489 constructs when transfected as episomes (Table 4-1). This demonstrates that the construct itself bears fully functional elements for *trans*-splicing of the reporter-encoded mRNAs and selectable markers. Thus, the deleted region lacks elements required for activity when integrated *in situ*.

Although these data present the possibility that the deleted region contains a *cis*-regulatory element important for transcription, it is also possible that placing two endogenous SA sequences in close genomic proximity could lead to interference and loss of function. To address this possibility, we generated a second construct containing a 489-bp substitution of the same region described above, using a gene-derived sequence from *T. brucei* that lacks homology to the *L. major* dSSR (Δ489 + S) (Fig. 4-6A; Supplemental Fig. 4-5A). Importantly, the structural properties of these sequences including DNA bendability, DNA curvature, melting temperature, and G/C content of these DNA sequences also do not share any apparent patterns (Supplemental Fig. 4-5B-D).

We were surprised to find that the Δ489 + S construct generated colonies that properly integrated the linearized DNA fragment at a frequency of 28 colonies per 10 µg of DNA, similar to that of the WT dSSR (11 colonies per 10 µg DNA) (Table 4-1). These data suggest that the core dSSR functions as a spacer between the endogenous splice acceptors, and that the sequence content of this region is not important. We quantified reporter gene activity using bioluminescence assays and qRT-PCR as performed previously and were again surprised to see that lines bearing this unrelated stuffer sequence were able to support reporter gene expression independent of which marker was selected for (Fig. 4-6B-C). We assessed 8 clones bearing the Δ489 + S deletion and found that on average the FLuc activity is 1.4 fold greater in these lines, and RLuc mRNA levels are 2.3 fold lower than lines compared to lines bearing the WT dSSR. Although the effect of substitution of the core dSSR on reporter gene expression is statistically significant (FLuc, p = 0.0499; RLuc, p = 0.0004), the effect is weak and only apparent because of the high reproducibility of these assays. These data demonstrate clearly that poly(dG:dC) tracts are not required for bidirectional transcription in dSSRs, and the ability to substitute the core dSSR with unrelated sequence strongly suggests that the capacity for bidirectional transcription in dSSRs may not depend on the sequence content or structural features present within the dSSR.

Discussion

The ubiquitous nature of *cis*-regulatory elements in the regulation of eukaryotic genes led many to speculate that *Leishmania* and other trypanosomatid protozoa contain DNA-encoded motifs required for transcription initiating within dSSRs. However, the lack of conserved motifs in these regions suggested that the mechanisms regulating transcription of protein-coding genes may be regulated using highly atypical *cis*-acting motifs or solely via epigenetic mechanisms, in contrast to most eukaryotes. To elucidate whether there is a requirement for *cis*-regulatory elements in dSSR-mediated transcription, we developed a bidirectional, dual-luciferase reporter system that integrates directly into a dSSR. Using this bidirectional reporter system, we focused our efforts on the dSSR of chromosome 1, which contains motifs which have been hypothesized to be important for transcriptional regulation. We demonstrate that the dSSR of chromosome 1 facilitates bidirectional transcription in a manner that does not depend on selective pressures applied during transfection and allelic replacement, indicating that this locus behaves as a *de facto* bidirectional promoter region. Importantly, the integrated bidirectional reporter construct is also marked by heterogeneous nuclease-hypersensitive sites as assayed by FAIRE, demonstrating that the integrated reporter effectively replicates the transcriptional and epigenetic features of endogenous dSSRs.

In light of the lack of conserved elements in dSSRs, we focused our efforts more broadly on the entire dSSR, rather than on specific elements within this region. To delineate the sequences required for normal *trans*-splicing of the dSSR-proximal reporter genes, we used an episome-based approach to identify the minimal functional splice acceptor sequences on both ends of the dSSR. We found that 182 bp of sequence on the *LmjF01.0315* side of the dSSR and 301 bp of sequence on the *LmjF01.0320* side of the dSSR were sufficient for usage of the major splice acceptors identified in the endogenous dSSR; the introduction of shorter DNA fragments resulted in either ablation reporter gene expression or in use of a minor SA dinucleotide. Once we delineated these sequences from those which could be deleted without affecting *trans*-splicing, we generated a 489 bp deletion that encompassed the entire region between minimal functional splice acceptor sequences. In a total of 5 independent experiments, we were unable to obtain colonies using this construct, suggesting a *cis*-regulatory element required for transcription was present in the deleted region. However, substitution of this region with an equivalently-sized, completely unrelated DNA sequence restored our ability to generate viable transfectants that properly integrated the construct and reporter gene activity was stunningly normal, suggesting that in fact a *cis*-regulatory element was not present in this locus. We anticipate that these lines maintain an epigenetic state similar to that of the endogenous dSSR, and we will quantify this using FAIRE as we described previously for lines bearing the WT dSSR reporter.

While these experiments provide strong evidence against the existence of *cis*-acting elements within a large portion of the dSSR, a number of possibilities remain which are the subject of ongoing experiments at this time. The data presented above suggest that some amount of space between divergently-oriented splice acceptor sequences is required for dSSR function, and the possibility that these endogenous splice acceptor sequences harbor a cryptic *cis*-acting element remains. Furthermore, it is possible that an unrelated *cis*-acting element was present within the stuffer used in these experiments. To address these possibilities, we have generated reporters containing a fully synthetic dSSR using two additional unrelated stuffer sequences and synthetic splice acceptor sequences, and analysis of transfectants from these lines is in progress. To characterize the requirement for a spacer between splice acceptors in dSSRs, we have generated constructs bearing smaller stuffers in the context of the synthetic dSSR. We have successfully generated parasites integrating the reporter with as little as 115 bp of "stuffer" sequence between these synthetic splice acceptors, and we are currently assessing the reporter gene expression of these lines. These preliminary experiments suggest that the only remaining sequences in dSSRs that could function as *cis*-acting elements are the splice acceptors themselves, an intriguing possibility, as the presence of divergently-oriented splice acceptors in dSSRs distinguishes them from the rest of the genome. However, the ubiquitous nature of these sequences throughout the genome and the presence of unidirectional transcription initiation regions within polycistronic gene clusters (9) suggest that the mechanisms controlling transcription from these loci are more complex this, possibly containing a significant epigenetic component.

Materials and Methods

Plasmid generation

The bidirectional reporter plasmid was cloned in two sections using the GeneArt High Order Genetic Assembly Kit (Life Technologies), with extensive modifications after the assembly. Briefly, DNA fragments encoding the genes and intergenic regions were amplified with Phusion polymerase (NEB) using the primers and templates described in Table 4-2. Individual fragments were run on an agarose gel and were extracted by gel purification (Qiagen); the purified fragments were pooled with linearized pYES1L vector according to the manufacturer's protocol in the combinations described in Table 4-2. The pooled DNA fragments were transfected into *S. cerevisiae* MaV203, and cells successfully assembling the plasmid were selected on Trp drop-out medium according to the manufacturer's recommendations. The same oligonucleotides used to amplify the DNA fragments were used to screen colonies for proper assembly, as indicated by the presence of a PCR amplicon of the expected size. Properly assembled plasmids were transformed into electrocompetent OneShot *E. coli*

(Life Technologies), and the entire insert was sequenced. This assembly generated the pYES1L-RLuc plasmid and the pYES1L-FLuc plasmid, which were subsequently modified to allow the utilization of additional restriction sites and to allow the use of a high-copy plasmid for propagation in *E. coli*.

The HindIII fragment from pYES1L-RLuc was cloned into HindIII-digested pUC19 to generate the high-copy plasmid pUC19-RLuc ½ (B6863). Similarly, the HindIII fragment from pYES1L-FLuc was cloned into HindIII-digested pUC19 to generate the high-copy plasmid pUC19-FLuc ½. The 5' end of the RLuc "half" insert was modified to introduce an optimized Kozak sequence CCACC upstream of the RLuc ORF by amplifying the RLuc gene with primers SMB4816 and SMB4992 (Table 4-3), digesting the PCR product with SacI, and cloning it into SacI-digested pUC19-RLuc ½ to generate pUC19-RLuc ½ (CCACC) (B6979). The 5' end of the FLuc "half" insert was modified similarly by amplifying the 5' end of the FLuc ORF with primers SMB4816 and SMB4992 (Table 4-2), digesting the PCR product with PacI and PsiI, and cloning it into PacI/PsiI-digested pUC19-FLuc ½ to generate pUC19-FLuc ½ (CCACC) (B6974). Note that introduction of the endogenous dSSR into the plasmid pUC19-RLuc ½ (CCACC) generates an upstream ORF that overlaps the RLuc start codon, ablating translation of the gene (Supplemental Fig. S4-2A). In contrast, cloning of dSSR framgents into the predecessor of this plasmid pUC19-RLuc ½ generates plasmids containing an upstream ATG that is in frame with the RLuc ORF, and RLuc activity is readily detected from derivatives of this plasmid. The bidirectional reporter plasmid "pLUC" (B6993) was generated by cloning the BamHI-XbaI fragment from pUC19-RLuc ½ (CCACC) into BamHI/XbaI-digested pUC19-FLuc ½ (CCACC). Because the original FLuc insert contained a defective splice acceptor between the *FLuc* and *PAC* genes, this construct was repaired. Briefly, DNA fragments consisting of the 3' end of the *FLuc* gene, the α-tubulin interORF region from *L. major* FV1, and the *SAT* gene were amplified using primers and templates described in Table 4-3. These fragments were assembled using fusion PCR (25) and cloned into pCR-Blunt (Life Technologies). The insert was confirmed by restriction digestion and sequencing. This insert was released with PsiI and BglII digestion and cloned into PsiI/BglII-digested pLUC to generate pLUC v2

(B7129), which was used for all studies using the integrated bidirectional reporter. This plasmid is depicted graphically in Supplemental Fig. S4-1.

For splice acceptor mapping studies, fragments of the dSSR were amplified using primers described in Table 4-3, digested with XbaI, and cloned into XbaI-digested pUC19-RLuc ½ to generate the episomes described in Table 4-3. For the remaining studies, dSSR derivatives were amplified using fusion PCR (25) using the primers described in Tables 4-3. The inserts were cloned into pCR-Blunt (Life Technologies) and were confirmed by restriction digestion and sequencing. The dSSR derivatives were released by XbaI digestion and were cloned into XbaI-digested pLUC v2 to generate the plasmids pLUC-SSR (B7138), pLUC-Δ489 (B7190), and pLUC-Δ489+S (B7191).

Cell culture and generation of lines bearing bidirectional reporter construct

Cell culture and transfections were performed as described in Chapter 2. The targeting fragments from all pLUC v2 derivatives were released by digestion with BamHI and BlpI. Selection on semisolid medium was performed using 12-50 µg/mL hygromycin B (Calbiochem) and/or 100-110 µg/mL nourseothricin (Werner BioAgents). Integration of the targeting fragment into the dSSR was confirmed using the primers described in Table 4-4.

Luciferase assays

Logarithmic-phase cells (2-4 x 10^6 cells/mL) were pelleted and resuspended at a concentration of 1 x 10^7 (FLuc) or 2 x 10^7 cells/mL (RLuc) in DMEM lacking phenol red (Gibco). 200 µL of cell suspension was aliquotted in 96 well plates with black walls (Fisher) in triplicate, and 1 µL of 30 mg/mL D-luciferin (Gold Bio) or 500 µM native coelenterazine (Gold Bio) was added to each well. Cells were incubated for 10 minutes at room temperature and were imaged with a 10 second exposure time using a Xenogen IVIS photoimager (Caliper Biosciences), and luciferase activity was quantified in photons/second (p/s).

Quantitative reverse-transcriptase PCR (qRT-PCR) and splice acceptor dinucleotide mapping

Total RNA was isolated from logarithmic-phase cells as described in Chapter 2. Reverse transcription was performed according to the manufacturer's instructions using SuperScript III first-strand reverse transcriptase (Life Technologies) using 1 µg total RNA in a 20 µL reaction volume. Reactions in which reverse transcriptase was omitted were performed in parallel to rule out DNA contamination. Real-time PCR was performed using Sybr Green PCR master mix (Life Technologies); 20 µL reactions were prepared according to the manufacturer's instructions, containing 1 µL of cDNA or no RT control reactions and primers specific to the *RLuc, FLuc,* or *DHFR* genes (Table 4-4). Quantitative analysis was performed using an ABI Prism 7000 (Applied Biosciences) using the default annealing temperatures and extension times, and the specificity of each amplicon was assessed by melt curve analysis. All no RT controls showed Ct values greater than 35, indicating little DNA contamination. Reporter gene abundance was calculated relative to the *DHFR* reference using the $2^{-\Delta\Delta Ct}$ method (26). Splice acceptor dinucleotide mapping was performed using a SL-specific primer (Table 4-4) and a reverse primer specific to the *FLuc* or *RLuc* gene (Table 4-4); reactions were amplified using Taq DNA polymerase (New England Biolabs) using cDNA or no RT control reactions. Amplicons were assessed by agarose gel electrophoresis or by Sanger sequencing.

Formaldehyde-assisted isolation of regulatory elements (FAIRE) and quantitative PCR

Formaldehyde-assisted isolation of regulatory elements was performed on logarithmically-growing cells as described in Chapter 3. Quantitative PCR reactions were prepared using Sybr Green PCR Master Mix according to the manufacturer's instructions, using primers described in Table 4-4 and 5 ng of DNA. Real-time PCR was performed as described previously using the default parameters, and the specificity of the amplicon was quantified using the melt curve. The enrichment relative to input DNA was quantified by the $2^{-\Delta\Delta Ct}$ method (26), normalizing to the dSSR-distal gene *LmjF01.0400*.

Acknowledgments

We thank Drs. Jenny Duncan and David Rudnick at Washington University for the use of their Bioruptor for FAIRE experiments. We also thank Dr. David Piwnica-Worms for the gift of the pRLuc-N3 plasmid. This work was funded by the National Institutes of Health (R0129646 to SMB, JSS, and BAA; T32 GM007067 to BAA) and the Washington University Berg-Morse and Schlesinger Graduate Fellowships to BAA.

Figure Legends

Figure 4-1. Description of the dSSR and planned integration of the bidirectional dual-luciferase reporter into chromosome 1. (A) Depiction of chromosome 1 dSSR. The dSSR-proximal region of chromosome 1 is depicted in the top of the panel, and a more detailed view of relevant landmarks within the dSSR is depicted below. Open reading frames (ORFs) are represented as blue and red arrows, and the color and orientation of the arrows indicate the coding strand of the ORF. The dSSR-proximal genes *LmjF01.0315* and *LmjF01.0320* are indicated in the bottom panel. Splice acceptors are depicted as black boxes, located at the 5' ends of genes. The core dSSR, defined as the region between the minimal functional splice acceptors, is depicted as a green box in both parts of the figure; the region containing the poly(dG:dC) tract is shown as a wider green box containing $(G)_n$ in the lower part of the figure. Transcription start sites within the dSSR identified in (5) are depicted as black, trianglular flags, and the strandedness is indicated by the placement above (positive strand) or below the green bar (negative strand); not all transcription start sites are shown, and these events are not drawn to scale. (B) Genes, splice acceptors, and the dSSR are depicted as described in (A). The gray boxes in the bottom part of the figure represent inter-ORF regions derived from *Leishmania* expression vectors or from the α-tubulin locus in *L. major* and are described in detail in Table 4-2; these contain sequences sufficient for *trans*-splicing and polyadenylylation of the upstream and downstream genes. Homologous recombination events between the targeting fragments derived from the dSSR-flanking genes *LmjF01.0315* and *LmjF01.0320* and the

endogenous locus, which result in integration of the construct into the chromosome, are depicted with black crossover lines.

Figure 4-2. Mapping of the minimal functional splice acceptor in the dSSR of chromosome 1. (A) Depiction of the RLuc ½ construct used for identification of the minimal functional splice acceptor. The *RLuc*, *HYG*, and *LmjF01.0315* genes are depicted as red arrows in the plasmid; the inter-ORF regions derived from pX-series or pIR-series vectors are shown as black arcs. Amp (ampicillin resistance gene) and ori (origin of replication) sequences required for propagation in bacteria are depicted in gray. Splice acceptor dinucleotides are designated with "AG". Fragments from the dSSR were cloned upstream of the RLuc ORF in the plasmid, as depicted at the top. In the presence of sufficient *trans*-splicing signals, mature mRNAs for the *RLuc* and *HYG* genes are generated (depicted as red lines; the 5' cap is depicted as a black circle, SL-derived sequence is depicted as a yellow box, and polyadenylylation site is indicated with AAAA). (B, C) Graphical representation of dSSR fragments used for minimal functional splice acceptor mapping of the "right" (B) and "left" (C) sides of the dSSR. Fragments are depicted in gray boxes, and the location of the *RLuc* ORF and plasmid backbone are shown in red and light gray boxes, respectively. The missing fragments of the dSSR are represented with dashed lines, and the location of the major splice acceptor AG dinucleotide is indicated with a vertical line across the fragments. (D, E) RLuc activity for transfectants bearing RLuc ½ episomes containing dSSR fragments depicted in (B) and (C). Y-axis indicates RLuc activity quantified in photons per second per 10^6 cells. Vertical bars represent individual clones bearing each construct, which is indicated below the graph. Error bars indicate the standard deviation of three technical replicate wells. WT indicates untransfected WT *L. major* FV1; "No SA" indicates a line transfected with the RLuc ½ backbone, which lacks a splice acceptor at the 5' end of the *RLuc* ORF.

Figure 4-3. Confirmation of planned integration of reporter constructs into chromosome 1 dSSR. (A) Description of PCR-based assay to validate the proper integration of the bidirectional reporter in the expected locus. Briefly, primers specific to regions outside of the targeting fragments used for homologous recombination (the forward primer in the 5' integration PCR and the reverse primer in the 3' integration PCR) were paired with primers specific to the 3' ends of the selectable markers to amplify across the sites where homologous recombination occurred to validate proper localization of the construct. Similarly, primers specific to the *FLuc* and *RLuc* ORFs were used to amplify across the middle of the construct to ensure that both sides of the reporter were present in the same chromosome. The bidirectional reporter was described in Figure 4-1. The PCR amplicons are depicted by horizontal brackets, and primer locations are indicated with black line arrows. (B-D) PCR analysis of all lines used in this study; primer pairs described in (A) were used to amplify across the 5', middle, and 3' junctions in the reporter construct, and the PCR products were resolved on an agarose gel relative to a 1 kb+ double-stranded DNA ladder (Life Technologies). Positive controls were generated previously using the "half" reporters, and the pLUC plasmid is used for positive controls in the middle PCR. WT *L. major* FV1 and a no-template control were also performed as negative controls. These reactions demonstrate amplification of bands of the expected size at the 5' junction (B), across the middle of the reporter (C), and at the 3' junction (D). Lines failing to demonstrate an amplicon of the expected size in any of these assays are marked with a yellow X and were excluded from additional study.

Figure 4-4. Reporter gene data for integrated bidirectional reporter clones containing the WT dSSR. (A) Graphical representation of bidirectional reporter, as described in Figure 4-1. (B) FLuc activity data, represented in photons per second per 2 x 10^6 cells. Y-axis indicates FLuc activity; individual bars in the X-axis indicate individual clones. Boxes below the X-axis indicate the selection applied during transfection, allelic replacement, and maintenance in culture. Error bars represent the standard deviation of three separate wells in one experiment. (B) RLuc mRNA levels, quantified relative to the *DHFR* mRNA and normalized to pLUC-WT SSR clone 1.1. Boxes below the X-axis and individual bars are as described in (A).

Figure 4-5. FAIRE-qPCR indicates maintenance of the epigenetic state of the dSSR. (A) Location of dSSR-proximal and "far right" amplicons relative to known FAIRE peaks are depicted with red boxes. FAIRE and acetylated histone H3 data are reprised from Chapter 3, Figure 3-4. Relative fragment depth values for FAIRE were calculated using input DNA as the normalization control; acetylated H3 data was obtained from the Gene Expression Omnibus series GSE 13415, sample GSM338433 and was converted to WIG format for display in IGV. Normalized fragment depth plots of FAIRE are plotted using input DNA as the normalization control. X-axis indicated physical location on the chromosome. Y-axis indicates normalized fragment depth per base pair (FAIRE) or acetyl-H3 to H3 ratios. Red and blue arrows below panel indicate the location and length of polycistronic gene clusters in chromosome 1. (B) Representation of the location of qPCR amplicons used in FAIRE analysis. Values in parentheses represent the distance from the dSSR. (C) FAIRE-qPCR data, represented as the enrichment in FAIRE DNA over an equivalent amount of input DNA. Values are normalized to the dSSR-distal gene *LmjF01.0400* (Far right). The directions of polycistronic gene clusters are indicated with blue and red arrows below the graph.

Figure 4-6. dSSR deletion and substition constructs demonstrate that the core dSSR of chromosome 1 lacks *cis*-regulatory elements. (A) Depiction of the deletions and substitutions present in the pLUCv2-Δ489 and pLUCv2-Δ489 + S constructs. Dotted arrows indicate the location of the 489 bp deletion; the gray box indicates the location of the 489 bp substitution from *T. brucei*-derived genic sequence. The black boxes indicate the location of the minimal functional splice acceptor sequences identified in Fig. 4-2. (B) FLuc activity of Δ489 + S clones. Graph is plotted identically to those in Fig. 4-3, with the exception that the error bars represent the average of all clones containing the WT dSSR or all clones containing the Δ489 + S substitution. (C) RLuc mRNA levels, quantified as described in Fig. 4-3. Error bars indicate the same as in panel (B). Statistical analysis was done using a Student's T-test; *, $p < 0.05$; **, $p < 0.005$.

Supplemental Figure S4-1. Plasmid maps for pLUC v2. ORFs are indicated with red or blue arrows, which correspond to the labeling of the ORFs in the reporter in Figure 4-1. InterORF regions, which contain sequences required for *trans*-splicing, are indicated with black arcs. The ampicillin resistance gene and origin of replication required for propagation of the plasmid in bacteria are indicated in gray. The XbaI site located between the *RLuc* and *FLuc* genes is indicated on the figure; the dSSR is cloned in at this locus in all derivatives of this plasmid.

Supplemental Figure S4-2. Correlation of reporter mRNA levels with FLuc, but not RLuc activity. (A) Comparison of FLuc mRNA levels with observed FLuc activity in episomal and integrated pLUC clones. Line indicates the linear fit; the slope, intercept, and R^2 values are indicated. (C) RLuc activity data from integrated and episomal pLUC v2 constructs containing the WT dSSR demonstrates that the RLuc activity from integrated and episomal lines bearing the pLUC v2 constructs is nearly at background levels and is much lower than the v1 constructs depicted in Fig. 4-2. Graph is labeled as described in Figure 4-3. The signal over untransfected *L. major* FV1 is described at the top of the graph.

Supplemental Figure S4-3. Identification of the splice acceptor dinucleotides in studies mapping the minimal functional splice acceptor by RT-PCR. (A) Schematic of splice acceptor mapping studies. The top portion depicts the genomic or extrachromosomal *RLuc* gene, flanked by splice acceptor dinucleotides (depicted as AG). Polycistronic transcription generates a pre-mRNA containing the AG dinucleotide. This pre-mRNA is *trans*-spliced with the spliced leader (SL) RNA, depicted as a yellow box, which contains the 5' cap, depicted as a black circle. In the mature transcript, the identity of the splice acceptor dinucleotide is determined by generating cDNA and using PCR primers complementary to the SL sequence and to the ORF of interest. Usage of the major splice leader dinucleotide is confirmed if the PCR amplicon is the expected size, or using sequencing-based methods. (B) Assessment of PCR amplicons generated from SL-primed PCR of cDNA derived from lines bearing the RLuc ½ episomes described in Figure 4-2. Relevant bands in the dsDNA standard are labeled on the left; no reverse transcriptase controls shown that the product is specific to cDNA and is not present in genomic DNA. A no cDNA control is present as a negative control. One clone from each construct depicted in Figure 4-2 was screened; note the presence of an approximately 386 bp amplicon in R-series constructs R301 and R401, and an amplicon of approximately 248 bp in L-series constructs. (C-E) Sanger sequencing data of selected PCR amplicons shown in (B). PCR amplicons were purified by gel extraction and sequenced using the reverse, RLuc-specific primer. Shown are alignments of the actual amplicon to the expected mRNA sequence, generated from *in silico* reconstruction based on the major splice acceptor dinucleotide of the endogenous dSSR-flanking genes. Data are shown for RLuc-L84 (C), RLuc-L182 (D), and RLuc-R301 (E).

Supplemental Figure S4-4. Splice leader (SL)-primed RT-PCR reveals the utilization of the expected splice acceptor dinucleotides for the *FLuc* and *RLuc* mRNAs in the integrated, WT dSSR, but identifies a cryptic splice acceptor dinucleotide in some Δ489 lines. We utilized the spliced leader-primed RT-PCR technique described in Supplemental Figure S4-3 to identify the splice acceptor dinucleotides utilized in the integrated and episomal bidirectional reporters. The expected size of the band is indicated above the gel. A second band identified in the Δ489 + S lines maps to a cryptic splice acceptor dinucleotide present within the stuffer sequence; clones utilizing this splice acceptor were omitted from further studies.

Supplemental Figure S4-5. Comparisons of the *T. brucei*-derived genic stuffer and the WT dSSR sequence between minimal functional SA sequences. (A) ClustalW alignment of the WT dSSR and the *T. brucei*-derived stuffer sequence. Black regions indicate homology between these sequences. (B) Predicted curvature (red) and bendability (green) of WT dSSR and *T. brucei*-derived stuffer sequences, plotted using the bend.it algorithm (27). Y-axis indicates the degrees per helical turn or the arbitrary bendability of the DNA sequence; X-axis indicates the position along the sequence. (C) Melting temperature analysis of WT dSSR (blue) and *T. brucei*-derived stuffer (red) sequences, calculated using the emboss dan algorithm (28). Y-axis indicates the Tm (in degrees Celsius) of a 10 bp sliding window; X-axis is as described in (B). (D) (G+C) content was calculated using the emboss dan algorithm, calculated across a 10 bp sliding window (28). Y-axis indicates the percent (G+C); X-axis and labels are as described in (B).

Table 4-1. Transfection efficiencies for episomal and linearized constructs used in this study. The number of colonies was normalized to the number of colonies obtained per 10 µg of episomal or linearized DNA transfected. The ratio of linearized to episomal transfectants is shown in the last column; numbers below the numbers transfected indicates the number of independent transfections performed.

Table 4-2. Primer sequences used for construction of pLUC plasmids. All sequences are listed from 5' to 3'; the templates and intended amplicon are also listed. The destination plasmid indicates the pools of DNA fragments used for genetic recombination in *S. cerevisiae*.

Table 4-3. Primer sequences used for modification of pLUC plasmids or for amplification of dSSR fragments. Primers are listed as described in Table 4-2.

Table 4-4. Accessory primers used for validation of proper integration, for splice acceptor dinucleotide mapping, or for quantitative PCR and reverse transcriptase PCR. Primers are listed as described in Table 4-2; see the Materials and Methods section for information regarding the usage of these primers

References

1. Günzl A. The pre-mRNA splicing machinery of trypanosomes: complex or simplified? Eukaryot Cell. American Society for Microbiology; 2010 Aug;9(8):1159–70.

2. Gunzl A, Vanhamme L, Myler P. Transcription in trypanosomes: a different means to the end. In: Barry J, Mottram J, McCulloch R, Acosta-Serrano A, editors. Trypanosomes - After the Genome. Horizon Bioscience, Wymondham, Norfolk, UK; 2007. p. 177–208.

3. LeBowitz J, Smith H, Rusche L, Beverley S. Coupling of poly(A) site selection and trans-splicing in Leishmania. Genes Dev. 1993 Jun 1;7(6):996–1007.

4. Matthews K, Tschudi C, Ullu E. A common pyrimidine-rich motif governs trans-splicing and polyadenylation of tubulin polycistronic pre-mRNA in trypanosomes. Genes Dev. 1994 Feb 15;8(4):491–501.

5. Martínez-Calvillo S, Yan S, Nguyen D, Fox M, Stuart K, Myler P. Transcription of Leishmania major Friedlin chromosome 1 initiates in both directions within a single region. Mol Cell. 2003;11(5):1291–9.

6. Kolev N, Franklin J, Carmi S, Shi H, Michaeli S, Tschudi C. The transcriptome of the human pathogen Trypanosoma brucei at single-nucleotide resolution. PLoS Pathog. 2010 Jan;6(9):e1001090.

7. Siegel TN, Hekstra DR, Wang X, Dewell S, Cross G a M. Genome-wide analysis of mRNA abundance in two life-cycle stages of Trypanosoma brucei and identification of splicing and polyadenylation sites. Nucleic Acids Res. 2010 Aug;38(15):4946–57.

8. Ruan J, Arhin G, Ullu E, Tschudi C. Functional Characterization of a Trypanosoma brucei TATA-Binding Protein-Related Factor Points to a Universal Regulator of Transcription in Trypanosomes. Mol Cell Biol. 2004;24(21):9610–8.

9. Thomas S, Green A, Sturm N, Campbell D, Myler P. Histone acetylations mark origins of polycistronic transcription in Leishmania major. BMC Genomics. 2009 Jan;10:152.

10. Ivens A, Peacock C, Worthey E, Murphy L, Aggarwal G, Berriman M, et al. The genome of the kinetoplastid parasite, Leishmania major. Science. 2005 Jul 15;309(5733):436–42.

11. McDonagh PD, Myler PJ, Stuart K. The unusual gene organization of Leishmania major chromosome 1 may reflect novel transcription processes. Nucleic Acids Res. 2000 Jul 15;28(14):2800–3.

12. Smircich P, Forteza D, El-Sayed NM, Garat B. Genomic analysis of sequence-dependent DNA curvature in Leishmania. PLoS One. 2013 Jan;8(4):e63068.

13. Siegel T, Hekstra D, Kemp L, Figueiredo L, Lowell J, Fenyo D, et al. Four histone variants mark the boundaries of polycistronic transcription units in Trypanosoma brucei. Genes Dev. 2009 May 1;23(9):1063–76.

14. Struhl K. Naturally occurring poly(dA-dT) sequences are upstream promoter elements for constitutive transcription in yeast. Proc Natl Acad Sci USA. 1985;82(December):8419–23.

15. Iyer V, Struhl K. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. EMBO J. 1995 Jun 1;14(11):2570–9.

16. Arnone MI, Dmochowski IJ, Gache C. Using reporter genes to study cis-regulatory elements. Methods Cell Biol. 2004 Jan;74:621–52.

17. Curotto de Lafaille M, Laban A, Wirth D. Gene expression in Leishmania: analysis of essential 5’ DNA sequences. Proc Natl Acad Sci USA. 1992 Apr 1;89(7):2703–7.

18. Dubessay P, Ravel C, Bastien P, Crobu L, Dedet J-P, Pagès M, et al. The switch region on Leishmania major chromosome 1 is not required for mitotic stability or gene expression, but appears to be essential. Nucleic Acids Res. 2002 Sep 1;30(17):3692–7.

19. Sherf BA. Dual-luciferase reporter assay: an advanced co-reporter technology integrating firefly and Renilla luciferase assays. Promega Notes. 1996;57:2–9.

20. Robinson K, Beverley S. Improvements in transfection efficiency and tests of RNA interference (RNAi) approaches in the protozoan parasite Leishmania. Mol Biochem Parasitol. 2003 May;128(2):217–28.

21. Murta S, Vickers T, Scott D, Beverley S. Methylene tetrahydrofolate dehydrogenase/cyclohydrolase and the synthesis of 10-CHO-THF are essential in Leishmania major. Mol Microbiol. 2009 Mar;71(6):1386–401.

22. Hogan G, Lee C, Lieb J. Cell cycle-specified fluctuation of nucleosome occupancy at gene promoters. PLoS Genet. 2006 Sep 22;2(9):e158.

23. Giresi P, Kim J, McDaniell R, Iyer V, Lieb J. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res. 2007 Jun;17(6):877–85.

24. Song L, Zhang Z, Grasfeder L, Boyle A, Giresi P, Lee B, et al. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res. 2011 Oct;21(10):1757–67.

25. Kuwayama H, Obara S, Morio T, Katoh M, Urushihara H, Tanaka Y. PCR-mediated generation of a gene disruption construct without the use of DNA ligase and plasmid vectors. Nucleic Acids Res. Oxford University Press; 2002;30(2):e2.

26. Livak K, Schmittgen T. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods. 2001 Dec;25(4):402–8.

27. Vlahovicek K. DNA analysis servers: plot.it, bend.it, model.it and IS. Nucleic Acids Res. 2003 Jul 1;31(13):3686–7.

28. Rice P, Longden I, Bleasby A. The European Molecular Biology Open Software Suite EMBOSS : The European Molecular Biology Open Software Suite. Trends Genet. 2000;16(6):2–3.

.

Figure 4-1.

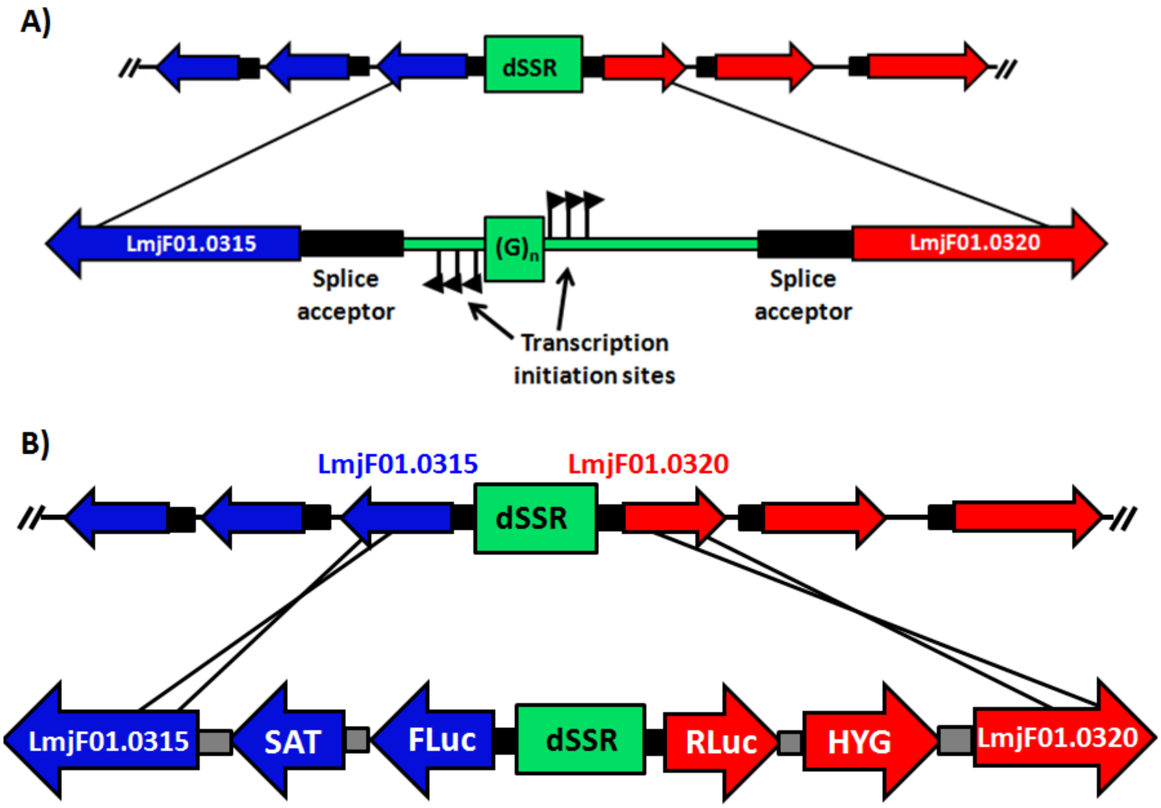
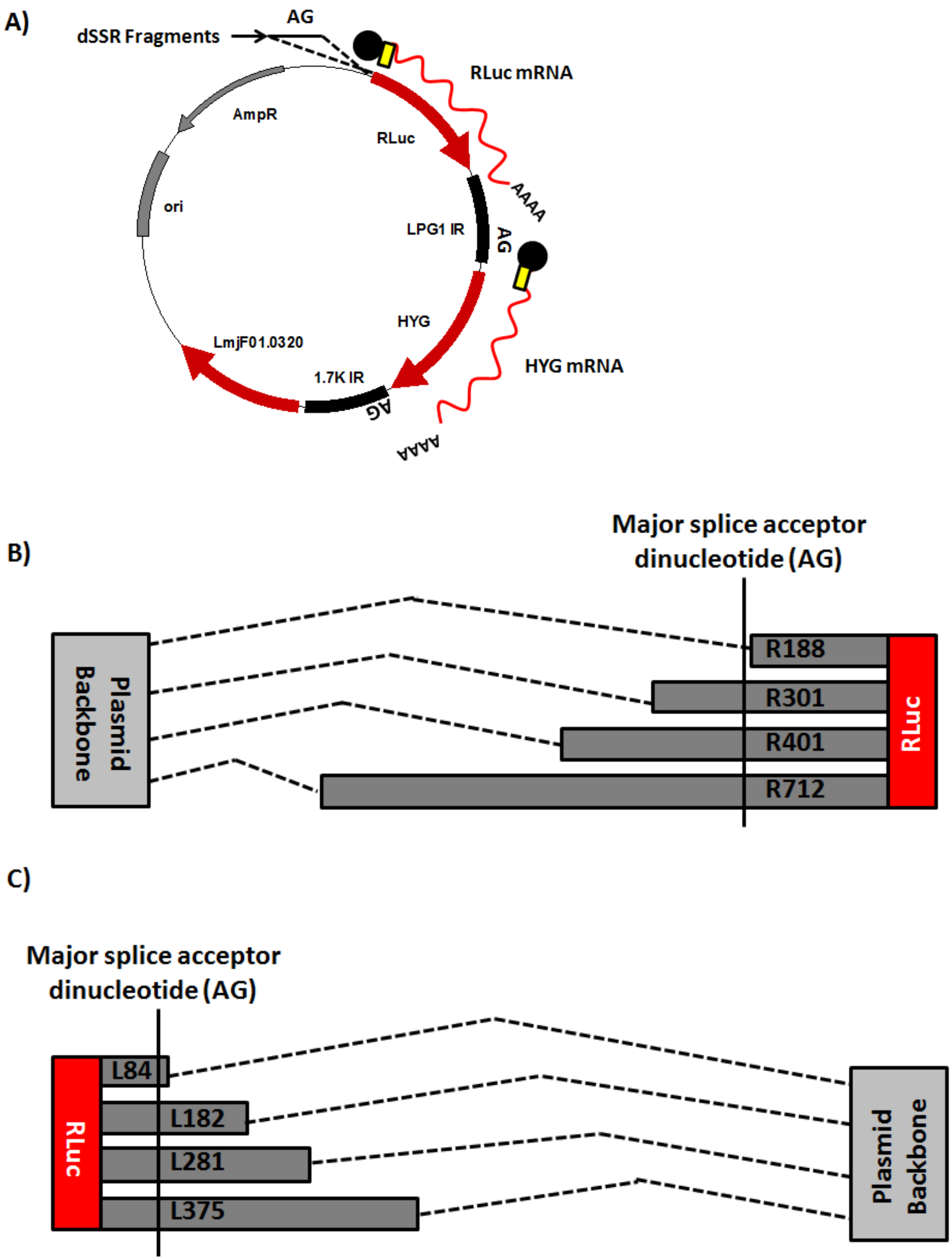

A)
dSSR
LmjF01.0315
LmjF01.0320
(G)$_n$
Splice acceptor
Splice acceptor
Transcription initiation sites
B)
LmjF01.0315
LmjF01.0320
dSSR
LmjF01.0315
SAT
FLuc
dSSR
RLuc
HYG
LmjF01.0320

Figure 4-2.


A)
dSSR Fragments
AG
RLuc mRNA
AmpR
RLuc
AAAA
ori
LPG1 IR
AG
HYG
LmjF01.0320
HYG mRNA
1.7K IR
AG
AAAA
B)
Major splice acceptor dinucleotide (AG)
Plasmid Backbone
R188
R301
R401
R712
RLuc
C)
Major splice acceptor dinucleotide (AG)
RLuc
L84
L182
L281
L375
Plasmid Backbone

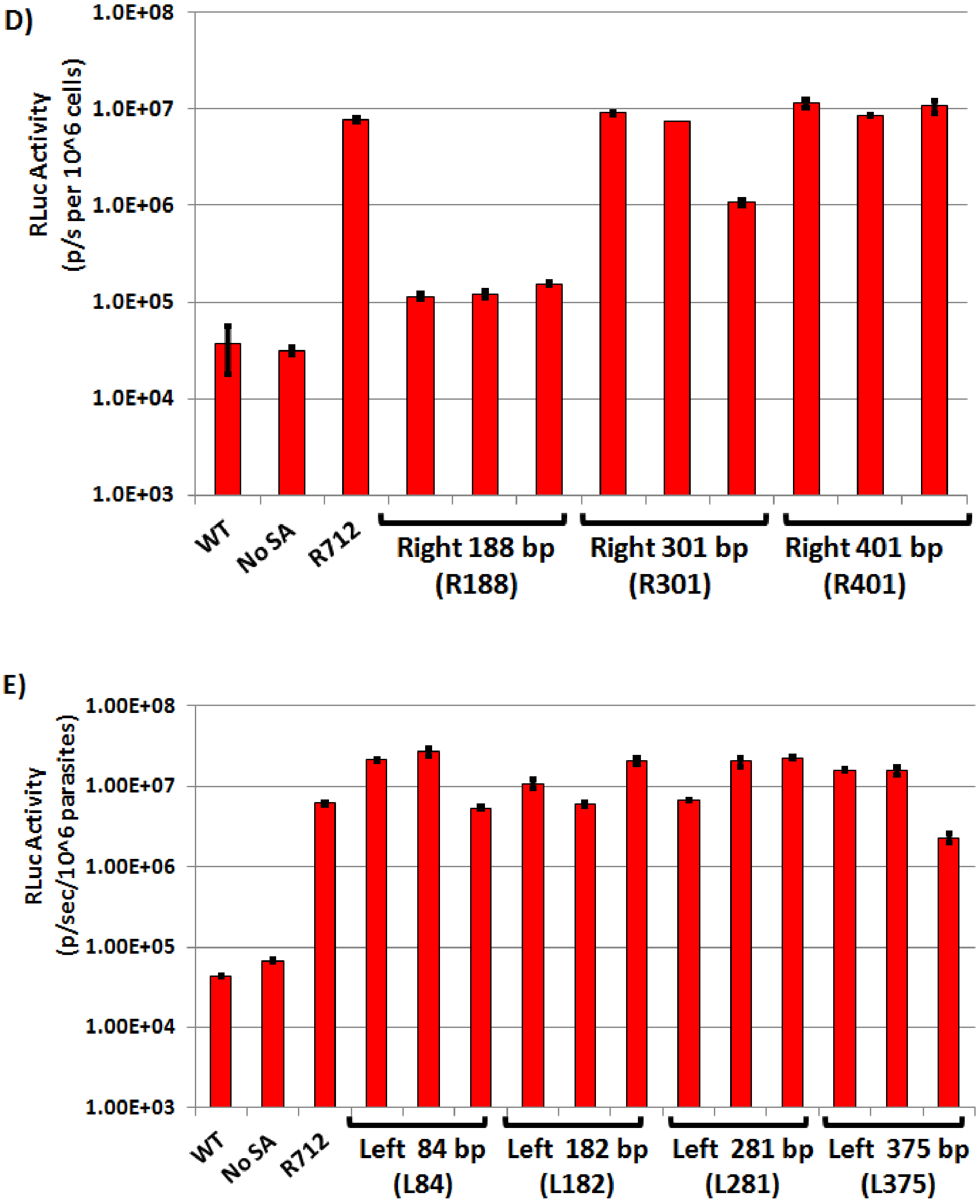
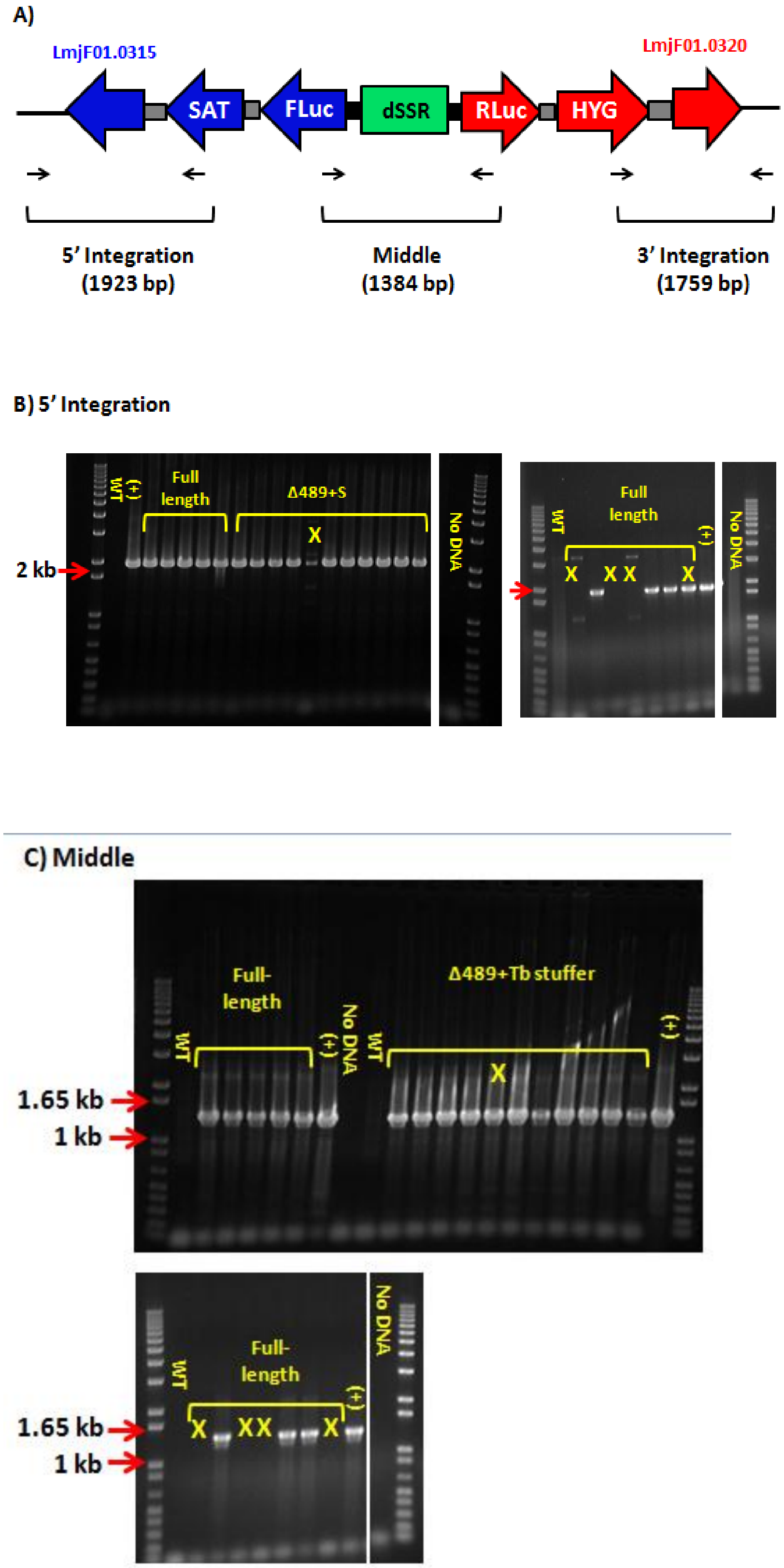

D)
RLuc Activity
(p/s per 10^6 cells)
1.0E+08
1.0E+07
1.0E+06
1.0E+05
1.0E+04
1.0E+03
WT
No SA
R712
Right 188 bp
(R188)
Right 301 bp
(R301)
Right 401 bp
(R401)
E)
RLuc Activity
(p/sec/10^6 parasites)
1.00E+08
1.00E+07
1.00E+06
1.00E+05
1.00E+04
1.00E+03
WT
No SA
R712
Left 84 bp
(L84)
Left 182 bp
(L182)
Left 281 bp
(L281)
Left 375 bp
(L375)

Figure 4-3.



A)

B) 5' Integration

C) Middle

D) 3' Integration

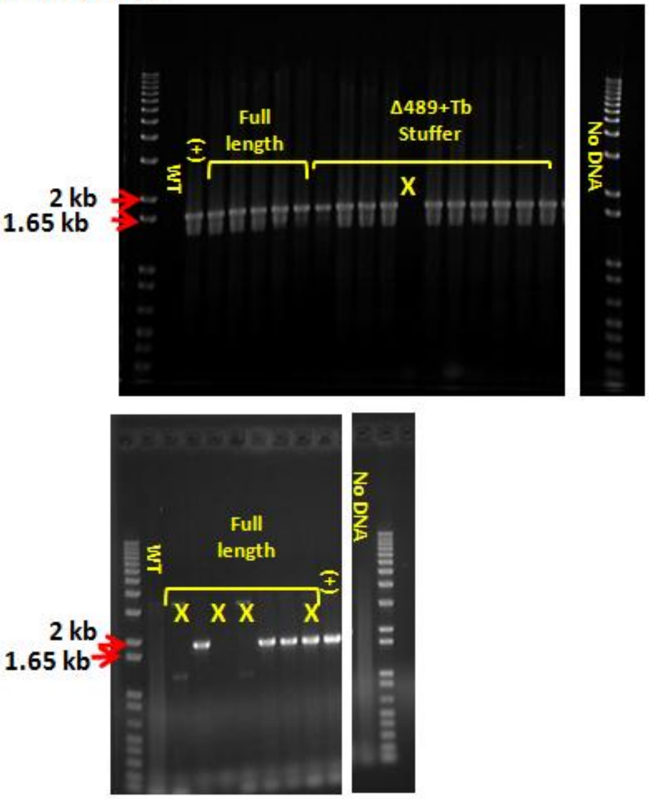


Figure 4-4.

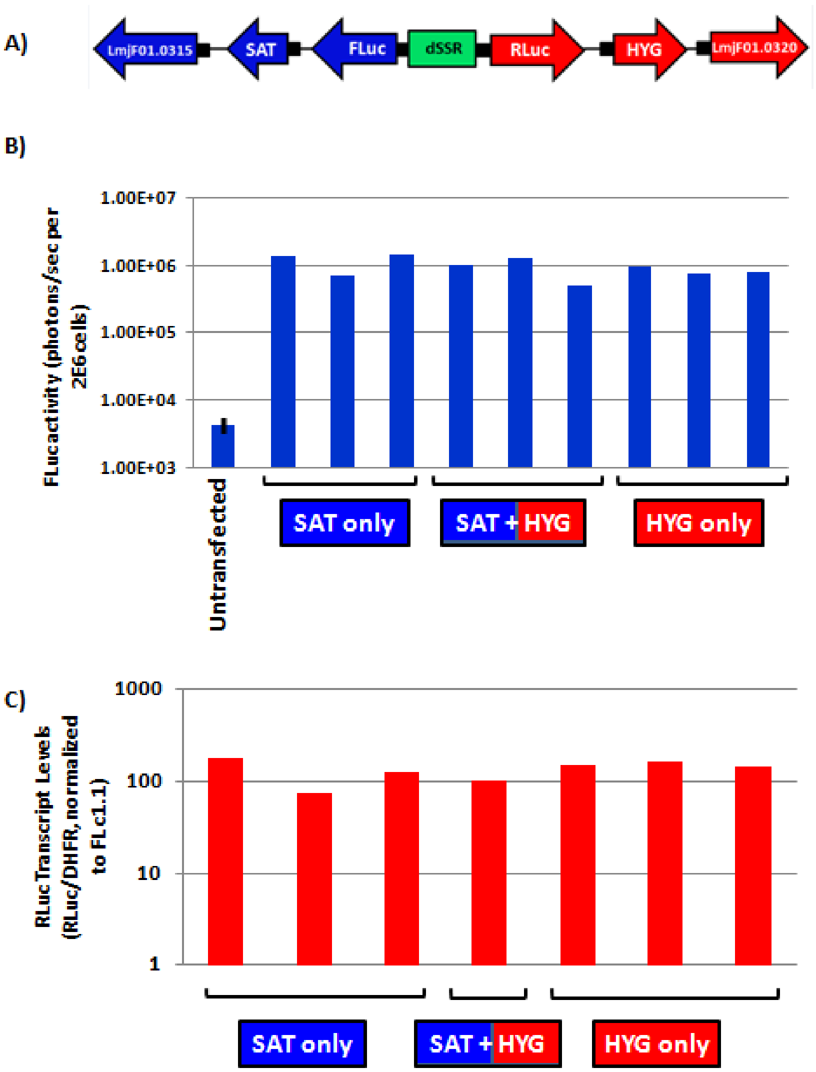

A)
LmjF01.0315
SAT
FLuc
dSSR
RLuc
HYG
LmjF01.0320
B)
FLuc activity (photons/sec per 2E6 cells)
1.00E+07
1.00E+06
1.00E+05
1.00E+04
1.00E+03
Untransfected
SAT only
SAT + HYG
HYG only
C)
RLuc Transcript Levels (RLuc/DHFR, normalized to FL c1.1)
1000
100
10
1
SAT only
SAT + HYG
HYG only

Figure 4-5.

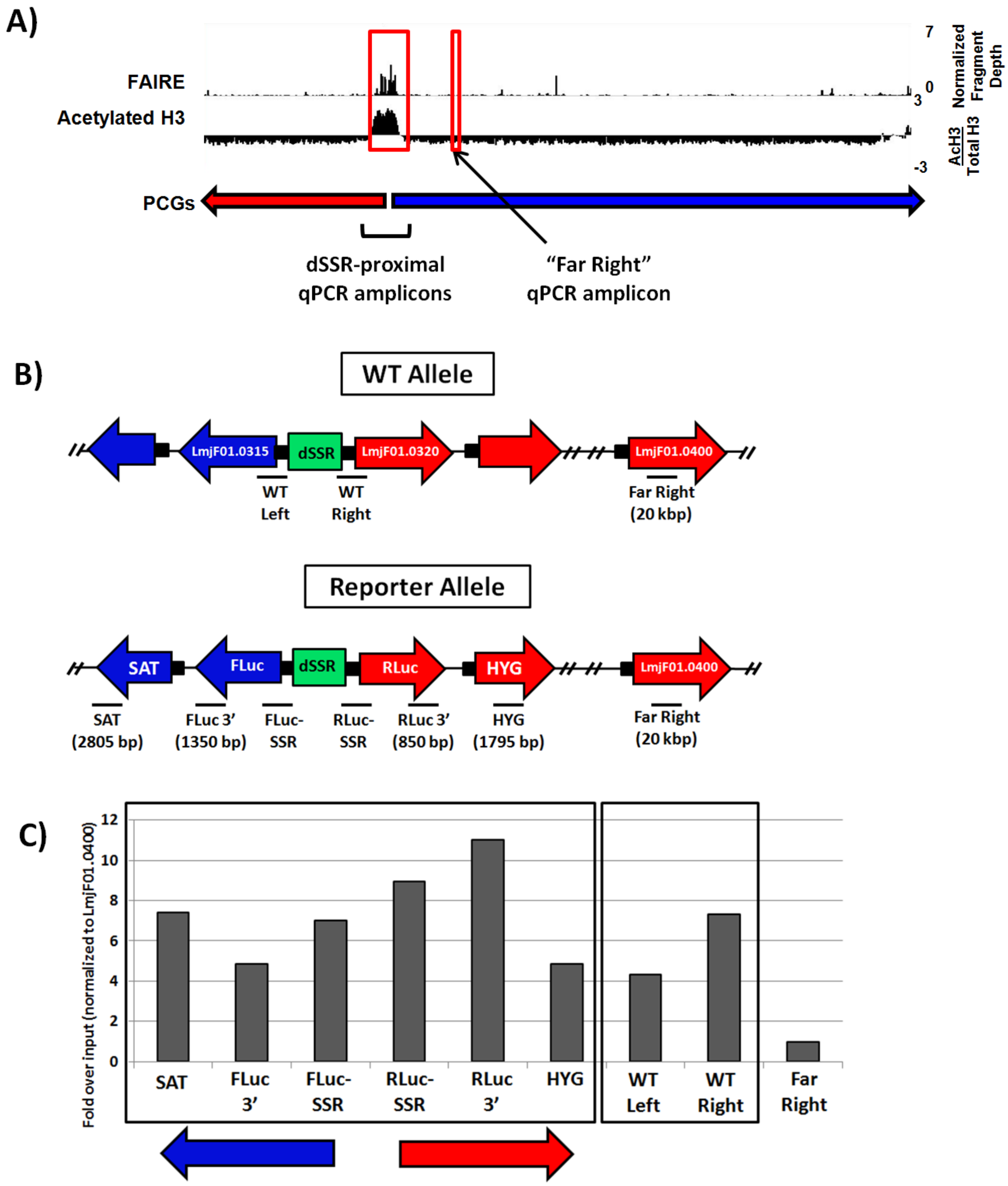
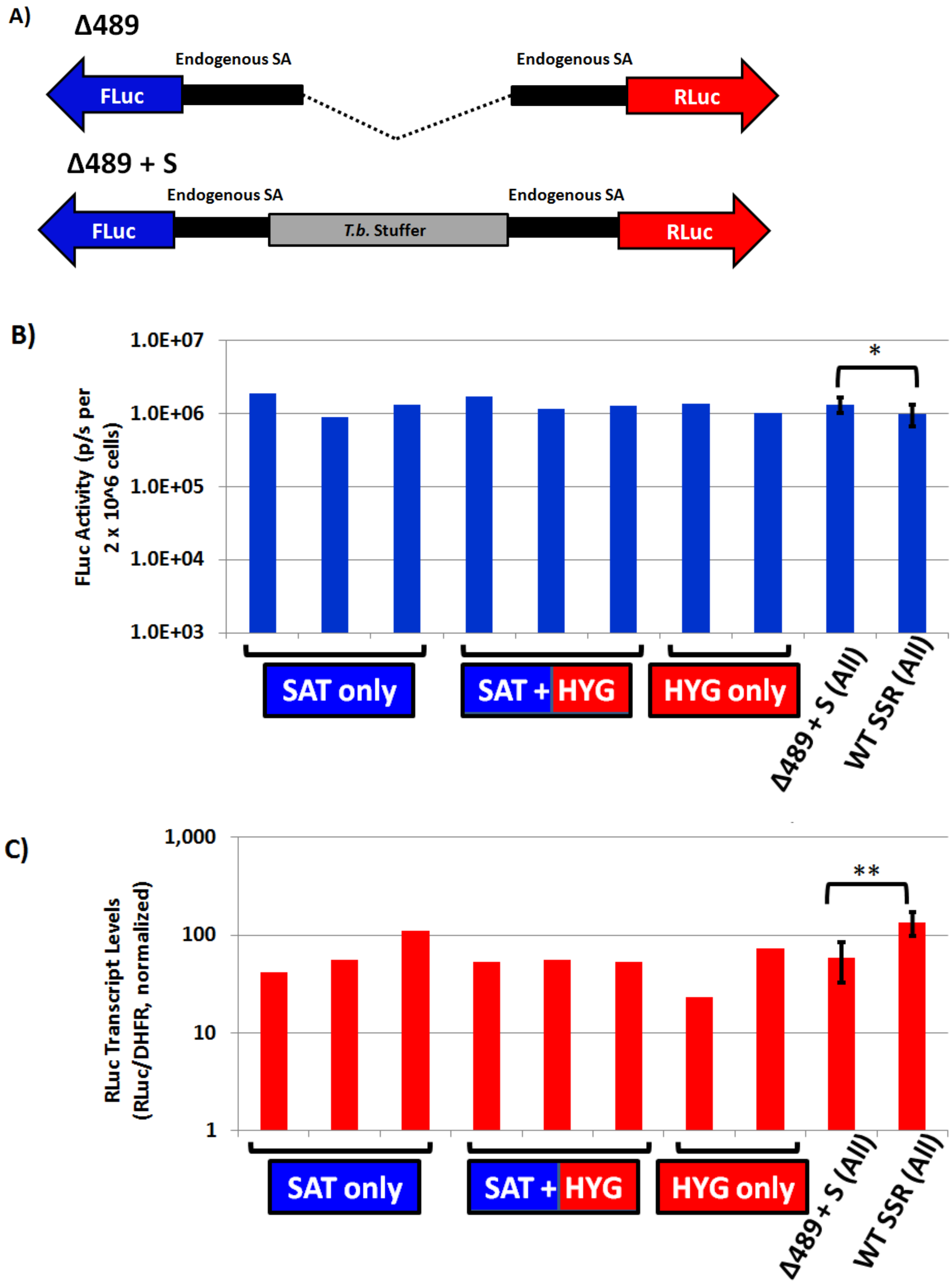
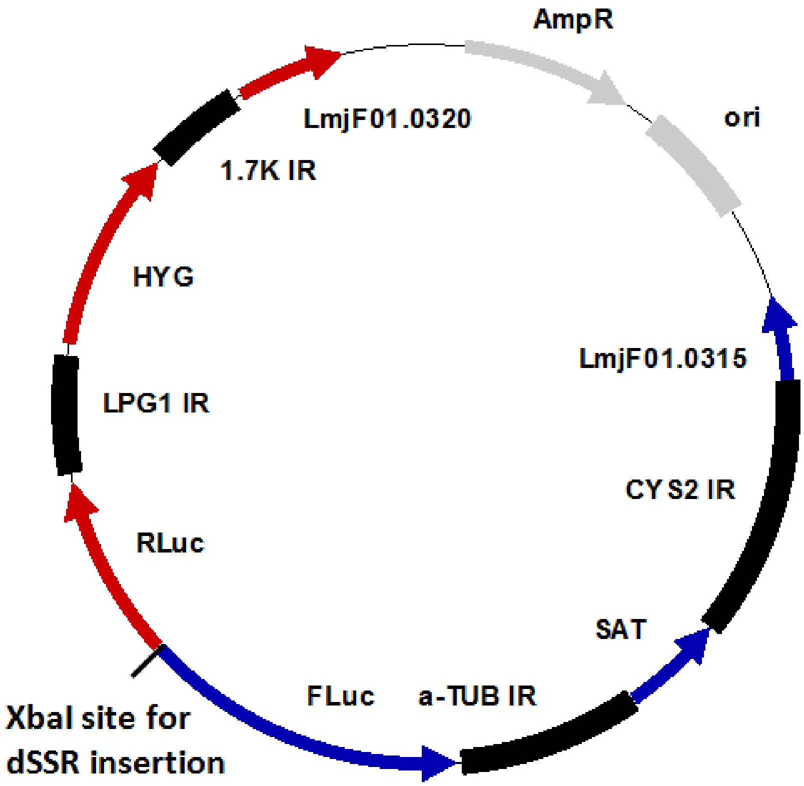

A)
FAIRE
Acetylated H3
7
0
3
-3
Normalized Fragment Depth
AcH3 / Total H3
PCGs
dSSR-proximal qPCR amplicons
"Far Right" qPCR amplicon
B)
WT Allele
LmjF01.0315
dSSR
LmjF01.0320
LmjF01.0400
WT Left
WT Right
Far Right (20 kbp)
Reporter Allele
SAT
FLuc
dSSR
RLuc
HYG
LmjF01.0400
SAT (2805 bp)
FLuc 3' (1350 bp)
FLuc-SSR
RLuc-SSR
RLuc 3' (850 bp)
HYG (1795 bp)
Far Right (20 kbp)
C)
Fold over input (normalized to LmjF01.0400)
12
10
8
6
4
2
0
SAT
FLuc 3'
FLuc-SSR
RLuc-SSR
RLuc 3'
HYG
WT Left
WT Right
Far Right

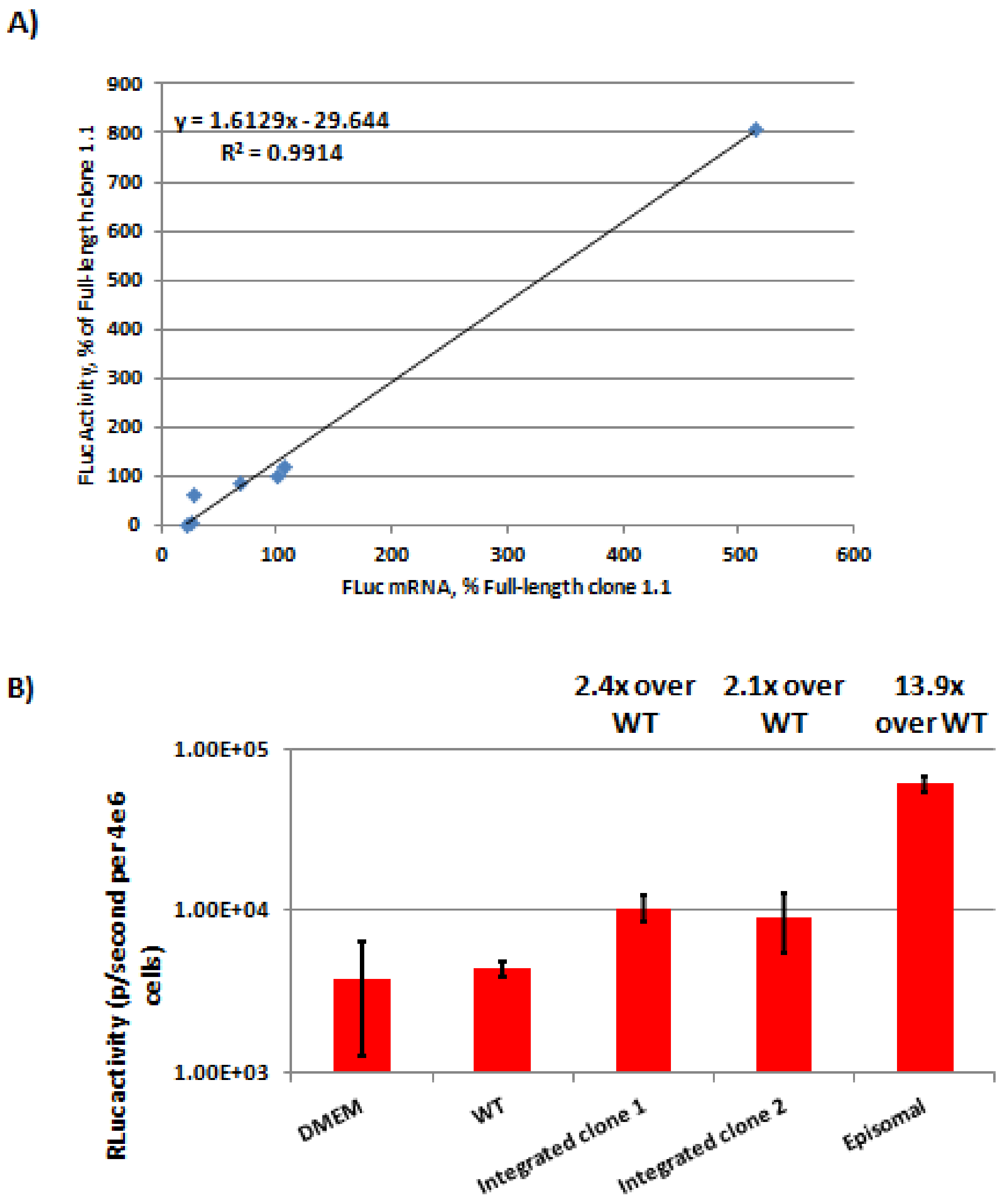
Figure 4-6.


A)
Δ489
Endogenous SA
Endogenous SA
FLuc
RLuc
Δ489 + S
Endogenous SA
Endogenous SA
FLuc
T.b. Stuffer
RLuc
B)
FLuc Activity (p/s per 2 x 10^6 cells)
1.0E+07
1.0E+06
1.0E+05
1.0E+04
1.0E+03
*
SAT only
SAT + HYG
HYG only
Δ489 + S (All)
WT SSR (All)
C)
RLuc Transcript Levels (RLuc/DHFR, normalized)
1,000
100
10
1
**
SAT only
SAT + HYG
HYG only
Δ489 + S (All)
WT SSR (All)

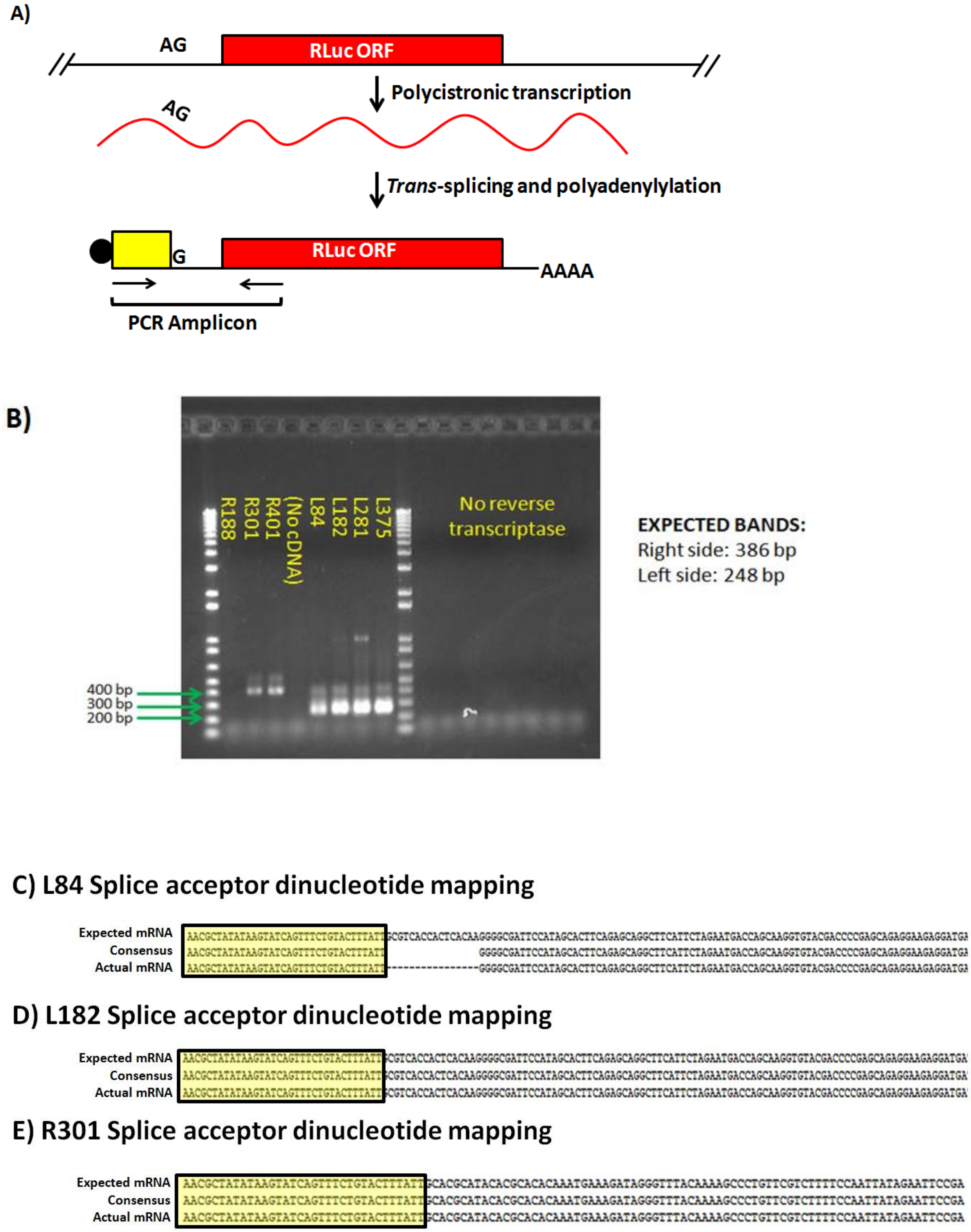
Supplemental Figure 4-1.



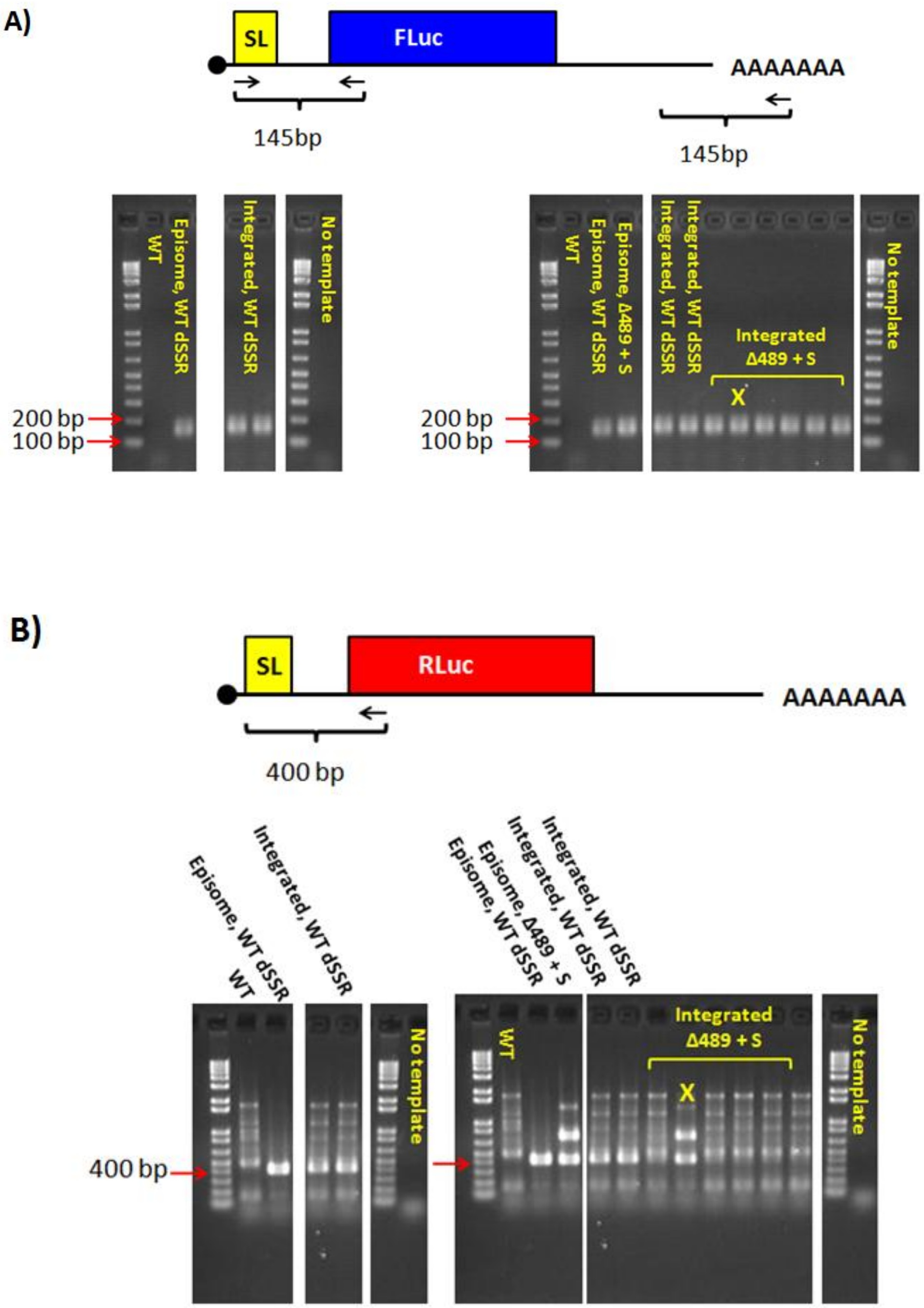
Supplemental Figure S4-2.



Supplemental Figure S4-3.

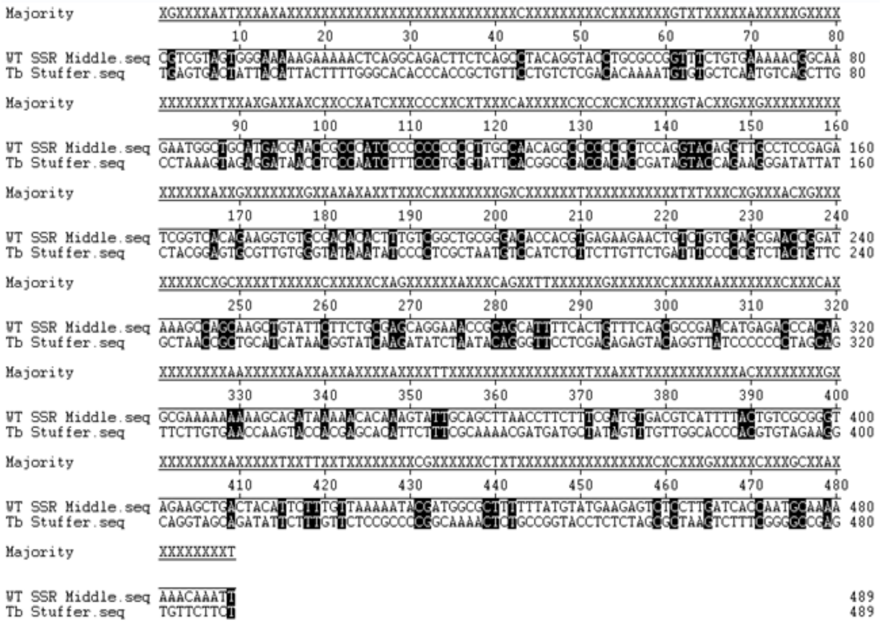

A)
AG
RLuc ORF
Polycistronic transcription
AG
Trans-splicing and polyadenylylation
G
RLuc ORF
AAAA
PCR Amplicon
B)
R188
R301
R401
(No cDNA)
L84
L182
L281
L375
No reverse transcriptase
400 bp
300 bp
200 bp
EXPECTED BANDS:
Right side: 386 bp
Left side: 248 bp
C) L84 Splice acceptor dinucleotide mapping
Expected mRNA AACGCTATATAAGTATCAGTTTCTGTACTTTATTGCGTCACCACTCACAAGGGGCGATTCCATAGCACTTCAGAGCAGGCTTCATTCTAGAATGACCAGCAAGGTGTACGACCCCGAGCAGAGGAAGAGGATGA
Consensus AACGCTATATAAGTATCAGTTTCTGTACTTTATT GGGGCGATTCCATAGCACTTCAGAGCAGGCTTCATTCTAGAATGACCAGCAAGGTGTACGACCCCGAGCAGAGGAAGAGGATGA
Actual mRNA AACGCTATATAAGTATCAGTTTCTGTACTTTATT----------------GGGGCGATTCCATAGCACTTCAGAGCAGGCTTCATTCTAGAATGACCAGCAAGGTGTACGACCCCGAGCAGAGGAAGAGGATGA
D) L182 Splice acceptor dinucleotide mapping
Expected mRNA AACGCTATATAAGTATCAGTTTCTGTACTTTATTGCGTCACCACTCACAAGGGGCGATTCCATAGCACTTCAGAGCAGGCTTCATTCTAGAATGACCAGCAAGGTGTACGACCCCGAGCAGAGGAAGAGGATGA
Consensus AACGCTATATAAGTATCAGTTTCTGTACTTTATTGCGTCACCACTCACAAGGGGCGATTCCATAGCACTTCAGAGCAGGCTTCATTCTAGAATGACCAGCAAGGTGTACGACCCCGAGCAGAGGAAGAGGATGA
Actual mRNA AACGCTATATAAGTATCAGTTTCTGTACTTTATTGCGTCACCACTCACAAGGGGCGATTCCATAGCACTTCAGAGCAGGCTTCATTCTAGAATGACCAGCAAGGTGTACGACCCCGAGCAGAGGAAGAGGATGA
E) R301 Splice acceptor dinucleotide mapping
Expected mRNA AACGCTATATAAGTATCAGTTTCTGTACTTTATTGCACGCATACACGCACACAAATGAAAGATAGGGTTTACAAAAGCCCTGTTCGTCTTTTCCAATTATAGAATTCCGA
Consensus AACGCTATATAAGTATCAGTTTCTGTACTTTATTGCACGCATACACGCACACAAATGAAAGATAGGGTTTACAAAAGCCCTGTTCGTCTTTTCCAATTATAGAATTCCGA
Actual mRNA AACGCTATATAAGTATCAGTTTCTGTACTTTATTGCACGCATACACGCACACAAATGAAAGATAGGGTTTACAAAAGCCCTGTTCGTCTTTTCCAATTATAGAATTCCGA

Supplemental Figure S4-4.



Supplemental Figure S4-5.

A)



```
Majority           XGXXXXAXTXXXAXAXXXXXXXXXXXXXXXXXXXXXXXXXXCXXXXXXXXXCXXXXXXXGTXTXXXXXAXXXXXGXXXX
                            10        20        30        40        50        60        70        80
WT SSR Middle.seq  CGTCGTAGTGGGAAAAAGAAAAACTCAGGCAGACTTCTCAGCCTACAGGTACCTGCGCCGGTTTCTGTGAAAAACGGCAA  80
Tb Stuffer.seq     TGAGTGACTATTACATTACTTTTGGGCACACCCACCGCTGTTCCTGTCTCGACACAAAATGTGTGCTCAATGTCAGCTTG  80

Majority           XXXXXXXTXXAXGAXXAXCXXCCXATCXXXCCCXXCXTXXXCAXXXXXCXCCXCXCXXXXXGTACXXGXXGXXXXXXXXX
                            90       100       110       120       130       140       150       160
WT SSR Middle.seq  GAATGGCTGCATGACGAACCGCCCATCCCCCCCCCCCTTGCCAACAGCCCCCCCCCTCCAGGTACAGGTTGCCTCCGAGA  160
Tb Stuffer.seq     CCTAAAGTAGAGGATAACCTCCCAATCTTTCCCTGCGTATTCACGGCGCACCACACCGATAGTACCAGAAGGGATATTAT  160

Majority           XXXXXXAXXGXXXXXXXGXXAXAXAXXTXXXCXXXXXXXXGXCXXXXXXTXXXXXXXXXXXTXTXXXCXGXXXACXGXXX
                           170       180       190       200       210       220       230       240
WT SSR Middle.seq  TCGGTCACAGAAGGTGTGCGACACACTTTGTCGGCTGCGGGACACCACGTGAGAAGAACTGTCTGTGCAGCGAACCGGAT  240
Tb Stuffer.seq     CTACGGAGTGCGTTGTGGGTATAAATATCCCCTCGCTAATGTCCATCTCTTCTTGTTCTGATTTCCCCCGTCTACTGTTC  240

Majority           XXXXXCXGCXXXXTXXXXXCXXXXXCXAGXXXXXXAXXXCAGXXTTXXXXXXGXXXXXXCXXXXXAXXXXXXXCXXXCAX
                           250       260       270       280       290       300       310       320
WT SSR Middle.seq  AAAGCCAGCAAGCTGTATTCTTCTGCGAGCAGGAAACCGCAGCATTTTCACTGTTTCAGCGCCGAACATGAGACCCACAA  320
Tb Stuffer.seq     GCTAACCGCTGCATCATAACGGTATCAAGATATCTAATACAGGGTTCCTCGAGAGAGTACAGGTTATCCCCCCCTAGCAG  320

Majority           XXXXXXXXAAXXXXXXXAXXAXXAXXXXAXXXXTTXXXXXXXXXXXXXXXXXTXXAXXTXXXXXXXXXXXACXXXXXXXXGX
                           330       340       350       360       370       380       390       400
WT SSR Middle.seq  GCGAAAAAAAAAGCAGATAAAAACACAAAGTATTGCAGCTTAACCTTCTTTCGATGTGACGTCATTTTACTGTCGCGGGT  400
Tb Stuffer.seq     TTCTTGTGAACCAAGTACCACGAGCACATTCTTTCGCAAAACGATGATGCTATAGTTTGTTGGCACCCACGTGTAGAAGG  400

Majority           XXXXXXXXAXXXXXTXXTTXXTXXXXXXXXXCGXXXXXXCTXTXXXXXXXXXXXXXXXXCXCXXXGXXXXXCXXXGCXXAX
                           410       420       430       440       450       460       470       480
WT SSR Middle.seq  AGAAGCTGACTACATTCTTTGTTAAAAATACGATGGCGCTTTTTTATGTATGAAGAGTCTCCTTGATCACCAATGCAAAA  480
Tb Stuffer.seq     CAGGTAGCAGATATTCTTTGTTCTCCGCCCCGGCAAAACTCTGCCGGTACCTCTCTAGCGCTAAGTCTTTCGGGGCCGAG  480

Majority           XXXXXXXXT


WT SSR Middle.seq  AAACAAATT                                                                         489
Tb Stuffer.seq     TGTTCTTCT                                                                         489
```

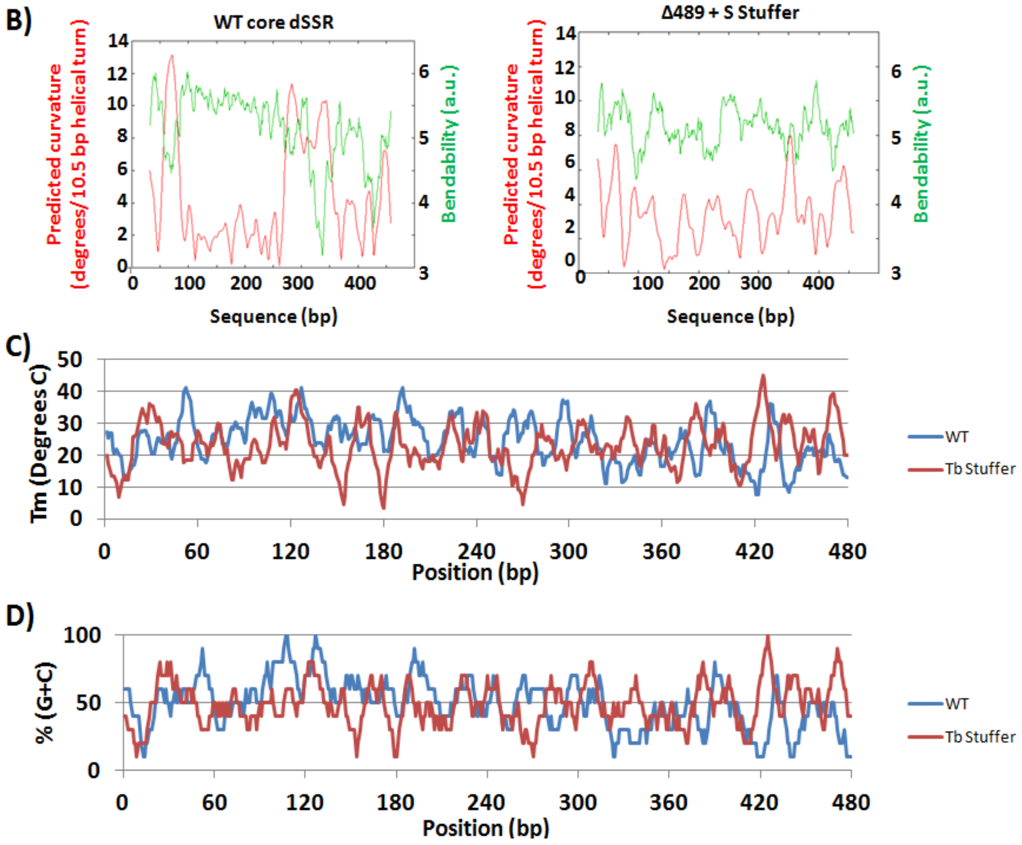
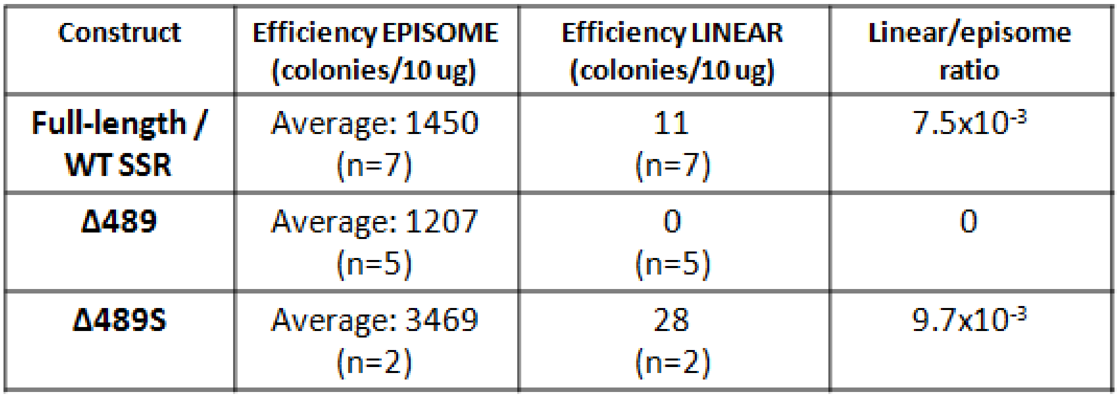
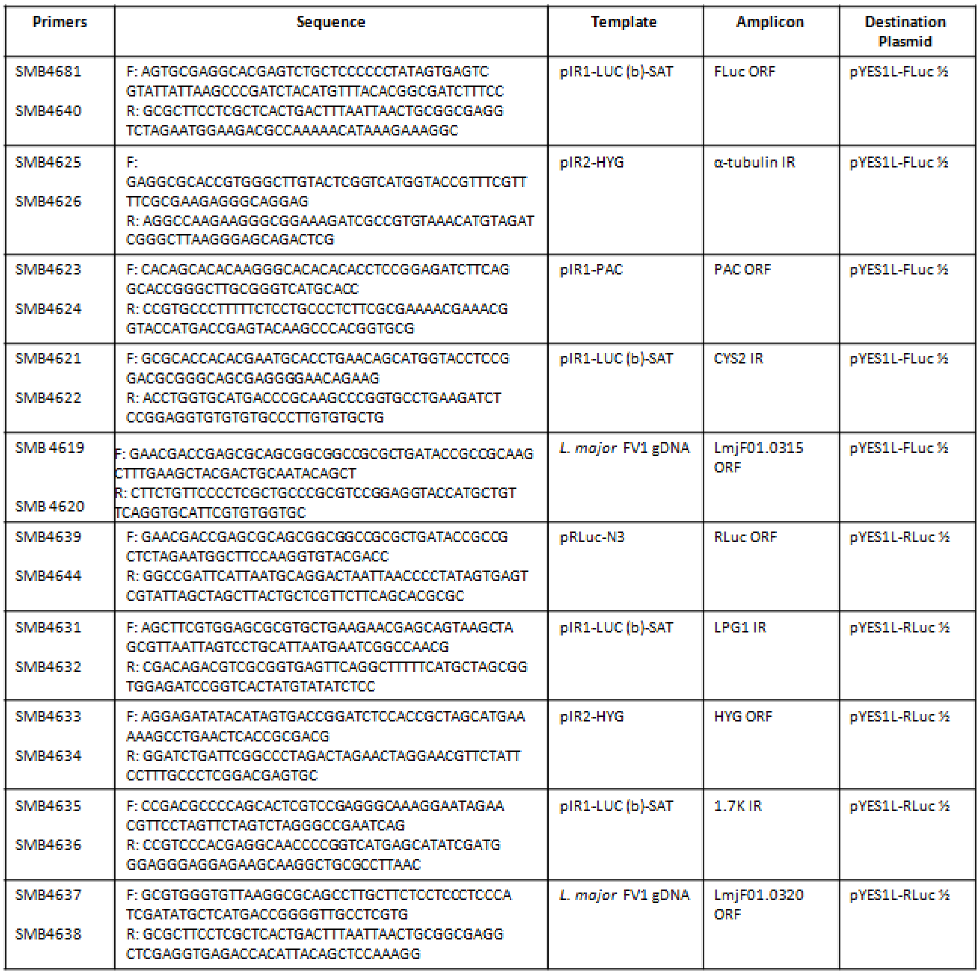


Table 4-1.



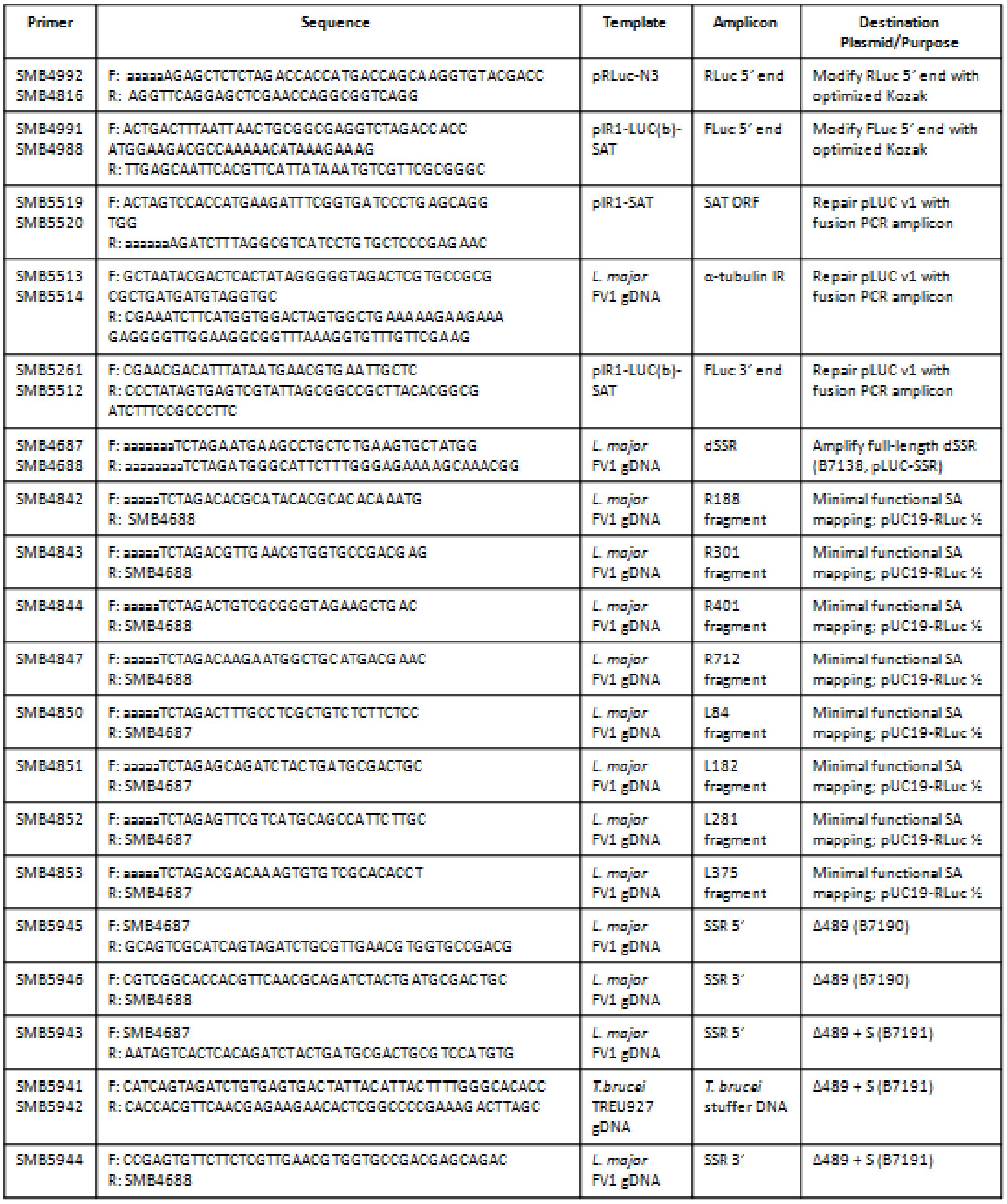
Construct	Efficiency EPISOME (colonies/10 ug)	Efficiency LINEAR (colonies/10 ug)	Linear/episome ratio
Full-length / WT SSR	Average: 1450 (n=7)	11 (n=7)	$7.5x10^{-3}$
Δ489	Average: 1207 (n=5)	0 (n=5)	0
Δ489S	Average: 3469 (n=2)	28 (n=2)	$9.7x10^{-3}$

Table 4-2.



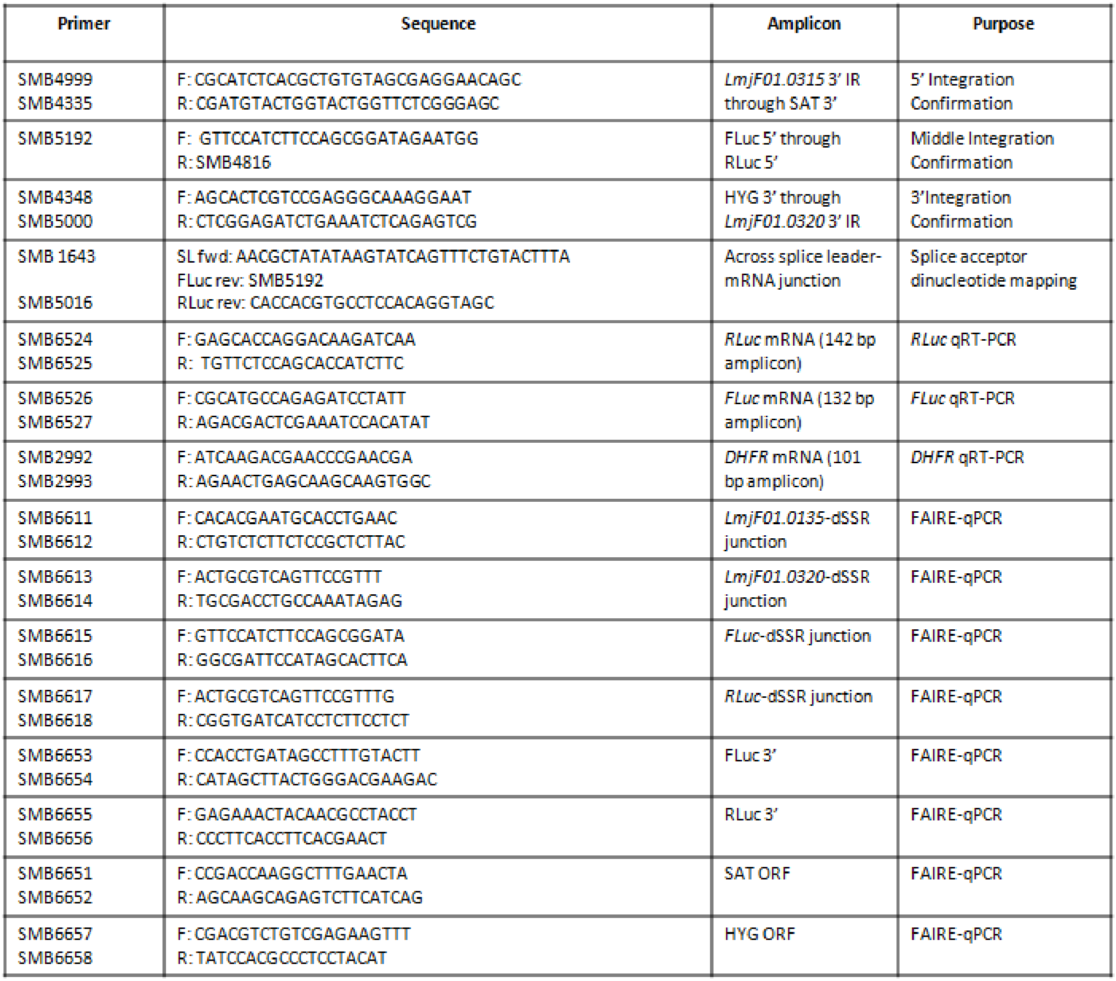
Primers	Sequence	Template	Amplicon	Destination Plasmid
SMB4681 SMB4640	F: AGTGCGAGGCACGAGTCTGCTCCCCCCTATAGTGAGTCGTATTATTAAGCCCGATCTACATGTTTACACGGCGATCTTTCC R: GCGCTTCCTCGCTCACTGACTTTAATTAACTGCGGCGAGGTCTAGAATGGAAGACGCCAAAAACATAAAGAAAGGC	pIR1-LUC (b)-SAT	FLuc ORF	pYES1L-FLuc ½
SMB4625 SMB4626	F: GAGGCGCACCGTGGGCTTGTACTCGGTCATGGTACCGTTTCGTTTTCGCGAAGAGGGCAGGAG R: AGGCCAAGAAGGGCGGAAAGATCGCCGTGTAAACATGTAGATCGGGCTTAAGGGAGCAGACTCG	pIR2-HYG	α-tubulin IR	pYES1L-FLuc ½
SMB4623 SMB4624	F: CACAGCACACAAGGGCACACACACCTCCGGAGATCTTCAGGCACCGGGCTTGCGGGTCATGCACC R: CCGTGCCCTTTTTCTCCTGCCCTCTTCGCGAAAACGAAACGGTACCATGACCGAGTACAAGCCCACGGTGCG	pIR1-PAC	PAC ORF	pYES1L-FLuc ½
SMB4621 SMB4622	F: GCGCACCACACGAATGCACCTGAACAGCATGGTACCTCCGGACGCGGGCAGCGAGGGGAACAGAAG R: ACCTGGTGCATGACCCGCAAGCCCGGTGCCTGAAGATCTCCGGAGGTGTGTGTGCCCTTGTGTGCTG	pIR1-LUC (b)-SAT	CYS2 IR	pYES1L-FLuc ½
SMB 4619 SMB 4620	F: GAACGACCGAGCGCAGCGGCGGCCGCGCTGATACCGCCGCAAGCTTTGAAGCTACGACTGCAATACAGCT R: CTTCTGTTCCCCTCGCTGCCCGCGTCCGGAGGTACCATGCTGTTCAGGTGCATTCGTGTGGTGC	*L. major* FV1 gDNA	LmjF01.0315 ORF	pYES1L-FLuc ½
SMB4639 SMB4644	F: GAACGACCGAGCGCAGCGGCGGCCGCGCTGATACCGCCGCTCTAGAATGGCTTCCAAGGTGTACGACC R: GGCCGATTCATTAATGCAGGACTAATTAACCCCTATAGTGAGTCGTATTAGCTAGCTTACTGCTCGTTCTTCAGCACGCGC	pRLuc-N3	RLuc ORF	pYES1L-RLuc ½
SMB4631 SMB4632	F: AGCTTCGTGGAGCGCGTGCTGAAGAACGAGCAGTAAGCTAGCGTTAATTAGTCCTGCATTAATGAATCGGCCAACG R: CGACAGACGTCGCGGTGAGTTCAGGCTTTTTCATGCTAGCGGTGGAGATCCGGTCACTATGTATATCTCC	pIR1-LUC (b)-SAT	LPG1 IR	pYES1L-RLuc ½
SMB4633 SMB4634	F: AGGAGATATACATAGTGACCGGATCTCCACCGCTAGCATGAAAAAGCCTGAACTCACCGCGACG R: GGATCTGATTCGGCCCTAGACTAGAACTAGGAACGTTCTATTCCTTTGCCCTCGGACGAGTGC	pIR2-HYG	HYG ORF	pYES1L-RLuc ½
SMB4635 SMB4636	F: CCGACGCCCCAGCACTCGTCCGAGGGCAAAGGAATAGAACGTTCCTAGTTCTAGTCTAGGGCCGAATCAG R: CCGTCCCACGAGGCAACCCCGGTCATGAGCATATCGATGGGAGGGAGGAGAAGCAAGGCTGCGCCTTAAC	pIR1-LUC (b)-SAT	1.7K IR	pYES1L-RLuc ½
SMB4637 SMB4638	F: GCGTGGGTGTTAAGGCGCAGCCTTGCTTCTCCTCCCTCCCATCGATATGCTCATGACCGGGGTTGCCTCGTG R: GCGCTTCCTCGCTCACTGACTTTAATTAACTGCGGCGAGGCTCGAGGTGAGACCACATTACAGCTCCAAAGG	*L. major* FV1 gDNA	LmjF01.0320 ORF	pYES1L-RLuc ½

Table 4-3.



Primer	Sequence	Template	Amplicon	Destination Plasmid/Purpose
SMB4992 SMB4816	F: aaaaaAGAGCTCTCTAG ACCACCATGACCAGCAAGGTGTACGACC R: AGGTTCAGGAGCTCGAACCAGGCGGTCAGG	pRLuc-N3	RLuc 5′ end	Modify RLuc 5′ end with optimized Kozak
SMB4991 SMB4988	F: ACTGACTTTAATTAACTGCGGCGAGGTCTAGACCACC ATGGAAGACGCCAAAAACATAAAGAAAG R: TTGAGCAATTCACGTTCATTATAAATGTCGTTCGCGGGC	pIR1-LUC(b)-SAT	FLuc 5′ end	Modify FLuc 5′ end with optimized Kozak
SMB5519 SMB5520	F: ACTAGTCCACCATGAAGATTTCGGTGATCCCTG AGCAGG TGG R: aaaaaaAGATCTTTAGGCGTCATCCTG TGCTCCCGAG AAC	pIR1-SAT	SAT ORF	Repair pLUC v1 with fusion PCR amplicon
SMB5513 SMB5514	F: GCTAATACGACTCACTATAGGGGGTAGACTCG TGCCGCG CGCTGATGATGTAGGTGC R: CGAAATCTTCATGGTGGACTAGTGGCTG AAAAAGAAGAAA GAGGGGTTGGAAGGCGGTTTAAAGGTGTTTGTTCGAAG	*L. major* FV1 gDNA	α-tubulin IR	Repair pLUC v1 with fusion PCR amplicon
SMB5261 SMB5512	F: CGAACGACATTTATAATGAACGTG AATTGCTC R: CCCTATAGTGAGTCGTATTAGCGGCCGCTTACACGGCG ATCTTTCCGCCCTTC	pIR1-LUC(b)-SAT	FLuc 3′ end	Repair pLUC v1 with fusion PCR amplicon
SMB4687 SMB4688	F: aaaaaaaTCTAGAATGAAGCCTGCTCTGAAGTGCTATGG R: aaaaaaaaTCTAGATGGGCATTCTTTGGGAGAAAAGCAAACGG	*L. major* FV1 gDNA	dSSR	Amplify full-length dSSR (B7138, pLUC-SSR)
SMB4842	F: aaaaaTCTAGACACGCATACACGCAC ACAAATG R: SMB4688	*L. major* FV1 gDNA	R188 fragment	Minimal functional SA mapping; pUC19-RLuc ½
SMB4843	F: aaaaaTCTAGACGTTG AACGTGGTGCCGACG AG R: SMB4688	*L. major* FV1 gDNA	R301 fragment	Minimal functional SA mapping; pUC19-RLuc ½
SMB4844	F: aaaaaTCTAGACTGTCGCGGGTAGAAGCTG AC R: SMB4688	*L. major* FV1 gDNA	R401 fragment	Minimal functional SA mapping; pUC19-RLuc ½
SMB4847	F: aaaaaTCTAGACAAGAATGGCTGC ATGACG AAC R: SMB4688	*L. major* FV1 gDNA	R712 fragment	Minimal functional SA mapping; pUC19-RLuc ½
SMB4850	F: aaaaaTCTAGACTTTGCCTCGCTGTCTCTTCTCC R: SMB4687	*L. major* FV1 gDNA	L84 fragment	Minimal functional SA mapping; pUC19-RLuc ½
SMB4851	F: aaaaaTCTAGAGCAGATCTACTGATGCGACTGC R: SMB4687	*L. major* FV1 gDNA	L182 fragment	Minimal functional SA mapping; pUC19-RLuc ½
SMB4852	F: aaaaaTCTAGAGTTCG TCATGCAGCCATTC TTGC R: SMB4687	*L. major* FV1 gDNA	L281 fragment	Minimal functional SA mapping; pUC19-RLuc ½
SMB4853	F: aaaaaTCTAGACGACAA AGTGTG TCGCACACCT R: SMB4687	*L. major* FV1 gDNA	L375 fragment	Minimal functional SA mapping; pUC19-RLuc ½
SMB5945	F: SMB4687 R: GCAGTCGCATCAGTAGATCTGCGTTGAACG TGGTGCCGACG	*L. major* FV1 gDNA	SSR 5′	Δ489 (B7190)
SMB5946	F: CGTCGGCACCACGTTCAACGCAGATCTACTG ATGCGACTGC R: SMB4688	*L. major* FV1 gDNA	SSR 3′	Δ489 (B7190)
SMB5943	F: SMB4687 R: AATAGTCACTCACAGATCTACTGATGCGACTGCG TCCATGTG	*L. major* FV1 gDNA	SSR 5′	Δ489 + S (B7191)
SMB5941 SMB5942	F: CATCAGTAGATCTGTGAGTGACTATTACATTAC TTTTGGGCACACC R: CACCACGTTCAACGAGAAGAACACTCGGCCCCGAAAG ACTTAGC	*T.brucei* TREU927 gDNA	*T. brucei* stuffer DNA	Δ489 + S (B7191)
SMB5944	F: CCGAGTGTTCTTCTCGTTGAACG TGGTGCCGACGAGCAGAC R: SMB4688	*L. major* FV1 gDNA	SSR 3′	Δ489 + S (B7191)

Table 4-4.



Primer	Sequence	Amplicon	Purpose
SMB4999 SMB4335	F: CGCATCTCACGCTGTGTAGCGAGGAACAGC R: CGATGTACTGGTACTGGTTCTCGGGAGC	*LmjF01.0315* 3' IR through SAT 3'	5' Integration Confirmation
SMB5192	F: GTTCCATCTTCCAGCGGATAGAATGG R: SMB4816	FLuc 5' through RLuc 5'	Middle Integration Confirmation
SMB4348 SMB5000	F: AGCACTCGTCCGAGGGCAAAGGAAT R: CTCGGAGATCTGAAATCTCAGAGTCG	HYG 3' through *LmjF01.0320* 3' IR	3'Integration Confirmation
SMB 1643 SMB5016	SL fwd: AACGCTATATAAGTATCAGTTTCTGTACTTTA FLuc rev: SMB5192 RLuc rev: CACCACGTGCCTCCACAGGTAGC	Across splice leader-mRNA junction	Splice acceptor dinucleotide mapping
SMB6524 SMB6525	F: GAGCACCAGGACAAGATCAA R: TGTTCTCCAGCACCATCTTC	*RLuc* mRNA (142 bp amplicon)	*RLuc* qRT-PCR
SMB6526 SMB6527	F: CGCATGCCAGAGATCCTATT R: AGACGACTCGAAATCCACATAT	*FLuc* mRNA (132 bp amplicon)	*FLuc* qRT-PCR
SMB2992 SMB2993	F: ATCAAGACGAACCCGAACGA R: AGAACTGAGCAAGCAAGTGGC	*DHFR* mRNA (101 bp amplicon)	*DHFR* qRT-PCR
SMB6611 SMB6612	F: CACACGAATGCACCTGAAC R: CTGTCTCTTCTCCGCTCTTAC	*LmjF01.0135*-dSSR junction	FAIRE-qPCR
SMB6613 SMB6614	F: ACTGCGTCAGTTCCGTTT R: TGCGACCTGCCAAATAGAG	*LmjF01.0320*-dSSR junction	FAIRE-qPCR
SMB6615 SMB6616	F: GTTCCATCTTCCAGCGGATA R: GGCGATTCCATAGCACTTCA	*FLuc*-dSSR junction	FAIRE-qPCR
SMB6617 SMB6618	F: ACTGCGTCAGTTCCGTTTG R: CGGTGATCATCCTCTTCCTCT	*RLuc*-dSSR junction	FAIRE-qPCR
SMB6653 SMB6654	F: CCACCTGATAGCCTTTGTACTT R: CATAGCTTACTGGGACGAAGAC	FLuc 3'	FAIRE-qPCR
SMB6655 SMB6656	F: GAGAAACTACAACGCCTACCT R: CCCTTCACCTTCACGAACT	RLuc 3'	FAIRE-qPCR
SMB6651 SMB6652	F: CCGACCAAGGCTTTGAACTA R: AGCAAGCAGAGTCTTCATCAG	SAT ORF	FAIRE-qPCR
SMB6657 SMB6658	F: CGACGTCTGTCGAGAAGTTT R: TATCCACGCCCTCCTACAT	HYG ORF	FAIRE-qPCR

Chapter Five

Mutation of poly(dG:dC) tracts in the dSSR core suggests that transcription directionality can be biased in an irregular manner suggestive of epigenetic control

Preface

BA designed and performed all experiments, analyzed data, and wrote the first draft of this chapter. SMB supervised these studies and provided comments that were incorporated into the final version presented here. This chapter is not intended for publication.

Abstract

In Chapter 4 I described an integrated, bidirectional dual-luciferase reporter system for the identification of determinants of transcription controlled by divergent strand switch regions (dSSRs). Tests of a single core dSSR deletion and a single core dSSR substitution in the dSSR of chromosome 1 showed little alteration in the expression of reporter genes in this system, demonstrating that *cis*-regulatory elements within the dSSR core are not required to drive expression from these regions. In parallel to these experiments, we performed tests focused on the poly(dG:dC) tracts within this core dSSR, as these motifs were hypothesized to be involved in defining the directionality of transcription in related trypanosomatid protozoa. In most constructs tested there was little effect on bidirectional reporter expression, consistent with the full deletion studies. However, for all but one mutant construct tested, variants were observed in which the RLuc reporter was elevated, usually accompanied by a reduction in FLuc patterns that are suggestive of a switch from bidirectional to unidirectional initiation. These changes occurred without mutations in the dSSR reporter of these lines, and thus we favor a model invoking some kind of epigenetic alteration. The low FLuc, high RLuc lines showed sensitivity to the FLuc-linked SAT marker, and by selection we were able to induce SAT resistance and concomitant FLuc expression, again in the absence of dSSR reporter mutations or copy number variations. Preliminary assessment of the epigenetic state of the dSSR in original FLuc-defective and switcher, FLuc-expressing lines by FAIRE demonstrated some differences in the dSSR-proximal regions of the reporter allele, although additional experiments are needed to define the nature of these differences. Thus, there appears to be some influence by elements within the core dSSR on the epigenetic stability of bidirectional transcription, which could play a role in PGC-internal regions of transcription initiation or in dSSRs of other chromosomes.

Introduction

In the previous chapter, I described the development of an integrated, bidirectional dual-luciferase reporter system that allows the detailed interrogation of genetic determinants of transcription initiation events within divergent SSRs (dSSRs). These studies yielded convincing evidence that *Leishmania* do not require *cis*-regulatory elements within the core dSSR to effectively drive bidirectional transcription. Prior to discovering that the core dSSR was not required, I generated a number of more focused deletion and substitution experiments designed to characterize the potential role of poly(dG:dC) tracts in *Leishmania* dSSRs, as these had been hypothesized to play important roles in defining the directionality of transcription in trypanosomes (1). Because the general rationale for these studies was reviewed in detail in the previous chapter, I will instead remind the reader of several key points in the remainder of this introduction, focusing on the potential role for poly(dG:dC) tracts in dSSR function.

Despite a number of attempts to identify putative *cis*-regulatory motifs among dSSRs using comparative genomics, a motif that was universally conserved among all dSSRs and that is unique to these regions was not identified in *Leishmania* or in trypanosomes (1,2, Anderson and Beverley, unpublished data). Aside from the consensus elements involved in *trans*-splicing of the dSSR-proximal genes, which include a polypyrimidine tract and downstream AG dinucleotide (termed here the splice acceptor element), the sole motif identified in any trypanosomatid protozoa was that of a poly(dG:dC) tract greater than 9 bp, which was identified using metagene analysis of the dSSR-associated histone mark acetylated H4K10 in *T. brucei* (1). These poly(dG:dC) tracts present unique structural features which make them interesting motifs to consider in an organism lacking canonical eukaryotic promoter motifs: like poly(dA:dT) tracts, poly(dG:dC) homopolymers function as promoter elements in yeast independently of a *trans*-acting factor by virtue of their inflexibility, which inhibits nucleosome formation and creates an open chromatin environment (3). The effects of these sequences on chromatin structure fit well with observations that T7 promoter-induced open chromatin can facilitate RNAP II transcription events in silent loci in *T. brucei* (4), and suggest a possible mechanism by which an accessible chromatin state is conferred. Despite this attractive hypothesis, this motif is not present in all dSSRs, demonstrating that these cannot be absolutely required for the *de facto* promoter activity present in these regions. Moreover, nucleosome positioning studies in *Leishmania* demonstrated that these sequences are not nuclease-hypersensitive, suggesting that if they do play a role in transcription from these loci, it is likely not dependent on its ability to inhibit nucleosome formation (Anderson, Shaik, and Beverley, in preparation). Interestingly, Siegel and colleagues speculated that these poly(dG:dC) tracts might instead confer directionality to transcription events, as these sequences were primarily identified in the sense orientation relative to the polycistronic gene cluster (PGC) (1).

While metagene analysis of transcription-associated histone marks in *Leishmania* failed to show a similar overrepresentation of poly(dG:dC) tracts in dSSRs (5), transcription start site (TSS) mapping studies using the dSSR of chromosome 1 in *Leishmania* demonstrated that the innermost transcription start sites were only 73 bp apart, with two poly(dG:dC) tracts comprising a significant fraction of this region (6). The development of the integrated, bidirectional dual-luciferase reporter system described in Chapter 4 made it relatively easy to characterize the role of any sequence in the function of a dSSR, and we expected that mutation of the poly(dG:dC) tract would affirm our observations described in the previous chapter suggesting a lack of a requirement for any core dSSR element. However, the possibility that these tracts affected the directionality of transcription was an intriguing one; thus, we sought to characterize the role of poly(dG:dC) tracts in the context of *Leishmania* dSSR function. We generated a panel of mutants in the bidirectional reporter: a 73 bp deletion, encompassing the region between the innermost transcription start sites; substitutions of “scrambled” G/C tracts which disrupt the poly(dG:dC) homopolymer without altering G/C content; and poly(dA:dT) tract substitutions, which can be functionally complemented by poly(dG:dC) tracts in yeast promoters (3). For the most part, these experiments were consistent with the substitution experiments described in Chapter 4, in that bidirectional expression was maintained. However, in a subset (9 of 33 transfectants, for 6 of 7 constructs), asymmetric effects were seen. The implications of these data for models of gene expression and epigenetic control of transcription are discussed.

Results

Poly(dG:dC) tract mutants can maintain bidirectional reporter expression

While the general topology of the chromosome 1 dSSR was discussed in Chapter 4 with respect to the identification of the minimal functional splice acceptors, a more detailed discussion of landmarks within this locus is useful in the context of the experiments discussed in this chapter. We demarcate a number of relevant features in Figure 5-1A: the innermost transcription start sites and poly(dG:dC) tracts are annotated here, along with their location relative to the reporter genes in the bidirectional reporter. To characterize the role of the poly(dG:dC) tracts within the context of the chromosome 1 dSSR, we generated a panel of bidirectional reporters containing deletions and substitutions that remove, disrupt, or alter the nature of one or both of these homopolymer tracts (see Figure 5-1). In these studies, we utilized selection with only one antibiotic (SAT or HYG) to allow for the possibility that directionality of transcription may be altered; the expected outcomes for bidirectional and unidirectional transcription in the context of different combinations of selective pressure during transfection and maintenance in culture are depicted in Figure 5-2A. For the WT dSSR reporter, selection with SAT or HYG singly or simultaneously yield identical results (Chapter 4).

We began with a 73-bp deletion that encompasses the region between the innermost transcription start sites (Figure 5-1B), including both poly(dG:dC) tracts. Interestingly, we were only able to obtain colonies on HYG plates in 4 independent experiments, despite being able to easily select colonies with either antibiotic when this construct was transfected as an episome (Table 5-1). As described in Chapter 4, episomal transfections score only the functionality of the splice acceptor elements and the expression of the marker and reporters in a "context" and "promoter" independent fashion. The planned integration of the bidirectional reporter in these lines was confirmed by PCR, as described in Chapter 4 (data not shown). When we quantified reporter gene levels in lines containing an integrated Δ73 reporter, we observed that 3 of 4 clones screened had very low levels of FLuc activity, at roughly 2-fold over the WT background (Figure 5-3A). While quantitative analysis of *RLuc* mRNA levels has not yet been performed on all clones, we observed that the two clones tested that showed low FLuc activity also showed *RLuc* mRNA levels approximately 10-fold higher than lines bearing the WT dSSR (Figure 5-3B).

We performed similar experiments using constructs in which the homopolymeric nature of one or both poly(dG:dC) tracts was disrupted (scrambled) without altering the G/C content of the sequence (Figures 5-1C-E). Here, we were able to obtain integrated colonies using either antibiotic (Table 5-1), indicating that these sequences were able to drive transcription in either direction. In all lines containing one scrambled poly(dG:dC) tract and one intact poly(dG:dC) homopolymer, we saw no defect in reporter gene expression, regardless of the antibiotic used for selection [Figure 5-4; selection is indicated in red (HYG) and blue (SAT) circles below the bar].

Interestingly, three of the 12 clones screened showed 10-fold higher *RLuc* mRNA levels without showing an alteration in FLuc activity, suggesting that these phenotypes may not always be linked. In contrast, 3 of the 7 clones containing 2 scrambled poly(dG:dC) tracts demonstrated similar phenotypes to those observed previously—i.e. little FLuc activity, and higher *RLuc* mRNA levels, but only when selected for HYG resistance. In light of these observations, we wondered whether the homopolymer-derived structural properties might result in the generation of lines with only normal reporter gene expression. We readily generated lines containing one poly(dA:dT) and one poly(dG:dC) tract (Figure 5-1F) and a second one in which both poly(dG:dC) tracts were substituted with poly(dA:dT) tracts of an equivalent length (Figure 5-1G). Although the majority of lines containing one or two poly(dA:dT) substitutions showed the expected bidirectional reporter pattern, we observed one HYG-selected clone bearing two poly(dA:dT) substitutions that showed low FLuc activity and high *RLuc* mRNA levels (Figure 5-5), indicating that the presence of a generic homopolymer tract is not required for maintenance of a bidirectional promoter by itself, nor is it sufficient to "block" the definition of a functionally unidirectional reporter. Together, these data support the idea that poly(dG:dC) tracts may play a role in defining transcription directionality; however, it is clear that these regions are not absolutely required for bidirectional transcription, as transcriptionally-normal clones were generated with the same construct, often even in the same experiment. In comparison to some yeast promoters, in which poly(dA:dT) and poly(dG:dC) tracts are functionally interchangeable (3), we find that the function of these sequences differs somewhat in *Leishmania*.

Unidirectional poly(dG:dC) tract mutants can be converted to bidirectionally transcribing lines without genetic alterations

These experiments presented an interesting puzzle: although manipulation of the poly(dG:dC) tracts clearly altered transcriptional patterns, the phenotype was not completely penetrant, and sequence analysis did not reveal any mutations in FLuc-defective lines (data not shown). As a result, we wondered whether an epigenetic phenomenon might be responsible. We hypothesized that if it were the result of an epigenetic change, we might find that this phenotype was reversible with the application of selection toward the "off" half of the reporter allele—in this case, SAT (Figure 5-2B). We took two clones bearing the Δ73 deletion that showed very low FLuc activity in previous assays and grew them in medium containing only SAT. We found that although these clones were markedly delayed relative to a line bearing the full-length WT dSSR in their first passage in SAT, these clones eventually reached stationary-phase density (Figure 5-6A). While a delayed growth phenotype was observed for several passages in SAT (data not shown), after several passages (~50 cell doublings), a "switcher" population emerged that now grew normally under SAT selection (Figure 5-6A). Interestingly, these switcher populations maintained HYG resistance and could grow normally in medium bearing both antibiotics (Figure 5-6A). Moreover, the upstream FLuc and RLuc reporters for each marker now showed WT levels of bidirectional expression (Figure 5-6B). As before, no mutations were found in the dSSR reporter in these switcher populations (data not shown).

We favor the idea that these "switcher" lines re-acquired a permissive epigenetic state. However, other genetic alterations may also explain these phenotypic changes. Selective pressure has been demonstrated to alter chromosome copy number and may also produce extrachromosomal DNA elements [reviewed in (7)]. To test the possibility that a genetic alteration may have occurred in these "switchers", we first confirmed that the proper integration of this construct in the switcher lines was maintained by PCR (Figure 5-6C), which demonstrated that the reporter allele has not undergone genomic rearrangements. Furthermore, we subjected one clone and its parent line to copy number variation analysis by quantitative PCR (Figure 5-6D). Using primers detecting loci far from the dSSR on chromosome 1, primers on the disomic chromosome 5, as well as amplicons detecting the *RLuc* and *FLuc* genes, we find that the copy number of these genes is not increased in the switchers relative to the parents (Figure 5-6D).

FAIRE demonstrates different epigenetic states in unidirectional and bidirectional lines

Because these switcher lines appear genetically identical, we sought to characterize the dSSR-proximal epigenetic environment in these lines relative to a line bearing the WT dSSR within the bidirectional reporter allele. As we discussed in Chapter 4, formaldehyde-assisted isolation of regulatory elements (FAIRE) coupled to quantitative PCR integrates relevant epigenetic changes regardless of the specific marks involved and thus serves as a useful marker of the epigenetic state of this locus. We compared FAIRE-enriched DNA relative to input DNA for one Δ73 parent line and its switcher using the same allele-specific primers described in Chapter 4 (Figure 4-4A). Interestingly, we observe that while the WT dSSR shows highest FAIRE enrichment near the dSSR (data was previously discussed in Chapter 4), the FAIRE enrichment in the Δ73, FLuc-defective line is highest near the 5' end of the reporter construct, in the middle of the SAT ORF (Figure 5-7). Although the FAIRE enrichment at the 5' end of the reporter is returns to similar levels as the WT reporter in the switcher, the enrichment near the dSSR does not return to WT dSSR levels. In the absence of biological replicates of these experiments, it is not possible to quantify the biological variability of these lines to determine whether a true epigenetic difference is occurring. However, it would be very interesting to understand which, if any epigenetic marks differ between the parent lines and their switchers, as well as between these genetically identical lines exhibiting bidirectional and unidirectional phenotypes with respect to reporter gene expression.

Discussion

In the studies described in Chapter 4, we concluded that *cis*-regulatory elements within the core dSSR are not required to drive bidirectional transcription. Although poly(dG:dC) tracts, which are the sole motif identified in loci associated with transcription of protein-coding genes (1), fit well in models describing the requirement for an open chromatin environment in transcription initiation (4), these loci are clearly not required for this phenomenon, as large-scale substitutions lacking poly(dG:dC) tracts were capable of bidirectional transcription (Anderson and Beverley, in preparation; see Chapter 4). Siegel and colleagues instead propose a role for these sequences in defining the directionality of transcription (1); in light of this hypothesis and the relative ease in which we can examine the role of various genetic elements in dSSR function using an integrated, bidirectional dual-luciferase reporter, we sought to quantify the contribution of these sequences to dSSR function. We generated a panel of constructs bearing deletions and substitutions that alter the integrity of one or both poly(dG:dC) tracts within the dSSR of chromosome 1 in *L. major*, assaying reporter gene activity in a number of clones. As expected based on results described in Chapter 4, we observe that 24 of 33 clones screened using 7 different constructs showed the predicted pattern of bidirectional reporter gene expression. Unexpectedly, 6 of 7 constructs tested generated lines demonstrating unidirectional reporter activity that were not attributable to mutations in the dSSR reporter. This incompletely penetrant phenotype in genotypically identical clones suggests that although these sequences may play a role in defining the directionality of transcription, they do not play an essential role in dSSR function. However, the fact that any phenotype at all was observed was surprising, since large portions of the dSSR could be substituted without seeing similar variation among clones. Characterization of lines bearing the poly(dG:dC) tracts and the scrambled mutants in different orientations will clarify the role of these sequences in the context of dSSR biology and may provide interesting insight into the genetic factors which affect gene expression.

Because these phenotypes did not appear to arise from DNA sequence changes, we hypothesized that parasite lines demonstrating a unidirectional phenotype may be epigenetically different than those demonstrating bidirectional reporter gene expression. The integrated bidirectional reporter construct contains selectable antibiotic resistance markers on both sides of the reporter, allowing the selection of parasites expressing one side or the other, or both sides simultaneously. We used this capacity to revert the unidirectionally-transcribing clones to bidirectionally-expressing lines using antibiotic selection for the “off” side of the reporter; after a brief period of selection (~50 cell doublings), a SAT-resistant population emerged that retained HYG resistance and RLuc expression, while regaining FLuc expression levels comparable to the WT dSSR reporter (Fig. 5-6B). Again there was no evidence of mutation, genomic rearrangement, or changes in copy number, implying that the phenotypic change is of epigenetic origin. Preliminary data assessing the epigenetic state of parent and switcher lines relative to the WT dSSR show some alterations in FAIRE signal among these lines; however, further assessment of the biological variation between phenotypically-similar isolates will be needed to identify the nature of these epigenetic differences. Attempts to identify small molecule inhibitors which are capable of reversing this phenotype were not successful (data not shown); however, extended passage of these lines in these compounds may be required, as the reversion using SAT required many cell doublings.

Together, we believe that these data support a model in which *cis*-regulatory elements within the core dSSR are not explicitly required to drive bidirectional transcription. However, genetic elements such as poly(dG:dC) tracts may potentially contribute to the diversity of transcriptional patterns originating from these loci. It is important to consider that a small number of PGC-internal regions of transcription initiation, which are marked by the same transcription factor binding patterns and epigenetic marks as dSSRs, have been shown to function in unidirectional transcription of polycistronic gene clusters (5,8,9), demonstrating the potential relevance of unidirectional promoters in *Leishmania* and other trypanosomatids. These regions are often omitted from discussions regarding the factors controlling transcription initiation in trypanosomatid protozoa, as these regions are even more poorly understood than dSSRs. However, it is clear from these experiments that the nature of the interaction between genetic elements and the epigenetic networks that define dSSRs is complex and quite flexible, and we believe that the unidirectional clones generated in these experiments may facilitate the characterization of these interactions. Furthermore, characterization of the epigenetic state of these lines might aid in the identification of factors that distinguish bidirectional and unidirectional regions of transcription initiation in *Leishmania*, such as the differential positioning of histone modifications, histone variant incorporation, and DNA modification.

Materials and Methods

Plasmid generation

The development and characterization of pLUC v2 was described in Chapter 4. Variations in the dSSR of chromosome 1 were introduced using fusion PCR (10), using the primers described in Table 5-2. These amplicons were cloned into pGEM-T (Promega) or pCR-

Blunt (Life Technologies) to generate intermediate constructs which were sequenced completely to verify the integrity of the dSSR. The dSSR mutants were released by XbaI digestion and were cloned into XbaI-digested pLUC v2 (B7129). The orientation was validated by restriction digestion and sequencing, and the following plasmids relevant to this work were generated: pLUC-Δ73 (B7192); pLUC-SSR(ScrambleG9) (B7236); pLUC-SSR(ScrambleG11) (B7237), pLUC-SSR(Scramble PolyG) (B7238); pLUC-SSR(G9-A9); and pLUC-SSR(G9/11-A9/11).

Cell culture and generation of lines bearing bidirectional reporter construct

Cell culture and transfections were performed as described in Chapter 2. The targeting fragments from all pLUC v2 derivatives were released by digestion with BamHI and BlpI. Selection on semisolid medium was performed using 12-50 µg/mL hygromycin B (Calbiochem) and/or 100-110 µg/mL nourseothricin (Werner BioAgents). "Switcher" experiments were accomplished by growing cells in 50 µg/mL nourseothricin. Integration of the targeting fragment into the dSSR was confirmed using the primers described in Table 4-5.

Reporter gene analysis

Luciferases assays and quantitative reverse-transcriptase PCR (qRT-PCR) were performed as described in Chapter 4. Splice acceptor dinucleotide mapping was performed on all lines to confirm the utilization of the major splice acceptor dinucleotide (data not shown).

Copy number variation and FAIRE analysis

Copy number variation was assessed by quantitative PCR using protocols described in Chapter 4. Primer sets for the *RLuc*, *FLuc,* and *LmjF01.0400* genes are described in Chapter 4;

see Table 5-2 for remaining primers. FAIRE analysis was performed as described in Chapter 4, using the same primer sets.

Acknowledgments

We thank Drs. Jenny Duncan and David Rudnick at Washington University for the use of their Bioruptor for FAIRE experiments. We also thank Dr. David Piwnica-Worms for the gift of the pRLuc-N3 plasmid. This work was funded by the National Institutes of Health (R0129646 to SMB and BAA; T32 GM007067 to BAA) and the Washington University Berg-Morse and Schlesinger Graduate Fellowships to BAA.

Figure Legends

Figure 5-1. Landmarks of interest in chromosome 1 dSSR and description of dSSR mutants used in this chapter. (A) Landscape of the chromosome 1 dSSR, indicating sequences relevant to this work. Wide arrows indicate the dSSR-flanking ORFs; in the bidirectional reporter, these are FLuc (blue) and RLuc (red). The minimal functional splice acceptors defined in Chapter 4 are depicted as black boxes, located at the 5’ ends of genes. The core dSSR, defined as the region between the minimal functional splice acceptors, is depicted as a green box in both parts of the figure and is denoted at the top. The region containing the poly(dG:dC) tract is shown as a wider green box containing $(G)_n$ in the lower part of the figure. Transcription start sites within the dSSR identified in (5) are depicted as black, triangular flags, and the strandedness is indicated by the placement above (positive strand) or below the green bar (negative strand); not all transcription start sites are shown, and these events are not drawn to scale. The extension of the green box shown below indicates sequences of interest present within this region; the locations

of the poly(dG:dC) tracts described in the "scrambled" studies are shown here. (B) Schematic representation of the Δ73 deletion within the bidirectional reporter. Genes, splice acceptors, transcription start sites, and core dSSR are indicated as in (A). The location of the deletion is shown with dotted black lines.

Figure 5-2. Expected behavior of parasite lines containing mutations or deletions in the dSSR reporter. (A) Expected behavior of reporter genes and antibiotic resistance markers under selective pressure when the dSSR contains a unidirectional or bidirectional promoter. The panel above indicates the general orientation of the bidirectional reporter construct, and individual genes are represented by blue or red arrows. The orientation of the genes is depicted by the direction and color of the arrow. The levels of reporter gene expression depicted here are not intended to be quantitative, but to depict the general trends in expression for these classes of transcription.

Figure 5-3. Reporter gene activity differs in some but not all Δ73 clones. FLuc activity (blue bars) is depicted in photons per second per 2 x 10^6 cells; *RLuc* mRNA levels are quantified relative to the *DHFR* mRNA and are normalized to full-length dSSR clone 1.1, as performed in experiments in Chapter 4. Error bars indicate the average of three biological replicates (WT FV1) or independent clones (Full-length average). The circles at the bottom of the bars indicate that the HYG (red) or SAT (blue) antibiotics were used for selection. The dotted line within each graph indicates the level of FLuc activity or *RLuc* mRNA of lines bearing the WT dSSR. "ND" indicates lines which have not been assessed by quantitative RT-PCR at this time.

Figure 5-4. Reporter gene activity is also different in lines bearing poly(dG:dC) mutations. Graphs are labeled identically to those in Figure 5-3. The dSSR constructs present within the bidirectional reporter are labeled at the bottom and match descriptions in Figure 5-1; the order of clones is maintained in these graphs, allowing comparisons between the two reporters.

Figure 5-5. Substitution of poly(dA:dT) tracts does not protect against the unidirectional promoter phenotype. Graphs are labeled identically to those in Figure 5-3. The dSSR constructs present within the bidirectional reporter are labeled at the bottom and match descriptions in Figure 5-1.

Figure 5-6. Selective pressure can reverse the unidirectional phenotype without altering the genetic identity of the dSSR. (A) Growth curves for one representative Δ73 clones in medium containing HYG (blue) and SAT (first passage, red; SAT-adapted, green). Y-axis represents cell density (cells/mL), and X-axis indicates the time in hours. SAT-adapted line was grown in SAT for 40-50 cell doublings and a growth curve was repeated. (B) Firefly luciferase activity and Renilla luciferase mRNA levels from Delta73 parent and SAT-adapted switcher population. Axes and graph labels are the same as Fig. 5-3.

Figure 5-7. Δ73 switchers have no genomic rearrangements or copy number variations affecting the dSSR reporter relative to parent lines. (A) PCR re-validation of SAT-adapted Δ73 clones. The agarose gel analysis of PCR amplicons were described previously in Chapter 4, and the locations of the primer pairs relative to the bidirectional reporter are indicated with black arrows at the top. The expected band size is indicated above each gel, and relevant dsDNA standard

bands are indicated at the sides. (C) Copy number variation of loci on chromosome 1 and within the bidirectional reporter. Quantitative PCR was used to assess the relative copy number of two loci within chromosome 1 relative to a known disomic chromosome (chr. 5). Chromosome copy number or gene copy number was quantified using the $2^{-\Delta\Delta Ct}$ method. WT FV1 is represented with blue bars; the Δ73 clone 3 parent is represented in red, and the "switcher" of this clone is depicted in green.

Figure 5-8. FAIRE-qPCR analysis of "switchers" demonstrates differences in the epigenetic state of the reporter-associated dSSR. (A) Location of dSSR-proximal and "far right" amplicons relative to known FAIRE peaks are depicted with red boxes. FAIRE and acetylated histone H3 data are reprised from Chapter 3, Figure 3-4. Relative fragment depth values for FAIRE were calculated using input DNA as the normalization control; acetylated H3 data was obtained from the Gene Expression Omnibus series GSE 13415, sample GSM338433 and was converted to WIG format for display in IGV. Normalized fragment depth plots of FAIRE are plotted using input DNA as the normalization control. X-axis indicated physical location on the chromosome. Y-axis indicates normalized fragment depth per base pair (FAIRE) or acetyl-H3 to H3 ratios. Red and blue arrows below panel indicate the location and length of polycistronic gene clusters in chromosome 1. Figure is also described in Chapter 4. (B) Representation of the location of qPCR amplicons used in FAIRE analysis. Values in parentheses represent the distance from the dSSR. (C) FAIRE-qPCR data, represented as the enrichment in FAIRE DNA over an equivalent amount of input DNA. Values are normalized to the dSSR-distal gene *LmjF01.0400* (Far right). The directions of polycistronic gene clusters are indicated with blue and red arrows below the graph. Gray bars indicate FAIRE-qPCR data using the WT dSSR and was described in Chapter

4. Dark red bars indicate the FAIRE enrichment of FLuc-defective Δ73 clone described in Figs. 5-6 and 5-7. Green bars indicate the FAIRE enrichment of the SAT-adapted "switcher" population isolated from the same line.

Table 5-1. Quantitation of episomal and integrated colonies from pLUC poly(dG:dC) tract mutants. The numbers of colonies per 10 μg DNA are listed in the table, divided between lines selected on HYG only and lines selected on SAT only.

Table 5-2. Primer sequences used in this study. The primers and their mates are listed in this table, along with the amplicon generated and its purpose. The primer pairs at the top use primer mates described in Chapter 4, which allow them to be cloned into the pLUC v2 vector. The remaining primers were used in copy number variant analysis, described in the Materials and Methods section.

References

1. Siegel T, Hekstra D, Kemp L, Figueiredo L, Lowell J, Fenyo D, et al. Four histone variants mark the boundaries of polycistronic transcription units in *Trypanosoma brucei*. Genes Dev 2009;23(9):1063–76.

2. Ivens A, Peacock C, Worthey E, Murphy L, Aggarwal G, Berriman M, et al. The genome of the kinetoplastid parasite, *Leishmania major*. Science 2005;309(5733):436–42.

3. Iyer V, Struhl K. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. EMBO J 1995;14(11):2570–9.

4. McAndrew M, Graham S, Hartmann C, Clayton C. Testing promoter activity in the trypanosome genome: isolation of a metacyclic-type VSG promoter, and unexpected insights into RNA polymerase II transcription. Exp Parasitol 1998;90(1):65–76.

5. Thomas S, Green A, Sturm N, Campbell D, Myler P. Histone acetylations mark origins of polycistronic transcription in *Leishmania major*. BMC Genomics 2009;10:152.

6. Martínez-Calvillo S, Yan S, Nguyen D, Fox M, Stuart K, Myler P. Transcription of *Leishmania major* Friedlin chromosome 1 initiates in both directions within a single region. Mol Cell 2003;11(5):1291–9.

7. Beverley S. Gene amplification in *Leishmania*. Ann Rev Microbiol 1991;45:417–44.

8. Kolev N, Franklin J, Carmi S, Shi H, Michaeli S, Tschudi C. The transcriptome of the human pathogen *Trypanosoma brucei* at single-nucleotide resolution. PLoS Pathog 2010 Jan;6(9):e1001090.

9. Siegel TN, Hekstra DR, Wang X, Dewell S, Cross G. Genome-wide analysis of mRNA abundance in two life-cycle stages of *Trypanosoma brucei* and identification of splicing and polyadenylation sites. Nucleic Acids Res 2010;38(15):4946–57.

10. Kuwayama H, Obara S, Morio T, Katoh M, Urushihara H, Tanaka Y. PCR-mediated generation of a gene disruption construct without the use of DNA ligase and plasmid vectors. Nucleic Acids Res 2002;30(2):e2.

Figure 5-1. Landmarks of interest in the chromosome 1 dSSR and description of dSSR mutants used in this chapter.

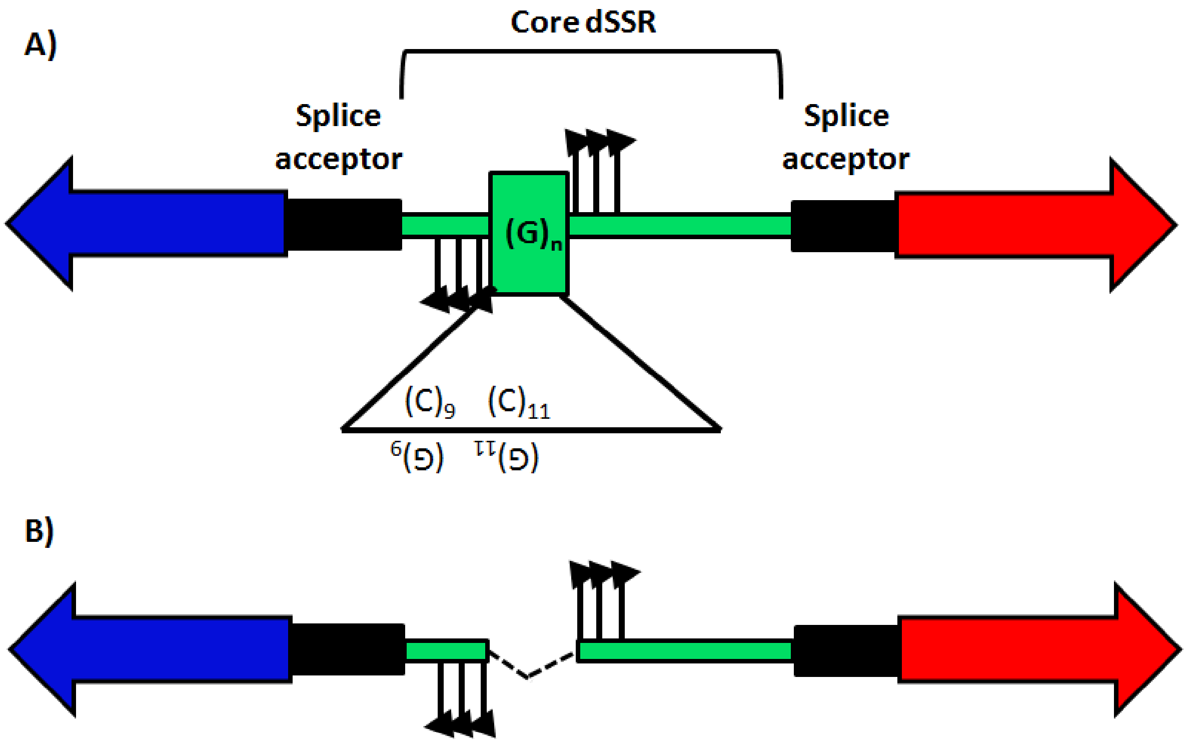
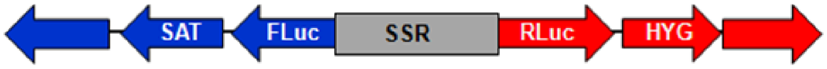


Figure 5-2. Expected behavior of parasite lines containing mutations or deletions in the dSSR reporter.



	SAT only	SAT + HYG	HYG only
Unidirectional	FLuc On, SAT^{R} **RLuc Off, HYG^{S}**	FLuc On, SAT^{R} RLuc On, HYG^{R}	**FLuc Off, SAT^{S}** RLuc On, HYG^{R}
Bidirectional	FLuc On, SAT^{R} RLuc On, HYG^{R}	FLuc On, SAT^{R} RLuc On, HYG^{R}	FLuc On, SAT^{R} RLuc On, HYG^{R}

Figure 5-3. Reporter gene activity differs in some but not all Δ73 clones.

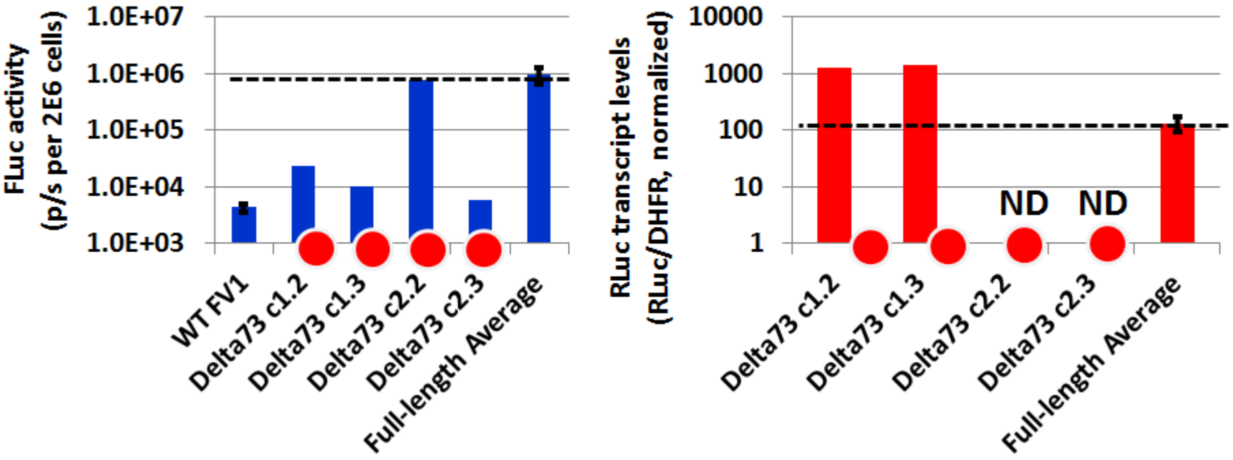


Figure 5-4. Reporter gene activity is also different in lines bearing poly(dG:dC) mutations.

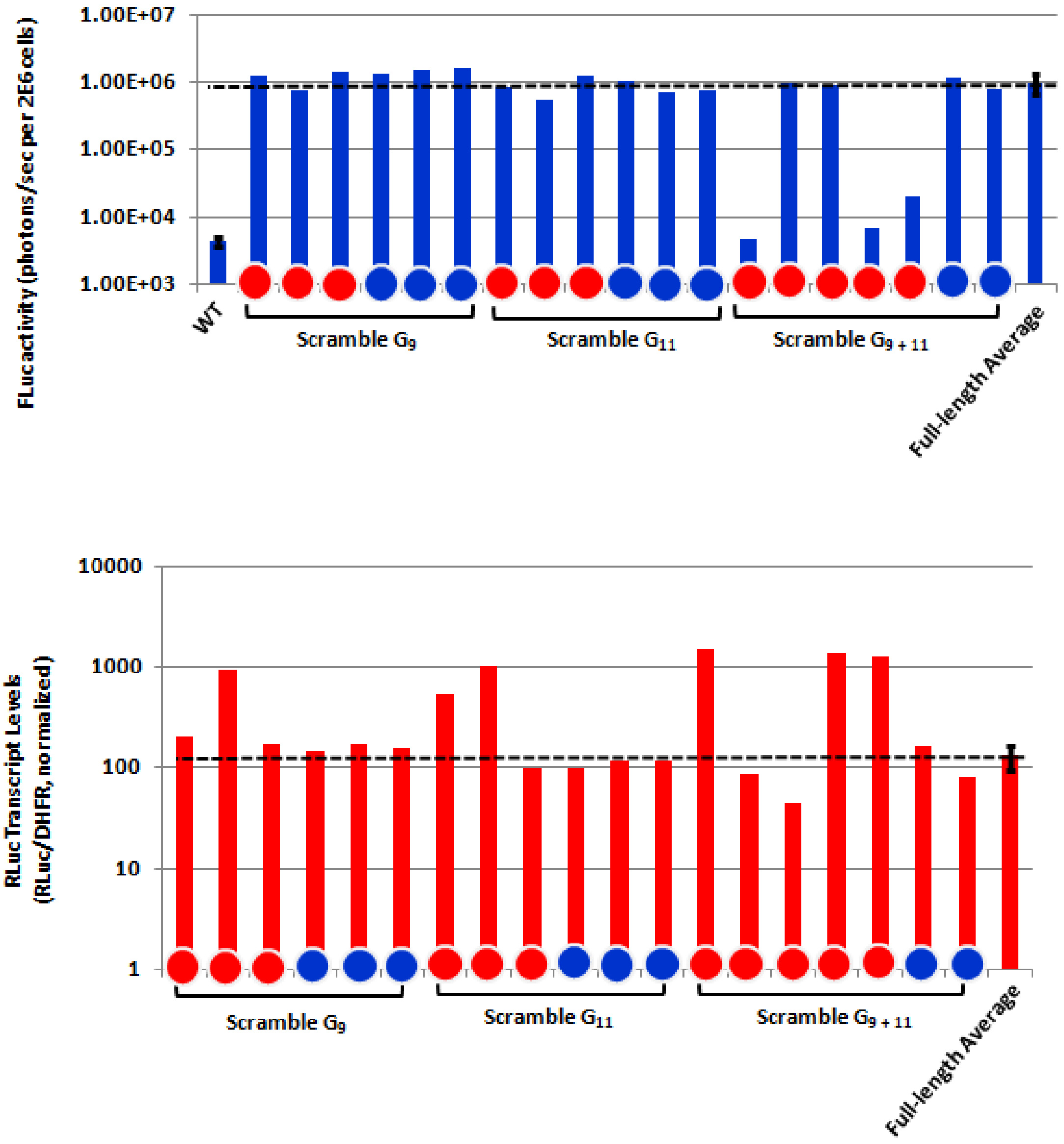


Figure 5-5. Substitution of poly(dA:dT) tracts does not protect against the unidirectional promoter phenotype.

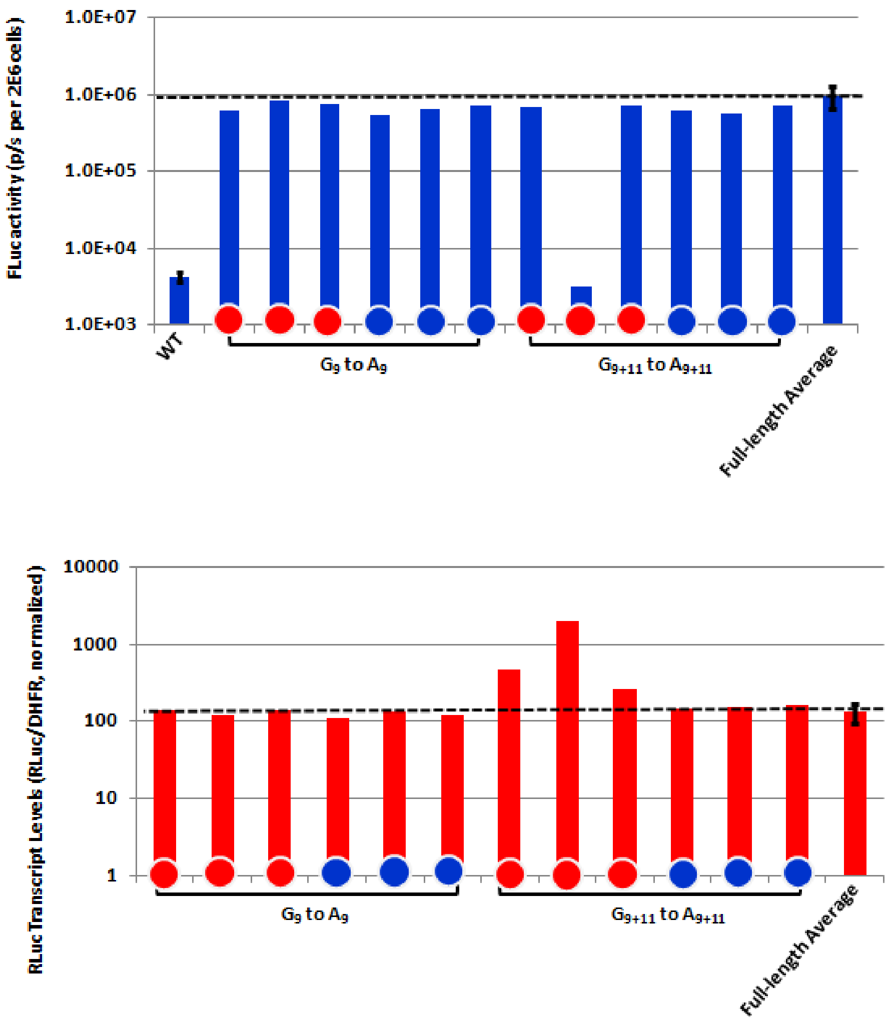


Figure 5-6. Selective pressure can reverse the unidirectional phenotype without altering the genetic identity of the dSSR.

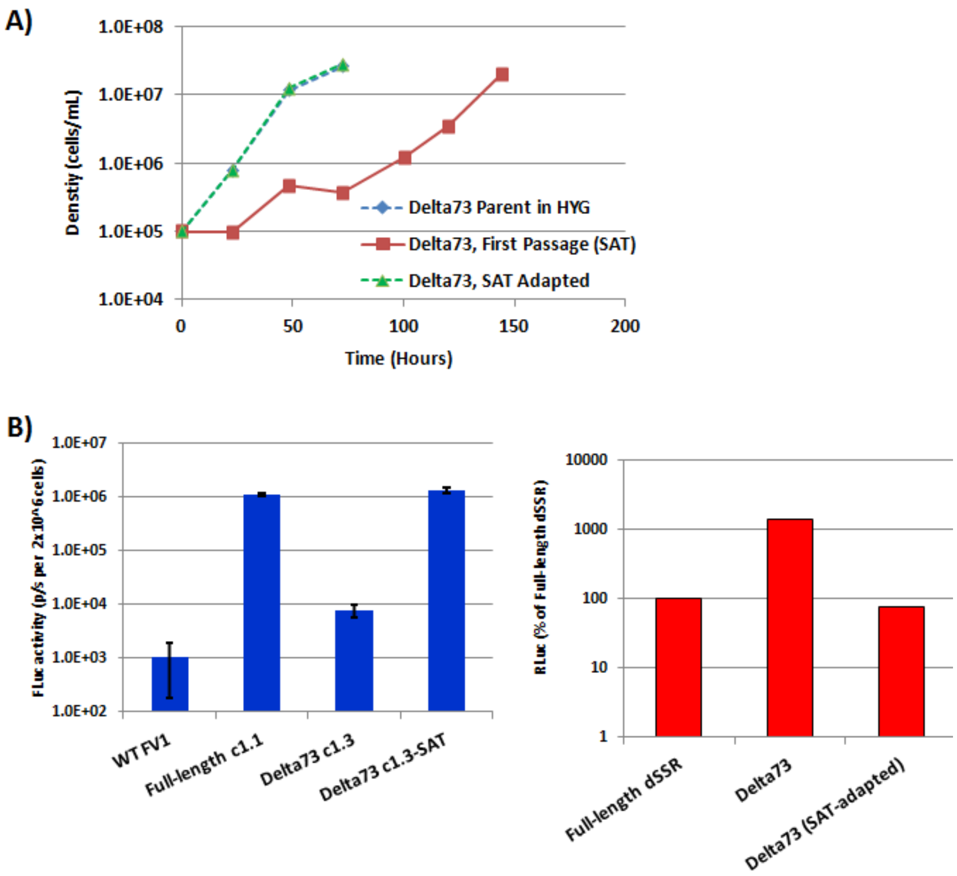


Figure 5-7. Δ73 "switchers" have no genomic rearrangements or copy number variations affecting the dSSR reporter relative to parent lines.

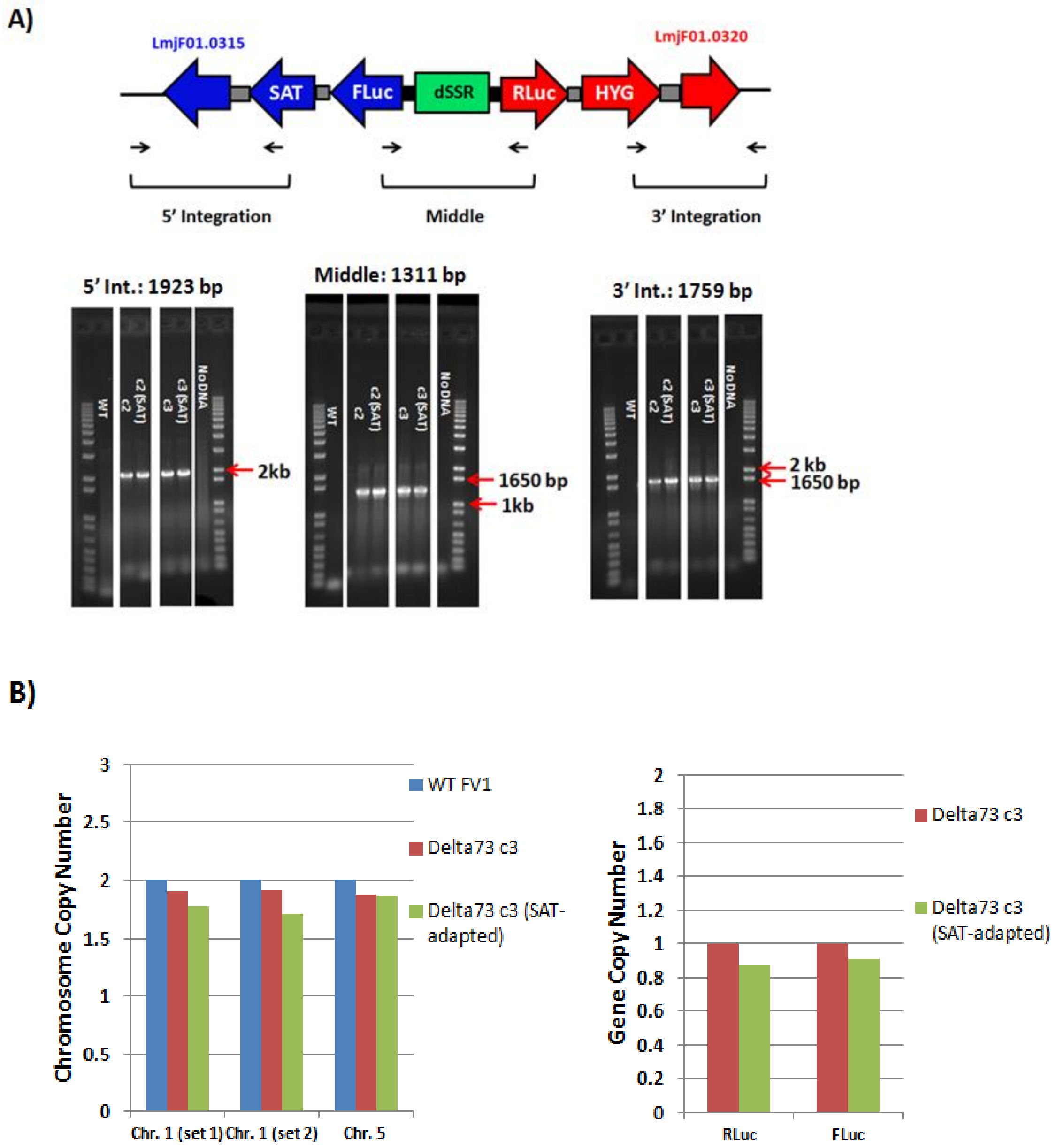


Figure 5-8. FAIRE-qPCR analysis of "switchers" demonstrates differences in the epigenetic state of the reporter-associated dSSR.

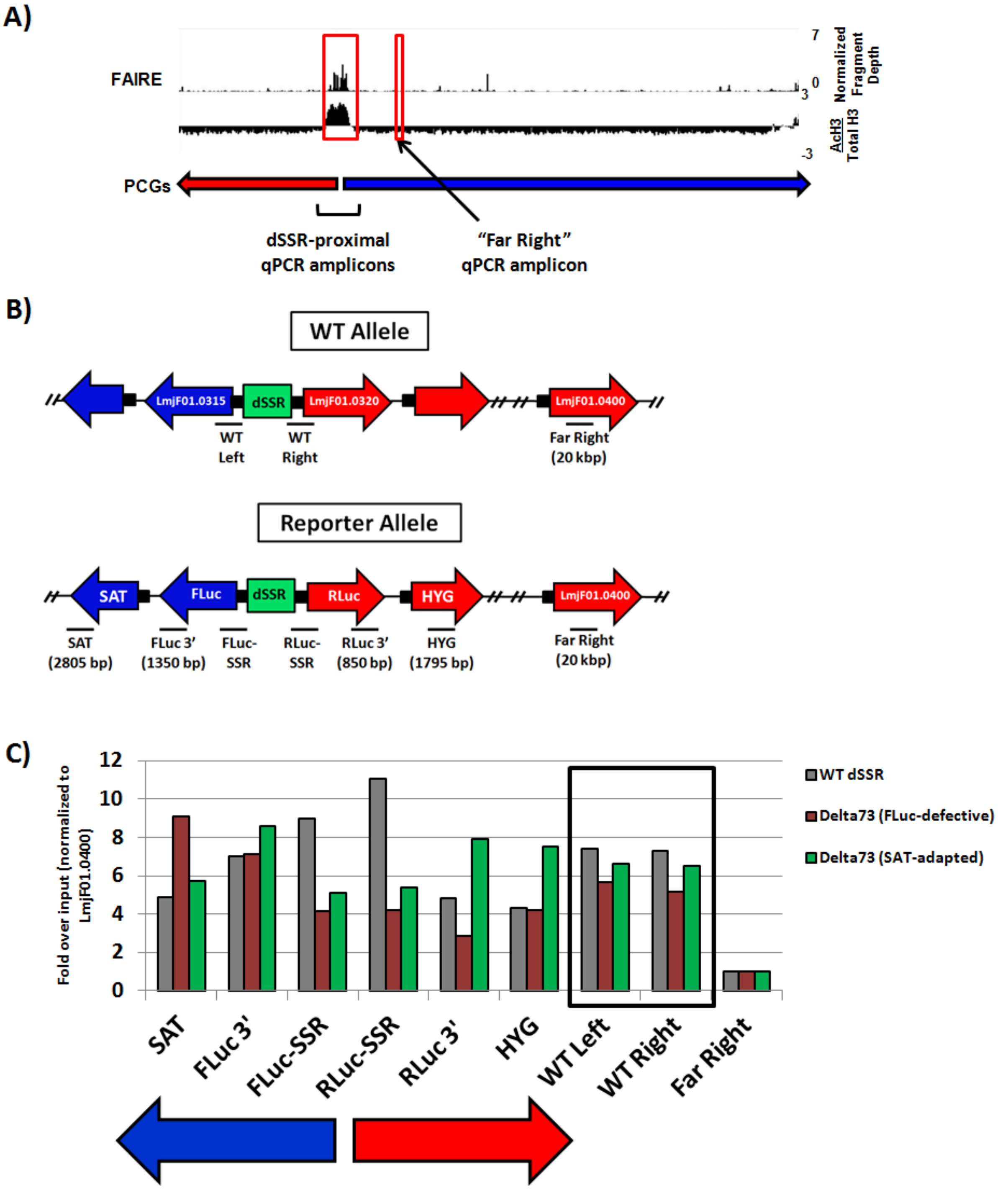
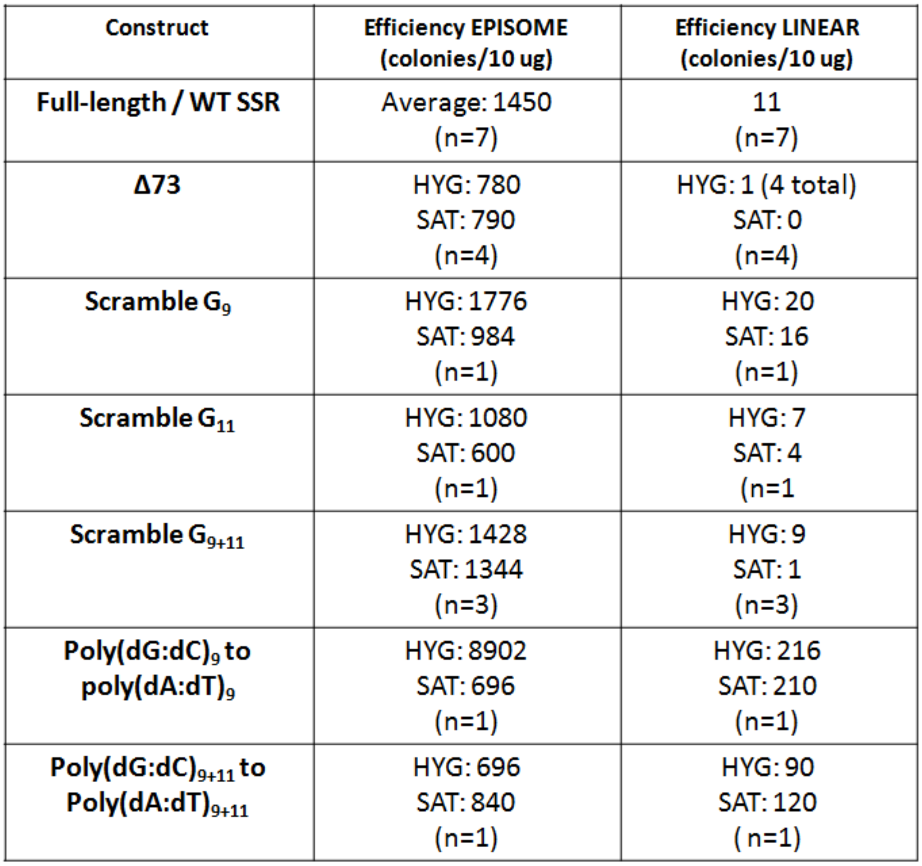


Table 5-1. Quantitation of episomal and integrated colonies from pLUC poly(dG:dC) tract mutants.



Construct	Efficiency EPISOME (colonies/10 ug)	Efficiency LINEAR (colonies/10 ug)
Full-length / WT SSR	Average: 1450 (n=7)	11 (n=7)
Δ73	HYG: 780 SAT: 790 (n=4)	HYG: 1 (4 total) SAT: 0 (n=4)
Scramble G_9	HYG: 1776 SAT: 984 (n=1)	HYG: 20 SAT: 16 (n=1)
Scramble G_{11}	HYG: 1080 SAT: 600 (n=1)	HYG: 7 SAT: 4 (n=1
Scramble G_{9+11}	HYG: 1428 SAT: 1344 (n=3)	HYG: 9 SAT: 1 (n=3)
Poly(dG:dC)$_9$ to poly(dA:dT)$_9$	HYG: 8902 SAT: 696 (n=1)	HYG: 216 SAT: 210 (n=1)
Poly(dG:dC)$_{9+11}$ to Poly(dA:dT)$_{9+11}$	HYG: 696 SAT: 840 (n=1)	HYG: 90 SAT: 120 (n=1)

Table 5-2. Primer sequences used in this study.

Primers	Sequences	Amplicon	Purpose
SMB5948	F: SMB4687 R: CGACAAAGTGTGTCGCACACCTTT CGTCATGCAGCCATTCTTGC	dSSR 5'	pLUC-Δ73
SMB5947	F: GCAAGAATGGCTGCATGACGAAAG GTGTGCGACACACTTTGTCG R: SMB4688	dSSR 3'	pLUC-Δ73
SMB6462	F: SMB4687 R: GGCAAGCGCGGCCGCGATGGGCG GTTCGTCATGCAGC	dSSR 5'	pLUC-SSR(ScrambleG9)
SMB6463	F: CCATCGCGGCCGCGCTTGCCAACAG CCC R: SMB4688	dSSR 3'	pLUC-SSR(ScrambleG9)
SMB6464	F: SMB4687 R: CCTGGAGCGCGGCCGCTGTTGGCAAGGG	dSSR 5'	pLUC-SSR(ScrambleG11)
SMB6465	F: CAGCGGCCGCGCTCCAGGTACAG GTTGCCTCCGAG R: SMB4688	dSSR 3'	pLUC-SSR(ScrambleG11)
SMB6466	F: SMB 4687 R: GCTGTTGGCAAGCGCGGCCGCGAT GGGCGGTTCGTCATGC	dSSR 5'	pLUC-SSR(ScrambleG9/11)
SMB6467	F: CTTGCCAACAGCGGCCGCGCTCC AGGTACAGGTTGCCTCCG R: SMB4688	dSSR 3'	pLUC-SSR(ScrambleG9/11)
SMB6619	F: SMB4687 R: CTGTTGGCAATTTTTTTTTTTATGGGC GGTTCGTCATGCAGC	dSSR 5'	pLUC-SSR(G9-A9)
SMB6620	F: CCGCCCATAAAAAAAAAAAATTGCCA ACAGCCCCCCCCCTCCAGGTACAG R: SMB4688	dSSR 3'	pLUC-SSR(G9-A9)
SMB6623	F: SMB 4687 R: CTGTTGGCAATTTTTTTTTTTATG GCGGTTCGTCATGCA	dSSR 5'	pLUC-SSR(G9/11-A9/11)
SMB6624	F: TTGCCAACAGAAAAAAAAAAATCCAGG TACAGGTTGCCTCCG R: SMB4688	dSSR 3'	pLUC-SSR(G9/11-A9/11)
SMB5221 **SMB5222**	F: CGTTGCCAACACGTTCACGCGGCTG R: CTCAAGCCTTCATGCCTTCATGCCTCTC	LmjF01.0190	Copy number analysis
SMB3289 **SMB3290**	F: ATGAGATAGACGTTTCCAAGGCA R: GATGGCGACCAAGGTGTCAC	LmjF05.0060	Copy number analysis

Chapter Six

Elements, strategies, and tactics oriented towards the development of a system for inducible transcription in *Leishmania*

Preface

The ideas described in this work arose in discussions with BA, SMB, and many members of the Beverley group over the years. The experiments were designed by SMB and BA. The experiments were performed and analyzed by BA with supervision and guidance from SMB, Katherine Owens, and George Lye. All *Leishmania* artificial chromosomes were developed and characterized by Jim Schwarz. The first draft of this chapter was written by BA, and comments from SMB were included in the final draft presented here. This chapter is not intended for publication but to set forth some interesting concepts and preliminary studies that may assist in the quest for a usable system of inducible transcription.

Abstract

Systems that facilitate controlled gene expression in response to a defined stimulus such as heat shock or addition of a small molecule have made a number of genetic studies possible in eukaryotic systems. In addition to regulating the expression of individual gene products, these systems can be used to control expression of stem-loop RNAs which are substrates for the RNA interference (RNAi) pathway, allowing rapid, specific, and robust knockdown of messenger RNAs (mRNAs) of interest without requiring additional genetic manipulations, which are often time-consuming. While these systems are used in other kinetoplastids, a robust, tightly-controlled system has not been developed for use in *Leishmania* species. The recent advances in RNAi technologies in *L. (Viannia)* subgenus have made the development of such a system a higher priority, as these technologies require stable transfection of RNAi transgenes and cannot be used to target essential genes at this time. In this chapter, we describe the development of a conditionally-expressed T7 RNA polymerase (T7RNAP) protein based on the previously utilized destabilization domain approach. Preliminary data demonstrates exceptional regulation of the destabilization domain-T7RNAP (ddT7RNAP) fusion protein, and assessment of T7RNAP activity using a T7 promoter-driven *LacZ* reporter demonstrates that transcription is specific to lines containing the *ddT7RNAP* transgene. However, we observe that the "off" state activity of the destabilized ddT7RNAP still generates significant production of β-galactosidase, demonstrating that very little ddT7RNAP protein is required to nearly saturate the T7 promoter. Despite this lack of reporter regulation in its current configuration, these preliminary experiments demonstrate that such a system is likely feasible in the context of an appropriate "off state", and in the context of the work presented throughout this thesis provide a useful perspective on modifications which may reduce background gene expression.

Introduction

Our ability to understand the functions of genes *in vivo* in diverse eukaryotic systems has been greatly enhanced by the development of systems that allow strict, inducible control of gene expression. In complex *in vivo* systems, experimentally-controlled gene expression allows the characterization of individual genes at discrete points during development, in response to controlled stimuli, or in specific tissues [reviewed in (1)]. In addition, these systems circumvent some of the restrictions that accompany classical genetic approaches using null mutants or hypomorphic alleles. First, inducible gene expression allows one to study genes that are essential for viability, as phenotypes can be monitored for some time after depletion of the gene product before viability is compromised. Second, inducible gene expression that is tightly controlled can sometimes reduce the need for complementation of null mutants, acting as a built-in experimental control; this is especially relevant for systems in which the generation of transgenic lines is difficult or time-consuming. Third, systems have been developed for inducible expression of short interfering RNAs (siRNAs), resulting in regulated knock-down of genes of interest using RNA interference (RNAi) in systems that may be less amenable to viral transduction methods (2).

At this time, few reagents allowing regulated gene expression are available for use in *Leishmania*, and their development has been complicated by the unusual nature of transcriptional regulation in these organisms. The *Leishmania* field would benefit greatly from a robust system for inducible gene expression, as traditional genetic approaches for *in vivo* characterization of gene functions are low-throughput, time-consuming due to the need to inactivate 2 or more alleles (see Chapter 1 for more details), and cannot be used for genes that are essential for viability. The potential utility of such a system became more evident upon the demonstration of a functional RNA interference (RNAi) pathway in the *Leishmania* (*Viannia*) subgenus (3). In this system, double-stranded RNAs (dsRNAs) are produced from 500-1000 bp stem-loop (StL) transgenes, which are then processed into short, 22-23 nucleotide RNA fragments called short interfering RNAs (siRNAs). These siRNAs are incorporated into the RNA-induced silencing complex (RISC), which identifies mRNAs bearing complementary sequences and degrades them, resulting in efficient, specific knock-down of the gene of interest (GOI) (3,4). Importantly, the targeting of genes using RNAi, much like the generation of null mutants, can only be accomplished for genes that are not essential for viability, as transfectants cannot be obtained for further study if an essential gene is targeted (Brettmann, Marcus, Lye, and Beverley, unpublished data). These challenges demonstrate a real need in the *Leishmania* field to develop a robust system for inducible expression of both protein-coding genes and RNAi transgenes.

In one approach, members of the Beverley lab adapted a reliable system for conditional gene expression using ligand-mediated alteration of protein stability for use in *Leishmania*. Here, the GOI is fused to a small engineered module derived from the FKBP12 protein called the destabilization domain (dd) (5). In the absence of stabilizing ligands, the dd is unfolded, resulting in targeted degradation of the fusion protein by the proteasome. The dd interacts specifically with the small molecule rapamycin and its derivatives FK506 and Shield1, and this interaction stabilizes the protein to effectively restore gene function (Fig. 6-1). Initial experiments in *Leishmania major* and *Leishmania braziliensis* demonstrated its utility using dd-tagged yellow fluorescent protein (ddYFP) and firefly luciferase (ddLUC) transgenes, as well as several endogenous genes, including enzymes required for lipophosphoglycan (LPG) biosynthetic and folate metabolism (6). In these studies, protein levels were tunable and rapidly altered by the addition or removal of the small molecule ligand, and both rapamycin and FK506 were well-tolerated by *Leishmania*. More importantly, enzymatic function also correlated with small molecule ligand concentration, demonstrating that this system may be used to regulate protein activity via modulation of its abundance.

Although this conditional protein stabilization method is the only inducible system which is tightly-regulated in *Leishmania* at this time, some aspects of this system limit its utility. First, the genetic manipulations required to establish conditional protein expression for endogenous genes are no faster than those previously used, still relying on multiple rounds of allelic replacement to generate strains bearing a single, dd-tagged transgenic copy of the gene of interest. Second, this system does not appear to work well for all proteins, especially those that are localized to subcellular compartments, such as the glycosome (6) or the mitochondrion [(5), Vickers and Beverley, unpublished data]. Finally, attempts to regulate essential proteins have not always yielded inducible "death", possibly due to the induction of stress chaperones, which stabilize the ligand-free destabilization domain, or other "leakiness". In my own attempts to utilize this system to further characterize the function of H2A.Z, an essential histone variant that we described in detail in Chapter 2. I was able to successfully generate parasites that relied on expression of a *ddH2AZ* fusion gene, and the levels of the ddH2A.Z fusion protein were dependent on the small molecule FK506 (Anderson and Beverley, unpublished data). However, removal of the stabilizing ligand from the medium did not result in complete ablation of ddH2A.Z protein levels in the absence of wild-type *H2AZ* genes, and approximately 10% of total H2A.Z remained, which appeared sufficient to confer viability in the absence of the FK506 inducer.

Potentially, this destabilization domain system could be used to regulate RNA levels through RNA interference (RNAi) or transcription through the control of RNA polymerase or repressors. For RNAi, this would require modulation of a RISC component, such as the catalytic protein Argonaute. *Ago1*- parasites are viable but lack RNAi, and we introduced a dd-AGO1 fusion protein into this line (Owens and Beverley, unpublished data). However, such "tunable RNAi" lines have not successfully regulated RNAi activity, as judged by a luciferase-based RNAi reporter assay (Owens and Beverley, unpublished data), perhaps due to "leakiness" or a saturation phenomenon described above.

In considering other strategies for inducible gene expression, it is apparent that the vast majority rely on transcriptional regulation for controlled gene expression. These systems typically use exogenous *cis*- and *trans*-acting elements that are not present in the host's genome but maintain the ability to interact with the cellular transcriptional machinery. In these systems, the gene of interest (GOI) is placed under the control of a *cis*-regulatory motif, which interacts with its cognate *trans*-acting factor to direct expression of the GOI. The inducible nature of this system can be conferred in a variety of ways, including by ligand-dependent interactions between these factors, by regulation of *trans*-acting factor levels using conditional destabilization, or by controlling *trans*-acting factor expression using a tissue-specific or environmentally-responsive promoter element [reviewed in (1)]. In addition, this system can be implemented with both transcriptional activators, which turn gene expression on when the correct *cis*-regulatory motif is recognized, and with transcriptional repressors, which turn gene expression off when the correct *cis*-regulatory motif is recognized.

Importantly, there is precedent for successful implementation of a system for transcriptionally-regulated inducible gene expression in kinetoplastids, despite the fact that they do not regulate individual protein coding genes at the level of transcription. In *T. brucei*, several variations of a multilayered system have been developed that facilitate inducible expression of protein-coding genes and RNAi transgenes (7–11). Here, two sets of *cis*-regulatory elements and their cognate *trans*-acting factors are implemented: the bacterial TetR repressor and the bacteriophage-derived T7 RNA polymerase. In one regulatory pair, the repressor protein TetR binds to its *cis*-regulatory motif the Tet operator (*tetO*) in a tetracycline-dependent manner and the interaction is disrupted in the presence of tetracycline. Therefore, introduction of the *tetO* sequence upstream of any promoter element effectively represses promoter activity when levels of TetR are high, and this repression is alleviated in the presence of tetracycline. The second regulatory pair in this system is that of the bacteriophage protein T7 RNA polymerase (T7RNAP), which requires the *cis*-regulatory element P_{T7} to initiate transcription. Importantly, the T7RNAP system can be used to generate dsRNAs using "dueling promoters" when P_{T7} is placed on either side of the transgene in opposite orientations, which greatly simplifies the preparation of libraries of RNAi transgenes for screening purposes.

In these systems, the tight regulation of inducible gene expression depends on the "leakiness" of transcription of the GOI. This can be affected not only by the cellular transcriptional machinery, which can be controlled by the locus used for GOI expression, but also by leaky transcription that alters the interactions between *cis*-regulatory elements and their *trans*-acting factors, which can be modified by the location, orientation, and combination of *cis*-regulatory motifs. In *T. brucei*, the GOI is usually integrated into the ribosomal RNA (rRNA) spacer sequence, which is not transcribed; moreover, all protein-coding sequences are integrated in the opposite orientation to the rRNA genes, minimizing productive transcription of these genes by RNA polymerase I (RNAP I). As a second layer of regulation, the expression of TetR and T7RNAP are often controlled. Typically, T7RNAP is expressed from a strong endogenous promoter, and the TetR is placed under the control of T7RNAP to generate high levels of this protein. In this system, regulated expression of firefly luciferase generated approximately 1 luciferase protein per cell in the “off” state, making it suitable for even the most toxic of gene products. However, these motifs may also be used to control T7RNAP levels—by placing the T7RNAP transgene under control of the TetR/*tetO* system, both transcription of T7RNAP and derepression of the GOI would require the addition of tetracycline, again resulting in tight regulation of GOI expression.

Although this T7RNAP-TetR hybrid system has been extremely effective in *T. brucei*, its implementation in *Leishmania* has been less successful (12). Part of the problem appears to be that unlike *T. brucei*, a transcriptionally silent region has yet to be identified. Notably, the rRNA spacer region differs significantly, and some data point to transcription across the 63 nt repeats comprising this region (13) A TetR/*tetO* inducible system was developed for use in *L. donovani*, in which a TetR/*tetO*-regulatable ribosomal RNA promoter was used to drive expression of a gene of interest. Although some regulation was obtained in this context using transient transfections, its dynamic range was much lower than that of *T. brucei*, and GOI repression in the absence of tetracycline did not correlate well with TetR levels. Moreover, expression of the GOI was relatively high in the absence of tetracycline, demonstrating that the “off” state was prone to leaky transcription of the GOI. Importantly, these differences in “off” state appear to be the key difference between a successful system in *T. brucei* and a poorly regulated one in *Leishmania*. Despite the fact that the current inducible systems in *Leishmania* leave much to be desired, the knowledge generated from these experiments and previously demonstrated inducible systems in *T. brucei* is tremendously useful going forward, especially when considered relative to the conditional protein regulation system more recently demonstrated in *Leishmania*.

In this chapter, we describe our efforts to improve upon current inducible systems using a dd-tagged T7RNAP to regulate gene expression at the transcriptional level. In this system, we anticipated we would be able to induce T7RNAP-dependent gene expression after the addition of rapamycin or FK506 to the growth medium, and additional layers of regulation could be added into the system using the TetR/*tetO* repressor system. To quantitatively assay T7RNAP activity in these lines, we used previously developed *Leishmania* artificial chromosomes (LACs) containing the *LacZ* reporter gene under the control of P_{T7}. These efforts, while still not perfect, demonstrate great potential for a layered strategy for inducible gene expression, and we describe a set of expression vectors which may improve the regulatory potential of this system. We believe that these data, as well as the knowledge generated in the work described in the previous chapters of this thesis will aid in the identification of the ideal "silent" locus for inducible transgene integration and development of a robust system for inducible gene expression in *Leishmania*.

Results

Destabilization domain-tagged T7 RNA polymerase is conditionally expressed in L. major

I began by generating *Leishmania* expressing a destabilization domain-regulated copy of T7 RNA polymerase and characterizing the conditional expression of the fusion protein using the small molecule FK506. The *Leishmania* vector pIR1 is designed to integrate into the ribosomal small subunit (SSU) after linearization, and this plasmid facilitates high levels of expression of up to 2 protein-coding genes using the ribosomal promoter (P_{rRNA}). A plasmid-encoded T7 promoter sequence appeared to complicate our efforts to clone T7RNAPnls into this plasmid, as we only obtained plasmids bearing T7RNAPnls in the antisense orientation. However, modification of pIR1 by inverting the vector backbone with SwaI digestion and re-ligation to generate the pIR1F corrected this problem, and we obtained pIR1F plasmids containing the ddT7RNAPnls fusion gene in the correct orientation (schematized in Fig. 6-2A). Transfection of linearized DNA into WT *L. major* in the absence of ddT7RNAP protein stabilization generated parasite lines containing this integrated construct without difficulty. The proper integration of this construct into the ribosomal SSU was confirmed by PCR using primers which span the 5' and 3' junctions of the construct (Fig. 6-3).

We next sought to quantify the degree of ddT7RNAPnls protein regulation in the presence and absence of FK506 (described in Fig. 6-1) by western blotting. We selected 4 clones for these tests, growing them in the presence and absence of 1 µM FK506 for 48 hours. Total cell lysates were resolved on a polyacrylamide gel, transferred, and probed with antisera against T7 RNA polymerase or against *Leishmania* histone H2A as a loading control (Fig. 6-4A). We observe significant induction of ddT7RNAPnls expression in lines grown in the presence of FK506, demonstrated by a T7RNAP-reactive band at the expected molecular weight. In lines grown in the absence of FK506 we observe no signal above background from the T7RNAP antibody, even using a highly sensitive western blot detection method. Quantitative analysis of these blots shows that ddT7RNAPnls levels in the absence of the stabilizing small molecule are below the limit of detection.

Quantitative analysis of T7RNAP activity in lines expressing regulatable ddT7RNAPnls

To understand whether the activity of T7RNAP was also regulatable in these lines, we turned to *Leishmania* artificial chromosomes (LACs) (Schwarz and Beverley, unpublished data). These constructs contain telomeric DNA at their ends and are propagated as 1-2 copies per cell, although they are shorter than the endogenous chromosomes. The constructs used in these studies contain the *LacZ* gene, which encodes the β-galactosidase protein, and the *NEO* selectable marker. The T7 promoter in these constructs is located either upstream of the RNA processing sequences associated with the *LacZ* gene, or those associated with the *NEO* gene (Fig. 6-2B). Importantly, the *LacZ* gene is only 300-400 base pairs from the telomeric DNA, and both LACs express very low levels of β-galactosidase in the absence of T7RNAP (Schwarz and Beverley, unpublished data). We chose to pursue additional studies using the integrated ddT7RNAPnls clone 4, as it had the highest levels of T7RNAPnls protein in the presence of FK506. We transfected linearized LAC constructs containing P_{T7} upstream or downstream of *LacZ* into both the ddT7RNAPnls clone and into WT *L. major*, which define a full panel of controls for T7RNAP-dependent transcription of the *LacZ* gene (Fig. 6-2B). To confirm the presence of the appropriate transgenes, we used PCR to amplify the appropriate selectable markers or to amplify across the junctions of transgene integration (data not shown).

To quantify the relative levels of T7RNAP activity in these lines, the ddT7RNAPnls-LAC clones and their parent line were grown in the presence or absence of 1 µM FK506 for 48 hours. The relative β-galactosidase activity was quantified as a proxy for T7RNAP activity using the β-galactosidase substrate 4-methylumbelliferyl-β-D-galactoside, which is converted into a fluorescent product by β-galactosidase (Fig. 6-5A). In this assay we observe very low levels of β-galactosidase activity in the ddT7RNAPnls parent lines without a LAC, as well as in WT parasites transfected with either the BG-T7 or T7-BG LAC. Similarly, we observed very low activity in the ddT7RNAPnls line containing the BG-T7 LAC, demonstrating that the ddT7RNAPnls fusion protein is not transcribing the *LacZ* gene non-specifically. Importantly, we observed robust activity in the ddT7RNAPnls line containing the T7-BG LAC when the line is grown in the presence of FK506, demonstrating that the ddT7RNAPnls fusion protein is active. However, this activity is only slightly reduced in the same line grown in the absence of FK506. Because these lines had been through additional rounds of transfection, it was possible that the dynamic regulation we observed in the parent lines was not being maintained in the ddT7RNAPnls/LAC clones. To confirm the proper regulation of the ddT7RNAPnls fusion protein, we collected total cell lysates at the same time as the β-galactosidase activity was performed. Western blot analysis of these lysates demonstrated similar FK506-dependent regulation of ddT7RNAPnls protein levels to those in the parent lines (Fig. 6-5B), suggesting that the minute levels of residual protein remaining under destabilization domain "off" conditions were still capable of nearly saturating the T7 promoter in this context. In the future, other efforts to reduce the baseline of ddT7RNAP protein expression could be useful. In these experiments we selected the clone bearing the highest ddT7RNAP protein level in the "on" state; selection of a clone with lower "on" state expression may improve this somewhat. Furthermore, integration of the dd-tagged transgene into an RNA polymerase II-transcribed locus would likely reduce the background levels of T7RNAP activity without significantly altering "on" state activity, as was shown for the dd-GLF fusion (6).

Discussion

Considerations for additional advances in regulatable gene expression in Leishmania

Although the T7RNAP activity did not regulate reporter gene expression as well as expected given the apparent on/off nature of the ddT7RNAPnls protein in this system, the preliminary data described here are promising and are very useful in considering additional modifications which may provide better regulation. Previous experiments using the LAC constructs led to the proposal of a "landing pad" model, in which the expression of genes is dependent on the distance from the telomere (Fig. 6-6; modified from Schwarz and Beverley, unpublished data). The data described throughout this thesis are consistent with this model, as it is apparent that chromatin state, rather than *cis*-regulatory motifs, define transcriptionally permissive loci (Chapter 4). It is clear that in addition to decreasing background levels of the ddT7RNAP fusion protein in the "off" state, the identification of a suitable, transcriptionally silent locus is necessary to prevent transcription of the gene of interest by the cellular polymerases. We have envisioned several loci that could be tested by others in the future, and I anticipate that a concerted effort to find such a locus will advance this system significantly.

In addition to the identification of a suitable locus for transgene integration, we believe that small, tractable modifications to the ddT7RNAP system may be sufficient to bring this inducible system into the mainstream. We demonstrated here that the levels of the ddT7RNAPnls protein regulate beautifully using the stabilizing ligand FK506, and the ddT7RNAPnls protein is active. Although the residual ddT7RNAPnls protein is sufficient for near-saturation of the T7 promoters in the context of the low-copy LAC, several modifications may decrease T7RNAP activity sufficiently for use in an inducible system. First, the T7RNAP protein used in these experiments contains a nuclear localization signal (NLS). Comparisons of the NLS-containing version with the diffusely-localized "wild type" version demonstrated that T7RNAP activity was approximately 10-fold lower when the NLS was removed, as some protein is transported into the nucleus (LeBowitz and Beverley, unpublished data). Second, it is possible that addition of "decoy" T7 promoters might decrease T7RNAP-dependent transgene expression from these LACs. Finally, the ddT7RNAPnls transgene is expressed from the highly active ribosomal promoter. It is likely that expression of the fusion protein from an RNAP II-transcribed locus, such as the tubulin array, may decrease the "off" state sufficiently while still allowing promoter-saturating ddT7RNAPnls levels in the presence of FK506.

More importantly, these experiments did not explore the TetR/*tetO* repressor system, which has been shown to regulate gene expression to some extent in *Leishmania*. In parasite lines expressing both ddT7RNAPnls and the TetR repressor, we expect that growth in the absence of both tetracycline and FK506 would result in extremely low levels of transgene expression if it is present in a transcriptionally-silent locus, similar to those shown in *T. brucei*. Addition of tetracycline or FK506 to the medium would result in leaky transgene expression, either by residual ddT7RNAPnls protein or by the cellular polymerases. However, addition of both tetracyline and FK506 would promote high levels of transgene expression, due to the lack of TetR-mediated repression and the stabilization of the ddT7RNAP fusion protein. These additional experiments build upon the groundwork that was laid out here, and we expect that these projects will be taken on by others in the laboratory in the near future.

Methods and Materials

Plasmid construction

The destabilization domain was excised from pGEM-BclI-dd-BglII (B6177) by BclI and BglII digestion and was ligated into BglII-digested and CIP-treated pIR1-phleo (B4054) to generate pIR1-dd(B)-phleo (B6392). This plasmid was digested with SwaI, and the resulting fragments were purified and re-ligated to invert the vector backbone, generating the plasmid pIR1F-dd(B)-phleo (B6395). The T7RNAPnls gene was amplified from the expression vector pX63-T7RNAPnls using the primers B3707 and B3708; the PCR product was digested with BglII and was ligated into BglII-digested B6392 to generate the plasmid pIR1F-ddT7RNAPnls

(B)-phleo (B6411). All constructs were confirmed by restriction endonuclease digestion and sequencing. The generation of the LAC constructs was described previously (Schwarz and Beverley, unpublished data).

Cell culture and transfection

Cell cultures and transfections were performed as described in Chapter 2. Transfections were selected using 10 µg/mL phleomycin (Invivogen) or with 10 µg/mL G418. Induction of ddT7RNAPnls stabilization was accomplished using 1 µM FK506 (LC Laboratories).

Western blotting

Total cell lysates were prepared as described in Chapter 2. Total cell lysates from 6 x 10^6 cells were resolved on a 4-16% polyacrylamide gel (Bio-Rad) by SDS-PAGE, electroblotted onto Hybond-ECL nylon membranes (Amersham Biosciences), and blocked with Odyssey blocking buffer (Li-Cor). Primary incubations were performed using 1:10,000 mouse anti-T7RNAP (Millipore) and 1:10,000 anti-H2A in Odyssey blocking buffer. Secondary incubations were performed with 1:10000 IR680-labeled goat anti-rabbit or IR800-labeled goat anti-mouse antibody (Li-Cor) and blots were analyzed and quantified using the Odyssey imaging system (Li-Cor).

B-galactosidase assay

Cells were collected from early logarithmic phase cultures (2-4 x 10^6 cells/mL), washed once with Dulbecco's phosphate-buffered saline, and resuspended in a reaction buffer containing 23 mM Tris-HCl, pH 7.5 (Fisher Scientific), 125 mM sodium chloride (Fisher Scientific), 2 mM magnesium chloride (Fisher Scientific), 12 mM β-mercaptoethanol (Sigma), and 1X cOmplete Protease Inhibitor Cocktail (Roche). Cells were lysed by adding sodium dodecyl sulfate to 0.5% (Sigma). A standard curve using recombinant β-galactosidase (Sigma) was prepared in enyzme buffer containing 0.5% SDS, and the β-galactosidase substrate 4-methylumbelliferyl-β-D-galactoside (Sigma) was added to a final concentration of 0.3 mM. Recombinant β-galactosidase and lysates were incubated for 4 hours at 37°C. Fluorescence was quantified using a Bio-Rad Fluoromark fluorimeter, with a 355 nm excitation filter and a 460 nm emission filter.

Figure Legends

Figure 6-1. Regulation of destabilization domain-gene of interest fusion proteins using the small molecule FK506. A transgenic copy of the destabilization domain-gene of interest (ddGOI) fusion is expressed from an integrated genomic locus or as an episomal copy. Transcription and translation of the fusion gene result in a ddGOI fusion protein, which is unstable and will be targeted for degradation by the proteasome. The fusion protein is stabilized upon addition of the small molecule FK506, restoring gene function.

Figure 6-2. Schematic representation of constructs described in this chapter. (A) Integration of pIR1-based constructs into the ribosomal small subunit (SSU). The genomic organization of the ribosomal RNA array is depicted at top, with LSU designating the large subunit components. Linearized pIR1 constructs integrate into the SSU using homologous recombination at the 5' and 3' ends of the linear fragment. Genes are depicted as colored boxes, and the dd-T7RNAP fusion protein is depicted in purple, and the selectable marker (PHLEO) is designated in charcoal. 5'splice acceptor sequences are depicted as black boxes. (B) Depiction of *Leishmania* artificial chromosomes (LACs) used for characterization of T7RNAP activity and specificity. Filled arrowheads indicate telomeric sequences, and genes and splice acceptors are designated as described in (A). The location and orientation of the T7 RNAP promoter element (P_{T7}) is designated with an angled arrow.

Figure 6-3. Validation of pIR1 integration into the ribosomal SSU. PCR primers outside of the targeting fragment were paired with vector-specific primers as depicted in the schematic at the top to validate integration. Both the 5' and 3' integration PCR controls are shown, using wild type (WT) *L. major* and H_2O as negative controls, and a previously validated pIR1 construct as a positive control (+).

Figure 6-4. Western blot confirmation of ddT7RNAPnls protein expression and regulation. Cells containing the ddT7RNAPnls transgene were incubated for 48 hours in the presence or absence of 1 µM FK506. Boiled lysates or known amounts of commercial T7 RNA polymerase were prepared in Laemmli buffer and 6 x 10^6 cells per lane were resolved on a 4-16% polyacrylamide gel. (A) Membranes were probed with anti-T7RNAP antibody or anti-*Leishmania* H2A antisera as described in the methods. (B) Quantitation was performed using the Licor Odyssey scanner.

Figure 6-5. Regulation of T7RNAP activity using ligand-mediated conditional protein stabilization. (A) β-galactosidase activity assay of strains expressing ddT7RNAPnls and a *Leishmania* artificial chromosome, as described in Fig. 5-2B. Cells were grown in the presence or absence of 1 µM FK506 for 48 hours prior to quantitation of β-galactosidase activity. (B)

Western blot verification of ddT7RNAPnls expression in the presence and absence of 1 μM FK506.

Figure 6-6. Proposed models for "silent" nature of LAC-based reporter systems. Description of a "landing pad" model describing telomere proximity-dependent transcription in *Leishmania* (adapted from Schwarz and Beverley, unpublished data).

References

1. Mills A. Changing colors in mice: an inducible system that delivers. Genes Dev 2001;15(12):1461–7.

2. Van de Wetering M, Oving I, Muncan V, Pon Fong MT, Brantjes H, van Leenen D, et al. Specific inhibition of gene expression using a stably integrated, inducible small-interfering-RNA vector. EMBO Rep 2003;4(6):609–15.

3. Lye L, Owens K, Shi H, Murta S, Vieira A, Turco S, et al. Retention and loss of RNA interference pathways in trypanosomatid protozoans. PLoS Pathog 2010;6(10):e1001161.

4. Atayde V, Shi H, Franklin J, Carriero N, Notton T, Lye L, et al. The structure and repertoire of small interfering RNAs in *Leishmania (Viannia) braziliensis* reveal diversification in the trypanosomatid RNAi pathway. Mol Microbiol 2013;87(3):580–93.

5. Banaszynski L, Chen L, Maynard-Smith L, Ooi A, Wandless T. A rapid, reversible, and tunable method to regulate protein function in living cells using synthetic small molecules. Cell 2006;126(5):995–1004.

6. Madeira da Silva L, Owens K, Murta S, Beverley S. Regulated expression of the *Leishmania major* surface virulence factor lipophosphoglycan using conditionally destabilized fusion proteins. Proc Natl Acad Sci U S A 2009;106(18):7583–8.

7. Wirtz E, Clayton C. Inducible gene expression in trypanosomes mediated by a prokaryotic repressor. Science 1995;268(5214):1179–83.

8. Biebinger S, Wirtz L, Lorenz P, Clayton C. Vectors for inducible expression of toxic gene products in bloodstream and procyclic *Trypanosoma brucei*. Mol Biochem Parasitol 1997;85(1):99–112.

9. Wirtz E, Leal S, Ochatt C, Cross G. A tightly regulated inducible expression system for conditional gene knock-outs and dominant-negative genetics in *Trypanosoma brucei*. Mol Biochem Parasitol 1999;99(1):89–101.

10. Alibu V, Storm L, Haile S, Clayton C, Horn D. A doubly inducible system for RNA interference and rapid RNAi plasmid construction in *Trypanosoma brucei*. Mol Biochem Parasitol 2005;139(1):75–82.

11. Wirtz E, Hartmann C, Clayton C. Gene expression mediated by bacteriophage T3 and T7 RNA polymerases in transgenic trypanosomes. Nucl 1994;22(19):3887–94.

12. Yan S, Myler PJ, Stuart K. Tetracycline regulated gene expression in *Leishmania donovani*. Mol Biochem Parasitol 2001;112(1):61–9.

13. Martínez-Calvillo S, Sunkin S, Yan S, Fox M, Stuart K, Myler P. Genomic organization and functional characterization of the *Leishmania major* Friedlin ribosomal RNA gene locus. Mol Biochem Parasitol 2001;116(2):147–57.

Figure 6-1. Regulation of destabilization domain-gene of interest fusion proteins using the small molecule FK506.

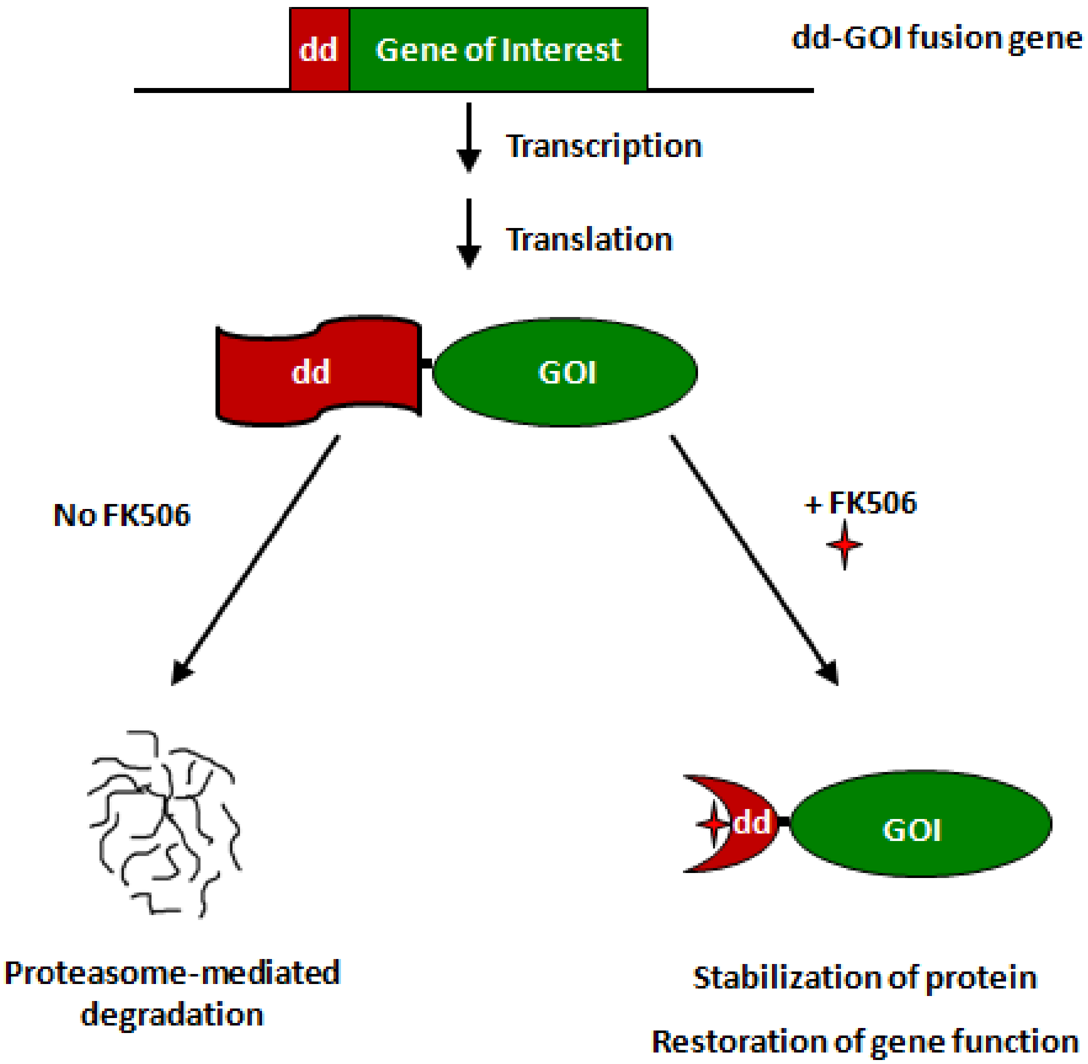
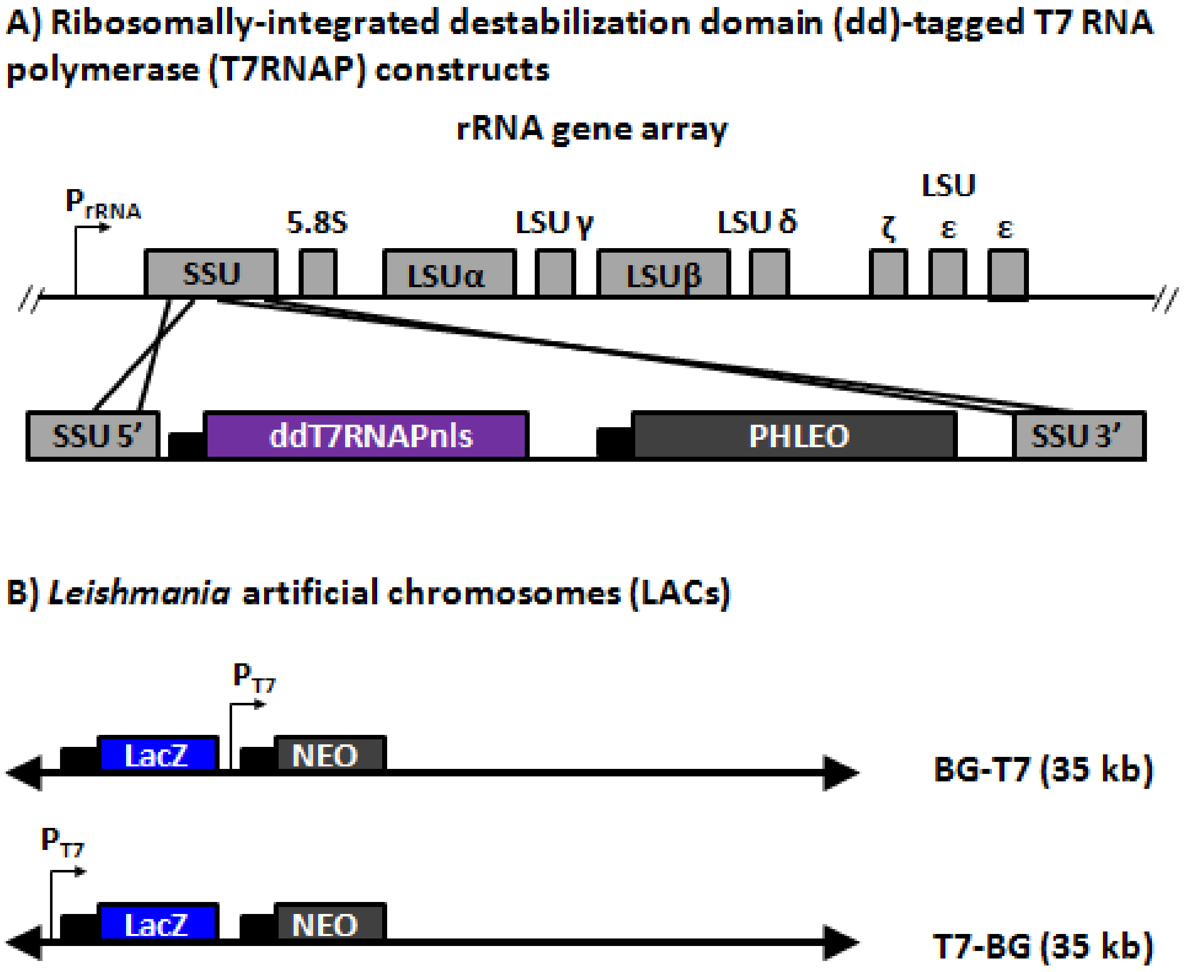


Figure 6-2. Schematic representation of constructs described in this chapter.



A) Ribosomally-integrated destabilization domain (dd)-tagged T7 RNA polymerase (T7RNAP) constructs

B) *Leishmania* artificial chromosomes (LACs)

Figure 6-3. Validation of pIR1 integration into the ribosomal SSU.

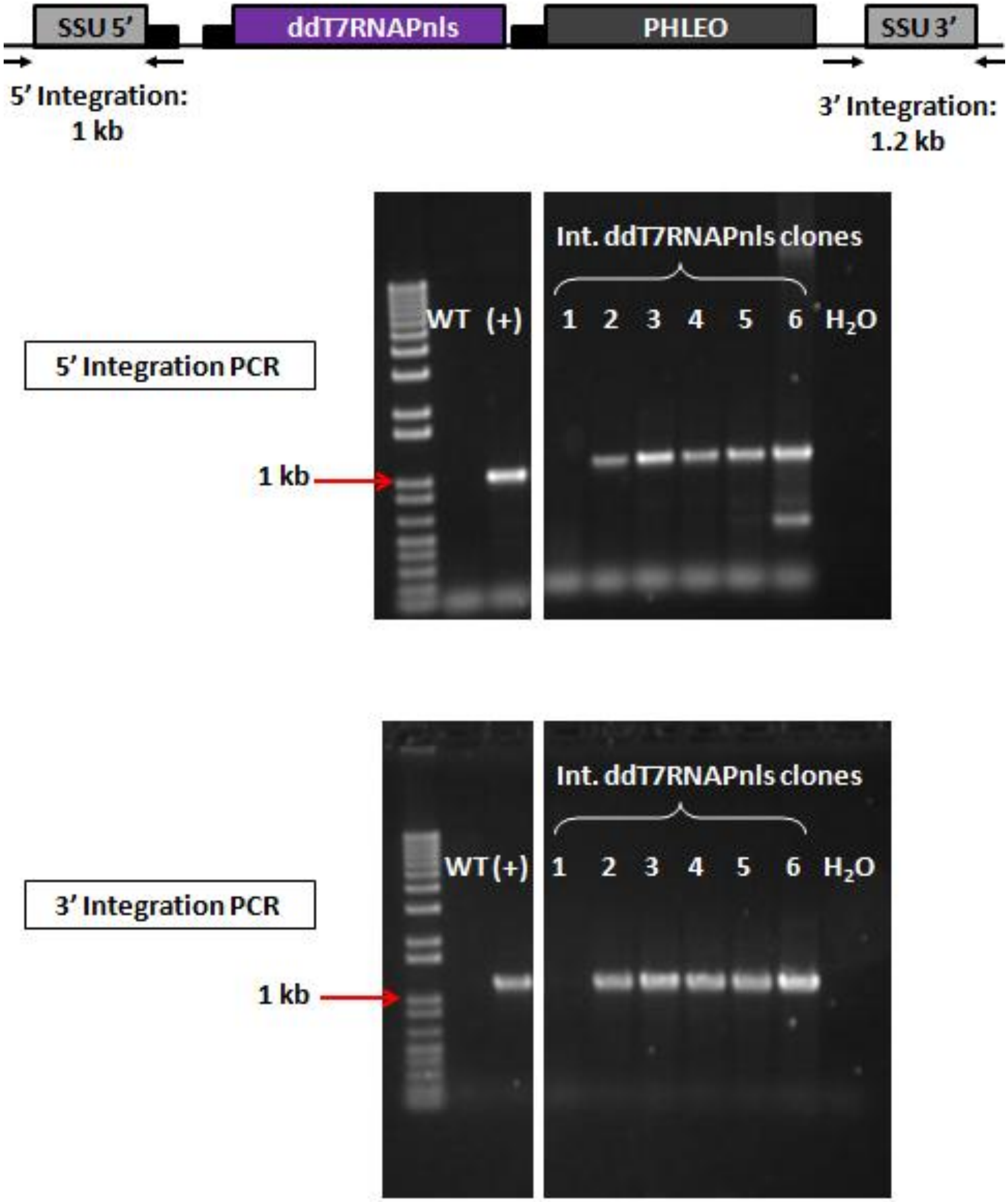


Figure 6-4. Western blot confirmation of ddT7RNAPnls protein expression and regulation.

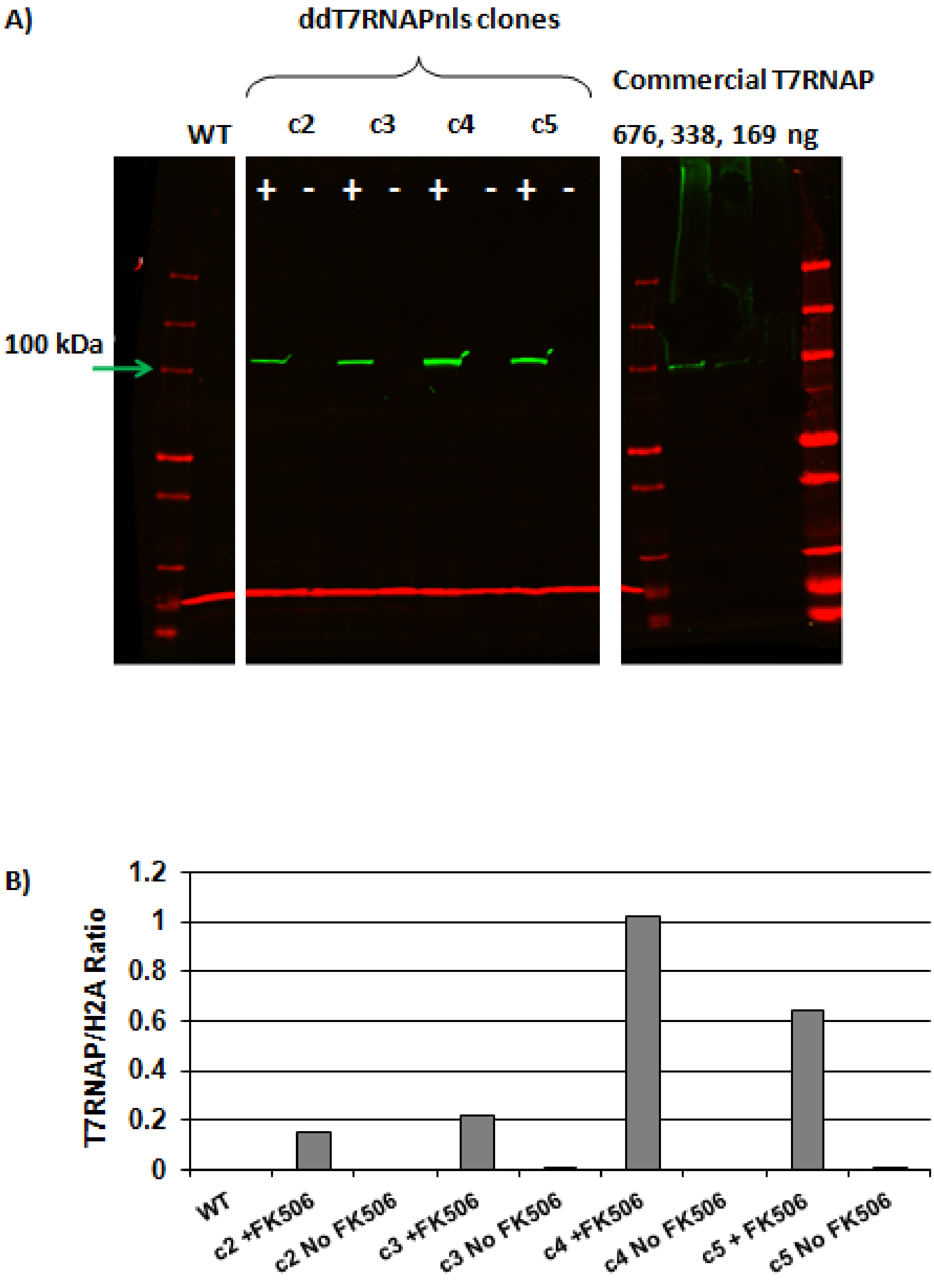


Figure 6-5. Regulation of T7RNAP activity using ligand-mediated conditional protein stabilization.

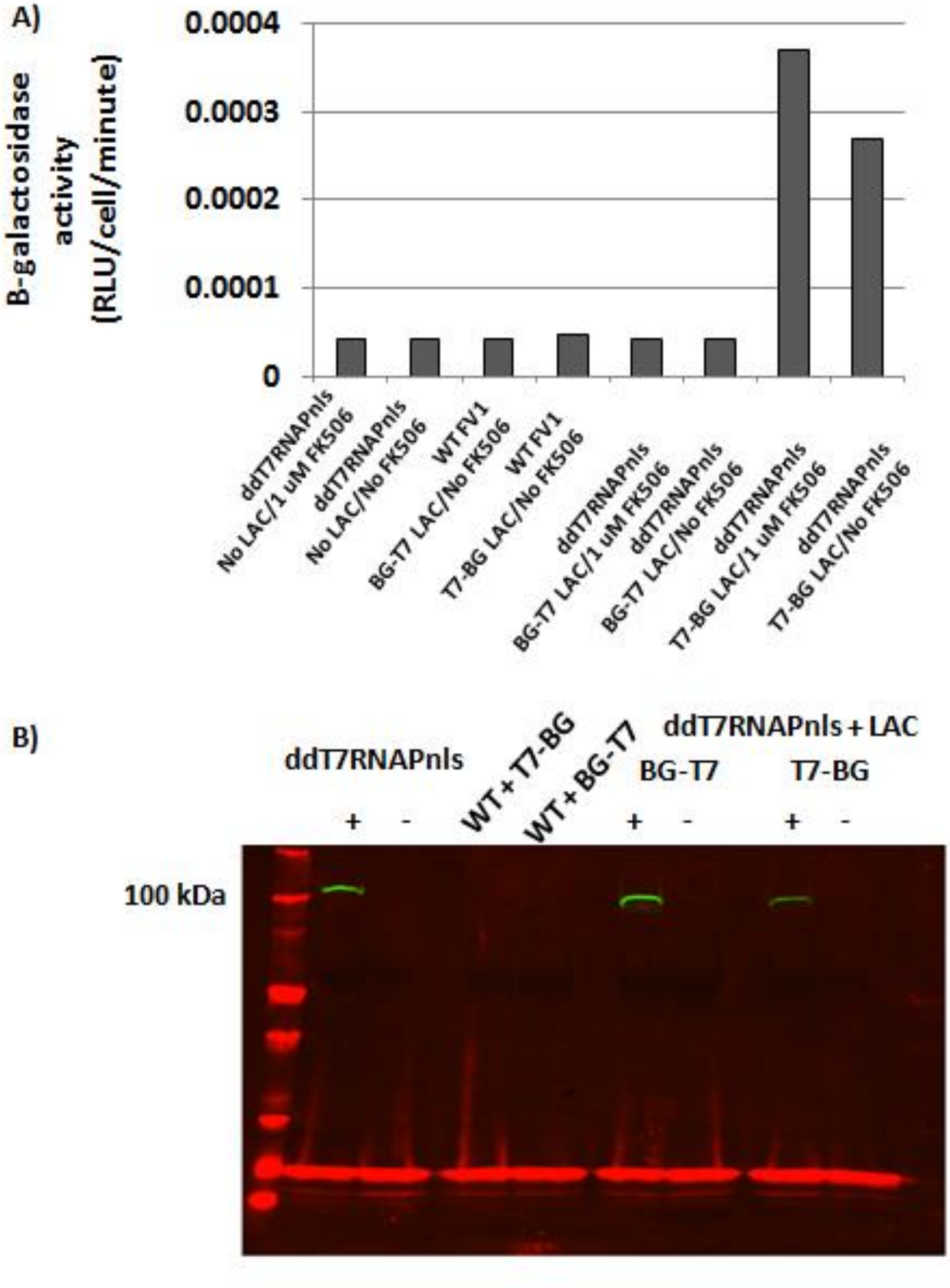
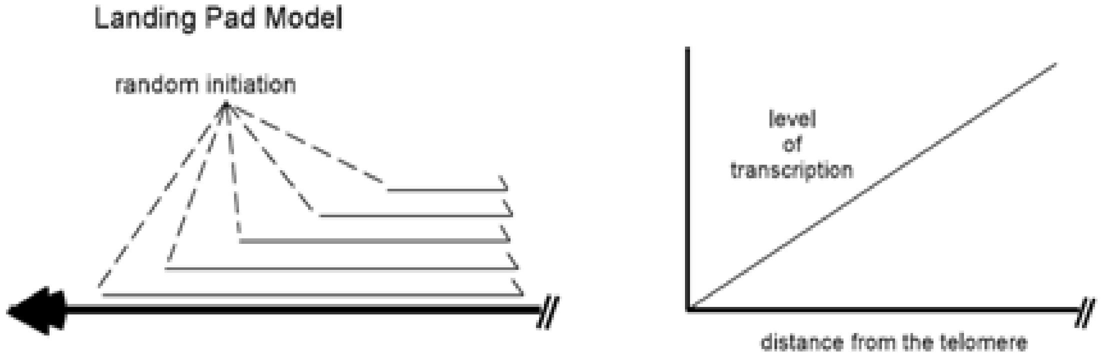


Figure 6-6. Proposed models for “silent” nature of LAC-based reporter systems.



Chapter Seven

Concluding remarks and future directions

Project goals

In most eukaryotes, DNA-encoded *cis*-regulatory elements and their cognate *trans*-acting factors are required for transcription of protein-coding genes. However, the levels of individual genes can be further controlled by alteration of the epigenetic state of the locus, which can be accomplished by the addition of chemical groups to histones or DNA or the incorporation of histone variants. Work in *Leishmania* and related trypanosomatid protozoa demonstrate that transcription of protein-coding genes is polycistronic, initiating in divergent strand switch regions (dSSRs) where polycistronic gene clusters (PGCs) are oriented head-to-head, and terminating in convergent strand switch regions (dSSRs) where PGCs meet tail-to-tail. Although a number of epigenetic marks have been localized to these regions, no DNA-encoded *cis*-regulatory motifs have been identified, and our understanding of the mechanisms controlling transcription initiation and termination are incomplete.

While it is widely appreciated that *Leishmania* and other trypanosomatid protozoa are highly unusual, it would be quite surprising if they did not require any DNA-encoded elements, as truly promoter-less genes have not been documented in other systems. In fact, RNA polymerase I (RNAP I), RNAP III, and the spliced leader (SL) locus transcribed by RNA polymerase II (RNAP II) more or less follow the eukaryotic paradigm [reviewed in (1)] However, trypanosomatid protozoa subject all polycistronic transcripts to *trans*-splicing, a processing step which defines the 5' end and is coupled to determination of 3' ends of the mature monocistronic messenger RNA (mRNA) (see Figure 1-2 in Chapter 1). This process circumvents the need for precise transcription initiation and termination, as internal transcription and termination events which would generate dysfunctional, deleterious gene products in other eukaryotes are corrected during mRNA maturation. Thus, it is conceivable that trypanosomatid protozoa may regulate transcription differently than model eukaryotes, relying solely on epigenetic mechanisms to ensure that the entire polycistronic gene cluster (PGC) is transcribed. The work described in this dissertation characterizes the contribution of genetic and epigenetic factors to the control of transcriptional events in *Leishmania*. We focus our efforts primarily on transcription initiation events in dSSRs, but our studies of histone variants in these networks also led us to characterize the role of the histone variant H3.V in transcription termination.

Analysis of genetic determinants of bidirectional transcription in divergent strand switch regions

In the introduction, we described a model in which nucleosome-disfavoring sequences such as poly(dG:dC) tracts might effectively function as *cis*-regulatory elements within dSSRs by influencing chromatin structural elements (Figure 1-4). These and other homopolymeric sequences function independently of *trans*-acting factors in yeast by inhibiting nucleosome incorporation (2), and the introduction of poly(dA:dT) tracts significantly alter the activity of weak promoters (3). The mechanism by which these sequences function in eukaryotic promoters fits well with what was previously known regarding RNA polymerase II-mediated transcription in trypanosomatid protozoa: these organisms lack specific transcriptional activator proteins, and canonical DNA-encoded promoter elements are absent from dSSRs (4); the general transcription factors demonstrate little sequence specificity and are capable of binding at low levels throughout the genome, with higher binding occurring within epigenetically-permissive regions of open chromatin (5,6); and transcription initiation events occur promiscuously in regions of open chromatin (7–9), which could be “nucleated” by these nucleosome-excluding sequences.

If this model sufficiently explained the phenomena controlling transcriptional events in dSSRs, we would expect that surveys of *in vivo* nucleosome positions would show that dSSRs contain nucleosome-depleted loci, which would likely coincide with poly(dG:dC) tracts and other previously unidentified nucleosome-excluding sequences, as not all dSSRs contain these motifs. To test this model, we interrogated the propensity of these sequences to disfavor nucleosomes in *Leishmania* using nuclease hypersensitivity (NH) assays. The locations of nucleosome-bound sequences can be surveyed genome-wide by using next-generation sequencing to assess DNA prepared using micrococcal nuclease (MNAse) digestion, and the identity of nucleosome-free sequences can be assessed similarly using formaldehyde-assisted isolation of regulatory elements (FAIRE), which preferentially removes histone-bound DNA. Using a novel computational pipeline to rigorously analyze datasets generated from these two classes of experiments and remove various forms of experimental, analytical, and technical artifacts, we found that *Leishmania* dSSRs lack well-positioned NH sites, instead demonstrating an abundance of heterogeneous, poorly-positioned NH sites scattered throughout regions marked with a transcriptionally permissive epigenetic state. This suggests that if poly(dG:dC) tracts do play a role in dSSR function, they operate differently in *Leishmania* than in other model eukaryotes.

While these data suggest that our previously described nucleosome-disfavoring model is not correct, they do not explicitly rule out the existence of a *cis*-regulatory element in dSSRs. We took these genome-wide experiments a step further, developing a novel integrated, bidirectional dual-luciferase reporter system that allows the identification of *cis*-regulatory elements within a specific dSSR in the proper *in situ* context. We were surprised to find that we were able to substitute the core dSSR of chromosome 1 with completely unrelated DNA sequences with no alteration in bidirectional reporter gene activity. Additional studies specifically manipulating the poly(dG:dC) tracts in this locus suggested a possible non-essential role for these sequences in defining the directionality of transcription, which was previously proposed by Siegel and colleagues (10). However, parasites expressing normal levels of dSSR-dependent reporter genes were generated with these poly(dG:dC) tract mutants, thus exhibiting incomplete penetrance, and lines demonstrating a unidirectional phenotype could be reversed with selection without genotypic alterations. These characteristics are suggestive of epigenetic events for which I obtained some supporting evidence (Chapter 5).

Although we find the data suggesting a lack of *cis*-regulatory elements in a large portion of the dSSRs to be quite convincing, additional experiments are needed to show that dSSRs lack *cis*-regulatory elements altogether. First, it is possible that the *T. brucei*-derived stuffer sequence, despite its lack of homology and structural similarity to the WT dSSR, contains a *cis*-regulatory element of unknown identity. I designed constructs bearing other, equally dissimilar sequences derived from other eukaryotes, and when tested I expect that these will validate our observations using the *T. brucei*-derived stuffer. Second, it is possible that the “minimal” endogenous splice acceptor sequences present within the dSSR contain a *cis*-regulatory element and that the Δ489 deletion construct failed to generate viable transfectants due to the close proximity of opposing splice acceptor sequences. I am currently assessing parasites bearing a completely artificial dSSR generated with synthetic splice acceptors, which will clarify the discrepancy between the Δ489 and Δ489 + S reporter lines. We expect that quantitative analysis of reporter gene activity in these lines along with examination of the epigenetic landscape of these dSSR mutants using FAIRE will make a strong case against the existence of *cis*-regulatory elements in these loci, suggesting that instead dSSR-mediated transcriptional activity is regulated epigenetically. These observations fit well with existing data supporting a model for a transcriptionally-permissive "initiation zone" in which transcription initiates promiscuously (Figure 7-1).

Importantly, the possibility that the splice acceptors themselves represent a *cis*-acting signal to the cell and function independently of the sequence content of the dSSR remains in all of these experiments (Figure 7-2A). This will be tested by the use of synthetic splice acceptors as discussed above. I have attempted to generate a dSSR *de novo* by transplanting a cassette containing a selectable marker and its associated *trans*-splicing sequences into a PGC in the opposing (antisense) orientation (Figure 7-2B) to test this possibility. We anticipate that if this model is correct, generation of a *de novo* dSSR in the middle of a PGC will spur the acquisition of a transcriptionally-permissive epigenetic state, effectively establishing this locus as a hub of bidirectional transcription initiation. At this time these experiments are not complete, but if we are successful in generating lines bearing this cassette, comparisons of the epigenetic state of the selectable marker in the "sense" and "antisense" orientation of the polycistronic gene cluster will determine if divergently oriented splice acceptors are all that are needed to form a *cis*-acting signal for bidirectional transcription in *Leishmania*.

While our models have focused on the large proportion of the transcription initiation regions in *Leishmania* described by dSSRs, it is unclear how PGC-internal regions of transcription initiation, which have been identified in *L. major* and in *T. brucei*, fit into this model. Interestingly, these loci are marked by bidirectional transcription initiation events (7,8) in *T. brucei*, suggesting that these regions are actually bidirectional promoters. However, assessment of *trans*-splicing sites genome wide has not identified antisense *trans*-spliced mRNAs at these loci in *L. major* (Chapter 2; Beverley and Myler, unpublished data), and it is unclear whether these loci contain divergently oriented splice acceptors. However, it is possible that antisense, *trans*-spliced mRNAs accumulate at low levels, and focused, strand-specific assessment of *trans*-splicing sites at these loci might yield independent confirmation of this hypothesis.

Acquisition of a transcriptionally-permissive epigenetic state at dSSRs

If these additional experiments show convincingly that *cis*-regulatory elements are not required for the bidirectional transcription, it is likely that the epigenetic state of the dSSR is the key determinant of this activity. However, this lands us in the middle of a "chicken or egg" conundrum, as in most eukaryotes the acquisition of a transcriptionally-permissive epigenetic state is intimately linked with and dependent on the presence of *cis*-regulatory elements. While nucleosome-disfavoring sequences can facilitate chromatin opening and improve the accessibility of *cis*-regulatory elements to their *trans*-acting factors, these motifs are not present in many genes and instead typically mark constitutively-transcribed "housekeeping" genes [reviewed in (11)]. For most other genes, the required *cis*-regulatory elements are bound up in chromatin and are inaccessible to most transcription factors; special "pioneer" transcription factors possess the ability to interact with nucleosome-bound *cis*-regulatory elements, effectively opening chromatin and permitting access to other regulatory elements [reviewed in (12)]. In both circumstances, the acquisition of transcriptionally-permissive epigenetic state is not absolutely required for transcription [see (13) for a recent demonstration of this phenomenon], but plays a role in establishing a more permissive environment for efficient transcription initiation. In the absence of both of these classes of *cis*-regulatory elements, what processes aid dSSRs in their acquisition of the appropriate epigenetic state?

In this situation, it may be that stable transmission of epigenetic marks leads to the definition of dSSRs and other regions as transcriptionally-permissive environments. In this case, DNA modifications, which can be heritably transmitted during DNA replication and cell division, are an obvious candidate to function as the initiating signal for the establishment of the correct epigenetic state. Little evidence supports a role for DNA methylation in trypanosomatid protozoa (14–17), but a major role for the DNA modification β-D-glucosylhydroxymethyluracil (base J) in transcriptional biology has been shown in *Leishmania* and in *T. cruzi* using mutants of the thymidine hydroxylases JBP1 and JBP2, which catalyze the first step of J biosynthesis (18–20). Although this DNA modification is localized to dSSRs and cSSRs in these species as well as in *T. brucei* (18,20,21), the phenotypes of JBP mutants among these species differ significantly and were discussed in detail in Chapter 2. This divergence suggests that under this model, the nature of transcriptional control would differ significantly among these species despite significant conservation in other aspects of transcriptional biology. More importantly, this modification is localized to both cSSRs and dSSRs, suggesting that additional processes would be needed to distinguish sites of transcription termination from sites of transcription initiation. However, the demonstration that reintroduction of JBP2 into J-null *T. brucei* resulted in site-specific reacquisition of J (22) confers an element of specificity that is missing in the epigenetically-focused models of transcriptional regulation, and additional studies characterizing these proteins and the recently identified glucosyltransferase involved in J synthesis (23) will prove to be valuable in this context.

Epigenetic determinants of transcription initiation and termination in *Leishmania*

Independent of whether *cis*-regulatory elements reside within dSSRs, the characterization of the chromatin landscape and epigenetic marks influencing transcriptional processes in *Leishmania* is a valuable, significant achievement. Epigenetic modifiers have the capacity to significantly alter gene expression, and the nature of these processes have made them ideal targets for therapeutic intervention across a wide range of human diseases, especially with respect to cancer [reviewed in (24)]. In light of the results presented in this dissertation it is likely that epigenetic modifiers of transcription may play an even more important role in trypanosomatid protozoa, as the epigenetic state of a dSSR may be the sole determinant of transcriptional activity stemming from these hubs of transcription initiation.

In Chapter 2, we discussed our use of powerful genetic approaches to study the role of histone variants in *Leishmania* biology. We find that much like other eukaryotes, the conserved histone variant *H2A.Z,* as well as the trypanosomatid-specific histone variant *H2B.V,* is essential in *L. major*. While the functional characterization of these proteins *in vivo* are difficult due to the lack of a robust inducible system at this time, chromatin immunoprecipitation (ChIP) studies of these proteins localized them to dSSRs in *T. brucei*, indicating an essential role for these proteins in transcription initiation. We anticipate that further advancements in inducible gene expression, which were discussed in depth in Chapter 6, would allow additional characterization of the roles of these histone variants in the acquisition and maintenance of a transcriptionally-permissive epigenetic state. In contrast, we were able to readily generate null mutants of the *H3.V* histone variant, which bears no similarity to other eukaryotic H3 variants and localizes to regions of transcription termination in *T. brucei*. Preliminary data suggests that this histone variant is similarly localized in *L. major* (R. Sabatini, personal communication), suggesting an additional role for this protein in transcription termination. In contrast to the DNA modification base J, which is required for normal transcription termination in *L. tarentolae*, we showed in Chapter 2 that H3.V is dispensable for this process. Additional studies of this protein in the epigenetic networks of cSSR are underway in the laboratory of Dr. Robert Sabatini (University of Georgia), incorporating several of the mutants I generated, and we are interested in learning what the potential roles of this protein might be in localization of other epigenetic marks to cSSRs.

While these genetic studies are extremely valuable, they are also time-consuming and may not easily facilitate the characterization of protein functions if the gene is essential. To more rapidly gain insight into broad categories of histone modifications that may be important in the epigenetic networks defining dSSRs, we have selected a panel of small molecules known to target a broad panel of epigenetic modifiers, focusing on those targeting proteins or domains that are conserved in *Leishmania*. Utilizing our previously developed bidirectional reporter, I have screened these compounds, controlling for parasite numbers using a metabolic assay to specifically identify compounds altering reporter gene expression. Despite selecting a limited number of compounds, we identified several which are toxic to *Leishmania* promastigotes, and others which have a significant effect on reporter gene expression. At this time we are unsure of the targets of these compounds or whether they affect bidirectional transcription, but we plan to use FAIRE to quantify the effects of these compounds on the epigenetic state at dSSRs.

Concluding remarks

This work provides useful insight into the factors regulating gene expression in *Leishmania*, demonstrating an especially important role for epigenetic modifiers in this process. The information described here will be especially useful when applied toward the development of inducible gene expression systems, as it may facilitate the identification of a suitable, transcriptionally-silent locus for transgene integration. In addition, the vast majority of these studies were performed in the highly versatile promastigote stage of *L. major* parasites, but we believe that additional characterization of dSSR function in other stages of the life cycle may be useful, as significant alterations in transcriptional rates and chromatin organization within the nucleus have been observed throughout the parasite life cycle (Akopyants and Beverley, in preparation). The significant resources available to study amastigote processes in axenic culture would be a particularly interesting avenue to pursue in future experiments, as very little information has been gleaned regarding epigenetic modifications in this stage of the *Leishmania* development. Finally, the identification of three small molecule inhibitors that directly affect gene expression in *Leishmania* promastigotes was accomplished in a relatively limited screen of epigenetic modifiers. The integrated, bidirectional dual-luciferase reporter serves as an ideal platform for larger-scale screening efforts to identify other small molecules that alter gene expression; we believe that these efforts will allow the identification of promising compounds for therapeutic use, but will also aid in the efforts to unravel the processes within the epigenetic networks of dSSRs, which can be validated using additional genetic approaches.

Figure Legends

Figure 7-1. Model for the definition of the *de facto* promoter activity of divergent SSRs in *L. major*. In the top panel, genes are depicted as blue and red box arrows, and polycistronic transcripts are indicated as blue and red line arrows. The dSSR is indicated with a green box. In the lower panel, the chromatin state of the chromosome is depicted; purple circles indicate

“ground state” nucleosomes lacking epigenetic signatures of active transcription, and green circles indicate nucleosomes containing these marks. The box indicates the presumed “initiation zone” where promiscuous transcription initiates within a permissive epigenetic environment.

Figure 7-2. Model for *trans*-splicing acceptor sites as genetic *cis*-acting signals in defining dSSRs and polycistronic gene cluster (PGC)-internal transcription initiation regions. Genes, transcripts, and the dSSR are indicated as described in Figure 7-1. (A) Gene and *trans*-splicing acceptor site (AG) orientations for dSSRs and for PGC-internal transcription initiation regions. (B) Development of antisense and sense cassettes for targeting of a selectable marker to the middle of a PGC. The cassette in the antisense orientation will generate a *de novo* dSSR arrangement of genes and splice acceptors, and the sense orientation serves as a control for effects from modification of the locus.

References

1. Martinez-Calvillo S, Vizuet-de-Rueda J, Florencio-Martinez L, Manning-Cela R, Figueroa-Angelo E. Gene expression in trypanosomatid parasites. J Biomed Biotechnol 2010;2010.

2. Iyer V, Struhl K. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. EMBO J 1995;14(11):2570–9.

3. Raveh-Sadka T, Levo M, Shabi U, Shany B, Keren L, Lotan-Pompan M, et al. Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. Nat Genet 2012;44(7):743–50.

4. Ivens A, Peacock C, Worthey E, Murphy L, Aggarwal G, Berriman M, et al. The genome of the kinetoplastid parasite, *Leishmania major*. Science 2005;309(5733):436–42.

5. Cribb P, Esteban L, Trochine A, Girardini J, Serra E. *Trypanosoma cruzi* TBP shows preference for C/G-rich DNA sequences *in vitro*. Exp Parasitol 2010;124(3):346–9.

6. Thomas S, Green A, Sturm N, Campbell D, Myler P. Histone acetylations mark origins of polycistronic transcription in *Leishmania major*. BMC Genomics 2009;10:152.

7. Kolev N, Franklin J, Carmi S, Shi H, Michaeli S, Tschudi C. The transcriptome of the human pathogen *Trypanosoma brucei* at single-nucleotide resolution. PLoS Pathog 2010;6(9):e1001090.

8. Siegel TN, Hekstra DR, Wang X, Dewell S, Cross G a M. Genome-wide analysis of mRNA abundance in two life-cycle stages of *Trypanosoma brucei* and identification of splicing and polyadenylation sites. Nucleic Acids Res 2010;38(15):4946–57.

9. Martínez-Calvillo S, Yan S, Nguyen D, Fox M, Stuart K, Myler P. Transcription of *Leishmania major* Friedlin chromosome 1 initiates in both directions within a single region. Mol Cell 2003;11(5):1291–9.

10. Siegel T, Hekstra D, Kemp L, Figueiredo L, Lowell J, Fenyo D, et al. Four histone variants mark the boundaries of polycistronic transcription units in *Trypanosoma brucei*. Genes Dev 2009;23(9):1063–76.

11. Struhl K, Segal E. Determinants of nucleosome positioning. Nat Struct Mol Biol 2013;20(3):267–73.

12. Zaret KS, Carroll JS. Pioneer transcription factors: establishing competence for gene expression. Genes Dev 2011;25(21):2227–41.

13. Zhang H, Gao L, Anandhakumar J, Gross DS. Uncoupling transcription from covalent histone modification. PLoS Genet 2014;10(4):e1004202.

14. Rojas M, Galanti N. DNA methylation in *Trypanosoma cruzi*. FEBS Lett 1990;263(1):113–6.

15. Rojas M, Galanti N. Relationship between DNA methylation and cell proliferation in *Trypanosoma cruzi*. FEBS Lett 1991;295(1-3):31–4.

16. Militello K, Wang P, Jayakar S, Pietrasik R, Dupont C, Dodd K, et al. African trypanosomes contain 5-methylcytosine in nuclear DNA. Eukaryot Cell 2008;7(11):2012–6.

17. Huff J, Zilberman D. Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes. Cell 2014;156(6):1286–97.

18. Ekanayake D, Minning T, Weatherly B, Gunasekera K, Nilsson D, Tarleton R, et al. Epigenetic regulation of transcription and virulence in *Trypanosoma cruzi* by O-linked thymine glucosylation of DNA. Mol Cell Biol 2011;31(8):1690–700.

19. Ekanayake D, Sabatini R. Epigenetic regulation of polymerase II transcription initiation in *Trypanosoma cruzi*: modulation of nucleosome abundance, histone modification, and polymerase occupancy by O-linked thymine DNA glucosylation. Eukaryot Cell 2011;10(11):1465–72.

20. Van Luenen H, Farris C, Jan S, Genest P, Tripathi P, Velds A, et al. Glucosylated hydroxymethyluracil, DNA base J, prevents transcriptional readthrough in *Leishmania*. Cell 2012;150(5):909–21.

21. Cliffe L, Kieft R, Southern T, Birkeland S, Marshall M, Sweeney K, et al. JBP1 and JBP2 are two distinct thymidine hydroxylases involved in J biosynthesis in genomic DNA of African trypanosomes. Nucleic Acids Res 2009;37(5):1452–62.

22. DiPaolo C, Kieft R, Cross M, Sabatini R. Regulation of trypanosome DNA glycosylation by a SWI2/SNF2-like protein. Mol Cell 2005 Feb 4;17(3):441–51.

23. Bullard W, Lopes da Rosa-Spiegler J, Liu S, Wang Y, Sabatini R. Identification of the glucosyltransferase that converts hydroxymethyluracil to base J in the trypanosomatid genome. J Biol Chem. 2014 (In press).

Figure 7-1. Model for the definition of the *de facto* promoter activity of divergent SSRs in *L. major*.

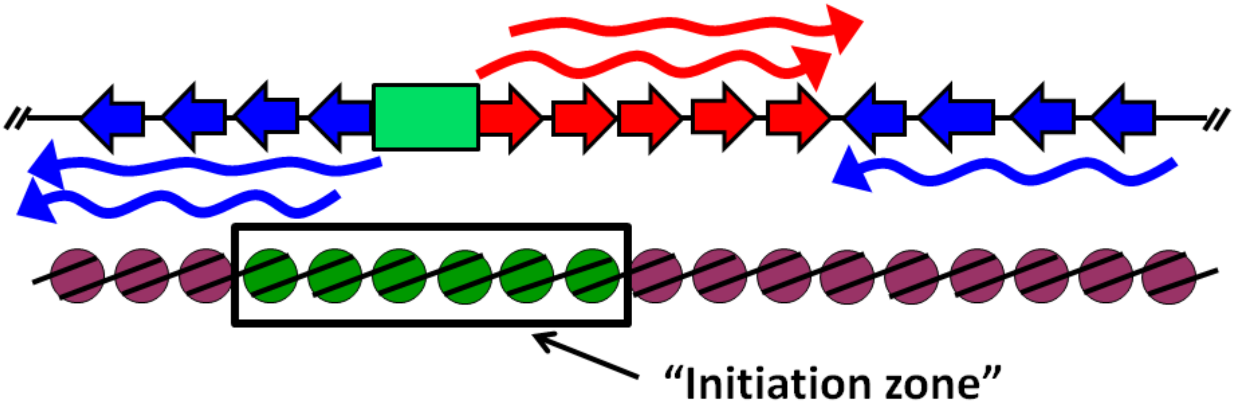
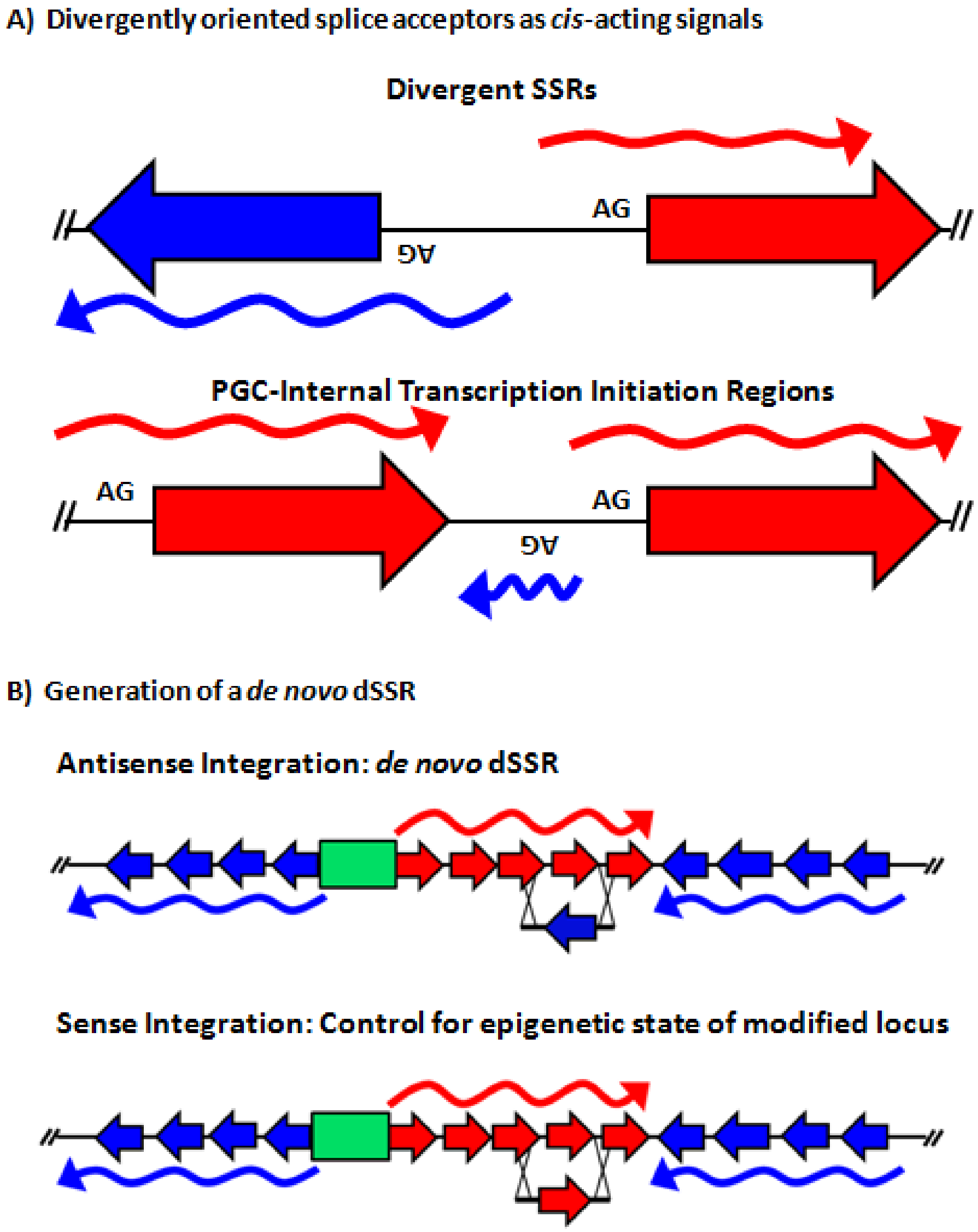


Figure 7-2. Model for *trans*-splicing acceptor sites as genetic *cis*-acting signals in defining dSSRs and polycistronic gene cluster (PGC)-internal transcription initiation regions.


A) Divergently oriented splice acceptors as *cis*-acting signals
Divergent SSRs
AG
AG
PGC-Internal Transcription Initiation Regions
AG
AG
AG
B) Generation of a *de novo* dSSR
Antisense Integration: *de novo* dSSR
Sense Integration: Control for epigenetic state of modified locus