Characterization of length-dependent GGAA-microsatellites in EWS/FLI mediated Ewing sarcoma oncogenesis

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Kirsten M. Johnson

Graduate Program in Biomedical Sciences

The Ohio State University

2018

Dissertation Committee:

Stephen L. Lessnick, MD, PhD, Advisor

Michael A. Freitas, PhD

Denis C. Guttridge, PhD

Charles E. Bell, PhD

Abstract

Objective: Ewing Sarcoma is a pediatric bone malignancy initiated by a t(11;22) chromosomal translocation that produces the EWS/FLI oncoprotein. EWS/FLI transcriptionally activates and represses its target genes to mediate oncogenic reprogramming. Expression of its up-regulated targets correlates with EWS/FLI binding to associated GGAA-microsatellites, which show length polymorphisms. These microsatellite polymorphisms may critically affect EWS/FLI-responsiveness of key gene targets. For example, *NR0B1* is necessary for EWS/FLI mediated oncogenic transformation, and we found a "sweet-spot" of 20-26 repeat length as optimal for EWS/FLI mediated transcriptional activity at *NR0B1* through clinical observations and *in vitro* studies. The mechanism underlying this optimal length, however, is unknown.

Methods: We explored the stoichiometry and binding affinity of EWS/FLI for different GGAA-repeat lengths through biochemical studies, including fluorescence polarization, ChIP-seq, and RNA-seq, combined with bioinformatics analysis. Additionally, use of EWS/FLI deletion constructs has been critical for elucidating the particular binding behavior of EWS/FLI at different microsatellite repeat lengths. Luciferase reporter assays, anchorage-independent growth and proliferation assays, as well as CRISPR technology have extended our findings to the *in vivo* setting. Finally, microscopy studies including use of confocal and transmission electron microscopy (TEM) have contributed visual characterization of the specific biochemical mechanisms we are investigating.

Results: CRISPR-mediated deletion of the *NR0B1* GGAA-microsatellite in Ewing sarcoma cells provided our field with the first *in vivo* evidence for the necessity of EWS/FLI binding at GGAA-microsatellites for anchorage dependent growth. Our biochemical studies, using recombinant Δ22 (a version of EWS/FLI containing only the FLI portion) demonstrate a stoichiometry of one monomer binding every two consecutive GGAA-repeats on shorter microsatellite sequences. Surprisingly, the affinity for Δ22 binding to GGAA-microsatellites significantly decreased, and was unmeasureable when the size of the microsatellite was increased to the "sweet-spot" length. In contrast, a fully-functional EWS/FLI mutant (Mut9, retaining approximately half of the EWS portion) showed low affinity for smaller GGAA-microsatellites, but instead significantly increased its affinity at "sweet-spot" microsatellite lengths. Single-gene ChIP and genome-wide ChIP-seq and RNA-seq studies extended these findings to the *in vivo* setting. Additionally, through bioinformatics analysis, we defined GGAA-microsatellites in a Ewing sarcoma setting, and showed GGAA-microsatellite length is predictive of EWS/FLI responsiveness (binding and transcriptional activation) at "promoter-like" EWS/FLI targets.

Conclusion: Together, these data demonstrate the necessity for EWS/FLI binding at GGAA-microsatellites in Ewing sarcoma, and characterize their role in oncogenesis. These data also reveal an unexpected novel role for the EWS portion of the EWS/FLI fusion in DNA-binding. Overall, our results suggest a length-dependent biochemical mechanism for EWS/FLI binding and transcriptional regulation at GGAA-microsatellites.

Dedication

This work is dedicated to my family (Paul, Marianne, Quinn, Hayden, and Lauren Johnson), Randy Stokes, and my best friend Katelyn Sheffield, who encouraged me, put up with me, and served as my rocks throughout this process. I could never have done it without them.

Acknowledgments

I would like to especially acknowledge and thank my principle investigator and mentor, Dr. Stephen Lessnick, who accepted me into his lab and transformed me into a researcher. He never stopped believing in me, even when I gave up on myself. He is a truly talented mentor with a gift for seeing what is really important in research and in life—the reason I followed him and his lab to transfer from the University of Utah to The Ohio State University and Nationwide Children's in Columbus, Ohio.

I would like to thank the many people who supported this research with their time, talents, expertise, and advice. Thank you to Kathleen Pishas and Ryan Roberts for all of the support and advice. Special thanks to Nathan Callender, Jesse Crow, and Cenny Taslim for all of their hard work and for making my last year in lab memorable. I would also like to thank Dr. Charles Bell and Cynthia McCallister for their expertise and research support and help in this work.

The National Cancer Institute (Grant F30 CA210588) provided financial support for my tuition fees, living expenses, research, and travel costs for presenting my research at conferences around the globe. I sincerely appreciate the support and invaluable opportunities this grant award has provided me in my training to become a physician scientist.

I would like to thank the MSTP and Biomedical Sciences Graduate Programs at The Ohio State University for accepting my transfer into their exceptional programs and the help and support they have provided along the way.

Vita

December 2011 .....B.S. Molecular Biology, Brigham Young University

2012 to present .....MSTP Trainee, The Ohio State University

Publications

Monument, Michael J., **Johnson, Kirsten M.,** Grossman, Allie H., Schiffman, Joshua D., Randall, R. Lor, and Lessnick, Stephen L. Microsatellites with Macro-Influence in Ewing Sarcoma. *Genes* 2012, *3*, 444-460.

Monument, M., **Johnson, K.,** McIlvaine, E., Abegglen, L., Watkins, W., & Jorde, L. et al. (2014). Clinical and Biochemical Function of Polymorphic NR0B1 GGAA-Microsatellites in Ewing Sarcoma: A Report from the Children's Oncology Group. *Plos ONE*, *9*(8), e104378. doi:10.1371/journal.pone.0104378

Johnson KM, Mahler NR, Saund RS, Theisen ER, Taslim C, Callender NW, Crow JC, Miller KR, Lessnick SL. Role for the EWS domain of EWS/FLI in binding GGAA-microsatellites required for Ewing sarcoma anchorage independent growth. Proc Natl Acad Sci U S A. 2017 Aug 28. pii: 201701872. doi: 10.1073/pnas.1701872114. PubMed PMID: 28847958

Johnson KM* and Taslim C* (*co-first authors), Saund RS, Lessnick SL. Identification of two types of GGAA-microsatellites and their roles in EWS/FLI binding and gene regulation in Ewing sarcoma. PLOSOne U S A. 2017. doi: 10.1371/journal.pone.0186275

Fields of Study

Major Field: Biomedical Sciences

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Ewing sarcoma

Ewing sarcoma is a bone malignancy primarily diagnosed in children and young adults that has seen little improvement in overall survival rates since the introduction of multi-agent chemotherapy 40 years ago. Although 5-year survival rates plateau at about 60-70% following surgery and chemotherapy, survival in patients with relapsed or metastatic disease plummets to 15-30%[1]. Ewing sarcoma is a rare disease, characterized histologically by small, round blue cells[2]. It is solely initiated by EWS/ETS translocations, 85% of which are EWS/FLI, resulting from a t(11;22)(q24;q12) chromosomal translocation that fuses the EWS gene on chromosome 22 to the FLI gene on chromosome 11[3–5] (Figure 1.1). The resulting fusion oncoprotein acts as an aberrant master transcription factor, initiating and maintaining oncogenic reprogramming via direct and indirect regulation of thousands of target genes[6–8].

Globally, sarcomas are rare, accounting for approximately 0.5% of all human malignancies. Proportionately, however, bone cancer is significantly more common in children than adults, with Ewing sarcoma as the second most common, after osteosarcoma. As of 1995, about 200 Ewing sarcomas (~34%) are diagnosed each year, with an incidence rate in the United States of one case per million[9]. The median age of

diagnosis is reported to be 13.7 years, with an average 5-year overall survival of approximately 63.5%[10] (Table 1.1). Though the long bones of the lower limb are the most frequent site of sarcoma development (~57%), Ewing sarcoma demonstrates a different tumor site proclivity. Ewing sarcoma frequently emerges in the central axis (45%), though can be found in any part of the body, with the pelvis, followed by the femur, tibia, humerus, and scapula as the most common tumor locations[9,11].

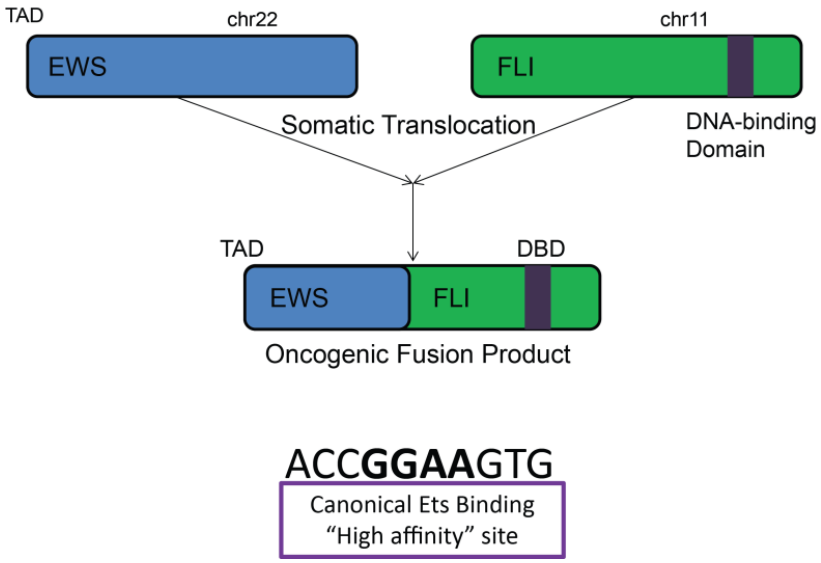


Figure 1.1 Chromosomal translocation creating the EWS/FLI fusion oncoprotein. The canonical Ets protein DNA binding site is also shown.

Clinically, Ewing sarcomas include Ewing and atypical Ewing's, which all bear the t(11;22) translocation[9]. Although extraosseous Ewing sarcoma and PNET tumors fall into the soft tissue sarcoma category, they still display the same pathognomonic translocation and subsequent fusion protein[9]. There is ongoing controversy whether Ewing sarcoma

cells are of mesenchymal or neuroectodermal (neural crest) origin[9]. However, CD99, an antigen normally expressed on the surface of human mesenchymal stem cells, is the most commonly used histological diagnostic marker for Ewing sarcoma[12].

# New Cases/Year	Gender ratio (M:F)	Peak Age	Incidence in Caucasians	Incidence in African Americans
200	1.2 : 1	10-15	0.155	0.017

Table 1.1 Epidemiology of Ewing sarcoma

Epidemiologically, there is a slightly higher incidence of Ewing sarcoma in males than females, with a ratio of 1.5:1[13]. Of note, Ewing sarcoma is ten times more common in Caucasian than African populations (Table 1.1)[14]. Intron 6 is at least 50% smaller (less Alu elements) in approximately ten percent of Africans. It has been proposed that this racial discrepancy could be attributed to differences in the number of Alu elements (short interspersed elements that are transposons comprising about ten percent of the genome), which seem to be preferential sites for cancer-associated genetic recombination[11]. These observed epidemiological differences are especially pertinent to this thesis, as we found EWS/FLI binds to polymorphic GGAA-microsatellite sequences that display distinct length heterogeneity between African and Caucasian individuals[15,16].

In the overall population, Ewing sarcoma has an incidence of three per million individuals under the age of 20 and is deadly when untreated[11]. Overall positive

prognostic factors for Ewing sarcoma include tumor size (for localized disease), absence of metastasis, and histological response to induction chemotherapy, regardless of tissue grade or tumor size[17]. Extraskeletal Ewing sarcoma is an additional favorable prognostic factor[18]. A clinical trial at St. Jude Children's Hospital risk stratified Ewing sarcoma patients based on patient outcome into the following four groups: 1) favorable: age <14, localized, non-pelvic tumor; 2) intermediate risk: localized, >/= 14yo, or pelvic tumors; 3) unfavorable-pulmonary: isolated lung metastasis; and 4) unfavorable-extrapulmonary: extrapulmonary metastases. Five-year overall survival (OS) for these groups was 88%, 65%, 54% and 27%, respectively[10]. Though the field has seen improvement in treatment of patients with localized disease, clearly little, if any, progress has been made for Ewing sarcoma patients with metastasis.

The current chemotherapy induction standard for Ewing sarcoma treatment is VIDE (vincristine, ifosfamide, doxorubicin, and etoposide) in Europe and VDC-IE (vincristine, doxorubicin, cyclophosphamide; ifosfamide + etoposide) in North America (Table 1.2). Marked improvement in survival was noted after the addition of ifosfamide and etoposide to the standard chemotherapy regimen, however, only in patients without metastasis at presentation[19–22].

Induction therapy is followed by surgical resection when possible, which has replaced single modality radiotherapy as the best method for localized tumor control[19,23]. Definitive radiotherapy is only used for inoperable lesions. Surgery or radiotherapy

without induction chemotherapy frequently results in metastasis[17]. Additionally, a small treatment study recently showed that patients treated primarily non-surgically showed overall neurological improvement, and that initial chemotherapy induction was vital for this overall outcome[24]. Numerous clinical trials have also evaluated intensified chemotherapy regimens and found these only demonstrate survival improvement in patients with localized disease[25–29].

Diagnosis	Biopsy
Assessment	Staging, MRI, CT chest, Bone scan
Initial Treatment	Neoadjuvant chemo: 6x cycles VIDE
Surgical Intervention	surgery and/or radiotherapy
Reassessment:	Depends on response & disease volume
Good	VAI x1, then VAC vs. VAI x7
Poor	VAI x1, then VAI x7 vs. high dose
Extensive Disease	High dose or phase II clinical trial

Table 1.2 Treatment of Ewing sarcoma

One of the unique molecular features of this pediatric malignancy is its relatively silent genomic background, making it a useful model for elucidating key aspects of pediatric oncogenesis without confounding factors, such as complex networks of signal transduction pathways seen in many other cancer types. Large-scale genomic sequencing efforts have demonstrated Ewing sarcoma possesses one of the lowest mutation rates amongst all cancers (0.15 mutations/Mb)[30]. Recurrent, though low frequency, mutations were consistently observed only in the cohesion complex subunit *STAG2* (21.5%), the tumor suppressor *TP53* (6.2%), and homozygous deletion of the cyclin-dependent kinase

inhibitor *CDKN2A* (13.8%)[31]. The EWS/FLI translocation appears to be a somatic event resulting in an oncogenic driver mutation, though recent data suggests the existence of germline susceptibility variants, such as GGAA-microsatellites associated with particular EWS/FLI targets such as *EGR2*[32]. It is difficult to study the origin of this disease, as neither mouse nor other animals acquire Ewing sarcoma spontaneously or even through genetic induction[33]. As a result, no model outside of cell lines exists to accurately recapitulate this disease.

Biological studies of this disease have led to the development of many new treatment approaches, however, integration of these for patients has been, and will continue to be difficult, due to the low volume of patients affected by this disease. To this point, most researchers in the field have studied Ewing sarcoma in the context of patient-derived cell lines, with a focus on downstream target genes of EWS/FLI, their impact on oncogenesis, and their overall therapeutic potential. Because Ewing sarcoma is a uniquely human disease not effectively recapitulated in animal models, knock-down and rescue experiments, combined with microarray, ChIP and RNA sequencing studies have been used extensively throughout the field to understand the mechanisms of EWS/FLI regulation.

EWS/FLI is considered undruggable due to its intrinsically disordered region within the EWS domain and lack of intrinsic enzymatic activity, which precludes small molecule targeting. More recently, numerous leaders in the field have turned their attention to

understanding the role of epigenetic regulation, seeking drug targets within these interactions. So far, EWS/FLI has proven a poor target for drug development, despite a few recent novel therapies focused on targeting of RANKL, IGFR-1, PARP1, VEGF, and epigenetic targets[17,34].

Although the EWS/FLI translocation is well-known to be the driver for Ewing sarcoma initiation and progression, few have studied the biochemical properties of EWS/FLI itself. A paucity of understanding regarding the mechanisms of EWS/FLI molecular biology and function constitute a potent barrier to therapeutic amelioration of this disease. The proposed work in this thesis is significant because it provides insight into the biochemical mechanism of EWS/FLI-mediated Ewing sarcoma development; knowledge which is crucial to discovery of therapeutically targetable transcriptional processes. This work is also anticipated to have a positive impact on our understanding of how EWS/FLI determines whether to transcriptionally activate or repress its target genes.

Additionally, because Ewing sarcoma is constituted by a relatively clean genetic background seen rarely in other cancer types, it provides an excellent simplified model for study of pediatric cancer development in general. The studies contained herein both explore and elucidate mechanisms of transcriptional regulation in an oncogenic setting. This work is also expected to alter understanding of the molecular biology underlying development and progression of Ewing sarcoma and broader aberrancies alike.

EWS/FLI Transcriptional Regulation

As mentioned previously, Ewing sarcoma is characterized by a chromosomal translocation t(11;22), resulting in the fusion oncoprotein EWS/FLI[3,8]. As an oncogenic driver, this aberrant transcription factor is absolutely essential for oncogenic transformation of Ewing cells[7]. The EWS/FLI fusion of Ewing sarcoma is created between exon 7 of EWS and exon 6 of FLI (60%) or exon 5 (20%), giving rise to type 1 and type 2 respectively[23]. The other two types of fusions (4%) consist of exon 10 of EWS fused to either exon 6 or 5, respectively, of FLI[35]. See Table 1.3 for the details of these fusions.

Fusion Type	EWS Exon	FLI Exon
Type I	7	6
Type II	7	5
Type III	10	6
Type IV	10	5

Table 1.3 Types of EWS/FLI fusions

Regardless of break-point location, EWS acts as an amino-terminal transcriptional activation domain linked to the carboxy-terminal DNA-binding domain of its ETS family member partner in crime[6,36]. The ETS portion of the fusion binds DNA near target genes it regulates[37]. Recent data has shown that although the EWS/ETS chimera binds at specific DNA sites, the protein's physical interaction with other transcription factors is both necessary and sufficient for effective oncogenic transformation[38]. For example,

activation of specific transcription factors, such as *NR0B1, NKX2.2, CAV1,* and *GSTM4,* results in regulation of their respective target genes, giving rise to aberrant expression of both direct and indirect EWS/ETS targets[39,40].

This EWS/FLI-mediated regulation of key target gene intermediates like *NKX2.2* helps to explain how three times more of its target genes are down-regulated than up-regulated[41,42]. Both the DNA binding and repressor domains of NKX2.2 are necessary for oncogenic transformation[41]. *NR0B1* (*DAX1*) has similarly been shown to be essential for oncogenic transformation of Ewing cells[43]. A member of the nuclear hormone receptor family, *NR0B1* is a critical EWS/FLI target, which allows mechanistic exploration of EWS/FLI transcriptional activation. Its importance in Ewing sarcoma was originally found through GSEA data showing correlation of *NR0B1* with EWS/FLI gene expression data across data sets of multiple Ewing sarcoma cell lines[43]. There is some experimental evidence to suggest the NR0B1 protein may physically interact with EWS/FLI, as ChIP-ChIP evaluation has shown the two occupy the same regions of genomic DNA at a subset of loci[39]. Though both have been shown necessary for oncogenic transformation of Ewing sarcoma cells *in vitro,* the combination of NR0B1 and NKX2.2 introduction into EWS/FLI deficient cells is unable to rescue the Ewing sarcoma phenotype[43].

Another example of EWS/FLI-mediated activation of key targets that further regulate downstream genes is MMP3. A member of the metalloprotease (MP) family, MMP3 functions in digestion of ECM (extracellular matrix) proteins to promote tumor invasion

and metastasis[38]. It is evident from these examples that rather than being implicated in transcriptional regulation of all its targets, EWS/FLI directly interacts with regulatory sequences of a smaller number of target genes, who in turn regulate the activities of additional targets[38].

Prior to the work outlined in this thesis, it was widely accepted that the FLI portion of the fusion allows EWS/FLI interaction with the DNA, while the EWS portion is necessary for transcriptional regulatory function. The EWS portion possesses both a transcriptional activation and an RNA binding domain. Though EWS/FLI deletion mutants demonstrate both of these domains are necessary for transformation, the exact function of the EWS protein is incompletely understood[44]. Its role as a transcriptional activator was originally suspected due to its high glutamine and proline content, often a hallmark for activation domains of transcription factors[3]. Moreover, EWS/FLI localizes to the nucleus, binding DNA in a sequence-specific manner. It has since been confirmed that the EWS portion functions as a potent transcriptional activator[8].

In contrast, there is also evidence for direct EWS/FLI-mediated repression. ChIP studies show that a subset of genes, such as *LOX* and *TGFBR2* are directly down-regulated EWS/FLI targets[34]. Forced expression of each of these genes in cell lines and *in vivo* xenograft models impairs tumor formation. Little, however, is known regarding how EWS/FLI distinguishes which transcriptional change to induce within any particular gene to either up or down regulate its expression.

Recent work by our laboratory highlighted the complexity surrounding EWS/FLI-mediated repression. Full repression by EWS/FLI necessitates HDAC activity, and in concert, EWS/FLI has been demonstrated to bind, or at least associate with, the NuRD complex[34]. Additionally, in recent years our laboratory aided in the development of a small molecule reversible inhibitor of the epigenetic modulator LSD1, which is an integral part of the NuRD complex. Pharmacological or genetic inhibition of this interesting new target interferes with EWS/FLI-mediated transcriptional activity in Ewing sarcoma cells[45]. Though this has significantly improved our understanding of the repressive function of EWS/FLI, specific binding sites on the DNA near repressed genes have yet to be characterized.

Prior to the studies described in this work, all that was known about EWS/FLI-mediated activation was data suggesting EWS/FLI binds to repetitive GGAA-motifs in noncoding DNA near genes it up-regulates[15,46]. Unlike EWS/FLI repressive function for which the above-mentioned associated repressive complex was delineated, associative activating protein complexes were not identified until very recently[47]. EWS/FLI is clearly critical for a complex network of functions requisite to engender oncogenic transformation of Ewing sarcoma cells.

ETS factors & FLI binding

ETS gene family members, like FLI-1, contain a conserved domain of 85 amino acids that is responsible for sequence specific DNA binding activity[37]. These proteins tend to bind a single core (GGAA) motif within the conserved high affinity 'ACC**GGAA**GTG' sequence by a monomeric DNA binding domain, which contains a winged helix-turn-helix component (Figure 1.2). Helix H3 serves as a DNA recognition helix, inserting into the major groove and allowing for specific base interactions via three highly conserved residues[48]. Overall, global genomic binding studies have demonstrated both highly redundant and specific binding across many ETS family members[49]. As many as five to fifteen percent of all 17,000 human promoters are redundantly occupied by ETS proteins[50].

In addition to their importance in development and a vast array of cellular functions, several ETS proteins regulate cancer development and tumorigenesis[50]. ETS family members phylogenetically related to FLI are capable of binding the aforementioned GGAA-microsatellites (tested up to 7 repeats) (Figure 1.2); however, transcriptional activation at these sites is an emergent property of EWS/ETS fusions[51]. Other ETS family members found fused to EWS in Ewing sarcoma are Erg (5-10%), ETV4, ETV1, and FEV1 (collectively less than 5% of cases)[33]. Structural studies analyzing ETS proteins cluster them into 4 different classes based on their ETS-binding profile[52]. Because class I ETS DNA binding domains (DBDs) are the only ones found in fusion onco-proteins, there is believed to be some degree of specificity in biological function distinguishing these classes. The binding specificities of these different proteins have since been

confirmed *in vivo* by ChIP-seq, demonstrating that small differences in transcription factor (TF) binding specificity contribute significantly to site selectivity[53,54].

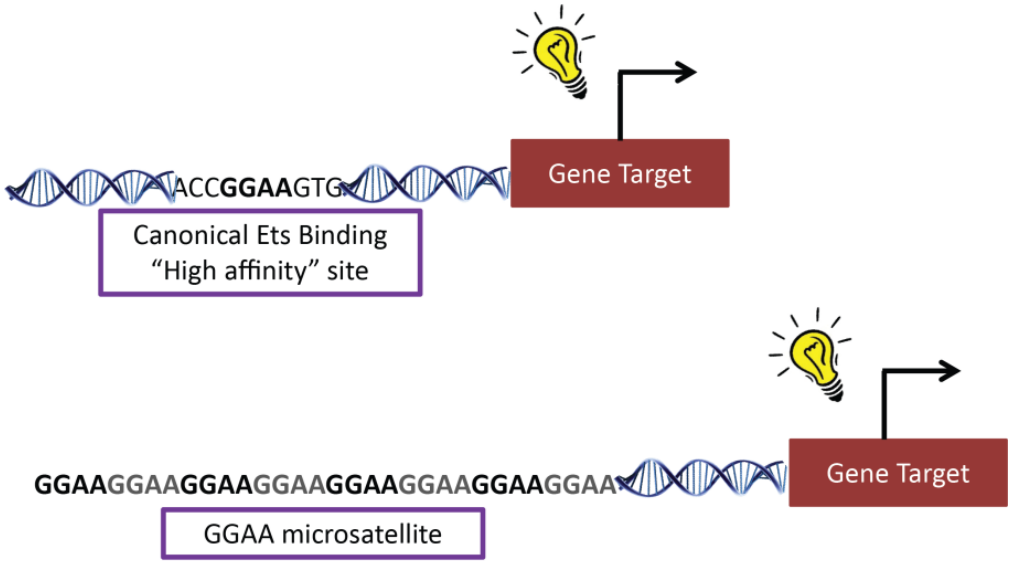


Figure 1.2 Two distinct EWS/FLI DNA binding sites

While ETS proteins bind as monomers to the conserved "high affinity" DNA sequence, they are also capable of both homo and hetero-dimeric interactions with other transcription factors[55]. For example, ETS1 has been shown in some situations to bind as a homodimer, while in others it can bind to similar sequences as a heterodimer with PAX5, FOXO1, or AML1 respectively[56]. Interestingly, 76% of human core promoters lack TATA-like elements, suggesting the possibility that repetitive elements like GGAA-microsatellites might serve as a scaffold for general transcription factors near the transcription start site (TSS) to make up for the lack of a TATA box at many gene targets[57]. In such a model, this repetitive GGAA scaffold would facilitate, for example,

the recruitment of the pre-initiation complex near the TSS. Taken together, it is possible that ETS functional specificity is determined by distinct interactions with regulatory proteins around a given binding site, as opposed to the actual direct sequence interaction of the ETS protein and DNA.

Homodimerization is also a classic feature of ETS-mediated transcriptional regulation[48,56,58,59]. At the homodimerization interface of FLI, for example, there is a high content of hydrophobic interactions and suspected helix swapping[59]. This dimerizing property of the FLI DNA-binding domain is thought to play a role in both transcriptional activation and repression.

At least 85% of EWS/ETS fusions found in Ewing sarcoma contain FLI[36,60,61]. This ETS family member is normally involved in hematopoietic development and is required for embryonic angiogenesis[51]. FLI knock-out is embryonically lethal and results in defects in megakaryopoiesis, with decreased numbers of hematopoietic colony forming units CFU-E and CFU-GM[62]. The DNA-binding part of FLI within the EWS/FLI fusion protein is not mutated, but conserved from wild-type FLI. Further, this conserved DNA binding domain is necessary and sufficient for EWS/FLI directed transformation and consequently transcriptional activation of Ewing's cells. Despite this conserved binding region, EWS/FLI has 10-fold higher transcriptional activity than wild-type FLI when compared in Gal4 reporter assays[63].

The FLI subfamily of ETS proteins is comprised of FLI, ERG, and FEV. The only differences between ERG and FLI are minor sequences at the N-terminus. This includes two amino acid residues, Ala295 and Ala297, which are serine residues in ERG[59]. FEV, conversely, differs from ERG and FLI in that it lacks a PNT (Pointed) domain[49]. Though not identical to FLI, the high degree of evolutionary conservation among many ETS family members suggest structural studies on one particular sub-type can be applied to conformational, stoichiometric, and potentially even functional properties of another.

ERG makes up about 5% of EWS/ETS fusions in Ewing sarcoma tumors[36,64]. The highest expressed ETS transcription factor in mature, quiescent endothelial cells, ERG is required for vascular development, angiogenesis, and vascular homeostasis. The wild type version contains both a PNT domain and the ETS domain, similar to FLI, which binds to the cognate sequence GGA(A/T)[65]. Both ERG and ETS-1 bind this conserved core sequence via their helix H3 at the major groove of the DNA[48]. The 'GGAA' core appears to contribute both sequence recognition and stability to the interactions between helix H3 and the ETS domain[66].

Specifically, crystallographic studies of ERG-DNA binding demonstrate direct DNA contact at 'GG' with Arg367 and Arg370, respectively, as well as the conserved Tyr371 contacting the first 'A' in the core recognition sequence[65]. These three residues are the only identified points of direct protein-DNA contact, implicating sequence-dependent flanking regions for conferring specificity to differentiate ETS family member binding[49].

In line with this conjecture, a number of hypothesized salt bridges in the region flanking the consensus sequence have also been confirmed via molecular dynamics simulations of ERG-DNA interactions[65]. For example, a salt bridge at Arg385 was reported as the most stable interaction for the ERG-DNA binding model system. Additional insight from molecular modeling suggests that the Arg367 and Arg370 residues serve as points of direct contact and facilitate order/disorder transitions upon DNA binding. In contrast, the Tyr371 residue is thought to be critical for both ETS factor recognition and auto-inhibitory regulation of ERG-DNA binding[65].

Interestingly, auto-inhibition is also a significant component of several ETS protein members[49]. Autoregulation suppresses protein function through two N-terminal ETS flanking regions in the absence of bound DNA. This inhibitory interaction has been shown in some instances to be interrupted by interaction with other, homo or heterologous transcription factors. For example, activation of *MMP3* at its promoter (which contains a palindromic EBS site) requires cooperative homodimeric binding[48]. Auto-inhibition has been perhaps most rigorously studied in ETS-1, where helix H1 is unfolding upon DNA binding, suggesting reduced binding affinity compensated for by an increase in thermodynamic instability. Other studies have shown that specific DNA sequences can modulate energy cost in twisting flexibility and bending[67]. The current model for this idea, conceptually, is that ETS-1 occurs in an equilibrium conformation between a rigid, inactive state, and a flexible DNA-binding competent state[49].

Wild-type FLI contains an autoinhibitory domain at its N-terminus[68]. Though both EWS/FLI and FLI associate with SRF(serum responsive factor) *in vitro*, only EWS/FLI binds to the ets-box of c-fos SRE automatically and without the presence of SRF to induce expression of transformation-associated genes including MMP3, cytokeratin 15, and CYP4F1. Deletion of the amino terminus of wild-type FLI, however, gives it the same autonomous binding capability of EWS/FLI in this process, suggesting the inhibitory domain located in the N-terminal part of FLI is absent or at least modified through EWS fusion in Ewing sarcoma oncogenesis[68]. Though this suggests intrinsic differences between EWS/FLI and FLI binding to specific ETS DNA sequences, both are capable of binding GGAA-microsatellites, as evident through experiments which demonstrated binding to 4-7 GGAA-repeats[51]. Interestingly, additional ETS proteins are also able to bind these sequences, however, lack the ability to transcriptionally induce reporter systems unless fused with EWS[51].

The mechanism by which proteins, especially transcription factors, precisely identify their functional binding sites remains unknown and is an area of active controversy. Current evidence suggests that DNA binding by transcription factors (TF) is dependent on a combination of TF 3D structure and flexibility, cofactor binding, cooperative TF-DNA binding, ability to access chromatin and nucleosomes, and DNA methylation[53]. Crystal structures of protein-DNA complexes have shown that preferential binding at a specific site seems to be established by physical interactions between the TF factor amino acid side chains and the DNA's accessibility[53].

DNAse I footprint experiments with EWS/FLI demonstrate a reproducible 14bp region of DNA protected from digestion, with a hypersensitivity site on the negative strand adjacent to the GGAA core for FLI binding to the conserved "high-affinity" site ETS sequence[51]. Increasing numbers of GGAA motifs, however, as with GGAA-microsatellites, show an altered pattern in these same footprint experiments. Four repeats, minimal for EWS/FLI homodimeric binding, contain a 28bp-protected region. Beyond this, an additional four bases are protected for each additional GGAA motif included[51]. The reason for EWS/FLI binding to GGAA-microsatellites rather than high affinity site regions remains uncharacterized. However, given what is known about the stoichiometry of EWS/FLI binding, early biochemical studies suggested the possibility of EWS/FLI flexibly binding in a "sliding fashion" to these repetitive GGAA sequences[15,51].

Many EWS/FLI transcriptionally activated targets are associated with GGAA-microsatellites within 5kb of the genes' transcriptional start site (TSS)[15]. However, recent studies have shown there are a number of microsatellites identified greater than 15kb from any EWS/FLI target suspected to facilitate EWS/FLI-mediated regulation[42,69,70]. An interesting model for this alternate regulatory mechanism is the idea that a ligand bound at one sequence can influence DNA structure at some distance from the binding site. A distal response is then modulated by the sequence's acceptance or resistance to conformational change[71]. This was originally described as the telestability hypothesis, but more recently emerging data suggest a "super-enhancer" model[69,72]. For example, the

nearest GGAA-microsatellite to EWS/FLI direct target *NKX2.2* is about 65kb away from its transcriptional start site, yet this particular sequence is strongly implicated in EWS/FLI–mediated activation of the gene[69]. Evidence for distal DNA binding implicates a model for transcription factor regulation via distinct mechanisms by which EWS/FLI regulates near versus distal gene targets.

Though structural and molecular modeling binding studies have been performed on a number of ETS proteins, including FLI and ERG, any DNA-binding structural characterization previously performed has been limited to the Ets-consensus sequence. Before this body of work, the preference for EWS/FLI binding to GGAA-microsatellites rather than high affinity regions remained uncharacterized.

It was, however, known that wild-type FLI binds the Ets consensus sequence with higher affinity than EWS/FLI[37] (Figure 1.2). Interestingly, proteins exhibiting the highest affinities for cognate sites are not necessarily the most specific. For example, phage λ-Cro protein binding O_R3 demonstrates slightly higher affinity but less specificity than for its binding of the lac repressor site[73]. Thus, there may be a specificity/affinity trade-off in protein-DNA binding. Additionally, charge and structure may also affect binding affinity while altering specificity.

Comparison of FLI with other ETS factors shows differential patterns of hydrogen bonds within two aforementioned conserved arginine residues of helix H3, causing a variation

in DNA bending from 11 to 28 degrees[74]. This suggests a possible "indirect readout" mechanism of protein-DNA recognition, where the ETS domain recognizes a sequence dependent structure, rather than particular base pairs in the DNA[53]. Considering this model, it is interesting to note that EWS-fusions to any ETS factor are capable of microsatellite binding, however, transcriptional activity requires fusion of EWS to the FLI (FLI, ERG, FEV1), or PEA3 (ETV1 & ETV4) families[51] (Figure 1.3).

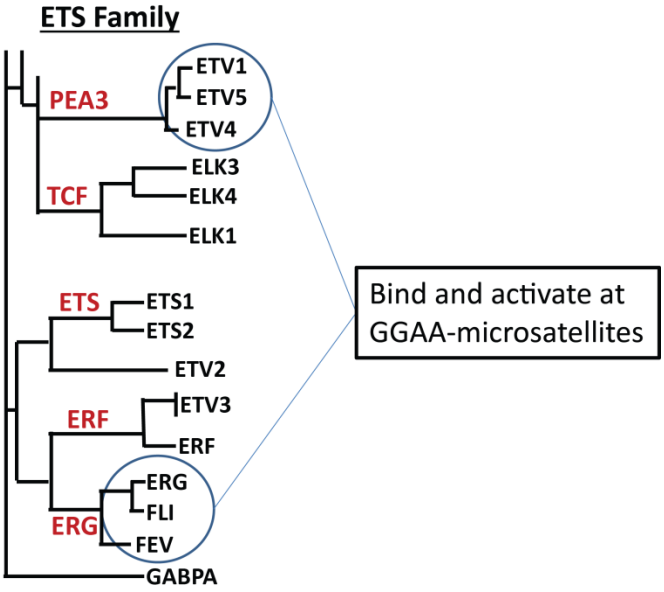


Figure 1.3 ETS protein family members capable of binding at GGAA-microsatellites

Understanding the binding mechanism of EWS/FLI on these unique response elements will likely inform on EWS/FLI functionality as an aberrant transcription factor. Additionally, the high propensity of ETS proteins for homo and hetero-dimeric interactions suggest a high likelihood of multimeric EWS/FLI binding at GGAA-microsatellites. Further study of these binding interactions, as in this thesis work, provide

useful insight for how EWS/FLI, and transcription factor DNA-binding broadly, modulates specific transcriptional regulatory repertoire.

EWS and its paralogs

It has long been known that EWS/FLI localizes to the nucleus and binds DNA in a sequence-specific manner, relying on EWS as a potent transcriptional activator[8,75]. EWS has two distinct regions, both necessary for oncogenic transformation of Ewing cells: one sufficient for most transformation (protein-protein interactions) and another sufficient for most of the transcriptional activity[44]. Both the activating and repressive activity of EWS/FLI requires amino acids 1-82 and 118-264 in the amino-terminal EWS portion of the fusion[34]. As yet, however, regions separating transcriptional activation and repression have not been found, though both are critical to EWS/FLI oncogenic function[44,63]. Additionally, EWS is an intrinsically disordered protein (IDP), containing a series of [G/S]Y[G/S] amino acid repeats[76,77]. These hexapeptide repeats comprise the LC (low-complexity) domain of EWS, and paralogous FET/TET RNA-binding proteins FUS/TLS and TAF15[78].

Prion-like N-terminal SYQG-rich domains are intrinsically aggregation prone sequences, shown by the high PONDR (predictor of naturally disordered regions) score for the LC region of EWS, and the other TET family members[76]. The field's current hypothesis is that polymerization of these intrinsically disordered regions (IDRs) enables formation of higher-order assemblies that allow for transcriptional activation[47,78–80]. Some groups

believe these triplet repeats are critical for transcriptional activation, some think for polymerization, and others hypothesize both[76,79]. IDP's lack fixed or three-dimensional ordered structures, and challenge the traditional paradigm that protein function depends on fixed structure. The umbrella term of IDP refers to all fully disordered proteins, as well as proteins containing IDR's. These are generally low hydrophobicity regions, thus enabling effective interactions with water due to their high polar and charged amino acid content[81,82]. The high net charge resulting from the accumulation of these polar residues promotes disorder, partially through the resultant electrostatic repulsions. Overall, this unique class of proteins has recently drawn a lot of attention, particularly in neurodegenerative pathologies. Thus, IDPs perhaps comprise systems of structured or "organized chaos," with significant biological functions and pathological implications[83].

Examining these IDR's further in a Ewing sarcoma setting, seventy-percent of the EWS activation domain (EAD) is composed of degenerate hexapeptide repeats (SYGQQS) with a highly conserved tyrosine residue[47,76]. Substitution of phenylalanine, but not alanine for some of these tyrosine residues still enables effective transcriptional activity for the EWS/ATF1 fusion, suggesting the structural need for an aromatic ring to confer EAD function. Retained activation of such a mutant also suggests that this activation occurs independently of hydroxyl phosphorylation at the tyrosine residue[76]. FUS and other EWS paralogs demonstrate a similar requirement for tyrosine residues within its LC-region[78,79,84].

Building on the previously mentioned model regarding the necessity of conserved SYQG-rich domains, it is further hypothesized that it is specifically the highly conserved tyrosine residues that are critical for EAD function[47]. Though the precise mechanism is not known, interestingly several of these tyrosine residues are putative phosphorylation sites[76]. In a series of polymerization studies, McKnight's group observed that phosphorylation of the LC domain of FUS prevents hydrogel retention[85].

Just as structural studies of closely-related ETS family members can be used to predict likely DNA binding mechanisms of FLI, studies of FUS and TAF15 may help generate hypotheses about the mechanisms involving the EWS portion of the fusion.

TET/FET proteins (EWSR1, TA15, and FUS/TLS) are a family of RNA-binding proteins frequently associated with disease pathogenesis, including both fusion onco-protein-mediated malignancies, as well as a number of neurogenerative diseases. The N-terminus of each of these TET proteins serves as a transcriptional activator when fused to a DNA-binding domain[63,86,87]. For example, FUS/TLS-CHOP is found in myoxid liposarcoma, a subtype of malignant adipose tumors[88]. Full length wild-type EWS, however, has decreased activation potential[89–91].

In normal tissue, these proteins function in cell growth, pre-mRNA splicing, and RNA polymerase II-mediated transcription[92]. Loss of FUS alters the distribution of RNA polymerase II across the genome, facilitating decreased RNA polymerase II C-terminal

domain (CTD) phosphorylation at serine residues located at the TSS[93]. As serine residue phosphorylation of the CTD is a critical component of RNA pol II-mediated transcriptional initiation and elongation, this implies a critical role for FUS in transcriptional regulation via its direct interaction with RNA pol II[94,95]. Additionally, fibrous assemblies of FUS have been demonstrated bound to the CTD of RNA polymerase II, suggesting the ability of FUS to polymerize may be required for this interaction[79,96]. Further evidence for this is supported by fluorescence microscopy studies demonstrating co-localization of FUS in a granular distribution with RNA polymerase II in human fibroblasts[97]. These include nuclear aggregates in ALS patient-derived fibroblasts, implicating FUS in a number of neurodegenerative diseases[80,97].

Interestingly, the CTD of RNA pol II consists of 52 hepta-peptide repeats (YSPTSPS), and makes up a part of the largest RNA polymerase II subunit, RPB1[98]. This domain is often bound by other proteins, such as transcription factors, to activate polymerase activity, and is by extension heavily involved in transcriptional initiation, RNA transcript capping, and even plays a role in spliceosome attachment[99,100]. Wild-type FUS binds to RNA pol II via its N-terminal domain, while simultaneously interacting with two serine-arginine (SR) splicing factors via its C-terminal domain. In contrast, the FUS-ERG leukemia fusion protein, lacking its C-terminal domain, is unable to recruit these splicing factors, resulting in both alternative splicing of CD44 mRNA, as well as E1A pre-mRNA splicing inhibition[101].

While homology of FUS to EWS renders these studies potentially suggestive of similar EWS-related structural and functional mechanisms, FUS interaction with the CTD appears to be more stably associated with the RGG-Zn-RGG domain than the proline-rich low-complexity domain (LCD) found in the EWS/FLI fusion[3,76,96,102]. While the hexapeptide-repeat containing N-terminus of both EWS and EWS/FLI has also been shown to interact with RNA polymerase II[101], co-immunoprecipitation studies suggest wild type EWS and EWS/FLI may play different roles in RNA Pol II mediated transcription as they associate with distinct factors of the RNA pol II complex [86]. Recent evidence, including the data demonstrated in this work, suggest the mechanism of EWS/FLI-mediated transcriptional activation as a consequence of its interaction with RNA polymerase II may be linked to EWS/FLI binding to GGAA-microsatellite sequences.

One of the barriers to studying EWS/FLI itself has been the difficulty of its isolation for *in vitro* experimentation. As discussed above, full-length EWS/FLI is a very disordered and aggregate protein[76]. This has made this fused oncoprotein difficult to study from a biochemical/biophysical perspective. The research contained herein utilizes a series of innovative truncated constructs of functional versions of EWS/FLI, which are unique to our lab. These allow us to more specifically characterize the mechanism of Ewing sarcoma development. Two such constructs are Δ22 (which comprises just the FLI-portion of EWS/FLI), and Mut9 (the FLI portion of the chimera fused with the minimal amount of EWS required for transcriptional regulation and oncogenic transformation).

These constructs have been instrumental for understanding the cooperation of EWS/FLI molecules with DNA at GGAA-microsatellites (Figure 1.4). Use of these constructs has also helped to elucidate the significance of length polymorphisms seen in clinical GGAA motifs, using a series of biochemical techniques, as well as cell culture applications.

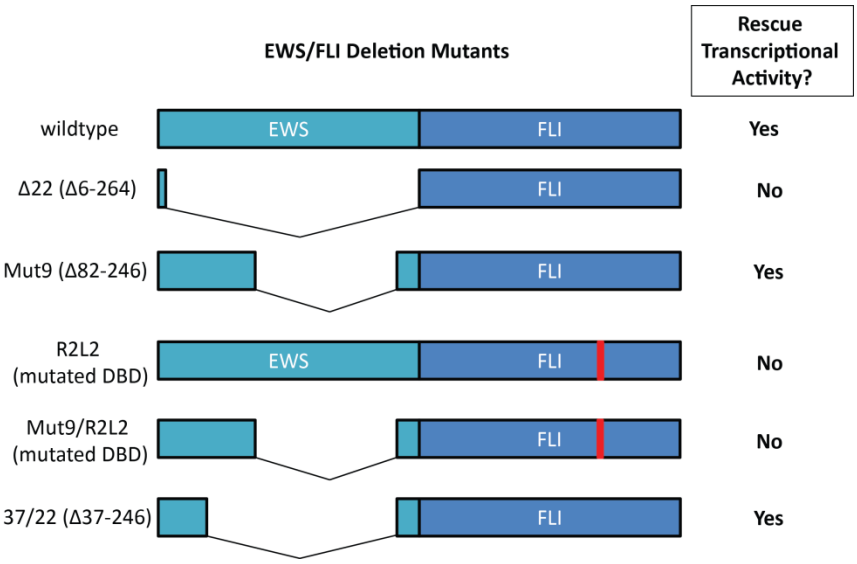


Figure 1.4 EWS/FLI deletion mutants used in the studies contained within this thesis.

Additional EWS/FLI mutants that have significantly aided this research are the R2L2 mutant, and our +37/+18 mutant. R2L2 is similar to our wild type/full-length EWS/FLI construct, with the exception of a mutated DNA-binding domain in the FLI region, rendering the mutant incapable of DNA binding to EWS/FLI (or ETS) target sequences. As such, it has been a useful negative control as we have sought to elucidate the mechanism of DNA binding for EWS/FLI. Our +37/+18 mutant, on the other hand, is believed to be a minimal construct for EWS/FLI activation and appears to rescue colony

formation in A673 Ewing sarcoma cells following RNAi-mediated knock-down of EWS/FLI (unpublished data).This more recent construct, though not used extensively in the outlined studies, represents a version of EWS/FLI even more whittled down in the EWS portion than our Mut9 mutant. It only contains three [G/S]Y[G/S] repeats, compared to five in Mut9 and twelve in full-length EWS/FLI. Figure 1.4 depicts the EWS/FLI mutants described in this work. If these repeats are indeed critical for EWS/FLI-mediated activation, and/or its ability to polymerize on microsatellites, such a minimal EWS-containing construct will help efficiently delineate the need for these conserved repetitive residues in future studies.

Microsatellites

DNA polymorphisms are variation in a given DNA sequence. One type of these, simple-sequence repeats (SSRs) are commonly known as microsatellites, or sequential tandem repeats of 1-13bp[103]. "Satellite" sequences are highly repetitive elements in the genome usually containing 100 or more nucleotides in any particular region[104]. They were originally identified through denaturation-renaturation experiments on mouse genomic DNA enabling C_0t-based DNA fractionation[105], because repetitive DNA sequences have a different AT/GC content than bulk DNA[106]. The term "satellites" was thus coined because this subset of repetitive DNA was found in a DNA fraction that would sediment as a strong, localized band versus the other DNA in the CsCl density gradients[105]. More recently, microsatellites are detected via PCR amplification and DNA sequencing[103].

Generally found in non-coding DNA regions and once regarded as junk DNA, a number of microsatellite sequences have been implicated in disease[15]. But how do these often vast repeat expansions arise in the genome? And why is a particular tetra-nucleotide (GGAA) microsatellite associated with Ewing sarcoma, while numerous tri-nucleotide microsatellites cause neurodegenerative diseases?

Direct sequencing evidence demonstrates that higher eukaryotes have genomes abundant in non-coding DNA. Because most human genes are expressed at relatively low levels, it appears there has been little selection pressure to reduce intron size for most human genes, thus accounting for the prevalence of repeat regions like microsatellites. Simple-sequence/satellite DNA makes up about 6% of the human genome, and may be generated by "backward slippage" of the daughter strand onto the template strand during replication. Part of this particular theory of microsatellite formation is that shorter repeats expand until a critical length, at which point they begin to contract[105]. Additionally, length variability (polymorphisms) observed between individuals within a species or a given population is likely due to unequal crossing-over during meiosis[103].

Mechanistically, however, meiotic recombination rate appears to be independent of the presence or absence of microsatellites (meiotic hotspots). Individual microsatellite polymorphisms and SNPs (small insertions/deletions in the DNA occurring about once every 1000 base pairs) have been especially useful in forensic medicine, positional cloning, and in disease genetics[103,105]. For example, microsatellite marker PCR assays

have been extensively used and validated in paternity testing and criminal investigation. They have also been used to detect and quantify transplant chimerism in allogeneic bone marrow transplants[107]. Additionally, the genetic heritability of both microsatellites and SNPs has enabled identification of DNA markers associated with disease susceptibility[103].

While microsatellites can contribute to a disease state regardless of whether they contain tri- or tetra-nucleotide repeats, it is believed that the ability of the sequence to form secondary structures and interfere with DNA replication/repair/recombination is more indicative of the contribution of genomic expansion to disease[105]. Of note however, only tri-nucleotide repeat expansions can occur in coding regions, keeping the sequence in frame. These occur in neurodegenerative disorders such as Huntington Disease and Spinocerebellar Ataxia. Longer microsatellites are generally more unstable. Expanded microsatellites acting as recessive mutations through interference with gene function or expression is relatively rare, while microsatellite expansion that behaves as dominant mutations are a much more frequent disease-contributor of these repetitive elements[103]. For example, in Huntington's disease, a CAG poly-glutamine expansion in the first exon of the HD gene creates toxic aggregates upon translation and protein synthesis. Duchenne's Muscular Dystrophy (DMD) is characterized by a CUG repeat that interferes with normal mRNA splicing and has deleterious effects on nerve and muscle cell function[103]. Other tri-nucleotide repeats in non-coding sequences contributing to disease

states include Fragile X syndrome, Friedrich's ataxia, Type I Myotonic Dystrophy (DM1), and progressive myoclonus epilepsy[105].

Disease-associated microsatellite expansions display a number of both interesting commonalities and disparaging differences that give rise to the respective pathophysiology for each. For example, sequence-specific (CCG vs. CTG) tri-nucleotide repeats have been shown to inhibit or enhance, respectively, nucleosome formation. Additionally, microsatellites occur in promoters or near promoters more than would be expected simply by chance alone, suggesting a potential role in transcriptional regulation[108]. To collectively demonstrate the spectrum of their functional contribution to these disease states, the following table provides a brief review of a number of disease-associated microsatellites (Table 1.4)[103,107,109].

Disease	**Microsatellite**	**Gene**	**Location**	**Normal Range**	**Disease Range**
Ewing sarcoma	GGAA	intergenic	non-coding	12-60	20-25
Huntington's Disease	CAG	*Huntingtin*	coding	6-35	40-121
Spinocerebellar Ataxia	CAG	*SCA1*	coding	6-44	39-82
Duchenne's Muscular Dystrophy	CA	*dystrophin*	coding		
Fragile X syndrome	CGG	*FMR1*	5'-UTR	5-55	> 200
Friedrich ataxia	GAA	*FDRA*	Intron 1	34-100	200-1700
Type I Myotonic Dystrophy	CTG	*DMPK*	3'-UTR	5-37	50-3000

Table 1.4 Microsatellite-associated diseases

As seen in these disease examples, most diseases with microsatellite repeat expansions are comprised of sequences enriched in G and C nucleotides. Ironically, G and C nucleotides contribute increased stability to DNA Watson-Crick base-pairing. These examples demonstrate the frequent correlation of increasing numbers of disease-associated microsatellite repeats with increasing disease severity. One mechanistic basis for this trend may be explained by microsatellite instability (MSI).

Microsatellite instability refers to the expansion or loss of repeats in microsatellite regions due to defective mismatch repair caused by "stuttering" of the DNA polymerase. This stuttering is commonly referred to in the literature as "slippage," though point mutations are actually more common and are frequently the cause of microsatellite alterations[104]. These repeat expansions and contractions generally occur in multiples of the repeat motif, as with the above examples[103,107]. MSI and CIN (chromosomal instability), or alterations in chromosomal number, can both result in cancer-initiating mutation events, however, they are mostly mutually exclusive in tumors[104]. Additionally, the few tumors that have actually been directly linked to microsatellite instability show little aneuploidy and hardly any CIN.

Microsatellite analysis to observe global genomic instability, LOH, CNV, and mapping of tumor suppressor genes is routinely performed in sarcoma research. In Ewing sarcoma, EWS/FLI preferentially binds GGAA-microsatellite repeat DNA regions embedded within the promoter region upstream from the transcriptional start site of genes directly

bound and activated by EWS/FLI[15]. Importantly, this enrichment has been observed near directly up-regulated, but not down-regulated genes. EWS/FLI and wild type FLI are capable of binding to GGAA-microsatellites *in vitro* with similar efficiency, though activation only occurs through the chimeric fusion protein[15]. Moreover, GGAA-microsatellite mediated transcription is not normally regulated by ETS family proteins in an *in vivo* setting.

From an epidemiological standpoint, the incidence of Ewing sarcoma in African populations, independent of geographical location, is ten times less than in European populations. As microsatellites are often polymorphic throughout populations, our laboratory PCR-amplified and commercially sequenced the *CAV1* and *NR0B1* microsatellite from the genomic DNA of 100 African and 100 European "normal" individuals[14]. The *CAV1* total microsatellite length and repeat number showed little variation between Africans and Europeans, though it did display a high polymorphic rate overall. In addition to *NR0B1* also being highly polymorphic, GGAA-microsatellites associated with this critical EWS/FLI target gene demonstrated significant differences between ethnic populations (Figure 1.5).

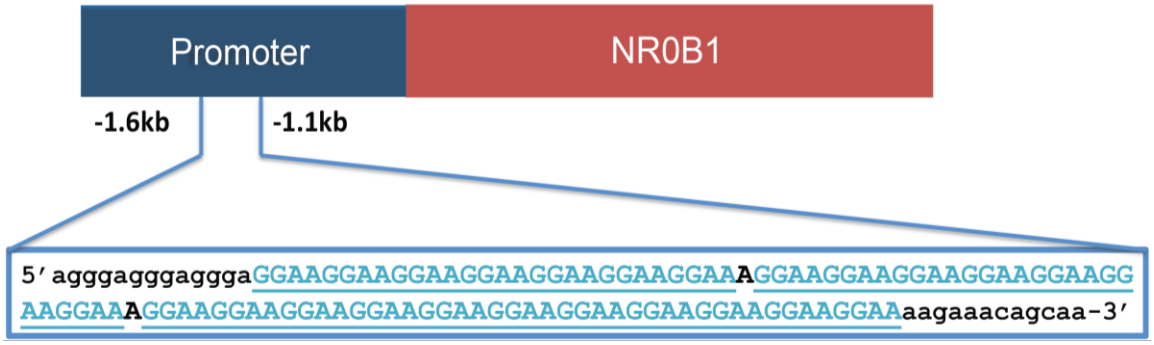


Figure 1.5 The NR0B1 GGAA-microsatellite, located about 1.5kb upstream of the NR0B1 transcriptional start site (TSS)

Overall, about 40% of the African samples contained microsatellites with greater than 30 GGAA-repeats, while the majority of the European samples contained a far more narrow range of 20-26 GGAA repeats, which we have dubbed the "sweet-spot"[14].

An interesting correlative of microsatellite length in Ewing sarcoma compared with other cancers and diseases is the suggestion of a length-dependent effect. While diseases such as Huntington's, Friedrich Ataxia, and Myotonic Dystrophy all contain repeat expansions within a particular gene or genetic loci, GGAA-microsatellites in Ewing sarcoma are found genome-wide, demonstrating a binding preference for nine or more repeats[15,46]. Further, GGAA-motifs are found in the promoter regions of a number of different genes and are thought to play a role in regulating proteins involved in cellular differentiation, immune response, and endocrine systems[50].

As we strive in this work to understand these unique GGAA-microsatellites in a Ewing sarcoma context, one of the questions regarding length becomes whether it is the total

number of repeats, or the number of *consecutive* GGAA-motifs that is significant for enabling EWS/FLI-mediated activation[15]. Our "sweet-spot" finding facilitates proposal of a unique length-dependent model with novel implications for our global understanding of transcription factor biology.

Goals of thesis

Microsatellite repeats characterized by the motif GGAA serve as binding sites for EWS/FLI within the promoters of upregulated target genes in Ewing sarcoma[15]. Previously thought of as "junk DNA," these microsatellites serve as response elements for EWS/FLI DNA binding with interesting genetic correlations and possible clinical implications. As an example, *NR0B1* transcriptional upregulation by EWS/FLI is necessary for oncogenic transformation of Ewing sarcoma cell lines, with the EWS/FLI binding occurring at a nearby GGAA-microsatellite[43,46]. At the start of the enclosed studies, it was hypothesized that the number of *consecutive* repeats rather than the total number of motifs determines binding and increased transcriptional activity[15]. In accordance with this hypothesis, it was previously observed that transcriptional activity increases with increasing numbers of GGAA-motifs, and that EWS/FLI binds DNA comprised of 4-7 of these repeats as a homodimer[51]. These initial data gave rise to a cooperative model for Δ22 binding, with the FLI domain binding to and interacting with microsatellites (Figure 1.6). We therefore conjectured that increasing numbers of FLI molecules bound would result in an overall increase in binding affinity on the DNA.

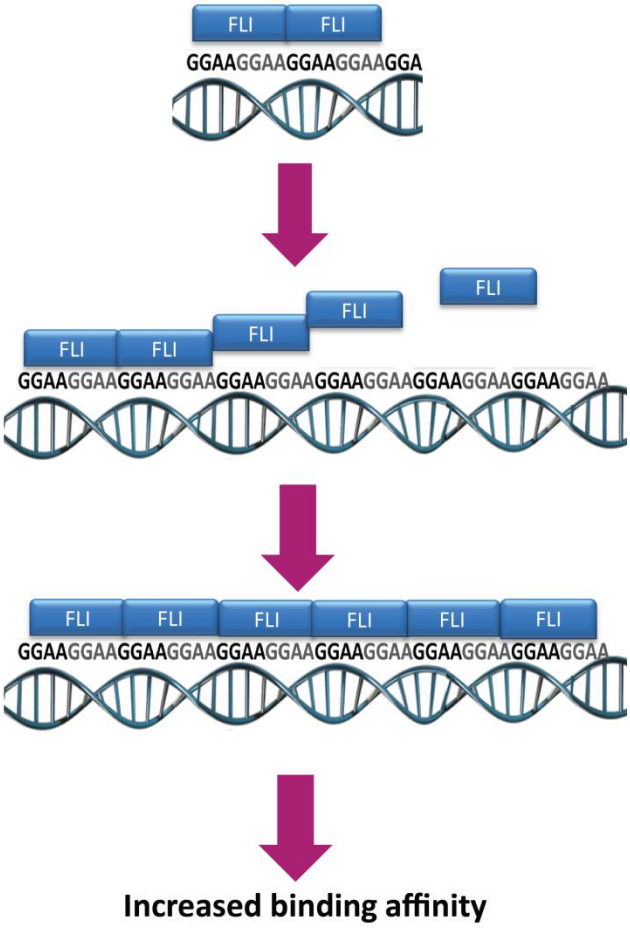


Figure 1.6 Cooperative model for FLI binding at GGAA-microsatellites

Prior to this body of work, both binding affinity and stoichiometric analysis of EWS/FLI at larger repeat regions had not been evaluated. Though the work presented herein has made strides in our understanding of this EWS/FLI-microsatellite interaction, additional exploration is still needed to characterize more precisely the biochemical mechanism of GGAA-microsatellites facilitating Ewing sarcoma oncogenesis, and to further evaluate whether clinical correlations exist.

Though it was previously shown *in vitro* that increasing numbers of contiguous GGAA-repeats correlates with an increased number of bound EWS/FLI proteins *in vitro*[51], no studies had been conducted on GGAA-microsatellites in Ewing sarcoma patients prior to this work.

The overall objective of this thesis was to investigate the biochemical properties that dictate how EWS/FLI regulates activation of its targets. Our main hypothesis is that EWS/FLI transcriptionally regulates oncogenic transformation and subsequent tumor progression through site-specific and length-dependent binding to specific response elements. Understanding how EWS/FLI functions is critical for developing adequate therapies to target this previously "undruggable" oncoprotein. Additionally, it is hoped that results from these in-depth studies may also provide insight into the oncogenic transcriptional regulation of other pediatric cancers driven by fusion oncoproteins, such as MLL-rearranged leukemia.

The work proposed herein evaluates the following aims:

1) Determine to what extent GGAA-microsatellites function as enhancer response elements. Although GGAA-microsatellites were identified near promoters of EWS/FLI up-regulated genes[15] and minimal clinical studies were performed comparing microsatellite length in African versus European populations[14], no one has definitely proven the role of these microsatellites in Ewing sarcoma as EWS/FLI-specific response

elements prior to the enclosed work. To this end, studies in this work clarify the role of GGAA-microsatellites in EWS/FLI-mediated transcriptional activation.

2) Elucidate the role of GGAA-motif length in driving transcriptional activation in Ewing sarcoma. Understanding why GGAA-microsatellite length correlates with increased disease susceptibility provides a unique and significant model for understanding oncogenesis in pediatric cancer. Further, biochemical characterization of EWS/FLI binding to GGAA-microsatellites suggests a potential sequence-specific mechanism by which EWS/FLI differentiates between transcriptional activation versus repression of its target genes. Additionally, study of a potential length-dependent effect in microsatellite function has broad biological implications that relate to a number of neurodegenerative and genetically devastating diseases involving microsatellite repeat regions.

3) Determine whether GGAA-microsatellite characteristics are predictive of EWS/FLI responsiveness across the genome. In addition to biochemical and cell culture-based studies, this work utilizes bioinformatics and statistical analysis to both define microsatellites and to determine whether repeat number and other microsatellite features inform both binding and transcriptional regulation at these sites.

Globally, this work expands our understanding of EWS/FLI-mediated transcriptional activation at GGAA-microsatellites. Additionally, it improves our mechanistic

knowledge concerning how GGAA-repeat length enables optimal binding and effector function. Specifically, we demonstrate that microsatellites are EWS/FLI bona fide activating response elements that are critical for Ewing sarcoma oncogenesis. Additionally, our data highlights a "sweet-spot" length of microsatellite that correlates with patient susceptibility, and elucidates a novel role for the EWS portion of the fusion in EWS/FLI binding of DNA. This study and future investigations relating to this body of work represent steps paramount to innovative therapeutic discovery as we strive to effectively combat Ewing sarcoma.

Chapter 2: Microsatellites with Macro-influence in Ewing Sarcoma (*Review Article*)

Michael J. Monument, Kirsten M. Johnson, Allie H. Grossmann, Joshua D. Schiffman R., Lor Randall, and Stephen L. Lessnick. (2012) *Genes*. 3(3), 444-460. Doi: 10.3390/genes3030444.

KMJ and MJM wrote the document. KMJ generated Figure 2.2 and MJM generated Figure 2.3. AHG, JDS, LR, and SLL reviewed the document.

Abstract

Numerous molecular abnormalities contribute to the genetic derangements involved in tumorigenesis. Chromosomal translocations are a frequent source of these derangements, producing unique fusion proteins with novel oncogenic properties. EWS/ETS fusions in Ewing sarcoma are a prime example of this, resulting in potent chimeric oncoproteins with novel biological properties and a unique transcriptional signature essential for oncogenesis. Recent evidence demonstrates that EWS/FLI, the most common EWS/ETS fusion in Ewing sarcoma, upregulates gene expression using a GGAA microsatellite response element dispersed throughout the human genome. These GGAA microsatellites function as enhancer elements, are sites of epigenetic regulation and are necessary for EWS/FLI DNA binding and upregulation of principal oncogenic targets. An increasing

number of GGAA motifs appear to substantially enhance EWS/FLI-mediated gene expression, which has compelling biological implications as these GGAA microsatellites are highly polymorphic within and between ethnically distinct populations. Historically regarded as junk DNA, this emerging evidence clearly demonstrates that microsatellite DNA plays an instrumental role in EWS/FLI-mediated transcriptional regulation and oncogenesis in Ewing sarcoma. This unprecedented role of GGAA microsatellite DNA in Ewing sarcoma provides a unique opportunity to expand our mechanistic understanding of how EWS/ETS fusions influence cancer susceptibility, prognosis and transcriptional regulation.

Introduction

Aberrant chromosomal translocations are common observations in cancer and in many instances these events give rise to chimeric fusion products with novel biological and cellular functions. Many of these chimeric fusion proteins function as oncogenic transcription factors, essential for cellular transformation and/or critical malignant cellular phenotypes[110,111]. Ewing sarcoma is a highly aggressive bone associated malignancy primarily affecting children and young adults, ubiquitously characterized by and derived from a balanced chromosomal translocation[1,112]. Ewing sarcoma belongs to a larger class of malignancies referred to as *sarcomas*, a term ascribed to a heterogeneous grouping of tumors derived from, or highly associated with connective tissue elements and mesenchymal precursors (Figure 2.1). Ewing sarcoma is an aggressive malignancy, with significant metastatic potential. Roughly 20% of patients present clinically with

detectable metastatic disease, where survival ranges from 60-75% in patients with localized disease and plummets to <20% in those with local recurrence or metastatic disease[21,112].

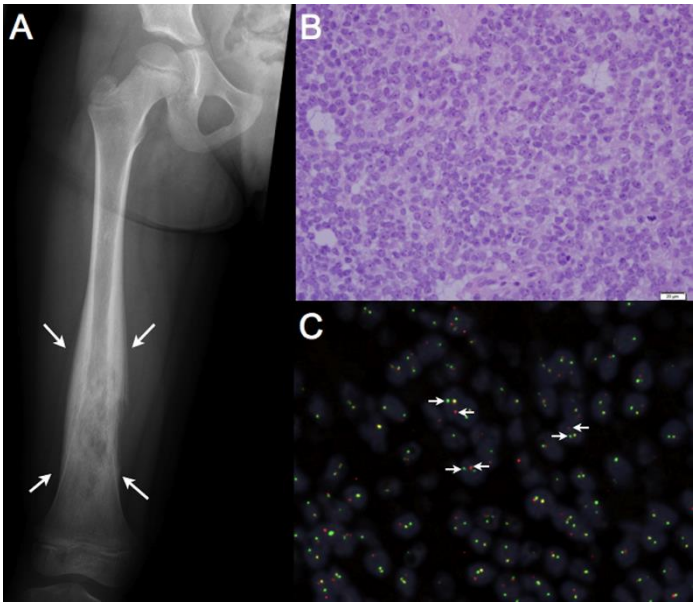


Figure 2.1 Ewing sarcoma is an aggressive bone associated malignancy characterized by chromosomal translocations.

(A) Classic radiographic appearance of Ewing sarcoma: an expansile, destructive lesion (outlined by white arrows) of the femoral diaphysis (shaft) in a skeletally immature patient. Ewing sarcomas can also present as an isolated soft tissue mass, although this is less common. **(B)** 400X magnification of a Hematoxylin and Eosin (H & E) stained section from a Ewing sarcoma tumor. Microscopically, these tumors are characterized by sheets of small round cells with a high nuclear-to cytoplasmic ratio. **(C)** Break-apart Fluorescence *in situ* Hybridization (FISH) showing EWSR1 rearrangements in 84% of tumors cells, confirming the diagnosis of Ewing sarcoma. Dual, non-overlapping, 5'-*EWSR1* probes (red) and 3'-*EWSR1* probes (green) detect the presence of a chromosomal rearrangement; when the red and green probes are split into two distinct signals (white arrows) a chromosomal rearrangement is identified, whereas an orange signal indicates an intact *EWSR1* locus.

Virtually all Ewing sarcoma tumors harbor a somatic translocation, fusing the *EWSR1* gene (encoding the EWS protein) on chromosome 22 with a member of the ETS family of transcription factors, most commonly *FLI1* (encoding the FLI protein), located on chromosome 11 [t(11;22)(q24;q12)]. The EWS/FLI fusion product is observed in 80-85% of cases, with highly related fusions such as EWS/ERG, EWS/FEV, EWS/ETV1 and EWS/ETV4 occurring less frequently (reviewed in Sankar and Lessnick, 2011)[113]. In Ewing sarcoma, chimeric EWS/ETS fusion products function as an aberrant oncogenic transcription factor, mediated by the transcriptional activating amino-terminus of EWS fused in frame to the DNA binding carboxy-terminus of the ETS transcription factor (Figure 2.2). Numerous studies have since confirmed that malignant transformation in Ewing sarcoma is dependent on EWS/ETS fusions and consequently, these chimeric oncoproteins are regarded as critical upstream regulators of the transcriptional hierarchy in this cancer[39,75,114]. The prevailing influence of EWS/FLI in Ewing sarcoma provides a unique opportunity to further characterize the oncogenic properties of EWS/ETS proteins, with hope that this growing body of knowledge will allow for a greater understanding of the molecular basis of oncogenesis and facilitate the development of more targeted, clinically efficacious therapy for this devastating malignancy.

ETS family of transcription factors

The ETS (E-twenty-six) transcription factors belong to a family of highly evolutionarily conserved DNA binding proteins instrumental for a variety of critical cellular processes including proliferation, cellular differentiation, angiogenesis, lymphoid cell development, apoptosis and cell migration (reviewed in ref[115]). Given these important functions, it is of

no surprise that dysregulation of numerous ETS family members is commonly observed in cancer. For example, in 50-70% of prostate cancers, chromosomal rearrangements involving ETS-members have been observed[116,117]. In many instances, these rearrangements position the androgen-receptor regulatory element, *TMPRSS2*, directly upstream of the ETS-member, *ERG*, resulting in a hormone-driven overexpression of this transcription factor in prostate cells[116]. In contrast, as this review will expand upon, fusion of the ETS-DNA binding to the transcriptional activating domain of EWS in Ewing sarcoma results in a transcription factor with unique biological properties responsible for oncogenic transformation[3,8,75].

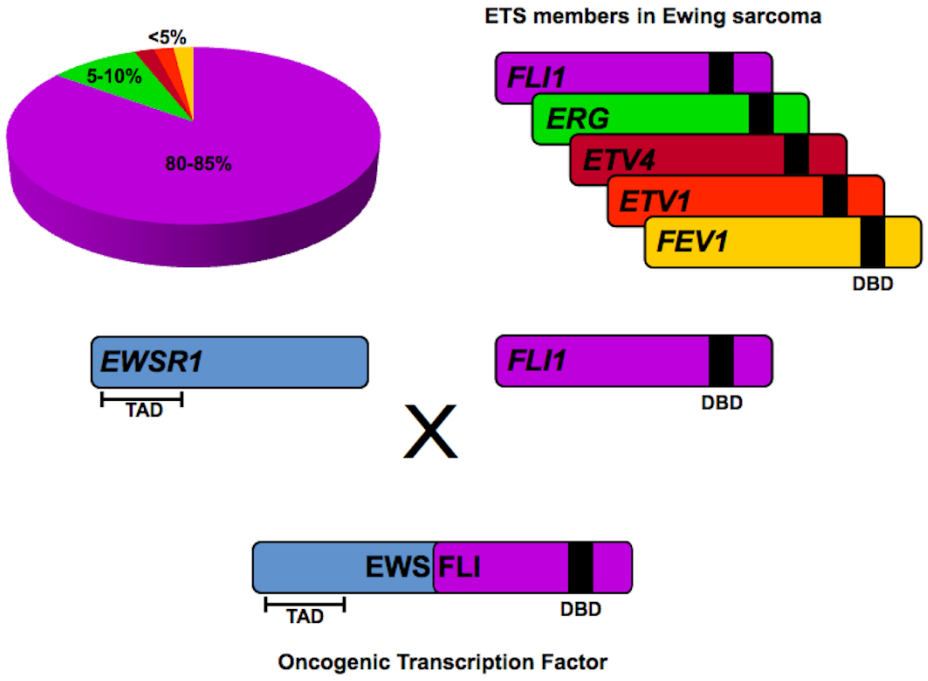


Figure 2.2 EWS/ETS fusions in Ewing sarcoma.

EWS/FLI fusions comprise 80-85% of all translocations in Ewing sarcoma. Translocations involving other ETS family members such as *ERG*, *ETV4*, *ETV1*

and *FEV1* are less common. In all instances, the transcriptional activating domain (TAD) in the N-terminus of EWS is fused to the C-terminal DNA binding domain (DBD) of the ETS family member. The resultant chimeric fusion protein functions as a potent oncogenic transcription factor responsible for tumorigenesis in Ewing sarcoma.

Twenty-eight distinct ETS-family members have been identified in humans, which are further categorized into four ETS-subfamilies of more highly related members[49,118]. Common to all ETS-family members is a highly conserved DNA binding domain referred to as the *'ETS domain.'* Structurally, this *'ETS domain'* is a winged helix-turn-helix DNA binding domain composed of about 85 amino acids[119]. This highly conserved DNA binding domain permits binding of ETS-family members to an invariable GGAA/T core DNA target, flanked by nucleotides which facilitate specific ETS-member targeting and cooperative protein-protein interactions[49,120,121]. Two general categories of ETS binding sites have been characterized, which include a high-affinity ETS consensus site located 20-40bp upstream of the transcriptional start site and a lower affinity consensus site further upstream in the promoter/enhancer element[49,118]. The high-affinity ETS consensus sites afford redundant ETS-member occupancy and gene regulation, are protected from DNA methylation and are associated with basal housekeeping genes. In comparison, the low-affinity sites are modified by simple flanking base substitutions, are frequently adjacent to binding sites for other cooperative transcription factors and are felt to provide a mechanism where individual ETS-members can regulate a distinct cell or tissue-specific transcriptional signature[49,54,118].

EWS/FLI in Ewing Sarcoma

EWS/FLI and EWS/ERG fusions compromise 80-85% and 5-10% of translocations observed in Ewing sarcoma, respectively[3,113]. Wild-type FLI and ERG are closely related proteins grouped within the ETS class I subfamily. As with other ETS-members, they bind DNA with preference for the traditional ETS high-affinity consensus sequence (ACCGGAAGT) via the highly conserved C-terminal 'ETS domain' and possess a weak N-terminus transcriptional activating domain[37,75]. Both function as important regulators of hematopoiesis, B-cell development and vasculogenesis[122–124]. Given the predominance of EWS/FLI fusions in Ewing sarcoma, the biology of wild type and fusion-associated FLI has been most thoroughly characterized. In contrast, the precise biology of wild type EWS remains ill-defined, however reports indicate wild type EWS functions as an RNA binding protein and participates in alternative RNA splicing[125–127].

Functional investigations over the last two decades clearly demonstrate that the biological properties of the EWS/FLI chimera are vastly distinct from wild-type FLI. For instance, while both FLI and EWS/FLI share affinity for the ETS consensus site, the EWS/FLI chimera is a substantially more potent transcriptional activator than wild-type FLI[8,75]. Additionally, ectopic expression of EWS/FLI in NIH 3T3 fibroblasts induces oncogenic transformation whereas wild-type FLI does not[8]. Silencing of EWS/FLI expression in patient-derived Ewing sarcoma cell lines reverses the oncogenic phenotype[43,114]. Interestingly, wild-type FLI is not expressed in Ewing sarcoma cells[114]. Furthermore, the transcriptional signature and genomic targeting of EWS/FLI in Ewing sarcoma is

markedly different from wild-type FLI[128], despite a shared affinity for ETS consensus sites[8,51].

EWS/FLI fusions mediate gene dysregulation via a GGAA microsatellite response element

Genome-wide microarrays have identified >1000 EWS/FLI-regulated genes, including indirect and direct gene targets[43,114,129]. Interestingly, ~80% of these are down-regulated targets. Subsequent chromatin immunoprecipitation approaches, including ChIP-chip and ChIP-seq have further characterized many direct EWS/FLI targets[15,46,128]. Many of the identified up- and down-regulated targets are associated with oncogenic processes described in a variety of other cancer models. However, the most highly regulated and bound target observed across multiple data sets is the gene *NR0B1* (also called *DAX1*)[43,46,114]. NR0B1 is an orphan nuclear receptor, a member of the sex-steroid receptor family, and is important for development of the hypothalamus-pituitary-adrenal-gonadal axis and sex determination[130,131]. NR0B1 has no prior associated role in oncogenesis, which is compelling given the results of the aforementioned microarray and ChIP-chip datasets. Interestingly, *NR0B1* is not bound or transcriptionally regulated by wild-type FLI[128,132]. Numerous independent reports have further validated that *NR0B1* is upregulated, a direct EWS/FLI target, and highly expressed in Ewing sarcoma. Additional functional assessments have shown that in patient-derived Ewing sarcoma cell lines, dysregulated *NR0B1* expression is necessary for oncogenic transformation[15,43,46,132,133].

Genome-wide localization studies have established that EWS/FLI highly occupies the *NR0B1* promoter. Mutational experiments have further demonstrated that a 500bp region, roughly -1.6kb upstream from the *NR0B1* transcriptional start site is required for EWS/FLI-mediated DNA binding and gene activation[15]. Within this 500bp region is a 102bp microsatellite characterized by a series of repetitive GGAA tetra-nucleotide repeats. Numerous investigations have demonstrated that EWS/FLI-mediated binding and activation of *NR0B1* is dependent on this repetitive element[15,46,133]. Interestingly, the highly enriched *NR0B1* promoter does not contain the traditional high-affinity ETS consensus site (ACCGGAAGT)[15,46]. Luciferase reporter constructs and electrophoretic mobility shift assays (EMSA) have further validated the *in vitro* specificity and affinity of EWS/FLI for both the 102bp *NR0B1* GGAA microsatellite and similar synthetic GGAA microsatellite constructs[15,51]. This data provides compelling evidence that the GGAA microsatellite of the *NR0B1* promoter functions as an "*EWS/FLI response element*," necessary for DNA binding and gene activation. Of the twenty-eight distinct ETS-members in humans, only 5 have been observed in chromosomal rearrangements with *EWS* in Ewing sarcoma (EWS/FLI, EWS/ERG, EWS/FEV, EWS/ETV1 and EWS/ETV4). All of these related fusion proteins are capable of binding the 102bp *NR0B1* GGAA microsatellite and activate gene expression[15,51]. Wild-type ETS-members can also bind the GGAA microsatellite; however, unlike EWS/ETS fusions, binding to these elements does not activate gene expression[51].

Microsatellite constitution influences EWS/FLI binding and gene activation

ETS family members are commonly known to bind DNA in a monomeric configuration with a characteristic DNAase I footprint of 14-15bp, although only 9-10bp are required for sequence specificity[134]. At the aforementioned high-affinity DNA sites, ETS-members bind as monomers, whereas at the lower-affinity, divergent DNA sites, ETS-members often bind as heterodimers in a cooperative fashion with other cell/lineage specific transcription factors[49]. In Ewing sarcoma, EWS/FLI appears to bind to GGAA microsatellites as a homodimer and requires a minimum of 4 consecutive GGAA motifs (16bp) for binding and gene activation[15,51]. Importantly, beyond a threshold of 4-6 repeats, an increasing number of GGAA motifs results in a proportional increase in EWS/FLI-mediated gene expression in both synthetic reporter constructs and *bona fide* targets, such as *NR0B1* (Figure 2.3)[15,46,51,128,132,133]. Genome-wide localization data further supports these observations, as sites of EWS/FLI enrichment are greatest in regions with microsatellite elements containing 12-14 consecutive GGAA motifs[46,128].

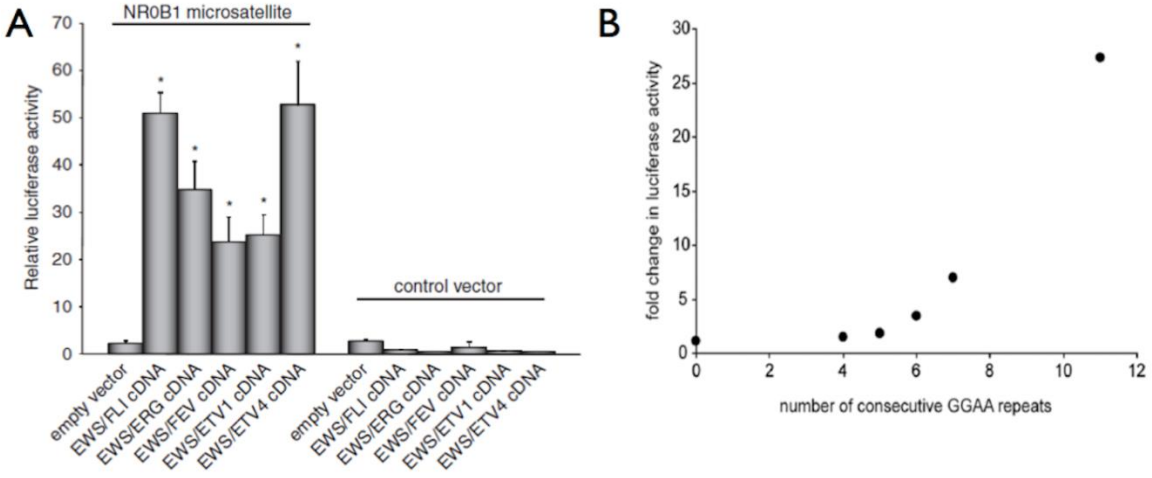


Figure 2.3 EWS/ETS fusion proteins bind DNA and regulate gene expression via a GGAA microsatellite response element.

(A) In luciferase reporter constructs, all five EWS/ETS fusions can activate gene expression via the 102bp *NR0B1* microsatellite. **(B)** Using similar reporter constructs, an increasing number of GGAA motifs, beyond a threshold of four, results in increased gene expression. Panel A reproduced with permission from *Gangwal et al., Genes Cancer. 2010 February 1; 1(2): 177–187*[51].

Collectively, these findings demonstrate an unprecedented role for microsatellite elements as direct EWS/FLI-transcriptional response elements in Ewing sarcoma. Because an increasing number of GGAA motifs substantially augments target gene expression, it is possible that the EWS/FLI chimeric protein has an increased affinity for larger microsatellites. Alternatively, larger microsatellites may facilitate the recruitment of additional EWS/FLI homodimers to produce a synergistic effect on transcriptional activation. Further studies are needed to evaluate these potential mechanisms (Figure 2.4).

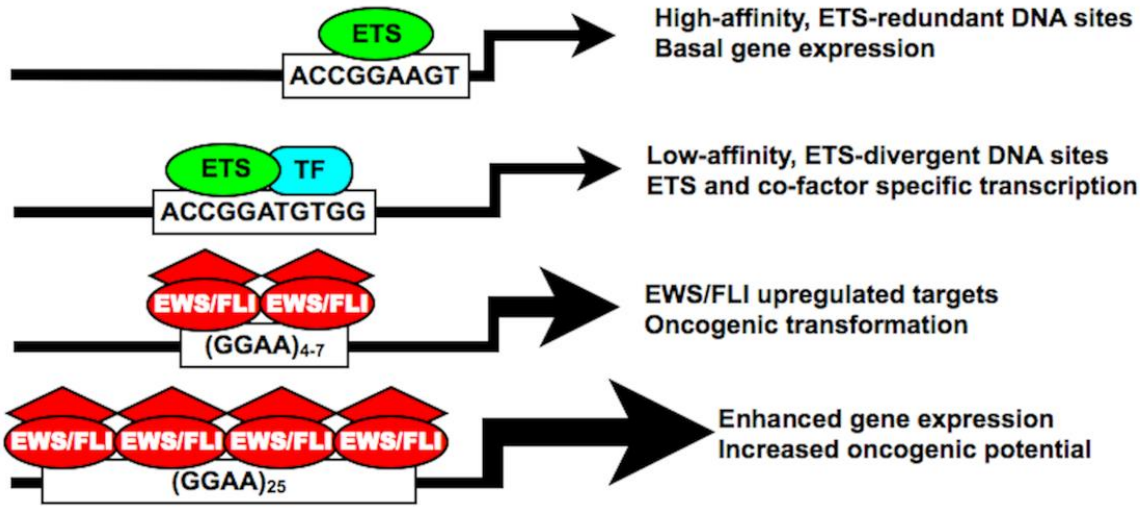


Figure 2.4 The EWS/FLI chimera possesses unique DNA binding affinities and biological properties distinct from native ETS family members.

Both high- and low-affinity ETS DNA binding sites are characterized by a core ACCGGAA/T consensus sequence facilitating both ETS-redundant and ETS-divergent transcriptional regulation. In Ewing sarcoma, EWS/FLI also binds the traditional ETS-consensus sequence, but shows increased preference for a GGAA-containing microsatellite. In certain upregulated targets, this GGAA microsatellite response element is required for DNA binding and gene activation, which proportionately increases with an increasing number of GGAA motifs. "TF" = transcription factor.

GGAA microsatellites identify other potential EWS/FLI targets and epigenetically regulated enhancer loci

The compelling evidence linking EWS/FLI-mediated transcriptional regulation of *NR0B1* in Ewing sarcoma to a GGAA microsatellite response element prompted the hypothesis that additional GGAA microsatellite containing genes may be critical targets for oncogenic transformation or other cancer-related phenotypes. By comparing EWS/FLI transcriptional microarray data-sets with genome-wide EWS/FLI localization data, numerous microsatellite-containing direct EWS/FLI targets have been identified[15,40,46,128]. For examples, in ChIP-chip experiments, a promoter microarray was used to assess ~ 17000 promoters spanning -5.5.kb to 2.5kb relative to the transcriptional start site, which identified ~ 900 direct targets. Of the top 134 EWS/FLI-bound genes, a GGAA microsatellite was identified in the promoter region of 12 genes[15]. As previously mentioned, the *NR0B1* promoter was the most highly enriched region, while the remaining GGAA microsatellite-containing genes were dispersed throughout the top 134 bound targets in no particular rank distribution. *Caveolin-1* (*CAV1*) was another GGAA microsatellite containing EWS/FLI target and encodes a critical membrane-associated protein involved in clathrin-independent endocytosis[135]. Dysregulation of *CAV1* has been associated with the metastases in other cancer models[136] and expression of CAV1 is necessary for maintenance of oncogenic transformation in patient-derived Ewing sarcoma cell lines[137]. Using a comprehensive computational mapping of the human genome screening for GGAA microsatellites, another GGAA microsatellite-containing, upregulated target, *GSTM4*, was identified. GSTM4 belongs to a family of glutathione detoxifying enzymes and in patient-derived Ewing sarcoma cell lines, GSTM4 expression is necessary for maintenance of oncogenic transformation[40]. Overexpression of this

protein also increases chemoresistance to a chemotherapeutic agent commonly used in Ewing sarcoma, etoposide[40]. Additionally, in a small clinical series, increased expression of GSTM4 in primary Ewing tumors was associated with a lower overall survival[40]. Other microsatellite-containing, direct EWS/FLI targets such as *CACNB2*, *FEZV1*, *FCGRT, FVT1/KDSR, ABHD6* and *KIAA1797* have also been identified, although the functional importance of these targets in Ewing sarcoma has not been determined[15,46].

In the ChIP-seq data-set reported by Guillon et al.,[46] a total of 246 EWS/FLI occupied regions were identified, 104 of which were characterized by a GGAA microsatellite. The vast majority of EWS/FLI occupancy was localized to intergenic regions (59%), with less frequent occupancy within gene introns, exons and promoter elements. Utilizing published transcriptional microarray data-sets, it was determined that 60% of EWS/FLI-specific binding was located within 2Mb upstream of the transcriptional start sites of upregulated EWS/FLI targets. Additionally, the distance of the GGAA microsatellite from the transcriptional start site did not correlate with the rank order of gene upregulation in these transcriptional microarrays. Instead, as predicted from numerous *in vitro* assays, the number of GGAA motifs within the microsatellite had a greater influence on EWS/FLI occupancy and gene expression, which was most pronounced at genomic sites with >9 GGAA motifs[46]. In a more recent genome-wide localization study by Patel et al.,[128] a combination of ChIP-seq and formaldehyde-assisted isolation of regulatory elements (FAIRE) produced a detailed mapping of EWS/FLI enrichment sites: 40% of EWS/FLI binding sites contained a GGAA microsatellite, >60% of these microsatellite elements were located within intergenic regions and global EWS/FLI-

enrichment favored microsatellite elements containing 8-14 consecutive GGAA motifs. Greatest enrichment was localized to a region containing a total of 25 GGAA motifs, which corresponded to the *NR0B1* promoter. A fascinating observation from this data-set was that EWS/FLI modifies the local chromatin structure at these GGAA microsatellites, characterized by a nucleosome-deplete enhancer-like signature. Silencing of EWS/FLI rapidly restored nucleosome occupancy and a closed chromatin configuration at these GGAA microsatellites[128].

Collectively, these experiments demonstrate three important mechanistic functions of GGAA microsatellites in Ewing sarcoma: first, as response elements instrumental for direct EWS/FLI-mediated transcriptional regulation of important oncogenic targets such as *NR0B1*, *CAV1* and *GSTM4*; secondly, the spatial relationship of these GGAA microsatellites to upregulated targets strongly suggests these elements possess an enhancer-like function; and finally, these microsatellite elements are regions of EWS/FLI-mediated chromatin modification, facilitating a unique transcriptional signature in Ewing sarcoma.

The *NR0B1* microsatellite: a functional assessment tool in Ewing sarcoma research

The affinity of EWS/FLI for the *NR0B1* GGAA microsatellite and subsequent gene activation mediated by this interaction is well established[15,46,51,128,132,133]. Consequently, the *NR0B1* GGAA microsatellite response element has become a useful molecular tool in Ewing sarcoma research. Since EWS/FLI is regarded as the principal upstream oncogenic transcription factor in Ewing sarcoma, it is a desirable target for drug development. High-

throughput drug and small peptide library screening protocols are effective strategies to simultaneously assess large numbers (10,000 – 50,000) of therapeutic agents potentially active against EWS/FLI. Reporter constructs using the *NR0B1* promoter are now routinely used as a sensitive measure of EWS/FLI inhibition and have assisted in the identification and a more detailed assessment of new drugs and small peptide inhibitors[138–140]. Since the precise cell of origin in Ewing sarcoma remains obscure (reviewed in ref[141]), forced expression or repression of EWS/FLI in patient-derived Ewing sarcoma cell lines and other heterologous systems is commonly employed to assess various cellular pathways of perceived importance in transformation and malignant phenotypes. The *NR0B1* promoter provides an ideal positive control for various systems of inducible EWS/FLI expression (unpublished data).

Microsatellite DNA in Cancer pathogenesis

Microsatellite DNA constitutes roughly 3% of the human genome, mostly in non-coding regions[142]. Traditionally, these repetitive elements have been regarded as "junk DNA," with an undetermined genetic function. Microsatellite DNA has been previously investigated as a potential marker of cancer susceptibility, genomic instability, and prognosis. However, the direct influence of GGAA microsatellite response elements on EWS/FLI-mediated transcriptional regulation of critical targets genes defines a completely novel role of microsatellite DNA in oncogenesis.

Microsatellite instability (MSI) refers to a change in repeat length of microsatellite DNA, typically due to loss of heterozygosity in genes coding for the DNA mismatch repair

(MMR) system. In hereditary non-polyposis colorectal carcinomas (HNPCC) and sporadic colorectal carcinomas, inherited or acquired alterations of the DNA mismatch repair system give rise to a mutator phenotype characterized by length expansions or contractions of multiple mono- and di-nucleotide microsatellites, respectively[143–145]. MSI-positive colorectal tumors possess defined biological attributes, such as a more common location in the proximal colon, increased patient survival and favorable patterns of chemosensitivity[144–146]. Detection of MSI and defects in the DNA mismatch repair system in colorectal cancer has become instrumental for the diagnosis of HNPCC, whereas in sporadic colorectal carcinomas, MSI provides an important prognostic molecular marker[147,148]. However, instability of these microsatellite sequences is more a manifestation of cancer-related genomic instability and these genetic elements do not appear to mediate specific oncogenic transcriptional signatures. Microsatellite instability has also been assessed in Ewing sarcoma, although with discordant findings[149–151]. Since it is now known that the number of GGAA motifs clearly influences EWS/FLI-mediated gene expression in Ewing sarcoma, the determination of MSI in these EWS/FLI microsatellite response elements warrants renewed assessment.

In addition to MSI, microsatellite polymorphisms associated with various genetic loci have also been associated to cancer susceptibility and pathogenesis. In breast cancer for example, overexpression of the epidermal growth factor receptor, EGFR, is a common finding in invasive ductal carcinomas, where EGFR-positive tumors represent an adverse prognostic marker[152,153]. A dinucleotide CA-microsatellite within intron 1 of EGFR has been identified and length-polymorphisms of this microsatellite have been shown to

correlate with basal transcription levels of EGFR[154]; however, a direct mechanistic understanding of this association remains unclear. In prostate cancer, a CAG tri-nucleotide has been identified in the first exon of the androgen receptor gene, coding for a polyglutamine tract in the translated protein. An increasing number of CAG motifs has been shown to reduce the transcriptional activity of the androgen receptor[155]. Polymorphisms of this polyglutamine tract in the androgen receptor also appear to be predictive of cancer susceptibility and prognosis: androgen receptors with a CAG microsatellite of ≤16 CAG motifs are associated with a lower disease incidence and less aggressive tumor biology in those with the disease[156,157]. One of the most common tumor suppressors, p53 has been shown to regulate the transcriptional regulation of one of its targets, *PIG3* using a microsatellite response element. However to date, no functional role for PIG3 has been defined in tumorigenesis[158,159].

Polymorphic EWS/FLI GGAA microsatellites: a novel approach to ethnic patterns of Ewing sarcoma susceptibility and prognosis

At present, compared to many other cancer models, the genetic and environmental risk factors for the development of Ewing sarcoma remain obscure[11]. For unknown reasons, considerable ethnic variation exists in the incidence of Ewing sarcoma: the incidence of Ewing sarcoma is greatest in European populations, which is 10- and 2-fold greater than populations of African and Asian descent, respectively[160,161]. This discrepancy is independent of geographic location, suggesting a strong genetic influence for these observations[160]. Additionally, a recent database of >1700 patients with Ewing sarcoma

demonstrated lower overall survival rates in African and Asian populations[162]. To date, no studies have conclusively explained these epidemiological patterns[11,163,164].

By virtue of the repetitive constitution of microsatellite DNA and the predilection of these repetitive elements for non-coding locations, mutational events have rendered microsatellite DNA highly polymorphic in the human population[142,165]. Microsatellite polymorphisms are routinely used in the assessment of heredity, and phylogenetic mapping of ethnically distinct human populations[166]. Given the mechanistic importance of GGAA microsatellites in EWS/FLI-mediated gene regulation, we hypothesized that polymorphic GGAA microsatellites within and between ethnically distinct human populations may exist, providing a potential explanation for the aforementioned patterns of Ewing sarcoma susceptibility and prognosis. The GGAA microsatellites of the *NR0B1* and *CAV1* promoters were sequenced from 100 unaffected subjects of European and African descent. Our initial hypothesis favored larger GGAA microsatellites in Europeans given the disproportionately high incidence of Ewing sarcoma in this population.

Results from this study demonstrated that the *NR0B1* and *CAV1* GGAA microsatellites were highly polymorphic in both European and African populations. The *NR0B1* microsatellite was substantially more polymorphic than *CAV1* in both populations, where the number of GGAA motifs ranged from 16-60 and 14-72 in Europeans and Africans, respectively. Additionally, while the characteristics of the *CAV1* promoter microsatellites were similar across both populations, the *NR0B1* microsatellite in African subjects was

significantly larger, harboring more repeat motifs, a greater number of repeat segments, and longer consecutive repeats, than in European subjects. The vast majority (>85%) of European *NR0B1* microsatellites were tightly clustered around smaller repeats ranging from 16-26 GGAA motifs, whereas 40% of African microsatellites were characterized by large, multi-segment repeats ranging from 30-72 GGAA motifs (Beck et al., Cancer Genetics, *in press*). These results were opposite to our original hypothesis, but considering the transcriptional implications of an increasing number of GGAA motifs in these EWS/FLI response elements, these results provoke several biologically intriguing hypotheses: It is possible that the massive *NR0B1* microsatellites commonly observed in Africans do not permit a stoichiometrically favorable environment for EWS/FLI binding and are therefore protective of EWS/FLI-mediated *NR0B1* gene activation. Alternatively, these large repeats may facilitate a toxic level of NR0B1 expression and permit premature cellular termination in the presence of EWS/FLI. It is also possible that the increased number of GGAA motifs observed in Africans has no influence on Ewing sarcoma susceptibility but instead supports an enhanced oncogenic potential of affected cells, contributing to the lower survival rates observed in African populations.

Polymorphisms of the *NR0B1* GGAA microsatellite have been observed across the various Ewing sarcoma cell lines, ranging from 16 – 26 motifs, which approximates the distribution repeats observed in Europeans. *NR0B1* mRNA levels in the various cell lines is tightly correlated with the number of GGAA motifs[133]. Based on this information, it is possible that EWS/FLI has preference for a narrow range of GGAA repeats in the *NR0B1* microsatellite, a so-called "sweet spot" with a GGAA-configuration conducive to

maximal EWS/FLI-mediated gene up-regulation. Given the highly polymorphic nature of the *NR0B1* microsatellite within and across ethnically distinct populations, functional assessment of these massive repeats is needed. Correlating polymorphisms of the *NR0B1* GGAA microsatellite in tumor samples with clinical parameters such as overall survival, metastatic burden, anatomic location and chemosensitivity may provide valuable information and lend to the development of GGAA microsatellite polymorphisms as prognostic biomarkers in Ewing sarcoma.

Conclusions

Chromosomal translocations are common molecular events in cancer, often producing novel fusion proteins with oncogenic properties. EWS/ETS chimeras in Ewing sarcoma are prototypical fusion products with unique DNA binding and regulatory properties responsible for tumorigenesis. A fascinating emergent property of EWS/ETS chimeras is their ability to directly modulate gene expression and the local chromatin environment via a tetra-nucleotide, GGAA microsatellite. This not only highlights how chimerism vastly alters the biological attributes of involved ETS-members, but also brings to attention a completely unappreciated role of microsatellite DNA in oncogenic transcriptional regulation. GGAA microsatellites have enabled the identification of novel target genes and have become important molecular tools in Ewing sarcoma research. These GGAA microsatellites are also highly polymorphic in human populations, and given that EWS/ETS-mediated gene expression is highly dependent on the length of these repetitive elements, GGAA microsatellite polymorphisms may also provide a unique

opportunity to improve our mechanistic understanding of disease susceptibility and prognosis in Ewing sarcoma. Certainly, in Ewing sarcoma, elements once regarded as "genomic junk," are proving to play a fundamental role in EWS/ETS-mediated oncogenesis.

Acknowledgments

All listed authors acknowledge support from the Huntsman Cancer Institute via P30CA042014 and the Terri Anne Perine Sarcoma fund. MJM acknowledges support from the Orthopaedic Research and Education Fund (OREF); AHG is supported by the R.L. Kirschstein NRSA (2T32HL007576-26) from the NHLBI; and SLL acknowledges support from the NIH/NCI via R01 CA140394.

Chapter 3: Clinical and biochemical function of polymorphic *NR0B1* GGAA microsatellites in Ewing sarcoma: a report from the Children's Oncology Group.

Michael J. Monument, Kirsten M. Johnson, Elizabeth McIlvaine, Lisa Abegglen, W. Scott Watkins, Lynn B. Jorde, Richard B. Womer, Natalie Beeler, Laura Monovich, Elizabeth R. Lawlor, Julia A. Bridge, Joshua D. Schiffman, Mark D. Krailo, R. Lor Randall, and Stephen L. Lessnick. (2014) *PLoS ONE*. 9(8): e104378. Doi: 10.1371/journal.pone.0104378

KMJ and MJM designed and performed all experiments. MJM wrote the manuscript, and KMJ reviewed document.

Abstract

Background: The genetics involved in Ewing sarcoma susceptibility and prognosis are poorly understood. EWS/FLI and related EWS/ETS chimeras upregulate numerous gene targets via promoter-based GGAA-microsatellite response elements. These microsatellites are highly polymorphic in humans, and preliminary evidence suggests EWS/FLI-mediated gene expression is highly dependent on the number of GGAA motifs within the microsatellite.

Objectives: Here we sought to examine the polymorphic spectrum of a GGAA-microsatellite within the *NR0B1* promoter (a critical EWS/FLI target) in primary Ewing

sarcoma tumors, and characterize how this polymorphism influences gene expression and clinical outcomes.

Results: A complex, bimodal pattern of EWS/FLI-mediated gene expression was observed across a wide range of GGAA motifs, with maximal expression observed in constructs containing 20-26 GGAA motifs. Relative to white European and African controls, the *NR0B1* GGAA-microsatellite in tumor cells demonstrated a strong bias for haplotypes containing 21-25 GGAA motifs suggesting a relationship between microsatellite function and disease susceptibility. This selection bias was not a product of microsatellite instability in tumor samples, nor was there a correlation between *NR0B1* GGAA-microsatellite polymorphisms and survival outcomes.

Conclusions: These data suggest that GGAA-microsatellite polymorphisms observed in human populations modulate EWS/FLI-mediated gene expression and may influence disease susceptibility in Ewing sarcoma.

Introduction

Ewing sarcoma is a prototypical chromosomal translocation-associated malignancy, in which virtually all cases harbor a balanced somatic translocation fusing the *EWSR1* gene (EWS) to a member of the (E- twenty six) ETS-family of transcription factors, most commonly *FLI1* (FLI)[5,113]. In fact, EWS/FLI and related EWS/ETS fusions are considered pathognomonic for the diagnosis of Ewing sarcoma. The EWS/FLI chimera product is a potent oncogenic transcription factor, characterized by fusion of a transcriptional-regulatory domain of EWS to the DNA binding domain of FLI[113].

EWS/FLI is considered the master-regulator of oncogenesis in Ewing sarcoma, regulating numerous critical gene targets necessary for oncogenic transformation[114,129].

Genome-wide localizations studies utilizing ChIP-seq and ChIP-chip strategies have identified many direct EWS/FLI targets. A remarkable observation derived from these studies was a previously unrecognized affinity of the EWS/FLI chimera for a repetitive GGAA-microsatellite element embedded within promoter/enhancer regions of numerous upregulated gene targets[15,46,128,133]. Forty to fifty percent of genomic EWS/FLI binding sites are associated with these GGAA-microsatellites[128] and EWS/FLI-mediated DNA binding and gene expression is dependent on these repetitive GGAA response elements[15,46,51]. These findings collectively demonstrate an unprecedented link between microsatellite DNA and transcriptional dysregulation in Ewing sarcoma.

Microsatellite DNA tracts represent ~3% of the human genome and are commonly located in non-coding extra-genic regions[167]. The repetitive nature and non-coding position of these elements allows microsatellite DNA to experience a higher baseline mutational rate than coding DNA. Consequently, these genetic elements are highly polymorphic at both an individual and population level[168]. Recently it has been shown that the GGAA-microsatellites within two critical upregulated EWS/FLI-target genes (*NR0B1* and *CAV1*) are highly polymorphic in healthy human subjects. Notably, significant length-dependent differences were observed comparing the *NR0B1* GGAA-microsatellite in white European and African populations[169]. This is significant as the

incidence of Ewing sarcoma is 10-fold less in African populations compared to white Europeans, irrespective of geographic location, suggesting a likely genetic influence[162]. Furthermore, *NR0B1* is among the most highly upregulated EWS/FLI targets and is essential for oncogenesis in Ewing sarcoma[39,133].

Initial studies characterizing the biochemical properties of these GGAA-microsatellite response elements demonstrated EWS/FLI DNA binding and subsequent transcriptional activation is highly dependent on the number of GGAA motifs within in the microsatellite: A minimum of 4 GGAA motifs is required for initial DNA binding, and gene expression markedly increases in a length-dependent manner with additional GGAA motifs[15,51,170]. Importantly, these early biochemical studies only characterized the relationship of EWS/FLI DNA binding and gene expression over a small and narrow range of 1-11 GGAA motifs. It remains unclear how the substantially larger spectrum of GGAA-microsatellite polymorphisms, observed in human populations influences EWS/FLI-mediated transcriptional activity. The goal of the present study was to characterize the polymorphic spectrum of the *NR0B1* GGAA-microsatellite in Ewing sarcoma tumors, define the biochemical properties of these GGAA length polymorphisms and to determine whether clinical outcomes are influenced by variations in these genetic elements.

Materials and Methods

Ethics statement

This study was approved by the University of Utah, Office for Research Integrity and Compliance prior to commencement. All patients enrolled in the Children's Oncology Group (COG) study AEW0031 (or legal guardians) provided written informed consent prior to study enrollment, which included the use of patient samples and tissues for molecular studies. All patient samples analyzed in the present study were de-identified and re-identification of samples was strictly reserved for the COG Statistics and Data Center to perform the appropriate clinical outcomes analysis. This study was carried out in accordance with the Declaration of Helsinki.

Patient samples

Ewing sarcoma tissue samples were obtained from the Biopathology Center (Columbus, OH), which serves as the specimen bank for the Children's Oncology Group. Patient demographics such as age, sex and race were self-reported by the patient (or legal guardians) at the time of study enrollment. Patients were instructed to identify their race as Caucasian, African American, Asian, Pacific Islander, American Indian or other. DNA from these tissue samples was extracted from OCT embedded tissue blocks or snap frozen tumors courtesy of Dr. Julie Bridge (University of Nebraska Medical Center, Omaha, NE). Approximately 20 nanograms of extracted genomic DNA also were commercially amplified using Qiagen's REPLI-g service (Qiagen Genomic Services, Hilden, Germany) for whole genomic amplification (WGA).

A second cohort of 20 Ewing sarcoma tumor samples and matching bone marrow aspirates collected at our local institute were also obtained. Tissues were stored in FFPE blocks and 5-micron scrolls were cut from each block in triplicate. DNA was extracted using the RecoverAll^{Tm} Total Nucleic Acid Isolation Kit (Life Technologies, Carlsbad, CA).

PCR sequencing

Forward and reverse primers, flanking the *NR0B1* GGAA-microsatellite loci were designed using promoter sequences obtained from the University of California Santa Cruz Human Genome Browser (http://genome.ucsc.edu/cgi-bin/hgGateway). All polymerase chain reaction (PCR) amplifications were performed using Pfx polymerase (Invitrogen, Grand Island, NY) in accordance with established laboratory protocols for microsatellite DNA. Each 25 mL PCR reaction consisted of 40-80 ng of genomic DNA, 0.3 mM of forward and reverse primers, 1U of Pfx polymerase, 0.8 mM of each deoxyribonucleotide triphosphate, 1X Pfx buffer and 1X Pfx enhancer solution. PCR products were subcloned into competent DH5a E. coli, with each bacterial colony representing an individual PCR-amplification clone. Twelve clones for each subject were selected and commercially sequenced (Beckman Coulter Genomics, Danvers, MA).

PCA analysis

NR0B1 GGAA-microsatellite sequence data for all samples were aligned using clustalx2. Because computational methods perform poorly on repetitive sequence, manual refinement was also necessary. Alignments in the repetitive regions were anchored on eight different single nucleotide adenosine residues that partition the contiguous GGAA repeats from the largest observed GGAA-microsatellite. The first 29 and last 50 bases of each raw sequence file were considered non-repetitive. For each contiguous GGAA segment, the number of GGAA repeats was counted and the count of the base differences between each non-repetitive region and the consensus sequence was determined (gap weight = 0.25). The pairwise distances between haplotypes were calculated as the squared Euclidian distance based on the 11 variable segments. Principal components analysis was performed using the MATLAB software package (The Mathworks, Natick, MA).

Luciferase Experiments

The pGL3 promoter luciferase vector (Promega, Madison, WI) was used for all experimental and control conditions. Human-derived *NR0B1* GGAA-microsatellite polymorphisms or synthetic GGAA constructs were cloned directly upstream of the SV40 minimal promoter element. 293EBNA cells were transfected with experimental reporter plasmid constructs or control plasmids, the Renilla plasmid and plasmids with and without EWS/FLI cDNA. Firefly luciferase activity was normalized to *Renilla* luciferase activity to control for transfection efficiency. Each experimental condition was performed in triplicate. Two-tailed Student's *t* tests were used for statistical comparisons.

Quantitative reverse-transcriptase polymerase chain reaction

Total RNA from established Ewing sarcoma cell lines (A673, COG-E-352, RDES, TC71 and SKES1)[171–174] and 293EBNA cells was amplified and detected using SYBR green fluorescence for quantitative analysis[175]. Normalized fold *NR0B1* expression in each of the Ewing sarcoma cell lines was calculated by determining the fold-change of each cell line relative to 293EBNA cells (negative control), with the data in each condition normalized to an internal housekeeping control gene *RPL21*. All experiments were performed in triplicate. Two-tailed Student's *t* tests were used for statistical comparisons.

Microarray data

Total RNA was extracted from fresh-frozen tumor specimens using miRNAeasy columns (Qiagen). RNA was then processed and hybridized to Affymetrix HuEx 1.0 arrays in the Genome Core at Children's Hospital Los Angeles according to standard Affymetrix protocols. Data for core probeset regions were quantile-normalized using robust multi-chip averaging in the Partek Genomics Suite software platform (Partek, St. Louis, Mo). *NR0B1* transcript level data were derived from normalized exon data using median summarization. Two-tailed Student's *t* tests were used for statistical comparisons.

Clinical outcomes analysis

Biological specimens were obtained from tissue submitted with consent for banking from eligible patients enrolled on COG study AEWS0031[176]. The primary objective of that

trial was to compare two chemotherapy regimens with respect to risk for an analytic event (EFS). Enrollment of 4.5 years with an additional year of follow-up provided for the detection of a hazard ratio (HR) of 0.64 in the failure rate with a probability of 0.80 when using a two-sided test with size 0.05. Four instances of interim monitoring were planned.

The primary study endpoint was event-free survival (EFS) defined as the time from entry into the study until the occurrence of an event (disease progression, second malignant neoplasm, or death) or until the last contact with the patient, whichever came first. Patients who did not experience an event by the time of last contact were considered censored for EFS-event. The method of Kaplan and Meier was used to estimate the probability of an event as a function of time since enrollment. Equality of risk for EFS-event across the various *NR0B1* GGAA-microsatellite haplotypes was assessed using the log-rank test[177]. All p-values are calculated using the chi-squared approximation and are therefore two-sided. EFS was assessed separately in males and females. Patient sex was not associated with risk for events in AEWS0031[176].

Results

Primary Ewing sarcoma tumor specimens

Ewing sarcoma tissue samples were obtained from the Biopathology Center (Columbus, OH), which serves as the specimen bank for the Children's Oncology Group (COG). All tumor samples were from patients with a pathologically confirmed diagnosis of primary

Ewing sarcoma who were enrolled in a large multicenter COG protocol, AEWS0031[176,178]. AEWS0031 was opened for enrollment on May 2001 and closed in August 2005. Data current through March 2009 (7.8 years after first enrollment) were used in this analysis. Patients presenting with clinically detectable metastatic disease were excluded. As part of protocol AEWS0031, enrolled patients were prospectively randomized into two different treatment arms: one group received the standard chemotherapy dosing schedule (cycles every 21 days) while the other group received interval compressed dosing (cycles every 14 days) of the same chemotherapeutic regimen, consisting of vincristine (2 mg/m^2), doxorubicin (75 mg/m^2), and cyclophosphamide (1.2 g/m^2) alternating with ifosfamide (9 g/m^2) and etoposide (500 mg/m^2). All other study protocols were standardized. Of 568 patients enrolled in AEWS0031, snap frozen (0.004-0.06 gram) or optimal cutting temperature (OCT) compound-embedded tissue (50-60 micron thickness) was available from 117 patients. Of this group, 5 patients were excluded: one patient was represented in duplicate, two patients were determined to have a final tissue diagnosis other than Ewing sarcoma, one patient presented with metastatic disease and one patient could not be properly identified due to a presumed clerical error. The final analytic data set included 112 patients (Figure 3.1). Ninety-percent (101/112) of patients were identified as Caucasian and only 2% (2/112) were identified as African American. The demographic characteristics of the included and excluded AEWS0031 patients were comparable (Table 3.1).

	Evaluated Patients		Not Evaluated		
Demographic	**n**(112)	**%**	**n**(456)	**%**	p value*
Age					0.04
< 9	38	34%	124	27%	
10-17	68	61%	271	59%	
>17	6	5%	61	14%	
Sex					0.8
Male	62	55%	246	56%	
Female	50	45%	210	46%	
Race					0.9
White European	101	90%	401	88%	
African American	2	2%	12	3%	
Other	3	3%	19	4%	
Missing	6	5%	24	5%	
Primary Tumor Site					0.4
Appendicular	44	39%	151	33%	
Thoracic	16	14%	73	16%	
Pelvic	20	18%	70	15%	
Other Axial	9	8%	66	14%	
Extraosseous	23	21%	96	21%	

Table 3.1 Patient demographics of included and excluded AEWS0031 patients
*Fisher's exact test

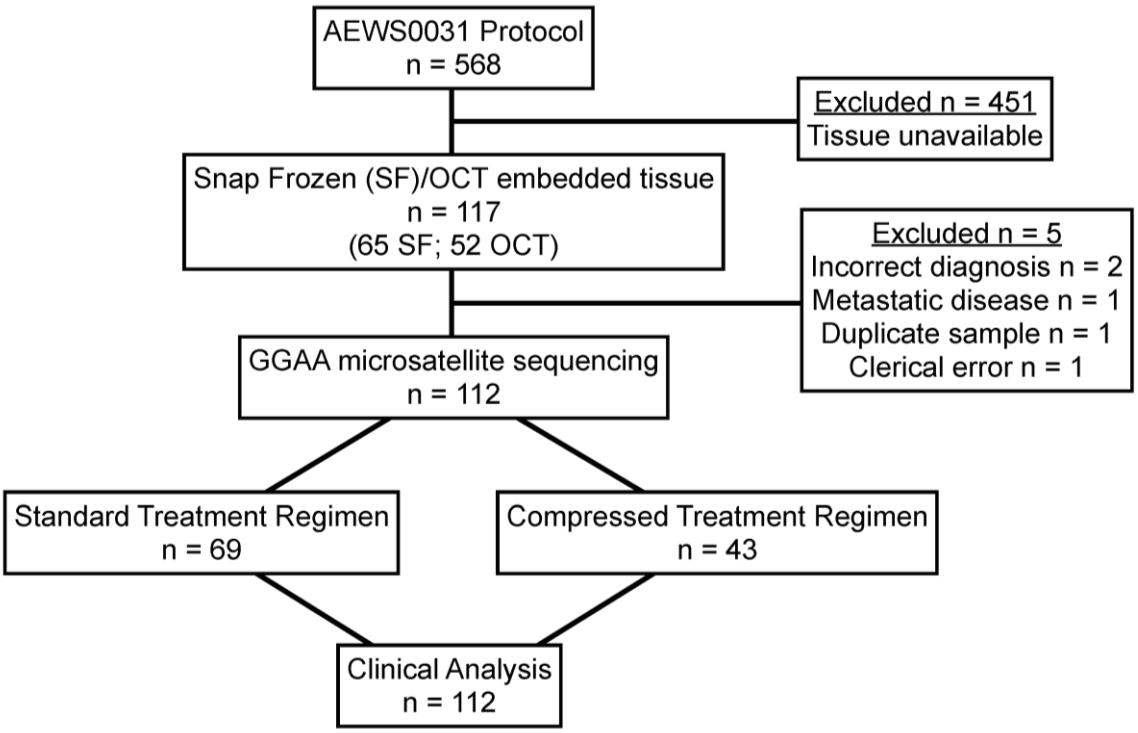


Figure 3.1 Flow diagram of COG study AEWS0031 patient samples included for GGAA-microsatellite sequencing and clinical analysis.

The NR0B1 GGAA-microsatellite is highly polymorphic in Ewing sarcoma tumors and significantly different than white European controls

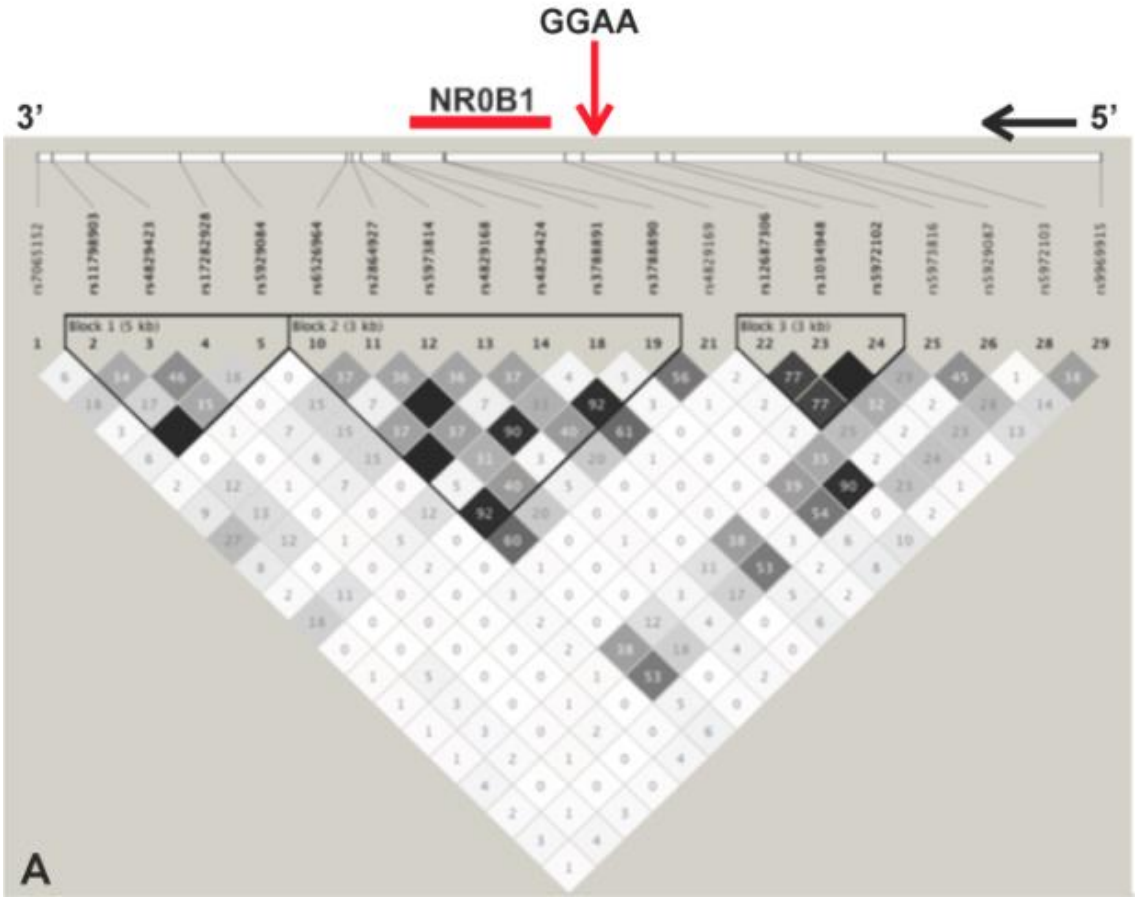
We have previously evaluated the polymorphic spectrum of three GGAA-microsatellite containing direct EWS/FLI targets: *NR0B1*, *CAV1* and *GSTM4*[169]. The GGAA microsatellites at these loci are polymorphic in human populations, although *NR0B1* was the most polymorphic loci with significant differences observed between African and Caucasian populations. Given the markedly different incidence of Ewing sarcoma in these populations and the role of the NR0B1 protein in sustaining the oncogenic phenotype of Ewing sarcoma, we elected to focus on *NR0B1* for this study. The *NR0B1*

GGAA-microsatellite, chrX:30328826 to chrX:30329008 (http://genome.ucsc.edu/cgi-bin/hgTracks; GRCh37/hg19) was amplified, cloned and sequenced in all 112 primary tumor samples. A subcloning strategy was used to sequence all microsatellites, ensuring in heterozygous patients that both alleles were accurately identified. A total of 143 haplotypes were identified, which was expected given 45% of the 112 patients were female. Sequence data were compared to a previously established data set of healthy African and white European controls[169]. It should be noted that in AEWS0031, white, non-Hispanic patients were classified as *Caucasian* and in the aforementioned data-set by Beck et al.[169], white, non-Hispanic subjects of northern European decent are referred to as *European*. For the purpose of clarity, in the present report all white, non-Hispanic patients are reported as *white Europeans*.

The *NR0B1* GGAA-microsatellite is located within the promoter region, roughly 1.5kb upstream of the transcriptional start site. This polymorphic microsatellite ranges in length from 80-240bp and is located within a defined haplotype block (International HapMap project[179], CEU reference population; Figure 3.2A). The GGAA-microsatellite is characterized by a series of contiguous GGAA motifs partitioned by a single adenosine base substitution (Figure 3.2B). Variability exists not only in the total number of GGAA motifs, but also in the number of contiguous segments and the number of GGAA motifs in each contiguous segment. In Ewing sarcoma tumors, *NR0B1* microsatellites ranged in size from small, two-segment repeats containing 16 GGAA motifs to larger multisegment repeats containing up to 61 GGAA motifs. The most frequent haplotype observed in

tumor samples was an intermediate sized, 3-segment microsatellite containing 24 GGAA motifs. A comparison of the pertinent microsatellite sequence characteristics in tumor samples and control white European and African populations is presented in Table 3.2. The descriptive statistical analyses demonstrate that the mean values for total number of GGAA motifs and longest consecutive GGAA segment in the tumor dataset were similar to that of the African data set. The white European dataset had lower mean values for the total number of GGAA motifs compared to both Africans and tumors. Raw sequence data of all included subjects is listed in supplemental Table 3.1 (published online[180]).

Figure 3.2 GGAA-microsatellite organization at the *NR0B1* locus.



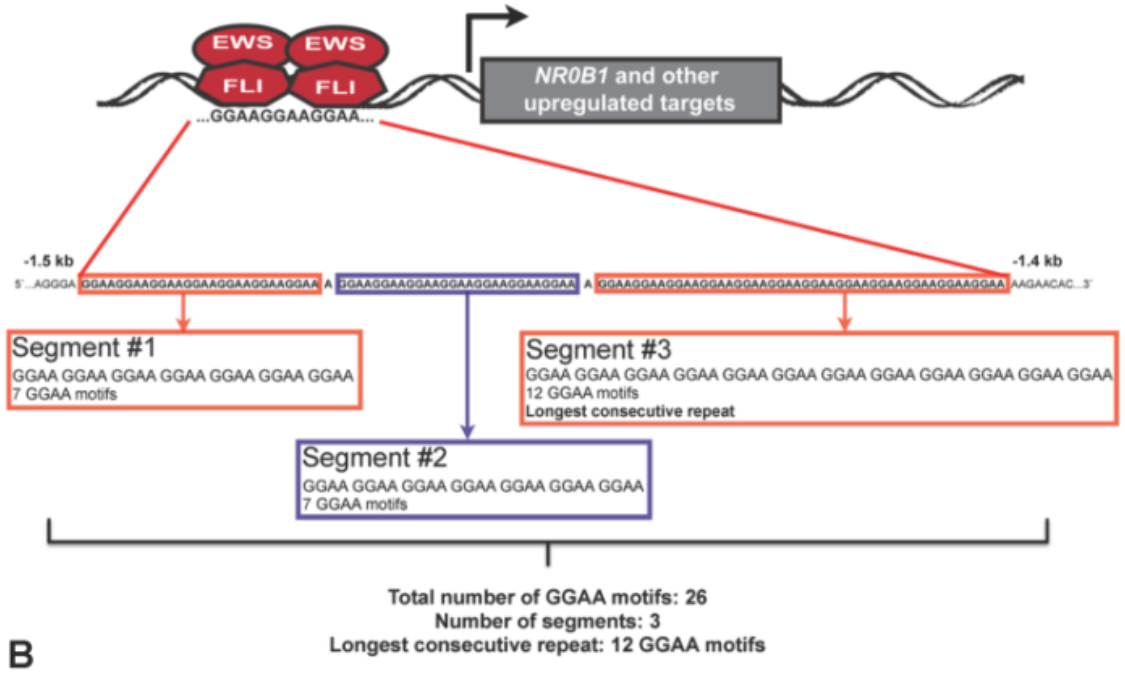


Figure 3.2 GGAA-microsatellite organization at the *NR0B1* locus.

(A) Using available single nucleotide polymorphism (SNP) data from the CEU reference population (northern and western European decent) of the International HapMap Project[179], the *NR0B1* GGAA-microsatellite is identified within a defined haplotype block. **(B)** For the *NR0B1* locus, the GGAA-microsatellite is located approximately 1.5kb upstream of the transcriptional start site (TSS) and is characterized a variable number of contiguous GGAA motifs, partitioned by single adenosine base substitutions**.** Sequence characteristics of interest include the total number of GGAA motifs, the total number of contiguous segments and longest consecutive GGAA segment. Figure panel adapted from[169].

Given that 90% of the Ewing sarcoma patients analyzed in this study were white Europeans, we sought to determine if the spectrum of *NR0B1* GGAA-microsatellite haplotypes in tumor samples were similar to a previously established control white European data set. To assess this a principal components analysis (PCA) was performed combining the raw *NR0B1* GGAA-microsatellite sequences from both tumor and control

white European data sets (Figure 3.3A). Repetitive regions of each sequence were manually aligned and GGAA repeat motifs were anchored by the single nucleotide adenosine residues that partition the contiguous GGAA repeat units observed in the largest haplotypes. For each GGAA track, the number of GGAA repeats units were counted. The counts of base differences between the flanking non-repetitive regions and the consensus sequence were also determined (gap weight = 0.25). Using this analysis, three distinct haplotype clusters were observed in tumors, which closely overlapped the distribution of haplotypes observed in the white European control data set.

	Average total number of GGAA motifs*	Most common number of GGAA motifs	Average longest consecutive GGAA segment*	Most common longest consecutive GGAA segment
Ewing sarcoma tumor samples	30 ± 14 Range: 16-61	24	11 ± 1 Range: 8-16	10
White European	24 ± 11 Range: 16-60	24	11 ± 1 Range: 8-16	11
African	32 ± 15 Range: 14-72	24	12 ± 1 Range: 8-21	11

Table 3.2 NR0B1 GGAA-microsatellite sequence characteristics in Ewing sarcoma tumors and healthy controls
*Mean values ± standard deviation

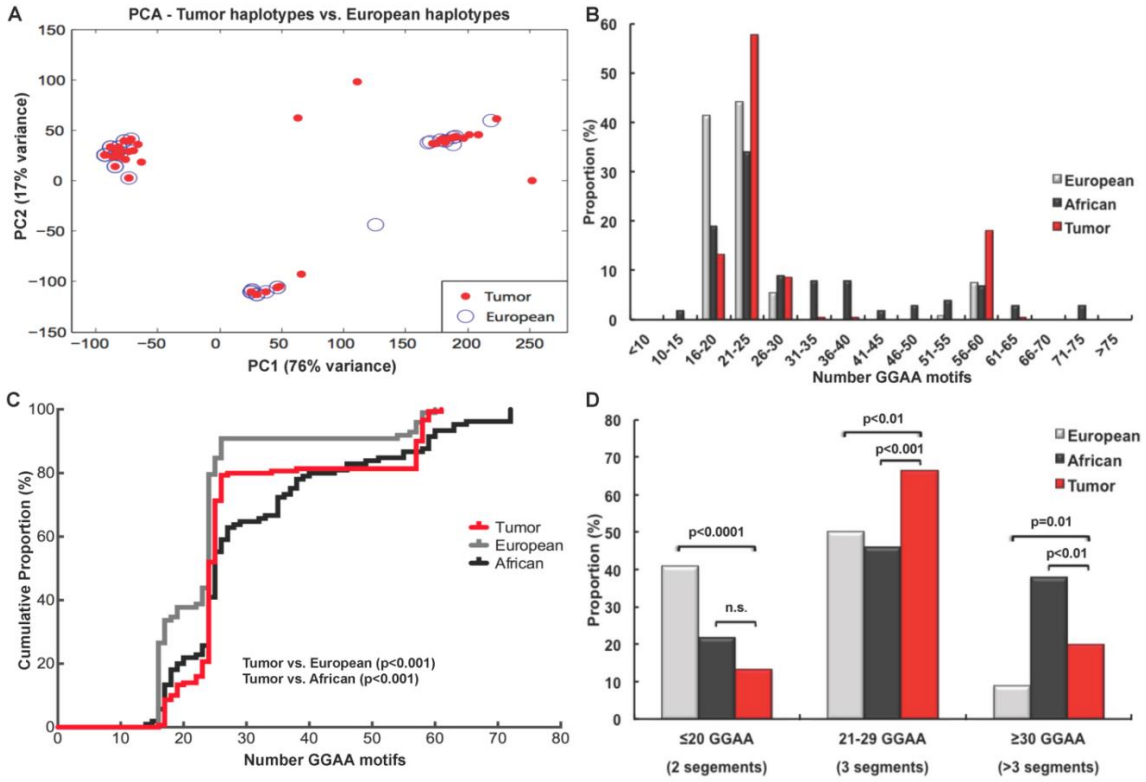


Figure 3.3 *NR0B1* GGAA-microsatellites are polymorphic in Ewing sarcoma tumors with an allelic distribution different than that of white European and African controls.

(A) Principal components analysis comparing unique microsatellite haplotypes in tumor samples and white European controls demonstrate three principal sequence clusters, with a high-degree of overlap between the two populations. **(B)** Histogram plots comparing the distribution frequency of *NR0B1* GGAA-microsatellite haplotypes in tumors and white European and African controls. Despite the overlapping PCA analysis, an enrichment of haplotypes containing 21-25 and 56-60 GGAA motifs was observed in tumor samples. Relative to white Europeans, a depletion of haplotypes containing 16-20 GGAA motifs was also noted in tumors. **(C)** Cumulative density plots for each study population similarly demonstrate the enrichment of haplotypes containing 21-25 and 56-60 GGAA motifs in tumors. The distribution of these haplotypes in tumors is significantly different from both white Caucasian and African populations (KS test, $p<0.001$). **(D)** Stratifying haplotypes according to the major sequence types identified in the PCA demonstrates that intermediate (3 segment) GGAA-microsatellites are more enriched in tumors and larger multisegment haplotypes (>3 segments) were also more enriched compared to white Europeans, although markedly less than Africans. Control white European and African population data from[169].

In contrast to the descriptive values reported in Table 3.2 and the PCA analysis, which examined relationships between unique haplotypes, when the frequency of GGAA-microsatellite haplotypes was plotted as a function of the total number of GGAA motifs, striking differences were observed (Figure 3.3B). Most notably, a strong enrichment for haplotypes containing 21-25 GGAA motifs was observed in the tumor population: 81/143 (58%) of tumor haplotypes, compared to 46/104 (44%) and 36/106 (34%) white European and African haplotypes, respectively, contained 21-25 GGAA motifs ($p=0.03$ and $p<0.001$, Chi-square), respectively. A second enrichment was also observed for tumor haplotypes containing 56-60 GGAA motifs: 27/143 (19%) of tumor haplotypes, compared to 8/104 (8%) and 7/106 (7%) of white European and African haplotypes, respectively ($p=0.03$ and $p=0.01$, Chi-square). Additionally, relative to white European controls, a depletion of haplotypes containing 16-20 GGAA motifs was also observed in tumors. The enrichment of tumor haplotypes containing 21-25 GGAA motifs, and the depletion of haplotypes with 16-20 GGAA motifs, contributes to the similar descriptive statistics shown in Table 3.2, despite the statistically different distribution of these data when more sophisticated techniques are used.

To circumvent some of the inherent bias associated with the arbitrary binning of data, a cumulative density function was performed for each population (Figure 3.3C). This figure recapitulates the trends observed in Figure 3.3B, showing a strong enrichment of haplotypes containing 23-26 GGAA motifs in tumor samples, while the European density function is represented by a larger shoulder at smaller GGAA haplotypes (16-20 GGAA

motifs) and a similar, although lower amplitude peak in the 23-26 GGAA range. The African density curve is more diffusely populated throughout the spectrum of GGAA motifs. Using a Kolmogorov-Smirnov test[181] to evaluate the haplotype distributions based on the total number of GGAA motifs across all three populations, the tumor data set was statistically dissimilar from both white European ($p< 0.001$) and African ($p<0.001$) populations (Figure 3.3C).

Using a slightly different approach, sequence data from all three populations was stratified based on the 3 major haplotype categories identified in the PCA analysis: 2 segment repeats with ≤20 GGAA motifs, 3 segment repeats with 21-29 GGAA motifs and a larger segmental repeats (4-8 segments) with ≥30 GGAA motifs (Figure 3.3D). Relative to both white European and African control populations, haplotypes containing 21-29 GGAA motifs were statistically over-represented, while haplotypes ≤ 20 GGAA motifs were under-represented in the tumor population ($p = 0.03$ and $p<0.0001$, respectively). These data demonstrate that the distribution of polymorphic *NR0B1* GGAA-microsatellite haplotypes in tumor samples were markedly different than white European populations, which is compelling given the higher incidence of Ewing sarcoma in white non-Hispanic patients of European descent. Such observations may represent a previously unidentified pattern of genetic susceptibility in Ewing sarcoma.

GGAA-microsatellites are genomically stable throughout oncogenesis and after whole genome amplification

Given the non-overlapping distribution of the *NR0B1* GGAA-microsatellite haplotypes in tumor samples compared to white European controls, we sought to determine if this difference could be attributable to microsatellite instability during the process of oncogenic transformation. Microsatellite instability has been observed in various other cancers, including sarcomas, although most commonly occurring at mono- and dinucleotide microsatellite loci[151,182,183]. To address this question, genomic DNA was extracted from 20 locally-archived primary or metastatic Ewing sarcoma FFPE tissue blocks, and the *NR0B1* GGAA-microsatellite sequence characteristics were compared to matched germ line DNA isolated from bone marrow aspirates. There was no evidence of microsatellite instability in any sample (Table 3.3). Microsatellite DNA stability is inversely proportional to the length of the microsatellite tract[167] and therefore it was important to assess the stability of the larger *NR0B1* GGAA haplotypes. There was no evidence of microsatellite instability in any of the haplotypes containing 55-60 GGAA motifs (n=4).

Patient ID	Tumor	Germline GGAA	Tumor GGAA	Alignment
EWS 17	Metastatic	25/57	25/57	Concordant
EWS 19	Primary	17/25	17/25	Concordant
EWS 22	Primary	25	25	Concordant
EWS 24	Primary	17	17	Concordant
EWS 29	Primary	24	24	Concordant
EWS 36	Primary	25	25	Concordant
EWS 41	Metastatic	25	25	Concordant
EWS 43	Metastatic	24	24	Concordant
EWS 44	Metastatic	25	25	Concordant
EWS 45	Metastatic	24	24	Concordant
EWS 46	Primary	17	17	Concordant
EWS 58	Primary	24	24	Concordant
EWS 59	Primary	24	24	Concordant
EWS 61	Primary	25/57	25/57	Concordant
EWS 62	Primary	24	24	Concordant
EWS 106	Primary	24	24	Concordant
EWS 107	Primary	58	58	Concordant
EWS 115	Primary	23	23	Concordant
EWS 116	Primary	57	57	Concordant
EWS 119	Primary	24	24	Concordant

Table 3.3 Comparison of germline and tumor *NR0B1* GGAA-microsatellites

In a complementary series of experiments, we also sought to determine if the process of whole genome amplification (WGA) altered the composition of these GGAA-microsatellites. Given the limited availability of tumor tissue, relatively small reserves of DNA are available for molecular studies in Ewing sarcoma; WGA provides an opportunity to amplify DNA from precious biological samples. Genomic DNA from all 112 Ewing sarcoma samples was commercially amplified using Qiagen's Repli-g WGA service (Qiagen Genomic Services, Hilden, Germany) and GGAA-microsatellite characteristics were compared to unamplified DNA. Repli-g WGA utilizes multiple displacement amplification technology and provides a highly unbiased and complete coverage of the genome[184]. A minimum of 10 WGA amplified tumor samples with an *NR0B1* GGAA-microsatellite sequence for each major sequence category (<20, 20-30, 50-60 GGAA motifs) were sequenced and compared to the unamplified, original DNA source. GGAA-microsatellite sequences were unaltered by the WGA process in 10/10 (100%) and 13/13 (100%) of small (<20 GGAA motifs) and medium (20-30 GGAA motifs) microsatellites, respectively. In the largest microsatellites (50-60 GGAA motifs), sequences were a perfect match in only 4/12 (42%) cases. However, of the 5/7 discordant cases, the WGA sequence was incorrect by only a single GGAA motif. In 2/12 samples, the WGA product did not yield a sequencable amplicon (Figure 3.4). These data suggest that for small and medium sized GGAA-microsatellites, the WGA process yields highly concordant sequences, although it introduces minor sequence perturbations in larger repeats containing 50-60 GGAA motifs.

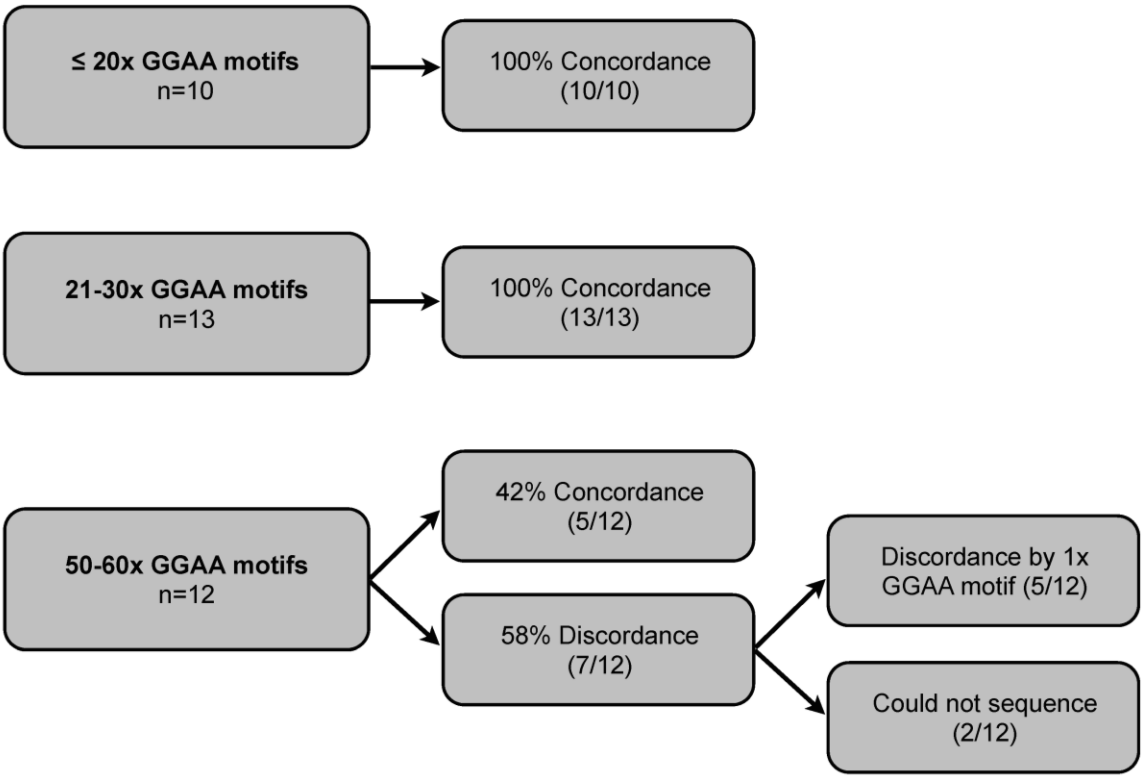


Figure 3.4 GGAA-microsatellites sequence characteristics after whole genome amplification (WGA). Microsatellites were sequences after WGA and compared to unamplified DNA.

A narrow range of GGAA motifs facilitates maximal EWS/FLI-mediated gene expression

Based on evidence from earlier studies, which initially characterized GGAA-microsatellites as EWS/FLI-response elements, it appeared that after a critical threshold of 4 GGAA motifs, DNA binding and subsequent *NR0B1* gene expression markedly increased with an increasing number of GGAA motifs. However, more recent data has demonstrated the polymorphic spectrum of these GGAA-microsatellites is well beyond the range tested in these earlier biochemical studies. To assess the potential length-dependent relationship between EWS/FLI-mediated gene expression and GGAA-

microsatellite polymorphisms, various polymorphic GGAA sequences identified in control populations ranging from 17-72 GGAA motifs were cloned into a luciferase reporter vector directly upstream of the SV40 minimal promoter element. 293 EBNA cells were co-transfected with the various experimental GGAA plasmids and a vector containing EWS/FLI or an empty vector control. All experiments were performed in triplicate and the luciferase data presented is a composite of two independent experiments.

In human-derived sequences (Figure 3.5A), a bimodal relationship of EWS/FLI-mediated gene expression across the spectrum of GGAA constructs investigated was observed. Gene expression was maximal in microsatellites containing 20-25 GGAA motifs, and values precipitously dropped in constructs ranging from 29-40 GGAA motifs followed by a second lesser peak in constructs ranging from 50-60 GGAA motifs. Relative to the 24 GGAA construct, the reduction in expression was maximal in constructs containing 17, 29 and 72 GGAA motifs (3-fold, 4.5-fold and 4.5-fold; $p< 0.0001$), respectively. Expression levels using the 58 GGAA construct were 1.5-fold less than the 24 GGAA construct ($p<0.001$).

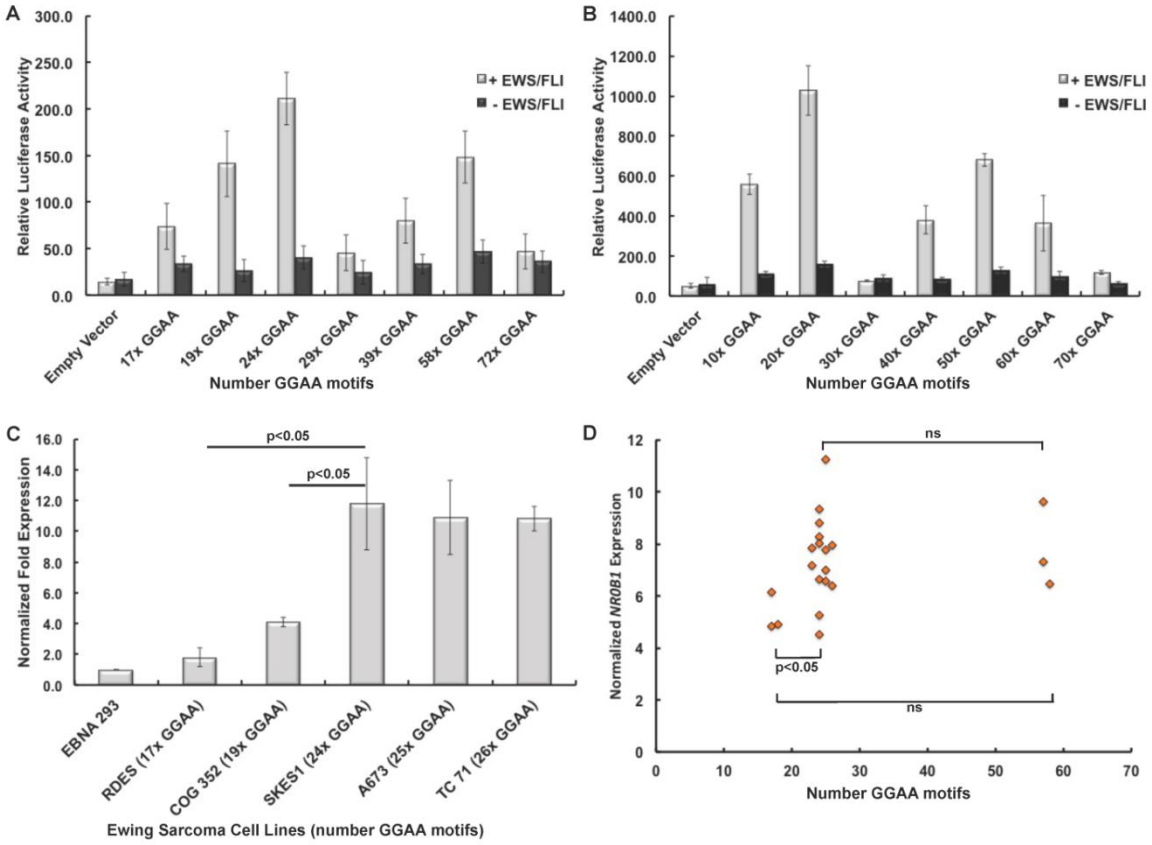


Figure 3.5 EWS/FLI-mediated gene expression is highly variable across various GGAA-microsatellite length polymorphisms.

(A) Polymorphic *NR0B1* GGAA-microsatellites from white European and African subjects were cloned into luciferase reporter vectors and co-transfected with EWS/FLI into 293 EBNA cells. A bimodal pattern of gene expression was observed, with greatest expression in constructs with 24 GGAA motifs and a lesser peak in constructs with 58 GGAA motifs. **(B)** A similar bimodal trend was observed using synthetic GGAA constructs identically cloned into the same luciferase reporter construct. **(C)** In patient derived Ewing sarcoma cell lines, RT-PCR quantified *NR0B1* mRNA expression was also maximal in cell lines containing an *NR0B1* microsatellite containing 24-26 GGAA motifs. **(D)** In primary Ewing sarcoma tumors, normalized *NR0B1* transcript levels were lowest in tumors with *NR0B1* GGAA-microsatellites containing 17-18 GGAA motifs, which was significant less than tumors with microsatellites containing 23-26 GGAA motifs (p=0.04).

The human-derived sequences contained varying combinations of single-base insertions and contiguous GGAA sequences, thus complicating the interpretation of these data (see

Figure 3.2B). To focus the evaluation on the overall length of the GGAA-microsatellite, synthetic GGAA-microsatellites constructs were synthesized, ranging from 10-70 contiguous GGAA motifs and cloned into the same luciferase vector in an identical fashion. Similar differences were observed in the assays using the synthetic constructs (Figure 3.5B). Additionally, average gene expression levels in the assays using the contiguous synthetic constructs were markedly elevated compared to the segmental constructs cloned from human DNA (458 ± 330 vs. 107 ±62, respectively, p=0.02). These trends suggest contiguous GGAA-microsatellites afford more optimal gene expression than partitioned repeats. Exemplifying this, gene expression in the smallest synthetic construct (10 GGAA) was 2.5-fold greater than the maximal expression observed from the partitioned 24 GGAA construct. Interestingly, the third segment of the 24 GGAA construct contains 10 contiguous GGAA motifs.

To assess the influence of these polymorphic GGAA-microsatellite response elements in a more native cellular context, *NR0B1* mRNA levels were quantified from various patient-derived Ewing sarcoma cell lines confirmed to be polymorphic at the *NR0B1* GGAA locus. Given the position of *NR0B1* on the X chromosome, to circumvent any issues associated with heterozygosity and potential X-linked inactivation, only cell lines either homozygous or hemizygous for a polymorphic *NR0B1* GGAA-microsatellite locus were included. Unfortunately none of the investigated cell lines were hemizygous or homozygous for a larger 50-60 GGAA motif allele; two cell lines (EWS502 and TC32) were heterozygous (20/58 GGAA and 24/58 GGAA, respectively), but a clear pattern of

allelic activation (or inactivation) could not be established for these cell lines. Figure 3.5C illustrates quantitative RT-PCR normalized expression levels of *NR0B1* mRNA transcripts relative to the number of GGAA motifs measured. Similar to the luciferase experiments, maximal gene expression was observed in cell lines ranging from 24-26 GGAA motifs. Negligible *NR0B1* levels were observed in the RDES cell line, which is hemizygous for a 17 GGAA-microsatellite. These results using a native cellular context strongly support the trends observed in both patient-derived and synthetic luciferase experiments of maximal gene expression in constructs containing 20-25x GGAA motifs.

When comparing the gene expression profiles (Figure 3.5) to the allelic distributions of the *NR0B1* GGAA-microsatellite sequenced from tumor samples (Figure 3.3), striking similarities are observed. Notably, the bimodal pattern of maximal gene expression and the amplitude of these peaks in constructs ranging from 20-25x and 56-60x GGAA motifs parallels the frequency and distribution of *NR0B1* GGAA-microsatellite in tumors.

To investigate if GGAA-microsatellite polymorphisms influenced *NR0B1* gene expression in Ewing sarcoma tumors, we quantified normalized *NR0B1* expression using microarray data from 31 Ewing sarcoma samples from which both PCR sequencing data and RNA were available. Ten of the 31 samples were heterozygous at the *NR0B1* GGAA-microsatellite locus, leaving 21 hemi- or homozygous tumors for analysis (Figure 3.5D). Consistent with the cell line data in Figure 3.5C, we observed lower normalized *NR0B1* expression levels in tumors containing only 17-18 GGAA motifs in their *NR0B1*

microsatellite. This was a statistically-significant diminished level as compared to tumors containing 23-26 GGAA motifs (p=0.04). Thus, the human tumor data is consistent with the cell line and *in vitro* luciferase studies.

Polymorphisms of the NR0B1 GGAA-microsatellite are not predictive of event free survival

Given the documented influence of GGAA length polymorphism on gene expression and mRNA levels, we next sought to determine if these polymorphisms influence tumor biology and clinical outcomes in patients with Ewing sarcoma. *NR0B1* is one the most highly upregulated, direct EWS/FLI targets, and expression of this gene is essential for transformation in Ewing sarcoma cell lines[43,133]. Clinical outcome data for at least 5 years (in surviving patients)[176] was available in all 112 samples used in the sequencing analysis. Of the 112 patients included in the analysis from AEWS0031, 69/112 were treated with the standard chemotherapy regimen as compared to 43/112 treated with compressed chemotherapy (Figure 3.1). It should be noted that the 5-year EFS was slightly improved in patients receiving compressed therapy (73% vs. 65%, p=0.048[21]). The aggregate outcome of patients who were considered in this analysis was similar to patients who were eligible for AEWS0031 but who were not included in the analysis (p = 0.21).

Given the biochemical data favoring optimal gene expression over a narrow range of GGAA motifs and the distribution of these haplotypes in Ewing sarcoma tumors, Kaplan-

Meier survival analyses were performed stratifying patients based on the presence or absence of *NR0B1* alleles containing 22-27 GGAA motifs (Figure 3.6). The presence of one or more *NR0B1* alleles containing 22-27 GGAA motifs did not influence EFS compared to patients without these alleles (Figure 3.6A and 3.6B). Given that females represented 45% of our cohort, numerous patients heterozygous for different length alleles were identified. Whether one or both of these alleles is active in tumor cells remains unclear; therefore, EFS was assessed separately in males and females (Figure 3.6C and 3.6D). In male and female patients, EFS was also not influenced by allele type. Furthermore, stratifying the patients based on assignment to standard vs. compressed chemotherapy arms also did not influence EFS survival based on allele type (Figure 3.6E and 3.6F). These results clearly demonstrate that despite biochemical data showing a strong relationship between GGAA-microsatellite length and gene expression levels, polymorphisms of the *NR0B1* GGAA-microsatellite do not influence clinically relevant outcomes.

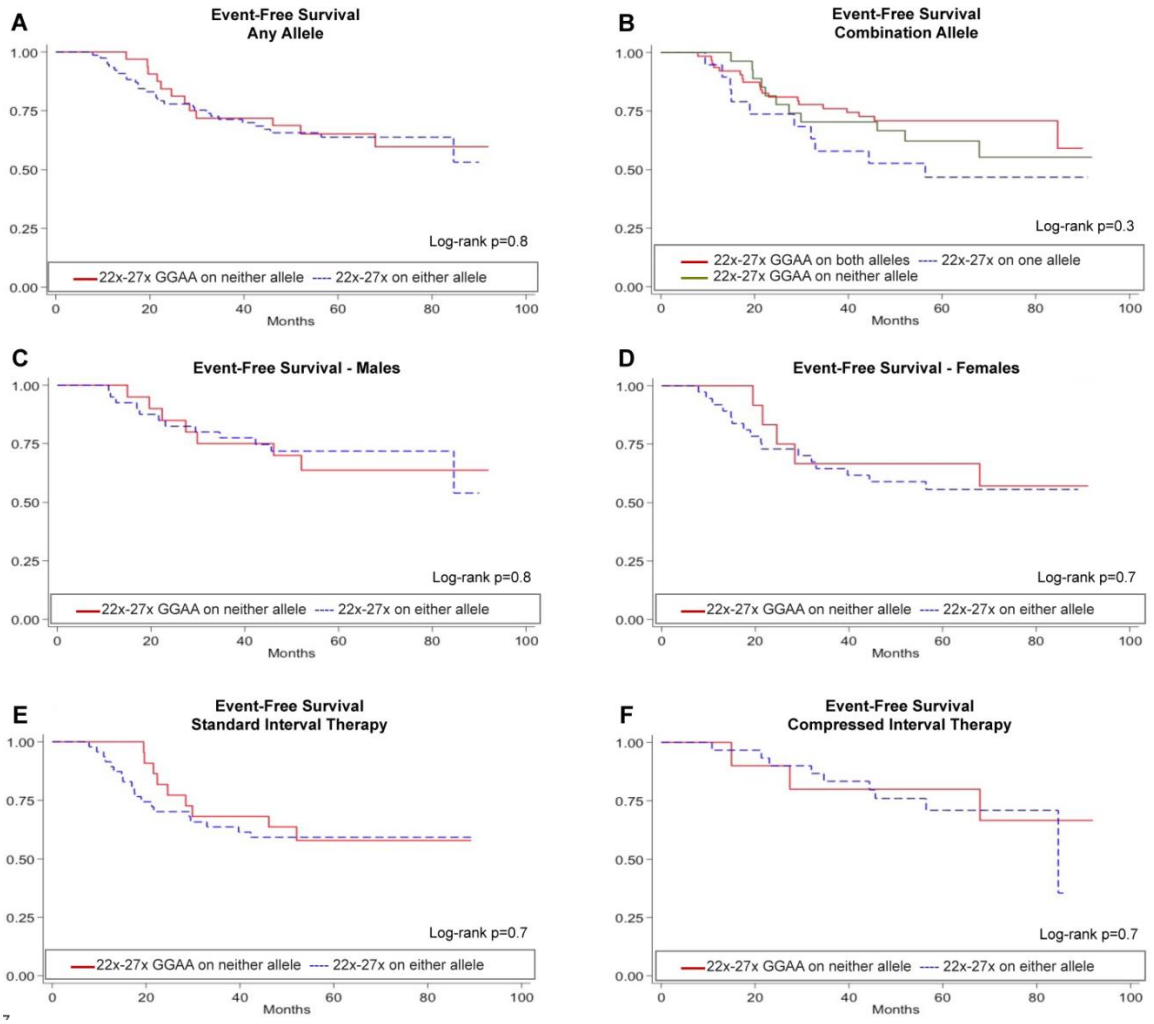


Figure 3.6 *NR0B1* GGAA-microsatellite polymorphisms do not influence event free survival (EFS) in Ewing sarcoma patients.

(A) EFS was compared in 112 patients from AEWS0031 based on the presence of absence of at least one *NR0B1* GGAA-microsatellite allele containing 22-27 GGAA motifs. This allele type was chosen based on the pattern of alleles present in tumor samples and the maximal EWS/FLI-mediated gene expression supported by alleles of this length category. **(B)** EFS was similarly assessed based on the presence of one or both alleles containing 22-27 GGAA motifs. Additional subgroup analyses were also performed in males **(C)** and females **(D)** and in patients receiving standard **(E)** or compressed **(F)** therapy.

Discussion

Transcriptional dysregulation via microsatellite DNA in Ewing sarcoma represents a fascinating and novel property of the EWS/FLI chimera. Microsatellite DNA is not subject to the same evolutionary pressures as coding DNA, rendering these sequences highly polymorphic across individuals and populations[165,168,169]. Furthermore, given that 40-50% of genomic EWS/FLI occupancy occurs at GGAA-microsatellites[128], these EWS/FLI-responsive elements provide a unique opportunity to examine Ewing sarcoma susceptibility and pathogenesis from an alternative genetic basis. In particular, our research group is interested in whether polymorphisms at transcriptionally important GGAA-microsatellites are biologically relevant in this context. In the present study, we have demonstrated that the *NR0B1* GGAA-microsatellite in primary Ewing sarcoma tumors is highly polymorphic, with an allelic distribution dissimilar from white European controls. Here we report the first series of biochemical experiments detailing the effect of GGAA-microsatellite polymorphisms on EWS/FLI-mediated transcriptional regulation demonstrating that the distribution of these *NR0B1* haplotypes in tumors is strongly biased towards a narrow range of microsatellite alleles that facilitate maximal EWS/FLI-mediated gene expression.

Traditionally viewed as "junk" DNA, microsatellite DNA is becoming increasingly recognized as an important cis-regulating genetic element[108,185]. The discovery of GGAA-microsatellites as a direct EWS/FLI-mediated transcriptional response element in Ewing sarcoma identified a novel function of microsatellite DNA in human cancer

development and a previously unrecognized ETS factor binding site[49]. We have demonstrated that across a large numeric range of GGAA motifs, EWS/FLI-mediated gene expression is highly variable. However, contrary to our preliminary understanding of these EWS/FLI-responsive elements, we did not observe a simple linear relationship of increasing gene expression as a function of an increasing number of GGAA motifs[15]. Instead, a bimodal relationship was observed. A mechanistic explanation for this bimodal relationship was not assessed in the present study, although similar findings have been observed in other model systems. For instance, in *Neisseria meningitides*, expression of a virulence factor, *NadA* is regulated by a polymorphic, promoter-based tetranucleotide microsatellite element with a similar pattern of transcript periodicity to that observed in our study[186,187]. The variations in *NadA* transcript levels were attributed to altered binding abilities of transcriptional cofactors across the various microsatellite polymorphisms[188].

The EWS/FLI chimera requires a minimum of 4 contiguous GGAA motifs (16bp) to effectively bind microsatellite DNA. Furthermore, EWS/FLI occupies these microsatellites in a ratio of 2 protein molecules for every DNA molecule in synthetic microsatellite constructs comprised of 4, 5, 6 and 7 contiguous GGAA motifs[51]. A potential explanation for the bimodal biochemical expression patterns observed in this study is that the stoichiometric occupancy of EWS/FLI and associated co-factors is most optimal across microsatellites containing 21-25 or 55-60 GGAA motifs. Another possibility is that certain GGAA-microsatellite polymorphisms are more (or less) likely to form inhibitory secondary DNA structures. Guanine-rich DNA sequences can

predispose to the formation of non-B-form DNA structures and G-quadruplexes[108,188], which may influence EWS/FLI and associated co-factor occupancy. Certainly, the results of the luciferase, cell line, and primary human tumor data detailed in the present study are compelling and warrant further investigations into the biochemical effects of GGAA content on EWS/FLI-mediated DNA binding in a native cellular and chromatin context.

The incidence of Ewing sarcoma in African populations is 10-fold less than that of white Europeans[162], but as of yet there is no concrete explanation for this difference[163,164,189]. The GGAA-microsatellite of two critical upregulated EWS/FLI-targets in Ewing sarcoma (*NR0B1* and *CAV1*) have been shown to be highly polymorphic in African and white European populations, with a predisposition for significantly larger GGAA-microsatellites in Africans, especially at the *NR0B1* locus[169]. This finding prompted further inquiry into the makeup of these elements in Ewing sarcoma tumor samples. Indeed, our results demonstrate that these GGAA elements are highly polymorphic in tumors, although the distribution of these haplotypes within primary tumors demonstrated compelling differences compared to both African and white European controls. The dissimilar distribution of tumor and white European control haplotypes is an important observation, given that 90% of the AWES0031 cohort was identified as white European. Our preliminary hypothesis was that the *NR0B1* GGAA-microsatellite sequence data set from white European controls would be very similar to the patient-derived tumor samples.

Compared to white European controls, a strong enrichment for *NR0B1* microsatellite haplotypes containing either 21-25 or 56-50 GGAA motifs and a bias against smaller alleles containing 17-20 GGAA motifs was observed in patient samples. Given the stability of these GGAA sequences as determined by the comparison of tumor and germline DNA sequences, the predilection for those two allele ranges does not appear to be a product of sequence evolution within tumor cells during the process of oncogenesis.

Given that *NR0B1* is among the most upregulated direct EWS/FLI targets, and is essential for maintenance of oncogenic transformation[15,39,133], two alternative hypotheses were proposed: the predilection for the selection bias of specific *NR0B1* GGAA-microsatellite haplotypes in tumors is a consequence of either superior oncogenic potential in tumors harboring 21-25 or 56-50 GGAA motifs at the *NR0B1* locus or conversely, these principal GGAA-microsatellite haplotypes observed in tumors are important for Ewing sarcoma susceptibility and transformation in progenitor cells harboring the EWS/FLI translocation. The luciferase assays and qRT-PCR data from human-derived cell lines clearly show that the most common GGAA-microsatellite allele observed in tumors also facilitates maximal EWS/FLI-mediated gene expression. The results from these experiments are further supported by patterns of *NR0B1* gene expression observed in tumor microarray data, wherein tumors harboring small GGAA microsatellites (<20 GGAA repeats, Figure 3.5D) are those that have the lowest levels of *NR0B1* gene expression. Although interpretation of tumor microarray experiments is limited by the small number of samples included, the limited number of samples available with smaller

numbers of GGAA repeats further supports the hypothesis that Ewing sarcoma tumor development is restricted by lower levels of *NR0B1*expression in the setting of these small GGAA-microsatellites. Thus, these results provide additional supportive data in a more biologically relevant context. Likewise, the clinical analysis of AEWS0031 patients demonstrates that EFS is not influenced by these *NR0B1* GGAA polymorphisms, which we believe also supports the latter hypothesis.

Assessing the clinical impact of these *NR0B1* GGAA polymorphisms was an important outcome measure in this study and consequently various statistical approaches were employed to sufficiently address this association. However, when subgroup analyses were performed to address potential confounding influences such as patient sex, zygosity, and chemotherapy our results clearly demonstrate that disease behavior in Ewing sarcoma is not influenced by GGAA-microsatellite polymorphisms at the *NR0B1* locus.

Integrating the results of this study we propose that in Ewing sarcoma, GGAA-microsatellite polymorphisms play an important role in disease susceptibility. It is generally accepted that the EWS/FLI translocation event is the driver oncogenic mutation in Ewing sarcoma. We suggest that in precursor cells exposed to the EWS/FLI chimera, cells with a more 'permissive' genetic constitution of GGAA-microsatellite polymorphisms are more likely to transform when exposed to the EWS/FLI chimera than cells with a non-permissive GGAA genotype (Figure 3.7). Further supporting this model is that Ewing sarcoma is believed by many experts to be exclusively a human condition;

spontaneous cases of Ewing sarcoma have not been observed in any other animal species (except for a single case report in a camel[190]), and inducible Ewing sarcoma models in murine progenitor cell and transgenic mice do not recapitulate the molecular hallmarks of disease[191–194]. Interestingly, the mouse orthologs of *NR0B1*, *CAV1*, *GSTM4, and FCGRT* (4 microsatellite-containing upregulated EWS/FLI targets in humans) do not possess a GGAA-microsatellite in their respective promoter/enhancer regions. Additionally, ectopic EWS/FLI expression in murine-derived NIH3T3 cells does not upregulate *Nr0b1*, further supporting observation that GGAA-microsatellites are necessary for regulation of *Nr0b1* in Ewing sarcoma[195]. Additional sequencing efforts are underway to better characterize a more comprehensive cohort of EWS/FLI-enriched GGAA-microsatellites in African, white Europeans and Ewing sarcoma patients.

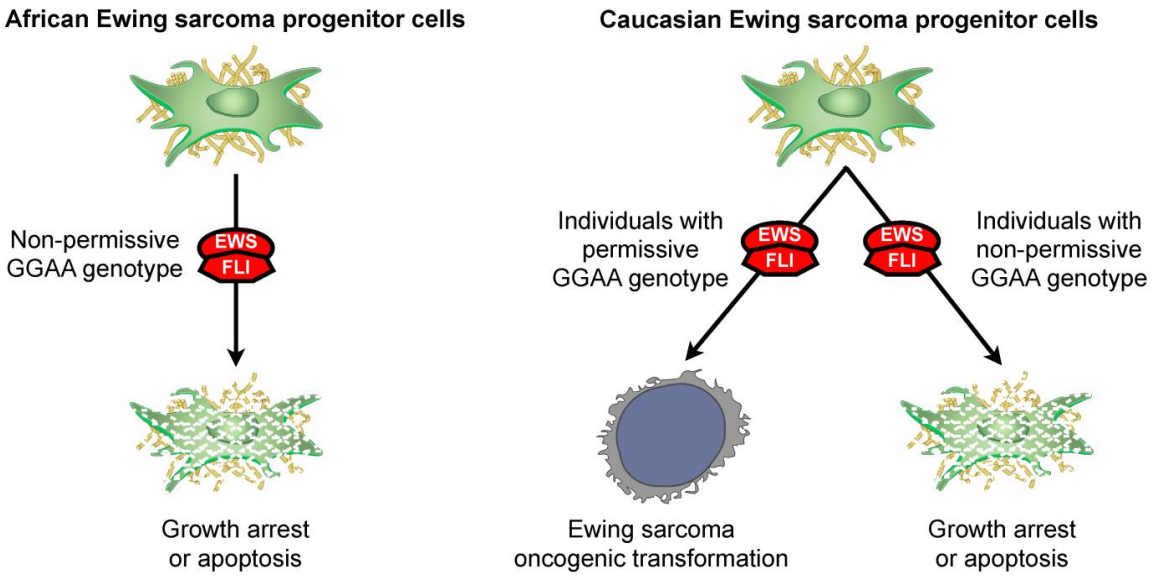


Figure 3.7 Model of GGAA-microsatellite polymorphism contributions to Ewing sarcoma susceptibility in African and white European populations.

In a recent genome-wide association study three candidate Ewing sarcoma susceptibility loci were identified using a comprehensive single nucleotide polymorphism (SNP) analysis[189]. The authors demonstrated a greater frequency of these susceptibility loci in white Europeans as compared to Africans. However, the oncogenic contribution of these identified susceptibility loci in the pathogenesis of Ewing sarcoma has yet to be clarified.

Furthermore, it does not appear that the observed differences in the frequency of these susceptibility loci will fully account for the 10-fold increase in Ewing sarcoma in white Europeans compared to Africans. EWS/FLI-responsive GGAA-microsatellites provide a complementary genetic approach to understand these discrepant patterns of disease incidence. These GGAA-microsatellites are highly polymorphic and are also direct genetic targets of the EWS/FLI chimera. Additionally, compared with white European and Asian populations, African populations are known to have increased genetic diversity for many microsatellite loci[166]. Based on our biochemical data, a greater diversity of GGAA motifs at important microsatellite loci may actually negatively impact EWS/FLI-mediated gene expression, which appears optimal over a narrow range of GGAA motifs. Additional work will be needed to discern the relative contributions of microsatellite polymorphisms and SNPs in the susceptibility to Ewing sarcoma development.

An additional important finding gleaned from this study is that GGAA-microsatellites are genetically stable during the process of oncogenic transformation. Consequently, tumor tissues are not required to obtain DNA for GGAA-microsatellite genotyping in

individuals with Ewing sarcoma. Given the current practice of CT-guided core biopsies and neoadjuvant chemotherapy, Ewing sarcoma tissue is infrequently available for genetic studies. Germline sources of DNA such as blood, bone marrow aspirates, saliva and buccal swabs are more readily available and based on the results presented here, can be used in future GGAA-microsatellite genotyping experiments. Additionally, our data also demonstrates that commercial WGA of tumor DNA does not erroneously expand or contract small or medium sized GGAA-microsatellites. Even in extremely large microsatellites (50-60 GGAA motifs) discordance was minimal (1 GGAA motif). Importantly, together these findings provide valuable insight into the stability of GGAA-microsatellites in Ewing sarcoma, providing an opportunity for prospective genotyping studies to progress beyond the barriers of limited tissue supplies.

In conclusion, this report is the first detailed examination of EWS/FLI-responsive GGAA-microsatellite polymorphisms in Ewing sarcoma. At the *NR0B1* locus, we have demonstrated that in primary Ewing sarcoma tumor samples, there is strong overrepresentation of a narrow range of GGAA haplotypes, which was discordant from healthy white European controls. We further demonstrated that maximal EWS/FLI-mediated gene expression is also highly dependent on a comparably narrow range of GGAA motifs. At the *NR0B1* locus, these polymorphisms do not influence clinical outcomes, favoring a model in which these GGAA polymorphisms may contribute to the elusive permissive cellular and genetic environment necessary for EWS/FLI-mediated transformation.

Acknowledgements

The authors would also like to thank Aimee Madsen and Xiao-qiong Liu for their technical assistance and Ken Boucher PhD, for statistical consultation.

Chapter 4: Role for the EWS domain of EWS/FLI in binding GGAA-microsatellites required for Ewing sarcoma anchorage independent growth

Kirsten M. Johnson, Nathan R. Mahler, Ranajeet S. Saund, Emily R. Theisen, Cenny Taslim, Nathan W. Callender, Jesse C. Crow, Kyle R. Miller, Stephen L. Lessnick. (2017) *PNAS*. 114(37), 9870-9875. Doi: 10.1073/pnas.1701872114

KMJ and SLL designed research; KMJ, NRM, RSS, ERT, NWC, JCC, and KRM performed research; KMJ contributed new reagents/analytic tools; KMJ, RSS, ERT, CT, and SLL analyzed data; and KMJ wrote the paper.

Abstract

Ewing sarcoma usually expresses the EWS/FLI fusion transcription factor oncoprotein. EWS/FLI regulates myriad genes required for Ewing sarcoma development. EWS/FLI binds GGAA-microsatellite sequences *in vivo* and *in vitro*. These sequences provide EWS/FLI-mediated activation to reporter constructs, suggesting that they function as EWS/FLI-response elements. We now demonstrate the critical role of an EWS/FLI-bound GGAA-microsatellite in regulation of the *NR0B1* gene, as well as for Ewing sarcoma proliferation and anchorage-independent growth. Clinically, genomic GGAA-microsatellites are highly variable and polymorphic. Current data suggest that there is an optimal "sweet-spot" GGAA-microsatellite length (of 18-26 GGAA repeats) that confers

maximal EWS/FLI-responsiveness to target genes, but the mechanistic basis for this remains unknown. Our biochemical studies, using recombinant Δ22 (a version of EWS/FLI containing only the FLI portion) demonstrate a stoichiometry of one Δ22-monomer binding to every two consecutive GGAA-repeats on shorter microsatellite sequences. Surprisingly, the affinity for Δ22 binding to GGAA-microsatellites significantly decreased, and ultimately became unmeasureable, when the size of the microsatellite was increased to the "sweet-spot" length. In contrast, a fully-functional EWS/FLI mutant (Mut9, which retains approximately half of the EWS portion of the fusion) showed low affinity for smaller GGAA-microsatellites, but instead significantly increased its affinity at "sweet-spot" microsatellite lengths. Single-gene ChIP and genome-wide ChIP-seq and RNA-seq studies extended these findings to the *in vivo* setting. Together, these data demonstrate the critical requirement of GGAA-microsatellites as EWS/FLI activating response elements *in vivo* and reveal an unexpected novel role for the EWS portion of the EWS/FLI fusion in binding to "sweet-spot" GGAA-microsatellites.

Introduction

Ewing sarcoma is an aggressive bone malignancy of children, adolescents, and young adults[3]. Disease pathogenesis is mediated by a t(11;22)(q24;q12) chromosomal translocation that creates the EWS/FLI fusion oncoprotein. This fusion protein functions as a transcription factor and master regulator of oncogenic transformation by activating and repressing thousands of target genes[34,38]. The amino-terminal EWS portion is a low-

complexity/intrinsically disordered domain that is indispensable for both transcriptional regulation and oncogenic transformation, but is not thought to contribute to DNA binding[8,44]. The carboxyl-terminal FLI portion contains the conserved ETS-type DNA-binding domain and binds with high affinity to the ETS consensus sequence ACCGGAAGTG[54,121]. The DNA binding domain is likewise necessary for EWS/FLI-induced oncogenesis. FLI and EWS/FLI each bind this high-affinity motif as a monomer with similar affinity and specificity[37].

We, and others, previously demonstrated the enrichment of GGAA-microsatellites near EWS/FLI-regulated target genes[15,46]. Our early studies found GGAA-microsatellites associated with genes transcriptionally activated, but not repressed, by EWS/FLI[15]. Many of these GGAA-microsatellites are bound by EWS/FLI *in vivo* and there is a correlation between EWS/FLI binding and EWS/FLI-mediated gene activation. Furthermore, introduction of GGAA-microsatellite sequences confer EWS/FLI-responsiveness to reporter constructs[16]. These data suggest GGAA-microsatellites serve as EWS/FLI-response elements *in vivo*, but this has not been definitively shown.

The human genome contains thousands of GGAA-microsatellites. These display a great deal of sequence variability arising from base transitions and transversions, indels, and variation in number of GGAA repeats. GGAA-microsatellites studied in detail demonstrate a high degree of polymorphism in populations[14]. For example, *NR0B1* is a critical EWS/FLI-regulated target gene required for oncogenesis in Ewing sarcoma[43].

NR0B1 contains a GGAA-microsatellite approximately 1500 bp upstream of the transcriptional start site that shows significant length-polymorphism across populations and between individuals[14]. Perhaps most interestingly, Ewing tumors demonstrate marked enrichment of a narrow-range of GGAA-microsatellite lengths in the *NR0B1*-associated microsatellite, with most containing 18-26 GGAA-repeats, suggesting a relationship between *NR0B1* microsatellite length and tumor development[16].

Our original biochemical studies focused on short microsatellite constructs containing 0-7 GGAA-repeats, and we found there was increasing EWS/FLI-mediated reporter gene activation as the number of GGAA-motifs increased[15]. However, subsequent work found this effect was maximal between 18-26 GGAA-repeats, and longer microsatellite lengths showed diminished EWS/FLI-responsiveness[16]. This led us to propose there is an optimal "sweet-spot" length of GGAA-microsatellite that provides maximal levels of EWS/FLI-mediated gene activation. The molecular basis for this "sweet-spot" maximal activity is not currently known.

To discover the mechanistic basis underlying optimal "sweet-spot" GGAA-microsatellite function, we combined *in vivo* studies of gene expression and oncogenic phenotype with *in vitro* biochemical evaluation of DNA-binding by EWS/FLI mutant alleles. We show EWS/FLI transcriptionally activates *NR0B1* through its associated GGAA-microsatellite. Additionally, this particular microsatellite is required for EWS/FLI-mediated Ewing sarcoma oncogenic transformation, as measured by anchorage-independent colony

formation. We also found smaller GGAA-microsatellites are only able to bind *in vitro* to versions of EWS/FLI that have near-complete deletions of the EWS portion of the fusion; in contrast, optimal "sweet-spot" microsatellites bind with higher affinity to versions of EWS/FLI that retain the EWS portion. Taken together, these data demonstrate an important and novel role for the transcriptional regulatory EWS-domain of EWS/FLI in contributing to binding of "sweet-spot" GGAA-microsatellites, and thus provide a biochemical basis for the enrichment of these microsatellite lengths in Ewing sarcoma.

Materials and Methods

Constructs and Retroviruses

Mammalian expression constructs included the following: Lentiviral vectors containing CRISPR/Cas9 cDNA and sgRNA (See Supplementary Methods); Retroviral vectors encoding Luc-RNAi and EF-2-RNAi and cDNAs for EWS/FLI, Δ22, R2L2, Mut9, and NR0B1 are previously described[43,44,114,175]; the Mut9/R2L2 construct was ordered as a gene block (IDT) and cloned into the pMSCV hygro vector between EcoRI and HindIII restriction sites. Luciferase reporter constructs included human-derived *NR0B1* GGAA-microsatellite polymorphic or synthetic GGAA constructs cloned upstream of the pGL3-promoter SV40 minimal promoter element (Promega Corporation), as described previously[15,16]. Bacterial expression constructs included cDNAs for 6xHis-Δ22 and Halo-Tagged-Mut9 in pET28a and pFN18K, respectively (EMD Chemicals; Promega Corporation).

Cell culture

HEK 293EBNA and Ewing sarcoma cell lines were grown as previously described[43,175]. Cells were infected with CRISPR/Cas9 lentiviral constructs for *NR0B1* microsatellite knockout experiments as previously described[114]. A673 cells were used for EWS/FLI knockdown/rescue experiments. Growth assays were performed on the IncucyteZoom live cell imager. Soft agar assays were performed as described previously[175].

CRISPR/Cas9

Two lentiviruses expressing Cas9 and distinct CRISPR sgRNAs (see Table 4.1 for sequences) and either puromycin and blastocidin resistance markers were used to infect target cells. Vectors were provided by the University of Utah MGD Core (http://cores.utah.edu/mutation-generation-detection/). A ~700bp region containing the *NR0B1* GGAA-microsatellite was deleted, and was the smallest that could be targeted with high-quality sgRNAs due to the region's repetitive nature. Control lentiviruses lacked the sgRNA sequences. Genomic DNA from drug-selected polyclonal cell populations was isolated within 10 days of infection, PCR amplified (using primers listed in Table 4.1), and sequenced to verify *NR0B1* microsatellite deletion (Figure S4.1). Results were validated by gel electrophoresis (Figure 4.1A). RNA and protein were collected within 2-3 weeks of CRISPR/Cas9 infection. A673 genomic DNA was collected weekly for 3 weeks to assess stability of CRISPR/Cas9-mediated knockout (Figure S4.2C).

Immunodetection

Antibodies used for immunodetection: anti-FLI (Abcam ab15289), anti-α-Tubulin (Calbiochem CP06), and anti-NR0B1 (Abcam ab97369).

Quantitative reverse-transcriptase polymerase chain reaction (qRT-PCR)

RNA was collected using an RNeasy Kit (Qiagen). Total RNA from cells was amplified and detected using SYBR green fluorescence for quantitative analysis[175]. Primer sequences are listed in Table 4.1.

FLI ChIP and ChIP-seq

FLI ChIP experiments were performed as previously described[118] using the anti-FLI antibody (sc-356X Santa Cruz Biotechnology, Inc.) and chromatin prepared from A673 and HEK 293 EBNA cells. ChIP DNA and input controls were sequenced with the Hi-Seq Illumina Genome Analyzer, and data was analyzed following the procedures previously described[198–200]. See Supplementary Methods for additional information.

RNA-seq Data Collection and Analysis

See Supplementary Methods for RNA-seq data collection and analysis.

Protein Purification

Recombinant proteins were expressed in BL21 competent cells from pET28a or pFN18K (EMD Chemicals, Promega) expression plasmids encoding Δ22 and Mut9, respectively. Batch purification conditions are available in the Supplementary methods.

Fluorescence Polarization

Fluorescein-labeled DNA duplexes were obtained from IDT (Integrated DNA Technologies). Sequences are listed in Table 4.2. Fluorescence polarization was performed using a BioTek Synergy2 fluorometer (Winooski, VT). Recombinant protein preparation is described in Supplementary Methods. DNA duplex (I) (containing a high-affinity ETS binding site) was used as a control for monomeric protein binding. Binding and stoichiometry assays were performed as before[51]. Affinity plots and curve fits were generated using the GraphPad Prism program (GraphPad Software). See detailed procedures in the Supplementary Methods.

Luciferase Assays

Luciferase reporter assays were performed by transfecting reporter constructs, as well as appropriate EWS/FLI expression constructs into HEK 293 EBNA cells. Luminescence was measured after 24 hours as described previously[16]. See Supplementary Methods for details.

Results

The NR0B1 GGAA-microsatellite is required for EWS/FLI-mediated transcriptional activation, Ewing sarcoma proliferation and oncogenic transformation

NR0B1 encodes an orphan nuclear receptor whose expression is necessary for oncogenic transformation of Ewing sarcoma cells[43]. There is a highly-polymorphic GGAA-microsatellite at approximately 1500 bp 5' to the *NR0B1* transcriptional start site which is bound by EWS/FLI in Ewing cells[15]. Knock-down of EWS/FLI expression causes a concomitant reduction in NR0B1 RNA and protein expression, suggesting *NR0B1* is regulated by direct binding of EWS/FLI to its microsatellite[43].

To explicitly test whether the GGAA-microsatellite is necessary for EWS/FLI-mediated activation of *NR0B1*, we utilized CRISPR/Cas9 to delete the region containing the *NR0B1* microsatellite in A673 Ewing cells. Genomic PCR and Sanger sequencing of isolated polyclonal cell populations demonstrated successful deletion of the GGAA-microsatellite in approximately 80% of the polyclonal population (Figures 4.1A, Supplementary Figure 4.5).

To determine the effects of microsatellite loss on *NR0B1* expression, we next evaluated mRNA and protein levels. Deletion of the microsatellite reduces *NR0B1* expression at both the RNA and protein level by 80% or more as compared to control cells (Figure 4.1B). Remaining NR0B1 expression is likely due to residual cells in the polyclonal population that did not undergo microsatellite excision, but we cannot exclude persistent low-level gene expression after microsatellite deletion. These data indicate the GGAA-microsatellite is required for full-level *NR0B1* expression in Ewing sarcoma cells.

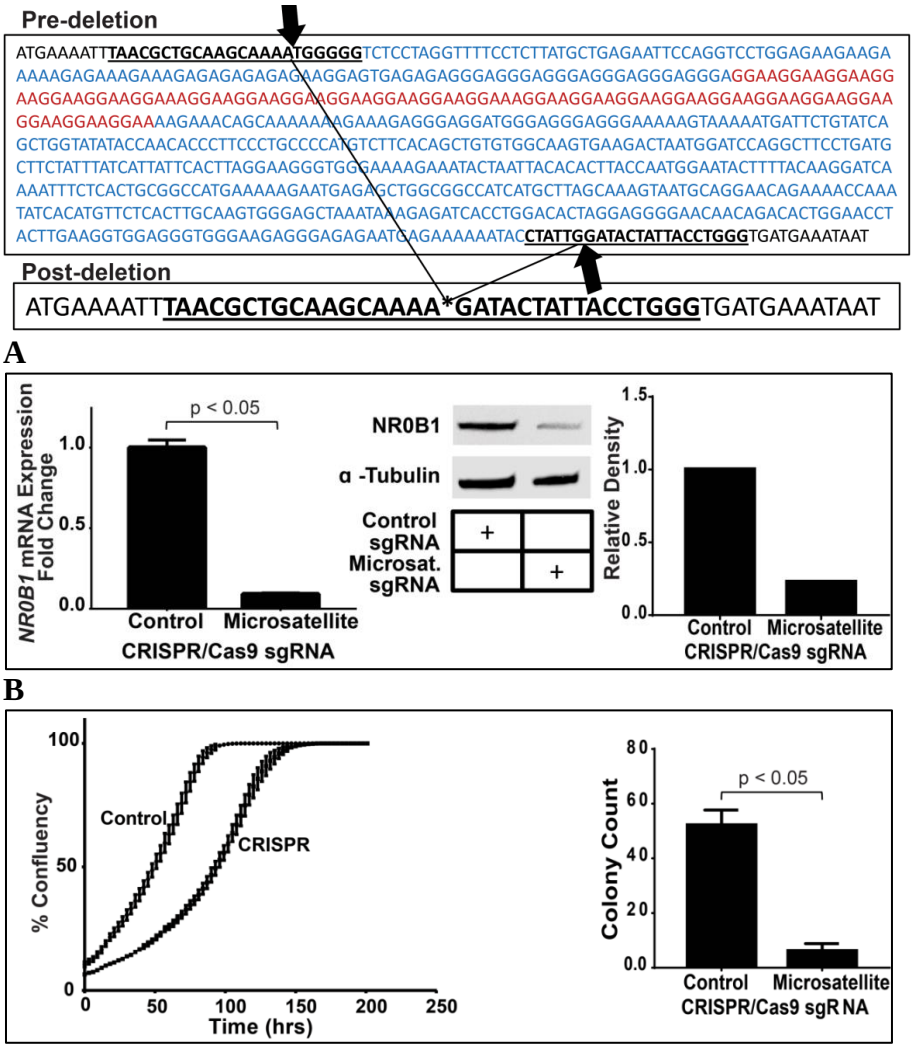


Figure 4.1 Deletion of the *NR0B1* microsatellite reduces *NR0B1* expression, impairs A673 cell growth and inhibits colony formation

(A) Sequencing results validating knock-out of the *NR0B1* GGAA-microsatellite about 1.5kb upstream of the *NR0B1* TSS in A673 cells. The sgRNAs targeted to either side of this region are underlined. GGAA-microsatellite is highlighted red and CRISPR/Cas9 deleted region is highlighted blue. Gel shows deletion of *NR0B1* microsatellite region compared to control (non-deleted), with densitometry quantification on the right ($p < 0.01$). Data represented as mean ± SEM (n=2) **(B)** NR0B1 mRNA ($p < 0.05$) and protein expression levels in control and CRISPR/Cas9-mediated knock-out of *NR0B1* microsatellite in A673 Ewing sarcoma cells, with western blot densitometry quantification on the right. Control CRISPR/Cas9 plasmids do not contain sgRNAs. Data represented as mean ± SEM (n=3) **(C)** Growth and colony formation assay quantification

of CRISPR/Cas9 control vs. *NR0B1* microsatellite knock-out in A673 cells ($p < 0.05$). Data represented as mean ± SEM (n=3).

We next evaluated the role of the *NR0B1* microsatellite on Ewing sarcoma cell behavior. Growth curves and anchorage-independent growth were determined for control, and microsatellite-deleted cells, using live cell imaging and soft-agar assays, respectively. We found cells harboring deletion of the *NR0B1* microsatellite exhibited impaired proliferation rate and near-complete loss of colony formation (Figure 4.1C). Similar results were observed using TC71 and EWS502 Ewing sarcoma cells (Supplementary Figures 4.6A-B). The effect was more subtle in these latter cells likely because of less-complete microsatellite deletion coupled with a propensity for non-deleted cells to outgrow CRISPR knock-out cells over time (Supplementary Figure 4.6C). In contrast to these Ewing sarcoma cells, deletion of the GGAA-microsatellite containing region in non-Ewing HEK293 cells did not cause a decrease in NR0B1 mRNA or protein levels; instead, there was a non-statistically significant minor increase in mRNA levels following microsatellite excision (Supplementary Figure 4.6D).

To determine if loss of oncogenic transformation of A673 cells following GGAA-microsatellite excision was due to loss of NR0B1 protein expression, we performed "rescue" experiments. We introduced an NR0B1 cDNA into A673 cells by retroviral infection and drug selection, and followed this by deletion of the GGAA-microsatellite using the lentiviral system described above. Enforced NR0B1 expression rescued both

NR0B1 protein levels and colony formation in soft agar (Supplementary Figure 4.6E), suggesting loss of transformation resulted from loss of NR0B1 protein expression, and not from an off-target or other non-specific effect. Collectively, these data indicate the GGAA-microsatellite is required for full-level expression of the *NR0B1* gene, tissue-culture proliferation, and anchorage-independent growth of Ewing sarcoma cells.

The FLI domain of EWS/FLI interacts with short GGAA-microsatellites as monomers via independent binding events

Our prior studies evaluating interactions between EWS/FLI and short GGAA-microsatellite sequences containing 0-7 consecutive GGAA repeats suggested that FLI exhibits homodimeric binding on elements containing 4 or more GGAA-repeats with a dissociation constant (K_D) of ~70 nM[51]. These data were based on a combination of electrophoretic mobility shift assays and fluorescence anisotropy studies using two recombinant mutants of EWS/FLI: the isolated 101 amino-acid FLI ETS domain, and the Δ22 mutant, containing all of the FLI portion of EWS/FLI and 6 amino acids from the EWS portion[15,51]. These studies did not evaluate larger microsatellite sequences and thus were unable to address the stoichiometry and potential for cooperative binding on longer GGAA-microsatellites.

To determine the stoichiometry of EWS/FLI binding on GGAA-microsatellites of more relevant lengths (i.e., longer microsatellites), we conducted fluorescence anisotropy studies using the same Δ22 construct (Supplementary Figure 4.7A) used in our earlier

studies[34]. Using fluorescein-labeled DNA duplexes containing increasing numbers of consecutive GGAA-motifs, ranging from 2-16 repeats, we found a highly consistent stoichiometric ratio of one Δ22 monomer binding for every two GGAA repeats (R^2 0.9881; Figure 4.2A, Supplementary Figure 4.7B). Additional evaluation revealed one Δ22 monomer binds two GGAA-repeats, and two monomers binds three repeats, the only ratio inconsistent with the established 1:2 pattern (Figure 4.2A, Supplementary 4.7C). These data differ slightly from our previous studies, which suggested Δ22 is unable to bind 3 or fewer GGAA-repeats[15]. The reason for this discrepancy is unknown, but may be related to differing sensitivities of EMSA (used in prior studies) and fluorescence anisotropy (used in this study). Overall, the stoichiometry experiments suggest a "head-to-tail" binding model whereby single molecules of Δ22 can "anchor" to Δ22 molecules already bound on the microsatellite, and extend a "chain" of bound Δ22 one molecule at a time.

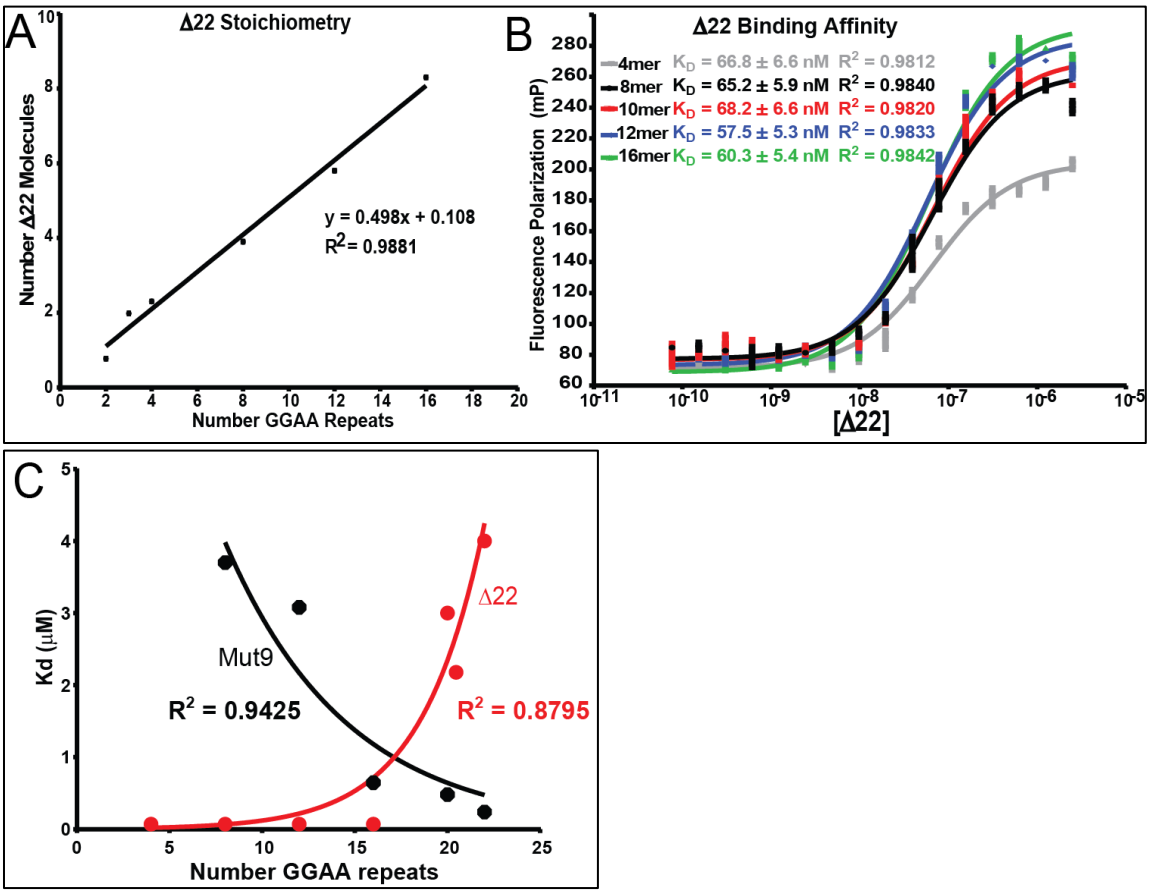


Figure 4.2 Characterization of Δ22 binding on DNA sequences of increasing GGAA-microsatellite numbers

(A) Fluorescence polarization (FP) was used to determine the stoichiometry of recombinant Δ22 protein binding fluorescein-labeled DNA probes from 2-16 consecutive GGAA repeats. Data represents mean of 2 independent experiments (each with 3 technical replicates) for each GGAA-repeat length. $R^2 = 0.9881$ **(B)** FP was used to assay binding affinity of recombinant Δ22 protein bound to fluorescein-labeled oligonucleotide probes of 4-16 consecutive GGAA motifs. K_D was determined to be approximately 70nM. Data represents mean ± SEM (n=3). **(C)** Summary of K_D (binding affinity) determined by fluorescence anisotropy for recombinant Δ22 vs. Mut9 proteins binding to fluorescein-labeled DNA oligonucleotides of increasing GGAA-microsatellite numbers. Data represent the mean of 2 independent experiments for each GGAA-repeat length. $R^2 = 0.9425$ and 0.8795 for Δ22 and Mut9, respectively.

EWS sequences are required for EWS/FLI binding to "sweet-spot" GGAA-microsatellite lengths in vitro

The analysis we performed above included GGAA-microsatellite sequences ranging from 2-16 consecutive GGAA-repeats. We next considered the possibility that "sweet-spot" microsatellite lengths might be optimal for gene expression because of improvements in affinity of the FLI DNA-binding domain for these longer repeat lengths. To test this possibility, we again used fluorescence anisotropy to evaluate binding of Δ22 to GGAA-microsatellites containing 18-22 consecutive GGAA-repeats. Unexpectedly, we found the affinity of Δ22 progressively worsened as the microsatellite length increased, and was unmeasureable at the longest microsatellite tested (22 repeats; Supplementary Figure 4.8A). It is known Δ22 is unable to rescue oncogenic transformation of Ewing sarcoma cells in which endogenous EWS/FLI has been knocked down[8]. This has been assumed to be due to the absence of a transcriptional regulatory domain contributed by the EWS portion of the fusion. The current data suggest an additional possibility: that Δ22 also fails to bind GGAA-microsatellites *in vivo*.

It is clear full-length EWS/FLI binds "sweet-spot" GGAA-microsatellites *in vivo*, and rescues oncogenic transformation of Ewing sarcoma cells[44,51]. We therefore sought to determine whether the inclusion of EWS sequence would change the binding characteristics of the fusion *in vitro*. Full-length EWS/FLI is challenging to purify as a recombinant protein in a fully-functional form as the low-complexity/intrinsically-disordered EWS portion tends to cause aggregation of the protein *in vitro*[196]. To

circumvent this challenge, we instead purified a mutant form (Mut9) containing an internal deletion of 164 amino acids that comprise much of the intrinsically disordered domains of the EWS portion of the EWS/FLI fusion. Mut9 fully rescues oncogenic transformation in Ewing sarcoma and regulates the limited number of genes tested in a manner nearly identical to full-length EWS/FLI (see below for Mut9 global transcriptional analysis)[44]. This construct is, however, readily purified as a recombinant protein, and we therefore used it in place of full-length EWS/FLI.

We analyzed recombinant Mut9 binding to a series of GGAA-microsatellite sequences using fluorescence anisotropy. In the case of suboptimal/shorter GGAA-repeat sequences, we found Mut9 binds with poor affinity (K_D in the 3-6 μM range for 8-12 GGAA-repeats; Figure 4.2C). Interestingly, the K_D significantly improves as the number of GGAA-motifs is increased into "sweet-spot" lengths: at 22 GGAA-repeats the K_D decreased to 805 nM (Supplementary Figure 4.8B). These data demonstrate a significant change in binding capacity to GGAA-microsatellites that is dependent on the EWS-portion of the EWS/FLI fusion: Δ22 binds well to short, but not "sweet-spot" microsatellites, while Mut9 binds to "sweet-spot," but not shorter, microsatellites (Figure 4.2C). Taken together, these data demonstrate the transcriptional regulatory EWS-domain plays an unanticipated, but critical role in binding of the EWS/FLI fusion to "sweet-spot" GGAA-microsatellite regulatory elements.

EWS sequences are required for EWS/FLI binding to "sweet-spot" GGAA-microsatellite lengths in vivo

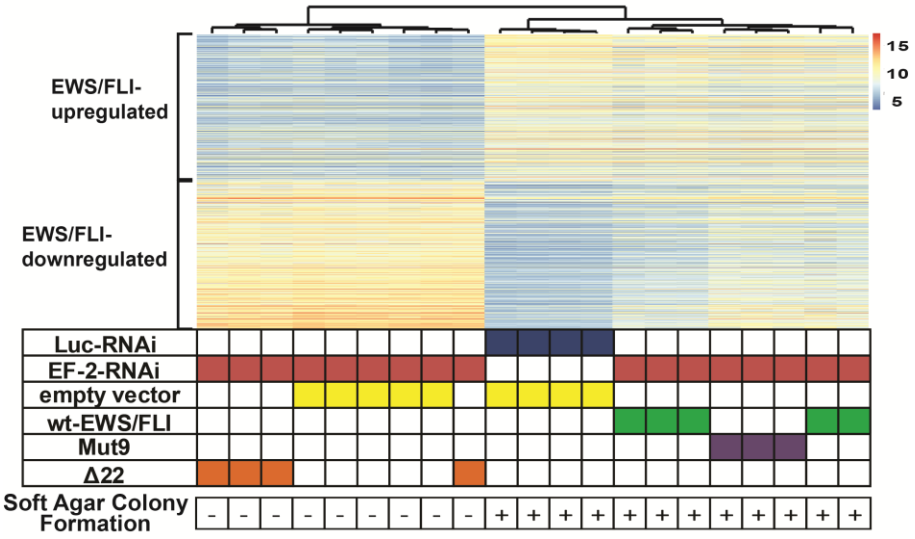
The work we presented on GGAA-microsatellite binding thus far used recombinant proteins and DNA duplexes in an *in vitro* setting. These *in vitro* studies are limited by their inability to account for the more complicated intracellular milieu and additional protein-protein interactions present in living Ewing sarcoma cells. To address this issue, we next performed *in vivo* experiments to test our model.

We previously demonstrated EWS/FLI activation of luciferase reporters containing GGAA-microsatellites reveals a "sweet-spot" of approximately 20-30 GGAA-repeats. To test whether Mut9 demonstrates a similar "sweet-spot" preference, we performed luciferase reporter assays using GGAA-repeat lengths from 10-70 repeats (Supplementary Figure 4.9A). We found peak Mut9 responsiveness at 40 repeats, and evidence of a second peak in the 70-repeat range (generally similar to what was previously observed with full-length EWS/FLI). These data indicate Mut9 functions in an analogous manner to wild-type EWS/FLI in this reporter system. In contrast, a Mut9/R2L2 mutant, which contains a two-amino acid substitution in the DNA-binding domain of the FLI portion (Supplementary Figure 4.7A) and cannot bind DNA[34], does not induce transcriptional activation, regardless of GGAA-repeat length (Supplementary Figure 4.9A). Thus, our data suggest a requirement for both the EWS and FLI portions of the fusion for DNA binding and transcriptional activation.

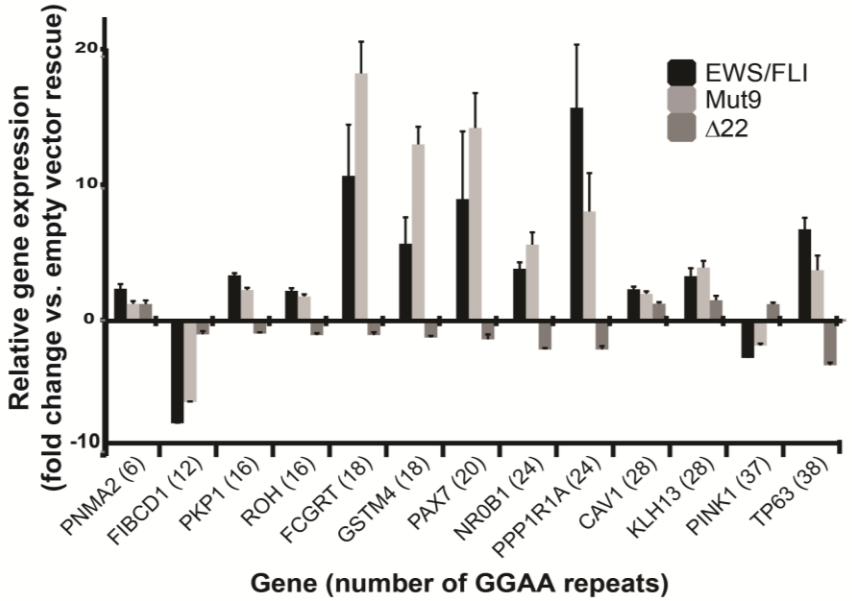
We next extended these findings to endogenous genes in patient-derived A673 Ewing sarcoma cells. We "knocked-down" EWS/FLI with a retrovirally-expressed shRNA (EF-2-RNAi), or used a control RNAi targeting luciferase (Luc-RNAi), and "rescued" expression with RNAi-resistant cDNAs expressing either wild-type EWS/FLI, Mut9 or Δ22, or an empty-vector control. As previously reported[34], both wild-type EWS/FLI and the Mut9 mutant rescue oncogenic transformation, while the Δ22 and empty-vector controls did not (Supplementary Figure 4.9B). RNA-seq was next performed on these polyclonal populations of cells. Unsupervised hierarchical clustering revealed that Mut9-rescued cells clustered (and intermixed) with the wild-type EWS/FLI-rescued cells (Figure 4.3A). These had gene expression patterns that were highly-similar to cells treated with the Luc-RNAi control. In contrast, the Δ22-rescued cells clustered (and intermingled) with empty-vector rescue cells (Figure 4.3A). These data indicate Mut9 shares a nearly-identical gene expression pattern with wild-type EWS/FLI, and therefore validates its use as an "EWS/FLI-equivalent" version.

To determine whether GGAA-microsatellite repeat number affects EWS/FLI- and Mut9-mediated gene activation, we next selected a handful of EWS/FLI-regulated genes containing nearby microsatellites of varying lengths (range: 6-38 GGAA-repeats) and plotted their EWS/FLI-induced gene expression from our RNA-seq data. Genes associated with microsatellites containing 18-26 GGAA-repeats had high levels of EWS/FLI- and Mut9-mediated gene activation (Figure 4.3B). In contrast, those with

fewer or greater numbers of repeats showed diminished EWS/FLI- and Mut9-mediated activation. As anticipated, Δ22 showed little if any gene regulatory activity.



A



B

Figure 4.3 EWS/FLI mediated differential gene expression in Mut9 vs. Δ22 rescue of EWS/FLI knock-down in A673 cells at different microsatellite lengths.

(A) Heat map of hierarchical clustering of the 500 most EWS/FLI up- and down regulated genes across cells expressing varying knockdown/rescue constructs. Each row represents one gene and each column represents one biological sample. Values used to determine differential expression were normalized count matrices (scale represents normalized counts). **(B)** Results comparing differential gene expression from RNA-seq data (Fig. 3A) of EWS/FLI-regulated genes in the context of rescue with wild type, Mut9, and Δ22 constructs. The number of GGAA motifs (according to UCSC hg19 reference genome) contained in their respective gene-associated microsatellites is indicated. Data represents mean ± SD (n=3).

We next asked whether the EWS portion of EWS/FLI is critical for binding "sweet-spot" GGAA-microsatellites *in vivo*. We performed directed ChIP-PCR experiments using an anti-FLI antibody. We used the A673 knock-down/rescue approach described above, and rescued expression with cDNAs expressing either wild-type EWS/FLI, Mut9, Δ22, or the R2L2 DNA-binding mutant[75]. Schematics of the EWS/FLI mutants are shown in Figure S4.3A. Following ChIP, we performed qPCR for the *NR0B1* microsatellite as an example "sweet-spot" microsatellite (containing 25 GGAA-repeats). We found wild-type EWS/FLI and Mut9 both demonstrate binding enrichment at this site, whereas Δ22 and R2L2 do not (Supplementary Figure 4.10A). Similarly, when introduced into the non-Ewing sarcoma cell line HEK293, both wild-type EWS/FLI and Mut9 show enrichment at the *NR0B1* GGAA-microsatellite while Δ22 and R2L2 do not (Supplementary Figure 4.10B).

We next compared the genomic localization of Mut9 to Δ22 using the knock-down/rescue strategy in A673 Ewing sarcoma cells. ChIP-seq demonstrated that Mut9 was globally-enriched at EWS/FLI binding sites across the human genome, while Δ22

binding was not significantly enriched over control cells expressing an empty-vector rescue construct (Figure 4.4, Supplementary Figure 4.10C). Residual FLI binding present in the empty-vector and Δ22 rescue samples reflects the incomplete knock-down observed with our EWS/FLI shRNA. Overall, these data extend our *in vitro* findings: the EWS-portion of EWS/FLI is required for *in vivo* DNA binding.

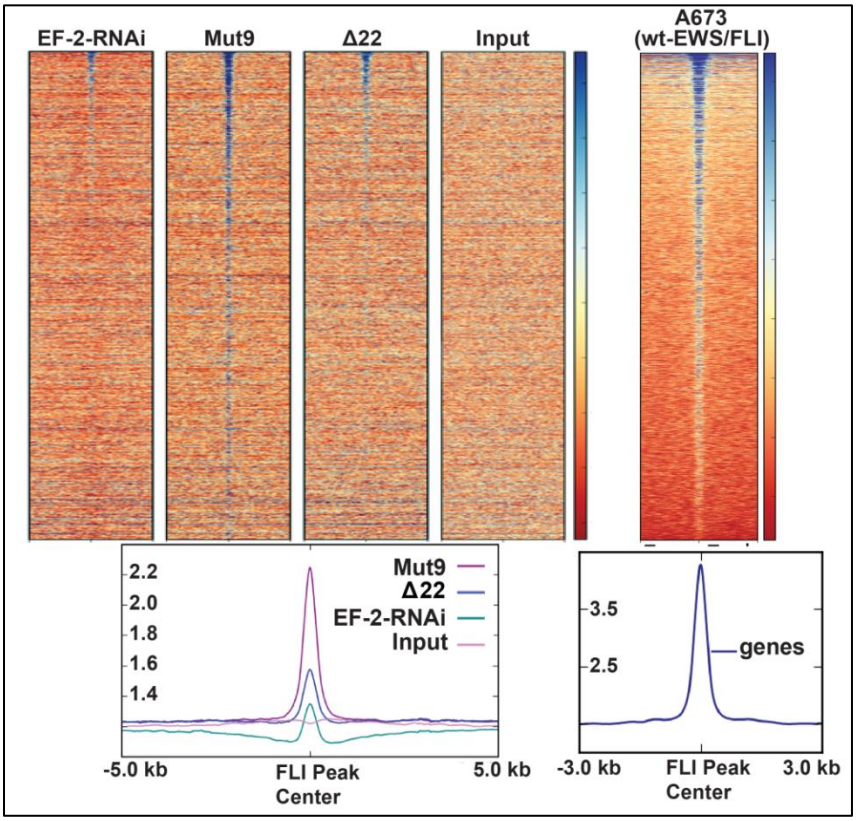


Figure 4.4 Genome-wide FLI-ChIP binding of Mut9 vs. Δ22. Heat map of genome-wide FLI-ChIP-seq data from A673 cells with EWS/FLI knock-down vs. Mut9 or Δ22 rescue compared to input and A673 wild-type EWS/FLI cells. See Supplementary methods for additional sorting information.

Taken together, these data highlight a previously-unrecognized requirement for sequences contributed by the EWS-portion of the EWS/FLI fusion in DNA binding and

gene activation. Furthermore, these data localize the portion of EWS required for this activity to the portion retained in the Mut9 construct (amino-acids 1-82 and 246-264 of the EWS portion of full-length EWS/FLI).

Discussion

EWS/FLI is a modular transcription factor containing an ETS-type DNA binding domain in the FLI-portion of the fusion, and a transcriptional activation/repression domain in the EWS-portion of the fusion[44]. We now demonstrate an unanticipated critical modulatory role for the EWS-portion in EWS/FLI binding to GGAA-microsatellites: at shorter microsatellite lengths the EWS-portion inhibits binding, and at longer "sweet-spot" lengths the EWS-portion enables binding. This finding has important implications for our understanding of transcriptional regulation and Ewing sarcoma development.

We and others first identified GGAA-microsatellites as potential EWS/FLI response elements using ChIP-chip, luciferase reporter, and *in vitro* DNA binding assays[15,16,46]. However, the data implicating GGAA-microsatellites as requisite EWS/FLI-response elements was circumstantial at best. In the current report, we demonstrate the region containing the GGAA-microsatellite adjacent to the *NR0B1* gene is required for activation of that gene in Ewing sarcoma, and genetic deletion of that region disrupts normal Ewing sarcoma cell growth and colony formation in soft agar. These studies are the first to explicitly link EWS/FLI-bound GGAA-microsatellites to cancerous phenotypes in Ewing sarcoma. These data support the notion that alterations in GGAA-

microsatellite function (for example, through microsatellite-length polymorphisms) can have significant effects on Ewing sarcoma development. This is important because we have shown significant population differences in GGAA-microsatellite length polymorphisms between African and European populations: Europeans have a significant enrichment in "sweet-spot" length GGAA-microsatellites at the *NR0B1* locus as compared to Africans, who have greater numbers of larger microsatellites (>30 GGAA-repeats)[16]. These data correlate with the incidence of Ewing sarcoma in these two populations: Europeans have a 10-fold higher incidence of Ewing sarcoma as compared to Africans[14]. Furthermore, patients who develop Ewing sarcoma have an even higher level of enrichment of "sweet-spot" microsatellites[16]. These data were certainly suggestive of a contribution of the *NR0B1* microsatellite polymorphisms in Ewing sarcoma susceptibility, but our current demonstration of the necessity for the adjacent GGAA-microsatellite in *NR0B1* gene expression provide an explicit linkage between these findings. These data also support a recent study suggesting a Ewing sarcoma-susceptibility locus creates a "sweet-spot" microsatellite in the risk allele and this leads to increased expression of the *EGR2* gene and Ewing sarcoma development[32].

The *in vitro* and *in vivo* studies presented in this report strongly corroborate one-another and indicate the EWS-portion of the EWS/FLI fusion is critical for binding of EWS/FLI to "sweet-spot" microsatellites. These data provide a mechanistic rationale for the presence of "sweet-spot" microsatellites: if GGAA-microsatellites are too short, EWS/FLI is not able to bind well. Thus, EWS/FLI binding and associated gene activation

is only possible at microsatellites exhibiting at least "sweet-spot" numbers of GGAA-repeats. We do not currently have the capability to assess microsatellite lengths longer than "sweet-spot" lengths in our *in vitro* studies, but we anticipate that longer microsatellite lengths would be inefficient at binding EWS/FLI.

Interestingly, recent published data indicate EWS/FLI-bound GGAA-microsatellites in Ewing sarcoma are in an open chromatin state, while knock-down of EWS/FLI results in a closed chromatin configuration at these loci[69]. Conversely, introduction of EWS/FLI into mesenchymal stem cells converts closed chromatin into an open state[69,128]. These data suggest an important role of EWS/FLI at GGAA-microsatellites is to convert closed chromatin to an open state to enable transcriptional activation. The implication is EWS/FLI might function at these loci as a "pioneer factor," binding DNA and recruiting chromatin-modifying complexes to induce an open-chromatin configuration[197]. An intriguing interpretation of our data is EWS/FLI may induce this chromatin-opening at GGAA-microsatellites via a "mechanical" biophysical mechanism: initial binding of EWS/FLI at "sweet-spot" microsatellites might facilitate the binding of additional EWS/FLI molecules, and essentially open the chromatin through a "coating" mechanism. Our data therefore implicate the EWS-portion of the fusion as critical in facilitating DNA accessibility.

Our data do not speak to the detailed mechanism by which the EWS-portion of EWS/FLI participates in fusion binding to "sweet-spot" microsatellites. It is tempting to speculate a

polymerization process, mediated by the EWS-portion, is involved. FUS, a paralog of EWS, is also involved in chromosomal translocations leading to oncogenic fusion transcription factors. FUS is capable of polymerizing and forming hydrogels under certain conditions[78]. When the amino-terminus of FUS is joined to the FLI DNA binding domain, addition of GGAA-microsatellite sequences appears to trigger polymerization of the fusion. This is thought to occur at a series of [G/S]Y[G/S] amino acid repeats[78]. Similar [G/S]Y[G/S] repeats are present in the amino-terminal portion of EWS that is included in the EWS/FLI fusion protein[78]. Future studies will be required to determine if polymerization, via the EWS-portion, is required for binding to "sweet-spot" GGAA-microsatellites.

In addition to GGAA-microsatellite binding, a significant portion of the EWS/FLI fusion is bound to high-affinity ETS sites[15,69]. One limitation of the current study is we did not evaluate the role of the EWS portion of the fusion on binding to these high-affinity sites. Studies addressing this question may shed additional light on the mechanism of EWS/FLI binding to, and activation of, its target genes.

In summary, we provide strong evidence for a critical role of the *NR0B1* GGAA-microsatellite in Ewing sarcoma development, and provide new mechanistic details for the ability of EWS/FLI to bind to GGAA-microsatellites at "sweet-spot" lengths. These data indicate the role of the EWS-portion of the fusion protein is not simply to interact with transcriptional co-regulators to mediate gene expression, but it is also required for

binding to GGAA-microsatellites. Furthermore, this work suggests new opportunities in targeting of the fusion: approaches that disrupt the DNA-binding modulatory role of the EWS-portion of EWS/FLI may be sought out as new therapeutic approaches for patients with Ewing sarcoma.

Acknowledgements

We thank Ryan Roberts for critical reading of the manuscript and members of the laboratory of S.L.L. for helpful discussions. This work was supported by funds awarded to S.L.L. from the National Cancer Institute (Grant R01 CA140394 and R01 CA183776), and funds awarded to K.M.J. from the National Cancer Institute (Grant F30 CA210588).

Supplementary Information

Supplemental Materials and Methods

Cell culture

HEK 293EBNA cells and Ewing sarcoma cell lines (A673, TC71, EWS/502) were infected with CRISPR/Cas9 lentivirus and A673 cells were retrovirally infected as previously described for EWS/FLI knockdown/rescue experiments[43,118]. Polyclonal cell populations were grown in the appropriate selection media for at least 4 days before any cells were seeded for collection[43,118]. Growth assays were performed in 96-well plates on the IncucyteZoom live cell imager within 7-10 days of lentiviral infection. Briefly, 8000 cells/well were seeded in triplicate and imaged every 4-6 hours for 7-10 days.

IncucyteZoom software pre-calibrated for Ewing sarcoma cells measured cell confluence levels as a percentage to assess cell growth over time.

Quantitative reverse-transcriptase polymerase chain reaction (qRT-PCR)

Total RNA from cells was amplified and detected using SYBR green fluorescence for quantitative analysis in triplicate for each sample[118]. Normalized fold enrichment was calculated by determining the fold-change of the mean of each condition relative to the control mean, with the data in each condition normalized to an internal housekeeping control gene *GAPDH*. Primer sequences used for qRT-PCR analysis for all target genes are included in supplementary Table 4.1. Two-tailed *t* tests were used for statistical comparison.

FLI ChIP and ChIP-seq

ChIP assays were carried out on A673 and HEK293 cultures of approximately 3 - 10 × 10^6 cells per sample and per epitope, following the procedures previously described[199,200]. Chromatin from formaldehyde-fixed cells was fragmented to a size range of 200-700 bases with a Misonix Sonicator. Solubilized chromatin was immunoprecipitated with antibodies against FLI (Santa Cruz, sc-356X Santa Cruz Biotechnology, Inc.) or IgG (Santa Cruz, sc-2027 Santa Cruz Biotechnology, Inc.). Antibody-chromatin complexes were pulled down with M-280 sheep anti-rabbit IgG Dynabeads (Thermo Scientific), washed (8 minutes each in 20mM Tris-Cl (pH 8.0) buffers containing 250mM NaCl, 500mM NaCl, and 250mM LiCl respectively) and then eluted (50mM Tris-Cl (pH 8.0),

50mM $NaHCO_3$, 1% SDS and 1mM EDTA). After crosslink reversal, RNAse A and Proteinase K treatment, immunoprecipitated DNA was extracted with the Mini-Elute PCR purification kit (Qiagen). ChIP DNA was quantified with Qubit. Quantitative RT-PCR was performed with *NR0B1* primers and normalized with 1% of starting chromatin, used as input, for each sample. ChIP analysis was performed using the % input method, and the fold enrichment method normalized to mock (IgG) control for each sample (Thermo Fisher Scientific). Remaining ChIP DNA samples were used to prepare sequencing libraries, and ChIP DNA and input controls were sequenced with the Hi-Seq Illumina Genome Analyzer. Peak calling was performed using USeq software (28). Heat maps and the profile plot were generated by individual sorting of the ChIP-seq signal from lowest to highest for each mutant rescue construct, and centered on FLI peaks demonstrated in previous A673 cell FLI ChIP-seq data. Integrated Genome Browser (IGB) was used to visualize individual FLI enrichment at specific genomic loci, such as at *NR0B1*. Raw sequence reads can be found in NCBI's Gene Expression Omnibus and are accessible through GEO SuperSeries accession number GSE94503.

RNA-seq Data Collection and Analysis

A673 cells were stably infected and selected for expression of a control Luc-RNAi or the EF-2-RNAi. Following 4 days of selection, cells were infected with either an empty pMSCV-hygro vector or vector expressing full length type IV EWS/FLI, the Δ22 EWS/FLI mutant, or Mut9 EWS/FLI mutant. Knock-down and rescue were verified by seeding cells in soft agar in duplicate for each condition. RNA was extracted with an

RNAeasy kit (Qiagen) using an on-column DNAse digestion protocol. Libraries for deep-sequencing were prepared according to the manufacturer's instructions (Illumina) and sequenced on an Illumina Hi-Seq 4000 to generate 150 bp paired-end reads. Sequences were aligned to the human genome build hg19 using version 2.5.0c of the aligner STAR. Raw sequence reads can be found in NCBI's Gene Expression Omnibus and are accessible through GEO SuperSeries accession number GSE94503. Raw read counts were generated using *htseq-count* with intersection-strict mode. Raw counts were normalized in DESeq2 with *rlog*. The 500 genes most up and down-regulated by EWS/FLI were used to generate the heatmap in *pheatmap* with default column clustering settings and row clustering off. Differential expression was calculated using *DEseq2* for each pairwise comparison of each EWS/FLI construct expressed over EWS/FLI knockdown as compared to rescue with an empty vector. Genes selected for GGAA-repeat number vs. mRNA expression level comparison were selected from a panel of microsatellite-associated genes (TSS of gene nearest to the center of the GGAA-microsatellite) that represented a spectrum of GGAA-repeat lengths. The TSS of nine of the 13 genes selected are within 7kb of a GGAA-microsatellite. The fold change values in Figure 4.3B represent the mean of at least 3 biological replicates for each condition, derived from our RNA-seq data (Figure 4.3A).

Recombinant Protein Purification

Recombinant Δ22 and Mut9 proteins were prepared from *E. Coli* BL21(DE3) cells transformed with pET28a or pFN18K (EMD Chemicals, Promega) expression plasmids,

respectively. Cultures were auto-induced in low-phosphate ZY-media at 37˚C for 7 hours, and then 24˚C for 20-22 hours. Harvested cells were lysed with 0.2mg/ml lysozyme in a lysis buffer containing a protein inhibitor cocktail (Roche, USA), 10% glycerol, 50mM Tris-HCl pH 8.0 (HEPES pH7.5 for Mut9), 500mM NaCl, 0.5mM BME, and 0.1mM PMSF for 45 min on ice, and then sonicated. The cell lysate was centrifuged at 45,000 RPM for 45min. The supernatant was then mixed for 2 hours at 4˚C with either Ni-NTA resin (Qiagen, USA) for His-tagged Δ22, or Halo-link resin (EMD Chemicals, Promega) for Halo-tagged Mut9. Resin-bound Δ22 was washed over a column (Fisher, USA) by gravity flow with 50mM Tris-HCl pH 8.0, 100mM NaCl, 10% glycerol, imidazole (40mM for 1^{st} wash, 20mM for 2^{nd} and 3^{rd} wash), and 20mM BME. Resin-bound Mut9 was washed with buffer containing 50mM HEPES pH 7.5, 150mM NaCl, 0.5Mm EDTA, and 10% glycerol. TEV protease (EMD Chemicals, Promega) was added to Halo-tagged Mut9 for 1hr at room temperature. TEV protease was removed via mix with Ni-NTA resin for 30min at 4˚C, followed by 5min centrifugation. Supernatant was collected and concentrated with Amicon Ultra centrifugal filters (Millipore, USA). Bound Δ22 was eluted from the resin with a buffer containing 20mM Tris-HCl (pH 8.0), 500mM NaCl, 200mM imidazole, 20mM BME, and 0.1mM PMSF. Mut9 was eluted from the resin with buffer containing 50mM HEPES pH 7.5, 150mM NaCl, 0.5Mm EDTA, and 10% glycerol. Purified proteins were concentrated with Amicon Ultra centrifugal filters (Millipore, USA) and concentrations were determined by absorbance at UV_{280}. Purified protein purities were assessed by SDS-PAGE.

Fluorescence Polarization

Fluorescence polarization was performed in 384-well format using a BioTek Synergy2 fluorometer (Winooski, VT) with fluorescein-labeled probes containing 0-22 consecutive GGAA motifs and 1x Gel Shift binding buffer (Promega Corporation, Madison, WI). Sequences of the consecutive GGAA motifs harboring probes and control sequences used in these assays are listed in Table 4.2. Recombinant proteins were prepared as described above.

For binding assays, recombinant Δ22 or Mut9 protein concentrations were varied in individual wells in triplicate for each concentration, and mixed with DNA at a final concentration at least 10-fold below the expected K_D. Samples were incubated at room temperature for at least 25 minutes prior to measuring polarization. Polarization values (mP) were measured and plotted as a function of Δ22 or Mut9 protein concentrations. The free and total protein concentrations were assumed to be equal because the DNA concentration is at least 10-fold lower than the K_D. The total fluorescence intensities were conserved across all protein concentrations tested, indicating that the characteristics of the probe were stable across different protein concentrations. The affinity plots and curve fits were performed using the GraphPad Prism program, using the one-site total binding equation (GraphPad Software, LaJolla, CA). Each probe was tested in at least two independent experiments.

For stoichiometry experiments, polarization measurements were performed as above. DNA was mixed in binding buffer solution at a concentration 20-fold above the

determined K_D. Δ22 protein was added to the DNA solution in triplicate in 384-well plate format at different final molar ratios and equilibrated for at least 25 minutes before measuring polarization. Each probe was tested in at least two independent experiments. Polarization values were plotted as a function of the concentration ratio of protein versus DNA. The protein:DNA ratio at which an inflection in the data occurs represents the binding stoichiometry, as this is the point at which DNA is saturated by protein and polarization values change minimally as protein concentration increases.

Luciferase Assays

The pGL3-promoter luciferase vector (Promega, Madison, WI) was used for all experimental and control conditions. HEK 293EBNA cells were transfected with experimental reporter plasmid constructs or control plasmids, the Renilla plasmid and plasmids with Mut9, empty vector, or Mut9/R2L2 cDNA as previously described[16]. Firefly luciferase activity was normalized to *Renilla* luciferase activity to control for transfection efficiency and is reported in figures as "relative luciferase activity." Each experimental condition was performed in triplicate. Two-tailed *t* test was used for statistical comparison.

Supplemental Figures

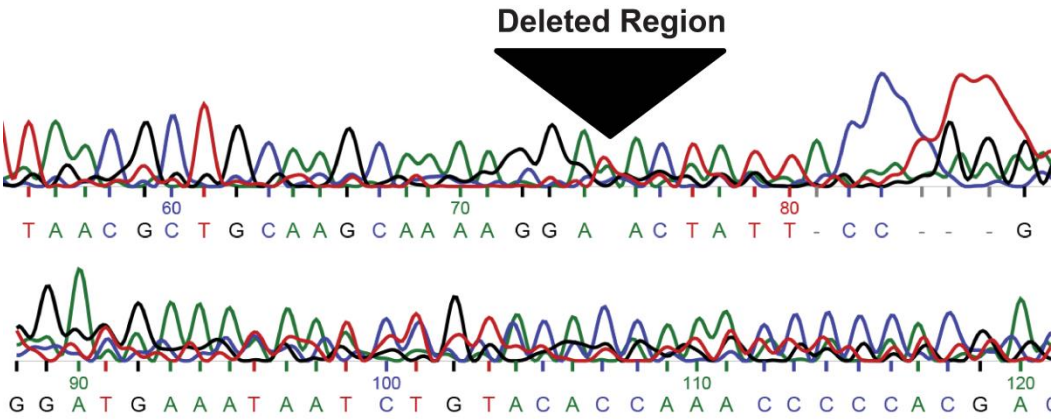


Figure 4.5 Sequencing results of the *NR0B1* microsatellite deletion. Sanger sequencing chromatogram of genomic DNA demonstrating deletion of the NR0B1 GGAA-microsatellite region in A673 cells

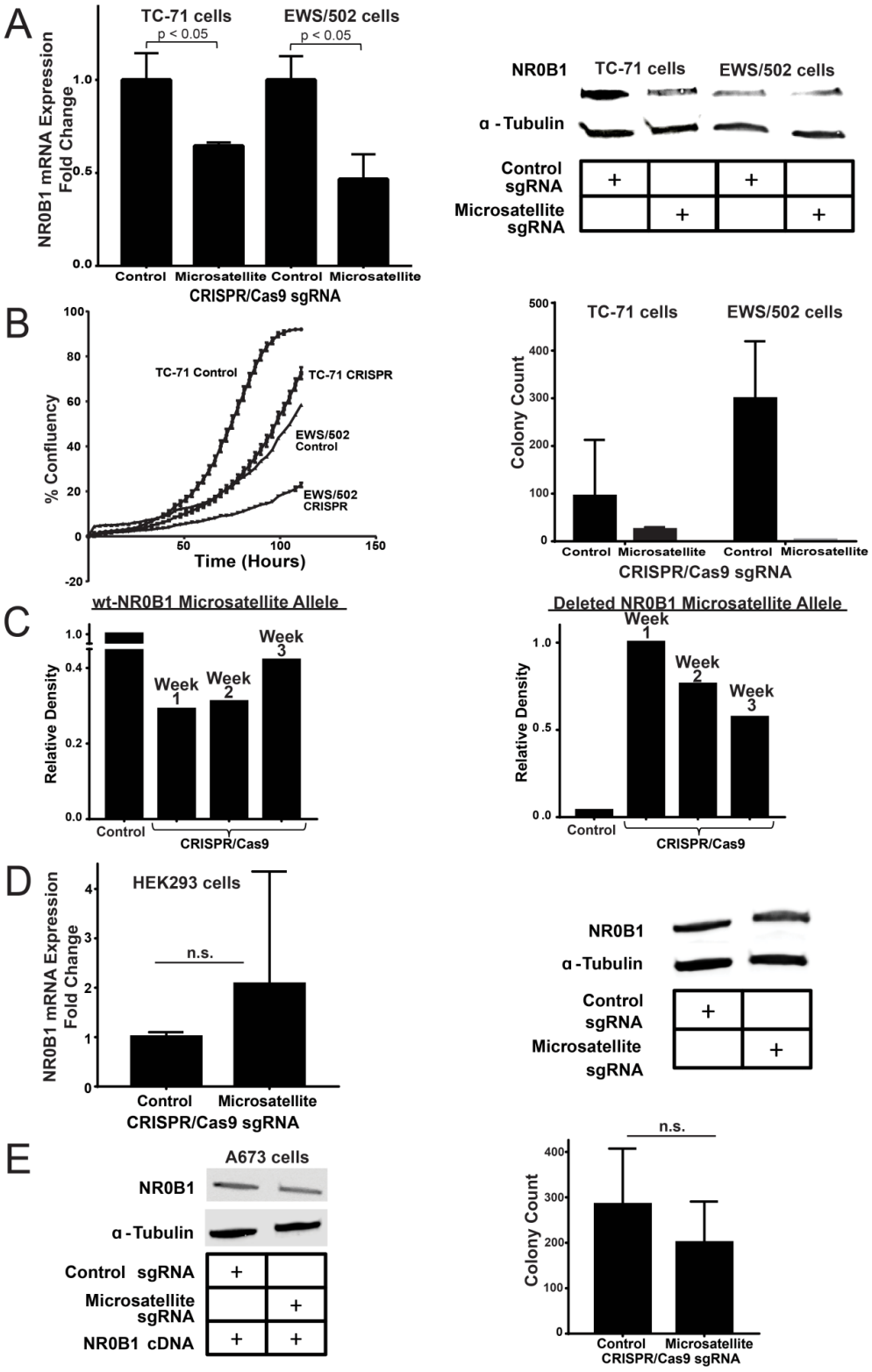


Figure 4.6 Deletion of the *NR0B1* microsatellite in other cell lines

Figure 4.6 continued

(A) NR0B1 mRNA and protein expression levels in control and CRISPR/Cas9-mediated knock-out of the *NR0B1* microsatellite in TC-71 and EWS/502 Ewing sarcoma cells ($p < 0.05$). Data is represented as mean ± SEM (n=3) **(B)** Growth curves and soft agar assay quantification for *NR0B1* microsatellite deletion in two other Ewing sarcoma cell lines (TC-71 cells and EWS/502 cells). Growth curve data is represented as mean ± SEM (n=4). Control vs. CRISPR for TC-71 and EWS/502 cells are each statistically significant ($p < 0.05$), denoted by asterisks. Soft agar data is represented as mean ± SEM (n=2) **(C)** Densitometry quantification of PCR-amplified *NR0B1*-microsatellite containing region for A673 control (wild type *NR0B1*-microsatellite) allele vs. CRISPR-Cas9 knock-out (deleted *NR0B1*-microsatellite) allele at different time points for up to 3-weeks post-lentiviral infection **(D)** NR0B1 mRNA and protein expression levels in control and CRISPR/Cas9-mediated knock-out of the *NR0B1* microsatellite in non-Ewing sarcoma HEK293 cells. Data is represented as mean ± SEM (n=3) (n.s. = not statistically significant) **(E)** NR0B1 protein levels and colony formation assay quantification for A673 cells with *NR0B1* cDNA rescue in CRISPR/Cas9 control vs. microsatellite knock-out (n.s. = not statistically significant). Data represented as mean ± SEM (n=2).

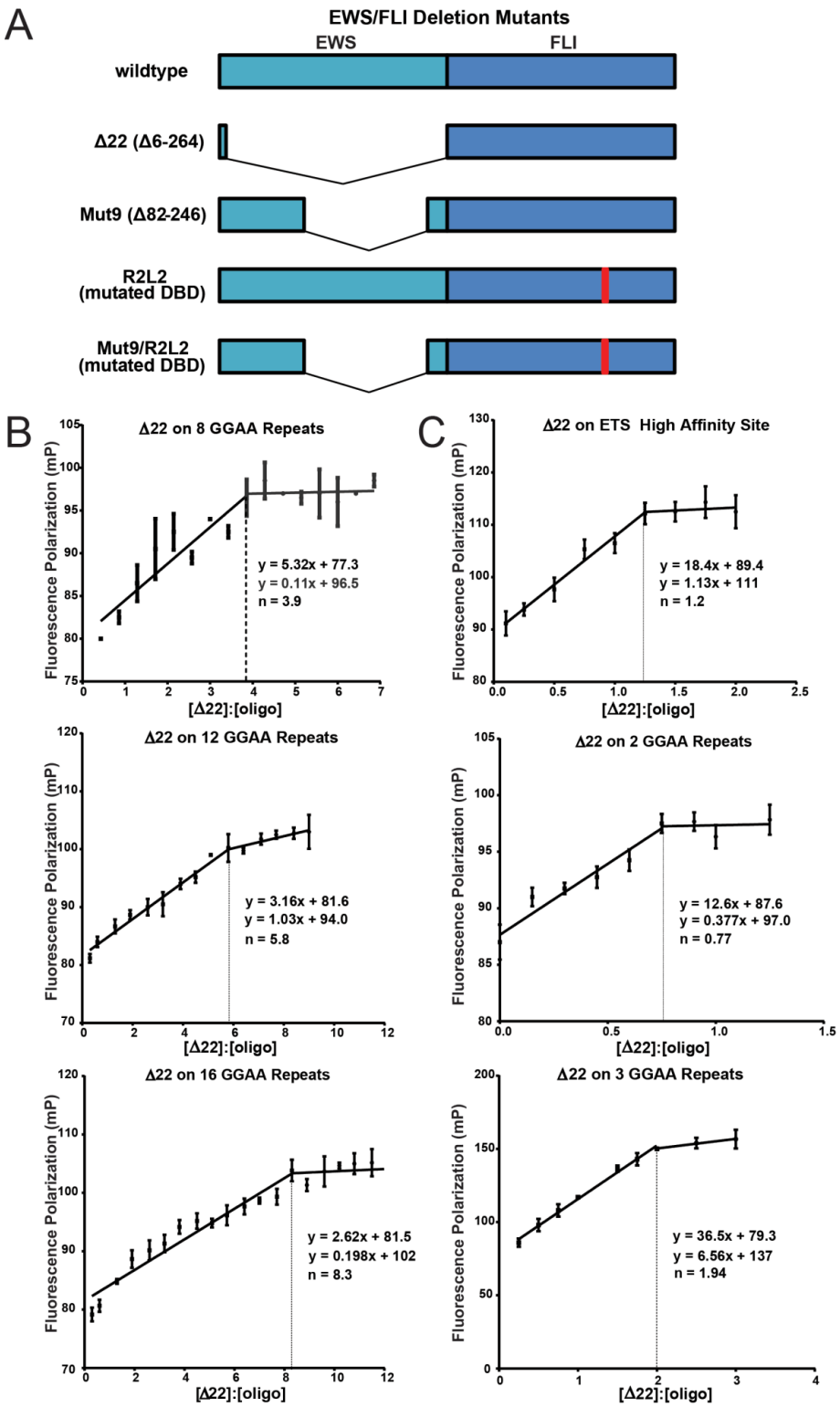


Figure 4.7 Stoichiometry of Δ22 binding on DNA sequences of increasing GGAA-microsatellite numbers.

Figure 4.7 continued

(A) Schematic of EWS/FLI constructs used in these studies **(B)** Fluorescent anisotropy results of recombinant Δ22 protein binding to fluorescein-labeled DNA probes containing 8, 12, and 16 consecutive GGAA-repeats, respectively. The inflection points represent stoichiometric ratios and are indicated by dashed lines (exact values indicated on graphs as n=3.9 for 8 repeats, n=5.8 for 12 repeats, etc.). Data is represented as mean ± SEM (n=3) **(C)** Stoichiometry of recombinant Δ22 protein binding at the conserved ETS high-affinity sequence (one 'GGAA'), 2, and 3 consecutive GGAA-repeats, respectively, to characterize minimal binding of Δ22. Dashed lines and n-values on graphs represent inflection points, or stoichiometric ratios. Data is represented as mean ± SEM (n=3).

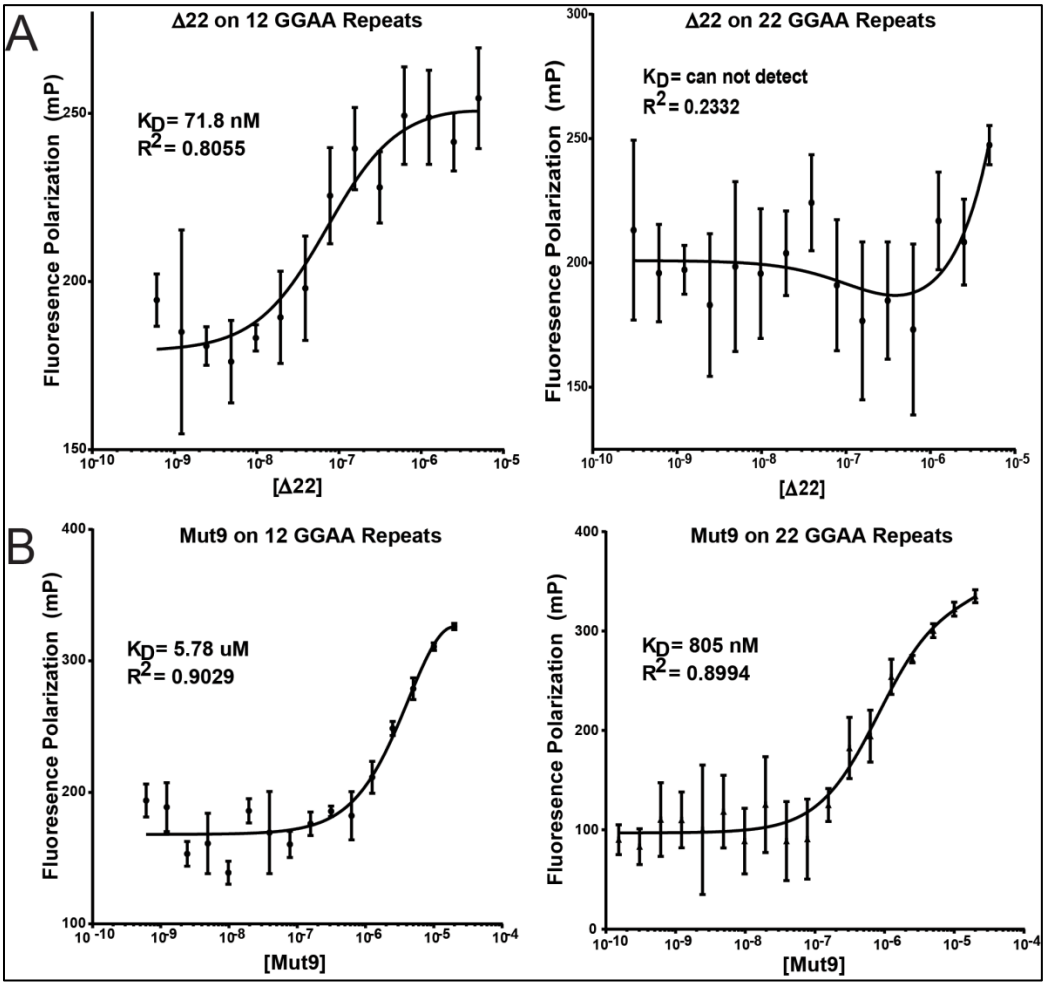


Figure 4.8 Mut9 vs. Δ22 binding at increasing GGAA-microsatellite lengths. Fluorescence anisotropy binding of recombinant **(A)** Δ22 or **(B)** Mut9 proteins on fluorescein-labeled DNA oligonucleotides containing 12 vs. 22 ("sweet-spot") consecutive GGAA repeats. Data is represented as mean ± SEM (n=3).

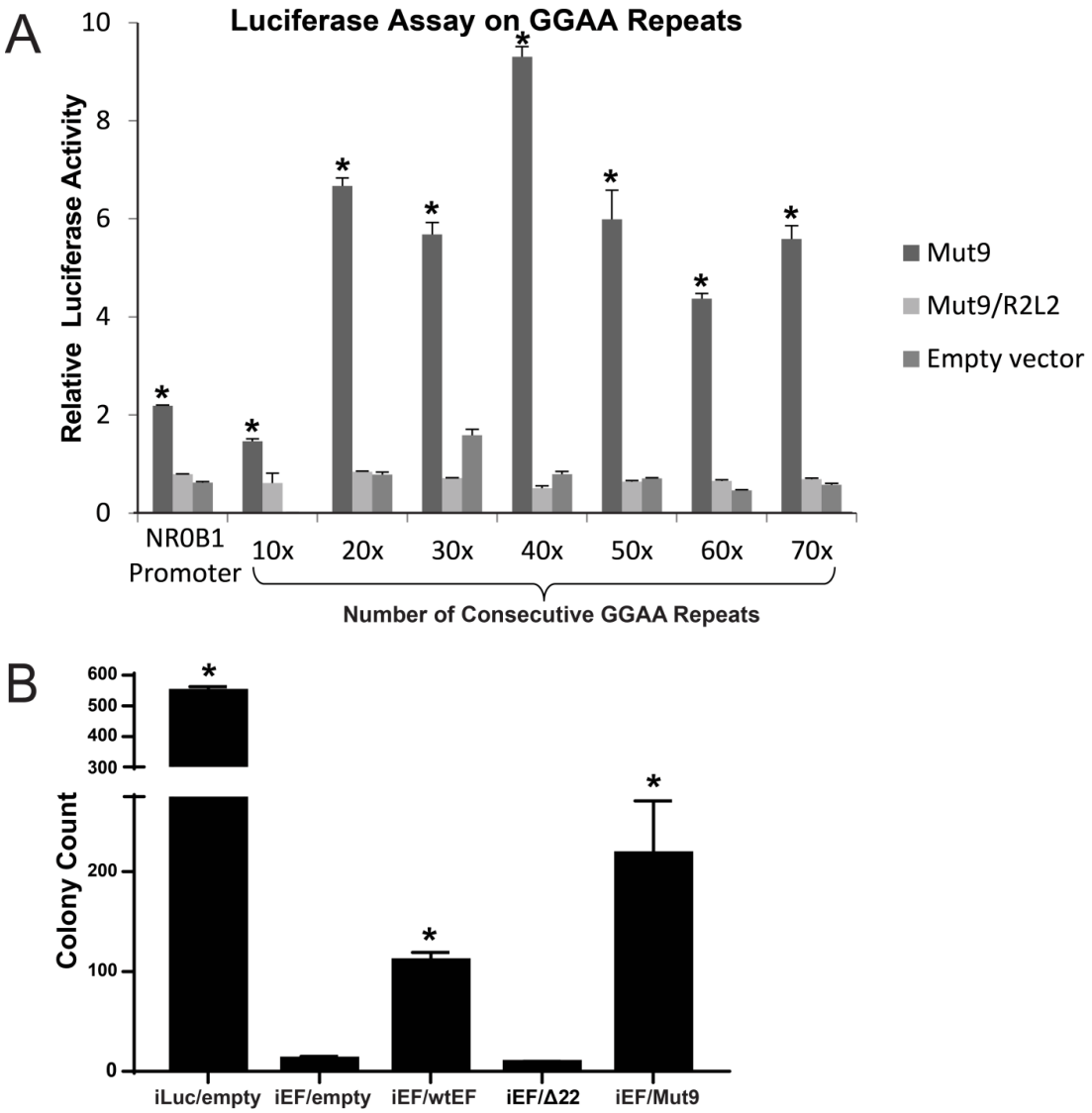


Figure 4.9 Mut9 transcriptional activation of increasing consecutive GGAA repeats

(A) Luciferase assay results showing transcriptional activity of Mut9 and Mut9/R2L2 vs. empty vector at microsatellites of increasing GGAA repeat numbers. Mut9 induced activity is significantly greater than Mut9/R2L2 and the empty vector control ($p < 0.01$), and are denoted by asterisks. Data is represented as mean ± SEM (n=3). **(B)** Quantification of colony formation assays for EWS/FLI knock-down and rescue cells used in RNA-seq experiments. Data represents mean ± SD (n=2). Paired 2-tailed t-tests were performed between constructs that rescue anchorage independent growth (Luc-RNAi, EWS/FLI, and Mut9) and those that do not (iEF-2-RNAi, Δ22, and R2L2). Asterisks denote statistical significance ($p < 0.05$).

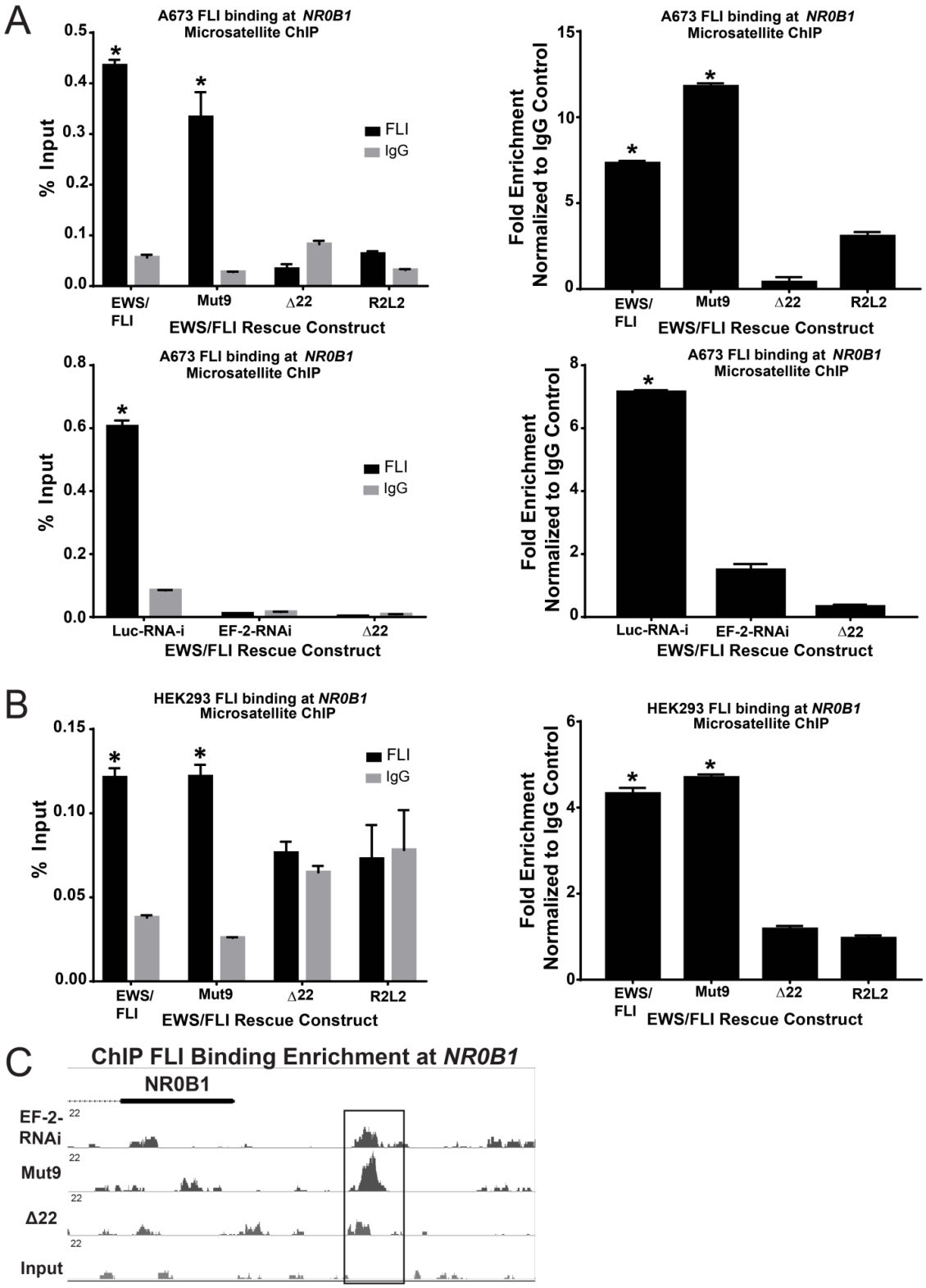


Figure 4.10 Mut9 and Δ22 binding at *NR0B1* microsatellite.

Figure 4.10 continued

qPCR shows binding enrichment expressed as % input and fold change normalized to IgG mock control at the *NR0B1* microsatellite following ChIP using an antibody against FLI vs. IgG in:

(A) A673 cells with control (Luc-RNAi) or EWS/FLI knockdown ("iEF-2-RNAi") rescued with the indicated EWS/FLI mutant constructs or empty-vector ("EF-2-RNAi") and **(B)** non-Ewing sarcoma HEK 293 EBNA cells infected with mutant EWS/FLI constructs. Data is represented as mean ± SEM (n=3). Asterisks denote statistical significance ($p < 0.05$) as assessed by paired two-tailed t-tests, between EWS/FLI wild type, Mut9, and Luc-RNAi respectively vs. iEF-2-RNAi, Δ22, and R2L2 respectively **(C)** A representative example from ChIP-seq data, showing FLI binding enrichment at the *NR0B1* microsatellite in EWS/FLI-depleted cells A673 cells rescued with empty vector, Mut9, or Δ22.

Supplementary Tables

Locus	Forward Primer	Reverse Primer
NR0B1 (ChIP)	GCTATTTAGGGCCTCTCACAGG	CCAGGACCTGGAATTCTCAGC
GAPDH (qRT-PCR)	CCGAGCCACATCGCTCAGACA	GCCTTCTCCATGGTGGTGAAG
NR0B1 (qRT-PCR)	GGGGACCGTGCTCTTTAACC	CTGACTGTGCCGATGATGG
NR0B1 (gDNA)	TGTCTATTTAGGGCCTCTCACA	TCAGGAGTACATGTGCGGGT
NR0B1 sgRNA	TAACGCTGCAAGCAAAATGGGGGG	CCCAGGTAATAGTATCCAATAGG
Control sgRNA	No sgRNA sequence	No sgRNA sequence

Table 4.1 Sequences for primers used in ChIP and qRT-PCR, and sgRNA sequences used in CRISPR/Cas9 experiment

DNA Oligo	Sequence (Forward Strand of Duplex)
FlrHighAffinity	56-FAM/TT TAC CGG AAG TGT TT
Flr2 Repeats	56-FAM/TT TTG GAA GGA ATT TT
Flr3 Repeats Scramble	56-FAM/TT GAG AGA GAG AGA TT
Flr3 Repeats	56-FAM/TTGGAAGGAAGGAATT
Flr4 Repeats	56-FAM/TTGGAAGGAAGGAAGGAA
Flr8 Repeats	56-FAM/TTGGAAGGAAGGAAGGAAGGAAGGAAGGAA
Flr12 Repeats	56-FAM/TTGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAA
Flr16 Repeats	56-FAM/TTGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAA
Flr20 Repeats	56-FAM/TTGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAA
Flr22 Repeats	56-FAM/TTGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAA

Table 4.2 Sequences for fluorescently-labeled oligos used in FP experiments

Chapter 5: Identification of two types of GGAA-microsatellites and their roles in EWS/FLI binding and gene regulation in Ewing sarcoma

Kirsten M. Johnson*, Cenny Taslim*, Ranajeet S. Saund, Stephen L. Lessnick. (2017) *PLoS ONE*. 12(11): e0186275. Doi: 10.1371/journal.pone.0186275. * These authors contributed equally to this work.

KMJ and CT designed and performed experiments, and wrote the document. RSS performed the ChIP-seq experiment. SLL helped design the experiments and reviewed the document.

Abstract

Ewing sarcoma is a bone malignancy of children and young adults, frequently harboring the EWS/FLI chromosomal translocation. The resulting fusion protein is an aberrant transcription factor that uses highly repetitive GGAA-containing elements (microsatellites) to activate and repress thousands of target genes mediating oncogenesis. However, the mechanisms of EWS/FLI interaction with microsatellites and regulation of target gene expression is not clearly understood. Here, we profile genome-wide protein binding and gene expression. Using a combination of unbiased genome-wide computational and experimental analysis, we define GGAA-microsatellites in a Ewing sarcoma context. We identify two distinct classes of GGAA-microsatellites and

demonstrate that EWS/FLI responsiveness is dependent on microsatellite length. At close range "promoter-like" microsatellites, EWS/FLI binding and subsequent target gene activation is highly dependent on number of GGAA-motifs. "Enhancer-like" microsatellites demonstrate length-dependent EWS/FLI binding, but minimal correlation for activated and none for repressed targets. Our data suggest EWS/FLI binds to "promoter-like" and "enhancer-like" microsatellites to mediate activation and repression of target genes through different regulatory mechanisms. Such characterization contributes valuable insight to EWS/FLI transcription factor biology and clarifies the role of GGAA-microsatellites on a global genomic scale. This may provide unique perspective on the role of non-coding DNA in cancer susceptibility and therapeutic development.

Introduction

Ewing sarcoma is the second most common pediatric bone malignancy, initiated by a chromosomal translocation t(11;22)(q24;q12), creating the fusion protein and oncogenic driver EWS/FLI. As an aberrant transcription factor, EWS/FLI plays a critical role in regulating genes involved in tumorigenesis[38]. Typically, FLI and other ETS family members bind DNA via their conserved DNA binding domain at the consensus sequence 'ACCGGAAGTG'[48,63]. This high affinity DNA binding site containing a single GGAA core motif is necessary for oncogenesis[6,64,66], with FLI and EWS/FLI displaying similar DNA binding affinity and specificity[37]. In Ewing sarcoma, however, EWS/FLI displays a

"gain-of-function" in its ability to also bind 'GGAA'-containing microsatellite (repeat) regions to regulate some of its targets, such as key oncogenic target *NR0B1*[15,74].

Microsatellites are tandem, or sequentially repeated DNA motifs, frequently found in or near gene promoters[105,108]. In Ewing sarcoma, repetitive "microsatellite" regions comprised of the motif "GGAA" have been identified as highly enriched EWS/FLI-bound sequences near transcription start sites of EWS/FLI up-, but not down-regulated genes[15,46]. We and others confirmed that these putative binding sites specifically confer EWS/FLI-mediated activation of their adjacent target[15,46,69,70]. Additionally, we recently demonstrated a relationship between the number of repeats in these regions and their ability to function as EWS/FLI-response elements: an 18-26 GGAA-motif "sweet-spot" repeat length provides maximal transcriptional function, and is significantly enriched in patients with Ewing sarcoma[33]**.** How polymorphisms of GGAA-microsatellites in Ewing sarcoma affect EWS/FLI binding and transcriptional regulation across the genome, however, remains unclear.

Although these GGAA-containing regions fall under the traditional definition of "microsatellites," this term has been loosely applied in a Ewing sarcoma context to include a wide-range of "GGAA" sequences and is somewhat arbitrary, especially given their polymorphic nature[16]. Clearly defining GGAA-microsatellites in a Ewing sarcoma relevant context is needed to understand their mechanistic role in EWS/FLI transcription factor regulation. Additionally, delineating a clear relationship between microsatellite

length, location and transcriptional regulation across the genome is essential. Together, these disparities represent a significant void in our understanding of EWS/FLI transcriptional biology, and remain a powerful barrier to potential therapeutic amelioration. Our previous demonstration of GGAA-microsatellites as EWS/FLI response elements, coupled with *in vitro* and clinical data indicating a "sweet-spot" length, suggest a relationship between EWS/FLI and these unique binding sites in transcriptional activation[15,16]. Here, we sought to define GGAA-microsatellites in a Ewing sarcoma context, and to understand their role across the genome.

To accomplish this, we use bioinformatics analysis of experimental data to first characterize GGAA-microsatellites, setting pre-determined parameters for an unbiased genome-wide approach. Once described, we then computationally link bound microsatellites to adjacent EWS/FLI regulated genes. Our data reveal two distinct types of GGAA-microsatellites: close-range ("promoter-like") and long-range ("enhancer-like"), and suggest differing mechanisms of EWS/FLI-mediated activation and repression at these elements. Classification of these clarifies the genome-wide presence of GGAA-microsatellites in Ewing sarcoma and their role in transcriptional regulation.

Materials and Methods

Cell culture

The Ewing sarcoma cell line A673 from ATCC was cultured, and retroviruses packaged in HEK293-EBNA cells, using standard procedures described previously[175,201]. For RNA interference experiments, cells were infected with pMSCV-puro retrovirus harboring shRNA constructs against luciferase (control) or EWS/FLI.

Searching for GGAA repeat regions

Human reference genome (hg19) was scanned to find the occurrences of GGAA and TTCC using Biostrings[202] and BSgenome[203] R packages. An in-house script was used to find a region that contains multiple GGAA-motifs not separated by more than 20 non-GGAA nucleotides. The region has to start and end with GGAA. The same procedure was used to find repeat regions with TTCC-motifs. Each region was then annotated with its nearest gene (pseudo genes were filtered from annotation database) using ChIPpeakAnno[204] R package.

ChIP-seq analysis

Chromatin immunoprecipitation (ChIP) was performed as previously described[205] using anti-FLI-1 (Santa Cruz, sc-356X Santa Cruz Biotechnology, Inc.). Briefly, chromatin from formaldehyde-fixed A673 cells was fragmented to a size range of 200-700 bases with a Misonix Sonicator. Solubilized chromatin was immunoprecipitated with anti-FLI-1 and antibody-chromatin complexes were pulled down with M-280 sheep anti-rabbit IgG Dynabeads (Thermo Scientific), washed and then eluted. After crosslink reversal, RNAse A and Proteinase K treatment, immunoprecipitated DNA was extracted with the

Mini-Elute PCR purification kit (Qiagen). ChIP DNA was quantified with Qubit, libraries prepared and sequenced with Illumina HiSeq 2500. Raw sequence reads can be found in NCBI's Gene Expression Omnibus database under GSE99959. Sequence reads were aligned to the human reference genome (hg19) using Novoalign (http://novocraft.com). Duplicate reads were removed using samtools[206]. Peaks were identified using MACS2[207] at FDR cut-off of 5%. To assess whether GGAA-repeat regions overlap with EWS/FLI binding sites more than one would expect by chance, we used permutation tests implemented in regioneR R library[208]. Overlap is defined as region with ≥ 1 bp overlap. Specifically, we compared the number of overlap in the actual EWS/FLI binding sites and GGAA-repeat regions (with at least 3 consecutive repeats) to that seen in a random sample of universe regions (i.e. resampleRegions strategy in regioneR library). Since GGAA-repeat regions with at least three consecutive motifs are a subset of all repeat regions in the genome, we used all repeat regions as the universe regions. This randomization strategy maintained the internal structure of GGAA-repeats. A different randomization strategy (i.e. randomizeRegions) which randomly places repeat regions along the mappable regions of the genome was also performed with similar results (data not shown). Although significant, the association between EWS/FLI binding sites and GGAA-repeat regions with at least three consecutive motifs might be indirect and based on the fact that both regions tend to cluster around gene-rich regions. In order to check whether this association is specifically linked to the relative position of these two regions with each other, we shifted the regions and evaluated the z-score for every shifted position. The sharp peak, as shown in S1C Fig, indicates this association is highly

dependent on the relative position of the two regions with each other, and the association is not regional. In order to do a correlation test, we associated each microsatellite with its nearest EWS/FLI peak binding sites (distance is calculated from the middle of the microsatellite to EWS/FLI peak summit location). Correlation coefficients (*r*) were calculated using Spearman's correlation.

RNA-seq analysis

The RNA-seq data set used in this work was previously published[209]. Briefly, RNA collected from A673 cells stably infected and selected for expression of a control Luc-RNAi or the EF-2-RNAi was extracted using the RNAeasy kit (Qiagen) with an on-column DNAse digestion protocol. Libraries for deep-sequencing were prepared according to the manufacturer's instructions (Illumina) and sequenced on an Illumina Hi-Seq 2000 with 50-bp single end reads. Sequences were aligned to the human genome build hg19 using Novoalign (http://novocraft.com). Raw sequence reads can be found in the NCBI SRA under SRA059239. Gene model used for counting reads/fragments were from Ensembl GRCh37 (release 75) GTF[210]. R packages GenomicAlignments[30], GenomicFeatures[30] and BiocParallel[211] were used to count the number of reads/fragments assigned to genomic features in each sample. Genomic features with total counts less than 2 across samples were removed. Data quality was assessed by clustering all samples. Normalized rlog (regularized log transformation) counts[212] and pheatmap[213] R package were used to do hierarchical clustering. Supplementary Figure 16 shows a heatmap of sample-to-sample distance. Differential gene analysis was done using DESeq2, which

uses negative binomial modeling and the empirical Bayes shrinkage method for fold-change estimation[212].

Correlations between EWS/FLI binding intensities, EWS/FLI-regulated gene expressions and microsatellites

All **c**orrelations were calculated using Spearman's rank correlation. LOESS regression (Local Polynomial regression fitting) line and its t-based approximation of 95% confidence bands were drawn using R library ggplot2[214]. We used Loess regression because of its advantage as robust to outliers and its ability to show non-linear association[215].

Data availability

RNA-seq raw sequence reads can be found in the NCBI SRA under SRA059239. ChIP-seq raw sequence reads can be found in NCBI's Gene Expression Omnibus database under GSE99959

Results

GGAA-motifs in a microsatellite occur on the same strand

EWS/FLI, the aberrant transcription factor in Ewing sarcoma, modulates gene expression by binding to GGAA-containing repetitive regions[15]. However, genome-wide characterization of these repeat regions is lacking, including whether microsatellites with GGAA-motifs are present on both strands of DNA. We first scanned the human reference

genome (hg19) on both strands for GGAA-motifs. We defined a repeat region as a sequence that starts and ends with a GGAA-motif and which has no more than 20 insertions (non-motif nucleotides) between two adjacent motifs (Figure 5.1A). Nearly 5 million repeat regions span the genome. Although the total number of motifs in any given region ranges from 2 to 266 motifs, 3.7 million regions contain less than 3 motifs (Figures 5.1B-C). These sparse repeat regions have an average GGAA content, or density, of around 50% (Figures 5.1D-E). Additionally, most of these repeat regions have no consecutive motifs (93.2%) and less than 0.6% has at least 3 consecutive motifs.

Given typical FLI binding within the major groove of DNA at GGAA-containing regions[55], we considered the possibility of GGAA-motifs existing on both strands of DNA within the same repeat region. Multiple EWS/FLI molecules could conceivably bind adjacent motifs on alternating strands of DNA and thereby avoid steric hindrance in binding[66]. We classified repeat regions that contain GGAA-motifs on the same strand as pure repeat regions, while regions that include GGAA on both forward and reverse strands are referred to as mixed repeat regions (Figure 5.1A). We determined more than half of the 5 million GGAA-repeat regions across the genome (55%) contain GGAA-motifs on the same strand (Figure 5.1F).

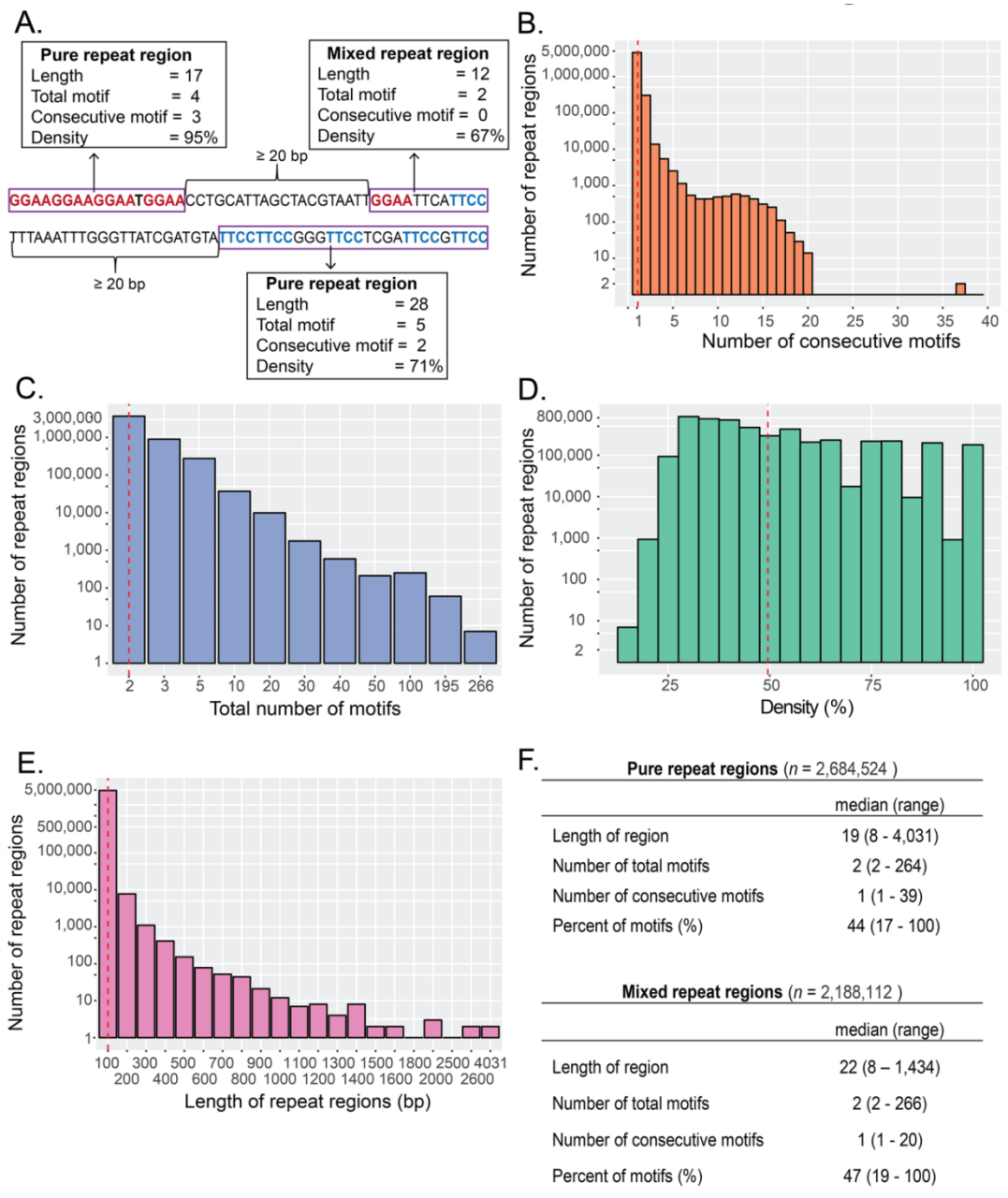


Figure 5.1 Schema and characteristics of repeat regions across genome

(A) Schema of repeat regions. Regions with only one type of motif are called pure repeat region while those with both GGAA and TTCC are called mixed repeat regions. Each repeat region (purple box) is separated by at least 20-bp consecutive non-motifs. **(B)** Histogram of maximum number of consecutive motifs. **(C)** Histogram of total number of motifs. **(D)** Histogram of motif density of repeat regions. $Density = \left(\frac{total\ number\ of\ motifs \times 4}{length\ of\ regions}\right) \times 100\%$. Bin width is 5%. **(E)** Histogram of

Figure 5.1 continued

length of repeat regions. Each bin is 100bp width (e.g., first bin is 0-100bp length). Bins with zero repeat regions are not shown. **(F)** The characteristics of repeat regions for pure and mixed repeat regions across the genome. Red line indicates the mean for each characteristic.

To determine whether a microsatellite can contain mixed motifs, we looked specifically at mixed repeat regions with at least 3 or more consecutive motifs. We found more than 81% of them contain only a single GGAA-motif on the opposite strand. While only 32 regions (1.2% of 2,589) have 2 or more consecutive GGAA-motifs on both strands in the same region, even in these rare examples motifs cluster together on the same strand. Additionally, only one region has more than 2 consecutive GGAA-motifs on both strands (Supplementary Table 5.2). Based on these observations, we deduced *bona fide* microsatellites with GGAA-motifs on both strands may not exist in the same region. This finding prompted us to re-process mixed repeat regions, separating clusters of GGAA-motifs as two distinct regions if they are on opposite strands. Thus we discounted repeat regions with only one GGAA-motif on each strand, leaving 3,321,889 repeat regions. We focus our downstream analysis solely on these homogenous (i.e. same strand) repeat regions. For ease of reading, henceforth we will refer to these GGAA-motifs simply as repeat regions.

Longer GGAA-regions are located near genes while shorter GGAA-regions are ubiquitous across the genome

Of the more than 3 million repeat regions in the genome, we found 99% of them contain only two consecutive motifs. In many of these regions, these motifs likely happen by chance and consequently have no function. A subset of these regions, however, may act as EWS/FLI response elements, driving regulation of critical oncogenic gene targets such as *NR0B1*[15]. To facilitate functional analysis of these repeat regions in an unbiased approach, we started by annotating each repeat region with its nearest genes and observing the distribution of these repeat regions in terms of both their nearest genes and genomic location. The nearest gene is the gene with the shortest distance from the center of the GGAA-microsatellite to the transcription start site (TSS), regardless of strand direction (Figure 5.2A). Most GGAA-regions occur within 3Mb of a gene. Notable exceptions include 1,355 regions with 1 or 2 consecutive motifs and a single 3-consecutive motif region that are greater than 30Mb away from a gene (Figure 5.2B).

Although many repeat regions with two or less consecutive motifs reside near the TSS, on average most are farther away from genes compared to regions with longer consecutive motifs (t-tests, $p < 0.05$) (Figure 5.2C). Conversely, we found that repeat regions with 10-11 consecutive motifs are closer on average to genes than other consecutive motifs and are slightly enriched in promoter regions of genes 2 to 3kb from the TSS (Figures 5.2C-D). This enrichment of repeat regions with 10-11 consecutive motifs near genes suggests possible preferential binding of EWS/FLI at these microsatellites.

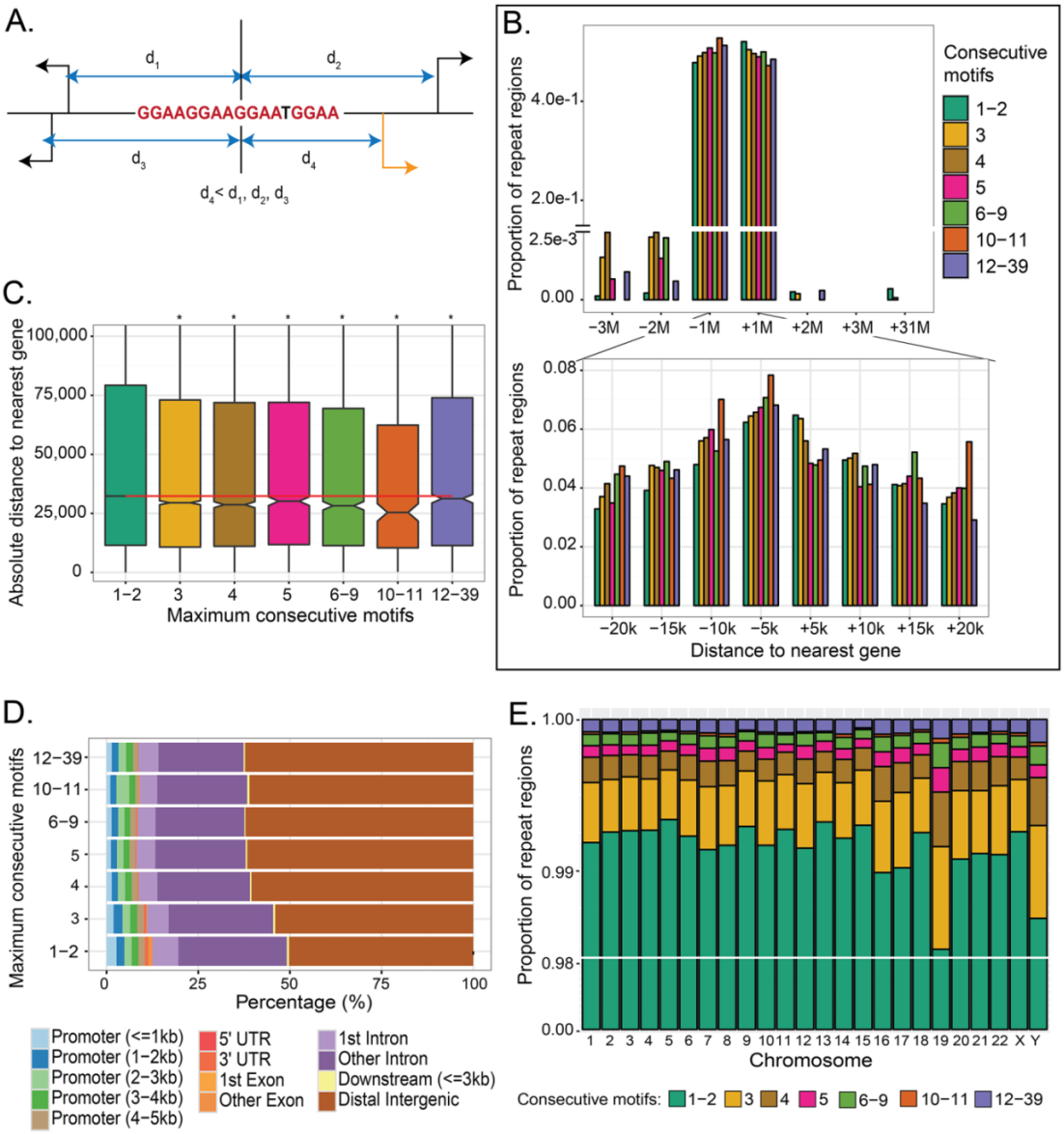


Figure 5.2 Nearest gene schema and genomic location of repeat regions

(A) Schema showing the nearest gene (orange) which is the gene with the shortest distance calculated from its TSS to the middle of the repeat region. **(B)** Distribution of distances to nearest genes for each repeat region grouped by number of consecutive motifs. The sum of percentages for each consecutive motif is 100%. **(C)** Comparisons of distance-to-nearest-gene for longer consecutive motifs to repeat regions with one to two consecutive motifs (i.e. '1-2'). * indicates the repeat regions are significantly closer to a gene than repeat regions with 1-2 consecutive motifs ($p < 0.05$). Red line represents the median distance-to-nearest gene for repeat regions with 1-2 consecutive motifs. **(D)** Feature distribution for each consecutive motif category. **(E)** Proportions of repeat regions in each chromosome grouped by the number of consecutive motifs.

Looking by individual chromosome, we demonstrated that repeat regions with 1-2 consecutive motifs account for 99% of GGAA-containing regions (Figure 5.2E). Interestingly, chromosome 19 has a higher proportion of longer consecutive motifs (more than 3 consecutive motifs) than the other chromosomes. We later found chromosome 19 also has a greater number of EWS/FLI peaks (see later discussion on EWS/FLI binding). Overall, our data indicate that short consecutive repeat regions (less than 3 consecutive motifs) may not have any EWS/FLI related function as they are ubiquitously scattered throughout the genome. We therefore investigated whether a GGAA-microsatellite needs to have a minimum number of motifs to allow EWS/FLI binding in a Ewing sarcoma context.

EWS/FLI bound GGAA-microsatellites contain three or more GGAA-repeats

Our previous in-vitro data indicated a minimum of three consecutive GGAA-motifs is required for EWS/FLI binding[170]. To test this requirement computationally across the genome, we addressed the following question: Does significant overlap exist between repeat regions with certain lengths and EWS/FLI binding sites? We investigated these relationships using ChIP-seq experiments in the A673 Ewing sarcoma cell line. Four paired-end ChIP-seq samples immunoprecipitated with a FLI-specific antibody were analyzed using Model Based Analysis for ChIP-seq (MACS2)[216]. 22,744 EWS/FLI binding sites were identified at a False Discovery Rate (FDR) cut-off of 0.05. Chromosome 19, which has more repeat regions at three or more consecutive motifs, also has an increased number of EWS/FLI binding sites per Mb compared to the other

chromosomes (Supplementary Figure 5.6A). This further supports defining repeats of 3 or more consecutive motifs as EWS/FLI response elements. The total repeat regions that overlap with EWS/FLI binding sites are 26,922 (Table 5.1).

	EWS/FLI		no EWS/FLI		Total
GGAA-motifs	**n**	**%**	**n**	**%**	
All Repeat Regions	26,922	0.81%	3,294,967	99.19%	3,321,889
1 motif	15,615	0.51%	3,023,699	99.49%	3,039,314
2 consecutive motifs	3,051	1.19%	252,809	98.81%	255,860
3 consecutive motifs	1,570	12.14%	11,359	87.86%	12,929
4 consecutive motifs	1,536	28.29%	3,894	71.71%	5,430
5 consecutive motifs	978	38.76%	1,545	61.24%	2,523
≥ 6 consecutive motifs	4,172	71.52%	1,661	28.48%	5,833

Table 5.1 Number of GGAA-repeat regions by number of consecutive GGAA-motif and EWS/FLI binding sites across the genome.

To evaluate whether the amount of overlap between repeat regions and EWS/FLI binding sites occurs by chance, we assessed statistical significance with a permutation test. We observed repeat regions with three or more consecutive motifs overlap significantly with EWS/FLI binding sites ($p < 0.001$, Figure 5.3A), while repeat regions with two or less consecutive motifs do not overlap significantly ($p = 1$, Supplementary Figure 5.6B). This finding is consistent with our experimental observation that a minimum of three consecutive GGAA-motifs is required for EWS/FLI binding[170].When we randomly move the locations of repeat regions with longer consecutive motifs across the genome, we observe a sharp decrease in the statistical significance of overlap with EWS/FLI binding sites. This decrease in overlap indicates that the association is not regional but is highly

dependent on motif location (Supplementary Figure 5.6C). Based on the combination of these observations, we now define GGAA-microsatellites as repeat regions with 3 or more consecutive motifs. Downstream analyses focus on GGAA-microsatellites according to this definition.

Increasing number of GGAA-motifs correlates with increased EWS/FLI binding intensity

Having defined GGAA-microsatellites, we next investigated EWS/FLI binding at these regions to determine whether GGAA-motif enrichment in a given microsatellite region affects binding of EWS/FLI at that genomic loci. We defined the closest microsatellite to the center of the EWS/FLI binding site as the putative EWS/FLI-bound microsatellite. We grouped the 6,031 EWS/FLI-bound microsatellites (Supplementary Table 5.3) by number of consecutive motifs (Supplementary Figure 5.1D). Since we found only 11 EWS/FLI-bound microsatellites with more than 20 consecutive motifs, statistical evaluation and inclusion of these data points were difficult and uninformative. We therefore excluded microsatellites with more than 20 consecutive motifs in this analysis.

Overall, we demonstrate a positive correlation between EWS/FLI binding intensity and the number of consecutive motifs contained by these EWS/FLI-bound microsatellites (r = 0.46, $p < 2.2\times10^{-16}$) (Figures 5.3B and Supplementary Figure 5.7). This genome-wide trend of overall increasing EWS/FLI binding enrichment with increasing number of consecutive motifs is consistent with our previous *in-vitro* study[51].

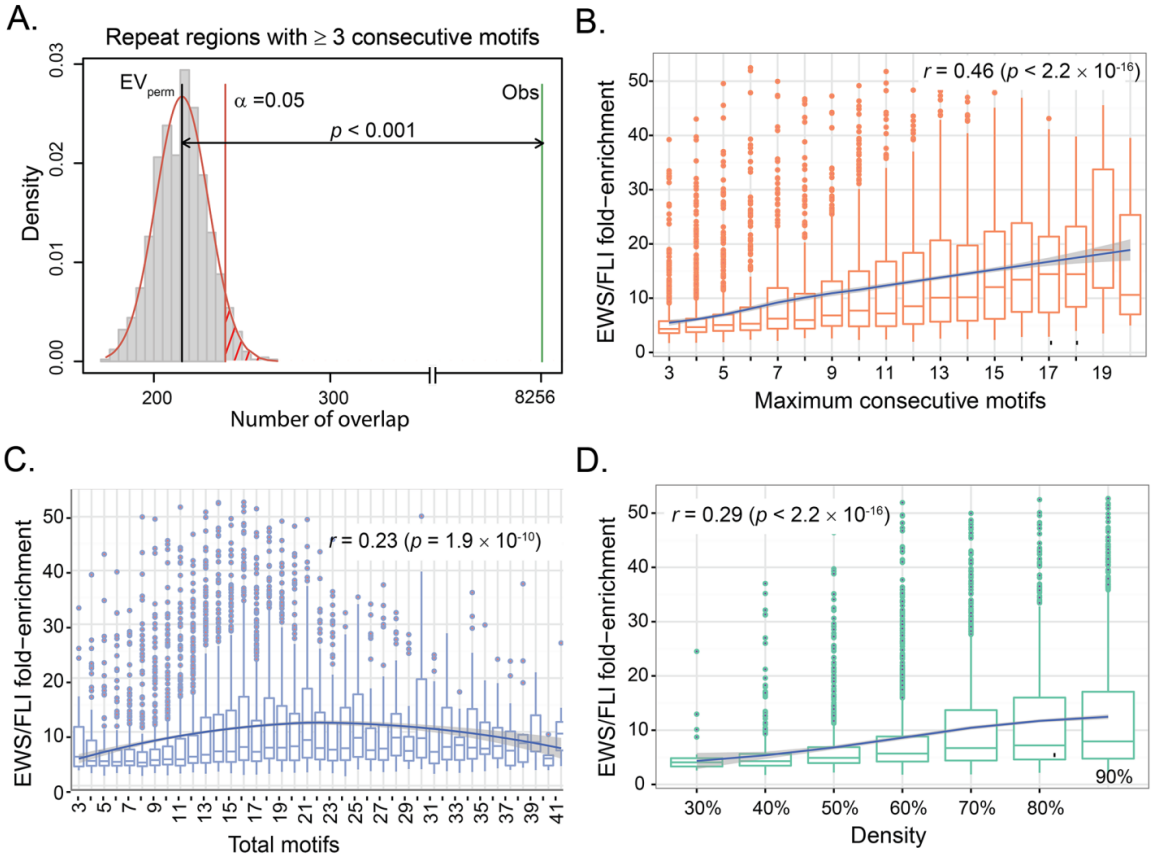


Figure 5.3 Characteristics of EWS/FLI-bound microsatellites

(A) Permutation test shows that the number of EWS/FLI binding sites that overlap with repeat regions (n = 8,256) with minimum of 3 consecutive motifs is significantly higher than random chance ($p < 0.001$). Red line denotes the significance limit ($\alpha = 0.05$). Gray bars represent the number of overlaps in the random regions with EWS/FLI binding sites in 1,000 permutations. The black line represents the mean of overlaps in random regions (EV_{perm}) and the green bar is the actual number of overlaps observed in repeat regions (Obs). **(B)** Boxplot of EWS/FLI fold-enrichment (relative to genomic background) and number of consecutive motifs in EWS/FLI-bound microsatellites showing statistically significant increasing trend ($p < 2.2 \times 10^{-16}$). The blue line is the estimated LOESS regression line of the mean with the estimated 95% confidence bands (shaded region). **(C)** Boxplot of EWS/FLI fold-enrichment and total number of motifs in EWS/FLI-bound microsatellites showing a positive correlation ($p = 1.9 \times 10^{-10}$) and a non-linear trend ($p < 0.05$). The blue line is the estimated LOESS regression line of the mean with the estimated 95% confidence bands (shaded region). **(D)** Boxplot of EWS/FLI fold-enrichment and Density $\left(= \frac{total\ motif \times 4}{length\ of\ microsatellite} \times 100\%\right)$ showing statistically significant positive correlation ($p < 2.2 \times 10^{-16}$). The blue line is the estimated LOESS regression line of the mean with the estimated 95% confidence bands (shaded region).

Most EWS/FLI-bound microsatellites have 11 to 19 total motifs with a maximum of 195 motifs (Supplementary Figure 5.8). We see a similar positive correlation between EWS/FLI binding and total motifs ($r = 0.23$, $p = 1.9 \times 10^{-10}$) as with consecutive motifs (Figure 5.3C). We also observe a non-linear relationship between EWS/FLI fold-enrichment and total motifs, with the EWS/FLI fold-enrichment increasing from 3 to about 16 total motifs, then decreasing again around 24-25 total motifs (LOESS regression) (Figure 5.3C and Supplementary Figure 5.9).

These data are in agreement with our recent finding that 18-26 motifs are the optimal length for EWS/FLI binding[16]. To see whether overall GGAA content within a microsatellite affects EWS/FLI binding, we then evaluated the relationship between GGAA-motif density within a microsatellite and EWS/FLI binding enrichment. We found that EWS/FLI fold-enrichment demonstrates a statistically significant positive correlation with GGAA-motif density ($r = 0.29$, $p < 2.2\times10^{-16}$) (Figure 5.3D). These EWS/FLI-bound microsatellite densities range from 30% to 90%, with most EWS/FLI-bound microsatellites (>1,500) having a density of 90% (Supplementary Figure 5.10).

Overall, our analysis shows a positive correlation between number of motifs and overall GGAA content, which increases with increased EWS/FLI binding. There is also a non-linear trend between EWS/FLI fold-enrichment and total motifs, implicating an optimal, or "sweet-spot", microsatellite length for EWS/FLI binding similar to our recent study[217]

.

EWS/FLI gene regulation at associated GGAA-microsatellites

In the previous section, we established the global correlation between microsatellite motif number and EWS/FLI binding intensities. Though this correlation allowed us to define GGAA-microsatellites in terms of length based on bound EWS/FLI, transcription factor binding is not always indicative of transcriptional regulation[217]. To determine whether GGAA-microsatellite characteristics are predictive of EWS/FLI responsiveness at a given genomic loci, we evaluated both the expression of EWS/FLI target genes and binding intensity associated with these microsatellites. We and others previously showed that EWS/FLI regulates its activated, but not repressed targets through binding at GGAA-microsatellites[170]. Prior analysis of these regions, however, has primarily focused on microsatellites located within about 5kb of associated EWS/FLI target promoters[15]. We therefore separately evaluated EWS/FLI activated and repressed targets associated with microsatellites both near (within 5kb) and distal (greater than 5kb) to the TSS of these genes. Differential gene expression profiles grouped based on these distinct categories of activated vs. repressed and close-range vs. distal microsatellites were integrated and stratified by distance of each GGAA-microsatellite to the nearest gene. Gene expression profiles were derived from six independent RNA-seq experiments on wild type vs. EWS/FLI knock-down A673 cells. DESeq2[212] identified 9,323 differentially expressed (4,278 activated and 5,045 repressed) genes between control and treatment cell lines at a FDR of 5%.

EWS/FLI binding and gene activation at promoter-like microsatellites is highly dependent on the length of GGAA-motifs

There are 114 microsatellites within 5kb of activated genes. To see if EWS/FLI binding at these close-range, promoter-like, microsatellites confer gene activation, we looked at EWS/FLI binding enrichment and gene expression for these microsatellites. As anticipated based on our previous studies, we found increased EWS/FLI binding correlates with expression of activated target genes (Figure 5.4A and Supplementary Figure 5.11) (r = 0.46, p = 3.3×10^{-7}). Furthermore, increasing number of consecutive GGAA-motifs correlates with increased EWS/FLI binding intensity (r = 0.43, p = 1.5×10^{-6}) and also increases in subsequent gene activation (r = 0.23, p = 0.01) (Figure 5.4B-C). EWS/FLI binding and total GGAA-motif number and density, demonstrate a trend toward positive correlation, though not significant for these microsatellites (Figures Supplementary Figure 5.12A-C).

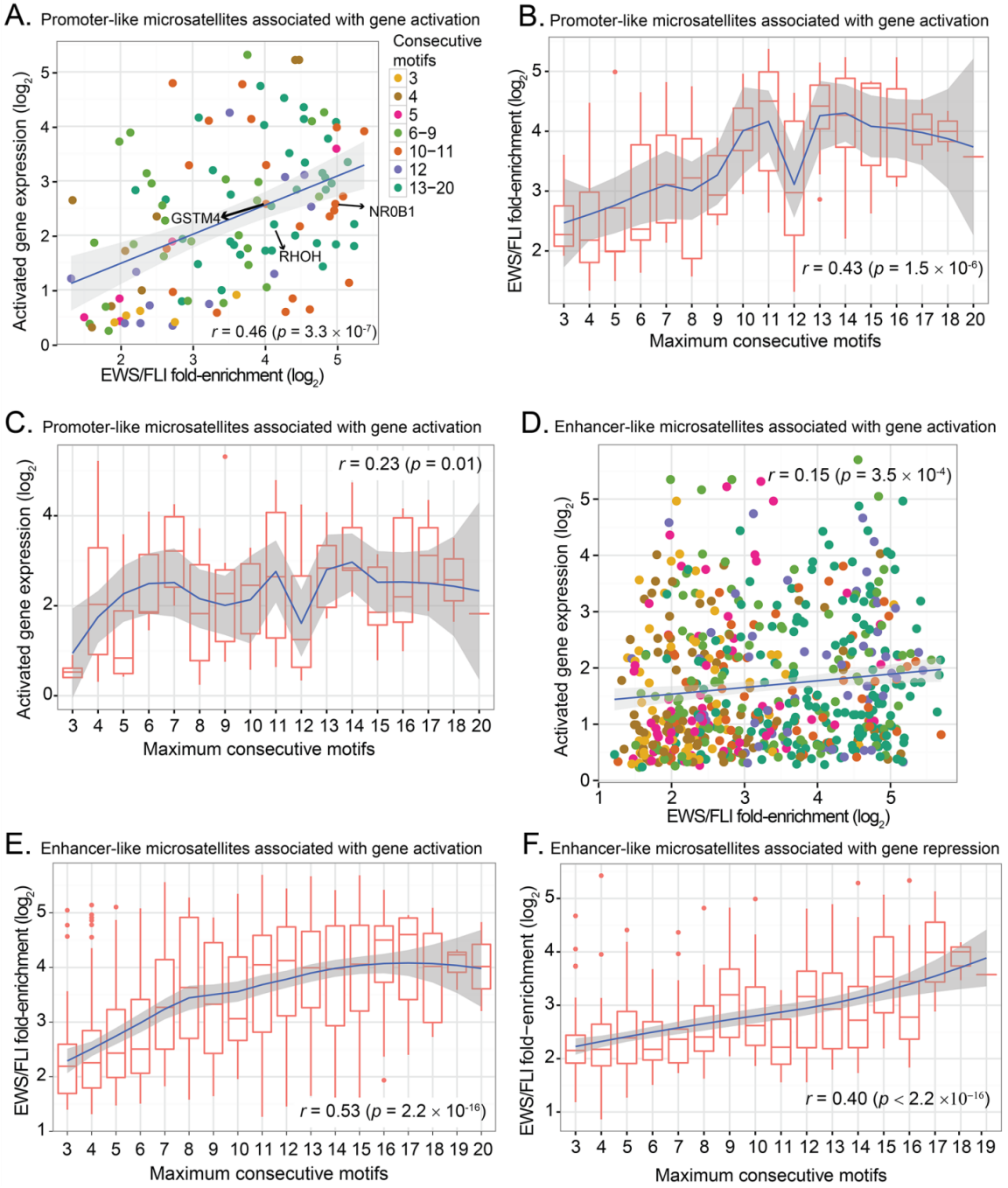


Figure 5.4 Correlation between EWS/FLI-bound microsatellites, GGAA-motif and gene expression.

(A) Scatter plot of expression of activated genes and EWS/FLI fold-enrichment at promoter-like microsatellites showing a positive correlation (r = 0.46, p = 3.35 × 10^{-7}). **(B)** Boxplot of EWS/FLI fold-enrichment and number of consecutive motifs of EWS/FLI-bound at promoter-like microsatellites for activated genes showing a non-

Figure 5.4 continued

linear trend. Blue line is the estimated LOESS regression line of the mean with the estimated 95% confidence interval (shaded region). Overall, there is statistically significant positive correlation (r = 0.43, $p = 1.5 \times 10^{-6}$). **(C)** Boxplot of EWS/FLI-activated gene expression and number of consecutive motifs at promoter-like EWS/FLI-bound microsatellites for gene activation showing a non-linear trend as seen in EWS/FLI binding intensities and a statistically significant positive correlation (r = 0.23, p = 0.01). The blue line is the estimated LOESS regression line of the mean with the estimated 95% confidence bands (shaded region). **(D)** Scatter plot of expression of activated genes and EWS/FLI fold-enrichment at enhancer-like microsatellites showing a positive correlation (r = 0.15, $p = 3.5 \times 10^{-4}$). **(E)** Boxplot of EWS/FLI fold-enrichment and number of consecutive motifs at EWS/FLI-bound enhancer-like microsatellites showing a positive correlation (r = 0.53, $p = 2.2 \times 10^{-16}$). Blue line is the estimated LOESS regression line of the mean and the standard error of the prediction shown as shaded region. **(F)** Boxplot of EWS/FLI fold-enrichment and number of consecutive motifs at EWS/FLI-bound enhancer-like microsatellites associated with gene repression showing positive correlation (r = 0.40, $p < 2.2 \times 10^{-16}$). The blue line is the estimated LOESS regression line of the mean with the estimated 95% confidence bands (shaded region).

We also observe, however, a non-linear pattern, with an increasing trend of EWS/FLI binding as consecutive motifs increase from 3 to 11, followed by a sharp decrease in binding at 12 consecutive motifs (Figure 5.4B). Binding then increases again at 13-14 consecutive motifs before a final overall decreasing trend (LOESS regression). Interestingly, we observe a similar non-linear pattern with the expression of genes activated by EWS/FLI (i.e. an increase in the activated gene expressions as the consecutive motifs increases from 3 to 11 and a decrease in gene expression from 11-12) (Figure 5.4C, LOESS regression).

We next sought to validate these findings using publically-available data from a different Ewing sarcoma cell line, SK-N-MC. The publically-available SK-N-MC data contained a

single ChIP-seq replicate and had significantly fewer EWS/FLI-bound peaks than we found in the A673 cell line (~3,900 versus ~22,000)[69]. Nevertheless, we found a low, but statistically significant correlation between consecutive GGAA-microsatellite length and EWS/FLI binding (Supplementary Figure 5.12D). Interestingly, we saw the same dip in EWS/FLI binding at 12-repeats as we observed in the A673 data, perhaps indicating an underlying biological mechanism worthy of future study. We also sought to correlate gene expression with microsatellite length and EWS/FLI occupancy; however, an insufficient number of genes passed the significance threshold used for our A673 data and so these correlations could not be performed. Overall, there is a positive correlation between EWS/FLI binding, activated gene expression, and microsatellite characteristics (i.e. consecutive motifs, total motifs and densities), though the correlation of microsatellite characteristics with EWS/FLI binding enrichment is consistently stronger than with gene expression (Supplementary Table 5.3). These observations demonstrate that as promoter-like microsatellite length (number of GGAA-motifs) increases, the EWS/FLI binding enrichment and expression of genes activated by EWS/FLI also increases. This finding also supports the "sweet-spot" model, suggesting there may be an optimal length of promoter-like microsatellites mediating EWS/FLI regulation of transcriptional gene activation.

At enhancer-like microsatellites, increased numbers of GGAA-motifs positively correlate with EWS/FLI binding but only minimally with gene activation

To determine whether longer-range, enhancer-like microsatellites also confer EWS/FLI activation associated with motif length, we next looked at the 580 microsatellites that are more than 5kb away from EWS/FLI activated genes. We observe a minimal (significant) positive linear correlation between EWS/FLI binding enrichment and gene expression for these long-range potential response elements (r = 0.15, p = 3.5×10^{-4}) (Figure 5.4D). Evaluating total number of motifs in these microsatellites, we observe a significant positive correlation with EWS/FLI binding enrichment (r = 0.25, p = 1.28×10^{-9}), and a minimal positive correlation with EWS/FLI activated gene expression (r = 0.10, p= 0.02) (See Supplementary Figure 5.13A-B and Supplementary Table 5.3). We also observe a significant positive linear correlation between EWS/FLI enrichment and the number of consecutive motifs of these microsatellites (r = 0.53, $p < 2.2\times10^{-16}$) (Figure 5.4E), and a minimal, non-significant positive trend with EWS/FLI activated gene expression (r = 0.07, p = 0.08) (Supplmentary Figure 5.13C). These observations suggest that although longer GGAA-motifs enhance EWS/FLI binding, it only minimally translates to an increase in the expression of activated gene targets. This is likely due to the complexity of long-range regulatory mechanisms.

At promoter-like microsatellites, number of GGAA-motifs demonstrates no length-dependency with EWS/FLI responsiveness for gene repression

To test whether GGAA-microsatellite characteristics affect EWS/FLI-mediated repression, we first looked at the 52 promoter-like microsatellites that are within 5kb of EWS/FLI-repressed genes. In contrast to EWS/FLI-activated genes, there is no

significant correlation between microsatellite characteristics (i.e. number of consecutive motifs and total number of motifs) and EWS/FLI binding enrichment or EWS/FLI-mediated gene repression. The correlation between EWS/FLI binding enrichment and EWS/FLI-regulated genes is 0.12 (p = 0.41) (Supplementary Figure 5.14A). Correlation of EWS/FLI binding enrichment with number of consecutive motifs is 0.18 (p = 0.19) and with total number of motifs is 0.06 (p = 0.66) (Supplementary Figures 5.14B-C). We also observed no correlation between EWS/FLI-repressed genes with the number of consecutive motifs (r = -0.22, p = 0.12) and total number of motifs (r = -0.04, p = 0.79) (See Supplementary Table 5.4 and Supplementary Figures 5.14D-E). Overall, promoter-like GGAA-microsatellites don't enhance either EWS/FLI binding or expression of repressed genes, supporting the model that EWS/FLI represses gene targets through an alternate regulatory mechanism.

At enhancer-like microsatellites, number of GGAA-motifs positively correlates with EWS/FLI binding but not gene repression

To test whether enhancer-like microsatellites confer EWS/FLI-mediated repression, we investigated EWS/FLI responsiveness at the 425 microsatellites that are more than 5kb away from EWS/FLI-repressed genes. Our data demonstrates increasing number of consecutive motifs positively correlates with EWS/FLI binding enrichment (r = 0.40, $p < 2.2\times10^{-16}$) (Figure 5.4F). Increased EWS/FLI binding enrichment is also shown to be positively correlated with total number of motifs (r = 0.22, $p = 3.0\times10^{-6}$) at these microsatellites (Supplementary Figure 5.15A). We found, however, there is no significant

correlation between EWS/FLI binding and expression of repressed genes more than 5kb from their associated microsatellite (r = -0.05, p = 0.33) (Supplementary Figure 5.15B). Accordingly, there is also no correlation between gene expression and number of consecutive motifs or total number of motifs (p = 0.43 and p = 0.68) for these EWS/FLI-repressed gene associated microsatellites (Supplementary Table 5.4 and Supplementary Figure 5.15C). In summary, EWS/FLI binding increases with increasing GGAA-motif length at long-range, enhancer-like microsatellites, however, there is no effect on concomitant gene repression of these EWS/FLI targets.

Discussion

Gene-associated GGAA-microsatellites serve as DNA response elements for EWS/FLI to bind and mediate transcriptional activation of its up-regulated targets[16,46,51,170]. In this study we describe microsatellites on a global genomic scale, and use ChIP-seq and RNA-seq analysis to computationally investigate EWS/FLI responsiveness at these repetitive elements. Overall, our genome-wide characterization of GGAA-microsatellites identifies two distinct classes of EWS/FLI-bound GGAA-microsatellites, demonstrating the integral relationship of microsatellite length and gene proximity to facilitate EWS/FLI binding and transcriptional activity in Ewing sarcoma (Figure 5.5).

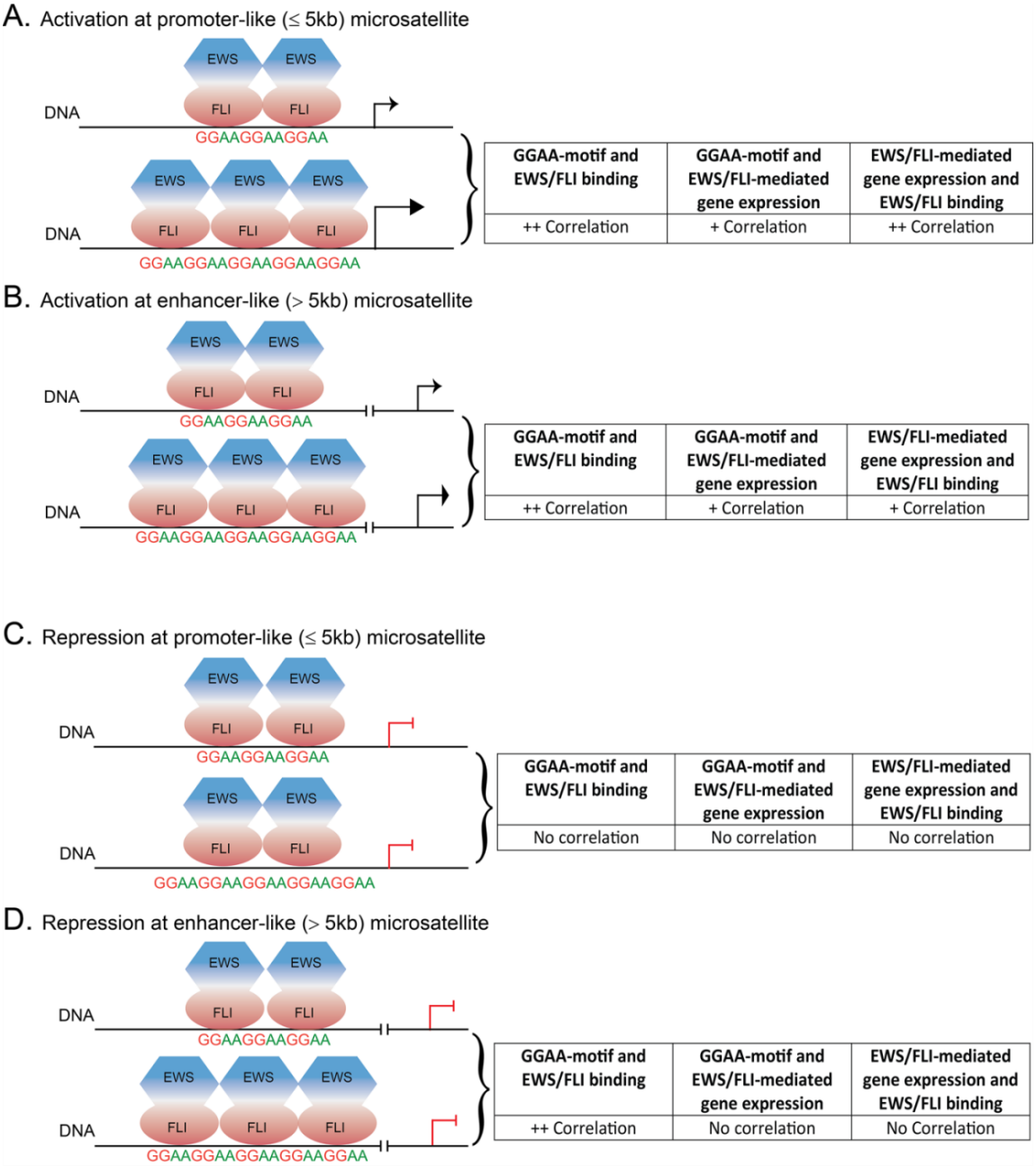


Figure 5.5 Schema of correlative associations between GGAA motifs in EWS/FLI-bound microsatellites for gene activation and repression. Schematic illustrating EWS/FLI responsiveness at given loci across the genome

(A) Promoter-like (close-range) GGAA-microsatellites positively correlate with EWS/FLI binding and activation of genes in a length dependent manner. **(B)** Enhancer-like (long-range) GGAA-microsatellites positively correlate with EWS/FLI binding but correlation with transcriptional regulation is only minimal for activated genes. **(C)**

Promoter-like GGAA-microsatellites display no correlation with EWS/FLI binding and transcriptional repression. **(D)** Enhancer-like GGAA-microsatellites positively correlate with EWS/FLI binding; however, they do not confer gene expression.

While we and others have previously described these GGAA-microsatellites, we recognized a paucity of definitive parameterization required for mechanistic understanding of EWS/FLI transcriptional modulation at these response elements. Pursuing an unbiased genome-wide approach, we found microsatellites of fewer than three consecutive GGAA-motifs do not significantly overlap with EWS/FLI binding sites, suggesting a minimum length of consecutive motifs is required for binding. This was in line with our previous experimental finding of multimeric EWS/FLI binding at a minimum of three consecutive GGAA-repeats[170]. Thus, our genome-wide description of GGAA-microsatellite regions in this study lays an unprejudiced groundwork upon which actual ChIP-seq and RNA-seq data can be overlain. For example, in describing genome-wide GGAA-microsatellite regions, we found an enrichment of longer consecutive GGAA-repeats on chromosome 19. When FLI-ChIP-seq data was applied to the analysis, we found a corresponding enrichment of EWS/FLI binding sites on the same chromosome. In this work we present, to our knowledge, the first attempt to determinately define GGAA-microsatellites across the genome in a Ewing sarcoma-relevant context.

We and others previously showed that EWS/FLI regulates its activated, but not down-regulated targets using GGAA-microsatellites as response elements [170]. The present

genome-wide analysis provides further support for length-dependency of EWS/FLI responsiveness near activated, promoter-like and enhancer-like microsatellites. Interestingly, we observe a significant correlation of GGAA-repeats associated with repressed targets and binding enrichment of EWS/FLI at enhancer-like microsatellites, illuminating a novel class of microsatellites with a potentially distinct function.

Transcriptional activation and repression are both critical for EWS/FLI-mediated oncogenic function, yet, the mechanism by which EWS/FLI differentiates these functions remains unknown. The association of EWS/FLI-bound microsatellites with only activated genes supports a likely molecular mechanistic difference in transcriptional modulation of EWS/FLI up vs. down-regulated targets. This model is further supported by our recent data that members of the chromatin remodeling NuRD complex interact with EWS/FLI near its repressed, but not activated targets[34].

Our findings in this study suggest the additional possibility that distance and overall chromatin landscape may be contributing factors in transcription factor activating vs. repressive functions. For example, recent studies have demonstrated evidence for super-enhancers, which function in long-range regulation and are associated with an enrichment of activating histone marks[219,220]. EWS/FLI binding in Ewing sarcoma cells has been shown to be bound in these super-enhancer regions[69,70,72]. To our knowledge, it has not yet been evaluated whether repressive regulatory domains exist on a similar genomic scale. The data from the present study suggests GGAA-microsatellites found in promoter-like regions convey EWS/FLI-mediated gene activation, while those found in enhancer-

like regions likely require more complex regulatory factors such as chromatin remodeling complexes to establish long-range interactions. Specifically, in association with gene repression, EWS/FLI may displace endogenous transcription factors disrupting enhancer activity, a mechanism proposed by Riggi et al.,[69] or these regions may naturally be more nucleosome-depleted to allow EWS/FLI binding.

An additional explanation for the mechanism by which EWS/FLI modulates activation or repression of its targets could be sequence specificity upon binding to length-dependent microsatellites[47]. We recently conducted a biochemical study to investigate the molecular reasoning behind EWS/FLI binding at "sweet-spot" microsatellites. We found that EWS/FLI binding affinity improves at "sweet-spot" microsatellites, and unexpectedly requires the EWS portion of the fusion to bind these optimal numbers of GGAA-motifs[217]. Our stoichiometric data further supports a model in which multiple EWS/FLI molecules bind across these GGAA-microsatellites. The "sweet-spot" finding evidenced in both our clinical and biochemical data implicate 18-26 GGAA repeats ("sweet-spot") as the length of GGAA-microsatellites that allow an optimal configuration of EWS/FLI binding at these sites. Our current study suggests EWS/FLI-responsive GGAA-microsatellites are enriched near activated, but not repressed EWS/FLI targets. Taken together, it is likely that our "sweet-spot" observation is due to the aforementioned biochemical mechanism, and that this multimeric EWS/FLI binding at repeat regions may facilitate EWS/FLI differentiation between activation and repression of its targets.

FLI, which contains the DNA-binding domain through which EWS/FLI directly associates with the DNA, is an ETS family member. ETS factor binding studies have demonstrated that small differences in transcription factor binding specificity contribute significantly to site selectivity[52]. While our "sweet-spot" finding supports this model of transcription factor binding site selectivity, it is not known whether total microsatellite number (microsatellite "length"), or number of *consecutive* GGAA-motifs confers this specificity. *Guillon et al.* determined that EWS/FLI shows a binding preference for 9 or more contiguous GGAA repeats, and postulated that binding at greater than 9 repeats is required for EWS/FLI-mediated activation of its up-regulated targets[46]. This is interesting in light of our present data demonstrating a peak in EWS/FLI DNA-binding and gene activation at 10-11 and 13-14 consecutive GGAA-motifs. Although minimal, due to few microsatellites longer than 20 consecutive repeats across the genome, our overall data nevertheless suggests that microsatellites with numbers of GGAA-motifs greater than the "sweet-spot" are not associated with EWS/FLI-mediated differential gene expression.

Further investigation will be required to also determine the role of consecutive motif number in relation to our "sweet-spot" finding. For example, EWS/FLI regulates *NR0B1* through a "sweet-spot" microsatellite of 24 *total* motifs in the A673 cell line, but this microsatellite region contains 11 *consecutive* motifs as its longest contiguous segment. As cited in the above results, we found this same repeat length enriched near genes compared to other consecutive motifs lengths in our genome-wide microsatellite characterization.

Our study should be considered in light of some limitations that may potentially mask the magnitude of EWS/FLI association with particular microsatellite characteristics. The first relates to the use of the human reference (hg19) genome instead of the A673 Ewing sarcoma genome as a reference. To evaluate the appropriateness of using the human reference genome, we selected a number of our favorite EWS/FLI activated genes, amplified the associated GGAA-microsatellites, and sequenced these regions. We found that some are very similar to the human reference genome (i.e. *NR0B1 and FIBCD1*), while others demonstrate significant alterations in GGAA-motif number (i.e. *PINK1*) (Johnson and Taslim, unpublished observation). Interestingly, *FCGRT* is a highly up-regulated EWS/FLI target, yet contains 12 consecutive motifs according to the human reference genome. This motif length was observed as the unexpected dip in our microsatellite-defining analysis for both EWS/FLI binding and gene-expression.

Sequencing the *FCGRT* microsatellite from A673 genomic DNA, however, revealed an *FCGRT*-associated microsatellite that is actually 9-consecutive GGAA-motifs in length. Together, these findings give us confidence that our data reflects the appropriate general trends in EWS/FLI responsiveness at microsatellites, but suggests a more accurate correlation will require A673 whole genome sequencing for reference. This concept may also be applied for consideration more broadly in other fields where the human reference genome has been used instead of relevant disease genomes.

Overall, our results reveal and characterize two classes of GGAA-microsatellites, suggesting EWS/FLI interacts with these unique binding sites via distinct regulatory mechanisms for distance-dependent activation and repression of its gene targets. Defining and characterizing GGAA-microsatellites is critical for understanding and prediction of EWS/FLI responsiveness across the genome. We also demonstrate the value of synergizing experimental and computational evaluation to better delineate the underlying molecular mechanisms of EWS/FLI transcription factor function and oncogenic re-programming.

Supporting Information

Supplementary Figures

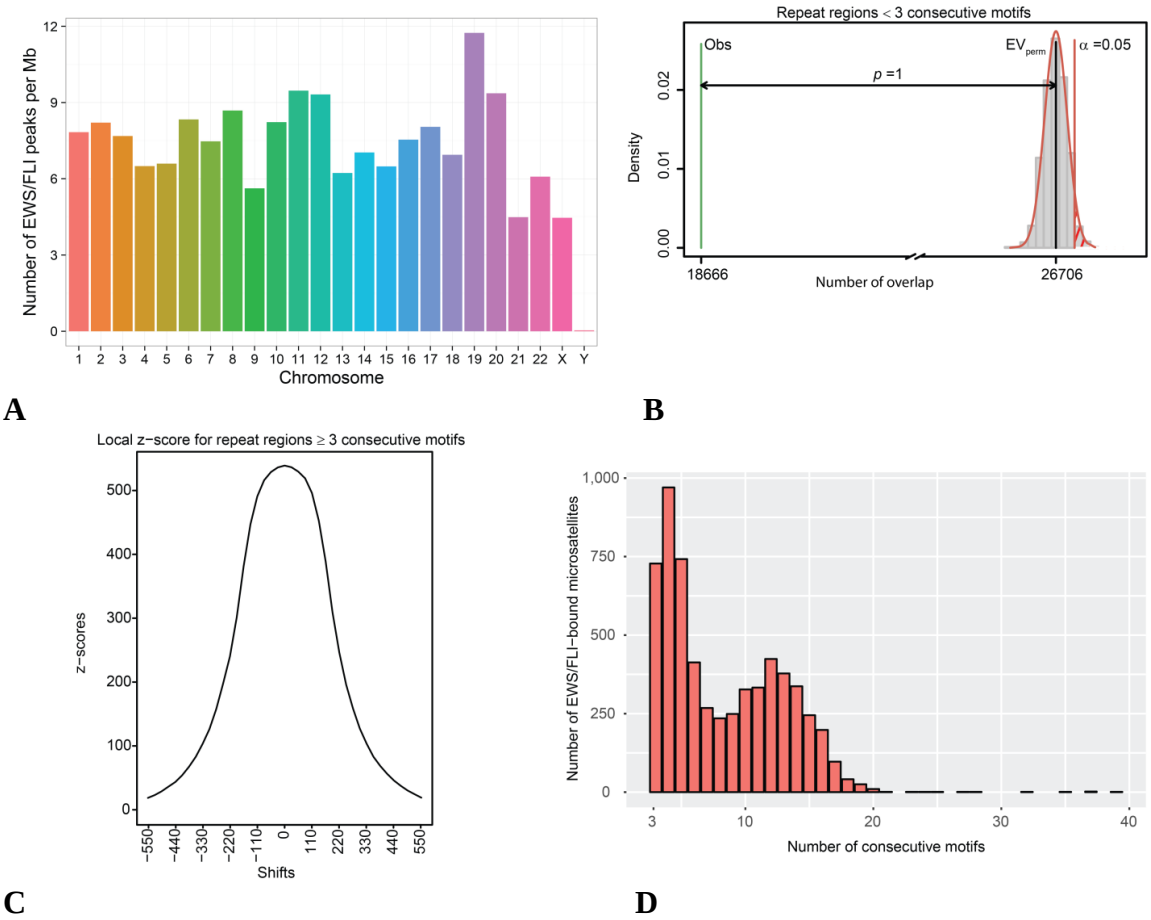


Figure 5.6 Characterization of GGAA-repeat regions across the genome

(A) Histogram of number of EWS/FLI peaks per Mb (normalized by chromosome length) in each chromosome. **(B)** Permutation test shows that the number of EWS/FLI binding sites that overlap with repeat regions with 2 or less consecutive motifs is not significantly higher than random chance ($p = 1$). The red line denotes the significance limit ($\alpha = 0.05$). Gray bars represent the number of overlaps of the random regions with EWS/FLI binding sites. The black line represents the mean and in green the number of overlaps of repeat regions with 2 or less consecutive motifs. EVperm is the expected value of the permutation (number of overlaps in random samples). Obs is the observed number of overlap. **(C)** Plot of shifted z-score for the association between EWS/FLI repeat regions ≥ 3 consecutive motifs and EWS/FLI binding sites showing that this association is highly dependent on the location of the regions. **(D)** Number of EWS/FLI-bound microsatellites and the number of consecutive motifs in these microsatellites.

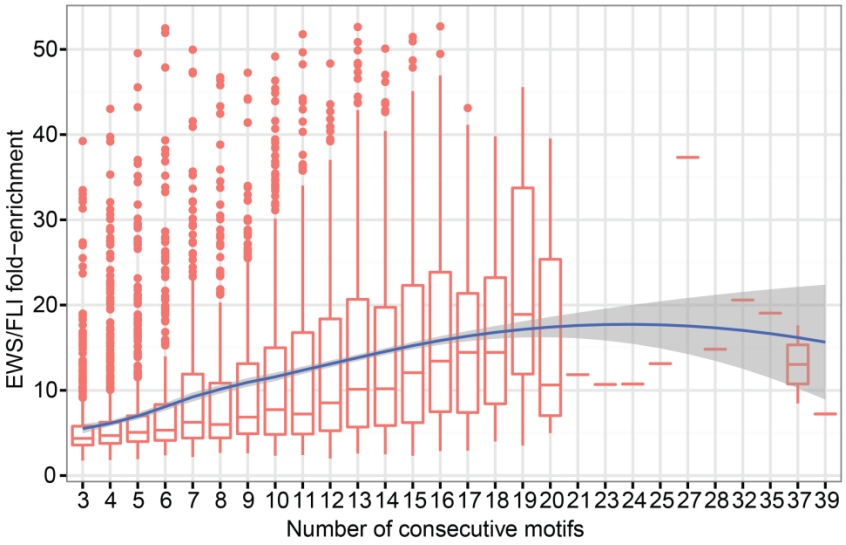


Figure 5.7 Boxplot showing the number of consecutive motifs of all EWS/FLI-bound microsatellites with the EWS/FLI fold-enrichment. The blue line is the estimated LOESS regression line of the mean with the estimated 95% confidence bands (shaded region).

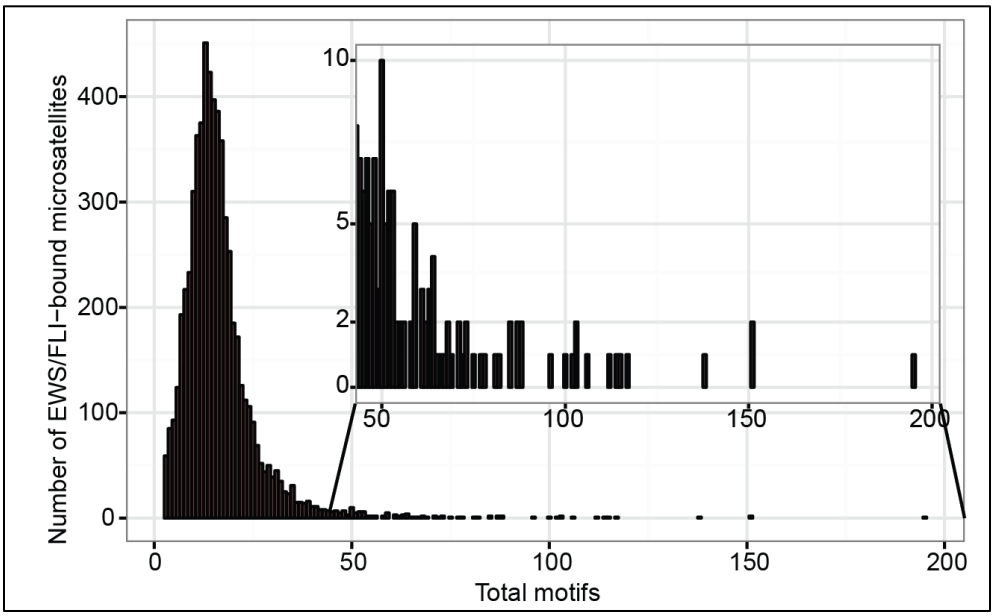


Figure 5.8 Histogram showing the number of EWS/FLI-bound microsatellites grouped by the total number of motifs

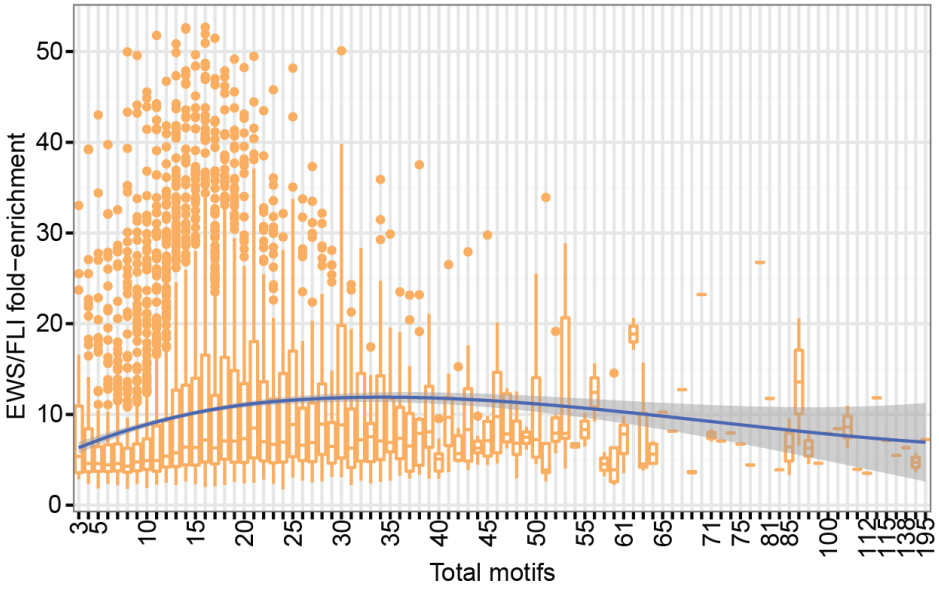


Figure 5.9 Boxplot showing the total motifs of all EWS/FLI-bound microsatellites with the EWS/FLI fold-enrichment. The blue line is the estimated LOESS regression line of the mean with the estimated 95% confidence bands (shaded region).

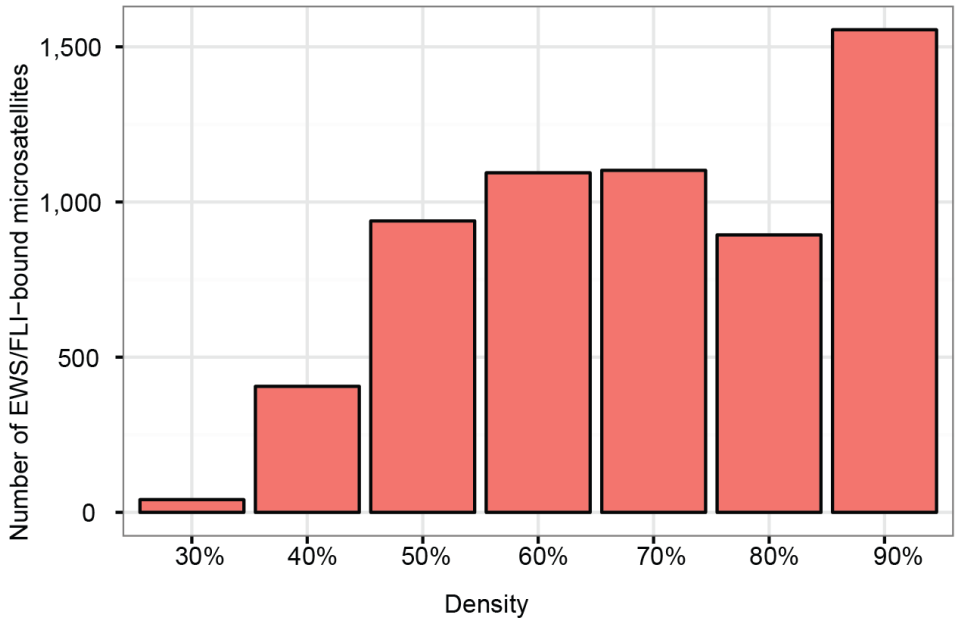


Figure 5.10 Histogram showing number of EWS/FLI-bound microsatellites with their densities

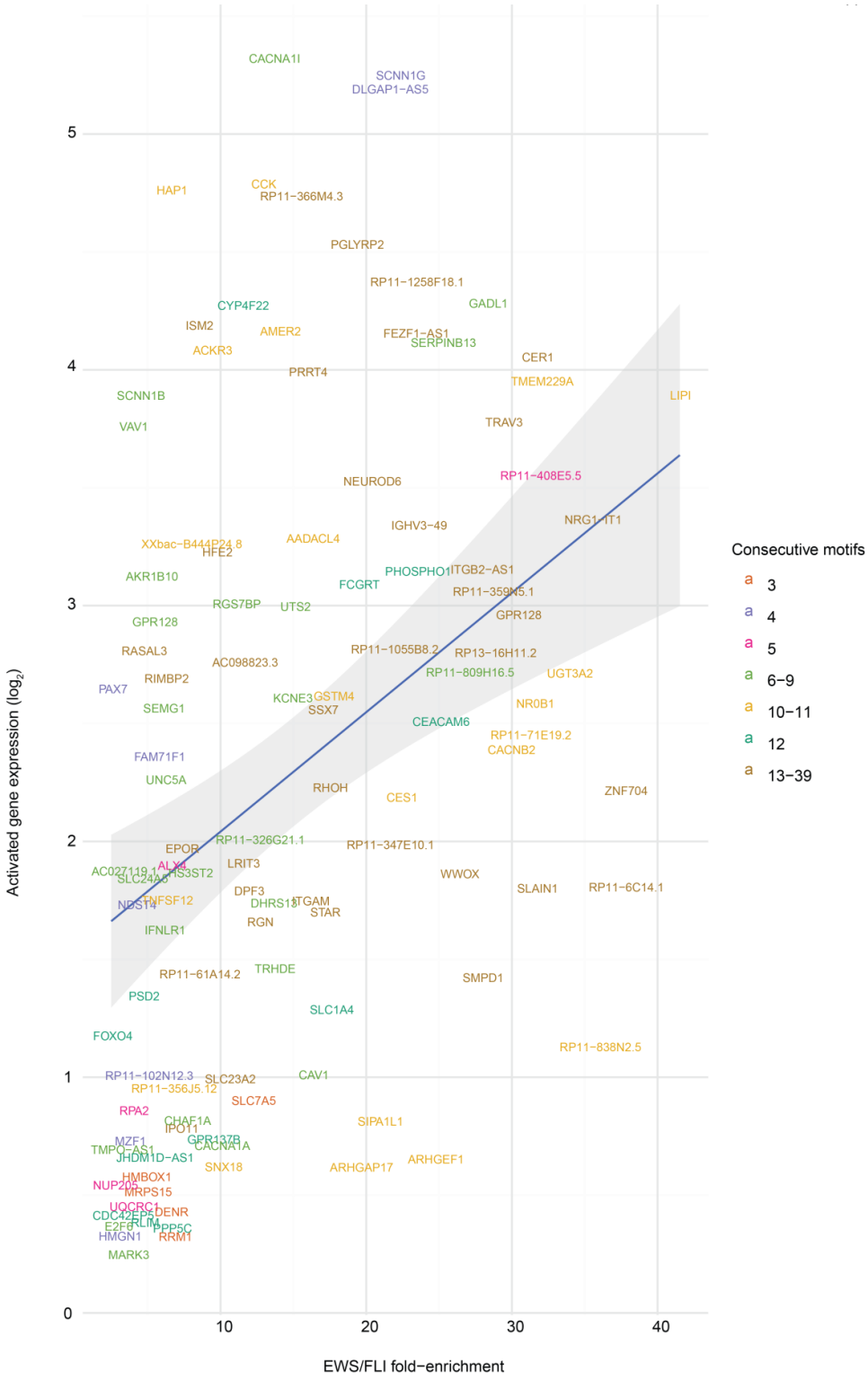
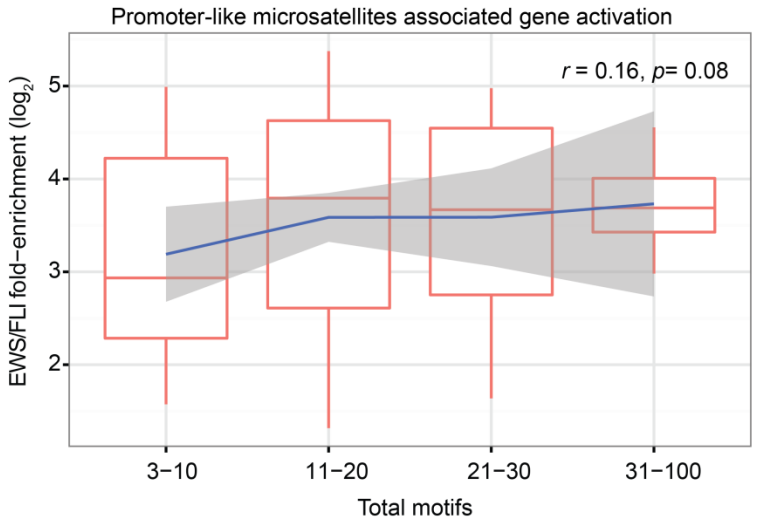


Figure 5.11 EWS/FLI responsiveness at promoter-like microsatellites near activated gene targets.

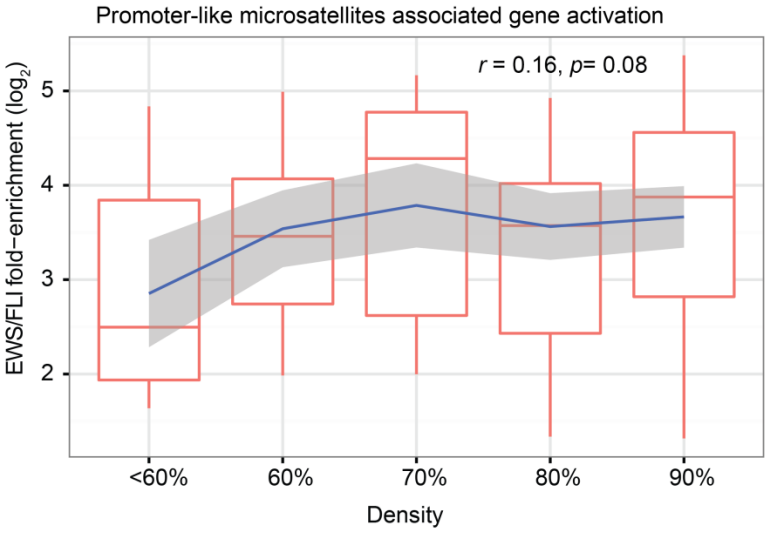
Figure 5.11 continued

Scatter plot showing activated gene names (False Discovery Rate (FDR) ≤ 5%) that are within 5kb of microsatellites with their EWS/FLI fold-enrichment and their corresponding gene expression (log_2). Note: some gene names are adjusted for readability.

Figure 5.12 Promoter-like microsatellites association with gene activation

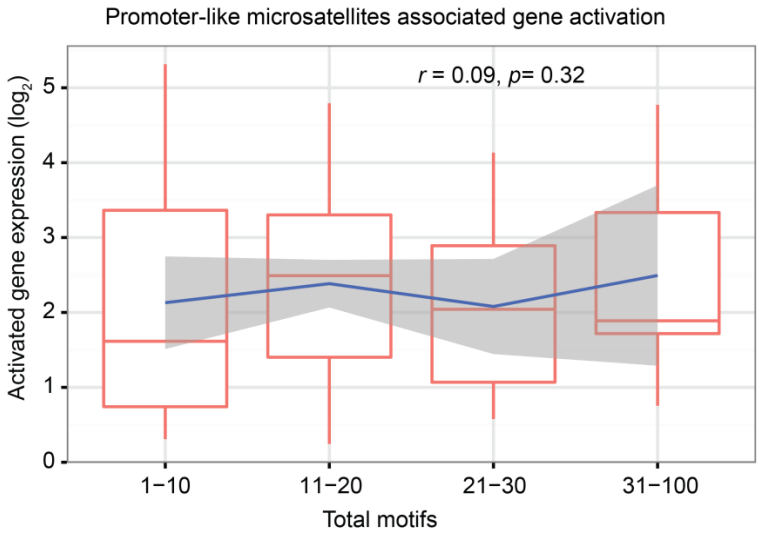


A

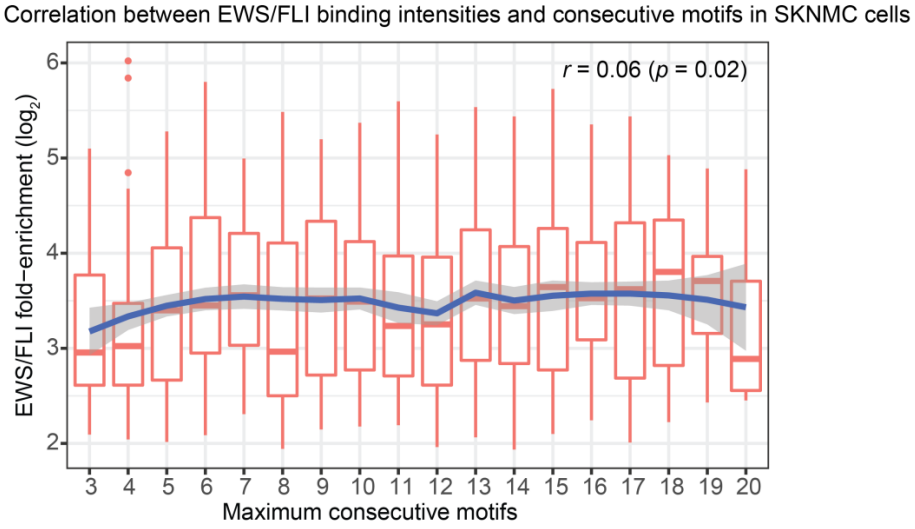


B

Figure 5.12 continued



C



D

(A) Trend toward positive correlation between total motifs of EWS/FLI-bound microsatellites and EWS/FLI fold-enrichment (log_2). **(B)** Trend toward positive correlation between densities of EWS/FLI-bound microsatellites and EWS/FLI fold-enrichment (log_2). **(C)** No significant correlation between total motifs of EWS/FLI bound microsatellites and activated gene expression (log_2). LOESS regression line is shown in blue. Shaded region is the estimated 95% confidence bands.**(D)** Trend toward positive correlation between EWS/FLI fold-enrichment (log_2) and number of consecutive motifs in SK-N-MC cells (r = 0.06, p = 0.02).

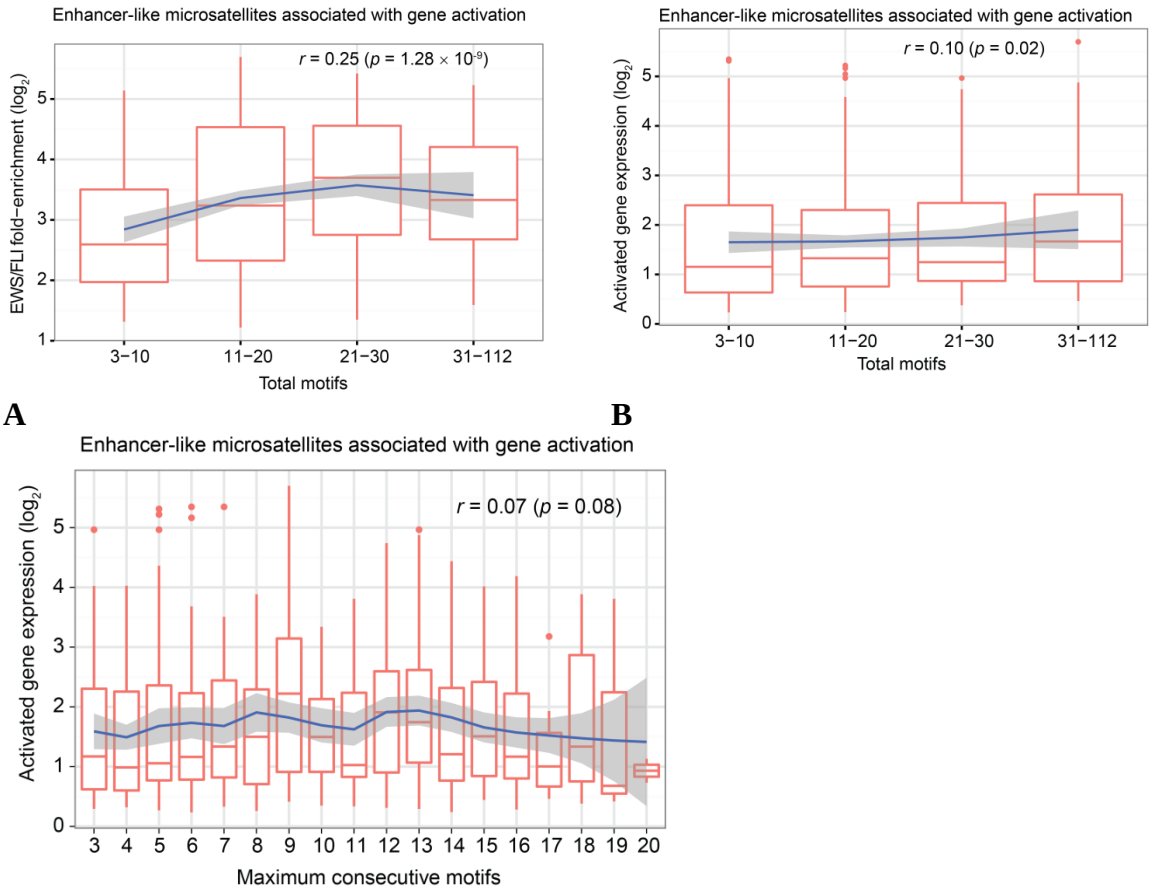


Figure 5.13 Enhancer-like microsatellites association with EWS/FLI activated genes.

(A) EWS/FLI fold-enrichment has a significant positive correlation with total number of motifs (r = 0.25, p = 1.28 × 10^{-9}). **(B)** Gene expression has significant but minimal positive correlation with total number of motifs (r = 0.10, p = 0.02). **(C)** Trend toward minimal positive correlation between activated gene expression and number of consecutive motifs (r = 0.07, p = 0.08).

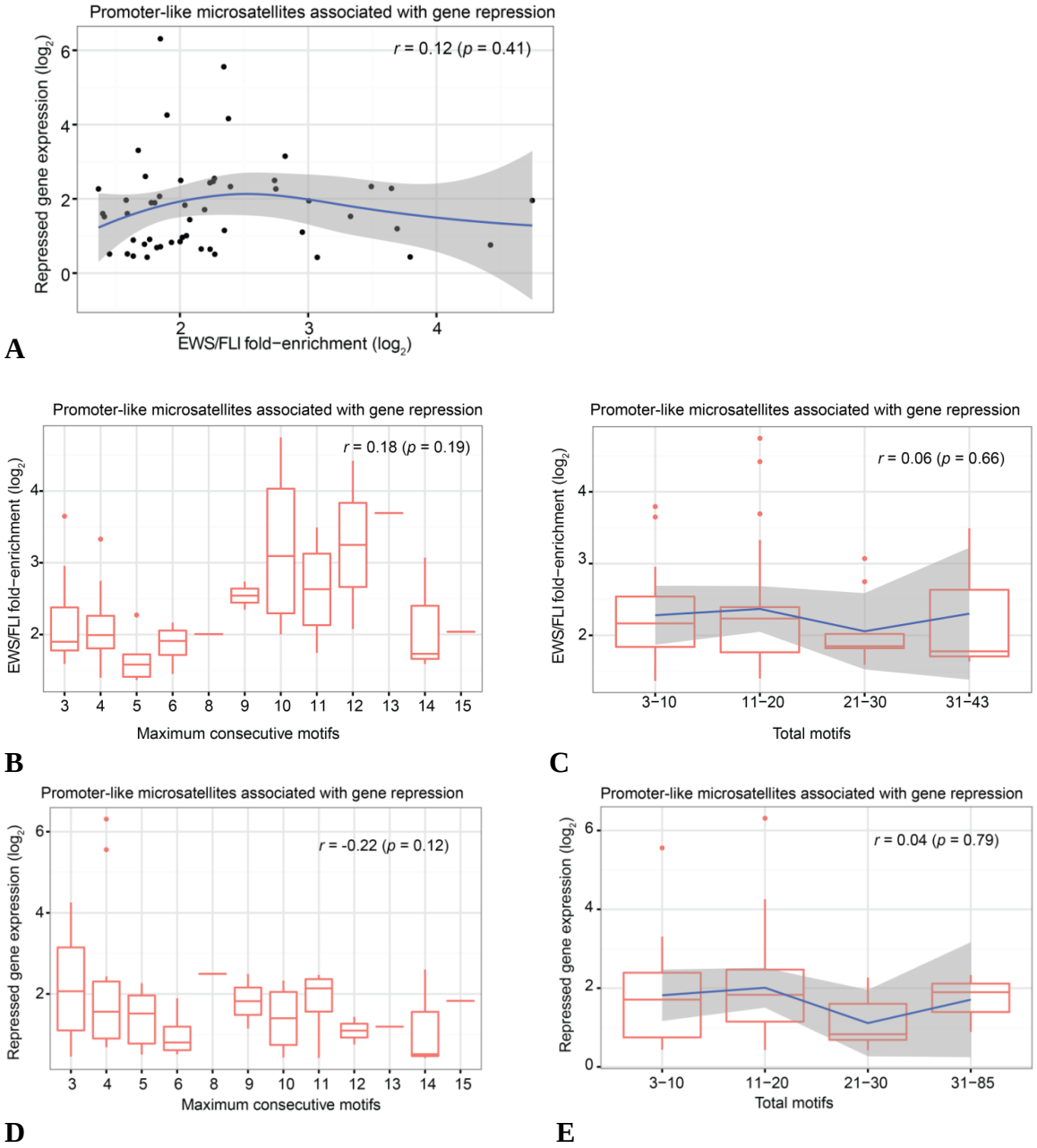
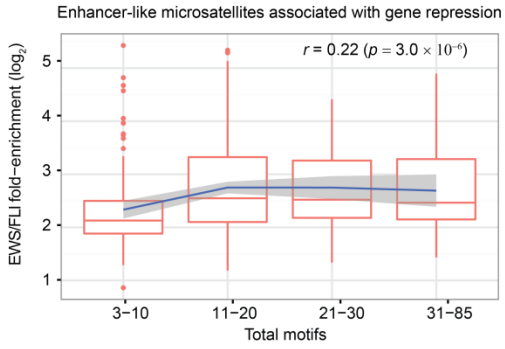
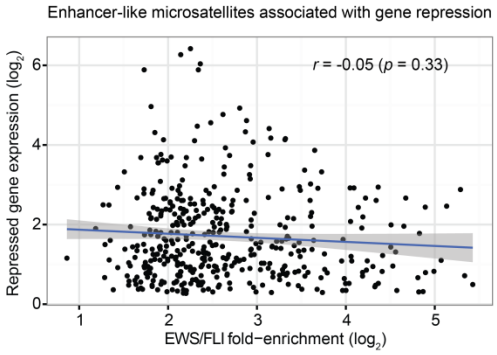


Figure 5.14 Promoter-like microsatellites association with gene repression.

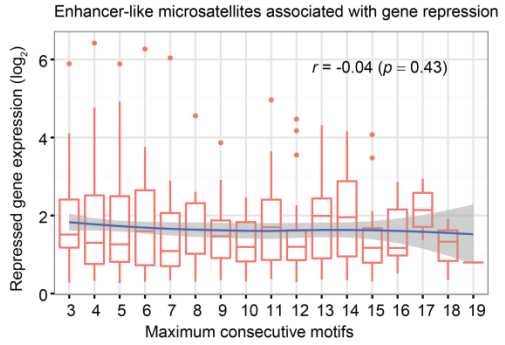
(A) No correlation between EWS/FLI fold-enrichment and gene expression (r = 0.12, p = 0.41). **(B)** No correlation between EWS/FLI fold-enrichment and number of consecutive motifs (r = 0.18, p = 0.19). **(C)** No correlation between EWS/FLI fold-enrichment and total motifs (r = 0.06, p = 0.66). **(D)** No correlation between gene expression and number of consecutive motifs (r = -0.22, p = 0.12). **(E)** No correlation between gene expression and total number of motifs (r = -0.04, p = 0.79). Shaded region is the 95% confidence interval.



A



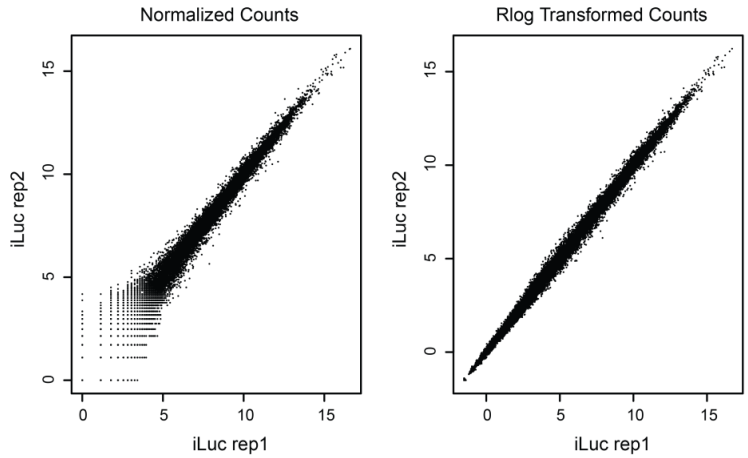
B



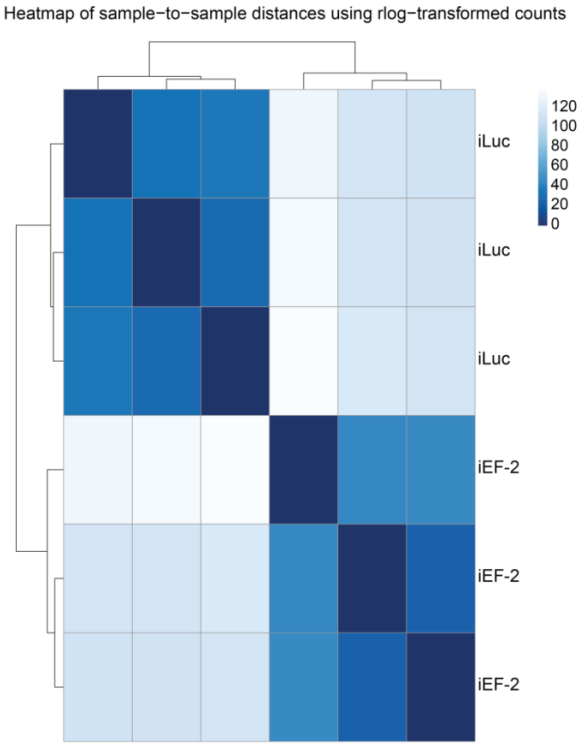
C

Figure 5.15 Enhancer-like microsatellites associated with gene repression.

(A) Significant positive correlation between EWS/FLI fold-enrichment and number of consecutive motifs ($r = 0.40$, $p < 2.2\times10^{-4}$). **(B)** No correlation between EWS/FLI fold-enrichment and gene expression ($r = -0.05$, $p = 0.33$). **(C)** No correlation between repressed genes' expression and number of consecutive motifs ($r = -0.04$, $p = 0.43$). LOESS regression line is shown in blue. Shaded region is the estimated 95% confidence bands.



A



B

Figure 5.16 RNA-seq normalization and samples similarities.

(A) Comparison of two different normalization methods. Left panel, counts normalized by sequencing depth. Right panel, counts normalized by rlog transformation. Sequencing depth normalization still showing bias toward highly expressed genes (i.e. high variance for low expressed genes), while rlog transformation no longer shows such bias (i.e. variances are stabilized across genes). **(B)** Heat map of sample-to-sample similarities using rlog transformed counts. Color represents distance between samples with dark blue indicating samples with high similarities.

Supplementary Tables

	Genomic location (hg19)	Sequence
1.	chr1:199,130,087-199,130,206	TTCCTTCCAAGGAAGGAAGGGAGGGAGGGAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAA
2.	chr4: 34,658,583-34,658,655	TTCCTTCCTCAGGAAAGGACCTTTGGGCAGCAAGGAAGGAAGGAAGGAAGGAAGGAAAGAAGGAAGAAAGGAA
3.	chr22: 45,138,923 - 45,138,773	GGAAGGGAGGGAGGGAGGAAGGAAGGAAAGGAAGGAAGGGAGGGACGAAGGGAGGAAGGAGAAAGAAAGGAAAGAAAGGAAAGATAGAGAAGGAAGGAAAGGAGAGAGGAAAGGACTCCTTCCTTCCTTGCAGTCCCTGTAGCTGCTTTCC
4.	chrY: 7,311,313- 7,311,439	GGAAATATGATTCTGGGAAGGAATTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTCCTTATTTCCTTCCTTCCTTCCTTCCTTCCTTCCTCCCACCTTTCC
5.	chr18: 30,686,220-30,686,285	TTCCTTCCTTCCTGCAAGTTTCCTCCTTTCAATTTTCCCAAGAAAATAAAGAAAGGAAGGAAGGAA

Table 5.2 Examples of mixed repeat regions (repeat regions that contain both GGAA and TTCC motifs).

Number of microsatellites	**6,031**
Corr. of peak fold change and number of consecutive motifs, r (p-value)	0.46 (2.2e-16)
Corr. of peak fold change and total number of motifs, r (p-value)	0.23 (2.2e-16)
Activated genes	
Number of genes (n)	533
Number of microsatellites (n)	694
Corr. of gene expression and peak fold change, r (p-value)	0.21 (1.6e-08)
Corr. of peak fold change and number of consecutive motif, r (p-value)	0.52 (2.2e-16)
Corr. of peak fold change and total number of motif, r (p-value)	0.24 (1.9e-10)
Corr. of gene expression and number of consecutive motif, r (p-value)	0.12 (1.5e-03)
Corr. of gene expression and total number of motif, r (p-value)	0.11 (5.0e-03)
Promoter-like microsatellites	
Number of genes (n)	113
Number of microsatellites (n)	114
Corr. of gene expression and peak fold change, r (p-value)	0.46 (3.3e-07)
Corr. of peak fold change and number of consecutive motif, r (p-value)	0.43 (1.5e-06)
Corr. of peak fold change and total number of motif, r (p-value)	0.16 (0.08)
Corr. of gene expression and number of consecutive motif, r (p-value)	0.23 (0.01)
Corr. of gene expression and total number of motif, r (p-value)	0.09 (0.32)
Enhancer-like microsatellites	
Number of genes (n)	440
Number of microsatellites (n)	580
Corr. of gene expression and peak fold change, r (p-value)	0.15 (3.5e-04)
Corr. of peak fold change and number of consecutive motif, r (p-value)	0.53 (2.2e-16)
Corr. of peak fold change and total number of motif, r (p-value)	0.25 (1.28e-09)
Corr. of gene expression and number of consecutive motif, r (p-value)	0.07 (0.08)
Corr. of gene expression and total number of motif, r (p-value)	0.10 (0.02)

Table 5.3 Correlation between microsatellites and EWS/FLI binding enrichment and EWS/FLI-activated genes.

Number of microsatellites	**6031**
Corr. of peak fold change and number of consecutive motifs, r (p-value)	0.46 (2.2e-16)
Corr. of peak fold change and total number of motifs, r (p-value)	0.23 (2.2e-16)
Repressed genes	
Number of genes	400
Number of microsatellites	477
Corr. of gene expression and peak fold change r (p-value)	-0.03 (0.51)
Corr. of peak fold change and number of consecutive motif, r (p-value)	0.39 (2.2e-16)
Corr. of peak fold change and total number of motif, r (p-value)	0.21 (3.1e-06)
Corr. of gene expression and number of consecutive motif r (p-value)	-0.06 (0.21)
Corr. of gene expression and number of total motif r (p-value)	-0.02 (0.62)
Promoter-like microsatellites	
Number of genes	49
Number of microsatellites	52
Corr. of gene expression and peak fold change, r (p-value)	0.12 (0.41)
Corr. of peak fold change and number of consecutive motif, r (p-value)	0.18 (0.19)
Corr. of peak fold change and total number of motif, r (p-value)	0.06 (0.66)
Corr. of gene expression and number of consecutive motif, r (p-value)	-0.22 (0.12)
Corr. of gene expression and total number of motif, r (p-value)	-0.04 (0.79)
Enhancer-like microsatellites	
Number of genes	357
Number of microsatellites	425
Corr. of gene expression and peak fold change, r (p-value)	-0.05 (0.33)
Corr. of peak fold change and number of consecutive motif, r (p-value)	0.40 (2.2e-16)
Corr. of peak fold change and total number of motif, r (p-value)	0.22 (3.0e-6)
Corr. of gene expression and number of consecutive motif, r (p-value)	-0.04 (0.43)
Corr. of gene expression and total number of motif, r (p-value)	-0.02 (0.68)

Table 5.4 Correlation between microsatellites, EWS/FLI binding enrichment and EWS/FLI-repressed genes.

Chapter 6: Allelic specificity in EWS/FLI-microsatellite binding

Introduction:

Allele-specific expression is a recently described phenomenon in which one of two heterozygous alleles is preferentially expressed compared to the other[221]. In gene regulatory regions, the presence of variants in the sequence (i.e. polymorphisms among individuals) may affect both protein binding and epigenetic regulation at that site[222]. Some classically noted examples are genomic imprinting and X-chromosome inactivation, both epigenetic processes. In each of these examples, epigenetic heterogeneity arises as a result of transcriptional silencing at affected genomic loci[221]. It is reasonable to surmise that transcription factors may play an important role in the mechanism underlying such intrinsic expression bias.

The incidence of Ewing sarcoma in Europeans is 10 fold greater than in Africans, independent of geographical location[11]. Reminiscent of allele-specific expression, critical EWS/FLI target *NR0B1* contains a GGAA-microsatellite upstream of its promoter whose length is genetically polymorphic[43]. These variable-lengths of non-coding DNA sequence are found upstream from the promoter region of *NR0B1* and other transcriptionally active targets where EWS/FLI preferentially binds. Interestingly, *NR0B1* GGAA-microsatellites in the African population are dispersed and often composed of larger repeats (16-72

repeats), while Europeans tend to have shorter repeats, typically 16-25[14]. Conversely, our clinical data demonstrates Ewing sarcoma patients have *NR0B1* microsatellites containing 20-25 GGAA-repeats in their germline *and* tumor cells (Chapter 3)[16]. This clinical "sweet-spot" observation of 20-25 GGAA-repeats has proven an optimal length for EWS/FLI mediated transcriptional activity in subsequent biochemical evaluation[16,217]. Thus, patients with this inherited length of GGAA-repeats appear to have a heightened preponderance for developing Ewing sarcoma.

This inherent transcriptional advantage of binding at particular GGAA-microsatellite lengths, coupled with the polymorphic nature of these response element regions within the population, suggests a potential for EWS/FLI allele specific binding. Characterizing EWS/FLI preferential binding at particular lengths of GGAA-microsatellites is important to achieving a better understanding of the specific mechanism by which EWS/FLI modulates transcriptional activity. However, experimentally assessing binding preferences of transcription factors, like EWS/FLI, remains a challenge.

Our recent bioinformatics study supplied a critical step forward in our characterization of GGAA-repeat regions in a Ewing sarcoma context (Chapter 5)[223]. Despite the helpful contribution of our analysis, an important limitation to consider is our use of the hg19 human reference genome, rather than bona fide Ewing sarcoma cells, such as A673 cells. This patient-derived Ewing sarcoma cell line is widely used in the field[34,43,69,70,114]. The

experiments in this chapter sought to evaluate this limitation prerequisite to advancing the study of EWS/FLI allelic specificity.

Materials and Methods

FLI ChIPseq data

See methods and supplementary methods of Chapter 4. Raw sequence reads can be found in NCBI's Gene Expression Omnibus and are accessible through GEO SuperSeries accession number GSE94503.

RNAseq data

See methods and supplementary methods of Chapter 4. Raw sequence reads can be found in NCBI's Gene Expression Omnibus and are accessible through GEO SuperSeries accession number GSE94503.

Variant calling verification

We performed variant calling using the GATK best practices principle pipeline, to detect alterations between our ChIPseq reads and the hg19 human reference genome to which the data is aligned. Six EWS/FLI bound regions with sequence changes detected by our algorithm were chosen. Primers were designed within the sequences flanking these regions. Primer sequences are listed in Table 1. PCR amplification from A673 genomic DNA and purification of the expected DNA product size were performed. These were

commercially sequenced and manually compared with the hg19 and variant calling sequences, respectively.

Locus	Forward Primer	Reverse Primer
chr1:1886019-1887019	GTGCACTGGCAGAAGGAC	CCACATACGCATCGAGATCCAG
chr5:105352851-105353851	CCTGACCTAGTTCTGACTAGGC	GCTCAGCGCTATCCGCTTC
chr5:105352875-105353875	CTA GTT CTG ACT AGG CCA CAT GAG	GTCCAGGGTGATTAAGCTCTGG
chr7:147993757-147994757	CAGACACTATGAGCACTGC	CTCCTAGTGGTCAGTGGCTG
chr20:17372051-17373051	GCTGAGTCATGAGCTACTAGTGG	CCTGAAGGATGAGCTACATGAGAC
chr20:17148414-17149414	GAG ATG TCT TAA CGG GCT CAG	GTGAACCACACTGTTGCC

Table 6.1 Variant calling validation primers

PCR-based microsatellite amplification and sequencing

Ten EWS/FLI genes were selected at random (including genes studied extensively in our lab, as well as microsatellites spanning the repeat number spectrum according to hg19). Primers were designed within the sequences flanking these microsatellites. Primer sequences are listed in Table 2. PCR amplification from A673 genomic DNA and purification of the expected DNA product size were performed. These were commercially sequenced and numbers and patterns of GGAA repeats were compared with corresponding microsatellite sequences of the hg19 human reference genome.

Associated Gene	Forward Primer	Reverse Primer
PNMA2	GCCACTGCACTCTAGCCTGG	CACAGCCAGTGGGCTAAGACC
FIBCD1	CACAGAGACAGAGACACACGG	CAGGCAGGTGCTTTCTTAAGAATGG
PKP1	CAGGATATGTCGGTGTGGACC	GGGAACTAGATGCAATTACACAGC
FCGRT	GCAGTGAGCCATGATCGCTC	CCCTGGCAACCATTCATCTGC
GSTM4	GATCGCACCATTGCACTCCAG	CCTTCCTGGATGGTCCACC
PINK1	GCCATGAGGAGCGCTTGAAC	CTAATGCCCCAGCCTGGAGAC
TCERG1L	GCCGGACATCAAGCTTGTCTG	GTCATCTGCATTCTTGTGAGTTCC
PCSK2	CAGACACTATGAGCACTGC	CTCCTAGTGGTCAGTGGCTG
CTD-2078B5.2	CACCGTGTTAGCCAGGATGGTC	GACTGCACCACTGCACTGC

Table 6.2 PCR amplification Sequence Validation Primers

Results

Because of the long length of many GGAA-microsatellites, a sequencing platform capable of accurately tiling across such repetitive regions requires longer than average read lengths, such as with the PacBio system. The expensive nature of this approach presents a further, though not insurmountable limitation. As such, our bioinformatician looked to utilize our existing data, and overlay the hg19 version of the human reference genome with our ChIP-seq results. An algorithm was developed to call variants between the two sequences.

To evaluate whether this computational algorithm provides accurate microsatellite sequence information through variant calling, we chose six microsatellite-containing regions with called variants. We designed primers and PCR-amplified these regions in genomic DNA isolated from A673 Ewing sarcoma cells. We then compared the resulting sequence to both the reference genome, and sequence predicted from our variant calling algorithm (Figure 6.1). As an example, we amplified and sequenced the nearest GGAA-microsatellite to *PCSK2*, which our variant calling suggests adds four consecutive GGAA-repeats to the ten indicated in the hg19 version of the human reference genome. Our validation sequencing matched exactly with the variant calling, adding four repeats for a total of 14 GGAA motifs in the A673 *PCKS2* GGAA-microsatellite (Figure 6.1).

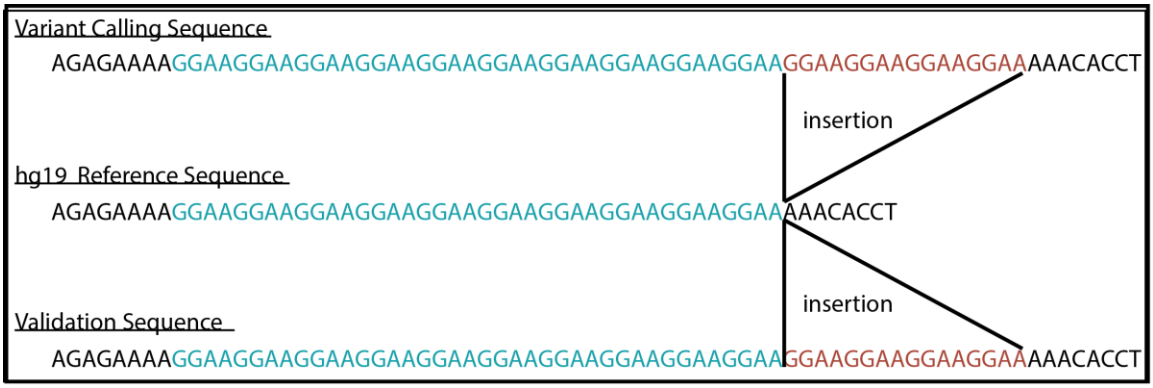


Figure 6.1 Sequencing validation of the PCKS2 GGAA-microsatellite

Interestingly, one of the other variant regions we examined predicted an addition of two more Ets consensus, or "high affinity" site sequences to only two tandem consensus sequences reported in the hg19 version. Our sequence validation for this variant call also

matched exactly, suggesting unexpectedly there may be tandem repeats of the Ets high affinity sites in addition to GGAA-microsatellites in Ewing sarcoma cells (Table 6.3).

Our variant call algorithm was also able to predict zygosity, with two of our six sample variant regions predicting heterozygous alleles with differing GGAA-repeat numbers. Our limited PCR-based methodology was able to verify the variant called for one, but not both alleles, as an exact match, for both heterozygous variant regions (Table 6.3).

Zygosity	**Genomic location**	**# GGAA repeats in hg19 genome**	**# GGAA repeats in variant call**	**# GGAA repeats in sequencing validation**	**Variant call match validation**
Homozygous	chr1: 1886487-1886559	2 High affinity sites	4 High affinity sites	4 High affinity sites	Yes
Homozygous	chr5: 105353332-105353420	13	18	9	No
Homozygous	chr5: 105353332-105353420	13	16	10	No
Homozygous	chr7: 147994226-147994325	10	14	14	Yes
Heterozygous	chr20:17372499-17372630	17	21	21	Yes
	2nd allele		22	23	No
Heterozygous	chr20: 17148855-17149002	11	10	10	Yes
	2nd allele		14		No

Table 6.3 Variant calling validation

Though these preliminary results indicate the relative robustness of our variant calling algorithm, we recognize the inherent bias in only evaluating regions called by this method. If the algorithm was unable to predict all the regions where microsatellite characteristics differ between the hg19 and A673 genomes, this approach would be unable to detect variants missed by our computational method. To test differences between microsatellite lengths in the hg19 versus A673 genomes, we chose ten well-characterized EWS/FLI regulated genes. After identifying the nearest microsatellite to these genes, we PCR-amplified and sequenced these regions. We then compared GGAA-repeat numbers between the hg19 genome and our sequencing results (Table 6.4).

Sequencing results were undeterminable for four of the genes, but the six that worked showed some differences in GGAA-repeat number. For each of these microsatellites we also included the ChIP peak binding fold change of EWS/FLI, as well as the log2 fold change of differential gene expression for the gene nearest to this microsatellite (Table 6.4). Importantly, both of these ChIP-seq and RNA-seq experiments were conducted in A673 Ewing sarcoma cells.

Gene	Microsatellite Location	# GGAA repeats in hg19 genome	# GGAA repeats in seq validation	FLI binding peak Fold Change	RNA expression log2 Fold Change
PNMA2	chr8: 26365042-26365070	6	-------	3.29	-1.02
FIBCD1	chr9: 133846346-133846444	14	14	4.44	2.6
PKP1	chr1: 201235502-201235577	17	-------	35.45	-1.69
FCGRT	chr19: 50014759-50014875	23	20	19.43	-3.07
GSTM4	chr1: 110196838-110196930	18	-------	16.19	-2.58
NR0B1	chrX: 30328875-30328976	25	25	31.47	-2.58
PINK1	chr1: 20952451-20952680	45	9	4.95	2.16
TCERG1L	chr10: 133145119-133145548	44	30	6.94	-2.59
CTD-2078B5.2	chr5: 39610634-39611090	68	-------	3.49	0.9
PCSK2	chr20: 17372553-17372661	17	21	15.8	-1.5

Table 6.4 PCR amplification sequence validation

To determine whether accurate GGAA-repeat length is a better predictor of FLI binding and EWS/FLI-mediated gene expression than using the human reference genome, we assessed binding and gene expression with respect to number of GGAA-repeats for each of these genes (Figure 6.2). FLI binding comparisons showed a “sweet-spot” of binding within the previously determined 20-26 GGAA-repeat range (Figure 6.2A). Notably, five

of the six genes remained in their respective "sweet-spot" or non-"sweet-spot" category, despite changing between the hg19 and A673 genomes. This suggests that although our microsatellite characterizing data as predictors of EWS/FLI responsiveness is limited by using the hg19 reference genome sequences, the correlations we observed are largely accurate trends[223]. *PCSK2*, however, was one gene whose peak binding fold change was similar to binding values of genes associated with "sweet-spot" microsatellite lengths, despite having only 17 GGAA repeats according to the hg19 sequence. Sequence validation at this site, however, demonstrated an actual GGAA-motif number of 21 in A673 sequence, falling within the "sweet-spot" repeat range (Figure 6.2A). The gene expression data for this gene fits the overall GGAA-microsatellite length-dependency model better with the accurate (A673) sequence length of microsatellite. This further supports the need for a reference A673 genome to provide more accurate correlations in our bioinformatics analysis.

To test whether this same pattern exists with microsatellite length as a predictor of gene expression, we examined hg19 versus A673 GGAA-repeat number for this same subset of genes, compared with their differential gene expression from our RNA-seq dataset (Figure 6.2B). We found expression changes vs. motif number between the two genomic sources paralleled our binding data (Figure 6.2C). Interestingly, the *PINK1*-associated microsatellite contains 45 repeats in the hg19 genome, but only 9 repeats in A673 cells. This drastic sequence difference has since been validated with two additional sequencing methods. Regardless of the change, these repeat numbers fall far on either side of the

"sweet-spot" range microsatellite length, reflected in poor FLI binding at the microsatellite (Figure 6.2A) and a significant loss of *PINK1* expression (Figure 6.2B-C).

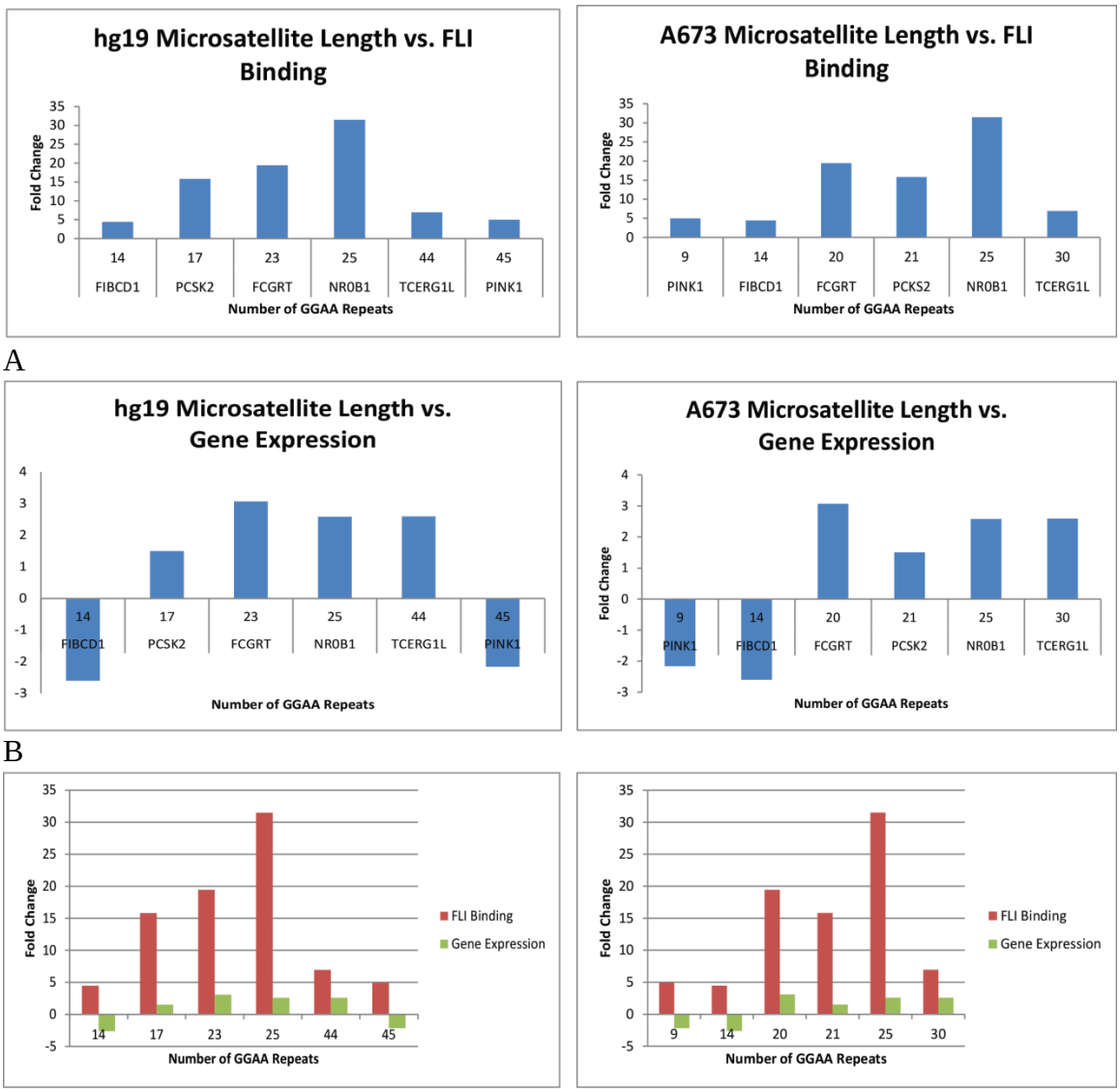


Figure 6.2 EWS/FLI-mediated binding and gene expression based on GGAA-repeat number

(A) GGAA-microsatellite repeat length as a predictor of FLI binding for the hg19 vs. A673 genome **(B)** GGAA-microsatellite length as a predictor of EWS/FLI-mediated gene expression for the hg19 vs. A673 genome **(C)** Comparison FLI binding and gene expression by increasing GGAA-repeat number for the hg19 and A673 genomes

Discussion

The implied significance of polymorphic heterozygous sequences is highly reminiscent of our studied GGAA-microsatellite response elements in Ewing sarcoma[16]. In chapters 3, 4, and 5, we demonstrated the importance of "sweet-spot" length in EWS/FLI binding, both *in vitro,* and *in vivo*[16,217,223]. In chapter 3, our clinical observation strongly supported the heritability of these sequence length polymorphisms, suggesting a significant role in Ewing sarcoma patient susceptibility[16]. After defining microsatellites and evaluating whether EWS/FLI responsiveness correlates with particular microsatellite characteristics, however, we recognized an important limitation in our study (chapter 5)[223]. Our data comparing microsatellite lengths, total repeat numbers, and maximum number of consecutive motifs, are comprised of microsatellite sequences pulled from the human reference genome. While we still saw significant, though minimal, correlations[223], we considered whether these might be more accurate using genomic sequences derived from a Ewing sarcoma patient.

The preliminary analysis presented in this chapter demonstrates that data using hg19 human reference genome sequences to test whether GGAA-microsatellite length correlates with EWS/FLI responsiveness at particular genomic loci is sufficient to represent the appropriate trends. However, the data also suggest re-evaluation of these correlations using A673 microsatellite sequences would make the correlations more accurate and stronger, as predicted. These computational studies of molecular

mechanisms provide valuable and unique insight into potential correlations at the mechanistic root of EWS/FLI-mediated Ewing sarcoma development.

To this end, there is a sufficient need for an A673 Ewing sarcoma reference genome. In collaboration with a genomics group, we recently made strides towards this using the 10x genomics sequencing platform. Preliminary data, however, while exciting, suggests a coverage disparity between 10x genomics, our ChIP-seq variant calling algorithm, and PCR-amplified sequencing limitations. In other words, our future directions include a need for the PacBio sequencing platform for longer read coverage, and therefore more reliably accurate sequence information.

Once this reference genome is obtained, it can be used to help determine whether EWS/FLI demonstrates allelic specificity. Earlier studies contained in this thesis (Chapter 3) presented clinical data suggesting at least on a population level, "sweet-spot" microsatellite length may be associated with increased disease susceptibility[16]. Furthermore, the *EGR2*-regulated microsatellite has been identified as a specific susceptibility locus in patients with Ewing sarcoma[32]. To test allelic specificity in EWS/FLI binding at microsatellites, we plan to combine genomic ChIP-seq FLI binding data with RNA-seq expression data to differentiate alleles through a SNP-based approach. Additional methodologies may also be adopted based on recent advances in the field linking bioinformatics and empirical evidence to test similar models.

For example, a number of groups have used ChIP-seq and a bioinformatics approach to test whether transcription factors exhibit binding preferences at particular allelic sequences. Specifically, differential distribution of sequence reads in ChIP-seq peaks across a heterozygous sample can be used to identify potential allele-specific bias of protein binding[222]. *Kasowski et al.* analyzed ChIP-seq data mapping RNA Polymerase II and *NFkB* binding in several humans and in one chimpanzee. They observed that differences in binding of these factors, and subsequent gene expression variation, were frequently associated with SNPs and other genomic variants[224]. Another group looked at genome-wide transcription factor binding in liver (HePG2) and cervical cancer (HeLa-S3) cell lines in association with GWAS-SNP identification[225]. They identified 3713 AS (allele-specific)-SNPs representing candidates of functional regulatory variants that could cause the observed expression differences.

The frequent association of SNPs with transcription factor binding suggests that *cis* elements, such as GGAA-microsatellites in Ewing sarcoma, are a critical heritable element of allelic-specific expression[16,224]. Given differential binding, two possibilities arise: that one allele provides more favorable binding for that specific protein, thereby enabling a gain-of-function effect, or, two, that the other allele disrupts or in some way interferes with binding[222]. Analyzing such specificity in protein-DNA binding contributes valuable insight to the possible functional consequences associated with binding at particular alleles.

While a valuable approach to address the question of allele-specific binding, most ChIP-seq variant methods for detecting allelic imbalance assume diploid genomes. This is not generally accurate for assessing transcription factor binding in cancer genomes, which frequently demonstrate copy number variation. Such a disparity generates a huge statistical limitation in detection of these specific alterations. A Bayesian statistical approach (BaalChIP) was recently developed to mitigate this limitation by jointly analyzing multiple ChIP-seq samples across a single variant[226]. A combination of ChIP-seq and FAIRE-seq samples were used as a proof of concept and demonstrate the power and effectiveness of their method.

An alternate method to address the allelic imbalance bias was developed with similar rationale. Known as the ABC method, it applies a binomial probability test for variant calling of allele-specific biases detected in ChIP-seq reads via comparison with the genomic allele ratio (gAR), or number of reads expected to evenly match to each allele, assuming no bias[227]. Two strand-specific read piles are generated around the genomic loci where the transcription factor binds and the Fisher's exact test is performed to compare strand distribution for both alleles, taking into account the SNV's position. Other statistical tests are also applied to account for potential read position bias.

Another interesting observation to come out of these studies is examples of allele-specific expression within non-coding DNA regulatory regions that give rise to differential gene expression[221]. In the ChIP-seq study of RNA Polymerase II and NFkB, both displayed

sequence variation differences in non-coding regulatory regions as significantly higher than in coding regions[224]. In contrast to EWS/FLI-regulated GGAA-microsatellites, similarly present in non-coding regions, however, it is not known whether these variants are heritable[221].

In conclusion, the purpose of the work and ideas presented in this chapter is to determine to what extent the GGAA-microsatellite sequences in the hg19 human reference genome differ from those in the A673 Ewing sarcoma cell line, and to evaluate how those differences affect the binding enrichment and differential gene expression correlations seen in our computational analysis. Our current data to this end is preliminary, yet provides strong evidence for continued investigation of these aims. Such analysis will allow us to expand on recent bioinformatics methodologies to computationally investigate whether EWS/FLI exhibits allelic-specific binding of GGAA-microsatellites. Such analysis would provide insight not only for Ewing sarcoma genetics and oncogenesis, but more broadly for our understanding of transcription factor gene regulation.

Chapter 7: Advances and new approaches to study EWS/FLI DNA binding at GGAA-microsatellites

Introduction

Normally, ETS family members like FLI bind a conserved sequence containing a single GGAA core motif by a monomeric DNA binding domain[63,48,52]. This DNA binding domain is necessary for oncogenesis[6,64,66], with FLI and EWS/FLI displaying similar DNA binding affinity and specificity[37]. In Ewing sarcoma, however, EWS/FLI preferentially binds low-affinity GGAA-microsatellite (repeat) regions upstream of the genes it directly activates rather than the normal conserved high affinity consensus sequence[121]. Previously thought of as "junk DNA," these microsatellites serve as response elements for EWS/FLI DNA binding with interesting genetic correlations and possible clinical implications. Characterizing EWS/FLI binding at length-dependent GGAA-microsatellites is important to achieving a better understanding of the specific mechanism by which EWS/FLI modulates transcriptional activity.

A main focus of the data and models presented within this work has been to determine whether microsatellite length facilitates biochemical properties of DNA-protein binding that engender EWS/FLI target specificity. The purpose of this section is to present new approaches and advances in investigation of the overall EWS/FLI-DNA binding objective. Our current data to this end highlights progress in the following areas:

- Immuno-precipitation assays to study the binding properties of EWS/FLI and EWS/FLI mutant constructs on various lengths of GGAA-microsatellites.
- Fluorescence polarization studies using full-length EWS/FLI recombinant protein to determine its binding affinity on increasing lengths of GGAA-microsatellites.
- Continued CRISPR/Cas9 studies with homologous recombination using donor templates of various GGAA-microsatellite lengths to determine the length-dependency of EWS/FLI microsatellite binding and oncogenic function *in vivo.*
- Molecular modeling of FLI binding on GGAA-microsatellites to computationally evaluate the structural binding mechanism of this protein-DNA interaction.

Materials and Methods

BioDIP Assay

For this IP, 20ul/sample of streptavidin beads (Dynabeads M-280 Streptavidin, Invitrogen) were pre-washed 3x with wash buffer consisting of 5mM Tris pH 8.0, 1M NaCl, and 0.5mM EDTA. After final wash, beads were resuspended in 100ul 2x wash buffer, followed by addition of 1uM annealed, biotinylated DNA oligo. Bead-DNA solution was rotated at room temperature for 30 minutes. 3x 5minute washes were then performed using 1x wash buffer, with a final resuspension in 300ul TE buffer. For the pull-down assay, 20ul of each oligo/bead mixture were aliquoted in separate tubes for each condition to be performed (time series, protein concentration variation, etc.). A 200ul 1% biotin/1%BSA solution was added to each sample, with rotation at room temperature for 30min to reduce non-specific binding. Bead/oligo mixtures were then

resuspended a total volume of 100ul binding buffer (10mM Tris pH 8.0, 50mM NaCl, 1mM $MgCl_2$, 0.5mM EDTA, 4% glycerol, 1% BSA, 4ug/ul poly-didc) plus recombinant protein, and rocked at room temperature for 1 hour (or different time points depending on experiment). 3x 5minute washes were performed with 1% BSA, 1% NP40, 1mM DTT, and 2mM EDTA in 1x PBS. After a final wash in TE buffer to remove residual detergents, the bead mixtures were resuspended in 20ul TE buffer. SDS was added and western blot performed for IP analysis, using FLI antibody (Abcam, USA).

ChIP-DIP Assay

Briefly, streptavidin beads (Dynabeads M-280 Streptavidin, Invitrogen) were washed in Dilution Buffer (1M Tris pH 8.0, 5M NaCl, 0.5M EDTA, 20% NP40, 1 PI tablet). Anti-FLI antibody was added (Santa Crus sc356x) and rotated at 4°C for a minimum of 6 hours. When the beads are ready, 2ul of 50uM DNA is added to 1uM of recombinant protein in 100ul of Binding Buffer (10mM Tris pH 8.0, 1mM EDTA, 2M NaCl). Binding reaction proceeds at room temperature for 15 minutes. Protein/DNA mix is then added to the bead/antibody mixture in Binding Buffer, with 500ul total volume). This reaction rotates at 4°C for 2-6 hours. Three, 5min washes are then performed with Wash Buffer (1% NP40, 2mM EDTA, 1mM DTT, 1% BSA, PBS). Protein is then eluted by adding 150ul fresh elution buffer (50mM Tris pH 8.0, 1% SDS, 50mM $NaHCO_3$, 1mM EDTA) and nutating at 37°C for one hour. Elution (containing desired product) is removed from beads, the salt concentration is brought to 200mM, and nutated overnight at 67°C.

Proteinase K and RNAse A are added. DNA is then purified (PCR purification kit, Qiagen) and electrophoresis performed to assess DNA product size.

Immunodetection

The following antibodies were used for immunodetection: anti-FLI (Abcam ab15289), anti-α-Tubulin (Calbiochem CP06).

Cell culture and CRISPR/Cas9 experiments

HEK 293EBNA cells and Ewing sarcoma cell lines (A673, TC71, EWS/502) were infected with CRISPR/Cas9 lentivirus and A673 cells were retrovirally infected as previously described for EWS/FLI knockdown/rescue experiments[43,114]. Single-stranded oligos of various lengths of GGAA-microsatellites used as donor templates were transfected or electroporated into the cells 2-4 hours following lentivirus infections. Polyclonal cell populations were grown in the appropriate selection media[43,114]. Growth assays were performed in 96-well plates on the IncucyteZoom live cell imager. Briefly, 8000 cells/well were seeded in triplicate and imaged every 4-6 hours for 7-10 days. IncucyteZoom software pre-calibrated for Ewing sarcoma cells measured cell confluency levels as a percentage to assess cell growth over time. Genomic DNA was harvested following antibiotic selection and used in PCR assays. The deleted region was detected with primers designed around the *NR0B1* microsatellite (Table 4.1).

Fluorescence Polarization

Fluorescence polarization was performed in 384-well format using a BioTek Synergy2 fluorometer (Winooski, VT) with fluorescein-labeled probes containing either 12 or 22 consecutive GGAA motifs and 1x Gel Shift binding buffer (Promega Corporation, Madison, WI). Sequences of the consecutive GGAA motifs harboring probes and control sequences used in these assays are listed in Table 4.1. Recombinant proteins were prepared as described above.

Results

Studying EWS/FLI-microsatellite binding via protein immunoprecipitation and detection

Our biochemical assays thus far have measured EWS/FLI binding kinetics at GGAA-microsatellites of increasing length (Chapter 4)[217], and transcriptional activity via reporter assays (Chapter 3)[16]. To test whether we could visually assess EWS/FLI binding at GGAA-microsatellites, we sought to develop an *in vitro* binding assay. Our immuno-precipitation-based approach, called BioDIP, utilizes streptavidin-coated beads incubated with biotin-labeled DNA oligos to specifically pull-down protein that binds these sequences. Following IP, protein denaturation and immuno-blotting are used for detection.

As a proof of concept, we first tested our EWS/FLI deletion construct Δ22, to see whether we could successfully pull-down recombinant protein using this method (data not shown). Though we achieved protein expression, we recognized the need for protein concentration optimization for each respective mutant. To optimize protein

concentrations required to visualize binding saturation, we performed the assay using increasing concentrations of Δ22 and Mut9, respectively (Figure 7.1A). We found the Mut9 construct required a nearly ten-fold higher micromolar concentration than Δ22 to achieve saturation with immunoblotting. This could be explained by a number of factors, such as the purity of our respective protein preps. An additional possibility is based on the inherent aggregate nature of the EWS-containing mutant, Mut9. Such a propensity to aggregate significantly increases the likelihood of purifying incompletely functional protein. For example, if only ten percent of the subsequent Mut9 protein prep was fully functional, it follows that ten-fold less protein would be capable of binding the DNA. As a result, though identical original starting concentrations were used for Δ22 and Mut9, the latter would appear ten times less saturated upon immunodetection. The possibility of this concentration discrepancy emphasizes the crucial role of careful optimization required for accurate interpretation of these experiments.

We next considered the possibility of non-specific protein binding to the streptavidin beads in our assay. To address this confounding factor, we sought to optimize the washing steps of our assay. Testing a variety of variables including salt stringency, time, and addition of molecules to saturate beads incompletely bound with biotinylated DNA, we determined the optimal wash conditions to give minimal background, or non-specific binding. This included incubation of the bead-DNA oligo mixture with 1-2.5% BSA prior to addition of recombinant protein (Figure 7.1B).

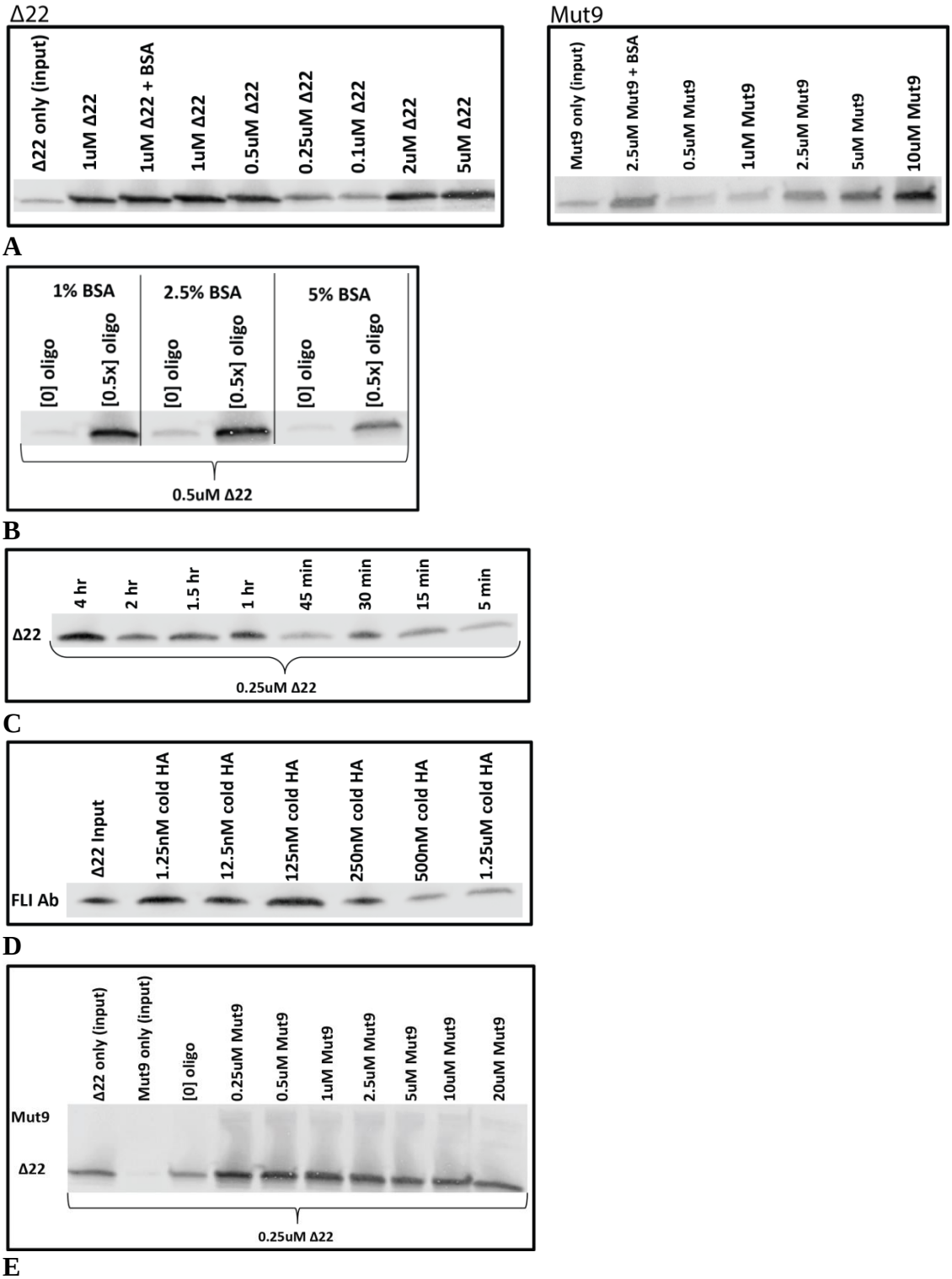


Figure 7.1 BioDIP assay trials testing Δ22 and Mut9 pull-down

Figure 7.1 continued

(A) Optimization of protein concentration for Δ22 and Mut9 **(B)** BSA concentration optimization for background reduction **(C)** Time course experiment to detect Δ22 on and off rate binding to GGAA-microsatellites **(D)** Competition assay with unlabeled ("cold") high affinity ETS DNA sequence **(E)** Competition assay with Mut9 protein to compete off Δ22 binding of microsatellites

A protein-DNA binding assay has a number of obvious potential applications, including kinetic experiments and binding specificity. Our previously conducted FP experiments to test binding affinity (K_D) for Δ22 and Mut9 were reproducible, yet lacked the visual assessment afforded by our BioDIP method. Having found Mut9 required a different concentration of protein to illicit the same binding saturation, we sought to use this assay as an alternate means of calculating K_D. This requires kinetic experiments to measure the on and off rate for protein binding to the DNA. Starting with a time point of 4 hours for protein and DNA binding, we collected a series of additional time points to measure the on-rate for Δ22 binding the DNA. We found Δ22 binding to GGAA-microsatellites is detectable at as early as 5 minutes post-incubation (Figure 7.1C). As such, shorter time points will be required to optimize this method of on-rate detection. Measurement of the off-rate requires knowing the definitive time point at which protein binding of the DNA is *just* saturated. Though we conducted a preliminary experiment to measure off-rate using a time point of 30 minutes based on our on-rate trial, more precise measurement of the on-rate will be prerequisite to further optimization for this application of our assay (data not shown).

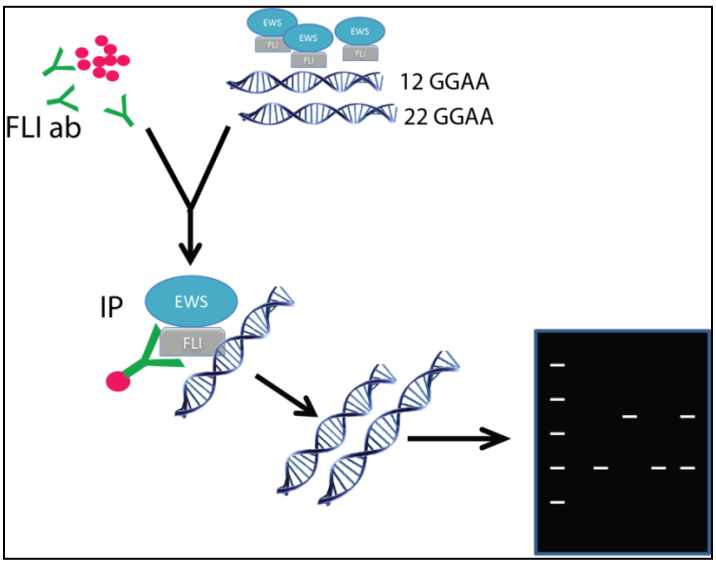
To determine whether this assay can assess binding specificity, we added increasing concentrations of cold competitor DNA to Δ22 incubated with GGAA-microsatellites. We used unlabeled high affinity Ets consensus sequence as competitor DNA, as this is reported to bind FLI with at least ten-fold higher affinity than GGAA-microsatellites[37]. We found Δ22 binding at the high affinity site appeared to compete off GGAA-microsatellite binding starting at a 1:1 competitor to microsatellite ratio (Figure 7.1D). Having successfully competed off microsatellite binding, we next asked whether similar experiments could be conducted with protein competitors.

While Δ22 binds shorter microsatellites with higher affinity than Mut9, we previously found the lack of the EWS portion results in failure to bind "sweet-spot" microsatellites while its inclusion significantly improves binding at this length (Chapter 4)[217]. To test whether this difference between Mut9 and Δ22 binding of "sweet-spot" microsatellites enables competitive binding between these EWS/FLI mutants, we performed a competition assay. Increasing concentrations of recombinant Mut9 protein were added to Δ22 binding of a 22-repeat GGAA-microsatellite. Our preliminary results show a decrease in Δ22 binding saturation with increasing Mut9 (Figure 7.1E). The extent of this competitive binding, however, is currently limited by the maximum concentration of Mut9 protein we are able to purify, considering the previously discussed limitation of Mut9 functionality. Taken together, this assay provides several useful applications for study of EWS/FLI binding at GGAA-microsatellites, however, further optimization of the technique is required.

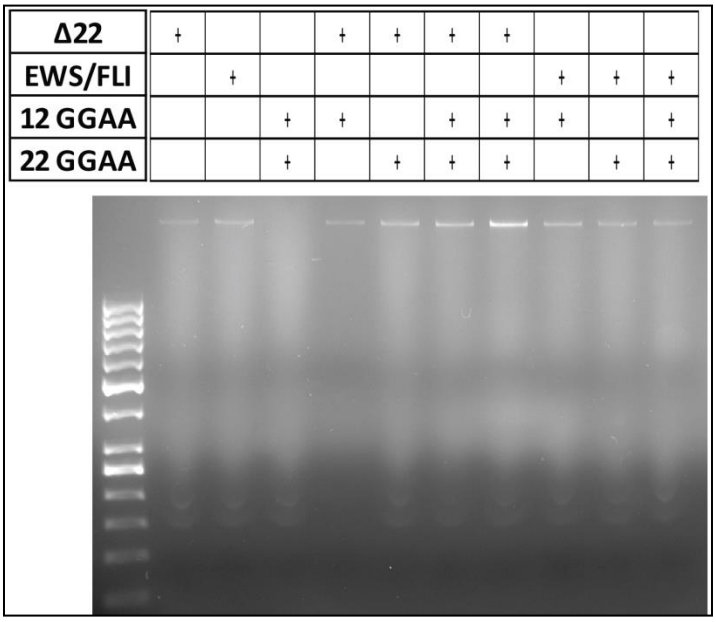
Studying EWS/FLI-microsatellite binding via DNA pull-down and detection

An intriguing model of EWS/FLI binding at microsatellites *in vivo* is the hypothesis of allelic specificity (discussed in Chapter 6). To see whether this model can be tested in an *in vitro* setting, we adapted our ChIP experiments. Opposite of BioDIP, this assay (called ChIP-DIP) relies on antibody-coated proteins to pull down the DNA, for a final detection that is nucleic acid-based (Figure 7.2A). In our preliminary trial, we added both 12 and 22 GGAA-repeat DNA oligos to Δ22 vs. EWS/FLI recombinant protein. These were combined with FLI antibody-coated beads. Following washes, elution, and protein digestion, the remaining DNA was assessed by gel electrophoresis. We hypothesized that if EWS/FLI does exhibit preferential binding for length-specific alleles, then we would detect more 22 GGAA-repeat DNA binding.

Our FP results of Δ22 demonstrated optimal binding at shorter (such as 12 repeats) with significantly reduced binding affinity at the sweet-spot range (Chapter 4)[217]. As such we predicted ChIP-DIP results for Δ22 binding would detect preferential 12-repeat, rather than 22 GGAA-repeat binding. Our first pass at this experiment was unsuccessful, showing nothing more than non-specific nucleic acid detection by electrophoresis across all samples (Figure 7.2B). If appropriately optimized, however, this technique would provide a valuable tool to assess EWS/FLI binding preferences on microsatellites in a length-dependent manner. An *in vitro* approach is expedient to verify the computational analysis and predictions of whether EWS/FLI truly displays allelic specificity in GGAA-microsatellite binding (Chapter 6).



A



B

Figure 7.2 ChIP-DIP assay **(A)** Work flow of assay **(B)** Assay preliminary trial

Binding affinity by fluorescence polarization

While Mut9 and Δ22 are invaluable constructs that enable isolation of key EWS/FLI components contributing to different aspects of its structural and functional behavior,

EWS/FLI is still best understood through study of its full-length protein. Because of a variety of informative studies comparing Mut9 and Δ22 (Chapter 4)[217], we anticipated that full-length EWS/FLI would display similar binding behavior to Mut9. To see whether EWS/FLI binds shorter GGAA-microsatellites with poor affinity, which improves with binding microsatellites of "sweet-spot" length, we repeated our fluorescence polarization (FP) studies using recombinant full-length EWS/FLI. We found that, as predicted, length-dependent EWS/FLI binding affinity appears to recapitulate Mut9 at microsatellites. The K_D for EWS/FLI binding improved from a poor affinity of 1.57uM on 12 GGAA-repeats, to a ten-fold tighter binding of 134nM K_D on 22 GGAA-repeats (Figure 7.3). This result is highly repeatable. Taken with previous studies in multiple types of assays (Chapter 4), this data validates Mut9 as an acceptable minimal construct for study of full-length EWS/FLI[217].

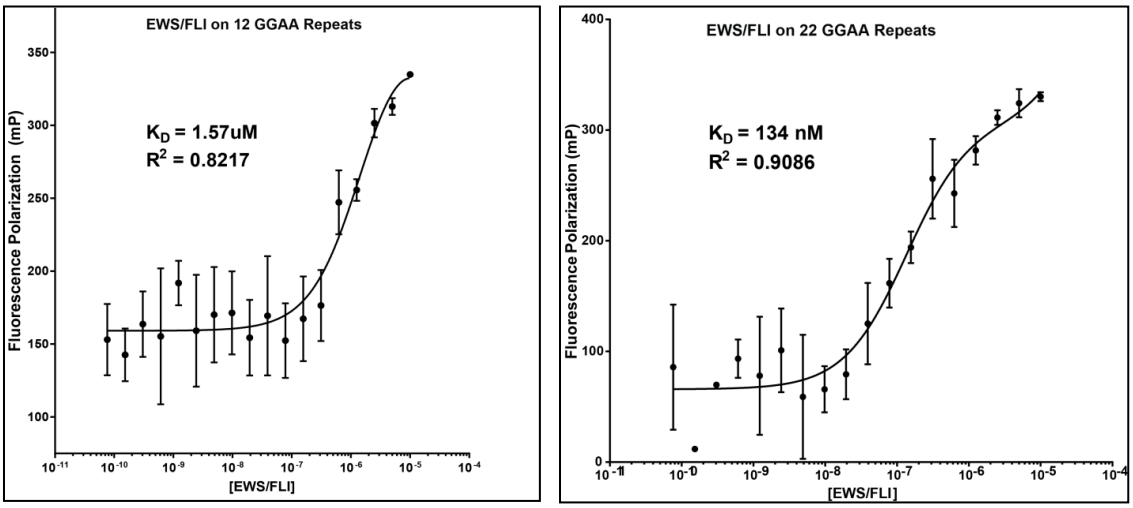


Figure 7.3 Fluorescence polarization assays measuring EWS/FLI binding on 12 vs. 22-repeat GGAA-microsatellites

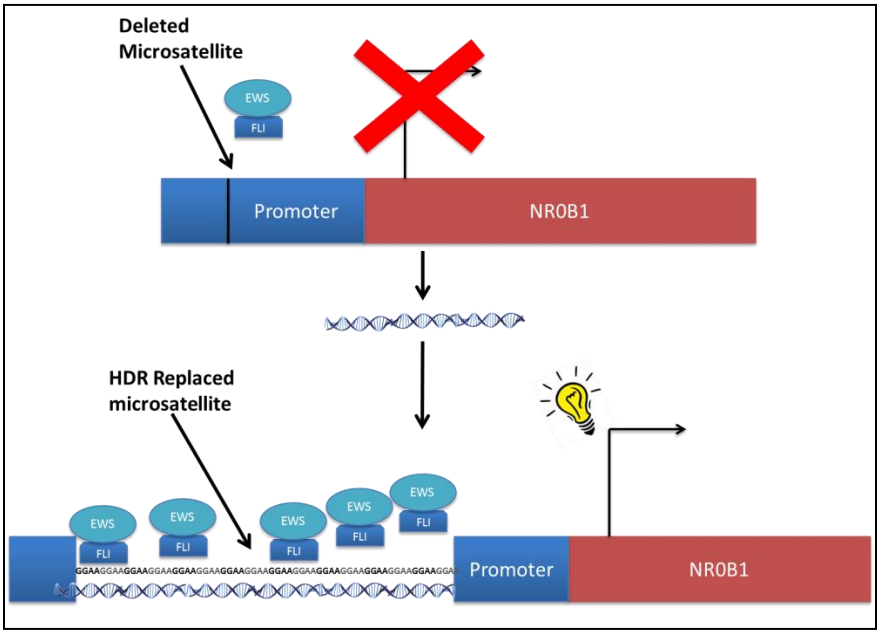
Additionally, this data presents, to our knowledge, the first assessment of K_D for full-length EWS/FLI microsatellite binding. In addition to providing valuable quantifiable data about protein-DNA binding, fluorescence polarization assays can also be used as a marker of protein functionality. Testing protein preps that have lost functionality due to sample age, flawed prep, etc. demonstrates non-readable binding affinities, with any detectable K_D's measured at several micromolar (data not shown). As such, the binding affinity readability and reproducibility for full-length EWS/FLI supports our achievement of functional, bona fide recombinant protein. This new, crucial tool in our arsenal will prove an invaluable contribution to future biochemical assays.

Genome-editing to evaluate EWS/FLI binding at GGAA-microsatellites

Though *in vitro* data suggests EWS/FLI exhibits binding preference for a particular length of GGAA-microsatellite, the translational value of this finding has yet to be explicitly evaluated in a Ewing sarcoma cell context (Chapter 3-5)[16,217,223]. We have shown in this thesis work through CRISPR/Cas9-mediated knock-out of the *NR0B1*-associated microsatellite, that EWS/FLI requires this region to transcriptionally activate the *NR0B1* gene (Chapter 4)[217]. Additionally, because *NR0B1* is required for EWS/FLI-mediated oncogenesis, deletion of this GGAA regulatory region also disrupts normal Ewing cell proliferation and colony formation ability, providing a helpful phenotypic manifestation of genome-editing[43,217].

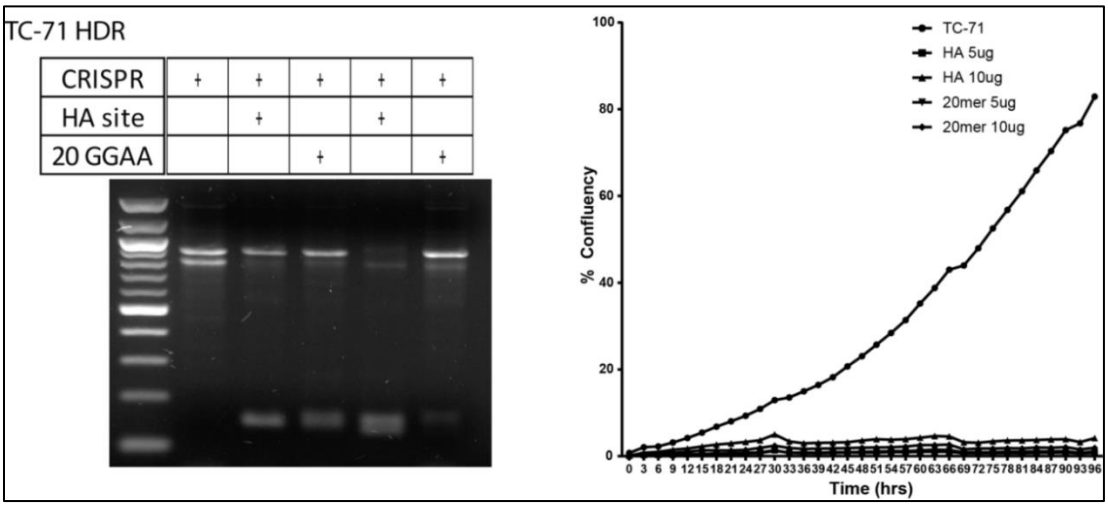
The CRISPR/Cas9 system provides an ideal means for evaluating the length-dependent effect of microsatellites on EWS/FLI gene regulation in Ewing sarcoma cells. To determine whether *in vivo* EWS/FLI transcriptional regulation of *NR0B1* changes with GGAA-microsatellite length, we knocked out this microsatellite as before, but also supplied a single-stranded oligo donor template of particular GGAA-microsatellite lengths to prompt homology directed repair (Figure 7.4A). We hypothesized that peak *NR0B1* gene expression would result from microsatellites of sweet-spot length, whereas GGAA-repeat donor templates shorter or longer than the sweet-spot would demonstrate reduced gene expression.

Figure 7.4 CRISPR/Cas9 strategy to induce HDR-mediated replacement of the deleted *NR0B1* microsatellite region with specific lengths of GGAA-microsatellite DNA template

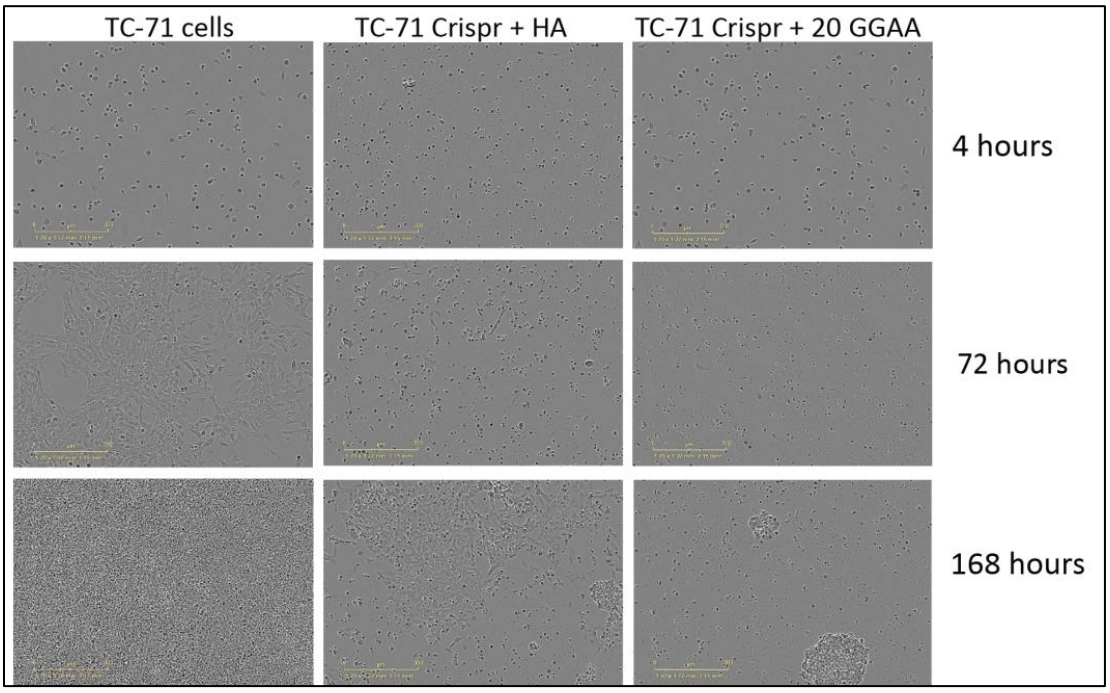


A

Figure 7.4 continued



B **C**



D

(A) Schema of GGAA-microsatellite length-replacement strategy by CRISPR/Cas9 **(B)** PCR amplification of microsatellite deletion and donor template replacement **(C)** Growth curve of proliferation assay results for each of these conditions **(D)** Images from proliferation assay

We first tried different concentrations of high-affinity site DNA template vs. 20 GGAA-repeat DNA template, and assessed for deletion and repair using PCR-amplified genomic DNA from each condition. We found that while this polyclonal cell population demonstrated deletion of the microsatellite region, we were not able to visualize whether the cells had incorporated the appropriate donor template supplied (Figure 7.4B). Because *NR0B1* is on the X chromosome, we performed these experiments in the male Ewing patient-derived TC-71 cell line, so as to increase the likelihood of successful recombination by only dealing with one allele. Proliferation assays showed complete growth inhibition, as seen in deletion of the *NR0B1* microsatellite, suggesting that our donor sequence was not incorporated (Figure 7.4C-D). Additional attempts in A673 Ewing sarcoma cells have also yielded similar results.

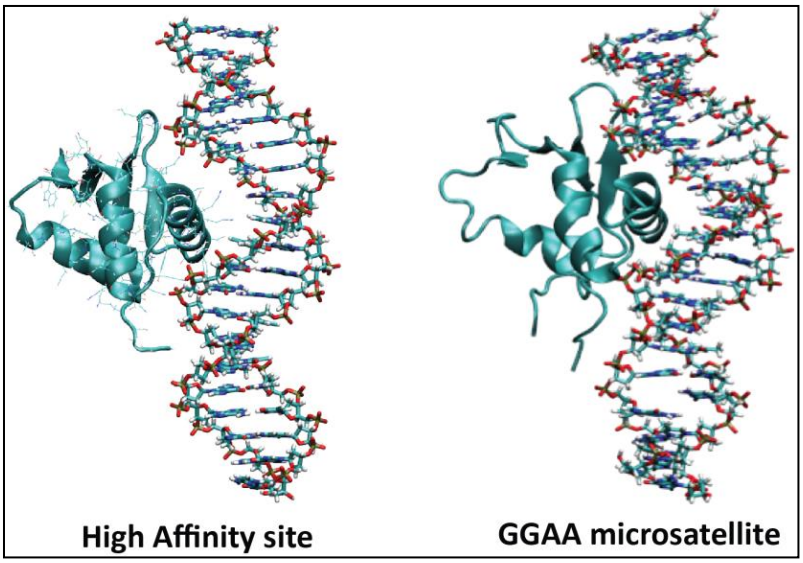
In general, HDR (homology-directed repair) is an extremely inefficient pathway, leaving non-homologous end-joining (NHEJ) as the default means of DNA repair. To combat this issue, there are a number of different methods that may improve HDR efficiency, including using a donor plasmid with 800bp homology arms instead of a single-stranded (ss)-oligo, linearizing said plasmid, inhibiting DNA ligase IV to block the NHEJ pathway (via shRNA or drug inhibition), and addition of Ad4 expression plasmids to increase the likelihood of HDR. Though these other strategies may prove helpful for CRISPR/Cas9-HDR broadly, some studies have suggested defects in DNA repair pathways in Ewing sarcoma[228,229]. Such defects may prove prohibitive to HDR-induced genome editing for study of this particular disease.

A recent study showed that Cas9 takes approximately 6 hours to dissociate from the DNA after making its enzymatic cut[230]. It may be that Cas9 is blocking the donor template from being incorporated into homology repair, as we have thus far introduced the Crispr/Cas9 plasmid and donor template simultaneously or within 2-4 hours of each other. Future attempts at this technique could try increasing the time between CRISPR/Cas9 knockout of the microsatellite and provision of the donor template. Inhibition of the NHEJ pathway and delayed addition of the donor template may also improve HDR efficiency. Though a difficult technique, successful HDR-induced GGAA-microsatellite replacement would significantly contribute to further elucidating the biology of both gene and allele-specific transcriptional activation mediated by EWS/FLI.

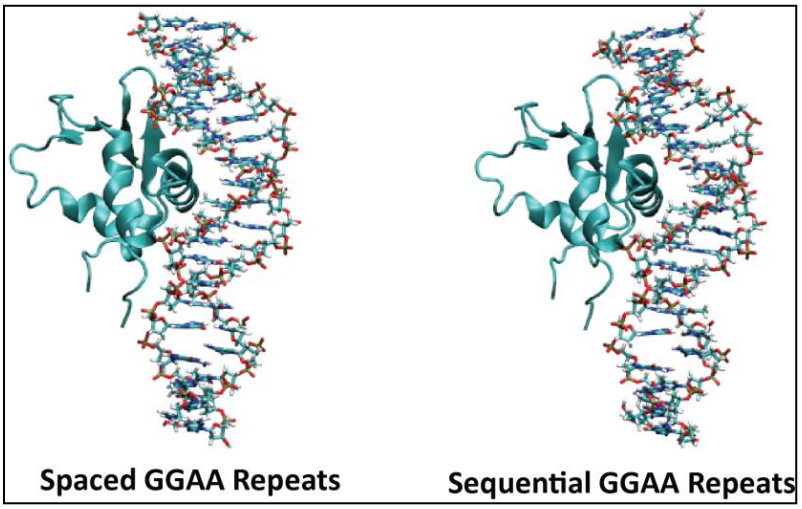
Molecular Modeling of EWS/FLI-GGAA-microsatellite binding

As discussed in chapters 5 and 6 of this thesis, computational approaches contribute informative and frequently unique insight to build on molecular studies[223]. In a brief exploratory collaboration, we examined a number of predicted FLI-DNA structural interactions using a molecular dynamics computational approach. To test if FLI is predicted to interact with DNA in a sequence-dependent manner, we modeled FLI binding on the Ets high-affinity site compared with GGAA-microsatellites of four repeats (as this was believed to be the minimal repeat-length required for EWS/FLI microsatellite binding)[51]. Though merely preliminary, our modeling results demonstrate that FLI bound to a GGAA-microsatellite appears to adopt a more bent configuration than when bound to the high affinity site (Figure 7.5A). This implies the likelihood that increased numbers of

GGAA-repeats lead to altered DNA structural configurations that may be essential for the curious observations of optimal EWS/FLI binding at "sweet-spot" GGAA-microsatellites.



A



B

Figure 7.5 Molecular modeling of FLI binding **(A)** Ets high-affinity site vs. 4-repeat GGAA-microsatellite DNA **(B)** 4 spaced vs. consecutive GGAA-repeats

One of the questions to arise from our bioinformatics analysis of EWS/FLI binding at microsatellites was whether number or consecutiveness of GGAA-repeats permitted optimal binding (Chapter 5)[223]. To computationally evaluate whether FLI binding at four spaced GGAA-repeats differs from binding at consecutive repeats, we ran simulations for each of these as well (Figure 7.5B). We found binding at these different sequences appear similar, however, this is only for monomeric FLI binding. It would be highly informative to model EWS/FLI, and even multimers of EWS/FLI binding to these same, and longer GGAA-repeat DNA sequences. Unfortunately, this requires a known crystal or NMR-based protein structure, which currently doesn't exist for EWS/FLI.

Discussion

Though now well established that GGAA-microsatellites function in Ewing sarcoma as EWS/FLI response elements, the mechanism by which EWS/FLI regulates transcriptional activation at these sites remains elusive[15,16,46,51,69,217]. Very recent advances in the field have demonstrated a neomorphic role for EWS/FLI in recruiting chromatin remodeling complexes to GGAA-microsatellites[47]. However, little progress has been made to explain the molecular reasoning behind both our clinical and biochemical observations of "sweet-spot" microsatellites enabling optimal EWS/FLI binding and effector function at these sites[16,217,223].

This chapter details new approaches and our latest advances in study of EWS/FLI-microsatellite binding. The variety of molecular methods utilized includes

immunoprecipation using both protein and DNA pull-down, fluorescence polarization, genome-editing, and even molecular modeling. Specifically, we developed a working DNA-protein assay that provides an alternate means of visually and quantifiably assessing EWS/FLI binding to GGAA-microsatellites. Additionally, we have begun work on a CRISPR/Cas9-mediated technique to look at GGAA-microsatellite length-specific binding *in vivo.* Though challenging, progress in this area would provide tremendous insight into the actual cellular mechanism of EWS/FLI-mediated transcriptional regulation. Extensive work is still needed in each of these areas of investigation.

Notably, the diversity of techniques described in this chapter also demonstrates the unique value of seeking molecular understanding through a multiplicity of experimental approaches. Synergizing data obtained from IP, genome-editing, FP, and computational modeling will inform on the mechanism of EWS/FLI binding at length-dependent GGAA-microsatellites. Further advances in each of these areas will also contribute significantly not only to our knowledge of EWS/FLI in Ewing sarcoma, but also broadly to the mechanisms by which transcription factors interact with and regulate DNA.

Chapter 8: Advances in understanding the biophysical mechanism of homotypic EWS interactions on GGAA-microsatellites

Introduction

As previously discussed, the EWS portion of the oncogenic fusion is critical for both the activation and repressive functions of EWS/FLI [8]. The LC (low-complexity) domain of EWS is enriched in [G/S]Y[G/S] repeats, which have been shown to be important for high-density-induced polymerization of other EWS paralogs, like TLS/FUS[78]. These prion-like, N-terminal SYQG-rich domains are intrinsically aggregation prone sequences[76]. The field's current hypothesis is that polymerization of these IDRs (intrinsically disordered regions) precipitates formation of higher-order assemblies that in turn enable transcriptional activation. Some groups believe these triplet repeats are critical for transcriptional activation, some think for polymerization, and others hypothesize both[76,79].

Full-length EWS/FLI contains 12 of these [G/S]Y[G/S] repeats, our Mut9 construct has 5 repeats, and Δ22 has none (Table 8.1). It is believed the tyrosine within these repeats is the most important residue for facilitating the hypothesized biochemical interactions.

Again, the aforementioned RNA-binding protein FUS is a paralog of EWS. In a landmark paper in the field, Steve McKnight's group pioneered hydrogel assays to look at possible

FET proteins binding to the RNA-Polymerase II C-terminal domain (RNA-pol II CTD), which contains SYS triplet repeats that correspond with the [G/S]Y[G/S] triplet repeats of EWS[79]. Additionally, when 25-repeat GGAA-microsatellite DNA is incubated with the FUS LC-domain fused to the FLI-DNA binding domain, spontaneous fiber formation is visualized by TEM. This same synthesized fusion protein was incapable of elongated fiber formation in the absence of DNA. Though FUS/FLI is not an endogenous fusion, and despite the homology of FUS and EWS, these studies were not also conducted for EWS/FLI.

	Number of [G/S]Y[G/S] repeats	Number of Tyrosine (Y) residues	Transcriptional Activity?
EWS/FLI	12	37	Yes
Δ22	0	1	No
Mut9	5	15	Yes

Table 8.1 Number of triplet repeats for EWS/FLI deletion constructs

A distinct, though not mutually exclusive, hypothesis of FET protein aggregation upon DNA binding is phase separation[80]. Phase separation is essentially a phenomenon where membrane-less organelles form liquid-like droplets. This state enables continued entropy within the phase-separated state, while also ensuring maximum energy preservation and efficacy[231]. Such compartmentalization may be the mechanism of intracellular biological functions requiring membrane-less components. These membrane-less organelles are often RNA/protein-rich bodies roughly spherical in shape, which are viscous and respond

to liquid-properties like wetting, dripping, & flow when sheer stress is applied[232]. Spontaneous formation of these droplets occurs following molecular supersaturation due to high concentration, charge state, temperature or salt concentration[233].

Computational and *in vitro* recapitulation of this droplet formation have shown that interaction domains of biopolymers are in several cases sufficient to drive phase separation, suggesting that multivalent motifs are important for signaling network regulation and organization[234]. LC (low-complexity), intrinsically disordered sequences enriched in polar side chains (G, Q, N, & S), positive or negatively charged side chains, or aromatic side chains (F & Y) often drive these intracellular phase transitions. This has been observed for all three FET proteins: FUS, EWS, and TAF15[80]. Further, FUS requires high concentration, low temperature conditions to condense in solution to form amyloid-like fiber containing hydrogels, as in the McKnight polymerization experiments (discussed above)[78].

FUS has additionally been shown to coalesce into liquid-like droplets, both *in vitro* and *in vivo*[235]. This process is accelerated in disease states, such as ALS and other neurodegenerative diseases, where aggregation propensity seems to directly correlate with the likelihood of disease[236]. The low-complexity domain has been demonstrated to stabilize this phase separation[237]. Overall, it seems that conformational heterogeneity, degree of sequence complexity, and charge patterning interplay to allow phase separation of particular protein sequences, like FUS.

We hypothesize that the aggregate nature of EWS gives rise to homotypic interactions that result in polymerization on "sweet-spot" microsatellites, and that enable optimal biochemical configuration to promote transcriptional activity at these sites. To characterize the nature of potential EWS-EWS homotypic interactions in Ewing sarcoma, we have begun to investigate whether EWS/FLI can polymerize on GGAA-microsatellites, and whether EWS/FLI binding at GGAA-microsatellites results in phase separation. Evaluation of the biophysical properties that govern the interplay of multimeric EWS/FLI molecules on GGAA-microsatellites may clarify the mechanism of binding to these response elements.

Materials and Methods

Constructs and Retroviruses

Bacterial expression constructs included cDNA's for the LC (low-complexity) domain FUS/FLI and EWS in the pHis-parallel1-mCherry vector (a generous gift from Steven McKnight's laboratory at UT Southwestern), in addition to Δ22, Mut9, and EWS/FLI constructs ordered as a gene block (IDT) and cloned within the multiple cloning sites of the pHis-parallel1-mCherry vector between EcoRI and HindIII restriction sites.

Protein Purification

His/mCherry FUS/FLI, His/mCherry LC domain recombinant proteins were expressed in *E. Coli* BL21(DE3) competent cells from pHis-parallel1-mCherry expression plasmids encoding Δ22, Mut9, EWS/FLI, LC domain FUS/FLI, and LC EWS, respectively. Batch

purification was performed according to previously published protocols, with Ni-NTA (QIAGEN, USA) resin.[78,85] Briefly, a single colony picked from each protein respectively transformed in BL21 cells was inoculated into an overnight LB/Amp culture at 37°C. 6 x 1L LB/Amp cultures each were inoculated with 8ml of pre-culture the next morning, shaking at 37°C until the OD reached 0.6-0.8 (~3-4 hours). Cultures continued to shake at 20°C for 45min, whereupon 0.5 mL of 1M IPTG is added to each culture for overnight shaking. Cells were harvested and lysed the following morning in lysis buffer containing 0.4mg/ml lysozyme, 50 mM Tris-HCl pH7.5, 500 mM NaCl, 20 mM BME, 1% Triton X-100 and protein inhibitor cocktail (Roche, USA) for 30 min on ice, and then sonicated. The cell lysate was centrifuged at 35,000 RPM for 1 hr. The supernatant was mixed for 1hr at 4°C with Ni-NTA resin (Qiagen, USA). The resin-bound protein was packed in a glass column, washed, and eluted.[78] Purified proteins were concentrated in Amicon Ultra centrifugal filters (Millipore, USA), and stored at -80°C until ready for biochemical assays.

Hydrogel Formation

Hydrogel droplets of mCherry tagged proteins were prepared as described before.[78,79] Briefly, concentrated mCherry fusion proteins were dialyzed in gelation buffer containing 20mM Tris-HCL pH7.5, 200mM NaCl, 20mM BME, 0.5mM EDTA, 0.1mM PMSF overnight. Dialyzed protein solution was sonicated at low power 3-5 times for 3 seconds using a micro probe. Centrifugation was used to eliminate precipitates. The protein was pipetted in 0.5ul drops onto a glass-bottomed dish (MatTek, USA) in triplet for each

protein sample. Damp Whatman filter paper (Sigma-Aldrich, USA) strips were used to line the dish edges, and dishes were sealed with parafilm to prevent the hydrogel drying out. Hydrogels were incubated at room temperature for 2-4 days. After taking initial confocal images of the droplets, hydrogel stability was tested by slow pipetting of 100ul gelation buffer proximal to the hydrogel, and observing whether the mCherry-tagged droplet remained intact or diffused into solution.

Turbidity Assay

Single-stranded DNA oligos containing GGAA-microsatellites of various lengths were ordered (IDT) and annealed. Sequences are listed in Table 8.2. Recombinant protein preparation was performed as described previously (Chapter 4)[217]. Turbidity assays were performed by making 200ul solutions of 1uM recombinant protein and 0.5uM DNA oligo in 1x Gel Shift binding buffer (10mM Tris pH 8.0, 50mM NaCl, 1mM $MgCl_2$, 0.5mM EDTA, 4% glycerol). Δ22 and Mut9 were tested with 8-repeat, 16-repeat, and 24-repeat GGAA-microsatellite oligos, respectively, as well as no-DNA controls. 100ul of each protein-DNA solution were added in duplicate wells to a clear 96-well plate (Corning, USA). The plate was incubated at room temperature for 15 minutes. OD measurements were taken on the spectrophotometer and images were taken on the confocal microscope. For competition assays, various reagents were added to respective wells, according to experiment conditions (ie. increased salt concentration or addition of high-affinity DNA oligo).

Transmission Electron Microscopy (TEM)

Annealed GGAA-repeat oligos were described previously (Chapter 4, 7)[217], with sequences included in Supplementary Table 8.2. Briefly, recombinant protein of EWS/FLI constructs (EWS/FLI, Mut9, and Δ22, respectively) were purified, and used in combination with GGAA-repeat oligos for DNA visualization. Following production (Chapter 4)[217], purified protein was sonicated on low power 4x (1 second on/5 seconds off). DNA was added and incubated with protein in 1x Gel binding buffer (described above) at room temperature for 2 hours. Formvar carbon 300 mesh copper TEM grids (Electron Microscopy Sciences, USA) were prepared by adding 10ul of protein-DNA solution to a new grid, suspended by tweezers, for 2 minutes. Excess liquid was blotted dry with a Kim wipe, followed by 2x 30 second washes with 10ul of RNAse free water. The grid was stained with 10ul of 1% uranyl acetate for no more than 5 seconds before also blotting dry. Grids were allowed to air dry for 30 minutes and then stored in TEM grid cassettes until microscope viewing. TEM grids were visualized using the transmission electron microscope (Hitachi, Germany) at 30-80kV resolution. Ten fields were randomly selected and imaged for each grid sample. Polymer/fiber lengths for each sample were quantified using Image J software (NIH, USA), and graphed in Microsoft Excel.

Results

EWS polymerization on GGAA-microsatellites

To test whether EWS/FLI can polymerize on GGAA-microsatellites, we have begun to conduct hydrogel assays, similar to the McKnight group. Using mCherry-tagged recombinant proteins, we tested the ability of EWS/FLI and FUS/FLI to form hydrogels with and without GGAA-microsatellite DNA. FUS/FLI droplets appeared to contain a concentrated ring of mCherry surrounding the droplet, which became observably less structured with addition of microsatellite DNA. In contrast, EWS/FLI formed more diffuse mCherry droplets (Figure 8.1). The EWS/FLI hydrogels appeared tighter and more spherical upon addition of "sweet-spot" vs. shorter GGAA-repeat or high affinity site DNA, compared with protein only hydrogels (Figure 8.1). These preliminary observations suggest EWS/FLI binding to microsatellite DNA of "sweet-spot" length may enhance hydrogel formation.

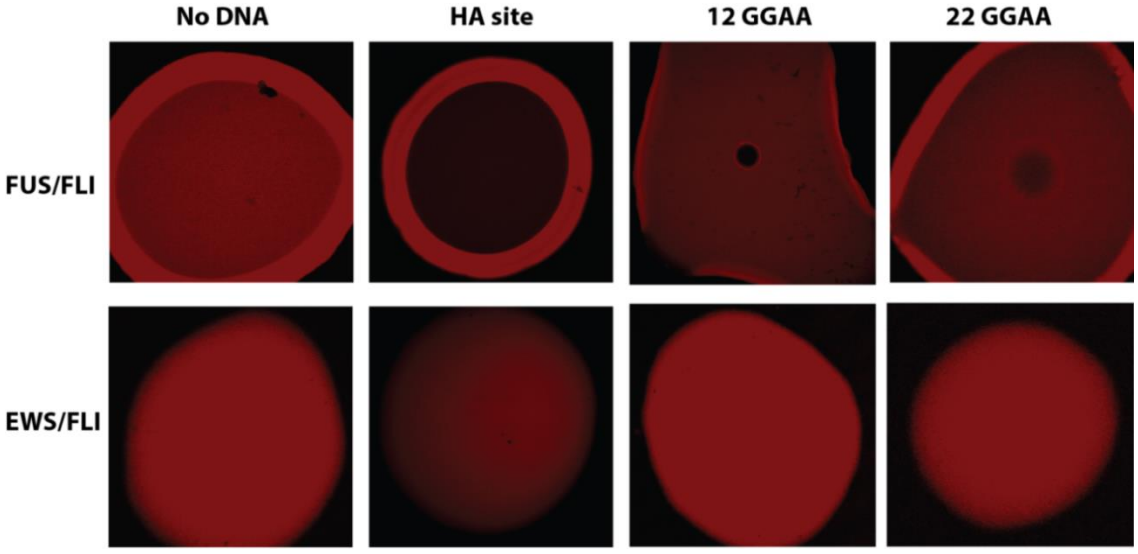


Figure 8.1 Hydrogel formation assays for EWS/FLI and FUS/FLI with and without DNA

To verify the stability of the observed hydrogel formation, buffer was added to pre-formed hydrogel droplets of protein-only samples. We then captured images pre and post addition of the liquid to measure stability, or conversely dispersion, of the hydrogels over time. Following addition of gelation buffer, the FUS/FLI droplets dispersed, suggesting instability in hydrogel formation (data not shown). Conversely, the FUS LC domain hydrogel remained intact for at least 5 minutes after addition of gelation buffer (Figure 8.2). This result recapitulates that observed by the McKnight group for this same construct, validating we have a working assay[78,79,85]. The EWS/FLI hydrogel also remained stable, though for just over 1 minute before diffusing into solution (Figure 8.2).

One explanation for this stability difference may be reflected in the respective protein concentrations achieved for this experiment. Recombinant FUS LC domain is much easier to purify than full-length EWS/FLI protein, and we were able to obtain much higher concentrations for the former. Based on our preliminary hydrogel observations with DNA (Figure 8.1), it may be that adding increasing GGAA-repeat microsatellites may improve EWS/FLI hydrogel stability.

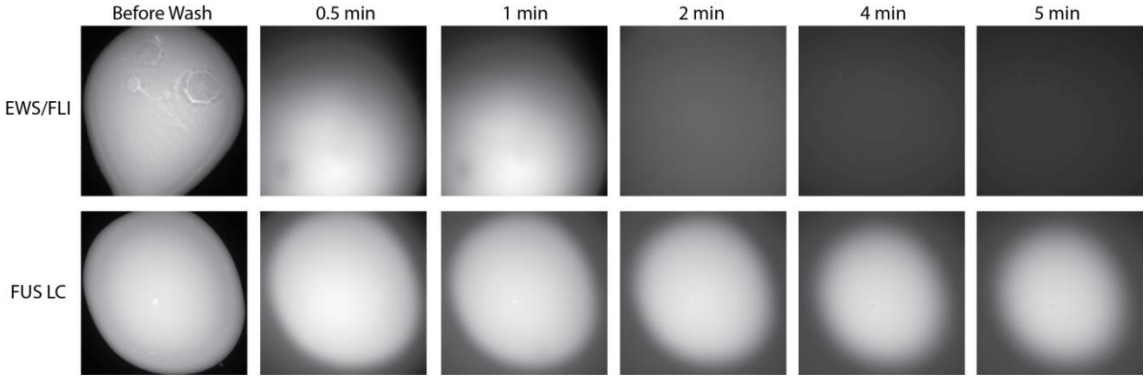


Figure 8.2 Hydrogel formation assay for EWS/FLI versus the FUS LC domain. Images demonstrate a time lapse before, and 0.5 to 5 minutes following addition of wash buffer to the hydrogel droplets.

Hydrogel formation assays assess whether a particular protein possesses the necessary biophysical properties to self-associate and form a stable multimeric structure. Though this hydrogel formation ability is suggestive of polymerization, substantiating this implication could be achieved through higher powered visual means, such as transmission electron microscopy (TEM). As mentioned previously, in the presence of sweet-spot length GGAA-repeat DNA, high concentrations of the FUS/FLI construct will spontaneously form fibers, visualized by TEM[79]. Given our preliminary hydrogel formation results, coupled with data from a number of other biochemical assays, we sought to determine whether the distinctions we previously observed for EWS/FLI constructs binding different lengths of GGAA-microsatellites is predictive of their respective abilities to form fibers *in vitro* on these same repeat lengths.

To test whether high concentrations of EWS/FLI recombinant protein spontaneously form fibers in the presence of sweet-spot GGAA-microsatellite DNA, we performed

TEM on EWS/FLI protein in the absence vs. presence of 22-repeat GGAA DNA oligos. We recognized the limitation that this assay would only be informative if we were able to achieve sufficiently high concentrations of recombinant EWS/FLI protein. Our initial trials demonstrated some evidence of disorganized aggregates for our EWS/FLI protein only samples (Figure 8.3), similar to that seen for FUS/FLI only samples by McKnight's group[79]. When 22 GGAA-repeat DNA was added to the EWS/FLI protein, however, we observed long, branching fibers at low magnification that appeared as more organized, fibrous clusters and strands with increasing magnification (Figure 8.3). This result was highly reproducible, permitting us to conclude that EWS/FLI displays spontaneous fiber formation in the presence of sweet-spot GGAA-microsatellite DNA *in vitro*.

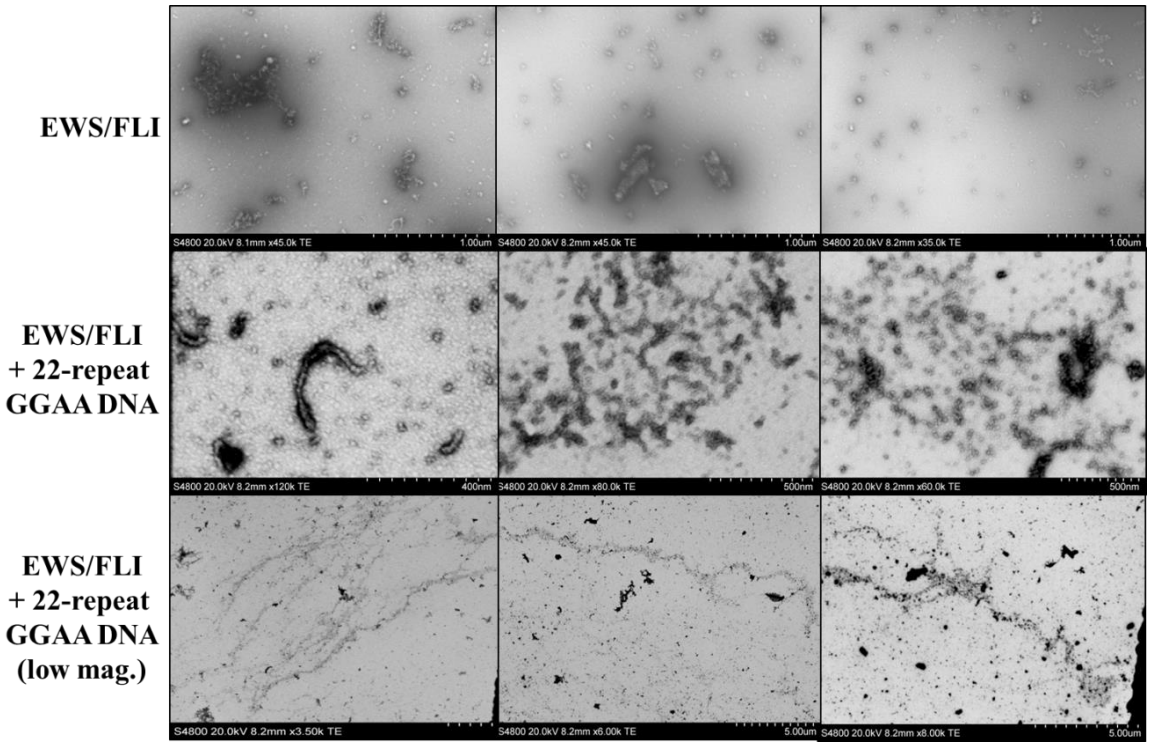


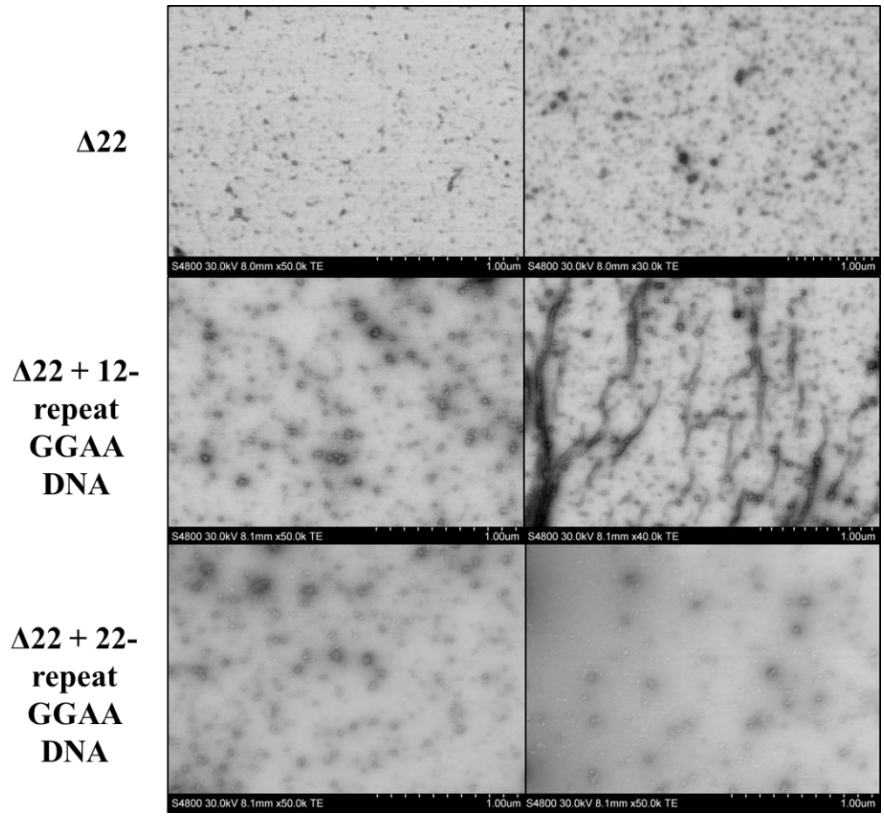
Figure 8.3 DNA-dependent enhancement of fiber formation of EWS/FLI

Given these results and our previous biochemical data, we considered utilizing TEM as a means of visually assessing the contribution of the EWS portion of the fusion to microsatellite binding (Chapter 4)[217]. Because our Δ22 construct does not contain EWS, and therefore the EWS LC domain, we expected little if any fiber formation for this EWS/FLI deletion construct. However, Δ22 is capable of binding with relatively high affinity to shorter GGAA-microsatellites (Chapter 4)[217]. This suggests some form of multimeric or cooperative Δ22 (FLI) binding at these microsatellites *in vitro*, that might manifest in some structurally visible effect. In contrast, Mut9 and EWS/FLI were predicted to demonstrate fiber formation that increases with increasing GGAA-repeat length.

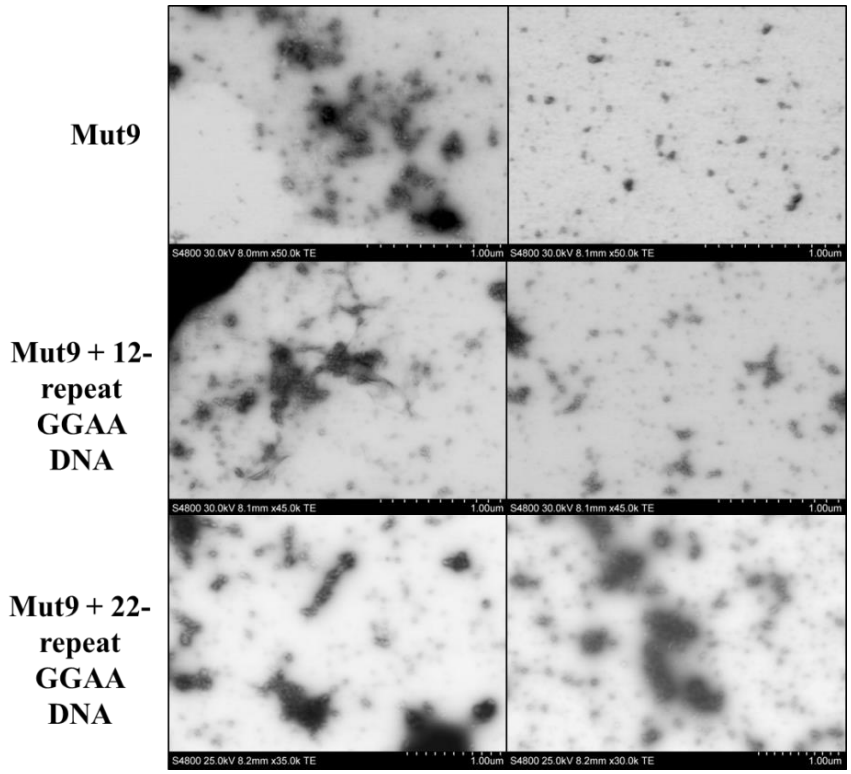
To test whether EWS/FLI, Mut9, and Δ22 spontaneously form fibers in the presence of GGAA-microsatellites in a length-dependent manner, we combined high concentrations of each of these recombinant protein constructs with no DNA, 12-repeat, or 22-repeat GGAA DNA, respectively. As expected, TEM of Δ22 only showed no fiber formation (Figure 8.4A). When added to 12 GGAA-repeat DNA, some stochastic nucleation was observed in most visual fields. Strikingly, a few fields contained branching fiber-like structures (Figure 8.4A). In contrast, when Δ22 was combined with 22 GGAA-repeat DNA, no long, definitive fibers were observed; most fields contained the short, dispersed nucleation structures observed in most visual fields for 12-repeats (Figure 8.4A). Though fiber formation was not expected for any of the Δ22 conditions, it is interesting to note the fibers observed with binding at 12-repeat microsatellites, which are absent at 22-repeats. This result is reminiscent of our prior fluorescence polarization data

demonstrating optimal Δ22 binding at 8-16 GGAA-repeats, which unexpectedly falls off in the sweet-spot range (Chapter 4)[217].

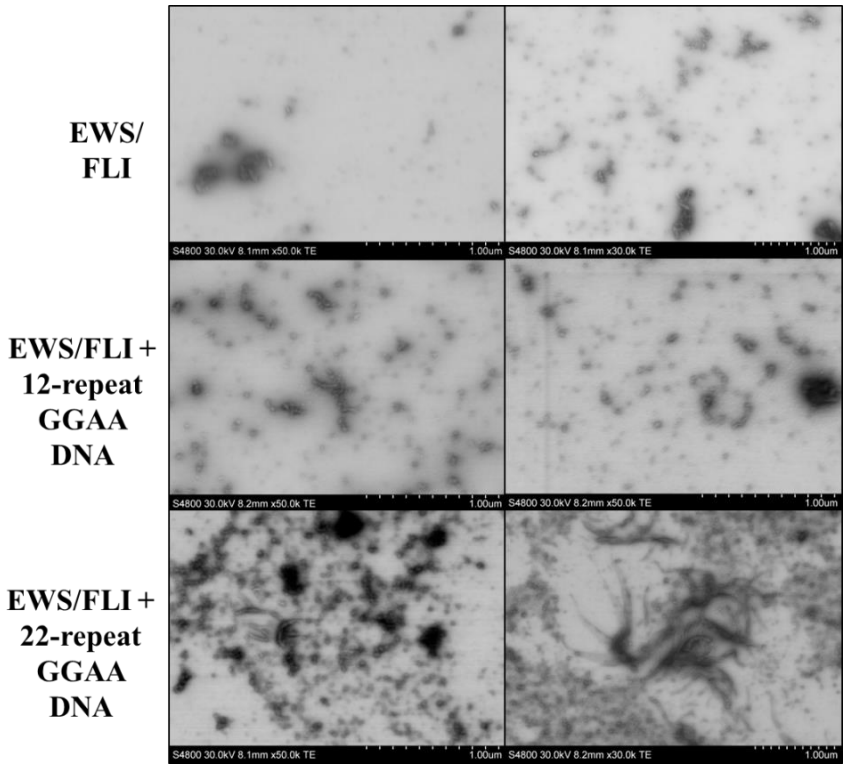
Figure 8.4 TEM images of fiber formation for **(A)** Δ22, **(B)** Mut9, and **(C)** EWS/FLI binding of GGAA-microsatellites



A


Mut9
Mut9 + 12-repeat GGAA DNA
Mut9 + 22-repeat GGAA DNA
S4800 30.0kV 8.0mm x50.0k TE
S4800 30.0kV 8.1mm x50.0k TE
S4800 30.0kV 8.1mm x45.0k TE
S4800 25.0kV 8.2mm x35.0k TE
S4800 25.0kV 8.2mm x30.0k TE
1.00um

B


EWS/ FLI
EWS/FLI + 12-repeat GGAA DNA
EWS/FLI + 22-repeat GGAA DNA
S4800 30.0kV 8.1mm x50.0k TE
S4800 30.0kV 8.1mm x30.0k TE
S4800 30.0kV 8.2mm x50.0k TE
S4800 30.0kV 8.1mm x50.0k TE
S4800 30.0kV 8.2mm x30.0k TE
1.00um

C

In contrast to our TEM results for Δ22, Mut9 protein exhibited aggregate clustering and minimal fiber formation in a few fields, with mostly unidentifiable background particulates (Figure 8.4B). These random clusters were not overly surprising as the intrinsically disordered region of Mut9 substantially increases the potential for this protein to aggregate, compared with the mostly FLI-only composition of Δ22. With addition of DNA, however, Mut9 spontaneous nucleation and fiber formation increased with increasing length of GGAA-microsatellite (Figure 8.4B). Further, Mut9 binding to 22 GGAA-repeat DNA displayed concentrated clusters of intertwining fibers, though none as robust as those fibers observed for full-length EWS/FLI (Figure 8.3, 8.4C).

Although Mut9 is a fully functional EWS/FLI construct *in vivo*[34,217], the missing low complexity domain components of Mut9 versus full-length EWS/FLI may give rise to differences in their biophysical properties. Nonetheless, TEM images observed for EWS/FLI displayed the same overall trends as for Mut9, notwithstanding markedly enhanced fiber formation for EWS/FLI binding of 12 and 22 GGAA-repeats (Figure 8.4C). Quantification of fiber length showed an overall increase with increasing GGAA-repeats for EWS/FLI and Mut9, while Δ22 fiber length decreased at the sweet-spot length (Figure 8.5). These trends support our prior biochemical data comparing binding of these EWS/FLI constructs on increasing GGAA-repeat DNA (Chapter 4, 7)[217].

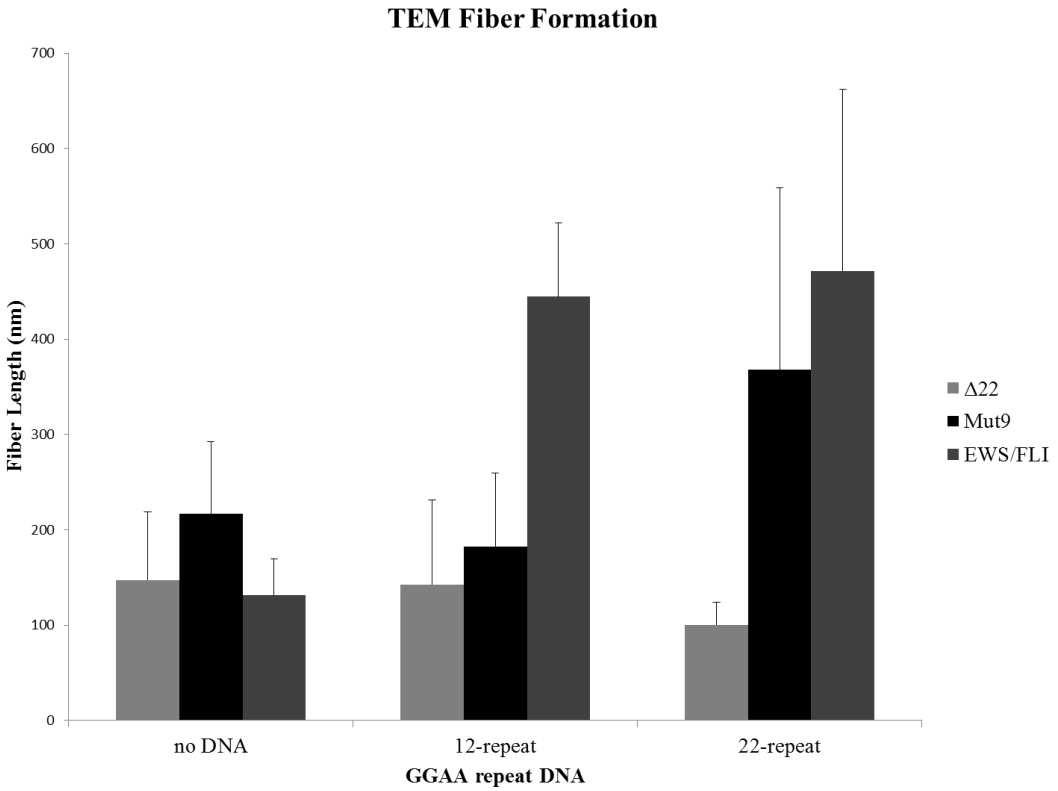


Figure 8.5 Quantification of fiber length for EWS/FLI, Mut9, and Δ22 binding of no DNA, 12-repeat, and 24-repeat GGAA DNA, respectively

Phase separation properties of EWS/FLI

Phase separation is a recently popular area of study, describing a liquid-like biophysical property that enables isolated biochemical reactions to occur via compartmentalization at liquid-liquid interphases[231–234]. These membrane-less organelles are dynamic structures, and are generally formed via concentration-dependent mechanisms[233]. Intrinsically disordered proteins (IDPs), like EWS, tend to have LC domains, contributing structural flexibility and inherent dynamism through constant rearrangements of multivalent weak interactions[80]. Our stoichiometric data provides evidence for multimeric binding of EWS/FLI at GGAA-microsatellites (Chapter 4)[217]. It reasonably follows that the

subsequent high density of EWS/FLI molecules brought into proximity with one another through this microsatellite binding may induce a phase separated state.

To determine whether EWS/FLI binding at GGAA-microsatellites results in phase separation, we conducted turbidity assays on recombinant protein incubated with various DNA oligos. Phase separation was undetectable for protein only samples, in the presence of DNA (Figure 8.6A). As GGAA-microsatellite DNA of increasing repeat number was added to recombinant Δ22, however, increasing degrees of phase separation were observed (Figure 8.6A). Consistent with previous FP studies conducted in this thesis, Δ22 binding appeared to "fall apart" at 24 GGAA-repeats, or "sweet-spot" lengths, as indicated by an inability to phase separate in one trial, and a reduction of phase separation in a second. Scramble-sequence controls (non-GGAA microsatellite repetitive sequence) of the same lengths, as well as the high affinity Ets site, also displayed no phase separation. This suggests the possibility that FLI, and perhaps EWS/FLI binding at microsatellites may undergo phase separation, however, additional trials and further characterization is required.

To test whether the observed phase separation likely caused by Δ22 binding the DNA could be disrupted, high salt was added to Δ22 binding at 24 GGAA-repeats. We observed dispersion of the phase separated "bubbles" under these conditions, however, no reduction in the number observed (Figure 8.6B). Addition of high affinity site DNA to compete for binding displayed a similar effect as with increasing the salt concentration

(Figure 8.6C). Most of these initial experiments were conducted using Δ22 recombinant protein; however, we also wanted to test our Mut9 construct.

Figure 8.6 Turbidity assays to evaluate whether recombinant Δ22 and Mut9 incubated with increasing lengths of GGAA-microsatellite DNA exhibits phase separation *in vitro*

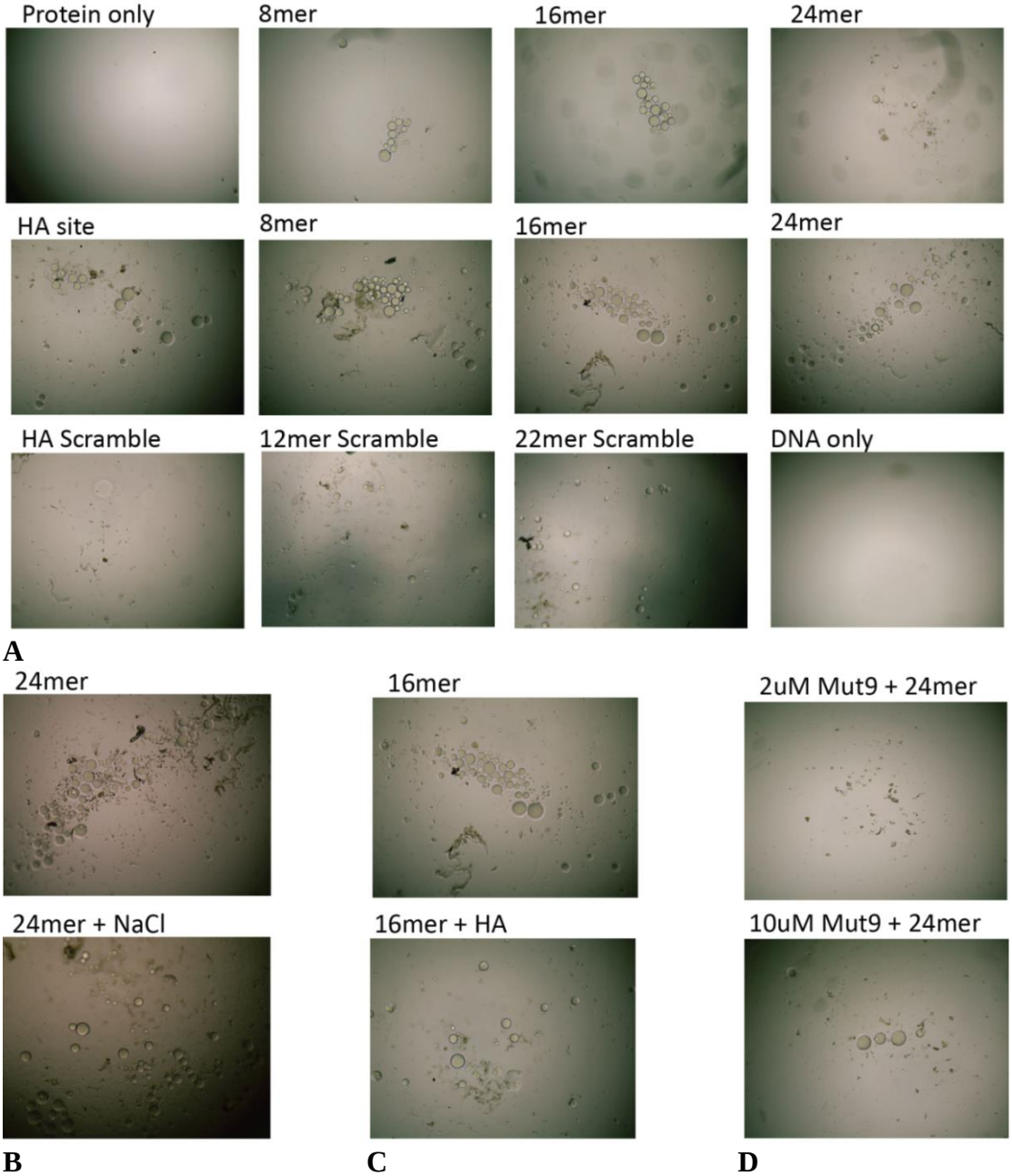
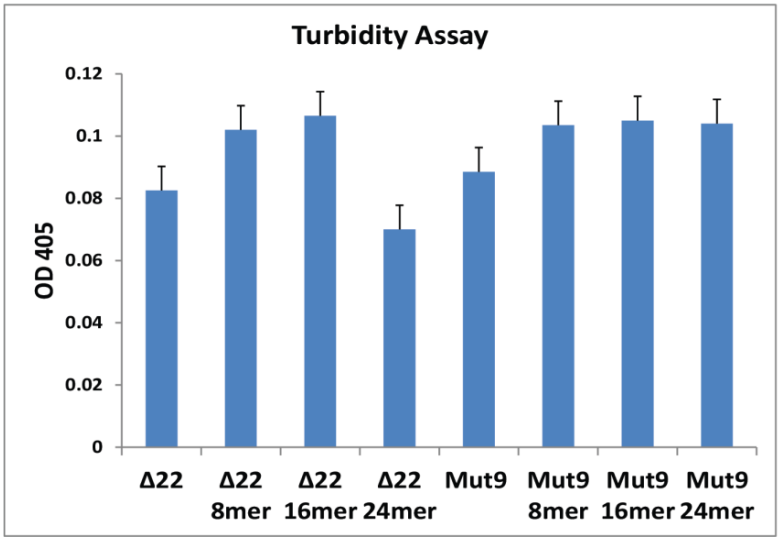


Figure 8.6 continued



E

(A) Δ22 incubated with increasing lengths of GGAA-microsatellite DNA (except where indicated for the protein only and DNA only) **(B)** Addition of high salt concentration to phase-separated Δ22 binding of 24-repeat DNA **(C)** Addition of Ets high-affinity site DNA to phase-separated Δ22 binding of 16-repeat DNA **(D)** Increased Mut9 concentration for Mut9 binding of 24-repeat DNA **(E)** OD_{405} turbidity assay measurements for Δ22 and Mut9 binding increasing lengths of GGAA-repeat DNA

Our initial trials with Mut9 yielded little to know phase separation formation. Notably, however, our starting yields of recombinant Mut9 protein are not particularly high. From our Mut9 trials in previous experiments discussed in this work (Chapter 4)[217], we considered that Mut9 may not be as pure or fully functional as our Δ22 protein preps, thereby requiring "higher" concentrations than the latter to test the same amount of actual purified, functional protein (Chapter 7). To test whether higher concentrations of Mut9 display phase separation, we added 5-fold higher Mut9 protein to 24 GGAA-repeat DNA. As anticipated, we observed some phase separation at the higher Mut9 concentration

(Figure 8.6D). This emphasizes the effect of protein concentration on the ability of EWS/FLI binding at microsatellites to undergo phase separation. Optical density quantification of these turbidity assays demonstrated increased density with increasing GGAA-repeat number for Δ22, but with a decrease at 24 repeats (Figure 8.6E). No change in optical density, however, was observed for Mut9 binding to increased lengths of GGAA-microsatellites. This uninformative result for Mut9 may be clarified by increased overall concentrations of Mut9 protein.

Discussion

Taken together, these data demonstrate promising evidence to support our hypothesis that EWS/FLI participates in homotypic interactions via the EWS portion to facilitate "sweet-spot" GGAA-microsatellite binding. We have shown success in our ability to purify recombinant full-length mCherry-tagged EWS/FLI for our hydrogel binding assays. Additionally, our TEM data demonstrates reproducible evidence for EWS/FLI fiber formation that increases in length with increasing number of GGAA-microsatellite repeats. Further, these results bolster our previous data showing the necessity of the EWS domain for EWS/FLI binding to GGAA-microsatellites (Chapter 4)[217]. Our TEM imaging presents, to our knowledge, the first attempt to assess EWS/FLI fiber formation, and offers visual evidence for EWS/FLI polymerization as the mechanism of GGAA-microsatellite binding.

Overall, we submit a model that EWS-EWS interactions facilitate EWS/FLI polymerization on GGAA-microsatellites, bending the DNA in such a conformation as to allow EWS-interaction with transcriptional machinery (such as RNA polymerase II). This optimal structural configuration and subsequent recruitment of transcription-associated factors precipitates transcriptional activation of the genes associated with these EWS/FLI bound microsatellites. A schema of this model is shown in Figure 8.7.

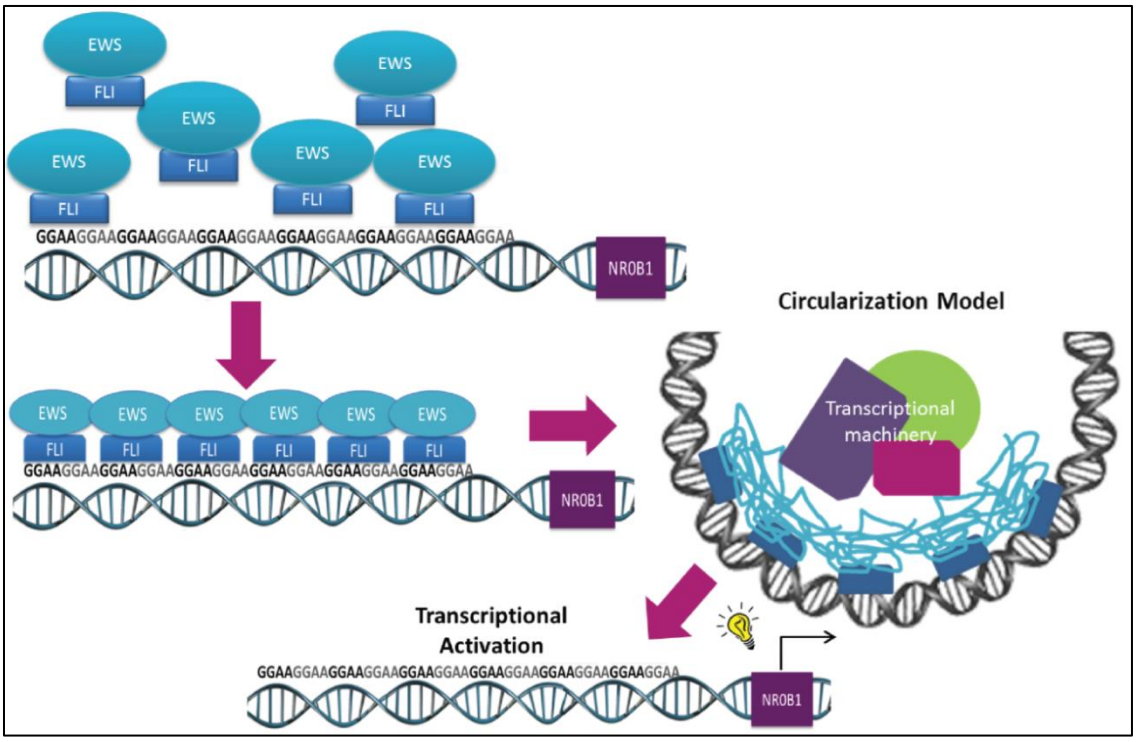


Figure 8.7 Schema of Cooperative model for EWS/FLI binding "sweet-spot" GGAA-microsatellites, allowing recruitment of transcriptional machinery for activation of its targets

The ensuing high density of EWS/FLI molecules interacting with GGAA-microsatellites corroborates the phase separation hypothesis. Our attempts at examining phase

separation, although preliminary, have already demonstrated visually discernible differences in length-dependent GGAA-microsatellites for our Δ22 vs. Mut9 recombinant proteins. This model is additionally strengthened by recent evidence proving EWS/FLI recruitment of the BAF chromatin remodeling complex to GGAA-microsatellites[47]. This complex is known to facilitate gene accessibility for transcriptional machinery.

Future directions for this project include testing whether [G/S]Y[G/S] repeats are necessary and sufficient for transcriptional activation, and for the ability of EWS/FLI to aggregate or polymerize. Understanding the mechanism of EWS/FLI interactions with DNA to facilitate binding and optimal effector function at "sweet-spot" microsatellites is critical. Such knowledge may elucidate means to disrupt these protein-protein and protein-DNA interactions, propounding promising therapeutic potential.

Supplementary Information

DNA Oligo	Sequence (Forward Strand of Duplex)
High Affinity	TT TAC CGG AAG TGT TT
8 Repeats	GGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAA
16 Repeats	GGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAA GGAAGGAAGGAA
24 Repeats	GGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAA GGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAA
Flr12 Repeats	56-FAM/TTGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAA GGAA
Flr22 Repeats	56-FAM/TTGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAA GGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAA

Table 8.2 Sequences for DNA oligos used in TEM and in turbidity assays

Chapter 9: Conclusion

Discussion

Ewing sarcoma is a pediatric bone malignancy initiated by a t(11;22) chromosomal translocation that produces the EWS/FLI oncoprotein. An aberrant transcription factor, EWS/FLI binds and transcriptionally regulates its target genes to mediate oncogenic reprogramming. GGAA-microsatellites are now well-established as heritable polymorphic EWS/FLI response elements unique to and critical for Ewing sarcoma oncogenesis. Studies over the last few years within and concurrent to this thesis have demonstrated a role for length-dependent GGAA-microsatellites in Ewing sarcoma as not only DNA binding sites, but also as regions of chromatin regulation and remodeling, enhancers to enable EWS/FLI-mediated activation, markers of Ewing sarcoma susceptibility, and as a scaffolding sequence for aggregation-induced phase separation to facilitate recruitment of transcriptional regulatory machinery. Here, each of these roles for GGAA-microsatellites is discussed. Additionally, we posit future directions requisite to further advance the field.

GGAA-microsatellites as EWS/FLI DNA binding sites

Our review article in this thesis (Chapter 2) provides the first reference in the literature to a "sweet-spot" GGAA configuration conducive to maximal EWS/FLI-mediated gene up-regulation[33]. The "sweet-spot" (20-26 repeats) microsatellite facilitates maximal

EWS/FLI-mediated gene expression, evidenced upon cloning polymorphic GGAA sequences into a luciferase reporter vector in the absence versus presence of EWS/FLI (Chapter 2)[16]. We saw the same bimodal pattern of maximal gene expression as in our clinical observation, with peak amplification in constructs from 20-25 and 60-65 GGAA-motifs. We also looked at *NR0B1* gene expression in 21 hemi- or homozygous patient Ewing tumors, and found a significant difference in gene expression for those tumors with 17-18 as opposed to 23-26 GGAA-motifs. A similar result was observed in Ewing sarcoma cell lines with various *NR0B1*-associated GGAA-repeat numbers[16].

Given the epidemiological differences previously demonstrated between *NR0B1* GGAA-microsatellite lengths in African vs. Caucasian populations[14], we sought to compare these control populations to individuals with Ewing sarcoma (Chapter 3). Data consisted of 112 tumor specimens (90% Caucasian vs. 2% African). Though the *NR0B1* microsatellite is also highly polymorphic among Ewing tumors, the length range for tumor patients represented a significantly more narrow range from that observed in white European controls. This held true for both primary and metastatic tumor samples[16].

Considering reasons for the "sweet-spot" length led us to evaluate the possibility of microsatellite instability (MSI) in GGAA-microsatellite polymorphisms. We decided this is unlikely, due to the large African microsatellites that appear stable and in non-disease individuals vs. the observed 20-26 GGAA-repeat lengths associated with Ewing sarcoma patients[33]. If there were evidence for MSI in Ewing sarcoma, we argue that the ~70x

GGAA-repeats found in both African and European individuals would be much more intrinsically unstable and linked to disease susceptibility or clinical outcome than the 20-26 repeat patients, as seen in a number of neurodegenerative diseases. Additionally, we found GGAA-repeats in patients are ubiquitous in length in both germline and tumor cell microsatellites[33].

Plausible explanations for the "sweet-spot" length then, include optimal stoichiometric occupancy of EWS/FLI at these lengths. Alternatively, certain GGAA polymorphisms might be more (or less) likely to form inhibitory secondary DNA structures. These include non-B-form DNA structures such as G-quadraplexes (G-rich DNA sequences) or triplex DNA. Alternatively, EWS/FLI-mediated non-canonical DNA structure formation could enable transcriptional regulation by bio-mechanical means. Evaluation of these possibilities will require in-depth biophysics methodology and possibly visual characterization via atomic force microscopy. Additionally, a solved NMR structure of EWS/FLI bound to a number of GGAA-repeats would enhance our understanding of why EWS/FLI seems to preferentially bind "sweet-spot" length GGAA-repeat DNA.

EWS/FLI acts as a pioneer factor at GGAA-microsatellites

Riggi et al. demonstrated that EWS/FLI-bound GGAA-microsatellites in Ewing sarcoma are in an accessible open chromatin state. EWS/FLI knockdown results in a closed chromatin state at these loci, while EWS/FLI introduction into mesenchymal stem cells (MSCs) converts closed chromatin into an open state[69]. Interestingly, Ewing sarcoma

cells are the only cell type with open chromatin at EWS/FLI-bound GGAA-microsatellites, and EWS/FLI actively depletes nucleosome occupancy of these regions[238]. This and other data suggest that EWS/FLI may function as a pioneer factor at GGAA-microsatellites to open chromatin and enable transcriptional activation at these response elements. However, it was not known whether FLI, which contributes the binding function of this chimeric fusion, is sufficient to open chromatin, or whether the EWS portion also plays a key role.

In Chapter 4 we examined the biochemical characteristics governing the binding of different aspects of the EWS/FLI fusion to various lengths of GGAA-microsatellites (Chapter 4)[217]. These studies began with the FLI-only portion of the fusion, and demonstrated a stoichiometry of one FLI molecule bound for every two GGAA-motifs in a microsatellite. As discussed in Chapter 1, studies foundational to this work suggested a cooperative binding model, with homodimeric interactions between the FLI domains of multiple EWS/FLI molecules, resting on the DNA in a Lego-like linking manner [51]. We expected to measure increasing binding affinity with increasing GGAA-repeat length, in accordance with this cooperativity. Contrary to this original model, however, binding affinity was identical for increasing repeats, suggesting independent FLI binding.

Moreover, we observed an even more unexpected phenomenon when testing "sweet-spot" microsatellite lengths. The FLI-only portion deletion construct (Δ22) failed to bind at "sweet-spot" lengths, while our EWS/FLI mutant with a sufficient portion of EWS to

rescue oncogenic transformation (Mut9), improved binding significantly. This suggested the EWS portion of the fusion is critical for EWS/FLI binding at "sweet-spot" lengths and prompted further studies, including RNA-seq and ChIP-seq using the same mutant constructs. These genome-wide binding and expression analyses lent further support to our biochemical findings. Overall, this study provides the first evidence for the necessity of the EWS portion of the EWS/FLI fusion for both *in vitro* binding of GGAA-microsatellites and for EWS/FLI binding targets genome-wide in Ewing sarcoma cells[218].

This data also reinforces the *Riggi et al.* study suggesting EWS/FLI may act as a pioneer factor[69]. As our work demonstrates an unexpected role for EWS in interacting with DNA, it may be that EWS interacts with the chromatin at GGAA-microsatellite regions to open the DNA at these sites. Reciprocal evidence for this model was concurrently published with our study (Chapter 4) by the Rivera group, demonstrating the physical properties of the EWS low-complexity (LC) domain are necessary for chromatin remodeling, complex retargeting, and EWS/FLI-mediated activation at these microsatellites[47]. Specifically, conserved [G/S]Y[G/S] repeat sequences within the LC domain contribute to the biophysical properties that enable multi-dimeric EWS/FLI binding at these sites (see below for further discussion).

This proven role for EWS in interacting with the DNA and facilitating critical EWS/FLI functions may prove a therapeutically targetable process. Conceivably, disruption of

EWS's ability to aggregate and access the DNA would inhibit the downstream chain of events mediating EWS/FLI's role as an oncogenic driver.

GGAA-microsatellites are EWS/FLI activating response elements

Multiple studies, including the ones discussed previously[16,217], provide convincing evidence of the association of GGAA-microsatellites with EWS/FLI up-regulated targets[15,46,51]. Additionally, we show these GGAA-motifs are sufficient for EWS/FLI-mediated activation in transcriptional reporter assays (Chapter 3)[16]. Despite mounting confirmation, no one had definitively proven the necessity of GGAA-microsatellites as EWS/FLI activating response elements in bona fide Ewing sarcoma cells prior to this work. We sought to accomplish this via the CRISPR/Cas9 system, using the *NR0B1* gene as an ideal model of EWS/FLI mediated activation[43]. We show for the first time, the direct need for these GGAA-microsatellites in EWS/FLI binding and subsequent activation of its associated up-regulated targets in Ewing sarcoma cells[218]. This was contemporaneously proven by others in the field through demonstration of EWS/FLI-mediated recruitment of the BAF chromatin remodeling complex to GGAA-microsatellites to transcriptionally activate EWS/FLI targets[47].

Having established GGAA-microsatellites as EWS/FLI-specific enhancer response elements, we sought to define these unique repetitive regions in a Ewing sarcoma context. Using FLI ChIP-seq and RNA-seq data of Ewing sarcoma cell lines, we utilized a computational approach to accomplish two major objectives. First, we performed an

unbiased genome-wide screen of GGAA-microsatellites to characterize and define these in an EWS/FLI-regulatory context. We then used our established definition and asked whether particular characteristics of these microsatellites, such as consecutive motif number, total length, or gene proximity, are predictive of EWS/FLI responsiveness at specific genomic loci. This turned out to be intricately complex and does not take into account a variety of factors, including other proteins and genomic interactions EWS/FLI may encounter *in vivo*. However, we successfully instituted a working definition of a GGAA-microsatellite. Further, we showed evidence of a minimal correlation for both binding and transcriptional regulation of EWS/FLI at "promoter-like" microsatellites for activated targets (Chapter 5)[223].

We also identified a category of GGAA-microsatellites dubbed "enhancer-like" microsatellites, characterized by GGAA-repeat regions located further than 5kb from the nearest gene (Chapter 5)[223]. Interestingly, there was no correlation between EWS/FLI activated targets and microsatellite length, though there was a correlation with EWS/FLI binding at these distant enhancer regions. This finding exposes the complexity of studying long-range transcription factor binding and regulation, and suggests additional factors (i.e. other transcription factors, protein complexes, etc.) are likely involved.

Epigenetic examination of similar long-range interactions led to proposal of a "super-enhancer" model in the field a of couple years ago, with complex regulatory mechanisms involving chromatin remodeling and cooperative epigenetic regulators[70]. For example,

chromatin conformation capture (3C) demonstrated long-distance physical interaction via DNA looping between a particular EWS/FLI-bound GGAA-microsatellite and the distant *NKX2.2* promoter[69]. Structurally, these homo-purine elements of a particular sequence length may offer the optimal function for EWS/FLI binding, potential DNA-looping for super-enhancer function, and even sequence inhibition near EWS/FLI-repressed targets[219,234]. Further clarification of this model could provide helpful insight not only for EWS/FLI regulatory mechanisms, but also broader transcriptional biological means of long-range gene regulation.

Our bioinformatics study especially highlights the significant value of combining biophysical, computational, and molecular investigation (Chapter 5-6). Many recent studies have produced large volumes of genome-wide datasets using Ewing sarcoma cell lines with genetic manipulation of EWS/FLI, as well as treatment with numerous epigenetic and novel targeted inhibitors[47,69,70]. Though computational approaches can't mechanistically elucidate every aspect of EWS/FLI's driving role in Ewing sarcoma, such rich and readily available datasets should be exploited for patterns, predictive models, and to generate previously unconsidered queries in Ewing sarcoma research. Concomitant molecular studies will continue to validate these computational models and to advance the field at an ever-increasing rate.

One such area that could be computationally exploited using existing datasets is the question of whether EWS/FLI exhibits allelic specificity in GGAA-microsatellite

binding[239]. In light of EWS/FLI *in vitro* binding preference for "sweet-spot' length microsatellites, EWS/FLI could conceivably exhibit allelic preference in binding heterozygous alleles, when one allele-containing microsatellite more closely resembles the "sweet-spot" length. Our preliminary bioinformatics analyses of data from the A673 Ewing sarcoma cell line are agreement with this hypothesis (Chapter 6). We have clearly shown multiple examples of EWS/FLI allele-specific binding, with one allele much more highly expressed than the other (Chapter 6). Complexities of this approach include a reliable Ewing sarcoma reference genome (currently in process of sequencing), in addition to the need to identify SNPs (single nucleotide polymorphism) associated with differentially expressed genes that are near enough to heterozygous GGAA-microsatellite alleles to enable linkages between our gene expression data. Such established connections are necessary to identify which allele contains the EWS/FLI bound GGAA-microsatellite associated with the observed differential gene expression.

GGAA-microsatellites as markers of Ewing sarcoma susceptibility

As GGAA-microsatellites appear to be uniquely regulated by EWS/FLI in Ewing sarcoma, it is conceivable that these polymorphic genetic elements could serve as clinical markers. However, our preliminary patient analysis to determine whether microsatellite length correlates with patient survival demonstrated that polymorphisms of the *NR0B1* GGAA-microsatellite are not predictive of event-free survival (Chapter 3)[16]. As an alternative, we postulated that GGAA-microsatellites might be more suited as diagnostic rather than prognostic markers of disease. While having a sweet-spot length GGAA-

microsatellite at the *NR0B1* locus doesn't imply an individual will develop Ewing sarcoma, patients with Ewing sarcoma tend to have the sweet-spot length. Therefore, individuals with sweet-spot GGAA-microsatellites have a heightened preponderance for developing Ewing sarcoma.

Grunewald et al. later identified a SNP within the *EGR2*-associated GGAA-microsatellite that links two adjacent runs of GGAA-repeats to create a risk allele found significantly more often in Ewing sarcoma patients[32]. This susceptibility locus supports our clinical observation of a length-dependent GGAA-microsatellite association with risk of disease development (Chapter 3)[16]. Also in accordance with our clinical findings and observed Ewing sarcoma epidemiology, sequence analysis using the 1000 Genomes Project revealed this *EGR2* risk allele is found at a significantly higher frequency in non-African populations[32]. As GGAA-microsatellites are response elements unique to Ewing sarcoma, these findings suggest identification of a disease-specific risk allele.

A scaffolding model for GGAA-microsatellites in enabling EWS/FLI-mediated transcriptional activation

The LC domain of EWS is enriched in [G/S]Y[G/S] repeats, which have been shown to be important for high-density-induced polymerization in other EWS paralogs, like FUS and TAF-15 (collectively known as FET proteins). Prion-like N-terminal SYQG-rich domains are intrinsically aggregation prone sequences. In a particularly landmark study, the McKnight group pioneered hydrogel assays to look at possible FET proteins binding

to the RNA-pol II CTD, which contains SYS triplet repeats that correspond with the [G/S]Y[G/S] triplet repeats previously mentioned present in EWS[78,85]. In a follow-up study, they visually demonstrated spontaneous fiber formation when 25-repeat ("sweet-spot") GGAA-microsatellite DNA is incubated with the FUS LC-domain fused to the FLI-DNA binding domain[79]. In the absence of microsatellite DNA, this fiber formation was significantly reduced. We have recently repeated these studies with different lengths of GGAA-microsatellite DNA and observed similar results (Chapter 8).

The field's current hypothesis is that polymerization of these intrinsically disordered regions (IDRs) enables formation of higher-order assemblies that spontaneously form spherical structures, via liquid de-mixing, at high concentrations[80]. This phase separation then enables the association of the necessary transcriptional machinery requisite for gene activation at a given promoter region[234]. EWS/FLI was recently shown to exhibit phase separation properties through sedimentation experiments demonstrating EWS/FLI, but not wild type FLI precipitates spontaneously at sufficient concentrations[47]. Moreover, a punctate vs. diffuse pattern staining of protein was observed by confocal imaging following GFP-tagged EWS/FLI vs. wild type FLI lentiviral expression, respectively, in mesenchymal stem cells (MSCs).

One explanation for the observed polymerization properties of FET proteins is the seed-model, where multiple FET proteins cooperatively bind along a particular sequence of DNA. This protein-nucleic acid complex then forms a "seed," capable of organizing non-

nucleic acid bound proteins (i.e. additional FET or other related proteins) into fibers capable of binding transcriptional machinery, such as the C-terminal domain (CTD) of RNA polymerase II[79,93]. This hypothesis is especially intriguing in light of recent evidence that EWS/FLI recruits the BAF chromatin remodeling complex to GGAA-microsatellites to activate associated target genes[47]. Both the aforementioned conserved tyrosine residues of the EWS LC-domain, and the domain's phase-transition capabilities, are necessary for this recruitment and activation process.

We have also explored the requirement for the conserved tyrosine residues found within the aforementioned [G/S]Y[G/S] triplet repeats for polymerization, through creating a series of tyrosine mutants. Our mutants include the regions mutated by the Rivera group, and similarly demonstrate the need for particular tyrosine residues[47]. Interestingly though, the Rivera group described some evidence of possible functional redundancy within the highly repetitive EWS LC-domain.

The length-dependent nature of GGAA-microsatellites suggests both inherent structural and functional mechanisms by which these repetitive elements may play an essential role in Ewing sarcoma. Given the aggregative propensity of EWS/FLI attributable to specific amino acid repeats in the EWS LC-domain, multimeric EWS/FLI binding at GGAA-microsatellites likely facilitates phase separation. Subsequent compartmentalization of EWS/FLI proteins, capable of inducing chromatin accessibility and recruiting specific activation complexes such as BAF, foments a conceivable model that EWS/FLI is a

master organizer of a veritable transcription factory (Figure 8.7). In providing conditions conducible to polymerization, "sweet-spot" GGAA-microsatellites may serve as a scaffold for EWS/FLI to bind and allow recruitment of critical transcriptional machinery, such as RNA polymerase II. Such a model implicates GGAA-microsatellites as fundamentally requisite for recruitment and initiation of EWS/FLI-mediated transcriptional activation in Ewing sarcoma.

Conclusion

The overall objective of this research was to investigate the biochemical properties that dictate how EWS/FLI regulates activation of its targets. This thesis has successfully elucidated the mechanism by which EWS/FLI transcriptionally activates direct target genes that mediate Ewing sarcoma oncogenesis 1) through *in vivo* demonstration of microsatellites as bona fide DNA response elements for EWS/FLI –mediated gene activation, 2) by biochemical and reporter assay demonstration of a particular GGAA-microsatellite length (the "sweet-spot"), as critical for maximal EWS/FLI responsiveness, and 3) via bioinformatics characterization of these GGAA-microsatellites showing computational predictability of EWS/FLI binding and transcriptional regulation based on distinguishing microsatellite features.

It is hoped that the work herein provides insight that will enable better understanding of Ewing sarcoma biology. Such mechanistic understanding of EWS/FLI-mediated regulation of this disease is critical to uncovering therapeutic means of targeting this

pathognomonic oncogenic driver. Furthermore, these insights will expand our knowledge of pediatric sarcoma biology more broadly, hopefully inciting new advances in other translocation or microsatellite-driven malignancies.

References

1. Balamuth NJ, Womer RB. Ewing's sarcoma. [Review] [75 refs]. *Lancet Oncol*. 2010;11(2):184-192. doi:http://dx.doi.org/10.1016/S1470-2045(09)70286-4

2. Llombart-Bosch A, Carda C, Peydro-Olaya A, et al. Soft tissue Ewing's sarcoma: Characterization in established cultures and xenografts with evidence of a neuroectodermic phenotype. *Cancer*. 1990;66(12):2589-2601. doi:10.1002/1097-0142(19901215)66:12<2589::AID-CNCR2820661223>3.0.CO;2-7

3. Delattre O, Zucman J, Plougastel B, et al. Gene fusion with an ETS DNA-binding domain caused by chromosome translocation in human tumours. *Nature*. 1992;359(6391):162-165. doi:10.1038/359162a0

4. Aurias A, Rimbaut C, Buffe D, Zucker JM, Mazabraud A. Translocation involving chromosome 22 in Ewing's Sarcoma. A cytogenetic study of four fresh tumors. *Cancer Genet Cytogenet*. 1984;12(1):21-25. doi:10.1016/0165-4608(84)90003-7

5. Delattre O, Zucman J, Melot T, et al. The Ewing family of tumors--a subgroup of small-round-cell tumors defined by specific chimeric transcripts. *N Engl J Med*. 1994;331(5):294-299. doi:10.1056/NEJM199408043310503

6. Ohno T, Rao VN, Shyam E, Reddy P. EWS/Fli-1 Chimeric Protein Is a Transcriptional Activator. *Cancer Res*. 1993;53(24):5859-5863.

7. Kovar H, Aryee DN, Jug G, et al. EWS/FLI-1 antagonists induce growth inhibition of Ewing tumor cells in vitro. *Cell Growth Differ*. 1996;7(4):429-437.

8. May WA, Lessnick SL, Braun BS, et al. The Ewing's sarcoma EWS/FLI-1 fusion gene encodes a more potent transcriptional activator and is a more powerful transforming gene than FLI-1. *Mol Cell Biol*. 1993;13(12):7393-7398. doi:10.1128/MCB.13.12.7393

9. Ries LAG, Smith MA, Gurney JG, et al. Cancer Incidence and Survival Among Children and Adolescents: United States SEER Program 1975-1995. *Natl Cancer Institute, SEER Program NIH Pub No 99-4649*. 1999:179. doi:10.1037/e407432005-001

10. Rodríguez-Galindo C, Liu T, Krasin MJ, et al. Analysis of prognostic factors in Ewing sarcoma family of tumors: Review of St. Jude Children's Research Hospital studies. *Cancer*. 2007;110(2):375-384. doi:10.1002/cncr.22821

11. Randall RL, Lessnick SL, Jones KB, et al. Is there a predisposition gene for ewing's sarcoma? *J Oncol*. 2010. doi:10.1155/2010/397632

12. Rocchi A, Manara MC, Sciandra M, et al. CD99 inhibits neural differentiation of human Ewing sarcoma cells and thereby contributes to oncogenesis. *J Clin Invest*. 2010;120(3):668-680. doi:10.1172/JCI36667

13. Parvizi J, Kim GK. *High Yield Orthopaedics*.; 2010. doi:10.1016/B978-1-4160-0236-9.00165-6

14. Beck R, Monument MJ, Watkins WS, et al. EWS/FLI-responsive GGAA microsatellites exhibit polymorphic differences between European and African populations. *Cancer Genet*. 2012;205(6):304-312. doi:10.1016/j.cancergen.2012.04.004

15. Gangwal K, Sankar S, Hollenhorst PC, et al. Microsatellites as EWS/FLI response elements in Ewing's sarcoma. *Proc Natl Acad Sci U S A*. 2008;105(29):10149-10154. doi:10.1073/pnas.0801073105

16. Monument MJ, Johnson KM, McIlvaine E, et al. Clinical and biochemical function of polymorphic NR0B1 GGAA-microsatellites in Ewing sarcoma: A report from the Children's Oncology Group. *PLoS One*. 2014;9(8):e104378. doi:10.1371/journal.pone.0104378

17. Gaspar N, Hawkins DS, Dirksen U, et al. Ewing sarcoma: Current management and future approaches through collaboration. *J Clin Oncol*. 2015;33(27):3036-3046. doi:10.1200/JCO.2014.59.5256

18. Cash T, McIlvaine E, Krailo MD, et al. Comparison of clinical features and outcomes in patients with extraskeletal versus skeletal localized Ewing sarcoma: A report from the Children's Oncology Group. *Pediatr Blood Cancer*. 2016;63(10):1771-1779. doi:10.1002/pbc.26096

19. Grier HE. The Ewing family of tumors. Ewing's sarcoma and primitive neuroectodermal tumors. *Pediatr Clin North Am*. 1997;44(4):991-1004. doi:10.1016/S0031-3955(05)70541-1

20. Obata H, Ueda T, Kawai A, et al. Clinical outcome of patients with Ewing sarcoma family of tumors of bone in Japan: The Japanese musculoskeletal oncology group cooperative study. *Cancer*. 2007;109(4):767-775. doi:10.1002/cncr.22481

21. Grier HE, Krailo MD, Tarbell NJ, et al. Addition of ifosfamide and etoposide to standard chemotherapy for Ewing's sarcoma and primitive neuroectodermal tumor of bone. *N Engl J Med*. 2003;348(8):694-701. doi:10.1056/NEJMoa020890

22. Wexler LH, DeLaney TF, Tsokos M, et al. Ifosfamide and etoposide plus vincristine, doxorubicin, and cyclophosphamide for newly diagnosed Ewing's sarcoma family of tumors. *Cancer*. 1996;78(4):901-911. doi:10.1002/(SICI)1097-0142(19960815)78:4<901::AID-CNCR30>3.0.CO;2-X

23. Van Doorninck JA, Ji L, Schaub B, et al. Current treatment protocols have eliminated the prognostic advantage of type 1 fusions in ewing sarcoma: A report from the children's oncology group. *J Clin Oncol*. 2010;28(12):1989-1994. doi:10.1200/JCO.2009.24.5845

24. Mirzaei L, Kaal SEJ, Schreuder HWB, Bartels RHM a. The Neurological Compromised Spine Due to Ewing Sarcoma. What First. *Neurosurgery*. 2015;0(0):1. doi:10.1227/NEU.0000000000000903

25. Smith MA, Ungerleider RS, Horowitz ME, Simon R. Influence of doxorubicin dose intensity on response and outcome for patients with osteogenic sarcoma and ewing's sarcoma. *J Natl Cancer Inst*. 1991;83(20):1460-1470. doi:10.1093/jnci/83.20.1460

26. Kolb EA, Kushner BH, Gorlick R, et al. Long-term event-free survival after intensive chemotherapy for Ewing's family of tumors in children and young adults. *J Clin Oncol*. 2003;21(18):3423-3430. doi:10.1200/JCO.2003.10.033

27. Granowetter L, Womer R, Devidas M, et al. Dose-intensified compared with standard chemotherapy for nonmetastatic Ewing sarcoma family of tumors: a Children's Oncology Group Study. *J Clin Oncol*. 2009;27(15):2536-2541. doi:10.1200/JCO.2008.19.1478

28. Paulussen M, Ahrens S, Dunst J, et al. Localized Ewing tumor of bone: Final results of the Cooperative Ewing's Sarcoma Study CESS 86. *J Clin Oncol*. 2001;19(6):1818-1829. doi:10.1200/jco.2001.19.6.1818

29. Bacci G, Ferrari S, Bertoni F, et al. Prognostic factors in nonmetastatic Ewing's sarcoma of bone treated with adjuvant chemotherapy: Analysis of 359 patients at the Istituto Ortopedico Rizzoli. *J Clin Oncol*. 2000;18(1):4-11. doi:10.1200/JCO.2000.18.1.4

30. Lawrence M, Huber W, Pagès H, et al. Software for computing and annotating genomic ranges. *PLoS Comput Biol*. 2013;9(8):e1003118. doi:10.1371/journal.pcbi.1003118

31. Brohl AS, Solomon DA, Chang W, et al. The Genomic Landscape of the Ewing Sarcoma Family of Tumors Reveals Recurrent STAG2 Mutation. *PLoS Genet*. 2014;10(7). doi:10.1371/journal.pgen.1004475

32. Grünewald TGP, Delattre O. Cooperation between somatic mutations and

germline susceptibility variants in tumorigenesis – a dangerous liaison. *Mol Cell Oncol*. 2016;3(3):e1086853. doi:10.1080/23723556.2015.1086853

33. Monument MJ, Johnson KM, Grossmann AH, Schiffman JD, Randall RL, Lessnick SL. Microsatellites with macro-influence in ewing sarcoma. *Genes (Basel)*. 2012;3(3):444-460. doi:10.3390/genes3030444

34. Sankar S, Bell R, Stephens B, et al. Mechanism and relevance of EWS/FLI-mediated transcriptional repression in Ewing sarcoma. *Oncogene*. 2013;32(42):5089-5100. doi:10.1038/onc.2012.525

35. Stegmaier S, Leuschner I, Aakcha-Rudel E, et al. Identification of various exon combinations of the ews/fli1 translocation: An optimized RT-PCR method for paraffin embedded tissue: A report by the CWS-Study Group. *Klin Padiatr*. 2004;216(6):315-322. doi:10.1055/s-2004-832338

36. Sorensen PH, Lessnick SL, Lopez-Terrada D, Liu XF, Triche TJ, Denny CT. A second Ewing's sarcoma translocation, t(21;22), fuses the EWS gene to another ETS-family transcription factor, ERG. *Nat Genet*. 1994;6(2):146-151. doi:10.1038/ng0294-146

37. Mao X, Miesfeldt S, Yang H, Leiden JM, Thompson CB. The FLI-1 and chimeric EWS-FLI-1 oncoproteins display similar DNA binding specificities. *J Biol Chem*. 1994;269(27):18216-18222.

38. Braun BS, Frieden R, Lessnick SL, May WA, Denny CT. Identification of target genes for the Ewing's sarcoma EWS/FLI fusion protein by representational difference analysis. *Mol Cell Biol*. 1995;15(8):4623-4630.

39. Kinsey M, Smith R, Iyer AK, McCabe ERB, Lessnick SL. EWS/FLI and its downstream target NR0B1 interact directly to modulate transcription and oncogenesis in Ewing's sarcoma. *Cancer Res*. 2009;69(23):9047-9055. doi:10.1158/0008-5472.CAN-09-1540

40. Luo W, Gangwal K, Sankar S, Boucher KM, Thomas D, Lessnick SL. GSTM4 is a microsatellite-containing EWS/FLI target involved in Ewing's sarcoma oncogenesis and therapeutic resistance. *Oncogene*. 2009;28(46):4126-4132. doi:10.1038/onc.2009.262

41. Owen LA, Kowalewski AA, Lessnick SL. EWS/FLI mediates transcriptional repression via NKX2.2 during oncogenic transformation in Ewing's sarcoma. *PLoS One*. 2008;3(4). doi:10.1371/journal.pone.0001965

42. Fadul J, Bell R, Hoffman LM, Beckerle MC, Engel ME, Lessnick SL. EWS/FLI utilizes NKX2-2 to repress mesenchymal features of Ewing sarcoma. *www.impactjournals.com/Genes&Cancer Genes & Cancer*. 2015;6:3-4.

doi:10.18632/genesandcancer.57

43. Kinsey M, Smith R, Lessnick SL. NR0B1 Is Required for the Oncogenic Phenotype Mediated by EWS/FLI in Ewing's Sarcoma. *Mol Cancer Res*. 2006;4(11):851-859. doi:10.1158/1541-7786.MCR-06-0090

44. Lessnick SL, Braun BS, Denny CT, May WA. Multiple domains mediate transformation by the Ewing's sarcoma EWS/FLI-1 fusion gene. *Oncogene*. 1995;10(3):423-431. http://www.ncbi.nlm.nih.gov/pubmed/7845667.

45. Sankar S, Theisen ER, Bearss J, et al. Reversible LSD1 inhibition interferes with global EWS/ETS transcriptional activity and impedes Ewing sarcoma tumor growth. *Clin Cancer Res*. 2014;20(17):4584-4597. doi:10.1158/1078-0432.CCR-14-0072

46. Guillon N, Tirode F, Boeva V, Zynovyev A, Barillot E, Delattre O. The oncogenic EWS-FLI1 protein binds in vivo GGAA microsatellite sequences with potential transcriptional activation function. *PLoS One*. 2009;4(3):e4932. doi:10.1371/journal.pone.0004932

47. Boulay G, Sandoval GJ, Riggi N, et al. Cancer-Specific Retargeting of BAF Complexes by a Prion-like Domain. *Cell*. 2017. doi:10.1016/j.cell.2017.07.036

48. Lamber EP, Vanhille L, Textor LC, Kachalova GS, Sieweke MH, Wilmanns M. Regulation of the transcription factor Ets-1 by DNA-mediated homo-dimerization. *EMBO J*. 2008;27(14):2006-2017. doi:10.1038/emboj.2008.117

49. Hollenhorst PC, McIntosh LP, Graves BJ. Genomic and Biochemical Insights into the Specificity of ETS Transcription Factors. *Annu Rev Biochem*. 2011;80(1):437-471. doi:10.1146/annurev.biochem.79.081507.103945

50. Uchiumi F, Miyazaki S, Tanuma SI. The possible functions of duplicated ets (GGAA) motifs located near transcription start sites of various human genes. *Cell Mol Life Sci*. 2011;68(12):2039-2051. doi:10.1007/s00018-011-0674-x

51. Gangwal K, Close D, Enriquez CA, Hill CP, Lessnick SL. Emergent Properties of EWS/FLI Regulation via GGAA Microsatellites in Ewing's Sarcoma. *Genes Cancer*. 2010;1(2):177-187. doi:10.1177/1947601910361495

52. Wei G-H, Badis G, Berger MF, et al. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO J*. 2010;29(13):2147-2160. doi:10.1038/emboj.2010.106

53. Slattery M, Zhou T, Yang L, Dantas Machado AC, Gordân R, Rohs R. Absence of a simple code: how transcription factors read the genome. *Trends Biochem Sci*. 2014;39(9):381-399. doi:10.1016/j.tibs.2014.07.002

54. Wei GH, Badis G, Berger MF, et al. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO J*. 2010;29(13):2147-2160. doi:10.1038/emboj.2010.106

55. Liang H, Mao X, Olejniczak ET, et al. Solution structure of the ets domain of Fli-1 when bound to DNA. *Nat Struct Biol*. 1994;1(12):871-875. doi:10.1038/nsb1294-871

56. Cooper CDO, Newman JA, Gileadi O. Recent advances in the structural molecular biology of Ets transcription factors: interactions, interfaces and inhibition. *Biochem Soc Trans*. 2014;42(1):130-138. doi:10.1042/BST20130227

57. Uchiumi F, Miyazaki S, Tanuma S-I. [Biological functions of the duplicated GGAA-motifs in various human promoter regions]. *Yakugaku Zasshi*. 2011;131(12):1787-1800. http://www.ncbi.nlm.nih.gov/pubmed/22129877.

58. Mayba O, Gilbert HN, Liu J, et al. MBASED: allele-specific expression detection in cancer tissues and cell lines. *Genome Biol*. 2014;15(8):405. doi:10.1186/s13059-014-0405-3

59. Hou C, Tsodikov O V. Structural Basis for Dimerization and DNA Binding of Transcription Factor FLI1. *Biochemistry*. 2015;54(50):7365-7374. doi:10.1021/acs.biochem.5b01121

60. Sankar S, Lessnick SL. Promiscuous partnerships in Ewing's sarcoma. *Cancer Genet*. 2011;204(7):351-365. doi:10.1016/j.cancergen.2011.07.008

61. Jeon IS, Davis JN, Braun BS, et al. A variant Ewing's sarcoma translocation (7;22) fuses the EWS gene to the ETS gene ETV1. *Oncogene*. 1995;10(6):1229-1234. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=7700648.

62. Oikawa T, Yamada T. Molecular biology of the Ets family of transcription factors. *Gene*. 2003;303:11-34. doi:S0378111902011563 [pii]

63. Bailly RA, Bosselut R, Zucman J, et al. DNA-binding and transcriptional activation properties of the EWS-FLI-1 fusion protein resulting from the t(11;22) translocation in Ewing sarcoma. *Mol Cell Biol*. 1994;14(5):3230-3241. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=358690&tool=pmcentrez&rendertype=abstract. Accessed February 14, 2017.

64. Sorensen PH, Triche TJ. Gene fusions encoding chimaeric transcription factors in solid tumours. *Semin Cancer Biol*. 1996;7(1):3-14. doi:10.1006/scbi.1996.0002

65. Regan MC, Horanyi PS, Pryor Jr. EE, Sarver JL, Cafiso DS, Bushweller JH. Structural and dynamic studies of the transcription factor ERG reveal DNA

binding is allosterically autoinhibited. *Proc Natl Acad Sci U S A*. 2013;110(33):13374-13379. doi:10.1073/pnas.1301726110

66. Obika S, Reddy SY, Bruice TC. Sequence specific DNA binding of Ets-1 transcription factor: molecular dynamics study on the Ets domain--DNA complexes. *J Mol Biol*. 2003;331(2):345-359.

67. Roychoudhury M, Sitlani A, Lapham J, Crothers DM. Global structure and mechanical properties of a 10-bp nucleosome positioning motif. *Proc Natl Acad Sci U S A*. 2000;97(25):13608-13. doi:10.1073/pnas.250476297

68. Magnaghi-Jaulin L, Masutani H, Robin P, Lipinski M, Harel-Bellan A. SRE elements are binding sites for the fusion protein EWS-FLI-1. *Nucleic Acids Res*. 1996;24(6):1052-1058. doi:10.1093/nar/24.6.1052

69. Riggi N, Knoechel B, Gillespie SM, et al. EWS-FLI1 utilizes divergent chromatin remodeling mechanisms to directly activate or repress enhancer elements in Ewing sarcoma. *Cancer Cell*. 2014;26(5):668-681. doi:10.1016/j.ccell.2014.10.004

70. Tomazou EM, Sheffield NC, Schmidl C, et al. Epigenome mapping reveals distinct modes of gene regulation and widespread enhancer reprogramming by the oncogenic fusion protein EWS-FLI1. *Cell Rep*. 2015;10(7):1082-1095. doi:10.1016/j.celrep.2015.01.042

71. Huang YQ, Rehfuss RP, Laplante SR, Boudreau E, Borer PN, Lane MJ. Actinomycin D induced DNAse I cleavage enhancement caused by sequence specific propagation of an altered DNA structure. *Nucleic Acids Res*. 1988;16(23):11125-11139. doi:10.1093/nar/16.23.11125

72. Kennedy AL, Vallurupalli M, Chen L, et al. Functional, chemical genomic, and super-enhancer screening identify sensitivity to cyclin D1/CDK4 pathway inhibition in Ewing sarcoma. *Oncotarget*. 2015;6(30):30178-30193. doi:10.18632/oncotarget.4903

73. Erie DA, Yang G, Schultz HC, Bustamante C. DNA bending by Cro protein in specific and nonspecific complexes: implications for protein site recognition and specificity. *Science*. 1994;266(5190):1562-1566. doi:10.1126/science.7985026

74. Szymczyna BR, Arrowsmith CH. DNA binding specificity studies of four ETS proteins support an indirect read-out mechanism of protein-DNA recognition. *J Biol Chem*. 2000;275(37):28363-28370. doi:10.1074/jbc.M004294200

75. May WA, Gishizky ML, Lessnick SL, et al. Ewing sarcoma 11;22 translocation produces a chimeric transcription factor that requires the DNA-binding domain encoded by FLI1 for transformation. *Proc Natl Acad Sci U S A*. 1993;90(12):5752-5756. doi:10.1073/pnas.90.12.5752

76. Ng KP, Potikyan G, Savene RO V, Denny CT, Uversky VN, Lee K a W. Multiple aromatic side chains within a disordered structure are critical for transcription and transforming activity of EWS family oncoproteins. *Proc Natl Acad Sci U S A*. 2007;104(2):479-484. doi:10.1073/pnas.0607007104

77. Erkizan H V., Uversky VN, Toretsky JA. Oncogenic partnerships: EWS-FLI1 protein interactions initiate key pathways of Ewing's sarcoma. *Clin Cancer Res*. 2010;16(16):4077-4083. doi:10.1158/1078-0432.CCR-09-2261

78. Kato M, Han TW, Xie S, et al. Cell-free formation of RNA granules: Low complexity sequence domains form dynamic fibers within hydrogels. *Cell*. 2012;149(4):753-767. doi:10.1016/j.cell.2012.04.017

79. Kwon I, Kato M, Xiang S, et al. Phosphorylation-regulated Binding of RNA Polymerase II to Fibrous Polymers of Low Complexity Domains. *Cell*. 2013;155(5):1049-1060. doi:10.1016/j.cell.2013.10.033

80. Altmeyer M, Neelsen KJ, Teloni F, et al. Liquid demixing of intrinsically disordered proteins is seeded by poly(ADP-ribose). *Nat Commun*. 2015;6:8088. doi:10.1038/ncomms9088

81. Dunker AK, Lawson JD, Brown CJ, et al. Intrinsically disordered protein. *J Mol Graph Model*. 2001;19(1):26-59. doi:10.1016/S1093-3263(00)00138-8

82. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol*. 2005;6(3):197-208. doi:10.1038/nrm1589

83. Uversky VN, Oldfield CJ, Dunker AK. Intrinsically disordered proteins in human diseases: Introducing the D2 concept. *Annu Rev Biophys*. 2008;37(1):215-246. doi:10.1146/annurev.biophys.37.032807.125924

84. Xiang S, Kato M, Wu LC, et al. The LC Domain of hnRNPA2 Adopts Similar Conformations in Hydrogel Polymers, Liquid-like Droplets, and Nuclei. *Cell*. 2015;163(4):829-839. doi:10.1016/j.cell.2015.10.040

85. Han TW, Kato M, Xie S, et al. Cell-free formation of RNA granules: Bound RNAs identify features and components of cellular assemblies. *Cell*. 2012;149(4):768-779. doi:10.1016/j.cell.2012.04.016

86. Bertolotti A, Bell B, Tora L. The N-terminal domain of human TAFII68 displays transactivation and oncogenic properties. *Oncogene*. 1999;18(56):8000-8010. doi:10.1038/sj.onc.1203207

87. Zinszner H, Albalat R, Ron D. A novel effector domain from the RNA-binding protein TLS or EWS is required for oncogenic transformation by CHOP. *Genes Dev*. 1994;8(21):2513-2526. doi:DOI 10.1101/gad.8.21.2513

88. Pérez-Losada J, Pintado B, Gutiérrez-Adán a, et al. The chimeric FUS/TLS-CHOP fusion protein specifically induces liposarcomas in transgenic mice. *Oncogene*. 2000;19(20):2413-2422. doi:10.1038/sj.onc.1203572

89. Rossow KL, Janknecht R. The Ewing's sarcoma gene product functions as a transcriptional activator. *Cancer Res*. 2001;61(6):2690-2695.

90. Li KKC, Lee KAW. Transcriptional activation by the Ewing's sarcoma (EWS) oncogene can be cis-repressed by the EWS RNA-binding domain. *J Biol Chem*. 2000;275(30):23053-23058. doi:10.1074/jbc.M002961200

91. Alex D, Lee KAW. RGG-boxes of the EWS oncoprotein repress a range of transcriptional activation domains. *Nucleic Acids Res*. 2005;33(4):1323-1331. doi:10.1093/nar/gki270

92. Tan AY, Manley JL. The TET family of proteins: Functions and roles in disease. *J Mol Cell Biol*. 2009;1(2):82-92. doi:10.1093/jmcb/mjp025

93. Schwartz JC, Ebmeier CC, Podell ER, Heimiller J, Taatjes DJ, Cech TR. FUS binds the CTD of RNA polymerase II and regulates its phosphorylation at Ser2. *Genes Dev*. 2012;26(24):2690-2695. doi:10.1101/gad.204602.112

94. Hsin J-P, Manley JL. The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes Dev*. 2012;26(19):2119-2137. doi:10.1101/gad.200303.112

95. Egloff S, Murphy S. Cracking the RNA polymerase II CTD code. *Trends Genet*. 2008;24(6):280-288. doi:10.1016/j.tig.2008.03.008

96. Schwartz JC, Wang X, Podell ER, Cech TR. RNA Seeds Higher-Order Assembly of FUS Protein. *Cell Rep*. 2013;5(4):918-925. doi:10.1016/j.celrep.2013.11.017

97. Schwartz JC, Podell ER, Han SSW, Berry JD, Eggan KC, Cech TR. FUS is sequestered in nuclear aggregates in ALS patient fibroblasts. *Mol Biol Cell*. 2014;25(17):2571-2578. doi:10.1091/mbc.E14-05-1007

98. Corden JL, Cadena DL, Ahearn JM, Dahmus ME. A unique structure at the carboxyl terminus of the largest subunit of eukaryotic RNA polymerase II. *Proc Natl Acad Sci U S A*. 1985;82(23):7934-7938. doi:10.1073/pnas.82.23.7934

99. Cramer P. RNA polymerase II structure: From core to functional complexes. *Curr Opin Genet Dev*. 2004;14(2):218-226. doi:10.1016/j.gde.2004.01.003

100. Sikorski TW, Buratowski S. The basal initiation machinery: beyond the general transcription factors. *Curr Opin Cell Biol*. 2009;21(3):344-351. doi:10.1016/j.ceb.2009.03.006

101. Yang L, Chansky HA, Hickstein DD. EWS·Fli-1 fusion protein interacts with hyperphosphorylated RNA polymerase II and interferes with serine-arginine protein-mediated RNA splicing. *J Biol Chem*. 2000;275(48):37612-37618. doi:10.1074/jbc.M005739200

102. Iko Y, Kodama TS, Kasai N, et al. Domain architectures and characterization of an RNA-binding protein, TLS. *J Biol Chem*. 2004;279(43):44834-44840. doi:10.1074/jbc.M408552200

103. Lodish HF, Berk A, Kaiser CA, et al. *Molecular Cell Biology, 7th Edition*. Vol 5.; 2013. doi:10.1016/S1470-8175(01)00023-6

104. Weinberg RA. *Biology of the Cancer*. Vol XXXIII.; 2014. doi:10.1007/s13398-014-0173-7.2

105. Richard G-F, Kerrest A, Dujon B. Comparative Genomics and Molecular Dynamics of DNA Repeats in Eukaryotes. *Microbiol Mol Biol Rev*. 2008. doi:10.1128/MMBR.00011-08

106. John HA, Birnstiel ML, Jones KW. RNA-DNA hybrids at the cytological level. *Nat Publ Gr*. 1969;223:582-587. doi:10.1038/223582a0

107. Kumar V, Abbas AK, Fausto N, Aster JC. *Robbins and Cotran Pathologic Basis of Disease*.; 2010.

108. Sawaya S, Bagshaw A, Buschiazzo E, et al. Microsatellite Tandem Repeats Are Abundant in Human Promoters and Are Associated with Regulatory Elements. *PLoS One*. 2013. doi:10.1371/journal.pone.0054710

109. Brouwer JR, Willemsen R, Oostra BA. Microsatellite repeat instability and neurological disease. *BioEssays*. 2009;31(1):71-83. doi:10.1002/bies.080122

110. Helman LJ, Meltzer P. Mechanisms of sarcoma development. *Nat Rev Cancer*. 2003;3(9):685-694. doi:10.1038/nrc1168

111. Martens JH, Stunnenberg HG. The molecular signature of oncofusion proteins in acute myeloid leukemia. *FEBS Lett*. 2010;584(12):2662-2669. doi:10.1016/j.febslet.2010.04.002

112. Maheshwari A V, Cheng EY. Ewing sarcoma family of tumors. *J Am Acad Orthop Surg*. 2010;18(2):94-107. http://www.ncbi.nlm.nih.gov/pubmed/20118326.

113. Sankar S, Lessnick SL. Promiscuous partnerships in Ewing's sarcoma. *Cancer Genet*. 2011;204(7):351-365. doi:10.1016/j.cancergen.2011.07.008

114. Smith R, Owen LA, Trem DJ, et al. Expression profiling of EWS/FLI identifies NKX2.2 as a critical target gene in Ewing's sarcoma. *Cancer Cell*. 2006;9(5):405-

416. doi:10.1016/j.ccr.2006.04.004

115. Hsu T, Trojanowska M, Watson DK. Ets proteins in biological control and cancer. *J Cell Biochem*. 2004;91(5):896-903. doi:10.1002/jcb.20012

116. Tomlins SA, Rhodes DR, Perner S, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science (80-)*. 2005;310(5748):644-648. doi:10.1126/science.1117679

117. Clark J, Attard G, Jhavar S, et al. Complex patterns of ETS gene alteration arise during cancer development in the human prostate. *Oncogene*. 2008;27(14):1993-2003. doi:10.1038/sj.onc.1210843

118. Hollenhorst PC, Shah AA, Hopkins C, Graves BJ. Genome-wide analyses reveal properties of redundant and specific promoter occupancy within the ETS gene family. *Genes Dev*. 2007;21(15):1882-1894. doi:10.1101/gad.1561707

119. Nye JA, Petersen JM, Gunther C V, Jonsen MD, Graves BJ. Interaction of murine ets-1 with GGA-binding sites establishes the ETS domain as a new DNA-binding motif. *Genes Dev*. 1992;6(6):975-990. http://www.ncbi.nlm.nih.gov/pubmed/1592264.

120. Seth A, Watson DK. ETS transcription factors and their emerging roles in human cancer. *Eur J Cancer*. 2005;41(16):2462-2478. doi:10.1016/j.ejca.2005.08.013

121. Szymczyna BR, Arrowsmith CH. DNA binding specificity studies of four ETS proteins support an indirect read-out mechanism of protein-DNA recognition. *J Biol Chem*. 2000;275(37):28363-28370. doi:10.1074/jbc.M004294200

122. Zhang XK, Moussa O, LaRue A, et al. The transcription factor Fli-1 modulates marginal zone and follicular B cell development in mice. *J Immunol*. 2008;181(3):1644-1654. http://www.ncbi.nlm.nih.gov/pubmed/18641300.

123. Hart A, Melet F, Grossfeld P, et al. Fli-1 is required for murine vascular and megakaryocytic development and is hemizygously deleted in patients with thrombocytopenia. *Immunity*. 2000;13(2):167-177. http://www.ncbi.nlm.nih.gov/pubmed/10981960.

124. Loughran SJ, Kruse EA, Hacking DF, et al. The transcription factor Erg is essential for definitive hematopoiesis and the function of adult hematopoietic stem cells. *Nat Immunol*. 2008;9(7):810-819. doi:10.1038/ni.1617

125. Ohno T, Ouchida M, Lee L, Gatalica Z, Rao VN, Reddy ES. The EWS gene, involved in Ewing family of tumors, malignant melanoma of soft parts and desmoplastic small round cell tumors, codes for an RNA binding protein with novel regulatory domains. *Oncogene*. 1994;9(10):3087-3097.

http://www.ncbi.nlm.nih.gov/pubmed/8084618.

126. Bertolotti A, Lutz Y, Heard DJ, Chambon P, Tora L. hTAF(II)68, a novel RNA/ssDNA-binding protein with homology to the pro-oncoproteins TLS/FUS and EWS is associated with both TFIID and RNA polymerase II. *EMBO J*. 1996;15(18):5022-5031. http://www.ncbi.nlm.nih.gov/pubmed/8890175.

127. Paronetto MP, Minana B, Valcarcel J. The Ewing sarcoma protein regulates DNA damage-induced alternative splicing. *Mol Cell*. 2011;43(3):353-368. doi:10.1016/j.molcel.2011.05.035

128. Patel M, Simon JM, Iglesia MD, et al. Tumor-specific retargeting of an oncogenic transcription factor chimera results in dysregulation of chromatin and transcription. *Genome Res*. 2011. doi:10.1101/gr.125666.111

129. Prieur A, Tirode F, Cohen P, Delattre O. EWS/FLI-1 silencing and gene profiling of Ewing cells reveal downstream oncogenic pathways and a crucial role for repression of insulin-like growth factor binding protein 3. *Mol Cell Biol*. 2004;24(16):7275-7283. doi:10.1128/MCB.24.16.7275-7283.2004

130. Niakan KK, McCabe ER. DAX1 origin, function, and novel role. *Mol Genet Metab*. 2005;86(1-2):70-83. doi:10.1016/j.ymgme.2005.07.019

131. McCabe ER. DAX1: Increasing complexity in the roles of this novel nuclear receptor. *Mol Cell Endocrinol*. 2007;265-266:179-182. doi:10.1016/j.mce.2006.12.017

132. Mendiola M, Carrillo J, Garcia E, et al. The orphan nuclear receptor DAX1 is up-regulated by the EWS/FLI1 oncoprotein and is highly expressed in Ewing tumors. *Int J Cancer*. 2006;118(6):1381-1389. doi:10.1002/ijc.21578

133. Garcia-Aragoncillo E, Carrillo J, Lalli E, et al. DAX1, a direct target of EWS/FLI1 oncoprotein, is a principal regulator of cell-cycle progression in Ewing's tumor cells. *Oncogene*. 2008;27(46):6034-6043. doi:10.1038/onc.2008.203

134. Graves BJ, Gillespie ME, McIntosh LP. DNA binding by the ETS domain. *Nature*. 1996;384(6607):322. doi:10.1038/384322a0

135. Martins AS, Ordonez JL, Amaral AT, et al. IGF1R signaling in Ewing sarcoma is shaped by clathrin-/caveolin-dependent endocytosis. *PLoS One*. 2011;6(5):e19846. doi:10.1371/journal.pone.0019846

136. Williams TM, Lisanti MP. Caveolin-1 in oncogenic transformation, cancer, and metastasis. *Am J Physiol Cell Physiol*. 2005;288(3):C494-506. doi:10.1152/ajpcell.00458.2004

137. Tirado OM, Mateo-Lozano S, Villar J, et al. Caveolin-1 (CAV1) is a target of

EWS/FLI-1 and a key determinant of the oncogenic phenotype and tumorigenicity of Ewing's sarcoma cells. *Cancer Res*. 2006;66(20):9937-9947. doi:66/20/9937 [pii]10.1158/0008-5472.CAN-06-0927

138. Grohar PJ, Woldemichael GM, Griffin LB, et al. Identification of an inhibitor of the EWS-FLI1 oncogenic transcription factor by high-throughput screening. *J Natl Cancer Inst*. 2011;103(12):962-978. doi:10.1093/jnci/djr156

139. Grohar PJ, Griffin LB, Yeung C, et al. Ecteinascidin 743 interferes with the activity of EWS-FLI1 in Ewing sarcoma cells. *Neoplasia*. 2011;13(2):145-153. http://www.ncbi.nlm.nih.gov/pubmed/21403840.

140. Erkizan H V, Scher LJ, Gamble SE, et al. Novel peptide binds EWS-FLI1 and reduces the oncogenic potential in Ewing tumors. *Cell cycle*. 2011;10(19):3397-3408. doi:10.4161/cc.10.19.17734

141. Kovar H. Context matters: the hen or egg problem in Ewing's sarcoma. *Semin Cancer Biol*. 2005;15(3):189-196. doi:10.1016/j.semcancer.2005.01.004

142. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860-921. doi:10.1038/35057062

143. Aaltonen LA, Peltomaki P, Leach FS, et al. Clues to the pathogenesis of familial colorectal cancer. *Science (80-)*. 1993;260(5109):812-816. http://www.ncbi.nlm.nih.gov/pubmed/8484121.

144. Thibodeau SN, Bren G, Schaid D. Microsatellite instability in cancer of the proximal colon. *Science (80-)*. 1993;260(5109):816-819. http://www.ncbi.nlm.nih.gov/pubmed/8484122.

145. Ionov Y, Peinado MA, Malkhosyan S, Shibata D, Perucho M. Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nature*. 1993;363(6429):558-561. doi:10.1038/363558a0

146. Popat S, Hubner R, Houlston RS. Systematic review of microsatellite instability and colorectal cancer prognosis. *J Clin Oncol*. 2005;23(3):609-618. doi:23/3/609 [pii]10.1200/JCO.2005.01.086

147. Pinol V, Castells A, Andreu M, et al. Accuracy of revised Bethesda guidelines, microsatellite instability, and immunohistochemistry for the identification of patients with hereditary nonpolyposis colorectal cancer. *JAMA*. 2005;293(16):1986-1994. doi:10.1001/jama.293.16.1986

148. Guastadisegni C, Colafranceschi M, Ottini L, Dogliotti E. Microsatellite instability as a marker of prognosis and response to therapy: a meta-analysis of colorectal cancer survival data. *Eur J Cancer*. 2010;46(15):2788-2798.

doi:10.1016/j.ejca.2010.05.009

149. Alldinger I, Schaefer KL, Goedde D, et al. Microsatellite instability in Ewing tumor is not associated with loss of mismatch repair protein expression. *J Cancer Res Clin Oncol*. 2007;133(10):749-759. doi:10.1007/s00432-007-0220-2

150. Ebinger M, Bock T, Kandolf R, Sotlar K, Bultmann BD, Greil J. Standard mono- and dinucleotide repeats do not appear to be sensitive markers of microsatellite instability in the Ewing family of tumors. *Cancer Genet Cytogenet*. 2005;157(2):189-190. doi:10.1016/j.cancergencyto.2004.08.008

151. Ohali A, Avigad S, Cohen IJ, et al. High frequency of genomic instability in Ewing family of tumors. *Cancer Genet Cytogenet*. 2004;150(1):50-56. doi:10.1016/j.cancergencyto.2003.08.014S0165460803003650 [pii]

152. Tzaida O, Gogas H, Dafni U, et al. Evaluation of the prognostic and predictive value of HER-1/EGFR in breast cancer patients participating in a randomized study with dose-dense sequential adjuvant chemotherapy. *Oncology*. 2007;72(5-6):388-396. doi:10.1159/000113148

153. Nogi H, Kobayashi T, Suzuki M, et al. EGFR as paradoxical predictor of chemosensitivity and outcome among triple-negative breast cancer. *Oncol Rep*. 2009;21(2):413-417. http://www.ncbi.nlm.nih.gov/pubmed/19148516.

154. Gebhardt F, Zanker KS, Brandt B. Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. *J Biol Chem*. 1999;274(19):13176-13180. http://www.ncbi.nlm.nih.gov/pubmed/10224073.

155. Chamberlain NL, Driver ED, Miesfeld RL. The length and location of CAG trinucleotide repeats in the androgen receptor N-terminal domain affect transactivation function. *Nucleic Acids Res*. 1994;22(15):3181-3186. http://www.ncbi.nlm.nih.gov/pubmed/8065934.

156. Stanford JL, Just JJ, Gibbs M, et al. Polymorphic repeats in the androgen receptor gene: molecular markers of prostate cancer risk. *Cancer Res*. 1997;57(6):1194-1198. http://www.ncbi.nlm.nih.gov/pubmed/9067292.

157. Giovannucci E, Stampfer MJ, Krithivas K, et al. The CAG repeat within the androgen receptor gene and its relationship to prostate cancer. *Proc Natl Acad Sci U S A*. 1997;94(7):3320-3323. http://www.ncbi.nlm.nih.gov/pubmed/9096391.

158. Contente A, Dittmer A, Koch MC, Roth J, Dobbelstein M. A polymorphic microsatellite that mediates induction of PIG3 by p53. *Nat Genet*. 2002;30(3):315-320. doi:10.1038/ng836

159. Kotsinas A, Aggarwal V, Tan EJ, Levy B, Gorgoulis VG. PIG3: A novel link

between oxidative stress and DNA damage response in cancer. *Cancer Lett*. 2011. doi:10.1016/j.canlet.2011.12.009

160. Polednak AP. Primary bone cancer incidence in black and white residents of New York State. *Cancer*. 1985;55(12):2883-2888. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=3857961.

161. Jawad MU, Cheung MC, Min ES, Schneiderbauer MM, Koniaris LG, Scully SP. Ewing sarcoma demonstrates racial disparities in incidence-related and sex-related differences in outcome: an analysis of 1631 cases from the SEER database, 1973-2005. *Cancer*. 2009;115(15):3526-3536. doi:10.1002/cncr.24388

162. Worch J, Matthay KK, Neuhaus J, Goldsby R, DuBois SG. Ethnic and racial differences in patients with Ewing sarcoma. *Cancer*. 2010;116(4):983-988. doi:10.1002/cncr.24865

163. Zucman-Rossi J, Batzer MA, Stoneking M, Delattre O, Thomas G. Interethnic polymorphism of EWS intron 6: genome plasticity mediated by Alu retroposition and recombination. *Hum Genet*. 1997;99(3):357-363. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=9050923.

164. Dubois SG, Goldsby R, Segal M, et al. Evaluation of polymorphisms in EWSR1 and risk of Ewing sarcoma: A report from the childhood cancer survivor study. *Pediatr Blood Cancer*. 2011. doi:10.1002/pbc.23263

165. Eckert KA, Hile SE. Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome. *Mol Carcinog*. 2009;48(4):379-388. doi:10.1002/mc.20499

166. Jorde LB, Rogers AR, Bamshad M, et al. Microsatellite diversity and the demographic history of modern humans. *Proc Natl Acad Sci U S A*. 1997;94(7):3100-3103. http://www.ncbi.nlm.nih.gov/pubmed/9096352.

167. Ellegren H. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet*. 2004;5(6):435-445. doi:10.1038/nrg1348

168. Bamshad MJ, Wooding S, Watkins WS, Ostler CT, Batzer MA, Jorde LB. Human population genetic structure and inference of group membership. *Am J Hum Genet*. 2003;72(3):578-589. doi:10.1086/368061

169. Beck R, Monument MJ, Watkins WS, et al. EWS/FLI-responsive GGAA microsatellites exhibit polymorphic differences between European and African populations. *Cancer Genet*. 2012;205(6):304-312. doi:10.1016/j.cancergen.2012.04.004

170. Gangwal kunal, lessnick stephen. Microsatellites are EWS / FLI response elements. *Cell Cycle*. 2008;7(October):3127-3132.

171. Martinez-Ramirez A, Rodriguez-Perales S, Melendez B, et al. Characterization of the A673 cell line (Ewing tumor) by molecular cytogenetic techniques. *Cancer Genet Cytogenet*. 2003;141(2):138-142. http://www.ncbi.nlm.nih.gov/pubmed/12606131.

172. May WA, Grigoryan RS, Keshelava N, et al. Characterization and Drug Resistance Patterns of Ewing's Sarcoma Family Tumor Cell Lines. *PLoS One*. 2013;8(12):e80060. doi:10.1371/journal.pone.0080060

173. Whang-Peng J, Triche TJ, Knutsen T, et al. Cytogenetic characterization of selected small round cell tumors of childhood. *Cancer Genet Cytogenet*. 1986;21(3):185-208. http://www.ncbi.nlm.nih.gov/pubmed/3004699.

174. Batra S, Reynolds CP, Maurer BJ. Fenretinide cytotoxicity for Ewing's sarcoma and primitive neuroectodermal tumor cell lines is decreased by hypoxia and synergistically enhanced by ceramide modulators. *Cancer Res*. 2004;64(15):5415-5424. doi:10.1158/0008-5472.CAN-04-0377

175. Lessnick SL, Dacwag CS, Golub TR. The Ewing's sarcoma oncoprotein EWS/FLI induces a p53-dependent growth arrest in primary human fibroblasts. *Cancer Cell*. 2002;1(4):393-401. doi:10.1016/S1535-6108(02)00056-9

176. Womer RB, West DC, Krailo MD, et al. Randomized controlled trial of interval-compressed chemotherapy for the treatment of localized Ewing sarcoma: a report from the Children's Oncology Group. *J Clin Oncol*. 2012;30(33):4148-4154. doi:10.1200/JCO.2011.41.5703

177. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data.* Second Edi. New York: John Wiley and Sons; 2002.

178. Borinstein SC, Beeler N, Block JJ, et al. A Decade in Banking Ewing Sarcoma: A Report from the Children's Oncology Group. *Front Oncol*. 2013;3:57. doi:10.3389/fonc.2013.00057

179. Consortium IH, Frazer KA, Ballinger DG, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature*. 2007;449(7164):851-861. doi:10.1038/nature06258

180. Monument MJ, Johnson KM, McIlvaine E, et al. Clinical and biochemical function of polymorphic NR0B1 GGAA-microsatellites in Ewing sarcoma: A report from the Children's Oncology Group. *PLoS One*. 2014;9(8):e104378. doi:10.1371/journal.pone.0104378

181. Stephens MA. EDF Statistics for Goodness of Fit and Some Comparisons. *J Am Stat Assoc*. 1974;69(347):730-737. doi:10.1080/01621459.1974.10480196

182. Arzimanoglou II, Gilbert F, Barber HR. Microsatellite instability in human solid tumors. *Cancer*. 1998;82(10):1808-1820. http://www.ncbi.nlm.nih.gov/pubmed/9587112.

183. Martin SS, Hurt WG, Hedges LK, Butler MG, Schwartz HS. Microsatellite instability in sarcomas. *Ann Surg Oncol*. 1998;5(4):356-360. http://www.ncbi.nlm.nih.gov/pubmed/9641458.

184. Dean FB, Hosono S, Fang L, et al. Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A*. 2002;99(8):5261-5266. doi:10.1073/pnas.082089499

185. Iglesias AR, Kindlund E, Tammi M, Wadelius C. Some microsatellites may act as novel polymorphic cis-regulatory elements through transcription factor binding. *Gene*. 2004;341:149-165. doi:10.1016/j.gene.2004.06.035

186. Martin P, van de Ven T, Mouchel N, Jeffries AC, Hood DW, Moxon ER. Experimentally revised repertoire of putative contingency loci in Neisseria meningitidis strain MC58: evidence for a novel mechanism of phase variation. *Mol Microbiol*. 2003;50(1):245-257. http://www.ncbi.nlm.nih.gov/pubmed/14507378.

187. Martin P, Makepeace K, Hill SA, Hood DW, Moxon ER. Microsatellite instability regulates transcription factor binding and gene expression. *Proc Natl Acad Sci U S A*. 2005;102(10):3800-3804. doi:10.1073/pnas.0406805102

188. Eckert KA, Yan G, Hile SE. Mutation rate and specificity analysis of tetranucleotide microsatellite DNA alleles in somatic human cells. *Mol Carcinog*. 2002;34(3):140-150. doi:10.1002/mc.10058

189. Postel-Vinay S, Veron AS, Tirode F, et al. Common variants near TARDBP and EGR2 are associated with susceptibility to Ewing sarcoma. *Nat Genet*. 2012. doi:10.1038/ng.1085

190. Weiss R, Walz PH. Peripheral primitive neuroectodermal tumour in a lumbar vertebra and the liver of a dromedary camel (Camelus dromedarius). *J Comp Pathol*. 2009;141(2-3):182-186. doi:10.1016/j.jcpa.2008.11.008

191. Lessnick SL, Ladanyi M. Molecular pathogenesis of Ewing sarcoma: new therapeutic and transcriptional targets. *Annu Rev Pathol*. 2012;7:145-159. doi:10.1146/annurev-pathol-011110-130237

192. Castillero-Trejo Y, Eliazer S, Xiang L, Richardson JA, Ilaria Jr. RL. Expression of the EWS/FLI-1 oncogene in murine primary bone-derived cells Results in

EWS/FLI-1-dependent, ewing sarcoma-like tumors. *Cancer Res*. 2005;65(19):8698-8705. doi:10.1158/0008-5472.CAN-05-1704

193. Riggi N, Cironi L, Provero P, et al. Development of Ewing's sarcoma from primary bone marrow-derived mesenchymal progenitor cells. *Cancer Res*. 2005;65(24):11459-11468. doi:10.1158/0008-5472.CAN-05-1696

194. Torchia EC, Boyd K, Rehg JE, Qu C, Baker SJ. EWS/FLI-1 induces rapid onset of myeloid/erythroid leukemia in mice. *Mol Cell Biol*. 2007;27(22):7918-7934. doi:10.1128/MCB.00099-07

195. Braunreiter CL, Hancock JD, Coffin CM, Boucher KM, Lessnick SL. Expression of EWS-ETS fusions in NIH3T3 cells reveals significant differences to Ewing's sarcoma. *Cell cycle*. 2006;5(23):2753-2759. http://www.ncbi.nlm.nih.gov/pubmed/17172842.

196. Uren A, Tcherkasskaya O, Toretsky J a. Recombinant EWS-FLI1 oncoprotein activates transcription. *Biochemistry*. 2004;43(42):13579-13589. doi:10.1021/bi048776q

197. Zaret KS, Carroll JS. Pioneer transcription factors: Establishing competence for gene expression. *Genes Dev*. 2011;25(21):2227-2241. doi:10.1101/gad.176826.111

198. Ku M, Koche RP, Rheinbay E, et al. Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet*. 2008;4(10). doi:10.1371/journal.pgen.1000242

199. Mikkelsen TS, Ku M, Jaffe DB, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*. 2007;448(7153):553-560. doi:10.1038/nature06008

200. Nix DA, Courdy SJ, Boucher KM. Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC Bioinformatics*. 2008;9. doi:10.1186/1471-2105-9-523

201. Hu-Lieskovan S, Zhang J, Wu L, Shimada H, Schofield DE, Triche TJ. EWS-FLI1 fusion protein up-regulates critical genes in neural crest development and is responsible for the observed phenotype of Ewing's family of tumors. *Cancer Res*. 2005;65(11):4633-4644. doi:10.1158/0008-5472.CAN-04-2857

202. Pagès H. Biostrings: String objects representing biological sequences, and matching algorithms. 2016.

203. Pagès H. BSgenome: Infrastructure for Biostrings-based genome data packages and support for efficient SNP representation. 2016.

204. Zhu LJ, Gazin C, Lawson ND, et al. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics*. 2010;11(1):237. doi:10.1186/1471-2105-11-237

205. Hollenhorst PC, Shah AA, Hopkins C, Graves BJ. Genome-wide analyses reveal properties of redundant and specific promoter occupancy within the ETS gene family. *Genes & Dev*. 2007;21(15):1882-1894. doi:10.1101/gad.1561707

206. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352

207. Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. *Nat Protoc*. 2012;7(9):1728-1740. doi:10.1038/nprot.2012.101

208. Gel B, Díez-Villanueva A, Serra E, Buschbeck M, Peinado MA, Malinverni R. regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics*. 2016;32(2):289-291. doi:10.1093/bioinformatics/btv562

209. Sankar S, Gomez NC, Bell R, et al. EWS and RE1-Silencing Transcription Factor Inhibit Neuronal Phenotype Development and Oncogenic Transformation in Ewing Sarcoma. *Genes Cancer*. 2013;4(5-6):213-223. doi:10.1177/1947601913489569

210. Flicek P, Amode MR, Barrell D, et al. Ensembl 2014. *Nucleic Acids Res*. 2014;42(Database issue):D749-55. doi:10.1093/nar/gkt1196

211. Morgan M, Obenchain V, Lang M, Thompson R. BiocParallel: Bioconductor facilities for parallel evaluation. 2016.

212. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi:10.1186/PREACCEPT-8897612761307401

213. Kolde R. pheatmap: Pretty Heatmaps. 2015.

214. Wickham H. *ggplot2: Elegant Graphics for Data Analysis (Use R!)*. New York: Springer; 2009.

215. Cleveland WS. Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting. *J Am Stat Assoc*. 1988;85:596-610.

216. Zhang Y, Liu T, Meyer C, et al. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol*. 2008;9(9):R137+. doi:10.1186/gb-2008-9-9-r137

217. Johnson KM, Mahler NR, Saund RS, et al. Role for the EWS domain of EWS/FLI

in binding GGAA-microsatellites required for Ewing sarcoma anchorage independent growth. *Proc Natl Acad Sci*. 2017:201701872. doi:10.1073/pnas.1701872114

218. Johnson KM, Mahler NR, Saund RS, et al. A novel role for the EWS domain of EWS/FLI in binding GGAA-microsatellites required for Ewing sarcoma anchorage independent growth. *Proc Natl Acad Sci*. 2017.

219. Hnisz D, Abraham BJ, Lee TI, et al. Super-enhancers in the control of cell identity and disease. *Cell*. 2013;155(4):934-947. doi:10.1016/j.cell.2013.09.053

220. Pott S, Lieb JD. What are super-enhancers? *Nat Genet*. 2015;47(1):8-12. doi:10.1038/ng.3167

221. Knight JC. Allele-specific gene expression uncovered. *Trends Genet*. 2004;20(3):113-116. doi:10.1016/j.tig.2004.01.001

222. Furey TS. ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions. *Nat Rev Genet*. 2012;13(12):840-852. doi:10.1038/nrg3306

223. Johnson KM, Taslim C, Saund RS, Lessnick SL. Identification of two types of GGAA-microsatellites and their roles in EWS/FLI binding and gene regulation in Ewing sarcoma. *PLoS One*. 2017;12(11):1DUMMY. doi:10.1371/journal.pone.0186275

224. Kasowski M, Grubert F, Heffelfinger C, et al. Variation in transcription factor binding among humans. *Science (80-)*. 2010;328(5975):232-235. doi:10.1126/science.1183621

225. Cavalli M, Pan G, Nord H, Wallén Arzt E, Wallerman O, Wadelius C. Allele-specific transcription factor binding in liver and cervix cells unveils many likely drivers of GWAS signals. *Genomics*. 2016;107(6):248-254. doi:10.1016/j.ygeno.2016.04.006

226. de Santiago I, Liu W, Yuan K, et al. BaalChIP: Bayesian analysis of allele-specific transcription factor binding in cancer genomes. *Genome Biol*. 2017;18(1). doi:10.1186/s13059-017-1165-7

227. Bailey SD, Virtanen C, Haibe-Kains B, Lupien M. ABC: A tool to identify SNVs causing allele-specific transcription factor binding from ChIP-Seq experiments. *Bioinformatics*. 2015;31(18):3057-3059. doi:10.1093/bioinformatics/btv321

228. Stewart E, Goshorn R, Bradley C, et al. Targeting the DNA repair pathway in Ewing sarcoma. *Cell Rep*. 2014;9(3):829-840. doi:10.1016/j.celrep.2014.09.028

229. Lee H-J, Yoon C, Schmidt B, et al. Combining PARP-1 Inhibition and Radiation

in Ewing Sarcoma Results in Lethal DNA Damage. *Mol Cancer Ther*. 2013;12(11):2591-2600. doi:10.1158/1535-7163.MCT-13-0338

230. Richardson CD, Ray GJ, DeWitt MA, Curie GL, Corn JE. Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. *Nat Biotechnol*. 2016;34(3):339-344. doi:10.1038/nbt.3481

231. Hyman AA, Weber CA, Jülicher F. Liquid-Liquid Phase Separation in Biology. *Annu Rev Cell Dev Biol*. 2014;30(1):39-58. doi:10.1146/annurev-cellbio-100913-013325

232. Brangwynne CP, Tompa P, Pappu R V. Polymer physics of intracellular phase transitions. *Nat Phys*. 2015;11(11):899-904. doi:10.1038/nphys3532

233. Mitrea DM, Kriwacki RW. Phase separation in biology; functional organization of a higher order. *Cell Commun Signal*. 2016;14(1):1. doi:10.1186/s12964-015-0125-7

234. Hnisz D, Shrinivas K, Young RA, Chakraborty AK, Sharp PA. A Phase Separation Model for Transcriptional Control. *Cell*. 2017;169(1):13-23. doi:10.1016/j.cell.2017.02.007

235. Patel A, Lee HO, Jawerth L, et al. A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. *Cell*. 2015;162(5):1066-1077. doi:10.1016/j.cell.2015.07.047

236. Aguzzi A, Altmeyer M. Phase Separation: Linking Cellular Compartmentalization to Disease. *Trends Cell Biol*. 2016;26(7):547-558. doi:10.1016/j.tcb.2016.03.004

237. Burke KA, Janke AM, Rhine CL, Fawzi NL. Residue-by-Residue View of In Vitro FUS Granules that Bind the C-Terminal Domain of RNA Polymerase II. *Mol Cell*. 2015;60(2):231-241. doi:10.1016/j.molcel.2015.09.006

238. Patel N, Black J, Chen X, et al. DNA methylation and gene expression profiling of ewing sarcoma primary tumors reveal genes that are potential targets of epigenetic inactivation. *Sarcoma*. 2012;2012. doi:10.1155/2012/498472

239. Furey TS. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nat Rev Genet*. 2012;13(12):840-852. doi:10.1038/nrg3306