INVESTIGATING THE MUSASHI-2 (MSI2) PROTEIN IN HEMATOPOIESIS

INVESTIGATING THE ROLE OF THE RNA-BINDING PROTEIN MUSASHI-2 (MSI2) IN NORMAL HEMATOPOIESIS AND LEUKEMIA

BY
NICHOLAS T. HOLZAPFEL, BSc (Honours)

A Thesis
Submitted to the School of Graduate Studies
In Partial Fulfillment of the Requirement
For the Degree
Doctorate of Philosophy

McMaster University

Descriptive Note

University of Toronto BACHELOR OF SCIENCE (2011) Toronto, Ontario (Honours Science)

TITLE:	Investigating the Role of the RNA-binding protein Musashi-2 (MSI2) in Normal Hematopoiesis and Leukemia
AUTHOR:	Nicholas T. Holzapfel
SUPERVISOR:	Dr. Kristin Hope
NUMBER OF PAGES:	198

Lay Abstract

The hematopoietic system is responsible for the production of billions of mature cells everyday. These mature cells are “differentiated”, meaning that they have gone through a process that has allowed them to become specialized to perform a very specific role. Throughout the process of differentiation, most functional cells lose their ability to proliferate. The continued production of these functional cells comes from a pool of rare, quiescent, hematopoietic stem cells (HSC). These cells maintain the production of mature cells throughout the lifetime of an organism. The Musashi-2 (MSI2) protein has been identified as a protein that is critical for the normal function of HSCs. By altering the levels of the MSI2, it is possible to greatly impair or enhance the activity of HSCs. Moreover, correlative studies implicate MSI2 as a contributor to aggressive Acute Myeloid Leukemia (AML), a disease that occurs when HSCs become dysregulated. Despite its important roles in normal and abnormal hematopoiesis, very little is known about how MSI2 functions and whether it actually has a functional role in AML. We set forth to identify mechanisms through which the MSI2 protein functions and to prove that MSI2 contributes to the maintenance of human AML.

We reveal that the MSI2 protein plays a critical role for the maintenance of human AML and identify novel pathways through which the protein functions. Importantly, MSI2 is known to interact with mRNA in order to alter post-transcriptional gene expression. We thoroughly characterize the RNA-binding characteristics of MSI2 and identify a plethora of MSI2 RNA targets. In an unbiased manner, we also identify a list of MSI2-protein interactors. We identify one MSI2 protein-binding partner,

Insulin-like growth factor 2 mRNA binding protein 2 (IGF2BP2) that is preferentially expressed in the most immature fraction of HSCs and is critical for the proper function of HSCs.

Abstract

Musashi-2 (MSI2), a member of the Musashi family of RNA-binding proteins, is thought to play a critical role in the maintenance of stem cell populations and in the formation of aggressive tumours. Multiple studies indicate that MSI2 plays an important role in the maintenance of hematopoietic stem cell (HSC) populations and recent studies in humans identify MSI2 as an independent prognostic factor for overall survival in patients with Acute Myeloid Leukemia (AML). Importantly, though correlative studies implicate MSI2 as a contributor to aggressive disease in human AML, no study to date has attempted to analyze the functional role of MSI2 in primary human AML samples. Furthermore, though MSI2 is critical for the maintenance of HSCs, the mechanisms through which MSI2 functions are unknown. The work presented in this thesis elucidates the biochemical mechanisms through which MSI2 functions and examines the functional role of MSI2 in human AML.

Using a lentiviral-mediated shRNA knockdown of MSI2, I demonstrate that MSI2 is critical for the maintenance of human AML. A loss of MSI2 greatly impairs the ability of AML samples to maintain disease in a xenotransplantation assay. MSI2 is an RNA binding protein that is thought to repress the translation of target mRNAs in the cytoplasm and prevent the maturation of microRNAs (miRNAs) in the nucleus. The targets of MSI2 are believed to be potent regulators of stem-ness and dysregulation of these targets could very well contribute to neoplastic transformation. Cross-linking immunoprecipitation followed by next generation sequencing (CLIP-Seq), revealed the RNA binding properties of MSI2 and the RNA

targets bound by MSI2. To identify novel MSI2 protein interactors, the MSI2 locus was endogenously tagged with the promiscuous biotin ligase BirA* and subjected to BioID analysis. When compared to appropriate controls, we were able to robustly identify proteins that associate with MSI2. The analysis of one of these protein binding partners, Insulin-like growth factor 2 mRNA binding protein 2 (IGF2BP2) reveals a critical role in the normal function of HSCs.

Acknowledgements

Thank you to all my friends and family for their ongoing support. Thank you for putting up with my absentmindedness and lack of social interactions over the last few years.

I would like to thank Kristin for her guidance and support throughout the last few years. She had much faith in me early on and gave me the opportunity to perform technically challenging experiments and chase novel ideas. She was always very approachable, kind, and quite supportive during all the frustrating moments when experiments were not working.

I would like to say a very special thank you to Dr. Brad Doble. He was instrumental in the development of the CRISPR/BirA* experiments. I did not even know about BirA* or CRISPR before Brad mentioned them to me during one of my first committee meetings. Brad was the first to suggest the creation of an endogenous MSI2-BirA* and was also the one who suggested the use of a P2A control. I really consider the CRISPR-BirA* experiments to be the finest work out of my PhD. None of this would have been possible without Brad's guidance and support.

A special thank you to Vickie Kwan for all the fun memories, laughter, blatant honesty, and words of support throughout the last few years. My PhD experience wouldn't have been the same without you. Thanks for spending countless hours reading through my thesis.

Table of Contents

Lists of Figures and Tables

Chapter1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of Abbreviations and Symbols

3'UTR	3' untranslated region
AHR	Aryl hydrocarbon receptor
ALL	Acute lymphoblastic leukemia
AML	Acute myelogenous leukemia
ASCT	Allogenic stem cell transplantation
ATRA	All-trans retinoic acid
BWT	Bhurrows-Wheeler Transform
Cas	CRISPR-associated
Cas9n	Cas9 nickase
cDNA	Complementary deoxyribonucleic acid
CFU	Colony forming unit
ChIP-Seq	Chromatin immunoprecipitation followed by sequencing
CLIP-Seq	Cross-linking immunoprecipitation followed by sequencing
CLP	Common lymphoid progenitor
CML	Chronic myeloid leukemia
CMP	Common myeloid progenitor
CNS	Central nervous system
CRISPR	Clustered-regularly interspaced short palindromic repeats
crRNA	CRISPR RNA
cRNP	CRISP-ribonuclease complex
CYP1B1	Cytochrome P450 Family 1 Subfamily B Member 1
DMEM	Dulbecco's Modified Eagle Medium
DMR	Differentially methylated region
DN	Dominant negative
eCLIP	enhanced-CLIP
ELN	European Leukemia Net
EMSA	Electrophoretic mobility shift assay
ESC	Embryonic stem cells
FBS	Fetal bovine serum
FDR	False discovery rate
GAR	glycine and arigine rich
G-CSF	Granulocyte colony stimulating factor
GFP	Green fluorescent protein
HEK293	Human embryonic kidney 293 cells
hiDAC	High dose cytarabine
HITS-CLIP	High-throughput sequencing cross-linking immunoprecipitation
HSC	Hematopoietic stem cells
HSCT	Hematopoietic stem cell transplantation
HSPC	Hematopoietic stem and progenitor cells
IGF1R	Insulin-like growth factor 1 receptor
IGF2BP2	Insulin-like growth factor 2 mRNA binding protein 2
IP	Immunoprecipitation

IRIC	Institute for research in immunology and cancer
KH	K-homology
LSC	Leukemic stem cell
LSK	$Lin^{-}Sca\text{-}1^{+}c\text{-}Kit^{+}$
LT-HSC	Long term hematopoietic stem cell
MNase	Micrococcal Nuclease
MPP	Multipotent progenitor
mRNA	Messenger ribonucleic acid
MSI1	Musashi-1
MSI2	Musashi-2
NEAA	Non-essential amino acids
NGS	Next generation sequencing
NMR	Nuclear magnetic resonance
NSG	NOD scid gamma
PAM	Protospacer adjacent motifs
PCR	Polymerase chain reaction
PRMT	Protein methyltransferase
PTM	Post-translation modification
PyPI	Python package index
qPCR	Quantitative polymerase chain reaction
RBD	RNA-binding domain
RBP	RNA-binding protein
RIPA	Radioimmunoprecipitation assay buffer
RIP	RNA-Immunoprecipitation
RNA	Ribonucleic acid
RNP	Ribonucleoprotein particle
RRM	RNA recognition motif
SCID	Severe combined immunodeficiency
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
shRNA	Short hairpin ribonucleic acid
SOP	Sensory organ precursor
SRC	SCID-repopulating cell
STAR	Spliced Transcript Alignment to a Reference
ST-HSC	Short term hematopoietic stem cell
STMN1	Stathmin-1
SUMO	Small ubiquitin-like modification
TBE	Tris/Borate/EDTA
tracrRNA	Trans-activating CRISPR RNA
TSPAN3	Tetraspanin-3
UV	Ulta-violet
WBM	Whole bone marrow
WHO	World health organization

Declaration of Academic Achievement

I performed all of the experiments highlighted in this thesis except for the following:

Chapter 2: NB4/ATRA experiments were performed by Muluken Belew (Hope Lab)

Chapter 3: Gabriel Pratt from the Yeo lab at the University of California San Diego performed bioinformatic analysis on CLIP libraries

Chapter 4: Derek Chan (Hope Lab) designed IGF2BP2 shRNAs and performed transplantation assays. The proteomics core at the Institute for Research in Immunology and Cancer at the University of Montreal performed Mass spectrometry analysis on the MSI2-BirA* samples

Chapter 1: Introduction

1.1 Early Studies in Hematopoiesis

1.1.1 Identification and characterization of hematopoietic repopulating cells

In the late 19th century, early clinicians and scientists debated the existence of a common hematopoietic stem cell that could give rise to all cell types of the blood. In 1909, hematologist Alexander A. Maximow proposed the Unitarian Hematopoietic Theory, which stated that all blood cells originate from a common hematopoietic stem cell[1,2]. Ernst Neumann, the pathologist who had previously demonstrated that the bone marrow was the site of hematopoiesis in mammals, supported this view[3]. Opponents of this theory, most notably, the physician Paul Ehrlich, believed in a 'Dualist' theory of hematopoiesis where bloods cells were maintained by two distinct hematopoietic populations, one existing in the lymph nodes that gave rise to lymphocytes and the other existing in the bone marrow that gave rise to leukocytes[3].

The dawn of the nuclear age gave birth to the field of hematopoietic stem cell transplantations; out of the devastation of Hiroshima and Nagasaki came new insights into the biological effects of radiation[4]. The observations of severe myelosupression in survivors of the atomic blasts ultimately led to the development of hematopoietic stem cell transplantation, a procedure that has saved countless lives and remains the pillar of stem cell therapy. Studies pioneered in the 1950's demonstrated that cells within the bone marrow could restore hematopoiesis after lethal irradiation in mice[5]. These studies showed that intravenously injected bone

marrow cells could rescue irradiated mice from lethality, strongly suggesting that stem-like cells existed in the bone marrow. Ultimate proof for the existence of multipotent HSCs came from the seminal work of Drs. James Till and Ernest McCulloch in the 1960's[6]. Till and McCulloch demonstrated that mouse bone marrow cells could be transplanted into irradiated recipients and form discrete myeloerythroid colonies in the spleen; the cell of origin for these spleen nodules was referred to as the colony forming unit-spleen or CFU-S. They demonstrated that 1 in 3000 bone marrow cells had spleen colony forming activity when intravenously injected into irradiated mice[7]. Importantly, these colonies were clonally derived and contained cells from various hematopoietic lineages proving the existence of multipotent hematopoietic cells[8]. Furthermore, when transplanted into secondary irradiated recipients, some of these nodules contained cells that could produce additional nodules. This demonstrated that certain CFU-S cells could "self-renew", a process that is now a defining feature of stem cells[9].

The pioneering work by Till and McCulloch gave rise to the concept that CFU-S were multipotent and in some cases, were able to self-renew. Importantly, CFU-S gave rise to myeloid and erythrocyte lineages but no lymphoid cells. Further experiments by Till and McCulloch used irradiated donor cells in order to genetically mark individual cells[10]. Transplantation assays demonstrated that single cells existed that could give rise to all myeloerythroid cells within a spleen nodule and could also generate lymphocytes supporting the concept that a single hematopoietic stem cell capable of giving rise to both myeloid and lymphoid cells existed. Over the next few decades, more complex hematopoietic transplantation

assays were developed allowing for the identification of a complex hematopoietic hierarchy.

1.1.2 Prospective Isolation of Murine Hematopoietic Cells

Importantly, although the pioneering assays by Till and McCulloch elegantly proved the existence of hematopoietic stem cells, they were unable to isolate these cells for further study. It was not until the advent of multiparametric flow cytometry and the development of monoclonal antibodies, that scientists were able to identify sets of cell surface markers that could enrich for specific hematopoietic cell populations.

The development of complex transplantation assays was critical for the identification of murine hematopoietic hierarchies. Boyse and colleagues developed mouse strains in the C57BL/6 background congenic for 2 alleles of the pan-hematopoietic cell surface marker CD45[11]. This allowed for the discrimination between host and donor cells following a transplantation of mouse hematopoietic stem and progenitor cells. Further advances in the search for murine stem cells came from the Weissman laboratory in the early 1980's. Importantly, while studying the clonogenic activity of B-Cells, researchers were able to show that B-Cell precursors lacked the canonical marker, B220[12]. This discovery led researchers to infer that progenitor cells might lack all markers of differentiated hematopoietic cells, which ultimately led to the development of a cocktail of antibodies that could divide the hematopoietic system into Lineage depleted (Lin^-) and Lineage positive (Lin^+) fractions[12]. Remarkably, it was shown that the Lin^- fraction contained all of

the clonogenic progenitors and greatly enriched for reconstituting cells. Cell surface labeling followed by multiparametric flow cytometry quickly took hold and greatly facilitated the identification of hematopoietic stem cell and progenitor populations.

Almost all murine HSC purification schemes revolve around the positive selection for c-Kit and Sca-1 and negative selection of lineage markers (B220, CD4, CD8, Gr-1, Mac-1, and Ter-119)[4]. Studies from the Weissman lab were able to demonstrate enrichment for radioprotective cells in $Lin^{-}Thy1^{lo}Sca1^{+}$ populations[13]. This population represents 0.05% of whole bone marrow (WBM) and when injected intravenously into lethally irradiated mice, 1 in 22 cells give rise to multilineage constitution. Transplantation assays carried out by the Weissman lab displayed interesting patterns of reconstitution; mice engrafted with $Lin^{-}Thy1^{lo}Sca1^{+}$ cells would either show myeloid and B-cell reconstitution by 6 weeks or they would display a transient multilineage reconstitution in which levels of donor cells would decline after 4 weeks and be undetectable after 8-10 weeks[14,15]. These studies demonstrated that this population was a mix of long-term HSCs (LT-HSCs), short-term HSCs (ST-HSCs), and multipotent progenitors (MPPs). Further purification schemes were developed allowing for a more intricate characterization of the murine hematopoietic hierarchy. Researchers in the Weissman lab discovered that cells committed to lymphoid fates resided in the $IL7R\alpha^{+}$ fraction of $Lin^{-}Thy1^{lo}Sca^{lo}cKit^{lo}$ fraction[16]. This population could rapidly give rise to lymphoid restricted cells but lacked all myeloid potential. Later, common myeloid progenitors were identified as being $FcyR^{lo}CD34^{+}IL\text{-}7R\alpha^{-}Lin^{-}cKit^{+}Sca1^{-}$[17]. Ultimately, these studies resulted in a model of hematopoiesis where largely quiescent long-term

HSCs gave rise to short-term HSCs and eventually multipotent progenitors. These multipotent progenitors were thought to differentiate into either committed progenitors capable of producing all myeloid cells of the blood (CMP) or into committed progenitors capable of producing all lymphoid cells of the blood (CLP). CMPs would give rise to erythroid cells, megakaryoctyes, dendritic cells, and granulocytes, whereas CLPs would give rise to the B and T cells[16,17].

Eventually, through the use of LSK and the SLAM family of markers (CD150, CD48, CD229, and CD244), researchers were able to distinguish multiple subsets of functionally distinct HSCs and MPPs to a high degree of purity[18]. Murine HSCs are usually characterized as $CD150^{+}CD48^{-}$LSK but can be further sub-divided into more quiescent and functionally distinct subpopulations[19,20]. Within this population, LT-HSCs sit at the apex of the hematopoietic hierarchy and possibly maintain hematopoiesis through asymmetric cell division. Functionally, LT-HSCs are defined as those cells that can give rise to long-term multilineage engraftment >3 months post transplant. ST-HSCs also exist within the $CD150^{+}CD48^{-}$ population and can give rise to multilineage engraftment upon transplantation but with a diminished self-renewal capacity when compared to LT-HSCs. Importantly ST-HSCs can be distinguished from MPPs *via* their enhanced duration and magnitude of repopulation. In contrast to LT- and ST-HSCs, MPPs do not possess a significant capacity for self-renewal and only show transient repopulation lasting no more than 4-6 weeks (Figure 1, p.73)[18].

Recent studies have shed light on the functional heterogeneity found within MPP, CLP, and CMP populations. Notably, by tracking individual MPPs, researchers

have revealed that some appear committed to specific blood lineages[21]. MPPs were initially characterized as a homogenous population of cells capable of transient repopulating and multilineage engraftment. Recently, an abundant subset of MPPs was identified as $FLK2^{+}LSK$[22]. These cells represent 60% of LSK cells and are considered to be fully multipotent but with a lymphoid bias[21]. Two other subsets of MPPs have recently been identified in the $FLK2^{-}$ LSK compartment. $FLK2^{-}CD150^{+}$ $CD48^{+}$ LSK and $FLK2^{-}CD150^{-}CD48^{+}$ LSK are both myeloid biased MPP subsets[20]. These studies highlight the complex nature of the hematopoietic hierarchy. MPPs are not a homogenous progenitor pool with multilineage differentiation potential; instead, the MPP population consists of a heterogeneous pool of progenitors that possess a pre-determined lineage bias. Furthermore, studies indicate that human HSCs also function in heterogeneous manner. Clonal tracking of human hematopoietic stem cells indicate that clonal fluctuations occur upon hematopoietic repopulation. Human HSCs have been demonstrated to contribute to hematopoiesis for a short period of time then exhaust. Furthermore, clones have been identified that remain dormant for a long period of time until becoming active. Interestingly, other clones have been identified that contribute to hematopoiesis, become dormant, then re-contribute to hematopoiesis at a later period of time. Importantly, studies suggest that the varied proliferation and self-renewal capacities of human HSCs are likely governed by unpredictable "stochastic" factors and are not determined inherently by the cell. Importantly, mouse models of human hematopoiesis have elegantly demonstrated that physically separated HSC daughter

cells behave in an unpredictable manner supporting a "stochastic" model of HSC function[23].

1.1.3 Immuncompromised Mouse Models for the Study of Human Hematopoiesis

The development of immunocompromised mouse models was critical for understanding human hematopoietic biology and for the prospective isolation of human hematopoietic stem and progenitor cells. In 1983, a seminal paper was published in *Nature* describing a mutation in mice that resulted in severe combined immunodeficiency (SCID)[24]. This recessive mutation occurring on chromosome 16 resulted in mice that lacked all B and T cells. Early pioneering experiments showed that human lymphoid cells could survive in SCID mice either after the intra-peritoneal injection of peripheral blood leukocytes or after the generation of humanized SCID (SCID-hu) mice[25,26]. Here human fetal thymus, liver, lymph node, and spleen were introduced surgically and resulted in mice that did not succumb to opportunistic infections, tolerized human and mouse tissues, and developed B and T cells. These mice developed a human thymus-like organ that could export mature human T-cells to the mouse circulation thus serving as a critical model for the study of human T-cell development. Though these mouse models allowed for unprecedented studies of human hematopoietic biology, they were ill-suited for the study of human HSC function. Notably, the aforementioned mouse models allowed for human lymphoid cell engraftment without human myeloid engraftment. Since lymphoid cells are long-lived and possess the ability to proliferate, these models were unable to assay HSC function. A groundbreaking advancement emerged out of

the lab of Dr. John Dick when researchers transplanted human bone marrow cells into immune-deficient bg/mu/xid mice that were continuously infused with the human myeloid growth factors interleukin-3 (IL-3) and granulocyte-macrophage colony-stimulating factor (GM-CSF)[27]. This technique allowed for the engraftment of human myeloid cells and thus allowed researchers to assay human HSC function. Additional transplantation assays were developed where SCID mice were transplanted with human bone marrow and ip-injected with cocktails of human growth factors allowing for the repopulation of the murine hematopoietic system with human myeloid, lymphoid, and erythroid cell lineages[28]. One main caveat of the SCID mouse model was the presence of an innate immune system; therefore further improvements to the model were developed. First, the SCID mouse was backcrossed onto the Non-Obese Diabetic (NOD) mouse, which had defects in innate immunity; this mouse could support higher levels of human engraftment[29]. Importantly, a polymorphism in the signal-regulatory protein alpha (SIRPa) gene allows NOD Sirpa to bind human CD47 preventing the destruction of human cells by NOD macrophages. In other mouse background, the inability of Sirpa to recognize human CD47 greatly impairs the repopulation of human cells[30]. However, limitations to the NOD-SCID model yet lingered. NOD-SCID mice still had NK cells that were able to resist engraftment and were prone to the development of thymic lymphomas, which greatly impaired long-term studies of human hematopoiesis. To improve upon this mouse model, NOG and NSG mice were created[31,32]. NOG mice resulted from NOD-SCID mice that had a truncation in the IL-2R common γ chain while NSG had a deletion in the γ common chain. These mice were demonstrated to have a complete

lack of B, T, and NK cells as well as severe defects in innate immunity permitting much more enhanced levels of human engraftment. Currently the vast majority of xenotransplantation assays are performed in NSG or NOG mice. Importantly, levels of CD34+ engraftment are 5-fold higher in NSG mice compared to NOD-SCID and defects in cytokine signaling prevent the formation of lymphomas allowing for the long-term analysis of human grafts[33].

1.1.4 Prospective Isolation of Human Hematopoietic Stem and Progenitor Cells

In 1984, a group of researchers was able to raise a mouse monoclonal antibody, anti-My-10, against the KG-1a human myeloid leukemia cell line that did not react with mature human granulocytes[34]. It was shown that anti-My-10 was specifically expressed on immature human marrow cells; anti-My-10 was later identified as targeting CD34[35]. $CD34^+$ cells were shown to reconstitute hematopoiesis in lethally irradiated baboons and in human autologous stem cell grafts[36,37]. These observations and numerous *in vitro* assays showed an association of immature hematopoietic cells with $CD34^+$ expression[34,38,39]. Further studies demonstrated that the $CD34^+$ fraction of bone marrow and cord blood cells was highly heterogeneous and further fractionation into the $Lin^-CD34^+CD38^-$ cells could isolate a rare population of cells that possessed sustained clonogenicity and extended long-term culture capabilities[40-42]. At the same time, the Dick lab demonstrated that intravenous transplantation of bone marrow and human cord blood cells into sub-lethally irradiated SCID mice could result in high levels of multilineage engraftment, including myeloid and lymphoid lineages; the engrafting

cells were defined SCID-repopulating cells (SRCs)[43]. Ultimately elegant studies out of the Dick lab demonstrated that fractionation of human bone marrow and cord blood on the combination of $Lin^{-}CD34^{+}CD38^{-}$ resulted in a 1500-fold increase in SRCs[44].

Further characterization of the human hematopoietic hierarchy continued and human counterparts of the CLP, CMP, GMP, and MEP were discovered[45-50] (Figure 2, p.74). Additionally, analysis of human multipotent progenitor populations revealed the existence of a lymphoid primed multipotent progenitor. This progenitor population was able to give rise to both myeloid and lymphoid cells but did not produce any megakaryocyte or erythroid cells suggesting that in a similar manner to the mouse system, the MPP population in humans is also heterogenous. Additionally, populations of multi-lymphoid progenitors (MLP) have been identified that are biased to, but not restricted to, the generation of lymphoid cells. MLPs give rise to B cells, T cells, NK cells, and macrophages but lack granulocyte and erythrocyte potential[48]. Lastly, human HSCs have recently been identified based on the expression of CD49f. Partitioning of the human HSC population on CD49f greatly enhances the chimerism and long-term repopulation of engrafted cells[49]. Interestingly, practically all long-term engrafting cells are contained within the CD49f+ fraction. The development of xenotransplantation assays was critical for the identification and description of the normal human hematopoietic hierarchy. Furthermore, the application of the xenotransplantation assay to the study of acute myelogenous leukemia (AML) has greatly enhanced our understanding of this disease.

1.2 Acute Myelogenous Leukemia & Abnormal Hematopoiesis

1.2.1 Brief History of Leukemia

AML is a heterogeneous disease characterized by the overproduction of immature blast cells that crowd the bone marrow and peripheral blood leading to bleeding, infection, and bone marrow failure. The first published clinical description of leukemia dates to 1827 by the French surgeon Alfred Velpeau[51]. In a case report, Velpeau noted a 63-year old patient presenting with abdominal swelling, fever, weakness, and urinary stones. Over a 9-year period the patient had faced recurrent fever and inflammation eventually succumbing to his disease. Upon autopsy, Velpeau noted a grossly enlarged liver and spleen along with an altered blood composition; the blood was described to be thick as gruel and filled with pus[51]. Velpeau had unknowingly published one of the first clinical descriptions of leukemia. It was the German pathologist Rudolf Ludwig Karl Virchow who would be the first to microscopically investigate leukemia. Virchow, who had pioneered the use of the light microscope, was able to elaborate on the clinical condition first described by Velpeau[51]. He noted excessive white blood cell production in these patients with a decrease in the amount of red blood cell corpuscles. He further elaborated on the condition of the white cells by stating that they were not completely regular. Notably, it was Virchow who coined the word 'Leukemia', Greek for 'white blood' in 1847[51]. In 1877 Paul Ehrlich developed novel staining techniques that could discern the characteristics of cells within dry blood films. Using this staining technique in combination with light microscopy, Ehrlich was able

to classify leukemia into two categories: the granulocytic myeloid group and the non-granular lymphoid group[52]. In 1900, the Swiss physician Otto Naegli supported Ehrlich's view when he described two distinct cellular entities: the lymphoblast and the myeloblast. Naegli described myeloblasts as being ancestors of granulocytes and lymphoblasts as being ancestors to lymphocytes. The presence of these 'blast' cells in the peripheral blood soon became a critical feature in the diagnosis of leukemia. In 1889, Ebstein introduced the term 'acute leukemia' in to describe the difference between rapidly fatal leukemia and chronic indolent leukemia[53].

1.2.2 Etiology and Diagnosis of AML

Acute myelogenous leukemia is a group of neoplasms that are thought to arise from either a transformed normal HSC or due to mutations in a more downstream progenitor that gives it the capacity to self-renew[54]. The transformation of these cells is thought to be a multistep process. Ultimately, these mutations result in the clonal proliferation of myeloid precursors in the bone marrow and prevent the full maturation of these cells. Within the bone marrow, immature blasts can crowd and impair normal hematopoietic cells resulting in severe anemia and leukopenia[55]. Organ infiltration can also result in the impaired function of that organ resulting in a variety of clinical presentations. Clinically, AML is diagnosed when a blast count >20% is seen in the bone marrow and the blasts can be identified as having a myeloid phenotype[55]. AML has a median age of diagnosis of 65 years and is the most common acute leukemia in adults with an incidence of 3 to 5 cases per 100, 000. AML accounts for less than 10% of acute leukemias in children.

AML remains an aggressive disease that is difficult to cure. The overall survival rate in young patients (<60 years) is 30% and drops to a dismal 5-15% in older adults[56].

The etiology of AML is complex with numerous factors contributing to the development of disease. Genetics, toxic exposures, radiation, and advanced age, all contribute to the development of disease. Inherited disorders including Down syndrome, Fanconi anemia, and neurofibromatosis predispose individuals to AML as well[57]**.** Notably, children with Down syndrome have a 20-30-fold increased likelihood of developing leukemia[58]. Numerous polymorphisms and germ-line mutations have been associated with the development of AML[57]. Additionally, chemicals such as benzene, alkylating agents, smoking, and exposure to radiation are associated with increased risks of AML. Ironically, exposure to anticancer drugs has long been known to increase the risk of AML; 10-15% of AML patients develop the disease after treatment with standard chemotherapy agents[57].

1.2.3 Genetic and Cytogenetic Abnormalities in AML

The genetic mutations and cytogenetic abnormalities underlying AML vary tremendously. A majority of AMLs are associated with non-random chromosomal translocations often resulting in the production of fusion proteins[59]. Notably, rearrangements involving loci encoding transcription factors are the most prevalent. Recurrent cytogenetic abnormalities have been identified across a variety of AML samples and are strongly associated with clinical subgroups and therapeutic outcome. Importantly, the cytogenetic analysis of metaphase chromosomes is a key first step in the evaluation of all patients diagnosed with AML[57]. Cytogenetic

abnormalities are found in 50 to 60% of newly diagnosed AML and remain the most important factor in predicting rates of relapse, remission, and overall survival[57]. The most common include translocations between chromosomes 8 and 21 (AML1-ETO), translocations between chromosome 15 and 17 (PML-RARa), inversions in chromosome 16 (CBF-MYH11), and abnormalities of 11q23 (MLL). In 1998, the Medical Research Council (MRC) developed a classification system that stratifies patients into 3 groups based on cytogenetic analysis: (1) those with a favourable outcome, (2) those with an intermediate outcome, and (3) those with a poor outcome[60] (Figure 3, p.75). Common cytogenetic changes found in patients with a favourable outcome include t(15;17), t(8;21), and inv(16). Patients lacking the aforementioned changes but possessing abnormalities in 3q (abn(3q)), deletions in 5q (del(5q)), monosomies of chromosome 5 and/ or 7 (-5/-7), or complex karyotypes (5 or more unrelated cytogenetic abnormalities) were classified as having a poor outcome[60]. The remaining patients were classified as having an intermediate outcome. Rates of overall survival for favourable-, intermediate-, and poor-outcome groups are estimated at 55, 24, and 5% respectively[61]. Approximately 40% of patients with AML have a normal karyotype and are therefore categorized as being in the intermediate risk subgroup; this represents a large and quite heterogeneous fraction of AML patients. It is thought that the heterogeneity within this group occurs due to variations in genomic mutations and given that the majority of AML patients fall within an intermediate subgroup, there is a need to further characterize the molecular changes that occur in AML[55]. Importantly, new

patient classification systems have begun to incorporate molecular biomarkers such as FLT3, NPM1, CEBPA, and cKIT[59].

In 2010, the European Leukemia Net (ELN) recommended that three molecular markers be used in clinical practice to further classify patients with cytogenetically normal AML into high, medium, and low risk groups. These markers are: NPM1, FLT3-internal tandem duplications, and CEBPA[55]. Here, patients with normal karyotypes and mutations in CEBPA or mutations in NPM1 without FLT3-ITD are classified as having a favourable outcome. Those with a normal karyotype but mutated NPM1 and FLT3-ITD, or mutated CEBPA with FLT3-ITD are categorized as having an intermediate risk. Those with normal cytogenetics but mutations in FLT3 are categorized as high risk (Figure 4, p.76). The effectiveness of this classification system has yet to be proven in numerous large-scale clinical studies. There is currently a large drive to understand the molecular landscape of AML with the hope that this will provide a better stratification of risk and ultimately more robust therapies.

Somatic mutations are found in 50-80% of AML cases. These mutations are broadly grouped into two categories: those that activate signal transduction (class I) and those that affect transcription factors or components of the cell cycle machinery (class II). Important class I mutations include mutations in KIT, FLT3, and NRAS. Important class II mutations include mutations in MLL, WT-1, CEBPa, DNMT3A, and NPM1. Recently the Cancer Genome Atlas Research Network was able to sequence 200 patient AML samples; 150 patient samples were sequenced through exome sequencing and 50 were sequenced through whole genome sequencing[62].

Surprisingly, AML samples were shown to have a limited number of genomic mutations; researchers observed that the average AML sample had 13 genic mutations with an average of 5 mutations occurring in genes that are recurrently mutated[62]. Commonly mutated genes included FLT3, NPM1, DNMT3A, IDH2, IDH1, TET2, and RUNX1. It is the hope that ongoing studies will allow clinicians and scientists to better understand the development of AML and that the identification of novel somatic mutations will help to further stratify patients based on prognosis and ultimately allow for the development of targeted therapy. One recent paper published in the New England Journal of Medicine highlight the importance of genomic classification of AML and its potential impact on patient prognosis[63]. In the analysis of 1540 AML patients, 5234 driver mutations were identified occurring across 76 genomic regions. Interestingly, based on genomic lesions (most of which were point mutations), 3 genomic categories could be identified: 1) those with mutations in genes for splicing and chromatin, 2) those with TP53 mutations, and 3) those with mutations in IDH2. Importantly, there were considerable differences in clinical presentation and overall survival across the genomic subgroups. Specifically, those in the chromatin-spliceosome subgroup tended to have poor overall survival, a low response to induction chemotherapy, and higher rates of relapse. Despite this, 84% of these patients would have been classified as being in the intermediate risk subgroup based on the current classification system even though their outcomes were similar to those patients classified as adverse risk. This highlights the important need to fully study the molecular and cytogenetic factors underlying the development of AML in order to better predict patient outcomes.

1.2.4 WHO and FAB Classification System

The current categorization of AML uses the World Health Organization (WHO) classification system to stratify AML into 4 different categories based on clinical features, and molecular and cytogenetic abnormalities[64] (Figure 5, p.77). This system was introduced in 1999 and replaced the French-American-British (FAB) classification system that had been used previously**.** The FAB system was introduced in 1976 and relied on morphology and cytochemistry to group AML into one of 8 subtypes[22]. The assumption here was that the morphological features of AML largely reflected the underlying genetic dysfunctions. Based on this assumption, morphologically similar AMLs could be grouped together with the belief that they would have similar prognosis and treatment strategies. The WHO classification system was introduced in 1999 in order to better stratify AML largely due to the newfound appreciation for the role of cytogenetic and molecular abnormalities in the development of AML[57]. The ultimate goal of the WHO classification system was to stratify AML samples into clinically unique diseases through the correlation of morphologic, genetic, and clinical data with the belief that this would aid clinicians in predicting patient prognosis and selecting an appropriate course of treatment.

According to the WHO classification system, AML is diagnosed when 20% of the cells in the bone marrow consist of immature blasts or when a patient presents with specific cytogenetic abnormalities[59]. Four AML subgroups are recognized by the WHO classification. They include (1) AML with recurrent genetic abnormalities,

(2) AML with multilineage dysplasia, and (3) therapy-related AML. Any case that cannot fit into one of the first three groups is classified into a fourth group (4) AML, not otherwise categorized[57]. There are seven recurring genetic abnormalities that are recognized by the WHO criteria, 3 of which are diagnostic for AML. A unique category recognized by the WHO is AML with myelodysplastic-related changes[57]; this category arose from the belief that dysplasia may relate to patient prognosis. Patients who have AML with accompanying dysplastic changes in 2 or more hematopoietic lineages or a prior history of myelodysplastic syndrome (MDS)- characterized by the production of dysfunctional blood cells belonging to one or more lineages- fall into this category and tend to have a poor prognosis. Regardless of the mutation or cytogenetic changes, AML is classified as a therapy-related AML when a patient's history reveals prior exposure to cytotoxic therapy. Therapy-related AML is classified based on the causative agent: (1) alkylating agent/ radiation-related or (2) topoisomerase II inhibitor-related. Therapy-related AML outcomes tend to be much worse when compared to *de novo* leukemia[57].

1.2.5 Treatment

The treatment regimen for individuals with AML varies dramatically and is dependent upon numerous factors such as the patient's age, health, cytogenetic factors, and molecular abnormalities[65]. Despite intensive research in the field, the treatment for AML has not varied dramatically in over 30 years. Healthy patients less than 60 years of age typically undergo two phases of therapy[65]. The first phase, referred to as 'induction therapy', is required to rapidly restore normal

hematopoietic function. The goal of induction therapy is to reduce the amount of leukemic blasts in the bone marrow to below cytologically detectable levels; at this point a patient is said to be in remission. Despite being in remission, it is recognized that most patients harbor a certain level of undetectable leukemic cells, referred to as 'minimal residual disease'[65]. If these residual cells are not targeted appropriately, almost all patients will undergo relapse. Therefore, the second phase of AML treatment aims to target residual AML cells; this treatment is greatly dependent on patient age, cytogenetic risk, and the presence of co-morbidities[65,66]. Overall, allogeneic stem cell transplantation is associated with the highest rates of disease-free survival but is associated with numerous co-morbidities, especially in older patients[65,66]. All clinicians must weigh the risks and benefits of ASCT in deciding upon an appropriate course of consolidation therapy.

The overall approach for leukemic patients is to treat the initial disease in order to achieve remission. Once remission is achieved the clinician must decide upon the best strategy to prevent relapse. Young patients typically go through a standard "7+3" induction regimen involving the administration of cytarabine for 7 days with the concomitant administration of an anthracycline for 3 days[65,66]. Cytarabine is a commonly used chemotherapy drug that inhibits DNA synthesis. The anthracyclines are a group of chemotherapy drugs that inhibit cancer growth through inhibition of DNA synthesis, inhibition of topoisomerase II, and the generation of cell-damaging reactive oxygen species (ROS). After this initial treatment, the bone marrow is evaluated for signs of remission and if necessary induction therapy may be repeated. Older patients often experience more

treatment-related morbidities and therefore induction therapy must be tailored based on the patient[66]. Healthy patients may receive standard cytarabine therapy or therapy with a slightly reduced dose, while others may be placed in clinical trials. Patients with serious co-morbidities are often referred to supportive care. Once remission is achieved, a clinician must decide on the best course of action to prevent relapse[66]. Typical post-remission therapies include consolidation chemotherapy, autologous stem cell transplantation, or allogeneic stem cell transplantation. Consolidation chemotherapy typically has a low treatment-related mortality but rates of relapse vary tremendously between patients (largely depending on cytogenetic factors)[66]. Autologous stem cell transplantation also has a low treatment-related mortality and is not associated with graft vs. host disease but relapse rates are typically high due to the lack of graft vs. leukemia effect and the presence of minimal residual disease[65,66]. Allogeneic stem cell transplantation (ASCT) is the gold standard for preventing disease relapse. ASCT destroys leukemic cells with the use of myeloablative therapy and through a graft vs. leukemia effect. Unfortunately, ASCT is associated with high rates of treatment-related mortality and is associated with graft vs. host disease[66]. For young patients with favourable risk, consolidation chemotherapy with high dose cytarabine (HiDAC) is the standard therapy[66]; this group of patients typically demonstrates low rates of relapse post-consolidation. Young patients with poor prognostic indicators, typically relapse despite consolidation chemotherapy. Therefore, patients in this group are typically referred to for heamtopoetic stem cell transplantation upon their first remission. The best course of treatment for young patients with an intermediate risk is still

disputed. Some clinicians prefer to treat with HiDAC upon the first remission and only refer for stem cell transplantation upon relapse[66], whereas others refer patients for stem cell transplantation immediately upon remission.

Post-remission therapy for older adults is complicated due to many confounding factors[66]. Older patients typically have numerous comorbidities that limit the extent of therapy and typically present with more aggressive disease. Older patients with favourable cytogenetics may undergo consolidation chemotherapy with a lower dose of cytarabine. For patients with adverse cytogenetics, ASCT offers better anti-leukemic effects. Older patients are typically not able to tolerate fully myeloablative HSCT but may be eligible for non-myeloablative HSCT[66]. Often times there is no good therapy option for older adults and the best course of action may be to enroll the patients into a well-designed clinical trial. Resultantly, there is a need to better understand the molecular mechanisms that underlie the development and maintenance of leukemia in order to identify more therapeutic targets. AML is a complex and heterogeneous disease but the identification of different mechanisms that support the survival of these cells may lead to the treatment of individual AMLs based on their unique molecular profiles.

1.2.6 Xenotransplantation Assays for AML

A better understanding of the pathogenesis of AML and the development of novel therapies are required to improve rates of disease-free survival. Murine xenotransplantation assays remain the gold standard for studying the biology of human AML due to their ability to identify and characterize leukemic stem cells

(LSCs). LSCs are thought to be a small population of leukemia cells with stem cell-like properties that are able to maintain leukemic growth. Xenotransplantation assays have been critical for the identification and characterization of LSCs and continue to play a critical role in developing novel anti-leukemic therapies.

It had been previously accepted that the vast majority of leukemic cells were non-proliferative when grown in culture or semi-solid media[67]. AML-CFU assays were generally used to assay AML function in culture; these assays suggested that in a similar manner to normal hematopoiesis, leukemic cells were maintained by a small fraction of highly proliferative cells that gave rise to large numbers of non-functional and partially differentiated blast cells. These assays demonstrated that only 1% of AML give rise to AML-CFU colonies[68]. AML-CFU however, only possessed limited self-renewal capabilities and thus a better assay was required to further characterize the cells capable of maintaining leukemic growth in vivo. In 1994, Lapidot et al. described the successful engraftment of human AML cells after transplantation into SCID mice allowing for the identification of LSCs[69]. Using a limiting dilution assay, they showed that SCID leukemia initiating cells (SL-ICs) were present in the peripheral blood of patients at a frequency of 1 in 250 000 cells. Further fractionation revealed that SL-ICs had a unique CD34+CD38- surface profile as it was demonstrated that only AML cells contained in this subset were able to engraft and give rise to AML-CFU progenitors. In the samples tested in this study, CD34+CD38+ cells and CD34- cells lacked these properties[69]. Lastly, the frequency of SL-ICs was demonstrated to be 1000-fold lower than the frequency of AML-CFU suggesting that AML was organized as a hierarchy maintained by SL-ICs that gave

rise to AML-CFU and eventually partially differentiated leukemic blasts (Figure 6, p.78). Importantly, when these initial SCID-leukemia transplantation studies were performed, secondary transplantation assays could not be performed due to the large cell doses required to repopulate SCID mice. Therefore it was not possible to prove beyond a doubt that these leukemic repopulating cells were truly stem cells.

Alternatively, the use of NOD-SCID mice for xenotransplantation assays proved superior, as the enhanced immunodeficiency of this strain required 10-20-fold less input cells for their engraftment.[70] Initial experiments showed that peripheral blood leukocytes from 15 of 18 AML patients were able to engraft NOD-SCID mice between 5-100%[70]. The enhanced engraftment of patient samples in these mice allowed for a greater serial dilution of patient mononuclear cells in order to quantify SL-IC numbers. Here, the frequency of SL-ICs in several patients was quantified and ranged between 0.2 and 100 cells per million mononuclear cells. In all cases, it was once again demonstrated that SL-ICs were only present within the CD34+CD38- population of human AML cells[70]. Importantly, the use of NOD/SCID mice in this work also made possible the demonstration of bona fide SL-IC self-renewal through performing secondary transplantation assays. Specifically, the researchers were able to demonstrate successful repopulation of secondary recipients using 4 different AML samples, where in each case, the leukemic morphology and cell surface phenotype remained unchanged when compared to the original sample. These early experiments were able to functionally demonstrate that LSCs could self-renew in order to maintain AML[70]. Further studies involving xenotransplantations into NOD-SCID mice demonstrated that the LSC fraction is

functionally heterogeneous[71]. In a study using lentiviral particles to mark individual HSCs, researchers were able to demonstrate that human AML samples contained different clones with different self-renewal capacities similar to that observed for the normal HSC pool. The existence of 3 types of LSCs was proposed by this group: long-term- (LT-), short-term-(ST-), and quiescent-LT-LSCs[71]. LT-LSCs persisted through a 12-week period in primary mice and LT-LSCs are able to repopulate secondary and even tertiary mice. ST-LSCs support clonal growth in primary mice but are unable to repopulate secondary recipients. Quiescent LT-LSCs were undetectable in primary mice but were identified in secondary recipients. The evidence that the LSC pool is organized in a hierarchical manner, similar to normal HSCs, supports the idea that the initial transforming mutation may occur within the HSC compartment.

More refined xenotransplantation assays have shed further light on the phenotypic heterogeneity and complexity of LSCs. Importantly, scientists have revealed that LSCs are not unique to the CD34+CD38- fraction of human AMLs[72-75]. It has been demonstrated that anti-CD38 antibodies have a severe inhibitory effect on AML engraftment mediated through the Fc-receptor. Importantly, the pre-incubation of human AML cells with anti-CD38 was shown to impair engraftment in NOD/SCID mice but the incubation with the anti-CD38 F(ab')2 did not. Furthermore when mice were treated with intravenous immunoglobulin (IVIG) or anti-CD122 to prevent the immune mediated clearance of human AML cells, engraftment levels were significantly enhanced[72]. Surprisingly, the effect of anti-CD38 was still present even in the more immune-deficient NSG strain of mice suggesting that other

immune mechanisms may play a role in the CD38-mediated clearance of AML cells. Notably when researchers transplanted CD34+CD38+ populations from a variety of AML samples into immunocompromised mice treated with IVIG or ant-CD122, all samples showed significant engraftment[72]. Furthermore, studies from the same group were able to show that certain subsets of leukemia contained LSCs within CD34- fractions[73]. Here researchers investigated NPM1-mutated human AML samples, which typically have very low levels of CD34 expression. When transplanted into NSG mice, it was demonstrated that half of the samples tested had LICs solely within the CD34- fraction[73]. These studies revealed that the phenotype of LSCs is much more heterogeneous than initially thought. In another study, when fractionated on CD34, CD38, and CD45RA, researchers showed that all fractions of human AML showed some ability to engraft NSG mice[74,75]. Surprisingly, though LSCs were found with the highest frequency in Lin-CD38- fraction, CD38+ and Lin+ LSCs were also shown to reconstitute NSG mice suggesting that AML LSCs may not necessarily arise solely from primitive hematopoietic cells. Ultimately, leukemic stem cells are quite heterogeneous when compared to their normal counterpart and cell surface markers are inadequate to define LSC function. Importantly, currently used cell surface markers are not functional and therefore are unable accurately predict self-renewal capabilities. Furthermore, LSCs may originate from HSCs or more differentiated progeny, which further complicate the use of these non-functional cell surface markers. For example, though LSCs are thought to arise from mutations within human HSCs, it is also possible that LSCs can arise from more differentiated cells that gained the capacity to self-renew. Do to the heterogeneous

nature of AML and the inability of cell surface markers to predict self-renewal capabilities, LSCs must be functionally validated in immunocompromised mice models.

1.3 Musashi-2 is a regulator of normal hematopoietic stem cells and is associated with aggressive leukemia

1.3.1 Functional assays of hematopoietic stem cell function

The use of *in vivo* transplantation assays has become the gold standard in assessing normal and leukemic HSC function. Importantly, these assays have made it possible to elucidate the function of individual proteins, RNAs, or therapeutic compounds in LSCs and HSCs. Numerous studies have used shRNA-lentiviral screens in order to identify critical regulators of hematopoiesis. One such study, used multiple shRNA constructs to target a list of 20 genes that were preferentially overexpressed in murine HSC populations[76]. A retroviral vector that co-overexpressed GFP was designed to introduce these shRNAs into HSC-enriched Lin-CD150+CD48- cells. These cells were transplanted into recipient mice and HSC activity was monitored by measuring the extent of GFP positivity in the peripheral blood samples taken at 4, 8, and 20 weeks post-transplant (Figure 7, p.79). This screen and subsequent downstream validations identified two hairpins, targeting the RNA-binding protein (RBP) Msi2 that could suppress HSC repopulation (Figure 8, p. 80).

1.3.2 Musashi family of RNA-binding proteins

Numerous RNA binding proteins (RNABPs) have been identified that play critical roles in the regulation of normal stem cell populations as well as malignant transformation. The Musashi family of RNA-binding proteins consists of two proteins: Musashi-1 (Msi1) and Musashi-2 (Msi2)[77]. These proteins are homologs of one another and are thought to have occurred due to a gene duplication event. They are highly conserved regulators of stem and progenitor cell populations across vertebrate and invertebrate species. Altered expression of Musashi family members can result in the dysfunction of embryonic and adult stem cell populations and can promote oncogenic transformation[77].

The Musashi protein was first described in Drosophila where it was shown to be a critical regulator of external sensory organ development[78] (Figure 9, p.81). The Drosophila mechanosensory bristle is generated from four successive asymmetric cell divisions from a common precursor cell called the sensory organ precursor (SOP)[78]. The resulting structure is comprised of four different non-neuronal support cells and one neuron. The first non-symmetric cell division yields two precursor cells referred to as IIa (non-neuronal precursor) and IIb (a neuronal precursor). The non-neuronal precursor is distinguished from the neuronal precursor by the expression of the Tramtrack69 (Ttk69) protein (a neuronal differentiation inhibitory protein)[78]. It was demonstrated that MSI can bind to the *Ttk69* mRNA in order to repress its translation[79]. This translational repression of *Ttk69* was further shown to be mediated by Notch signaling. Ultimately, Notch signaling can inhibit

MSI activity in the IIa precursor cell allowing for the translation of *Ttk69* mRNA[79]. In the IIb precursor cell however, Notch is inactivated and MSI prevents ttk69 translation resulting in the commitment to a neural fate.

Msi1 and *Msi2* homologs exist in Xenopus as *Nrp1* and *Xrp1* respectively[80]; both proteins share extensive sequence similarity within the RNA recognition motifs and moderate sequence similarity in the c-terminal tails (Figure 10, p.82). Northern blotting revealed that *Xrp1* is expressed in a variety of tissues including the brain, spleen, heart, lung, skeletal muscle, skin, and kidney[81]. In contrast, *Nrp1* expression is more restricted to the nervous system and oocytes. The function of *Nrp1* has been extensively studied in Xenopus where it plays a critical role in oocyte maturation by controlling the temporal expression of mRNAs[82]. The hormonally induced maturation of oocytes requires the precise translation of maternally derived mRNAs that control meiotic cell cycle progression. *Mos*, a proto-oncogene that facilitates cell cycle progression is translated immediately after hormonal stimulation[82] and a delay in the expression of *Mos* results in the termination of meiosis. Conversely, translation of the maternally derived *Wee1* mRNA occurs later on in oocyte maturation and the early expression of *Wee1* prevents meiotic cell cycle progression and oocyte maturation. The differential timing in the expression of maternally derived mRNAs is controlled by regulatory elements found within the 3'UTR. These elements direct the cytoplasmic polyadenylation of target mRNAs. Nrp1 was shown to be a *Mos* 3'UTR-interacting protein and its interaction with the 3'UTR of *Mos* was shown to regulate the early translational activation of this protein through its promotion of polyadenlation[82]. In fact, its was discovered that Nrp1 regulates the

polyadenylation of numerous xenopus mRNAs including *XBub3, Cdk2*, *Eg3*, *TATA binding protein 2*, and *Xotch*[82] (Figure 11, p. 83). All of these mRNAs are translated early during oocyte maturation and all contain the *Msi1* consensus motif (A/G)U_{1-3}(AGU) in their 3'UTR. Oocytes expressing a dominant-negative (DN) version of *Nrp1* blocked the progesterone-stimulated polyadenylation of these mRNAs[82]. *Nrp1* is also expressed throughout the developing Xenopus retina and shows a unique expression pattern in the mature structure. The Xenopus retina is composed of 6 types of neurons and 1 type of glia that all originate from a common multipotent pool of progenitors[83]. The retina is composed of 3 structural layers: an outer nuclear layer containing rods and cones, an inner nuclear layer consisting of glial cells and interneurons, and the ganglion layer that consists of ganglion cells. *Nrp1* is detected in early retinal progenitor cells and continues to be detected in the entire retina as these cells begin to differentiate[83]. Ultimately *Nrp1* continues to be expressed in a pool of BrdU-negative slowly dividing retinal stem cells, and also in mitotically active precursors and post-mitotic photoreceptors and pigment epithelium. It has been hypothesized that *Nrp1* may play a critical role in maintaining neural precursor in a similar manner that mammalian Msi1 maintains neural precursor populations.

In vertebrates, the Msi2 and Msi1 proteins are expressed throughout embryogenesis and in restricted populations of cells in the adult. Msi2 has been identified in the putative stem-cell compartments of numerous mouse systems including the small intestine, liver, lung, nervous system, ovary, and testis[77]. Similar studies in Zebrafish using *Msi2* reporter lines have reflected these findings[84].

Zebrafish models reflect two phases of *Msi2* expression, widespread expression in progenitor cells in the early embryo followed by a more restricted, tissue-specific expression in the adult. *Msi1* is also expressed in a variety of murine systems including the nervous system, small intestine, testis, and embryonic stem cells[77]. The expression of murine *Msi1* is less restrictive when compared to its Xenopus homolog, *Nrp1*. In all organ systems, the MSI proteins are thought to contribute to the maintenance of pools of stem and progenitor cells. In certain systems, both family members are expressed and are thought to act cooperatively; this has been demonstrated in the nervous system and intestinal crypts[85,86]. In other systems, such as the blood, testis, and ovaries, the MSI proteins are thought to have distinct functional roles[76,87,88].

The MSI proteins are highly expressed in the neural tissue of Zebrafish, mice, and humans. It is in fact considered one of the best markers of neural progenitors and stem cells in the vertebrate central nervous system (CNS)[77]. In the mouse, *Msi1* is expressed in the embryo throughout development but is at its highest level at embryonic day 12 (E12), when neurogenesis is actively occurring[89]. *Msi1* expression is first detected at E8 in the neural plate where it associates with the onset of neural tube formation. At E10, *Msi1* is expressed solely throughout the neural tube; no other tissues stain positive. *Msi1* expression levels drops off gradually as development continues and its expression becomes restricted in the adult[89]. Interestingly, as embryogenesis continues, mitotic neural precursors become restricted to the ventricular zone of the neural epithelium[89]. In a similar manner, *Msi1* expression becomes restricted to the ventricular zone at E12-E15[89]. These

Msi1 positive cells can be labeled with BrdU suggesting that *Msi1* is expressed in mitotic neural precursor cells during CNS development. In the post-natal CNS, *Msi1* is highly expressed in PCNA-positive proliferating cells in the subventricular zone that are thought to represent populations of neural precursor cells[89]. Msi2 is significantly co-expressed with Msi1 in the mammalian nervous system, where these proteins are thought to act co-operatively to regulate the proliferation and maintenance of neural stem and progenitor cells[90]. During embryonic development, Msi2 expression levels remain fairly constant. Despite this, Msi2 expression in the embryo is not ubiquitous but its expression is spatially and temporally regulated during the development of the brain and spinal cord[90]. In the embryonic CNS, Msi2 is predominantly detected in proliferating precursor cells located in the ventricular and subventricular zones in a manner similar to Msi1. In the postnatal brain, Msi2 expression is found in a population of subventricular zone cells surrounding the lateral ventricle. In the subventricular zone, Msi2 positive cells do not co-stain for oligodendrocyte or mature neuronal markers suggesting that Msi2 expression correlates with precursor cells[90].

Msi1 knockout *(Msi1-/-)* mice survive embryonic development but develop obstructive hydrocephalus resulting in death around 1 month of age[86]. This phenotype is not seen in *Msi2* knockout mice and is thought to result due to a unique role of Msi1 in murine ependymal cells. Neurospheres derived from *Msi1*-/- adult CNS cells are slightly smaller and less frequent than neurospheres derived from control cells[86]. However, secondary neurosphere repopulating activity is not impaired upon *Msi1* knockout indicating that a loss of *Msi1* may not have a severe

effect on CNS stem-cell self-renewal. Differentiation assays reveal that *Msi1-/-* neurospheres remain multipotent and can give rise to astrocytes, neurons and oligodendrocytes[86]. This data demonstrates that Msi1 is not an essential factor for the development of most tissues and implies that compensatory changes may occur during development. In the CNS specifically, the level of Msi2 expression is up-regulated 1.4-2.0 fold in *Msi1-/-* embryonic brains compared to wild type brains supporting the hypothesis that *Msi2* may compensate for *Msi1* deficiency[86]. Neurosphere cultures from *Msi1-/-* embryos incubated with peptide nucleic acids targeting MSI2 showed a significant reduction in the number of spheres in a dose-dependent fashion. This effect was not seen when wild type cells were treated with peptide nucleic acids (PNAs) targeting *Msi2*. Furthermore, though *Msi1-/-* neurospheres show no changes in cell proliferation as measured by Brdu incorporation, a dramatic reduction in Brdu positive cells is seen in PNA-treated *Msi1-/-* neurosphere cultures[86]. This data suggest that Msi2 may play a redundant role in the maintenance of murine stem and progenitor cells and that these proteins may be able to compensate for each other in this system.

Studies have shown that Msi1 can repress the translation of target mRNAs through specific recognition sequences in the 3'UTR (Figure 12, p.84). *In vitro* selection of high-affinity RNA ligands revealed that Msi1 binds to the specific consensus motif (G/A)U1-3(AGU)[91]. This Msi1 consensus motif was identified in the *m-numb* 3'UTR and UV cross-linking assays revealed that Msi1 could bind to the *m-numb* sequence in vitro[91]. *In vivo* assays revealed that Msi1 could bind the 3'UTR of *numb* in NIH-3T3 cells. Furthermore, overexpression of Msi1 in these cells was

shown to decrease Numb protein levels to 32% of control levels suggesting that Msi1 could repress the translation of *Numb* transcripts[91]. Additional studies revealed that poly(A)-binding protein (PABP) is a direct MSI1-binding protein[92]. These studies showed that Msi1 interacts strongly with the RNA recognition motif 1(RRM1) and RRM2 region of PABP, the same region where eIF4G is known to bind, suggesting that MSI1 may compete with eIF4G for PABP binding. eIF4G is a member of a group of proteins that bind to the mature mRNA 5' cap and facilitates ribosomal recruitment. A critical interaction between eIF4g and PABP stimulates the recruitment of the 48S and 80S ribosome. *In vitro* immunoprecipitation assays confirmed that Msi1 was able to decrease the interaction between eIF4G and PABP[92]. Because of its similarity to Msi1, Msi2 has been proposed to repress translation of target genes in a similar manner.

The mouse small intestine is made up of numerous mature cells that are thought to originate from multipotent stem cells located in the base of intestinal crypts. Mouse studies have characterized the localization of both Msi1 and Msi2 to these intestinal crypts[85]. The intestinal crypts consist of base columnar stem cells, paneth cells, and a rapidly dividing collection of cells known as the transit-amplifying zone. Overexpression of *Msi2* greatly increases the size of the proliferative zone and blocks the maturation of intestinal cells resulting in morbidity within 3-4 days[93]. In a similar manner to *Msi2*, *Msi1* is expressed in the base of intestinal crypts and its expression is restricted to LGR5+ and mTERT+ intestinal stem cells. Furthermore, in the adenomatous polyposis coli (APC) min

mouse model of adenocarcinoma, initial lesions that are thought to represent the initial stages of adenoma progression are strongly positive for MSI1[94].

Msi1 and Msi2 play critical roles in gametogenesis. Studies in Drosophila and mice indicate that the Musashi proteins play a critical role in the maintenance of testis stem cells and the regulation of meiosis. Spermatogenesis is a differentiation event occurring within the male testis where a pool of stem-like cells know as spermatogonia maintain the continuous production of spermatozoa throughout post-pubertal life[87]. Spermatogonia stem cells undergo mitotic divisions giving rise to spermatocytes that undergo two rounds of meiosis to form spermatids. Spermatids then undergo a post-mitotic differentiation to form spermatozoa. In Drosophila, MSI is found in both somatic cells and germ cells. A loss of MSI results in a loss of early germ-line stem cells in the adult testis[95]. Furthermore, MSI is important in late spermatogenesis where a loss of MSI results in defects in meiosis and cytokinesis. In the mouse, the MSI proteins show a unique temporal and spatial expression pattern suggesting a distinct role for each protein in murine spermatogenesis. *Msi1* is expressed early on in spermatogonial precursors and spermatogonia whereas *Msi2* is expressed later in spermatocytes and round spermatids[96]. Aberrant expression of *Msi1* or *Msi2* impairs normal spermatogenesis and is detrimental to cell health. Additionally, *Msi1* is expressed in murine sertoli cells where it plays a critical role in maintenance of the blood-testis barrier. In the mouse, *Msi1* and *Msi2* are differentially expressed during folliculogenesis. *Msi1* is expressed at higher levels in early primordial and primary follicles whereas *Msi2* is more abundant in pre-antral and antral follicles[88]. A KO mouse shows that *Msi2* is

essential for folliculogenesis. A *Msi2* genetrap mouse displays subfertility, a 41% reduction in total ovarian mass, and a 36% decrease in the total follicle number in mature ovaries compared to wild-type mice[88]. The follicles that do develop are of poor health and show impaired spermatozoa-zona pellucida binding.

Not surprisingly, numerous studies correlate the expression of MSI proteins with tumour aggressiveness. MSI proteins are active in glioma, medulloblastoma, myeloid leukemia, breast cancer, lymphoid leukemia, and gastric cancers[93,97-101]. *MSI1* is highly expressed in human medulloblastoma samples compared to normal tissue and correlates with poor prognosis[98]. *In vitro* and *in vivo* assays display impaired growth of medulloblastoma cells upon *MSI1* knockdown. Increased expression of *MSI2* correlates with poor clinical outcome in leukemia and both MSI1 and MSI2 promote gastric tumorignesis[93,99,102]. Additionally, both MSI2 and MSI1 are implicated in the process of intestinal tumorigenesis[85]. In human colorectal adenoma samples, *MSI1* is commonly overexpressed and correlates with poor survival and an increased risk of metastasis. Furthermore, knockdown of *MSI1* impairs tumor growth in xenograft models. Similarly, *MSI2* is consistently overexpressed in human colorectal adenocarcinomas where increased expression correlates with tumour grade[93]. *In vitro* and *in vivo* experiments demonstrate that MSI2 promotes cell proliferation and tumour growth and largely phenocopies APC loss in mouse models of intestinal adenomas.

1.3.3 The role of MSI2 in Hematopoiesis

MSI2 plays a critical role in the maintenance of murine and human hematopoietic stem and progenitor cells, whereas MSI1 is absent from this system[76,99]. Expression analysis has demonstrated that *Msi2* is most highly expressed in the LSK compartment of the murine hematopoietic system, which contains LT-HSC, ST-HSCs, and MPPs (Figure 13, p.85)[103]. In humans, *MSI2* is most highly expressed in the immature fraction of hematopoietic cells as well[104]. Its expression is highest in HSC fractions and decreases gradually as cells differentiate into multipotent progenitors, committed progenitors, and mature cells; its expression level is lowest in the post-mitotic myeloid cells. As mentioned previously, studies by Hope et al., were the first to characterize the functional impact of a loss of MSI2 in the mouse hematopoietic system[76]. Here they were able to show that a loss of MSI2 greatly impaired the repopulation potential of murine HSCs. The study by Hope et al., infected a Lin-CD150+CD48- LT-HSC enriched population with short hairpin RNAs (shRNAs) targeting *MSI2*. This population of cells was transplanted with congenic bone marrow into lethally irradiated mice and HSC activity was monitored by measuring the fraction of GFP+ cells in the Ly5.1 population in the peripheral blood from 4-20 weeks post transplantation. Notably, two hairpins targeting *MSI2* were consistently able to reduce the fraction of transduced cells from the donor graft. Furthermore, the decrease in repopulation was directly proportional to the decreased levels of the *MSI2* transcript. This study elegantly showed that this effect was not due to changes in homing capacity, apoptosis, or cell cycle status. Similarly, studies by Kharas et al., echoed these

results[99]. Here, a knockdown of *Msi2* in murine LSK cells, a population that contains LT-HSCs, ST-HSCs, and MPPs, was shown to impair their repopulation capabilities. Upon transplantation of sh*Msi2* LSKs into lethally irradiated mice, Kharas et al., revealed a significant loss of GFP+ cells in the peripheral blood and LT-HSCs of the bone marrow relative to input populations. They attributed this phenotype to a loss of hematopoietic progenitors, a phenotype that was supported by another study using an *Msi2* gene-trap mouse[105]. Using a gene-trap *Msi2* mouse, another group demonstrated that a loss of Msi2 greatly impaired the activity of ST-HSCs and MPPs while having a relatively benign effect on the activity of LT-HSCs under steady-state conditions. Upon competitive transplantation however, *Msi2* knockout bone marrow demonstrated a significant loss of LT-HSCs, suggesting that MSI2 may be critical for the appropriate function of LT-HSCs under stress conditions. Mice were killed 6 months post transplant and the bone marrow was analyzed. This revealed that MSI2 gene-trapped cells were out-competed even when they were in excess. Notably, at 8 weeks post transplantation, essentially no *Msi2* gene trapped cells of any lineage could be detected[105]. It was noted however, that *Msi2* gene-trapped cells were still intrinsically capable of engraftment and differentiation since 10 million BM cells from *Msi2* gene-trap mice were able to engraft and differentiate in mice up to 5 months post-transplant. Kharas et al., further characterized the function of *Msi2* in the murine hematopoietic system using a doxycycline-inducible *Msi2* transgenic mouse[99]. They demonstrated that under steady-state conditions, overexpression of *Msi2* caused a preferential increase in ST-HSCs and MPPs (LSK CD150-CD48-) and a proportional decrease in LT-HSCs (LSK CD150+CD48-). Under conditions of

competitive transplant, overexpression of *Msi2* caused an increase in LT-HSC numbers, once again suggesting that Msi2 may function to expand LT-HSCs under times of stressful repopulation. However, after a period of long-term and short-term engraftment (20 and 6 weeks respectively), a decreased contribution of Msi2 in the peripheral blood was observed, suggesting a differentiation block may occur with overexpression of *Msi2*. Both competitive and non-competitive transplants revealed a reduced contribution of *Msi2*-overexpressing cells in the peripheral blood of transplanted mice. Contrary to these findings, a retroviral overexpression of *Msi2* showed enhanced long-term engraftment with *Msi2* overexpression. Notably such results may occur due to different levels of *Msi2* overexpression. The retrovirus insertion was shown to elicit 2.8X up-regulation of *Msi2*, whereas the doxycycline inducible mouse model was shown to cause a 40X up-regulation of *Msi2*. Such results suggest that there may be an optimal *Msi2* dosage that is permissive for increased self-renewal and proliferation.

1.3.4 Musashi-2 maintains hematopoietic stem cell self-renewal and correlates with aggressive leukemia

Numerous studies implicate *Msi2* in the formation of aggressive leukemias[99,106-110]. Functional studies using a blast crisis model of chronic myelogenous leukemia (CML) indicate that *Msi2* promotes the development of aggressive disease. Notably, in mouse models of CML, the HSC-enriched LSK fraction, when infected with the CML-specific BCR-ABL translocation, induces a transplantable leukemia with a phenotype that closely resembles the chronic phase

of this disease[110]. Though BCR-ABL is known to be the causative agent in CML, other accessory mutations are required in order for this disease to progress into the more aggressive blast crisis phase, which closely mimics AML. One such accessory mutation is the NUP98-HOXA9 fusion. This translocation was first identified in human AML and it has been shown that mice transplanted with LSK cells infected with BCR-ABL and NUP98-HOXA9 develop a disease that closely resembles the blast crisis phase of CML[110]. It was demonstrated that in these more immature blasts, *Msi2* expression is 10X higher than in chronic phase blasts. Furthermore, HOXA9 is capable of binding to the *Msi2* promoter region and it is thus postulated that the NUP98-HOXA9 fusion protein may play a direct role in the up-regulation of *Msi2*[110]. Remarkably, leukemia burden is distinctly reduced when murine LSK cells from a *Msi2* gene-trap mouse transduced with blast crisis CML-inducing mutations are transplanted into wild type mice[110]. Such mice also show much greater survival compared to mice transplanted with control wild type LSK cells. Transplants of blast crisis bone marrow expressing shRNAs directed against *Msi2* have echoed these results[110]. In the clinical setting, an examination of 90 patient samples revealed an up-regulation of *MSI2* in all samples during CML progression [105,110]. Furthermore, *MSI2* expression correlated with patient relapse and death, suggesting that expression of *MSI2* may serve as a clinical marker of advanced CML.

In the context of AML, *MSI2* is an independent prognostic factor for overall survival[99] and strong correlations of *MSI2* with poor prognostic markers have also been reported[99]. In the risk stratification of AML, cytogenetics is the single most important predictor of outcome. Notably, anomalies involving core-binding factor

(CBP), such as inv16 or t(8;21) are the best prognostic markers and are associated with chronic remission rates >90%. Interestingly, expression profiling of 436 AML patient samples revealed a negative correlation between the presence of inversion 16 and *MSI2*[99]. Furthermore, amongst the worst cytogenetic markers are monosomies of chromosome 5 or 7. In the same analysis of 436 patient samples, it was noted that *Msi2* expression was significantly higher in patients with monosomy 7. Moreover, *MSI2* is up regulated in patients carrying the negative prognostic mutation FLT3-internal tandem duplication (ITD). Importantly, patients with normal karyotypes but FLT3-ITD only have a 30% survival rate, even after allogeneic transplantation.

Many key points in the scientific literature suggest a link between *Msi2* and LSCs. Notably, a knockout (KO) of *Msi2* in blast crisis mouse stem cells greatly impairs secondary repopulation[110]. Secondary transplantation assays yielded 4-fold higher mortality when mice were transplanted with bone marrow from wild type blast crisis primary mice vs. *Msi2* knockout blast crisis bone marrow. Furthermore, in a HoxA9-Meis1 driven model of murine acute leukemia, which can be divided into distinct fractions with tumour initiating capacity (TIC), it was discovered that *Msi2* is up regulated in fractions with TIC vs. those without[111]. In agreement with leukemic mouse models, preliminary data from a large-scale study of genes differentially expressed in LSC vs. non-LSC fractions of human AML indicates that *Msi2* is more highly expressed in the LSC-containing fraction (Personal Communication from Dr. John Dick). Lastly, an immunohistochemical (IHC) analysis of 120 patient AML samples supports the idea that *MSI2* expression may associate

with LSC function[109]. When researchers performed IHC staining of patient AML samples, they noted that 70% of samples showed some staining for MSI2 protein but that the percentage of positive cells was quite low. Despite this low expression pattern, *MSI2* levels negatively associated with patient outcome. Furthermore, the most significant association was seen when >1% of cells showed a very high level of MSI2[109]. The fact that the prognostic power of MSI2 can be attributed to the intense staining of a very small number of positive cells is indicative that MSI2 may associate with LSC activity.

1.4 Elucidating the Function of RNA binding proteins

1.4.1 Overview of RNA-binding proteins

RNA-binding proteins (RBP) are complex and mediate numerous cellular processes. Importantly, most RNAs are co-transcriptionally incorporated into RNA-protein complexes and remain in such complexes throughout their lifetime in the cell. Critical processes including RNA splicing, translation, transport, and stability, are all mediated by numerous RBPs that are often incorporated into large RBP complexes. Well known RBPs include splicing factors, ribosomal proteins, eukaryotic initiation factors, and those involved in RNAi silencing. Such proteins are ubiquitously expressed in the cell. Other RBPs are expressed in a much more restricted manner and bind very specific RNAs to effect their localization, stability, or translation.

RBPs are modular proteins that usually incorporate a unique combination of RNA-binding domains (RBD). These domains include RNA-recognition motifs

(RRMs), K-homology (KH) domains, Piwi domains, Pumilio domains, double stranded RNA (dsRNA)-binding domains, and the RNA-binding Zinc-finger domain[112]. RRMs are by far the most common RNA-binding domain. The unique arrangement and number of these RNABDs within a given RNABP is thought to determine the RNA-binding specificity of a given protein. Approximately 1% of human genes encode for an RRM[112]. This domain is composed of 90 amino acids that form a beta sheet with two helices packed against it[112]. RRM-RNA binding typically occurs on the beta-sheet in an association that usually involves the RNA interacting with three conserved residues: an arginine or lysine residue in combination with two aromatic ones. A single RRM is thought to bind between 4 and 8 nucleotides. The KH domain is another well characterized and relatively common RBD that can bind to both single-stranded DNA and single stranded RNA[112]. This domain is composed of approximately 70 amino acids and is structured as a three-stranded beta sheet packed against three alpha helices. The KH domain family can be further subdivided into two subfamilies (type I and type II) based on the arrangement of the beta-sheet and the alpha helices.

Isolated RBDs cannot typically interact with RNA in a sequence specific manner due to the small sequences (typically 4-8 nucleotides) that they recognize. Multiple domains must typically be combined onto a single polypeptide in order to create a more complex RNA-recognition sequence with greater specificity and affinity of RNA binding[112]. Typically two RBDs are found in a given RBP but this number can vary dramatically. Additionally, the linker region between two RBDs can have a significant effect of the RNA-binding properties of the protein[112]. This is

seen in protein such as HRP1, PABP, Nucleolin, and HuD. These proteins all contain two RRMs separated by a linker sequence[112]. Upon RNA binding, the RNA and protein undergo significant structural rearrangements resulting in the formation of a helical structure in the linker region, which creates more surfaces for the recognition of nearby stretches of RNA. In fact, one of the major determinants for RNA-binding specificity and affinity is the length and amino acid composition of the linkers found between RBDs[112]. In numerous proteins, the linker becomes ordered after RNA binding and this can increase the affinity of RNA binding up to 100 000-fold[112].

RBDs are not only important for RNA-binding, but also play critical roles in facilitating the protein-protein interactions between different RBPs. RRMs can typically facilitate protein dimerization through interactions between their c-terminal helices[112]. Such dimerization is critical to the function of numerous RBPs and only occurs in the presence of RNA since the c-terminal helix is bound to the beta sheet in the RNA-free proteins. Dimerization can act to strengthen the affinity of protein-RNA interactions and has also been shown to alter RBP function[112]. RBDs other than RRM can participate in protein-protein interactions as well. KH domains are known to dimerize and dsRBDs have also been shown to form protein-protein interactions[112]. Importantly, protein-protein interactions play a very important role in ribonucleoprotein (RNP) formation. In terms of protein dimerization, this effect has been shown to contribute to the formation of stable RNP granules. Studies involving the IGF2BP family of RBPs have elegantly displayed that these proteins are able to form homo- and heterodimer complexes and that these complexes greatly

contribute to the formation of stable RNA-protein complexes[113]. Specifically, studies have shown that IGF2BP2 and IGF2BP1 can form heterodimers on the 3'UTR of *Igf2* and *H19* mRNA. This binding occurs in a sequential manner through the third and fourth KH domains. The first recruitment step is fast but of low affinity and involves one of the IGF2BP family members binding to RNA. The second step involves the recruitment of a second IGF2BP family onto a nearby region on the same mRNA. This process allows for IGF2BP protein-protein interactions resulting in the formation of a highly stable protein-RNA complex[113].

RBPs typically undergo a variety of post-translation modifications (PTM) that can further affect the RNA-binding characteristics, function, and localization of the RBP. Common PTMs found on RBPs include phosphorylation, arginine methylation, and small ubiquitin-like modification (SUMO). Importantly, RBPs are rich in lysine and arginine residues; these positively charged amino acids are known to mediate hydrogen bonding and aromatic interactions. Furthermore, arginine residues located with glycine and arginine rich (GAR) motifs are targeted by arginine protein methyltransferases (PRMTs) that add one or two methyl groups onto the amino acid[114]. RBPs typically contain GAR motifs and are major targets for PRMTs. The arginine methylation of these RBPs is thought to play a variety of roles. Some studies indicate that arginine methylation may play a role as a maturation signal since some hypomethylated RBPs typically become mislocalized[115]. Other studies indicate that arginine methylation may facilitate the formation of RNP complexes. Furthermore, arginine residues tend to be critical players in binding to RNA, and it is thus postulated that methylation of these residues may decrease RNA affinities

due to the methyl group preventing the capacity for hydrogen bonding[114]. Other PTMs such as phosphorylation has been shown to impact the function of RBPs. Importantly, RBP phosphorylation has been shown alter the composition and function of RBP complexes. In Xenopus specifically, studies have characterized the role of CPEB1 as both a translational activator and repressor of its mRNA targets in neurons[116]. CPEB1 binds specific sequences in the 3'UTR of responsive mRNAs. CPEB1 can form a protein complex with the 4E-BP Maskin which binds eIF-4E, blocking the interaction with eIF-4G thus preventing the formation of the translation initiation complex[116]. However, the phosphorylation of CPEB1 can convert this repressive role into a translational activation role. Phosphorylation of CPEB1 results in its binding to the poly(A) polymerase Gld-2. This results in the elongation of the poly(A) and facilitates the binding of PABP, which recruits eIF4G to the 5'cap dislodging Maskin from eIF4E.

Several methods have been developed in order to characterize the RNA-binding characteristics of RBPs. Notably, cross-linking immunoprecipitation followed by sequencing (CLIP-Seq), also known as high-throughput sequencing cross-linking immunoprecipitation (HITS-CLIP), has provided an elaborate way to exhaustively characterize protein-RNA interactions in an unbiased manner[117].

1.4.2 CLIP-Seq allows for the identification of protein-RNA interactions

CLIP-Seq allows for the identification and exhaustive characterization of protein-RNA interactions. For many years, scientists have tried to characterize the interaction between protein and RNA. Such interactions were first characterized

using a variety of *in vitro* techniques including RNA Selection assays, RNase footprinting, and electrophoretic mobility shift assays (EMSA)[118]. These assays helped to characterize RBP affinities, kinetic stability, and were even able to identify *in vitro* consensus binding motifs for numerous RBPs; however, the *in vivo* application of this data was limited. RNA-Immunoprecipitaiton (RIP) assays were developed and involved the immunoprecipitation of protein-RNA complexes under non-denaturing conditions followed by qPCR or microarray analysis. However, this technique was limited due to a high signal to noise ratio; even under the most stringent conditions, RNA remains sticky, limiting the usefulness of RIP experimentation[119]. The post-lysis *in vitro* formation of RNA-protein complexes, as well as the binding of RNA to antibody and sepharose/ magnetic beads, all contribute to the large background signals generated by this technique. Moreover, the downstream analysis of RIP-derived RNAs through qPCR requires prior knowledge of the RNA target. Analysis of this qPCR data is further complicated due to differences in experimental protocols and controls. There is currently no agreed upon standard control and no consensus as to what constitutes a significant enrichment of target RNA. In regards to RIP-microarray analysis, this is only suitable for the most stable RNA-protein complexes and will fail to identify any transient interactions[119]. Furthermore, characterization of the actual RNA-binding motif is not possible through microarray analysis.

To characterize the *in vivo* binding characteristics of RBPs, low-throughput cross-linking immunoprecipitation was developed[120]. This technique was first used to identify the RNA targets of the RBP Nova in the brain, and it involved the

treatment of cells with UV-B radiation that would specifically crosslink closely associated protein-RNA species[120]. Only direct protein-RNA interactors interacting within the order of 1 angstrom are cross-linked by UV-B radiation. Immunoprecipitation of the Nova-RNA complex under stringent conditions, followed by RNA processing, allowed for the purification and cloning of Nova-specific RNA molecules. Notably, RNA processing with a dilute concentration of nuclease allowed for the trimming of RNA molecules down to an average size of 50 base pairs (bp) allowing for the precise identification of RNA binding sites[120]. Originally, Sanger sequencing of isolated RNA molecules was performed allowing for the identification of hundreds of unique reads. However, the advent of next generation sequencing has allowed for the exhaustive sequencing and analysis of millions of unique reads from a single CLIP experiment[117]. This "high-throughput" CLIP has been combined with extensive computational analysis allowing for the identification of target site consensus motifs, global analysis of RNA binding site location, and the generation of ranked lists of RNA binding targets based on sequence read intensity. The RBP Nova was the first protein to have its RNA-binding targets elucidated on a genome-wide scale[121]. Since then numerous other RBPs have been interrogated in this manner.

HITS-CLIP begins with the UV-crosslinking of suspension cells, cells grown in a monolayer, or a tissue sample. Cells are then lysed under very harsh conditions and the protein of interest is immunoprecipitated. Following stringent washes, attached RNAs are trimmed using an optimal concentration of a nuclease (commonly micrococcal nuclease, RNAse A, or RNAse T1)[122]. Trimming of the RNA

facilitates the identification of the actual RNA-binding site. A 3'RNA adapter is then ligated onto the RNA molecules, which are then 5' phosphorylated using radioactive P^{32}. This facilitates the identification of protein-RNA species *via* western blotting and development *via* autoradiography. After transfer to a nitrocellulose membrane, regions of the membrane containing the protein-RNA complexes of appropriate molecular weight can be cut out and treated with proteinase K in order to liberate the RNA. After phenol-chloroform isolation, RNA molecules then have a 5' RNA adapter ligated and undergo reverse transcription followed by limited rounds of PCR in order to amplify the input material and add appropriate sequences for next-generation sequencing (NGS). A typical HITS-CLIP experiment yields millions of sequencing reads[122]. To date, there is much variation in the number of reads that are sequenced in a given CLIP experiment with no consensus on the required sequencing depth[123]. However, the sequence complexity of a given CLIP-Seq library is usually not too complex and deeper sequencing of this library does not typically result in a more complex output. Numerous different bioinformatics pipelines exist for the analysis of CLIP-Seq reads[117,123]; all begin by filtering raw reads in order to remove low quality reads, adapter sequences, and poly(A) tails. These reads are then aligned back to a reference genome and identical reads are collapsed. Mapping back to a reference genome allows for the identification of binding sites within introns, which is critical if it is unknown whether a given protein binds mature mRNA or pre-mRNA[123]. It is also possible to align the reads against a transcriptome if the protein is thought to bind mature mRNA (i.e. the protein is uniquely located in the cytoplasm) or if the study simply wants to characterize the RBP binding sites on

mature mRNA. Collapsing identical reads aids in the removal of potential PCR duplicates[123]. This ensures that only reads originating from unique associations of protein-RNA complexes are analyzed. One drawback to this approach is that instances occur where identical reads originating from unique protein-RNA interactions are removed. Due to the nature of size selection and the use of nucleases that often have a bias for nucleotide cut location, it is possible to generate identical sequence reads from independent associations. Some studies have dealt with this through the introduction of barcodes in the 3'adapter sequence[123]. Therefore, identical reads originating from PCR duplication can be discerned from identical reads originating from unique protein-RNA associations. Those identical reads with identical barcode sequences are discarded as they likely originate from PCR duplication events whereas identical reads with separate barcode sequences likely result from independent protein-RNA capture events. This protocol has been estimated to significantly increase the identification of unique tags[124]. Unique "clusters" of CLIP reads are identified and ranked based on a Poisson distribution[117]. These clusters can be further analyzed by numerous software programs in order to identity consensus motifs, characterize the genome-wide location of these clusters, and elucidate patterns of RNA binding.

1.4.3 Alternative CLIP-Seq protocols

Other variations of HITS-CLIP exist. One closely related protocol, termed Photoactivatable-Ribonucleoside-Enhanced Cross-linking immunoprecipitation (PAR-CLIP), makes use of photoactivatable ribonucleoside analogs such as 4-

thiouridine (4SU) and 6-thioguanosine[125]. Here, cells are incubated with ribonucleoside analogs that become incorporated into RNA transcripts and are subsequently cross-linked using UV radiation with a wavelength of 365 nm. These analogs are used in order to overcome low cross-linking efficiencies that typically occur when using 254nm UV light. This variation in crosslinking allows for a higher yield of cross-linked RNAs using similar radiation intensities but it also allows for the precise identification of RNA-protein crosslinking sites[122]. Notably, when 4SU is cross-linked to protein, significant structural changes are thought to occur resulting in a T→C transition[125]. Clusters of sequence reads with a high percentage of these mutations, regardless of the number of sequence reads are therefore hypothesized to represent actual protein-RNA cross-linking sites. Clusters without significant T→C transitions represent background RNA species that were not cross-linked to the protein of interest. This protocol is limited to cultured cells that can be incubated with ribonucleoside analogs; it cannot be used to investigate RNA-protein interactions in tissue samples. Furthermore, to deal with the high frequency of mutations introduced upon UV-crosslinking of ribonucleoside analogs to proteins, a more relaxed bioinformatic analysis is required[122]. Typical experiments allow some flexibility when aligning sequence reads back to a human genome and allow for a small number of mismatches to be tolerated. In PAR-CLIP experiments, up to 80% of T residues may be mutated to C requiring a decreased alignment stringency, which can greatly decrease the number of uniquely mapped reads[125]. Individual-nucleotide resolution CLIP (iCLIP) is another HITS-CLIP variation that was developed in order to overcome the inefficiencies in reverse transcription[124]. Numerous studies have

indicated that when performing reverse transcription from the 3' universal adapter, the reverse transcriptase often stalls when it encounters the small polypeptides that remain attached to the RNA at the protein-binding site. This results in the formation of truncated cDNAs that lack the 5' universal adapter. Differences between the HITS-CLIP and iCLIP protocol occur after the RNA of interest is isolated from the nitrocellulose membrane. Instead of the addition of a 5' universal adapter, RNAs are reverse transcribed from the 3' universal adapter resulting in a variety of cDNA products (many of which are truncated due to reverse transcriptase stalling). The 3' adapter-ligated cDNA is then circularized and subsequently linearized *via* digestion of the 3'adapter sequence with BamHI[124]. This yields a suitable linear template for PCR amplification and allows for the high-throughput sequencing of truncated cDNAs. This protocol is more technically complicated when compared to the traditional HITS-CLIP technique and requires a more complicated bioinformatic workup due to the inclusion of truncated cDNAs.

CLIP protocols are technically demanding and important limitations must be recognized. Crosslinking RNA and protein using 254nm UV radiation occurs with very low efficiencies estimated between 1-5%[122]. Moreover, the use of numerous enzymatic steps with moderate efficiencies results in the loss of usable RNA at every step of the CLIP protocol. Additionally, the stalling of the reverse transcriptase at small polypeptides attached to RNA molecules is thought to greatly decrease the final yield of adapter-ligated cDNA. Consequently, these experiments typically require a large amount of input (typically greater than 50 million cells), in order to succeed. Even still, numerous PCR cycles are often required to amplify the CLIP-

derived cDNA to sufficient levels for NGS resulting in libraries with very low complexity. Importantly, retrospective analysis of 279 published CLIP experiments has revealed that 83.8% of CLIP reads are discarded due to PCR duplicates[126]. Despite these shortcomings, numerous CLIP studies are still able to generate between 100 000 and 4 million unique reads per experiment[123]. Limitations still exist in the analysis of these uniquely mapped reads. This is evident when comparing CLIP-Seq datasets to corresponding RNA-Seq data. Often, those CLIP-Seq reads with the greatest signal intensity correspond to mRNAs that are highly expressed in that cellular context[126]. More robust analysis is required in order to delicately pick apart true interactors from highly expressed background mRNA molecules and to appropriately rank the putative RNA targets while taking into account the expression of each individual RNA molecule.

Recently, an enhanced-CLIP (eCLIP), protocol has been developed; this protocol is technically simpler than traditional HITS-CLIP allowing for greater rates of successful library construction[126]. Furthermore, this protocol decreases the required number of PCR cycles for the amplification of cDNA, reducing the amount of PCR duplicates by an average of 60%. To better identify and rank RNA targets, eCLIP employs paired IgG and size-matched controls allowing for the normalization of read intensities to input RNA levels. Another great aspect of the eCLIP protocol is that it overcomes the decreased rates of reverse transcription that occur when the reverse transcriptase stalls at polypeptide sequences in cross-linked RNA. The authors of this protocol recognized that the standard reverse transcription performed in traditional HITS-CLIP was inefficient and they also recognized that the

circularization of 3'-adpater ligated RNA (as performed in iCLIP) was also inefficient. In the eCLIP protocol, reverse transcription is performed as in iCLIP resulting in the generation of a variety of full length and truncated cDNAs. Instead of circularizing the cDNA as in iCLIP, a single-stranded (ss) DNA oligonucleotide is ligated to the 5'end of cDNA molecules so that all cDNA products can be analyzed through NGS. Using high amounts of polyethylene glycol (PEG) and DMSO, the ligation efficiency has been determined to be greater than 70%. To increase library complexity, the ssDNA adapter contains a N5 randomer to determine whether a given sequencing read occurs due to PCR duplication. After the isolation of 5'adapter-ligated cDNA, the number of optimal PCR cycles is determined through a qPCR-based assay[126]. This has taken the guesswork out of determining the number of PCR cycles required to generate libraries for sequencing and ensures that only the minimal number of PCR cycles are used in order to minimize PCR duplication events. Compared to iCLIP, eCLIP yields a 100-fold increase in adapter-ligated pre-amplification cDNA products, a 54% increase in the number of uniquely mapped reads, and a greater signal-to-noise ratio due to the normalization of signals against size-matched input controls. The ranking of peaks based on size-matched input normalization has greatly facilitated the identification of true positive peaks. Strikingly, most genes showed a similar read density between size-matched input controls and eCLIP experiments[126]. Despite this, numerous genes had their read densities significantly increased in eCLIP datasets when compared to transcriptome frequency; these genes are thought to represent potent RNA targets.

1.5 Investigating protein-protein interactions

1.5.1 Early approaches in the identification of protein-protein interactors

Numerous RBDs such as RRMs and KH domains also function as protein-protein interacting domains[112]. Protein-protein interactions are thought to alter the structure of RBPs modifying their binding affinities and RNA-binding specificities. The identification of protein-protein interactions has typically relied on the use of protein libraries screened with bait proteins or through the isolation of protein complexes through co-immunoprecipitation (co-IP) followed by the identification of protein partners *via* mass spectrometry[127]. The yeast-two hybrid approach was critical in the early study of protein-protein interactions. Here, two recombinant proteins, one fused to a DNA binding domain and the other to a transcriptional activation domain are overexpressed in yeast along with a plasmid containing a reporter gene under the control of a promoter sequence recognized by the transcription factor[128]. Activation of the reporter gene is indicative of a protein-protein interaction occurring. Numerous limitations to this early technique exist. Notably, the overexpression of two proteins of interest does not indicate whether these proteins truly interact in vivo or whether an artificial interaction is occurring due to the non-physiological levels of each protein. Moreover, this system does not take into account the complexity that may underlie a protein-protein interaction. Importantly, numerous post-translational modifications and accessory proteins have been shown to facilitate the interaction of two proteins. Lastly, this screen is limited to those proteins that that can form robust complexes in the nucleus of a cell. The most commonly used technique for discerning protein-protein interactors is

through affinity purification followed by mass spectrometry[127]. Here high affinity tags (i.e. Flag) are attached to a protein of interest that is overexpressed in a cell line, and immunoprecipitation of the high affinity tag under non-denaturing conditions isolates the recombinant protein and associated protein-binding partners. This process has allowed for the characterization of numerous protein complexes, however, it also has limitations. The common need to overexpress a protein of interest may alter the normal physiological environment of the protein by promoting interactions that do not occur under normal physiological conditions. Furthermore, this protocol is designed to isolate robust protein-protein interactions and is likely to miss out on transient or weak interactors. Additionally, these assays are typically carried out in so called 'work-horse' cell lines such as HEK293 and HeLa cells. The cellular environments of these cell lines may differ significantly from the normal physiological environments in which a given protein functions[127]. This can therefore obscure protein-protein studies in numerous ways. First and foremost, tissue-specific interacting partners may simply not be expressed in a given cell line. Furthermore, critical accessory proteins or proteins required for the post-translational modification of a given protein may be absent in these cells preventing proper complex formation.

1.5.2 Identification of protein interactors using BioID

Recently, the development of the BioID system has allowed for the identification of proximally interacting proteins in an unbiased manner[129] (Figure 14, p.86). This technique allows for the identification of both transient and stable

direct and non-direct protein interactors. By combining this technique with the use of CRISPR-Cas9 technology, we are one of the first groups to apply the BioID system to a relatively un-perturbed cellular environment. The BioID system exploits the *Escheria Coli* (*E. coli*) BirA biotin ligase[129]. In *E. coli*, the BirA enzyme is required to transfer biotin to a specific lysine residue in the essential fatty acid synthetic protein, AccB, a component of the acetyl-CoA carboxylase complex[130]. When the levels of AccB acceptor protein are low or the concentration of biotin is high, BirA proteins can dimerize and bind to the operator sequence of the biotin operon resulting in repression of biotin synthetic gene transcripts.

This system was initially exploited through the development of biotin acceptor tags. A minimal 14-bp recognition motif termed the biotin acceptor tag (BAT), was initially identified that could be specifically biotinylated by BirA[131]. This tag was incorporated into fusion proteins and could be extensively biotinylated when transfected with BirA allowing for high affinity purification with streptavidin. This interaction has been thoroughly characterized and occurs in two steps. In the first step, BirA combines biotin and ATP to form a biotinoyl-5'-AMP (bioAMP) intermediate[132]. The BirA enzyme has a high affinity for this intermediate and it remains in the BirA active site until it comes in close proximity with a specific lysine residue within the context of the BAT. Further exploitation of the BirA enzyme, led to the discovery of a mutant form, R118G (termed BirA*), which has a reduced affinity for the bioAMP intermediate[132]. This mutant form therefore releases the bioAMP intermediate resulting in the promiscuous biotinylation of primary amines in a proximity-dependent manner[132].

Numerous proteins have had their binding partners elucidated through the use of BirA* fusion proteins[133,134]. This protocol usually involves the construction of a cell line expressing a BirA* fusion protein in an inducible manner, followed by incubation with 50uM biotin which stimulates massive levels of promiscuous biotinylation[132]. Cells are typically incubated with biotin for 24 hours and are subsequently lysed under stringent conditions. Biotinylated proteins are captured with streptavidin-coated beads, and washed extensively. Proteins are then digested with trypsin and peptides are identified *via* mass spectrometry. This protocol has been used to extensively study the interactome of numerous proteins. Additionally, the protein interactors identified through BioID studies have been compared to the protein interactors identified through affinity-purification mass spectrometry[135]. These studies have revealed that these two techniques identify largely distinct protein interactors[135,136]. In one study, the overexpression of a BirA* fusion protein led to the identification of 210 unique interactors and 90 unique interactors for the H2B and H3 proteins respectively[135]. However, only 21 and 10 of these proteins were identified by affinity purification mass the spectrometry. The differences in protein partner identification may occur due to the ability of the BioID system to recognize both stable and transient interactors as well as the ability of the BioID system to identify proximally interacting proteins in addition to direct protein interactors.

Proper controls are critical for the successful identification of true protein interactors in the BioID system. Importantly, in the previous study identifying H2B and H3 interactors, changes in the control cell lines could significantly alter the

number of proteins considered to be H2B and H3 interactors. Using the most relaxed controls, 385 H2B interactors and 183 H3 interactors could be identified[135]. Control experiments with NLS-BirA*-Flag and GFP-BirA*-Flag revealed that the overexpression of the BirA* moiety alone led to the detection of 1448 proteins in the NLS-BirA* line and 1534 protein in the GFP-BirA* cell line after streptavidin purification. Importantly, a number of proteins were uniquely identified in each control line. Although 1272 overlapping biotinylated proteins were discovered in both NLS-BirA*-Flag and GFP-BirA* controls, 176 and 262 unique proteins were found in the NLS-BirA* and GFP-BirA* controls respectively[135]. Additionally, 65 unique biotinylated peptides were identified in untreated cell lines. Three technical replicates were typically required in order to saturate the number of proteins detected by each control line. This suggests that in order to truly identify the vast number of non-specific biotinylated proteins, the inclusion of multiple controls and replicates is required.

The enhanced ability of the BioID system to detect protein-protein interactions in combination with the relative scalability of the BioID protocol makes it an appealing system to use in the elucidation of numerous protein interactomes. Importantly, co-IP assays must use very carefully selected buffers that can maintain normal complex formation[127]. This can impair the retrieval of insoluble proteins and proteins which are difficult to isolate under less stringent conditions (i.e. membrane proteins). Furthermore, co-IP assays rely on stable protein interactions and are unsuitable for the detection of transient protein interactions.

One possible way to increase the specificity of the BioID system would be to use CRISPR/Cas9 to create endogenous BirA* fusion proteins (Figure 15, p.87). It was our belief that these endogenously tagged proteins would precisely biotinylate only those proteins occurring at the exact physiological locations where our protein of interest resides. Furthermore, by using very specific controls in which a self-cleaving P2A sequence was inserted between our protein of interest and the BirA* protein, we aimed to better identify whether a perceived positive interactor was an actual positive signal by determining whether the biotinylated peptide count dropped when the BirA* protein was no longer physically connected to our protein of interest. We also hypothesized that the peptide counts derived from this system would more accurately reflect the relative abundance of proteins naturally occurring within the immediate vicinity of our protein of interest. This system is widely applicable to numerous different cell lines allowing for the interrogation of protein-protein interactions under relatively physiological conditions in numerous cellular contexts.

1.6 CRISPR-Cas9 Genome engineering

1.6.1 Description of the CRISPR locus

The clustered-regularly interspaced short palindromic repeats (CRISPR) are a bacterial and archaea adaptive defence system that protects these organisms from viruses and exogenous plasmids[137]. The CRISPR locus was first identified in *E. coli* and is present in 45% of bacteria. This locus contains regions of short repeated sequences that are separated by unique protospacer regions[138]. These protospacer

sequences are derived from the nucleic acids of viruses and exogenous plasmids. Adjacent to the CRISPR locus is a set of CRISPR-associated (Cas) genes[138]. The proteins encoded by these genes are critical for the adaptive immune process (Figure 16, p.88). The CRISPR system can be divided into 3 steps: 1) adaptation, 2) expression, and 3) interference. Adaptation involves the acquisition of short foreign DNA elements that are integrated into the CRISPR locus as spacers. The mechanism of adaptation is poorly understood but is dependent on the ubiquitously expressed Cas1 and Cas2 proteins[138]. The 'expression' step involves the transcription and translation of Cas genes and the transcription of the CRISPR locus into a long precursor RNA (pre-crRNA). This RNA is processed by Cas ribonucleases into mature CRISPR RNA (crRNA) that is incorporated into a CRISPR-ribonuclease complex (crRNP). In the last stage, interference, the combination of crRNA and Cas proteins are able to selectively target and destroy foreign DNA[138]. The degradation of foreign DNA is mediated by different Cas nucleases depending on the CRISPR system. Importantly, alignment of the protospacer regions to bacteriophage genomes has revealed that protospacer sequences are typically associated with adjacent sequence motifs in the phage genome, referred to as protospacer adjacent motifs (PAM). The PAM sites vary throughout the CRISPR systems of different bacteria[138]. They likely generate a recognition site for protospacer excision from the genome and are critical for interference to occur. Importantly, many Cas protein complexes will not bind to or cleave foreign protospacer sequences unless they are in the context of a PAM sequence.

A given bacterial genome may have single or multiple CRISPR loci and a variable composition of Cas proteins. The variability in the way different Cas proteins and CRISPRs mediate the recognition and cleavage of foreign DNA has led to the classification of 3 different CRISPR-Cas systems[138]. In type I and II systems, PAM motifs are critical in the acquisition and interference against foreign DNA. Type I and III systems share a similar feature where after processing of the pre-crRNA into mature crRNAs, a large multisubunit complex is required to mount the crRNA to recognize and cleave foreign genetic materal[138]. In the type II CRISPR-Cas system, a single protein, Cas9 is required for RNA-guided DNA recognition and cleavage of foreign DNA[137-139]. In this system, the Cas9 protein is guided to foreign DNA through its interactions with crRNA and another RNA called the trans-activating crispr RNA (tracrRNA). The tracrRNA is located upstream of the CRISPR-Cas locus in *S. pyogenes*. Site-specific cleavage occurs upon base pairing between the 20 base pair protospacer sequence in the crRNA and a target sequence immediately preceding a PAM motif.

1.6.2 CRISPR-Cas9 as a genome-engineering tool

The type II CRISPR system from *S. pyogenes* has been developed into a powerful genetic tool for genome engineering in eukaryotic cells[137,139,140]. The CRISPR-Cas9 system can be utilized in human cells through the overexpression of a human-codon optimized Cas9 protein and a single RNA molecule. Notably, the crRNA and tracrRNA can be fused together, generating a chimeric single-guide RNA (sgRNA). This shRNA has the 20-nucleotide crRNA protospacer sequence at the 5’

end that determines the DNA target site. The double stranded RNA structure at the 3' end of the sgRNA corresponds to the tracrRNA sequence and is responsible for Cas9 binding[139]. The Cas9 protein contains two nuclease domains, HNH and RuvC. Each domain cleaves a single DNA strand; the HNH domain cleaves the complementary strand and the RuvC domain cleaves the non-complementary strand[138]. This results in the formation of a double strand break that can be repaired through homologous recombination or non-homologous end joining (NHEJ) in eukaryotic cells. The induction of these repair pathways at a specific location in the DNA can be exploited in order to create gene knockouts or to insert a specific sequence at any location in the DNA. Repair of double stranded breaks by NHEJ pathways can introduce indel mutations at a given locus typically resulting in the introduction of nonsense mutations[140]. Targeted modifications can be introduced through the co-transfection of sgRNA, Cas9 nuclease, and a repair template in the form of a single-stranded oligonucleotide or plasmid[140]. Plasmid-based donor repair templates typically provide a template for homologous recombination by incorporating large homology arms (greater than 500 base pairs) flanking the modification of interest[138,140]. These are useful for the introduction of large inserts into a locus. For small modifications, single stranded oligonucleotides harbouring short homology arms (greater than 40 base pairs) are quite successful. Furthermore, an aspartate to alanine (D10A) mutation in the RuvC nuclease domain of the Cas9 protein transforms the Cas9 nuclease into a Cas9 nickase (Cas9n)[141]. This protein nicks the DNA resulting in a single stranded break. Importantly, targeting Cas9n proteins to cleave opposite strands of the DNA in close proximity to

one another can simulate a double stranded break allowing for the precise modification of a genetic locus and minimizing off-target double stranded breaks.

1.7 Thesis Objectives

Overall, numerous RBPs have been identified that play critical roles in the regulation of normal stem cell populations as well as malignant transformation. In particular, studies focusing on MSI2 show it to be a key regulator of human and mouse HSCs and provide evidence linking its dysregulation to the transformation of hematopoietic cells[76,99,104]. Numerous studies implicate MSI2 in the formation of aggressive leukemias and mouse models of CML have shown that MSI2 expression is 10-fold higher in the most immature blast populations[110]. In the clinical setting, an examination of 90 patient samples revealed an up-regulation of MSI2 in all samples during CML progression[110]. Furthermore, MSI2 expression correlates with relapse and death, suggesting that expression of MSI2 may serve as a clinical marker of advanced CML. In the context of AML, MSI2 is an independent prognostic factor for overall survival and strong correlations of MSI2 with markers of poor prognosis have also been reported[99,109]. Despite this knowledge, very little is known about biochemical mechanisms through which MSI2 functions and although correlative data implicates MSI2 in aggressive myeloid leukemia, no studies to date functionally demonstrate a role of MSI2 in human AML. With these facts in mind, the first goal laid out in my PhD project was to assess the function of the MSI2 protein in the development of human AML through the use of xenograft transplantation assays. Although some interesting results were gained from these experiments, they proved

rather difficult to complete. The heterogeneous nature of human AML meant that each sample behaved uniquely. Nevertheless, techniques were refined and xenotransplant assays were carried out and interesting information was gained from these studies. To elucidate the mechanism of action through which MSI2 functions, another goal of my PhD was to identify the RNA binding targets of MSI2 and to characterize the RNA binding characteristics of this protein. This involved the standardization and optimization of the complex technique known as CLIP-Seq. After extensive refinements to this technique, MSI2 CLIP-Seq libraries were constructed, sequenced, and subjected to exhaustive bioinformatic analysis yielding a plethora of novel information. The last goal of my PhD was to take advantage of CRISPR-Cas9 technology in order to create endogenously tagged MSI2-BirA* cell lines for the identification of MSI2 protein binding partners through the BioID system. The use of CRISPR-Cas9 in order to create an endogenously tagged BirA* fusion protein had not been previously reported. We were unsure whether this technique would even work. Traditional studies utilizing the BirA* protein relied on transient transfection in order to overexpress a fusion protein of interest but these experiments were plagued by high levels of background and inconsistencies between replicates. We hypothesized that the creation of an endogenously tagged MSI2 locus would aid in the detection of true interactors since the BirA* protein would be uniquely localized to the exact regions of the cell where MSI2 is localized. Furthermore, to robustly identify MSI2 interactors from background proteins that are endogenously biotinylated, we decided to generate a P2A control cell line where a P2A peptide was inserted immediately downstream of the final exon of MSI2

between MSI2 and BirA*. This cell line would generate a single transcript that contained both BirA* and MSI2 and the translation of this transcript would result in the production of two separate proteins. Since MSI2 and BirA* are both driven off the endogenous MSI2 promoter, the two proteins will be at an equal stoichiometric level in control cell lines and are likely to be very similar to the levels of MSI2-BirA* in experimental lines. We hypothesized that this would serve as an excellent experimental control since the levels of BirA* would be very similar between experimental and control cells lines and the only real difference between these two lines would be the difference in BirA* localization. This technique proved to be rather robust and we were able to identify numerous proteins that were preferentially biotinylated in MSI2-BirA* but not MSI2-P2A-BirA* cell lines. Importantly BioID performed in two independently derived control cell lines and two independently derived experimental cells lines revealed identical results. The downstream analysis of MSI2 protein interactors revealed a MSI2-interacting protein that is uniquely expressed in the most immature fraction of hematopoietic stem cells and early functional experiments implicate it as a critical regulator of normal heamtopetic stem cell function. Overall, the three aims laid out for my PhD project were to: 1) investigate the functional role of MSI2 in human AML, 2) discover the RNA binding targets of MSI2 and to characterize the RNA-binding characteristics of MSI2, and 3) Identify MSI2 protein-protein interactions through the use of the BioID system. The following body of work will independently address each aim. Chapter 2 will discuss the functional role of MSI2 in human AML, chapter 3

will discuss the RNA-binding properties of MSI2, and chapter 4 will discuss the identification of novel MSI2 protein interactors through the use of the BioID system.

1.8 References

1 Maehle, A. H. Ambiguous cells: the emergence of the stem cell concept in the nineteenth and twentieth centuries. *Notes Rec R Soc Lond* **65**, 359-378 (2011).

2 Cavaillon, J. M. The historical milestones in the understanding of leukocyte biology initiated by Elie Metchnikoff. *J Leukoc Biol* **90**, 413-424, doi:10.1189/jlb.0211094 (2011).

3 Cooper, B. The origins of bone marrow as the seedbed of our blood: from antiquity to the time of Osler. *Proc (Bayl Univ Med Cent)* **24**, 115-118 (2011).

4 Weissman, I. L. & Shizuru, J. A. The origins of the identification and isolation of hematopoietic stem cells, and their capability to induce donor-specific transplantation tolerance and treat autoimmune diseases. *Blood* **112**, 3543-3553, doi:10.1182/blood-2008-08-078220 (2008).

5 Eaves, C. J. Hematopoietic stem cells: concepts, definitions, and the new reality. *Blood* **125**, 2605-2613, doi:10.1182/blood-2014-12-570200 (2015).

6 Weissman, I. L. The road ended up at stem cells. *Immunol Rev* **185**, 159-174 (2002).

7 Till, J. E. & Mc, C. E. A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. *Radiat Res* **14**, 213-222 (1961).

8 Becker, A. J., Mc, C. E. & Till, J. E. Cytological demonstration of the clonal nature of spleen colonies derived from transplanted mouse marrow cells. *Nature* **197**, 452-454 (1963).

9 Siminovitch, L., McCulloch, E. A. & Till, J. E. The Distribution of Colony-Forming Cells among Spleen Colonies. *J Cell Physiol* **62**, 327-336 (1963).

10 Worton, R. G., McCulloch, E. A. & Till, J. E. Physical separation of hemopoietic stem cells differing in their capacity for self-renewal. *J Exp Med* **130**, 91-103 (1969).

11 Boyse, E. A., Miyazawa, M., Aoki, T. & Old, L. J. Ly-A and Ly-B: two systems of lymphocyte isoantigens in the mouse. *Proc R Soc Lond B Biol Sci* **170**, 175-193 (1968).

12 Muller-Sieburg, C. E., Whitlock, C. A. & Weissman, I. L. Isolation of two early B lymphocyte progenitors from mouse marrow: a committed pre-pre-B cell and a clonogenic Thy-1-lo hematopoietic stem cell. *Cell* **44**, 653-662 (1986).

13 Spangrude, G. J., Heimfeld, S. & Weissman, I. L. Purification and characterization of mouse hematopoietic stem cells. *Science* **241**, 58-62 (1988).

14 Morrison, S. J. & Weissman, I. L. The long-term repopulating subset of hematopoietic stem cells is deterministic and isolatable by phenotype. *Immunity* **1**, 661-673 (1994).

15 Morrison, S. J., Wandycz, A. M., Hemmati, H. D., Wright, D. E. & Weissman, I. L. Identification of a lineage of multipotent hematopoietic progenitors. *Development* **124**, 1929-1939 (1997).
16 Kondo, M., Weissman, I. L. & Akashi, K. Identification of clonogenic common lymphoid progenitors in mouse bone marrow. *Cell* **91**, 661-672 (1997).
17 Akashi, K., Traver, D., Miyamoto, T. & Weissman, I. L. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature* **404**, 193-197, doi:10.1038/35004599 (2000).
18 Oguro, H., Ding, L. & Morrison, S. J. SLAM family markers resolve functionally distinct subpopulations of hematopoietic stem cells and multipotent progenitors. *Cell Stem Cell* **13**, 102-116, doi:10.1016/j.stem.2013.05.014 (2013).
19 Kiel, M. J. *et al.* SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. *Cell* **121**, 1109-1121, doi:10.1016/j.cell.2005.05.026 (2005).
20 Wilson, A. *et al.* Hematopoietic stem cells reversibly switch from dormancy to self-renewal during homeostasis and repair. *Cell* **135**, 1118-1129, doi:10.1016/j.cell.2008.10.048 (2008).
21 Pietras, E. M. *et al.* Functionally Distinct Subsets of Lineage-Biased Multipotent Progenitors Control Blood Production in Normal and Regenerative Conditions. *Cell Stem Cell* **17**, 35-46, doi:10.1016/j.stem.2015.05.003 (2015).
22 Adolfsson, J. *et al.* Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential a revised road map for adult blood lineage commitment. *Cell* **121**, 295-306, doi:10.1016/j.cell.2005.02.013 (2005).
23 McKenzie, J. L., Gan, O. I., Doedens, M., Wang, J. C. & Dick, J. E. Individual stem cells with highly variable proliferation and self-renewal properties comprise the human hematopoietic stem cell compartment. *Nat Immunol* **7**, 1225-1233, doi:10.1038/ni1393 (2006).
24 Bosma, G. C., Custer, R. P. & Bosma, M. J. A severe combined immunodeficiency mutation in the mouse. *Nature* **301**, 527-530 (1983).
25 Mosier, D. E., Gulizia, R. J., Baird, S. M. & Wilson, D. B. Transfer of a functional human immune system to mice with severe combined immunodeficiency. *Nature* **335**, 256-259, doi:10.1038/335256a0 (1988).
26 McCune, J. M. *et al.* The SCID-hu mouse: murine model for the analysis of human hematolymphoid differentiation and function. *Science* **241**, 1632-1639 (1988).
27 Kamel-Reid, S. & Dick, J. E. Engraftment of immune-deficient mice with human hematopoietic stem cells. *Science* **242**, 1706-1709 (1988).
28 Lapidot, T. *et al.* Cytokine stimulation of multilineage hematopoiesis from immature human cells engrafted in SCID mice. *Science* **255**, 1137-1141 (1992).
29 Shultz, L. D. *et al.* Multiple defects in innate and adaptive immunologic function in NOD/LtSz-scid mice. *J Immunol* **154**, 180-191 (1995).

30 Yamauchi, T. *et al.* Polymorphic Sirpa is the genetic determinant for NOD-based mouse lines to achieve efficient human cell engraftment. *Blood* **121**, 1316-1325, doi:10.1182/blood-2012-06-440354 (2013).
31 Shultz, L. D. *et al.* Human lymphoid and myeloid cell development in NOD/LtSz-scid IL2R gamma null mice engrafted with mobilized human hemopoietic stem cells. *J Immunol* **174**, 6477-6489 (2005).
32 Ito, M. *et al.* NOD/SCID/gamma(c)(null) mouse: an excellent recipient mouse model for engraftment of human cells. *Blood* **100**, 3175-3182, doi:10.1182/blood-2001-12-0207 (2002).
33 Doulatov, S., Notta, F., Laurenti, E. & Dick, J. E. Hematopoiesis: a human perspective. *Cell Stem Cell* **10**, 120-136, doi:10.1016/j.stem.2012.01.006 (2012).
34 Civin, C. I. *et al.* Antigenic analysis of hematopoiesis. III. A hematopoietic progenitor cell surface antigen defined by a monoclonal antibody raised against KG-1a cells. *J Immunol* **133**, 157-165 (1984).
35 Schmitt, C., Eaves, C. J. & Lansdorp, P. M. Expression of CD34 on human B cell precursors. *Clin Exp Immunol* **85**, 168-173 (1991).
36 Andrews, R. G. *et al.* CD34+ marrow cells, devoid of T and B lymphocytes, reconstitute stable lymphopoiesis and myelopoiesis in lethally irradiated allogeneic baboons. *Blood* **80**, 1693-1701 (1992).
37 Berenson, R. J. *et al.* Engraftment after infusion of CD34+ marrow cells in patients with breast cancer or neuroblastoma. *Blood* **77**, 1717-1722 (1991).
38 Tindle, R. W. *et al.* A novel monoclonal antibody BI-3C5 recognises myeloblasts and non-B non-T lymphoblasts in acute leukaemias and CGL blast crises, and reacts with immature cells in normal bone marrow. *Leuk Res* **9**, 1-9 (1985).
39 DiGiusto, D. *et al.* Human fetal bone marrow early progenitors for T, B, and myeloid cells are found exclusively in the population expressing high levels of CD34. *Blood* **84**, 421-432 (1994).
40 Hao, Q. L., Thiemann, F. T., Petersen, D., Smogorzewska, E. M. & Crooks, G. M. Extended long-term culture reveals a highly quiescent and primitive human hematopoietic progenitor population. *Blood* **88**, 3306-3313 (1996).
41 Huang, S. & Terstappen, L. W. Lymphoid and myeloid differentiation of single human CD34+, HLA-DR+, CD38- hematopoietic stem cells. *Blood* **83**, 1515-1526 (1994).
42 Petzer, A. L., Hogge, D. E., Landsdorp, P. M., Reid, D. S. & Eaves, C. J. Self-renewal of primitive human hematopoietic cells (long-term-culture-initiating cells) in vitro and their expansion in defined medium. *Proc Natl Acad Sci U S A* **93**, 1470-1474 (1996).
43 Vormoor, J. *et al.* Immature human cord blood progenitors engraft and proliferate to high levels in severe combined immunodeficient mice. *Blood* **83**, 2489-2497 (1994).
44 Bhatia, M., Wang, J. C., Kapp, U., Bonnet, D. & Dick, J. E. Purification of primitive human hematopoietic cells capable of repopulating immune-deficient mice. *Proc Natl Acad Sci U S A* **94**, 5320-5325 (1997).

45 Fritsch, G. *et al.* Rapid discrimination of early CD34+ myeloid progenitors using CD45-RA analysis. *Blood* **81**, 2301-2309 (1993).
46 Galy, A., Travis, M., Cen, D. & Chen, B. Human T, B, natural killer, and dendritic cells arise from a common bone marrow progenitor cell subset. *Immunity* **3**, 459-473 (1995).
47 Manz, M. G., Miyamoto, T., Akashi, K. & Weissman, I. L. Prospective isolation of human clonogenic common myeloid progenitors. *Proc Natl Acad Sci U S A* **99**, 11872-11877, doi:10.1073/pnas.172384399 (2002).
48 Doulatov, S. *et al.* Revised map of the human progenitor hierarchy shows the origin of macrophages and dendritic cells in early lymphoid development. *Nat Immunol* **11**, 585-593, doi:10.1038/ni.1889 (2010).
49 Notta, F. *et al.* Isolation of single human hematopoietic stem cells capable of long-term multilineage engraftment. *Science* **333**, 218-221, doi:10.1126/science.1201219 (2011).
50 Notta, F. *et al.* Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science* **351**, aab2116, doi:10.1126/science.aab2116 (2016).
51 Kampen, K. R. The discovery and early understanding of leukemia. *Leuk Res* **36**, 6-13, doi:10.1016/j.leukres.2011.09.028 (2012).
52 Piller, G. Leukaemia - a brief historical review from ancient times to 1950. *Br J Haematol* **112**, 282-292 (2001).
53 Coller, B. S. Blood at 70: its roots in the history of hematology and its birth. *Blood* **126**, 2548-2560, doi:10.1182/blood-2015-09-659581 (2015).
54 Horton, S. J. & Huntly, B. J. Recent advances in acute myeloid leukemia stem cell biology. *Haematologica* **97**, 966-974, doi:10.3324/haematol.2011.054734 (2012).
55 Dohner, H. *et al.* Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. *Blood* **115**, 453-474, doi:10.1182/blood-2009-07-235358 (2010).
56 Sekeres, M. A. Treatment of older adults with acute myeloid leukemia: state of the art and current perspectives. *Haematologica* **93**, 1769-1772, doi:10.3324/haematol.2008.000497 (2008).
57 Vardiman, J. W. *et al.* The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. *Blood* **114**, 937-951, doi:10.1182/blood-2009-03-209262 (2009).
58 Lange, B. The management of neoplastic disorders of haematopoiesis in children with Down's syndrome. *Br J Haematol* **110**, 512-524 (2000).
59 Gollin, S. M. Mechanisms leading to nonrandom, nonhomologous chromosomal translocations in leukemia. *Semin Cancer Biol* **17**, 74-79, doi:10.1016/j.semcancer.2006.10.002 (2007).
60 Grimwade, D. *et al.* The importance of diagnostic cytogenetics on outcome in AML: analysis of 1,612 patients entered into the MRC AML 10 trial. The

Medical Research Council Adult and Children's Leukaemia Working Parties. *Blood* **92**, 2322-2333 (1998).

61 Byrd, J. C. *et al.* Pretreatment cytogenetic abnormalities are predictive of induction success, cumulative incidence of relapse, and overall survival in adult patients with de novo acute myeloid leukemia: results from Cancer and Leukemia Group B (CALGB 8461). *Blood* **100**, 4325-4336, doi:10.1182/blood-2002-03-0772 (2002).

62 Cancer Genome Atlas Research, N. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* **368**, 2059-2074, doi:10.1056/NEJMoa1301689 (2013).

63 Papaemmanuil, E. *et al.* Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N Engl J Med* **374**, 2209-2221, doi:10.1056/NEJMoa1516192 (2016).

64 Arber, D. A. *et al.* The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* **127**, 2391-2405, doi:10.1182/blood-2016-03-643544 (2016).

65 Estey, E. H. Treatment of acute myeloid leukemia. *Haematologica* **94**, 10-16, doi:10.3324/haematol.2008.001263 (2009).

66 Dombret, H. & Gardin, C. An update of current treatments for adult acute myeloid leukemia. *Blood* **127**, 53-61, doi:10.1182/blood-2015-08-604520 (2016).

67 Killmann, S. A., Cronkite, E. P., Robertson, J. S., Fliedner, T. M. & Bond, V. P. Estimation of phases of the life cycle of leukemic cells from labeling in human beings in vivo with tritiated thymidine. *Lab Invest* **12**, 671-684 (1963).

68 Minden, M. D., Till, J. E. & McCulloch, E. A. Proliferative state of blast cell progenitors in acute myeloblastic leukemia (AML). *Blood* **52**, 592-600 (1978).

69 Lapidot, T. *et al.* A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. *Nature* **367**, 645-648, doi:10.1038/367645a0 (1994).

70 Bonnet, D. & Dick, J. E. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat Med* **3**, 730-737 (1997).

71 Hope, K. J., Jin, L. & Dick, J. E. Acute myeloid leukemia originates from a hierarchy of leukemic stem cell classes that differ in self-renewal capacity. *Nat Immunol* **5**, 738-743, doi:10.1038/ni1080 (2004).

72 Taussig, D. C. *et al.* Anti-CD38 antibody-mediated clearance of human repopulating cells masks the heterogeneity of leukemia-initiating cells. *Blood* **112**, 568-575, doi:10.1182/blood-2007-10-118331 (2008).

73 Taussig, D. C. *et al.* Leukemia-initiating cells from some acute myeloid leukemia patients with mutated nucleophosmin reside in the CD34(-) fraction. *Blood* **115**, 1976-1984, doi:10.1182/blood-2009-02-206565 (2010).

74 Sarry, J. E. *et al.* Human acute myelogenous leukemia stem cells are rare and heterogeneous when assayed in NOD/SCID/IL2Rgammac-deficient mice. *J Clin Invest* **121**, 384-395, doi:10.1172/JCI41495 (2011).

75 Eppert, K. *et al.* Stem cell gene expression programs influence clinical outcome in human leukemia. *Nat Med* **17**, 1086-1093, doi:10.1038/nm.2415 (2011).
76 Hope, K. J. *et al.* An RNAi screen identifies Msi2 and Prox1 as having opposite roles in the regulation of hematopoietic stem cell activity. *Cell Stem Cell* **7**, 101-113, doi:10.1016/j.stem.2010.06.007 (2010).
77 Sutherland, J. M., McLaughlin, E. A., Hime, G. R. & Siddall, N. A. The Musashi family of RNA binding proteins: master regulators of multiple stem cell populations. *Adv Exp Med Biol* **786**, 233-245, doi:10.1007/978-94-007-6621-1_13 (2013).
78 Nakamura, M., Okano, H., Blendy, J. A. & Montell, C. Musashi, a neural RNA-binding protein required for Drosophila adult external sensory organ development. *Neuron* **13**, 67-81 (1994).
79 Okabe, M., Imai, T., Kurusu, M., Hiromi, Y. & Okano, H. Translational repression determines a neuronal potential in Drosophila asymmetric cell division. *Nature* **411**, 94-98, doi:10.1038/35075094 (2001).
80 Arumugam, K., Macnicol, M. C. & Macnicol, A. M. Autoregulation of Musashi1 mRNA translation during Xenopus oocyte maturation. *Mol Reprod Dev* **79**, 553-563, doi:10.1002/mrd.22060 (2012).
81 Good, P. J., Rebbert, M. L. & Dawid, I. B. Three new members of the RNP protein family in Xenopus. *Nucleic Acids Res* **21**, 999-1006 (1993).
82 Charlesworth, A., Wilczynska, A., Thampi, P., Cox, L. L. & MacNicol, A. M. Musashi regulates the temporal order of mRNA translation during Xenopus oocyte maturation. *EMBO J* **25**, 2792-2801, doi:10.1038/sj.emboj.7601159 (2006).
83 Amato, M. A. *et al.* Comparison of the expression patterns of five neural RNA binding proteins in the Xenopus retina. *J Comp Neurol* **481**, 331-339, doi:10.1002/cne.20387 (2005).
84 Hochgreb-Hagele, T., Koo, D. E., Das, N. M. & Bronner, M. E. Zebrafish stem/progenitor factor msi2b exhibits two phases of activity mediated by different splice variants. *Stem Cells* **32**, 558-571, doi:10.1002/stem.1583 (2014).
85 Li, N. *et al.* The Msi Family of RNA-Binding Proteins Function Redundantly as Intestinal Oncoproteins. *Cell Rep* **13**, 2440-2455, doi:10.1016/j.celrep.2015.11.022 (2015).
86 Sakakibara, S. *et al.* RNA-binding protein Musashi family: roles for CNS stem cells and a subpopulation of ependymal cells revealed by targeted disruption and antisense ablation. *Proc Natl Acad Sci U S A* **99**, 15194-15199, doi:10.1073/pnas.232087499 (2002).
87 Sutherland, J. M., Siddall, N. A., Hime, G. R. & McLaughlin, E. A. RNA binding proteins in spermatogenesis: an in depth focus on the Musashi family. *Asian J Androl* **17**, 529-536, doi:10.4103/1008-682X.151397 (2015).
88 Sutherland, J. M. *et al.* Knockout of RNA Binding Protein MSI2 Impairs Follicle Development in the Mouse Ovary: Characterization of MSI1 and MSI2 during

Folliculogenesis. *Biomolecules* **5**, 1228-1244, doi:10.3390/biom5031228 (2015).
89 Sakakibara, S. *et al.* Mouse-Musashi-1, a neural RNA-binding protein highly enriched in the mammalian CNS stem cell. *Dev Biol* **176**, 230-242 (1996).
90 Sakakibara, S., Nakamura, Y., Satoh, H. & Okano, H. Rna-binding protein Musashi2: developmentally regulated expression in neural precursor cells and subpopulations of neurons in mammalian CNS. *J Neurosci* **21**, 8091-8107 (2001).
91 Imai, T. *et al.* The neural RNA-binding protein Musashi1 translationally regulates mammalian numb gene expression by interacting with its mRNA. *Mol Cell Biol* **21**, 3888-3900, doi:10.1128/MCB.21.12.3888-3900.2001 (2001).
92 Kawahara, H. *et al.* Neural RNA-binding protein Musashi1 inhibits translation initiation by competing with eIF4G for PABP. *J Cell Biol* **181**, 639-653, doi:10.1083/jcb.200708004 (2008).
93 Wang, S. *et al.* Transformation of the intestinal epithelium by the MSI2 RNA-binding protein. *Nat Commun* **6**, 6517, doi:10.1038/ncomms7517 (2015).
94 Sureban, S. M. *et al.* Knockdown of RNA binding protein musashi-1 leads to tumor regression in vivo. *Gastroenterology* **134**, 1448-1458, doi:10.1053/j.gastro.2008.02.057 (2008).
95 Siddall, N. A., McLaughlin, E. A., Marriner, N. L. & Hime, G. R. The RNA-binding protein Musashi is required intrinsically to maintain stem cell identity. *Proc Natl Acad Sci U S A* **103**, 8402-8407, doi:10.1073/pnas.0600906103 (2006).
96 Sutherland, J. M. *et al.* Developmental expression of Musashi-1 and Musashi-2 RNA-binding proteins during spermatogenesis: analysis of the deleterious effects of dysregulated expression. *Biol Reprod* **90**, 92, doi:10.1095/biolreprod.113.115261 (2014).
97 Dahlrot, R. H. *et al.* Prognostic value of Musashi-1 in gliomas. *J Neurooncol* **115**, 453-461, doi:10.1007/s11060-013-1246-8 (2013).
98 Vo, D. T. *et al.* The RNA-binding protein Musashi1 affects medulloblastoma growth via a network of cancer-related genes and is an indicator of poor prognosis. *Am J Pathol* **181**, 1762-1772, doi:10.1016/j.ajpath.2012.07.031 (2012).
99 Kharas, M. G. *et al.* Musashi-2 regulates normal hematopoiesis and promotes aggressive myeloid leukemia. *Nat Med* **16**, 903-908, doi:10.1038/nm.2187 (2010).
100 Wang, X. Y. *et al.* Musashi1 regulates breast tumor cell proliferation and is a prognostic indicator of poor survival. *Mol Cancer* **9**, 221, doi:10.1186/1476-4598-9-221 (2010).
101 Mu, Q. *et al.* High expression of Musashi-2 indicates poor prognosis in adult B-cell acute lymphoblastic leukemia. *Leuk Res* **37**, 922-927, doi:10.1016/j.leukres.2013.05.012 (2013).
102 Kuang, R. G. *et al.* Expression and significance of Musashi-1 in gastric cancer and precancerous lesions. *World J Gastroenterol* **19**, 6637-6644, doi:10.3748/wjg.v19.i39.6637 (2013).

103 Bagger, F. O. *et al.* BloodSpot: a database of gene expression profiles and transcriptional programs for healthy and malignant haematopoiesis. *Nucleic Acids Res* **44**, D917-924, doi:10.1093/nar/gkv1101 (2016).

104 Rentas, S. *et al.* Musashi-2 attenuates AHR signalling to expand human haematopoietic stem cells. *Nature* **532**, 508-511, doi:10.1038/nature17665 (2016).

105 de Andres-Aguayo, L. *et al.* Musashi 2 is a regulator of the HSC compartment identified by a retroviral insertion screen and knockout mice. *Blood* **118**, 554-564, doi:10.1182/blood-2010-12-322081 (2011).

106 Taggart, J. *et al.* MSI2 is required for maintaining activated myelodysplastic syndrome stem cells. *Nat Commun* **7**, 10739, doi:10.1038/ncomms10739 (2016).

107 Kaeda, J. *et al.* Up-regulated MSI2 is associated with more aggressive chronic myeloid leukemia. *Leuk Lymphoma* **56**, 2105-2113, doi:10.3109/10428194.2014.981175 (2015).

108 Barbouti, A. *et al.* A novel gene, MSI2, encoding a putative RNA-binding protein is recurrently rearranged at disease progression of chronic myeloid leukemia and forms a fusion gene with HOXA9 as a result of the cryptic t(7;17)(p15;q23). *Cancer Res* **63**, 1202-1206 (2003).

109 Byers, R. J., Currie, T., Tholouli, E., Rodig, S. J. & Kutok, J. L. MSI2 protein expression predicts unfavorable outcome in acute myeloid leukemia. *Blood* **118**, 2857-2867, doi:10.1182/blood-2011-04-346767 (2011).

110 Ito, T. *et al.* Regulation of myeloid leukaemia by the cell-fate determinant Musashi. *Nature* **466**, 765-768, doi:10.1038/nature09171 (2010).

111 Gibbs, K. D., Jr. *et al.* Decoupling of tumor-initiating activity from stable immunophenotype in HoxA9-Meis1-driven AML. *Cell stem cell* **10**, 210-217, doi:10.1016/j.stem.2012.01.004 (2012).

112 Lunde, B. M., Moore, C. & Varani, G. RNA-binding proteins: modular design for efficient function. *Nat Rev Mol Cell Biol* **8**, 479-490, doi:10.1038/nrm2178 (2007).

113 Nielsen, J., Kristensen, M. A., Willemoes, M., Nielsen, F. C. & Christiansen, J. Sequential dimerization of human zipcode-binding protein IMP1 on RNA: a cooperative mechanism providing RNP stability. *Nucleic Acids Res* **32**, 4368-4376, doi:10.1093/nar/gkh754 (2004).

114 Blackwell, E. & Ceman, S. Arginine methylation of RNA-binding proteins regulates cell function and differentiation. *Mol Reprod Dev* **79**, 163-175, doi:10.1002/mrd.22024 (2012).

115 Cote, J., Boisvert, F. M., Boulanger, M. C., Bedford, M. T. & Richard, S. Sam68 RNA binding protein is an in vivo substrate for protein arginine N-methyltransferase 1. *Mol Biol Cell* **14**, 274-287, doi:10.1091/mbc.E02-08-0484 (2003).

116 Wells, D. G. RNA-binding proteins: a lesson in repression. *J Neurosci* **26**, 7135-7138, doi:10.1523/JNEUROSCI.1795-06.2006 (2006).

117 Lovci, M. T. *et al.* Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat Struct Mol Biol* **20**, 1434-1442, doi:10.1038/nsmb.2699 (2013).
118 Panwar, B. & Raghava, G. P. Identification of protein-interacting nucleotides in a RNA sequence using composition profile of tri-nucleotides. *Genomics* **105**, 197-203, doi:10.1016/j.ygeno.2015.01.005 (2015).
119 Darnell, R. B. HITS-CLIP: panoramic views of protein-RNA regulation in living cells. *Wiley Interdiscip Rev RNA* **1**, 266-286, doi:10.1002/wrna.31 (2010).
120 Ule, J. *et al.* CLIP identifies Nova-regulated RNA networks in the brain. *Science* **302**, 1212-1215, doi:10.1126/science.1090095 (2003).
121 Licatalosi, D. D. *et al.* HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**, 464-469, doi:10.1038/nature07488 (2008).
122 Kishore, S. *et al.* A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat Methods* **8**, 559-564, doi:10.1038/nmeth.1608 (2011).
123 Wang, T. *et al.* Design and bioinformatics analysis of genome-wide CLIP experiments. *Nucleic Acids Res* **43**, 5263-5274, doi:10.1093/nar/gkv439 (2015).
124 Konig, J. *et al.* iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* **17**, 909-915, doi:10.1038/nsmb.1838 (2010).
125 Hafner, M. *et al.* Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**, 129-141, doi:10.1016/j.cell.2010.03.009 (2010).
126 Van Nostrand, E. L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* **13**, 508-514, doi:10.1038/nmeth.3810 (2016).
127 Zhang, X. F., Ou-Yang, L., Hu, X. & Dai, D. Q. Identifying binary protein-protein interactions from affinity purification mass spectrometry data. *BMC Genomics* **16**, 745, doi:10.1186/s12864-015-1944-z (2015).
128 Bruckner, A., Polge, C., Lentze, N., Auerbach, D. & Schlattner, U. Yeast two-hybrid, a powerful tool for systems biology. *Int J Mol Sci* **10**, 2763-2788, doi:10.3390/ijms10062763 (2009).
129 Roux, K. J., Kim, D. I. & Burke, B. BioID: a screen for protein-protein interactions. *Curr Protoc Protein Sci* **74**, Unit 19 23, doi:10.1002/0471140864.ps1923s74 (2013).
130 Chakravartty, V. & Cronan, J. E. Altered regulation of Escherichia coli biotin biosynthesis in BirA superrepressor mutant strains. *J Bacteriol* **194**, 1113-1126, doi:10.1128/JB.06549-11 (2012).
131 Beckett, D., Kovaleva, E. & Schatz, P. J. A minimal peptide substrate in biotin holoenzyme synthetase-catalyzed biotinylation. *Protein Sci* **8**, 921-929, doi:10.1110/ps.8.4.921 (1999).

132 Roux, K. J., Kim, D. I., Raida, M. & Burke, B. A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *J Cell Biol* **196**, 801-810, doi:10.1083/jcb.201112098 (2012).
133 Kim, D. I. *et al.* Probing nuclear pore complex architecture with proximity-dependent biotinylation. *Proc Natl Acad Sci U S A* **111**, E2453-2461, doi:10.1073/pnas.1406459111 (2014).
134 Chan, P. K. *et al.* BioID data of c-MYC interacting protein partners in cultured cells and xenograft tumors. *Data Brief* **1**, 76-78, doi:10.1016/j.dib.2014.10.001 (2014).
135 Lambert, J. P., Tucholska, M., Go, C., Knight, J. D. & Gingras, A. C. Proximity biotinylation and affinity purification are complementary approaches for the interactome mapping of chromatin-associated protein complexes. *J Proteomics* **118**, 81-94, doi:10.1016/j.jprot.2014.09.011 (2015).
136 Coyaud, E. *et al.* BioID-based Identification of Skp Cullin F-box (SCF)beta-TrCP1/2 E3 Ligase Substrates. *Mol Cell Proteomics* **14**, 1781-1795, doi:10.1074/mcp.M114.045658 (2015).
137 O'Connell, M. R. *et al.* Programmable RNA recognition and cleavage by CRISPR/Cas9. *Nature* **516**, 263-266, doi:10.1038/nature13769 (2014).
138 Rath, D., Amlinger, L., Rath, A. & Lundgren, M. The CRISPR-Cas immune system: biology, mechanisms and applications. *Biochimie* **117**, 119-128, doi:10.1016/j.biochi.2015.03.025 (2015).
139 Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821, doi:10.1126/science.1225829 (2012).
140 Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* **8**, 2281-2308, doi:10.1038/nprot.2013.143 (2013).
141 Ran, F. A. *et al.* Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154**, 1380-1389, doi:10.1016/j.cell.2013.08.021 (2013).

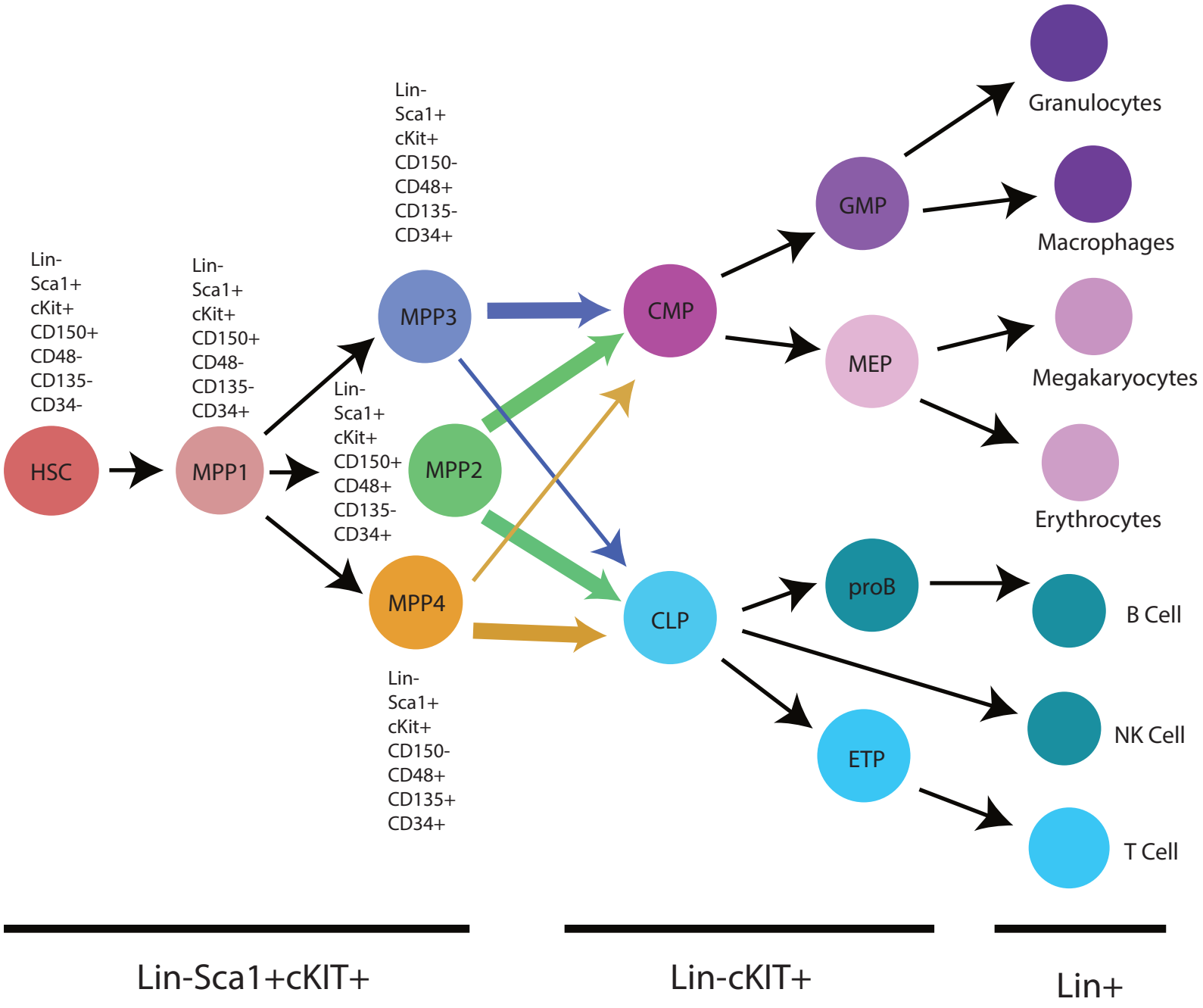


Figure 1. Mouse Hematopoietic Hierarchy. The mouse hematopoietic system is organized as a hierarchy with LT-HSCs at the apex. LT-HSCs are mostly quiescent. These cells can give rise to robust multilineage reconstituion of both primary and secondary mice. ST-HSCs (MPP1) give rise to robust multilineage engraftment (>16 weeks) in primary mice but show impaired reconstituion of secondary mice. MPP2, MPP3, and MPP4 represent multipotent progenitor popualtions. Transplantation assays reveal that these populations are fully multipotent but in vitro and in vivo analysis reflects a lymphoid bias in the MPP4 population and a myeloid bias in the MPP3 popualtion. MPP2, MPP3, and MPP4 popualtions can only maintain myeloid engraftment for ~ 1 month in vivo

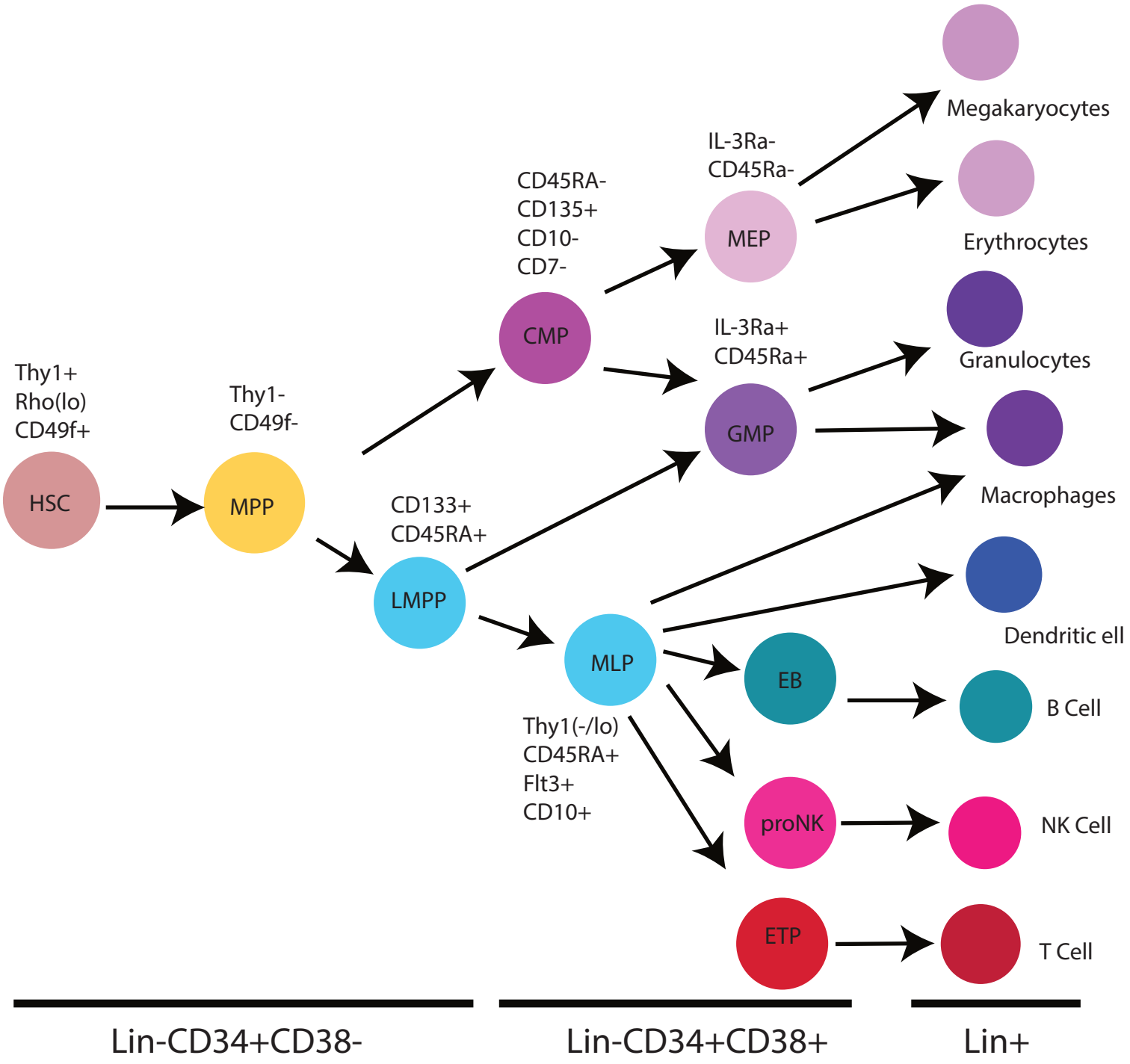


Figure 2. Human Hematopoietic Hierarchy. The human heamtopoetic system is organized as a hierarchy with HSCs at the apex. HSCs give rise to multipotent progenitors that have a limited capacity for self renewal. Relatively little is known about human MPP popualtions when comapred to the mouse system. However, in a similar manner different MPP populations are thought to exist that have differences in lymphoid and myeloid reconstituion capabilities. The lymphoid primed multipotent progenitor (LMPP) is one recently identified MPP population that has both lymphoid and myeloid developmental capacities but lacks the abilitiy to produce megakaryocytes and erythrocytes. Committed myeloid progenitors (CMPs) and committed lymphoid progenitors (MLPs) have been identified in the human system. CMPs only give rise to myeloid lineage cells. MLPs give rise to B cells, T cells, NK cells, macrophages, and dendritic cells.

Medical Research Council AML Stratification		
Favourable	Intermediate	Adverse
-Approximately 16% of newly diagnosed patients -t(8;21) -inv(16)(p13;q22) -t(16;16)(p13;q22) -t(15;17)(q24.1;q21.1)	-Approximately 60% of newly diagnosed patients -Normal Karyotype -Abnormalities not described in favourable or unfavourable	-Approximately 25% of newly diagnosed patients -del(5q) -add(5q) -del(7q) -add(7q) -monosomy 5 -monosomy 7 -inv(3)(q21;q26) -t(3;3)(q21;q26) -t(6;11)(q27;q23) -t(10;11)(p11-13;q23) -t(9;22)(q34;q11) -monosomy 17 -complex karyotype (at least 4 abnormalities)

Figure 3. Medical Research Council stratification of AML. The MRC has developed a classification system that divides AML patients into 3 groups: (1) those with favourable prognosis, (2) those with an intermediate prognosis, and (3) those with a poor prognosis. Patients are categorized based on the presence and/ or absence of cytogenetic abnormalities. The rate of overall survival in the favourable group are estimated at 55%. Rates of overall survival in the intermediate group are 24%. The rate of overall survival in the poor prognisis group is 5%.

European Leukemia Net AML Stratification			
Favourable	Intermediate-I	Intermediate-II	Adverse
-t(8;21)(q22;q22) -inv(16)(p13.1q22) -t(16;16)(p13.1;q22) -Mutated NPM1without FLT3-ITD (and normal karyotype) -Mutated CEBPA (and normal karyotype)	-Mutated NPM1 and FLT3-ITD (and normal karyotype) -Wilt-type NPM1 and FLT3-ITD (and normal karyotype) -Wild-type NPM1 without FLT3-ITD (and normal karyotype	-t(9;11)(p22;q23) -Cytogenetic abnormalities that are not classified as favourable or adverse	-inv(3)(q21q26.2) -t(3;3)(q21;q26.2) -t(6;9)(p23;q34) -t(v;11)(v;q23) -monosmy 5 -deletion 5q -monosomy 7 -abnormal(17p) -Complex karyotype (three or more chromosomal abnormalities)

Figure 4. European Leukemia Net AML Classification. The ELN has developed a system for the classification of AML that takes into account both cytogenetic abnormalities and three molecular markers: NPM1, FLT3, and CEBPA. This classification system allows for the further stratification of patients with cytogenetically normal AML into different subgroups based on the presence of molecular abnormalities recognized by the World Health Organization (WHO). Recent studies have indicated a prognostic utility of the ELN classification, especially in younger patients (<60 years). Significant differences in overall survival and disease-free survival are seen in younger patients. Those in the favourable group have the best outcomes but importantly, those patients in the intermediate-II group have better outcomes than those in the intermediate-1 group. Patients in the adverse group have the worst outcomes.

World Health Organization Classification of Acute Myelogenous Leukemia	
Class	Description
1) AML with recurrent genetic abnormalities	-t(8;21)(q22;q22) -inv(16)(p13.1q22) -t(16;16)(p13.1;q22) -t(15;17)(q24.1;q21.1) -t(9;11)(p22;q23) -t(6;9)(p23;q34) -inv(3)(q21;q26.2) -t(1;22)(p13;q13) -AML with mutated NPM1 -AML with mutated CEBPA
2) AML with MDS-related features	-AML that evolved from a previous case of MDS -AML with multilineage dysplasia -AML with MDS-related cytogenetic abnormalities (monosomy 5, del(5q), monosomy 7, del(7q))
3) Therapy-related AML	-Therapy-related AML is classified when a patient deveops AML after prior exposure to cytotoxic agents
4) AML, not otherwise specified	-Cases of AML that cannot be categorized as having a recurrent genetic abnormality, MDS-related features, nor therapy-related, are classified into this group

Figure 5. World Health Organziation (WHO) AML Classification. The WHO AML classification system organizes AML patients into one of four subgroups based on cytogenetic, molecular, and clinical features. The WHO classification system attempts to seperate AML into distinct groups based on factors that are known to effect prognosis. Once categorized into a subgroup, clinicians can better tailor therapy based on the features of the AML. For example, some recurrent genetic abnormalities are known to have good outcomes while other recurrent abnormalities have poor outcome. Recognizing this, clinicians can more agressively treat those with poor outcomes. Additionally, patients with therapy -related AML tend to have a very poor outcome compared to those with *de novo* AML. This stratification allows clinicians to more closely monitor these patients and tailor therapy appropriately

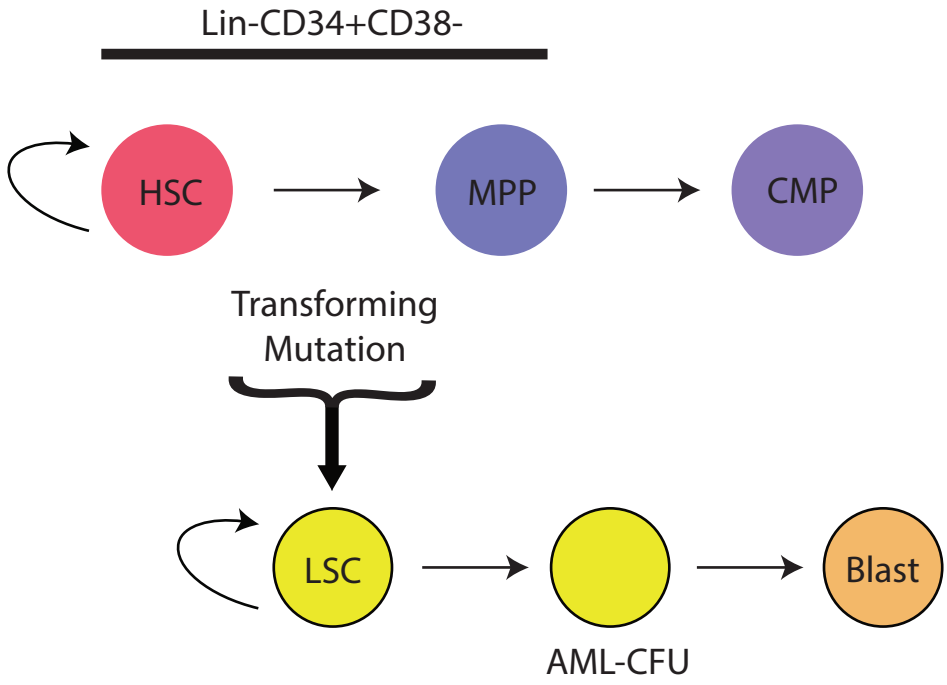


Figure 6. AML vs. HSC hierarchy. In a similar manner to the normal hematopoietic system, AML is also thought to be organized as a hierarchy. A transforming mutation is thought to occur either in a hematopoietic stem cell or a more downstream progenitor that results in the formation of a leukemic stem cell (LSC) that is characterized by abberant self-renewal. LSCs are thought to be relatively quiescent and responsible for relapse after treatment with standard chemotherapy. Importantly, due to their relatively quiescent nature, it is thought that standard chemotherapy regimens, which target proliferating cells, are unable to effectively target LSCs. LSCs are thought to give rise to leukemic progenitor cells that have an enhanced proliferative capacity and eventually differentiate into partially differentiated, non-functional, non-proliferative leukemic blasts

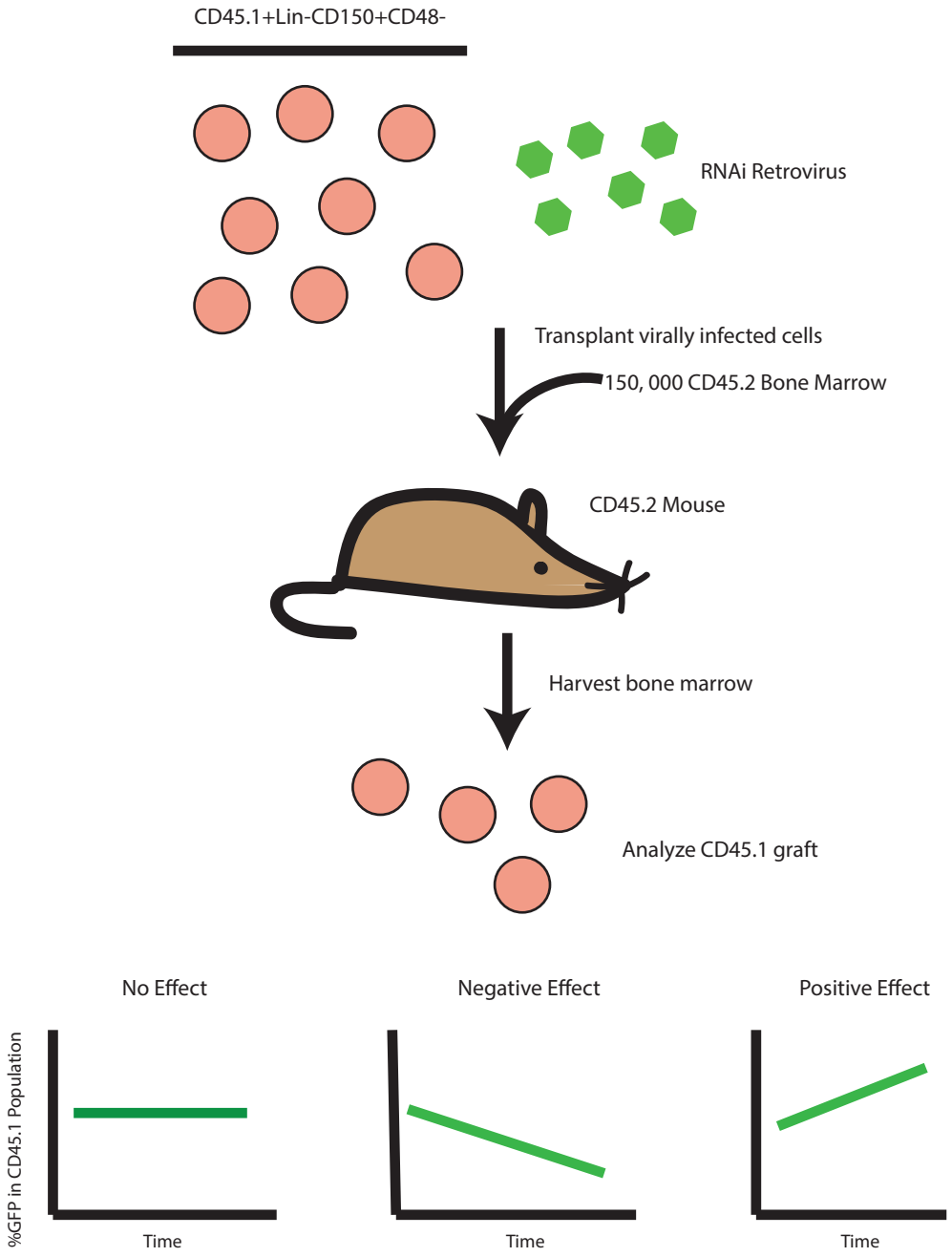


Figure 7. RNAi screen identifies MSI2 as a regulator of hematopoietic self-renewal. shRNAs were designed against a variety of target genes and packaged into retroviral particles. Individual retroviruses were used to infect Lin-CD150+CD48- CD45.1 positive donor cells. Importantly donor mice and recipient mice were identical except for their expression of the pan-hematopoietic marker CD45. Donor mice expressed the CD45.1 allele while recipient mice expressed the CD45.2 allele. This allowed for the detection of the donor cells in the graft after transplantation. The shRNA vector also expressed a GFP construct allowing for the tracking of infected cells. Importantly, the ratio of GFP+ to GFP- cells in the donor graft was used to indicate the effect of the hairpin on transduced cells. Those hairpins that have no effect on the cells would be expected to maintain similar GFP levels compared to input levels. If a hairpin is having a negative effect on the cells, we would expect GFP- cells to perform better than GFP+ cells resulting in a decrease in GFP levels in the donor graft compared to input levels. This protocol identified two hairpins, targeting the MSI2 gene, that had a negative impact on engraftment.

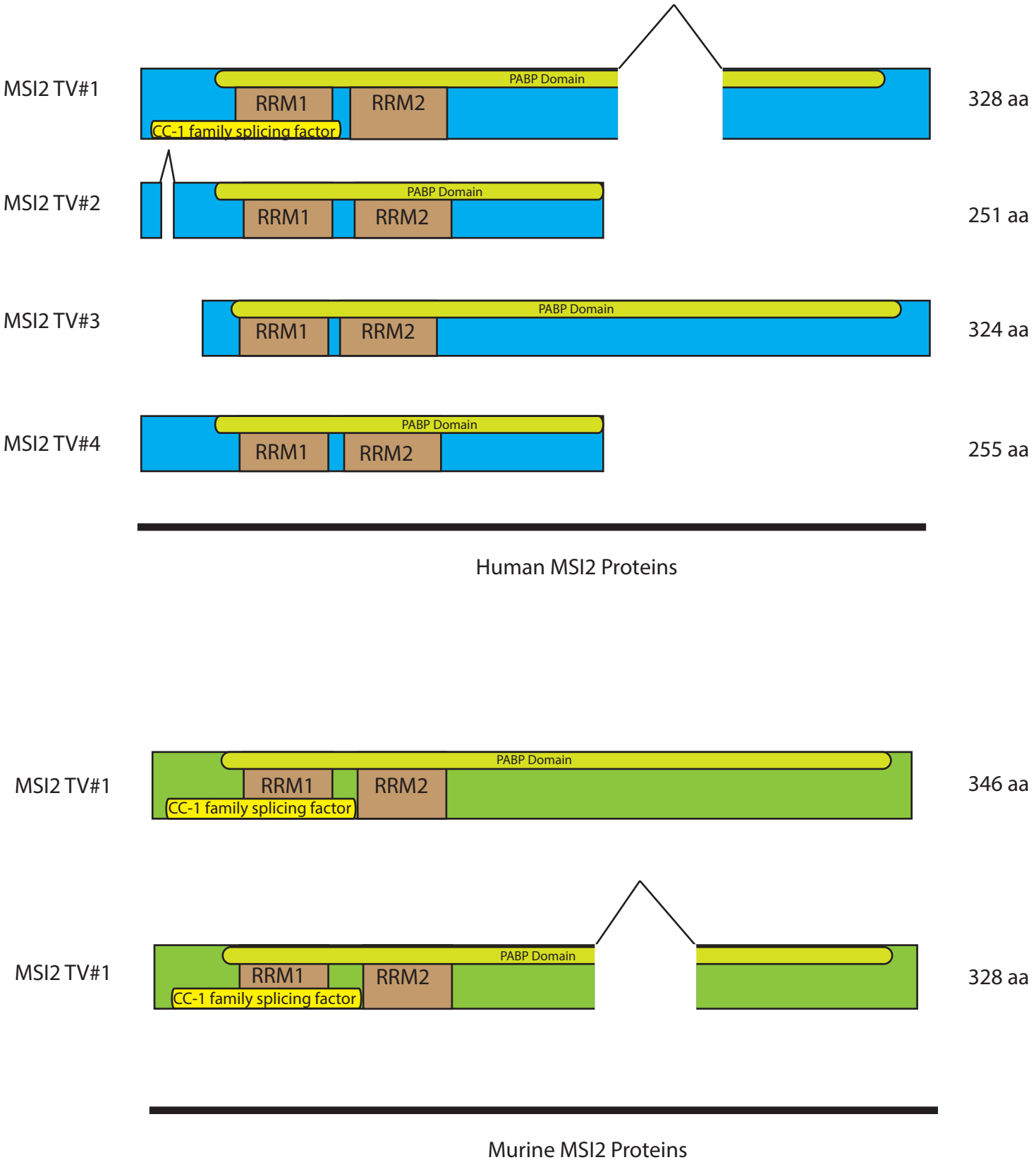


Figure 8. Mouse and Human MSI2 isoforms. 4 MSI2 transcript variants are expressed in human cells and two MSI2 transcript variants are expressed in mouse cells. All MSI2 proteins have two RNA-recognition motifs (RRMs) in their N-terminal region and a general poly-A binding protein domain spanning the C-terminal region of the protein. Interestingly, a CC-1 family splicing factor domain is found in the N-terminal region of many of the MSI2 isoforms. This is closely related to the U2AF family of splicing factors and suggests a possible role for MSI2 in control pre-mRNA processing

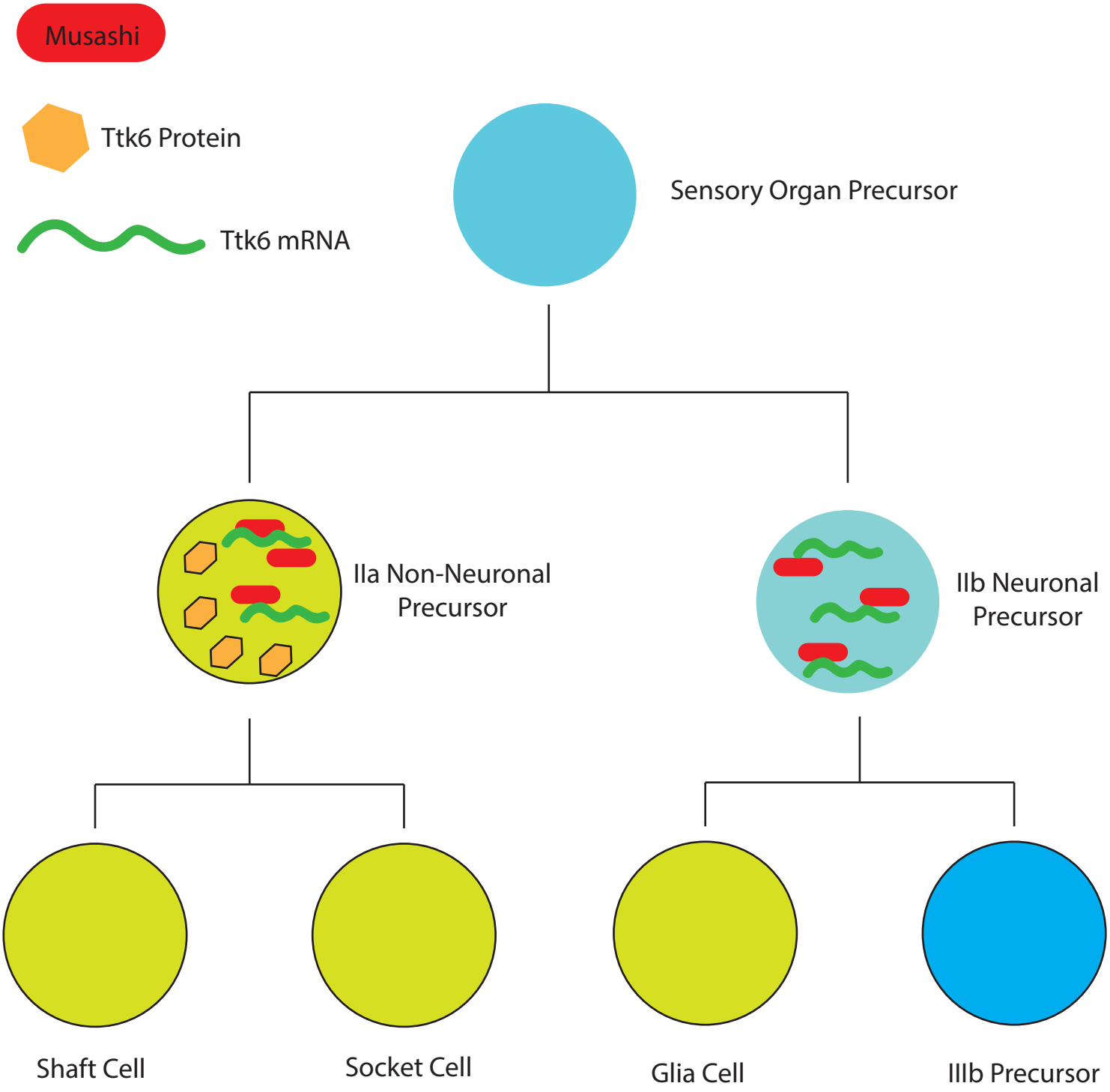


Figure 9. Musashi in the Drosophila mechanosensory bristle. The drosophila mechanosenory bristle consists of 4 non-neuronal support cells and one neuronal cell. These cells originate from a series of assymetric cell division from a common sensory organ precursor (SOP) cell. The first assymetric division results in the formation of a IIa non-neuronal precursor and a IIb neuronal precursor. tramtrack69 (ttk69) is a transcriptional repressor that is sufficient to specify a non-neuronal identity. ttk69 protein is expressed solely in IIa precursors despite there being equal levels of ttk69 mRNA in both IIa and IIb precursors. In IIb precursor cells, Musashi binds to a 3'UTR in the ttk69 mRNA and represses its translation. MSI translational repression does not occur in IIa cells possibly due to the regulation of MSI by Notch signaling.

Musashi Proteins Across Species	
Species	Musashi Family Members
Danio Rerio	Musashi-1 Musashi-2a Musashi-2b
Caenorhabditis Elegans	Musashi
Drosophila Melanogaster	Musashi
Mus Musculus	Musashi-1 Musashi-2
Homo Sapiens	Musashi-1 Musashi-2
Xenopus Laevis	Nrp-1 (Musashi-1) Xrp-1 (Musashi-2)

Figure 10. Musashi proteins across different species. Invertebrate species possess a single Musashi gene while vertebrate species posess both a Musashi-1 and Musashi-2 gene. Evolutionary analysis suggests that these genes arose due to a gene duplication event. Zebrafish have two copies of the MSI2 gene- MSI2a and MSI2b. These are thought to have arisen due to a gene duplication event in teleosts.

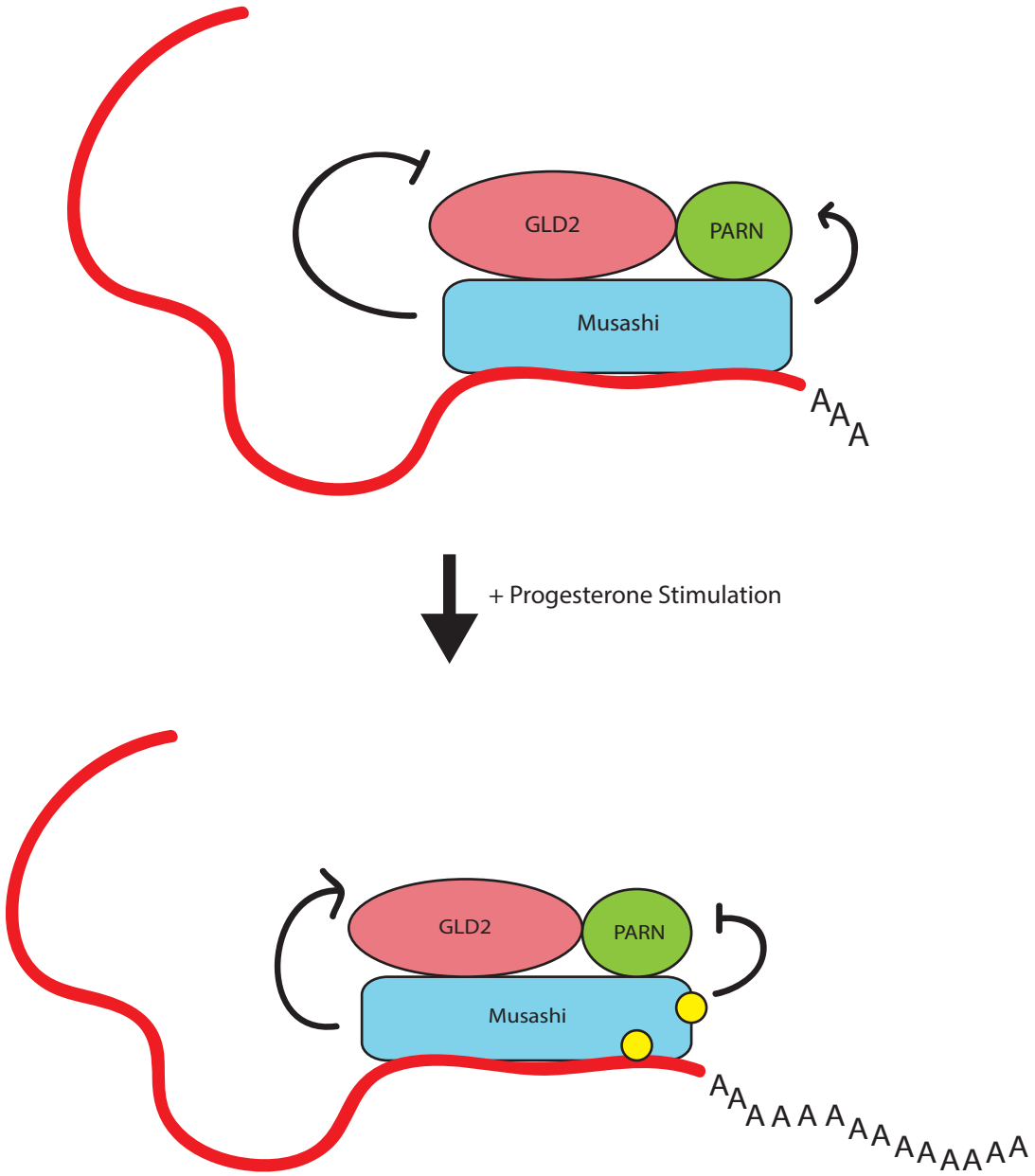


Figure 11. Musashi-1 mechanism of translational activation. Studies in Xenopus identify MSI1 in a complex with the poly(A) polymerase GLD2 and the deadenylase PARN. Phosphorylation of MSI1 triggers the polyadenylation of several target mRNAs trhough the GLD2. When MSI1 is not phosphorylated, the translation of target mRNAs is inhibited to due to action of PARN. The phosphorylation of MSI2 is though the control the activation of GLD2 and inhibition of PARN through unknown mechanisms. The phosphorylation status of MSI1 does not effect the binding of MSI1 to GLD2 or PARN

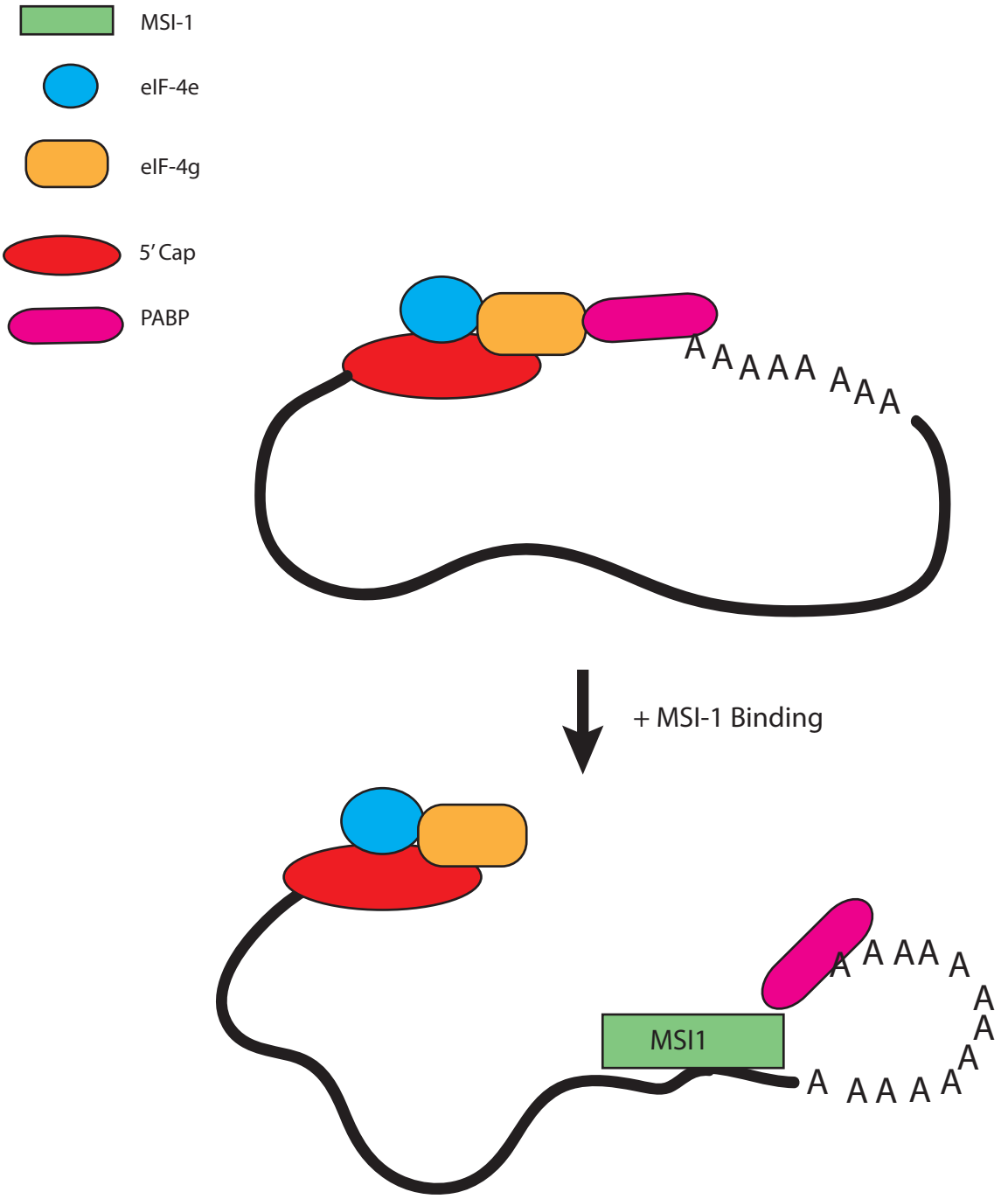


Figure 12. Mechanism of MSI1-mediated translational repression. Translation is critically dependent on the interaction between the poly(A) binding protein (PABP) and eIF4g. This interaction greatly enhances the rates of ribosomal assembly and circularizes the mRNA allowing for a phenomenom known as 'ribosome recycling'. When a ribosome reaches the end of the circularized mRNA it can rapidly reassemble onto the 5' cap and translate a new protein. The MSI1 protein binds to the 3'UTR of target mRNAs where it interacts with the PABP preventing the interaction between eIF4g and PABP.

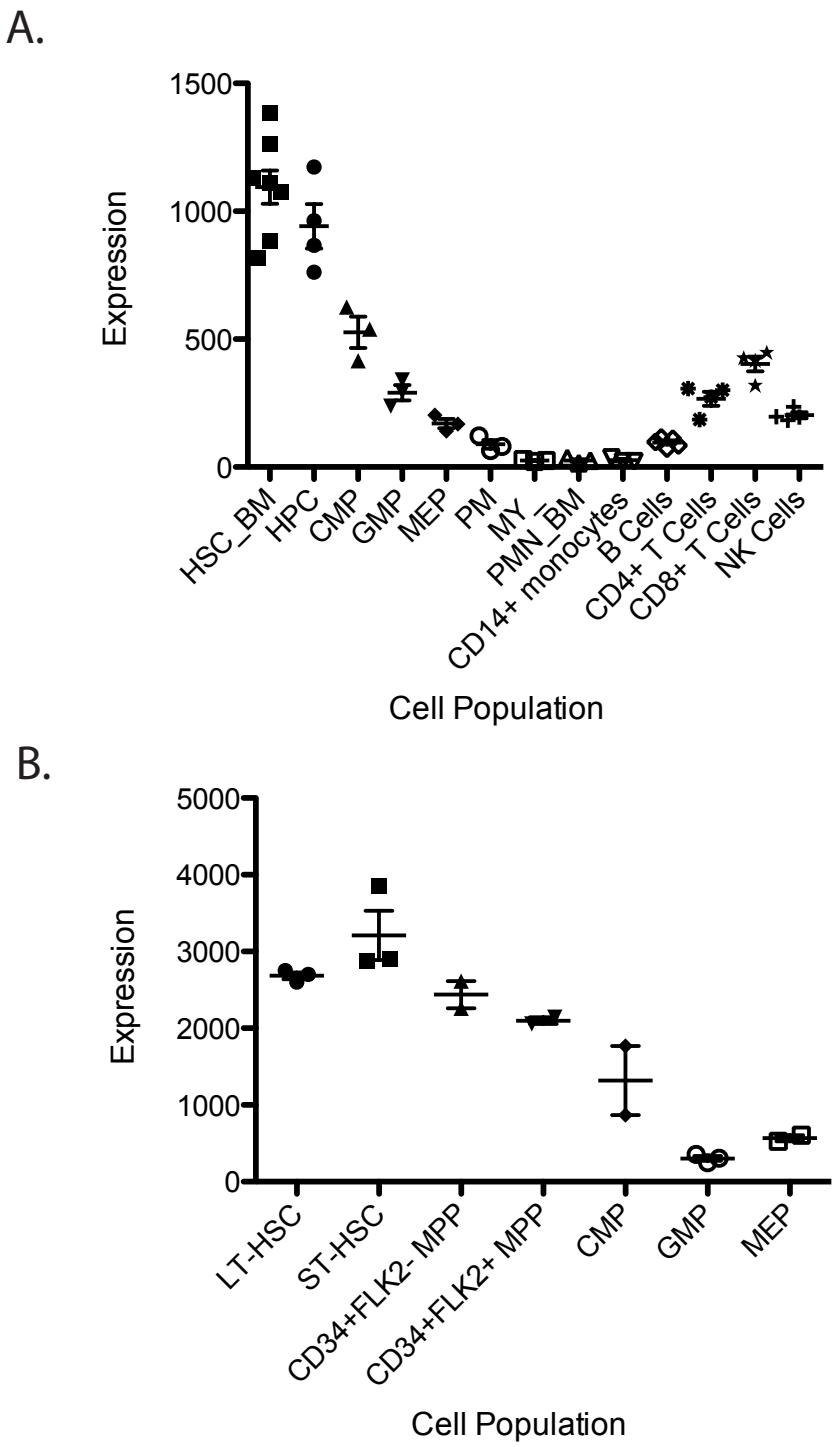


Figure 13. Musashi-2 expression in mouse and human hematopoietic systems. Microarray analysis reveals that Musashi-2 is highly expressed in (A) human and (B) mouse heamtopoetic stem and progenitors cells; HSC_BM=bone marrow heamtopoetic stem cell; HPC=hematopoetic progenitor cells; MPP=multipotent progenitor; CMP= common myeloid progenitor; GMP= granulocyte-monocyte progenitor; MEP=megakaryocyte-erythroid progenitor. For more information regarding these cell populations please refer to Bagger et al., Nucleic Acids Research, 2015.

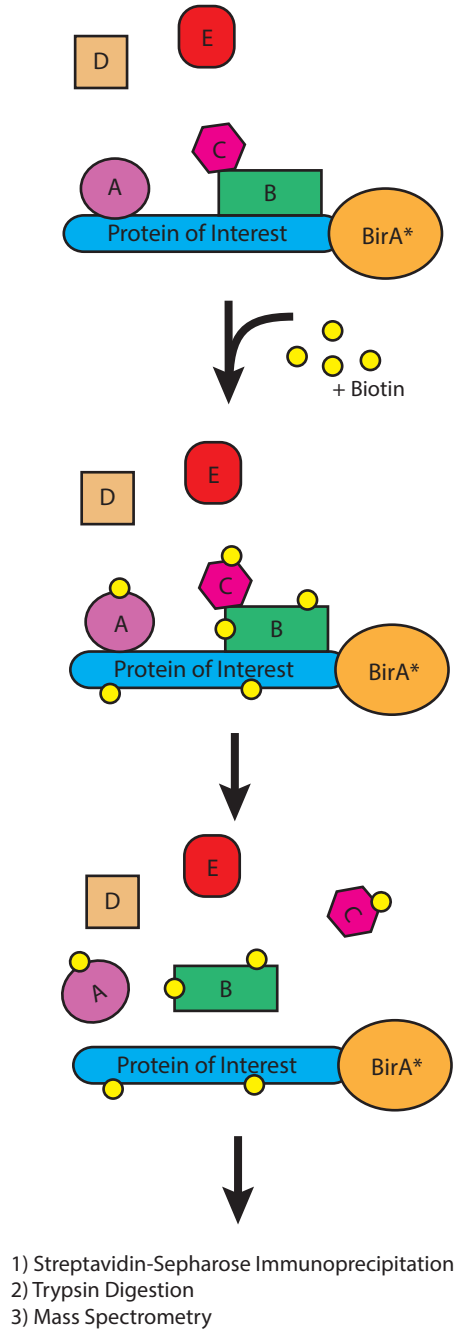


Figure 14. Overview of BioID. A protein of interest is fused to the promiscuous biotin ligase BirA*. After the addition of 50uM biotin, BirA* will biotinylate proteins in a proximity-dependent fashion. BirA* will biotinylate both direct (A, B) and indirect (C) protein binding partners. It will not biotinylate proteins that are not within a close proximity (D, E). After a 24-hour incubation with biotin, cells are lysed in RIPA buffer and biotinylated proteins are immunoprecipitated using streptavidin-sepharose beads. Proteins are digested on-bead and subjected to mass spectrometry analysis

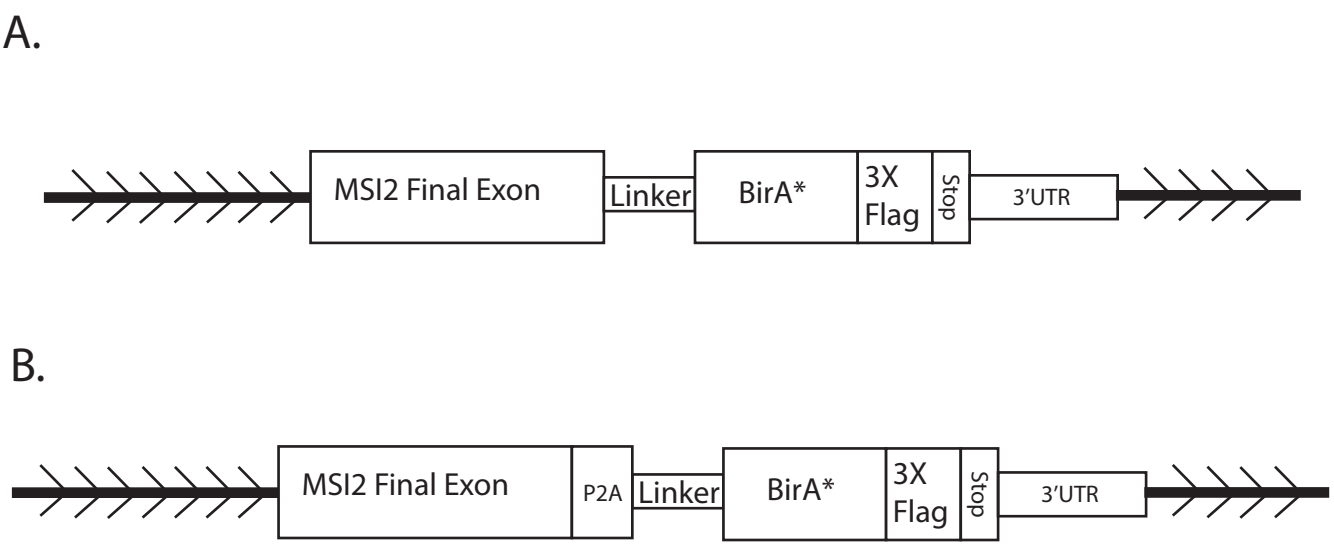


Figure 15. Structure of the endogenous MSI2-BirA* locus. The CRISPR-Cas9 system was used to insert a linker-BirA*-3X-Flag moiety immediately downstream of the final exon of murine MSI2. The stop codon was removed from the MSI2 sequence. Control cell lines had a P2A seqeunce inserted immediately downstream from the final exon of MSI2 (with the stop codon removed) immediately before the linker-BirA*-3Xflag moiety. Upon translation of the MSI2-BirA* mRNA, the P2A sequence will cause the ribosome to skip a peptide bond resulting in the formation of two seperate proteins.

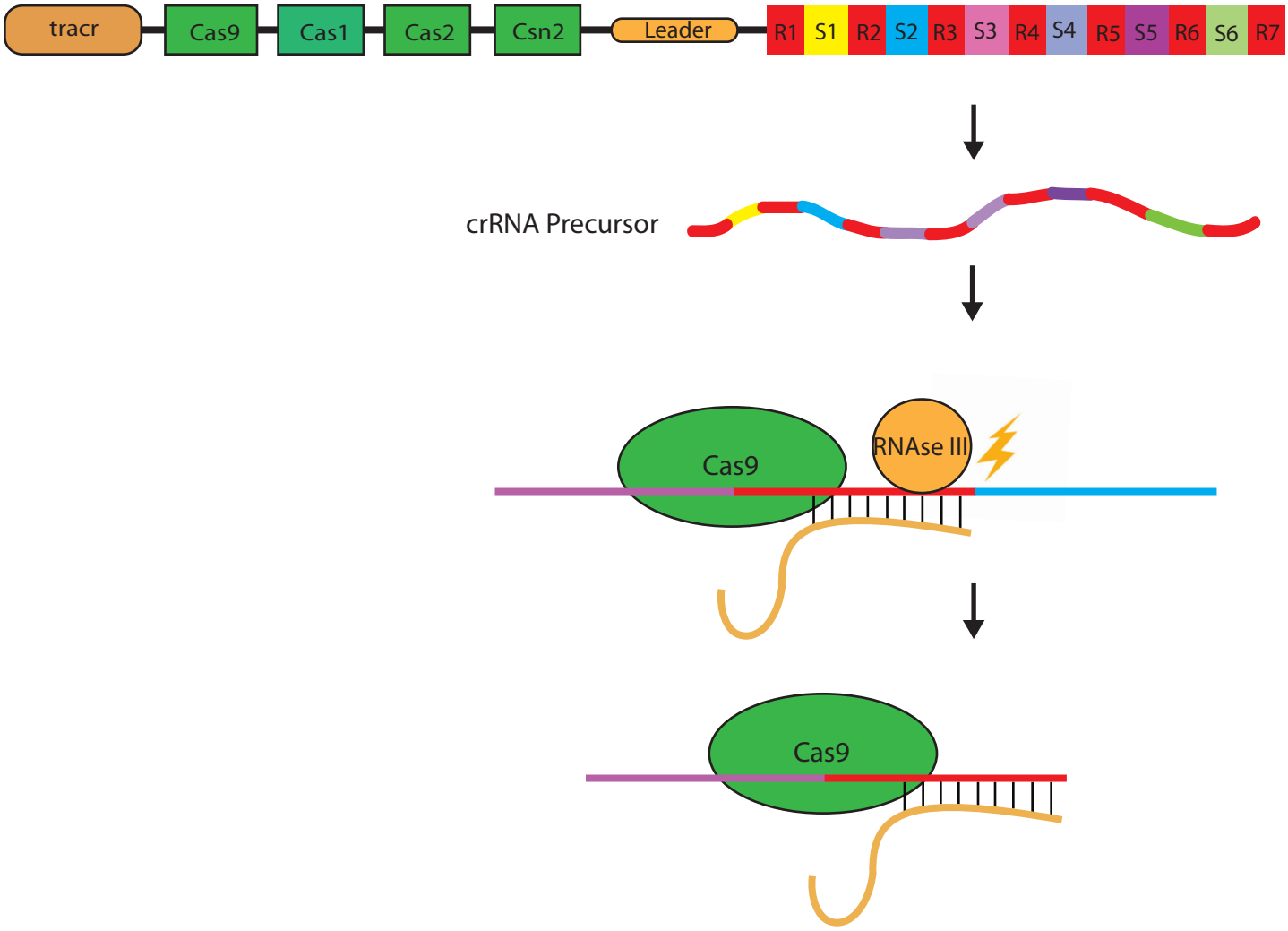


Figure 16. Schematic of the CRISPR locus. The SF370 strain of S. pyogenes contains four protein coding crispr-associated (Cas) genes (Cas1, Cas2, Csn2, and Cas9), a trans-activating CRISPR RNA (tracrRNA), and seven repeated regions surrounding six spacer regions. Transcription of the CRISPR locus is driven from a leader sequence in front of the array. This results in the formation of crRNA precursor (pre-crRNA). Repeats in the pre-crRNA form a double stranded RNA complex with the tracrRNA. This double-stranded RNA complex is bound by Cas9 and RNAseIII resulting in the cleavage of the pre-crRNA in the repeat region. This results in a Cas9-crRNA complex that is free to bind to and cleave target DNA.

Chapter 2:
Investigating the functional role of Musashi-2 in human AML

Abstract

Current chemotherapy regimens used in the treatment of acute myelogenous leukemia (AML) are unable to sufficiently eliminate leukemic stem cells (LSCs). The inability to effectively target these cells is thought to be the driving factor resulting in patient relapse. As a result, the development of targeted chemotherapy that can effectively kill or promote the differentiation of LSCs is an active field of research. The Musashi-2 (MSI2) protein is highly expressed in the most immature fraction of the mouse and human hematopoietic hierarchy and functional studies identify MSI2 as a critical regulator of murine hematopoietic stem cells (HSCs). Though correlative studies exist that implicate MSI2 in the formation of aggressive myeloid leukemia in humans, no functional studies to date have characterized the role of MSI2 in primary human AML.

Here we demonstrate that *Msi2* is preferentially expressed in LSC vs. leukemic blasts and a loss of MSI2 can greatly impair the function of leukemic cells in a preclinical xenotransplantation model of human AML. Moreover, progenitor cell assays reveal the severe functional defects resulting from a loss of MSI2 protein may be unique to their role in LSCs. This work demonstrates the critical role of MSI2 in human LSCs and suggests that the MSI2 protein/ pathway may be an important target for the development of LSC-effective chemotherapy.

Introduction

AML is a cancer of the hematopoietic system characterized by the accumulation of immature and dysfunctional hematopoietic cells. AML is thought to arise due to a series of transforming mutations in either an HSC or early progenitor that results in the cell gaining aberrant self-renewal capabilities[1]. This cell is referred to as a leukemic stem cell and is thought to play a critical role in the maintenance of leukemic disease. In a manner akin to the normal hematopoietic system, AML is organized as a hierarchy with LSCs at the apex[2]. LSCs are thought to differentiate into highly proliferative progenitor cells (AML-CFUs) and eventually into partially differentiated and dysfunctional blast cells. These LSCs are relatively quiescent and are thought to be responsible for relapse due to their inability to be effectively targeted by standard chemotherapy[3]. In the murine setting, the RNA binding protein MSI2 is a critical regulator of normal HSCs[4,5]. Furthermore, mouse studies implicate MSI2 in the development of aggressive leukemic disease and have functionally demonstrated that a loss of MSI2 can impair leukemic growth[6]. Although high levels of MSI2 correlate with AML patient outcomes and overall survival, no study to date has characterized the functional role of MSI2 in human AML[4,7]. MSI2 is a post-translational regulator of gene expression[8]. It is thought to bind specifically to a select pool of RNA targets in order to control their rates of translation. Importantly, by demonstrating a critical role for MSI2 in human LSCs, one can propose that RNA targets bound by MSI2 may represent critical regulators of leukemogenesis. Thus, the MSI2 protein may serve as a stepping-stone in understanding the pathways and proteins that are critical for the process of normal and leukemic stem cell self-renewal. Notably, studies analyzing the common

molecular programs downstream of MSI2 in the murine setting have identified TSPAN3 as a MSI2 target that is a critical protein required for the development and propagation of AML in mouse models[9].

In the human setting, we recently reported on the role of MSI2 in human cord blood-derived HSCs[10]. Here we showed that *Msi2* is most highly expressed in the immature fraction of the hematopoietic hierarchy and its levels decrease upon differentiation. We further showed that an overexpression of *Msi2* promotes numerous self-renewal phenotypes including the *in vitro* expansion of short-term repopulating cells, enhanced progenitor cells activity, and a 23-fold *ex vivo* expansion of LT-HSCs, likely through an attenuation of the aryl hydrocarbon signalling axis. We suggest that MSI2 functions in a similar manner to promote enhanced self-renewal in the context of AML.

In the current body of work, we show that MSI2 is preferentially expressed in LSCs compared to leukemic blasts and that its knockdown significantly impairs LSC engraftment capacity. We further carried out progenitor assays that demonstrate a unique role of MSI2 in the LSC compartment specifically.

Materials and Methods

Mice

NOD-*scid-IL2Rγc*$^{-/-}$ (NSG) (Jackson Laboratory) mice were bred and maintained in the Stem Cell Unit animal barrier facility at McMaster University. All procedures received the approval of the Animal Research Ethics Board at McMaster University

AML Samples

All AML patient samples were obtained with informed consent and with the approval of the local human subjects research ethics board at the University Health Network. AML samples are described in Figure 1.

Generation of shMSI2 lentiviral hairpins

MSI2 shRNAs were designed with the Dharmacon algorithm (http://www.dharmacon.com). Predicted sequences were synthesized as complimentary oligonucleotides, annealed and cloned downstream of the H1 promoter of the modfied cppt-PGK-EGFP-IRES-PAC-WPRE lentiviral expression vector[11]. Sequences for the MSI2 targeting and control RFP targeting shRNAs were as follows: shMSI2, 5'-gagagatcccactacgaaa-3'; shRFP, 5'-gtgggagcgcgtgatgaac-3' (Designed by Muluken Belew). All lentivirus was prepared by transient transfection of 293FT cells with pMD2.G and psPAX2 packaging plasmids (Addgene) to create VSV-G pseudotyped lentiviral particles. All viral preparations were titrated on HeLa cells before use. Standard SDS-PAGE and western blotting procedures were performed to validate the effect of knockdown on transduced HeLa cells and the human promyelocytic leukemia cell line, NB4 (Figures 4 and 5). Immunoblotting was performed with anti-MSI2 rabbit monoclonal IgG (EP1305Y, Epitomics) and β-actin mouse monoclonal IgG (ACTBD11B7, Santa Cruz Biotechnology) antibodies. Secondary antibodies used were IRDye 680 goat anti-rabbit IgG and IRDye 800 goat anti-mouse IgG (LI-COR).

Lentiviral Infection of AML Samples and Xenotransplantation

Prior to transduction with MSI2 knockdown lentivirus, AML cells were flow sorted based on CD34 and CD38 expression to enrich for LSC activity. Transplantation assays performed previously had identified LSC-containing fractions and frequencies for all AML samples used. AML cells were infected at an MOI of 50 for 24-hours in StemSpan medium (StemCell Technologies) supplemented with growth factors Interleukin 6 (IL-6; 20 ng/ml, Peprotech), Stem cell factor (SCF; 100 ng/ml, R&D Systems), Flt3 ligand (FLT3-L; 100 ng/ml, R&D Systems) and Thrombopoietin (TPO; 20 ng/ml, Peprotech). Post infection cells were washed and transplanted intra-femorally at 50,000-200,000 cells (cell dose matched for each experiment between shMSI2 and shControl) per sub-lethally irradiated (315 cGy) NSG mouse. Three months post transplant, mice were sacrificed and bone marrow from tibias, femurs and pelvis was harvested. Human AML engraftment was analyzed by first blocking reconstituted mouse bone marrow with mouse Fc block (BD Biosciences) and human IgG (Sigma), followed by staining with fluorochrome-conjugated antibodies against human CD45 (HI30), CD33 (P67.6, BD Biosciences), CD14 (HCD14), and CD15 (MMA).

CFU Assays

AML-CFU assays were performed in semi-solid ColonyGel media (Human Complete Media 1102; reachBio) with flow-sorted GFP+ MSI2 knockdown AML cells. 10000 GFP+ cells were plated in duplicate in each well of a 24-well plate and loose colonies consisting of 10 or more cells were counted 7 and 14 days post plating.

Intracellular Flow cytometry

Primary AML cells were initially stained with anti-CD34 PE (581, BD Biosciences) antibody and LIVE/DEAD Fixable Violet (Invitrogen) and then fixed with the Cytofix/Cytoperm kit (BD Biosciences) according to the manufacturer's instructions. Fixed and permeabilized cells were immunostained with anti-MSI2 rabbit monoclonal IgG antibody (EP1305Y, Abcam) and detected by Alexa-488 goat anti-rabbit IgG antibody (Invitrogen).

NB4 ATRA Treatment

NB4 cells were cultured in IMDM 10% FBS over 10 days with 1 μM ATRA or 0.1% DMSO. At days 0, 5, 6, 7 and 10 cells were sampled for anti-MSI2 immunobloting and cytospinning followed by Wright-Giemsa staining to detect differentiation. Wright-Giemsa prepared cytopsins were imaged with an Aperio CS2 slide scanner (Leica) at 20x magnification and subsequent image processing was performed with ImageJ software (NIH) (*This experiment was performed by Muluken Belew)

qRT-PCR Analysis of mobilized peripheral blood and AML Samples

Healthy MPB samples were obtained with informed consent and with the approval of the local human subject research ethics board at McMaster University. Primary samples were thawed in PBS 10% FBS with 100 μg/ml DNAse. For qRT-PCR analysis total cellular RNA was isolated with Trizol LS reagent (Invitrogen) according to the manufacturer's instructions and cDNA was synthesized using qScript cDNA Synthesis Kit (Quanta Biosciences). qRT-PCR was done in triplicate with PerfeCTa

qPCR SuperMix Low ROX (Quanta Biosciences) with gene specific probes (Universal Probe Library, UPL, Roche) and primers: MSI2 UPL-26, F- ggcagcaagaggatcagg, R-ccgtagagatcggcgaca and GAPDH UPL-60, F-agccacatcgctcagacac, R-gcccaatacgaccaaatcc;. The mRNA content of samples compared by qRT-PCR was normalized based on the amplification of GAPDH.

Results and Discussion

MSI2 expression correlates with the immature fraction of AML

To characterize the expression of MSI2 in human AML samples, we analyzed *Msi2* transcript levels across a panel of 22 AMLs of varying subtypes *via* qPCR (Figure 1, p.104). MSI2 expression levels across these AML samples were normalized to the expression of MSI2 in Lin- mobilized peripheral blood (Figure 2, p.105). Surprisingly, we saw no significant enrichment in MSI2 levels when profiled in bulk AML. Instead, it appeared that many AML samples had lower MSI2 levels when compared to Lin- mobilized peripheral blood. As mentioned however, AMLs are organized in a hierarchical manner and are driven by a small population of LSCs which generate the bulk of the AML cell population comprised of non-proliferative, partially differentiated blast cells[2]. Importantly, many AML blasts display markers of differentiation on their surface such as CD13 and CD33[1]. It is likely that these partially differentiated cells display decreased levels of MSI2 when compared to the more immature leukemic cells that are thought to maintain this disease. Importantly, any clinical significance of MSI2 is likely to derive from its presence in LSCs and not in bulk AML as it is the LSCs that are thought to maintain leukemia and

drive its progression[7]. To first answer the question whether MSI2 is enriched in LSCs, we examined whether MSI2 levels correlate with stemness in a cohort of AML samples where immunophenotypically distinct subsets had been experimentally validated for LSC content (unpublished dataset from Dr. John Dick). Briefly, human AML samples were fractionated based on CD34 and CD38 expression and transplanted into immunocompromised mice. Those AML fractions that were able to repopulate immunocompromised mice were said to contain LSCs and microarray analysis was performed on this "LSC" fraction separately from those fractions that did not yield any repopulation upon transplantation (the "non-LSC" fractions). MSI2 was found to have a 1.6-fold increase in transcript expression in LSC-enriched populations relative to non-LSC containing cell populations. We further performed intracellular flow cytometry to probe for MSI2 protein levels in CD34+ and CD34- fractions of human primary AML samples, since CD34 is a common marker that enriches for LSC activity[12]. Across 21 AML samples, we saw a significant increase in the percent of MSI2 positive cells between CD34+ and CD34- populations (Figure 3, p.106). In addition to performing these studies with primary samples, we also examined the human promyelocytic leukemia cell line, NB4, to further correlate the expression of MSI2 with immature AML cells. Here we treated the NB4 human promyelocytic leukemia cell line with all-trans retinoic acid (ATRA), a molecule that is known to stimulate the granulocytic maturation of these cells[13]. An impressive down-regulation of MSI2 protein levels were observed over a 10-day period of differentiation suggesting that MSI2 expression is preferentially elevated in immature leukemic cells (Figure 4, p.107). This down regulation of MSI2 in response to a known AML treatment

strategy in combination with its preferential expression in the most immature fraction of human AMLs led us to investigate what functional role, if any, the MSI2 protein had in human AML.

Knockdown of MSI2 in human AML impairs repopulation in xenotransplants

To characterize the functional effect of a loss of MSI2 in human primary leukemic cells, we infected primary AML samples with lentiviral particles containing short hairpins targeting MSI2 (shMSI2) or RFP (shControl). The lentiviral plasmid was designed to drive a hairpin of interest from an H1 promoter and drive the expression of eGFP from a PGK promoter (Figure 5, p.108). AML samples were consistently infected to a minimum of 60% and immediately transplanted into sub-lethally irradiated NSG mice 24 hours post transduction. Three months post-transplantation, mice were culled, bone marrow cells were isolated and the human graft was analyzed. GFP levels were examined in the human myeloid graft in order to elucidate the impact that a hairpin targeting MSI2 had on leukemic cell growth (Figure 6, p.109). All AMLs were infected at a rate of greater than 60% (Figure 7, p.110). One AML sample, #090191 showed exceptional results, indicating that a loss of MSI2 in this human AML greatly impairs the ability of these cells to maintain the disease (Figure 8, p.111). When human CD33+ cells were analyzed in control (shRFP)-infected cells, GFP levels were similar to initial infection rates. However when cells were transduced with a hairpin that knocked down MSI2 (sh332) there was a dramatic decrease in GFP levels suggesting that cells infected with shMSI2 are unable to repopulate immunocompromised mice. This effect was observed when the

same AML sample was infected with another MSI2-specific hairpin (sh541) suggesting that the impairment in leukemic growth is in fact due to a decrease in MSI2 protein levels. Three other AML samples, #0840, #100091, and #0596, were also efficiently transduced and engrafted in NSG mice. One startling effect seen in these samples however, was the silencing of GFP expression. Despite >70% of AML cells being initially infected, the CD45+CD33+ populations in these grafts showed almost no GFP expression, 3 months post-transplantation (Figure 9, p. 112). Despite the lack of GFP expression, when the percentage of CD45+CD33+ cells in each graft was analyzed 3-months post transplantation, the initial knockdown of MSI2 levels in these 3 primary AMLs appeared to significantly decrease their ability to repopulate NSG mice when compared to control (Figure 9, p.112). Altogether, across these 3 AML samples, we saw a significant ($p<0.01$) 6.8 fold higher level of human engraftment in shControl infected AML cells when compared to shMSI2-infected cells. Overall, across a diverse set of 4 AMLs, a loss of MSI2 appeared to significantly impair *in vivo* reconstitution implicating this protein as a functional regulator of leukemic growth.

To investigate the impact that a loss of MSI2 has on AML progenitor activity, 4 AML samples, #0596, #110751, #100753, and #090191 were infected with shRFP or shMSI2 (sh332) lentiviral particles, sorted 3 days post-infection, and 25, 000 GFP-positive cells were plated in duplicate. AML colonies (defined as clusters of 10 or more loosely packed cells) were counted at 7 and 14 days post-plating; normal colonies were disregarded. Remarkably, no significant changes in AML-CFU colony forming potential was observed suggesting that MSI2 does not significantly alter the

function of AML progenitors (Figure 10, p. 113). Despite the expression of MSI2 in these samples, it does not appear to be required for their clonogenicity or viability. The ability of MSI2 knockdown to impact the repopulating activity of primary AML samples despite the absence of any appreciable effect of its reduction on AML-CFU activity suggests that MSI2 plays a unique role in the maintenance of LSC activity.

Silencing of transgene expression in xenotransplanted human AML samples

The silencing of GFP expression in transduced AML samples proved to be a recurring issue that impaired the analysis of lentivirally transduced AML cells. Importantly, promoter hypermethylation is known to play a defining role in altering gene expression patterns in numerous cancers; it has been described as a hallmark of cancer[14]. It is thus a possibility that aberrant methylation programs that are active in certain AML samples may contribute to the silencing of transgene expression *in vivo*. Furthermore, several studies have indicated that lentiviral silencing can occur in long-term cultures and upon differentiation[15-17]. It is a possibility that GFP expression remained throughout the early stages of the xenotransplantation but eventually became silenced throughout the latter stages of the 3-month transplant. The choice of promoter may be a critical factor in order to maintain long-term transgene expression and the optimal promoter may differ across the various AML samples. The heterogeneity of AML samples cannot be overstated. At the molecular and clinical levels, different AML samples can be viewed as individual diseases each with their own unique pattern of gene expression that reflects altered states of differentiation. Importantly, studies in

mouse embryos have demonstrated the necessity to optimize the choice of lentiviral promoter based on the differentiation state[15]. Furthermore, other work demonstrates that the differentiation process can impair transgene expression even if a transgene can optimally be expressed in the more differentiated cell population[15]. In one study researchers demonstrated that in mESCs infected with a GFP driven by an EF1a promoter, GFP levels were dramatically diminished upon the differentiation of these cells into neural progenitors and neural cells *in vitro*[15]. Despite this, neural progenitor cells that were transduced with an EF1a-GFP construct showed sustained expression of GFP for at least 3 weeks. This demonstrates that even though differentiated cells can express high levels of a lentiviral transgene, the expression of such a transgene can be dramatically silenced when a cell's fate is being altered. In a similar manner, many of our AML samples were transduced at high efficiency and showed significant levels of transgene expression upon initial transduction. Although the majority of the AML blasts that were initially transplanted showed stable transgene expression, it is possible that the AML cells that repopulated NSG mice were derived from LSC and progenitor cells that underwent aberrant differentiation resulting in transgene silencing.

Importantly, studies have demonstrated that robust expression of a lentiviral transgene is dependent on the promoter driving the transgene expression. In one study, VSV-G pseudotyped lentiviral particles expressing Luciferase under the control of a CMV promoter (LV-CMV-Luc) or an MHCII promoter (LV-MHC-IILuc) were intravenously injected into C57BL/6 mice[18]. Results demonstrated a potent silencing of Luciferase expression in the spleen of these mice when driven off of the

CMV promoter but no difference in Luc expression in the spleen when driven off of the MHCII promoter, suggesting that the MHCII promoter was more persistent in maintaining transgene expression in the spleen. Perhaps the use of a more biologically relevant promoter could help in the maintenance of long-term transgene expression in AML samples. Another study demonstrated that efficient lentiviral transgene expression in bone marrow-derived endothelial progenitor cells was highly dependent on the appropriate gene promoter[19]. Here, GFP levels were driven off of EF1a, PGK, and CMV promoters. The expression of GFP from the CMV promoter was remarkably enhanced when compared to EF1a and PGK promoters. Furthermore, *in vitro* culture conditions were shown to dramatically affect the expression of cells transduced with a CMV-GFP construct. By altering the presence of growth factors when cultured *in vitro*, a 100-fold difference in the level of eGFP expression was noted that was not due to differences in lentiviral transduction efficiency[19]. Evidently, the *in vitro* transduction conditions can play a critical role in the stable expression of a lentiviral transgene. Perhaps optimizing the culture conditions in which AML samples are transduced may lead to a more robust and long-term transgene expression.

MSI2 impairs human leukemic stem cell function

Despite technical difficulties with these AML xenotransplantation assays, interesting observations were made. We showed that by knocking down *Msi2* expression levels MSI2 in human AML samples, we could impair their growth in xenotransplantation assays. Furthermore, a loss of MSI2 in AML samples did not

appear to inhibit AML progenitor capacity suggesting that a loss of MSI2 has a specific role in impairing LSCs. Interestingly, this data suggests that AML-CFU reduction can not act as a proxy for LSC killing and supports the idea that *in vivo* assays are crucial for measuring the impact of a treatment on LSC function. Our data is the first to show a functional role of MSI2 in primary human AML. Similar studies have demonstrated a functional role of MSI2 in mouse models of leukemia. This in conjunction with numerous correlative studies indicating that MSI2 promotes aggressive myeloid leukemia and poor patient outcomes implicates the MSI2 protein as a critical regulator of LSC biology. Further elucidation of the pathways and regulators involved in MSI2 function are likely to uncover novel and potent regulators of leukemic transformation. Of important, note, MSI2 is thought to regulate the translation of numerous target mRNAs. It is likely that the functional role of MSI2 in LSCs is due, at least in part, to the regulation of these target mRNAs. Perhaps MSI2 is regulating the expression of a set of functionally regulated genes that are critical for self-renewal and the maintenance of LSC quiescence. By identifying these target mRNAs, we may be able to identify sets of proteins that are critical regulators of leukemic stem cells. This will not only help to identify mechanisms that are active in the maintenance of LSCs but could perhaps identify novel drug targets that can effectively target LSCs.

References

1 Dohner, H., Weisdorf, D. J. & Bloomfield, C. D. Acute Myeloid Leukemia. *N Engl J Med* **373**, 1136-1152, doi:10.1056/NEJMra1406184 (2015).

2 Hope, K. J., Jin, L. & Dick, J. E. Acute myeloid leukemia originates from a hierarchy of leukemic stem cell classes that differ in self-renewal capacity. *Nat Immunol* **5**, 738-743, doi:10.1038/ni1080 (2004).

3 Pollyea, D. A., Gutman, J. A., Gore, L., Smith, C. A. & Jordan, C. T. Targeting acute myeloid leukemia stem cells: a review and principles for the development of clinical trials. *Haematologica* **99**, 1277-1284, doi:10.3324/haematol.2013.085209 (2014).

4 Kharas, M. G. *et al.* Musashi-2 regulates normal hematopoiesis and promotes aggressive myeloid leukemia. *Nat Med* **16**, 903-908, doi:10.1038/nm.2187 (2010).

5 Hope, K. J. *et al.* An RNAi screen identifies Msi2 and Prox1 as having opposite roles in the regulation of hematopoietic stem cell activity. *Cell Stem Cell* **7**, 101-113, doi:10.1016/j.stem.2010.06.007 (2010).

6 Ito, T. *et al.* Regulation of myeloid leukaemia by the cell-fate determinant Musashi. *Nature* **466**, 765-768, doi:10.1038/nature09171 (2010).

7 Byers, R. J., Currie, T., Tholouli, E., Rodig, S. J. & Kutok, J. L. MSI2 protein expression predicts unfavorable outcome in acute myeloid leukemia. *Blood* **118**, 2857-2867, doi:10.1182/blood-2011-04-346767 (2011).

8 Sakakibara, S., Nakamura, Y., Satoh, H. & Okano, H. Rna-binding protein Musashi2: developmentally regulated expression in neural precursor cells and subpopulations of neurons in mammalian CNS. *J Neurosci* **21**, 8091-8107 (2001).

9 Kwon, H. Y. *et al.* Tetraspanin 3 Is Required for the Development and Propagation of Acute Myelogenous Leukemia. *Cell Stem Cell* **17**, 152-164, doi:10.1016/j.stem.2015.06.006 (2015).

10 Rentas, S. *et al.* Musashi-2 attenuates AHR signalling to expand human haematopoietic stem cells. *Nature* **532**, 508-511, doi:10.1038/nature17665 (2016).

11 Doulatov, S. *et al.* PLZF is a regulator of homeostatic and cytokine-induced myeloid development. *Genes Dev* **23**, 2076-2087, doi:10.1101/gad.1788109 (2009).

12 Bonnet, D. & Dick, J. E. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat Med* **3**, 730-737 (1997).

13 Idres, N., Benoit, G., Flexor, M. A., Lanotte, M. & Chabot, G. G. Granulocytic differentiation of human NB4 promyelocytic leukemia cells induced by all-trans retinoic acid metabolites. *Cancer Res* **61**, 700-705 (2001).

14 Herman, J. G. & Baylin, S. B. Gene silencing in cancer in association with promoter hypermethylation. *N Engl J Med* **349**, 2042-2054, doi:10.1056/NEJMra023075 (2003).

15 Hong, S. *et al.* Functional analysis of various promoters in lentiviral vectors at different stages of in vitro differentiation of mouse embryonic stem cells. *Mol Ther* **15**, 1630-1639, doi:10.1038/sj.mt.6300251 (2007).

16 Herbst, F. *et al.* Extensive methylation of promoter sequences silences lentiviral transgene expression during stem cell differentiation in vivo. *Mol Ther* **20**, 1014-1021, doi:10.1038/mt.2012.46 (2012).
17 Pfaff, N. *et al.* A ubiquitous chromatin opening element prevents transgene silencing in pluripotent stem cells and their differentiated progeny. *Stem Cells* **31**, 488-499, doi:10.1002/stem.1316 (2013).
18 Kimura, T. *et al.* Lentiviral vectors with CMV or MHCII promoters administered in vivo: immune reactivity versus persistence of expression. *Mol Ther* **15**, 1390-1399, doi:10.1038/sj.mt.6300180 (2007).
19 Liu, J. W. *et al.* Promoter dependence of transgene expression by lentivirus-transduced human blood-derived endothelial progenitor cells. *Stem Cells* **24**, 199-208, doi:10.1634/stemcells.2004-0364 (2006).

AML ID	%CD34	De Novo vs. Secondary	Source	Molecular Abberations	Cytogenetics	Engrafting	LSC in +/-
596	53	De Novo	Diagnosis	NPM1-FLT3-		Yes	Yes
840	56	De Novo	Diagnosis	Unknown		Yes	Yes
100091	10	De Novo	Relapse	NPM1+FLT3-TKD+		Yes	Yes
90191	95	De Novo	Diagnosis		46,XY,ider(7)(q10)del(7)(q21)	Yes	Yes
5732	93				Abnormal 5 AND Abnormal 7 AND 43~45,X,-Y,t(1;5)(q21;p13),t(1;6)(q21;p25),-2,add(2)(p21),add(3)(q27),del(3)(q27),add(4)(p16),del(4)(q33),del(5)(q31),add(6)(p25),del(6)(q21)	Unknown	
80043	50				47,XY,+4[20]	No	
80554	<1					No	
90202	<1					No	
90501	44					No	
90543	38	De Novo	Diagnosis		46,XY,inv(3)(q21q26.2),t(9;22)(q34;q11.2)[9]/46,XY[1] AND 9;22, Ph+	Yes	No
526386	97					No	
90620	61				45,X,-Y[13]/46,XY[7]	No	
90703	88					No	
90765	80					No	
90784	<1			NPM1+ FLT3-ITD + (high)		No	
100596	16			NPM1+ FLT3-TKD+		No	
100622	46				46,XY,inv(16)(p13.1q22)[10]/46,XY[6] AND INV (16)	No	
100753	83					No	
100808	<1					No	
100857	90				Trisomy 8 AND 46,XY,+8[9]	No	
110751	30					No	
120150	60					No	

Figure1. Description of AML Samples. The following table is a list of the AML samples that were used in these studies. 4 of the samples engrafted in immune-compromised mice and were thus used for transplantation studies. Molecular abberations and cytogenetics are provided, otherwise AMLs are cytogenetically normal.

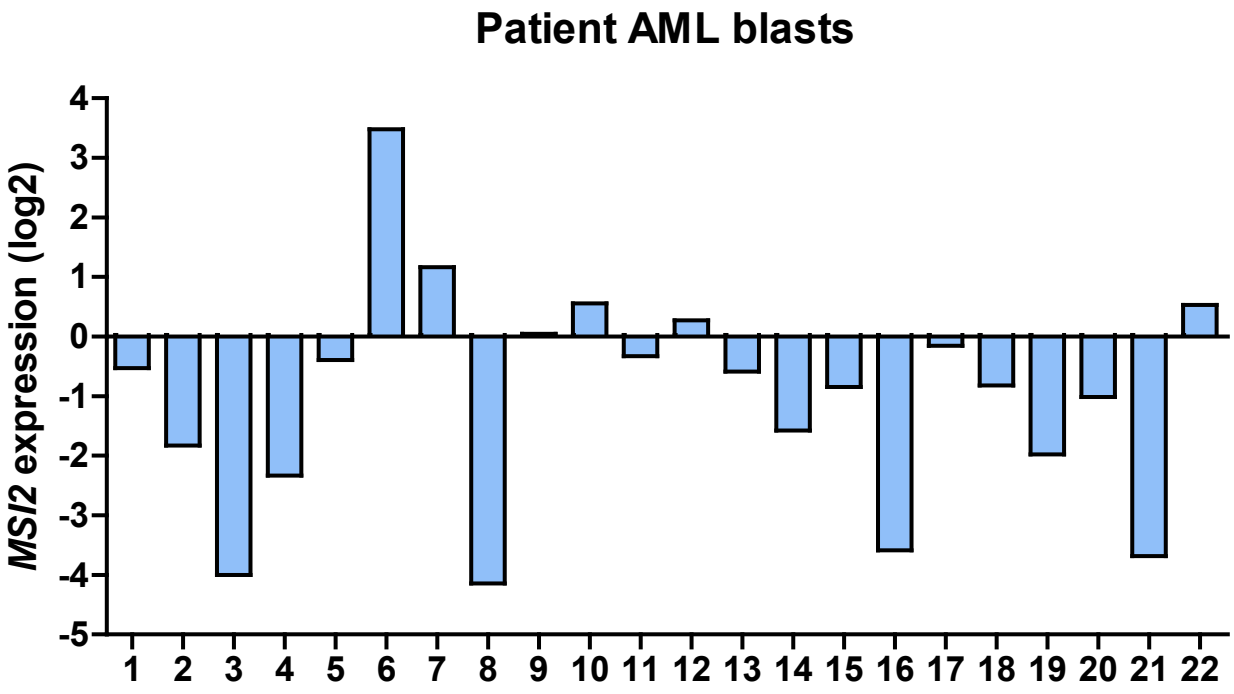


Figure2. qRT-PCR analysis of MSI2 levels in human AML samples. MSI2 qRT-PCR analysis was performed across 22 different AML samples and normalized to the expression of GAPDH. Expression was graphed relative to MSI2 levels in Lin- mobilized peripehral blood.

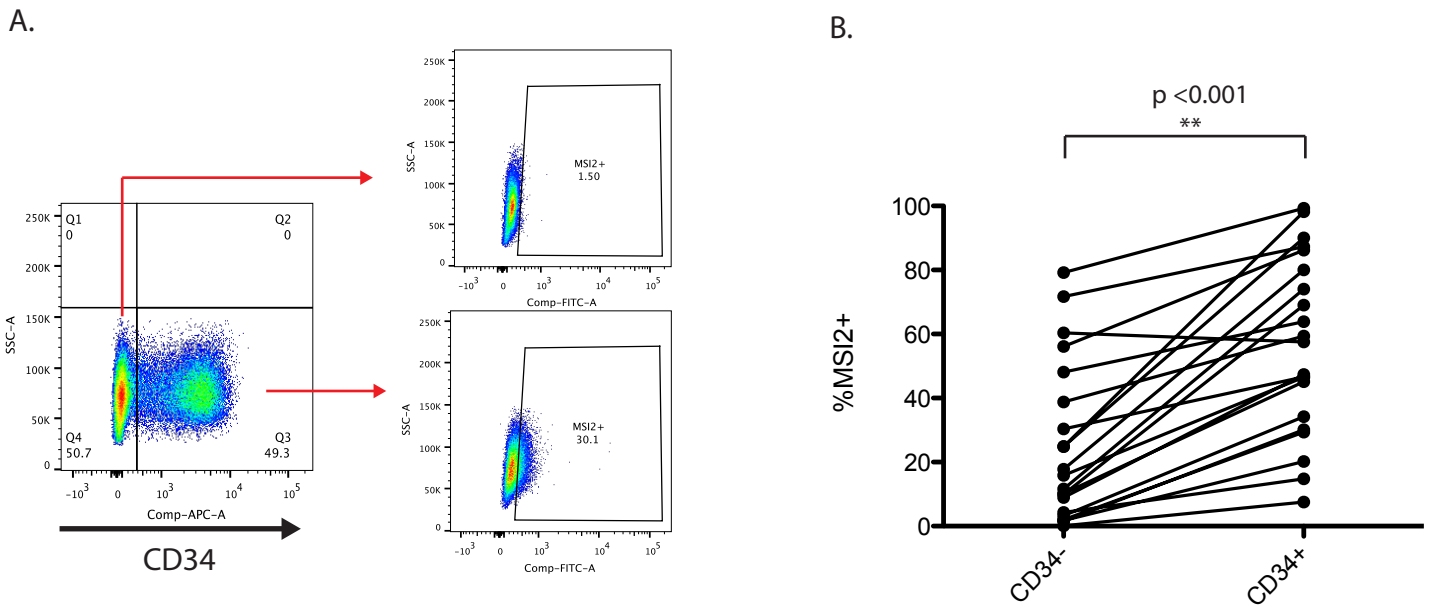


Figure 3. MSI2 correlates with CD34+ expression in primary human AML. Primary human AML samples were stained with CD34-APC, fixed, permeabilized, and stained with rabbit-anti-MSI2. Anti-rabbit-FITC was used to detect MSI2 staining. The percent of MSI2-positive cells was calculated in CD34- and CD34+ AML fractions. We saw a significant ($p<0.001$) increase in MSI2 levels in the CD34+ fraction of AMLs. (A) A typical gating strategy used to investigate MSI2 expression in the CD34 fractions of AML. (B) A plot of the MSI2 levels between CD34- and CD34+ fractions. A solid line indicates the changes in a single AML samples. A two-tailed, paired t-test was performed

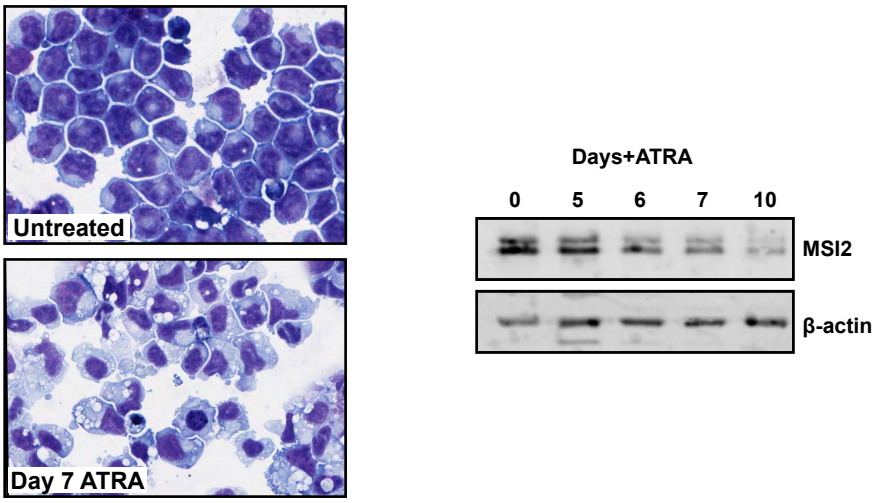
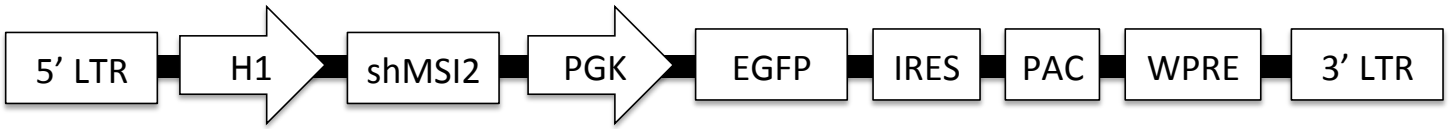


Figure 4. MSI2 levels in ATRA-treated NB4 cells. NB4 cells were treated with 1uM ATRA or 0.1% DMSO for 10 days. At 7 days post treatment, cells appeared highly differentiated vs control cells. The granulocytic maturation of NB4 cells was accompanied by a significant downregulation of MSI2 protein levels.

A.



B.

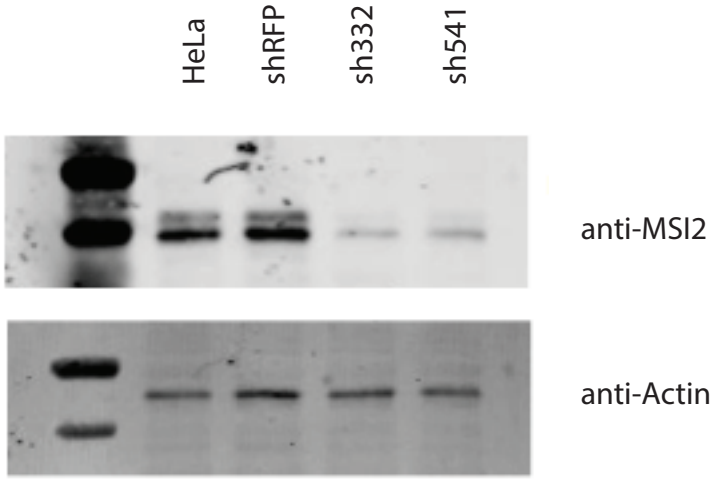


Figure 5: Structure of the shMSI2 H1GIP vector and knockdown validation. (A) The MSI2 shRNA (or shRFP) was driven off of the H1promoter. The PGK promoter drove the expression of an EGFP construct. An IRES seqeunce on the EGFP transcript allowed for the expression of a puromycin cassette. (B) Two hairpins- sh541 and sh332 were able to efficiently knockdown MSI2 levels in HeLa cells.

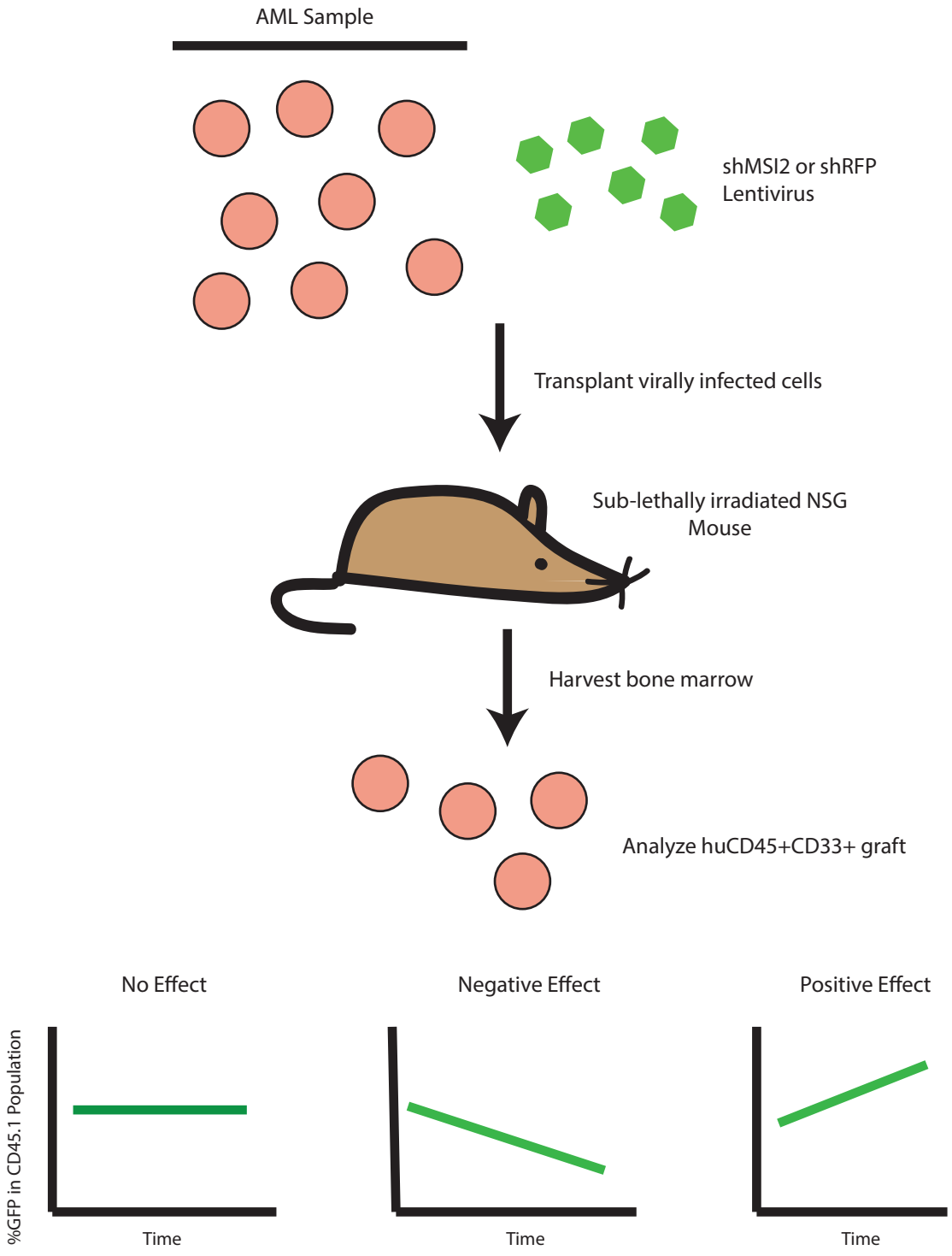


Figure 6. AML Transplantation Assay. AML cells were flow sorted to enrich for LSCs and infected with shMSI2 or shRFP lentiviral particles for 24 hours. Cells were washed extensively in PBS and depending on the AML, 50, 000- 200, 000 cells were intrafemorally injected per sub-lethally irradiated mouse. A subset of cells were kept in culture and initial GFP transduction levels were assayed 3-days post transduction. 3-months post transplant, bone marrow was harvested and GFP expression in the CD45+ CD33+ human myeloid graft was analyzed for GFP expression

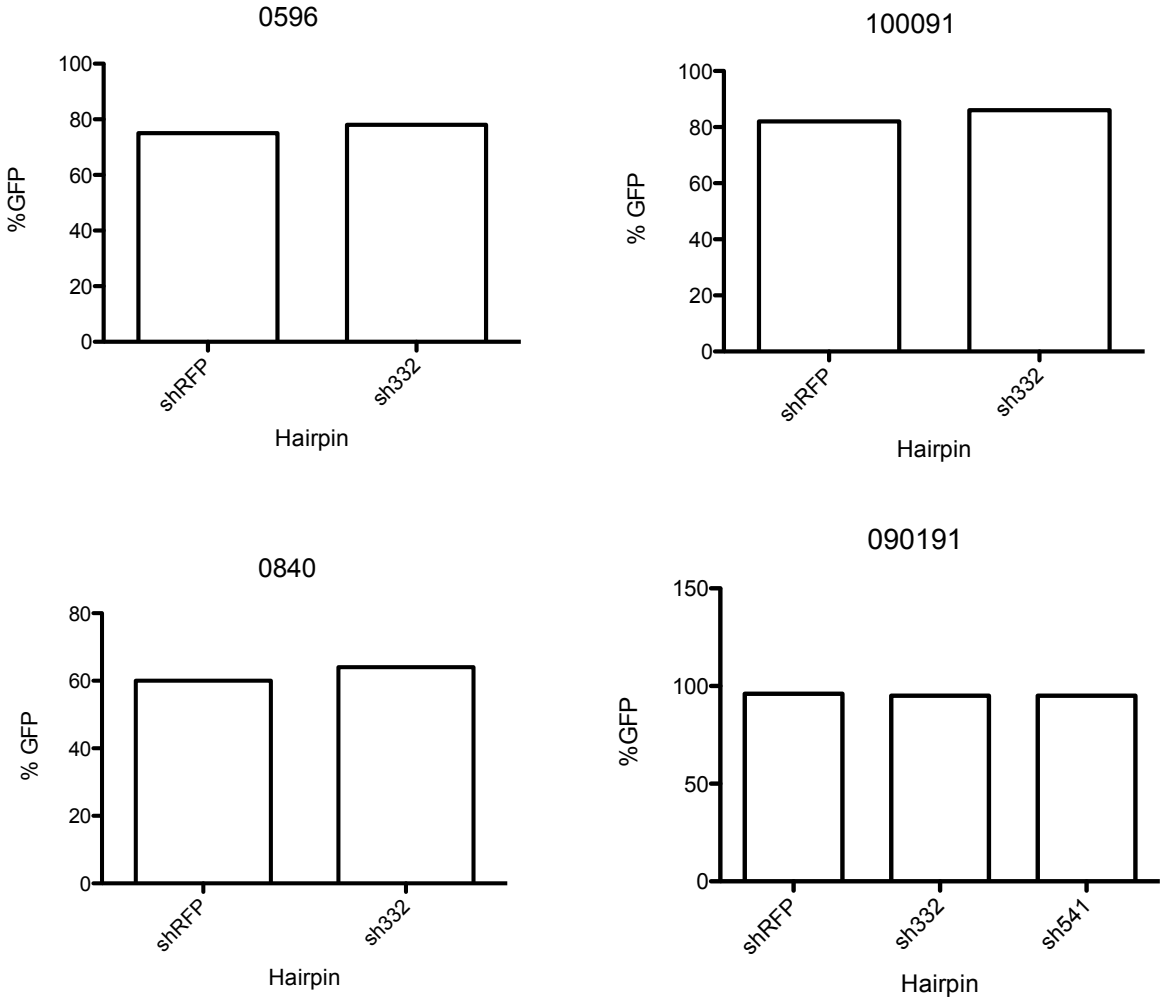
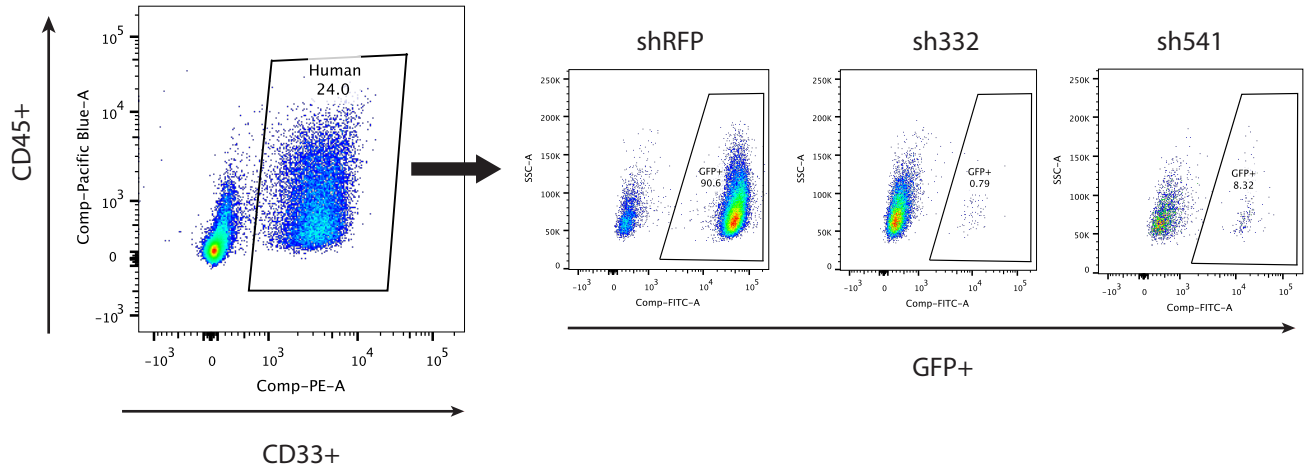


Figure 7. Initial AML transduction rates. Four AML samples were infected with either shRFP or shMSI2. Cells were infected in StemSpan media supplemented with 20ng/ mL Il-6, 100ng/ mL SCF, 100ng/ mL FLT-3L, and 20ng/ mL TPO. The virus was removed from the cells by washing with PBS after 24 hours. GFP levels were analyzed 3-days post transduction. All AML samples were transduced between 60 and 95%.

A.



B.

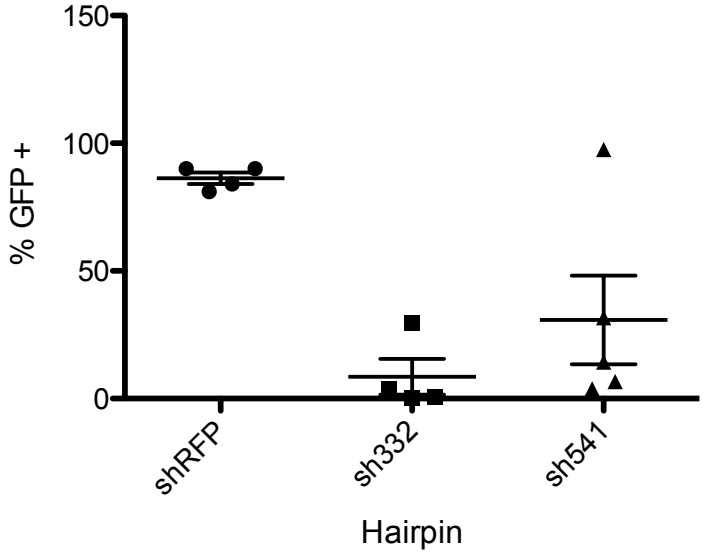
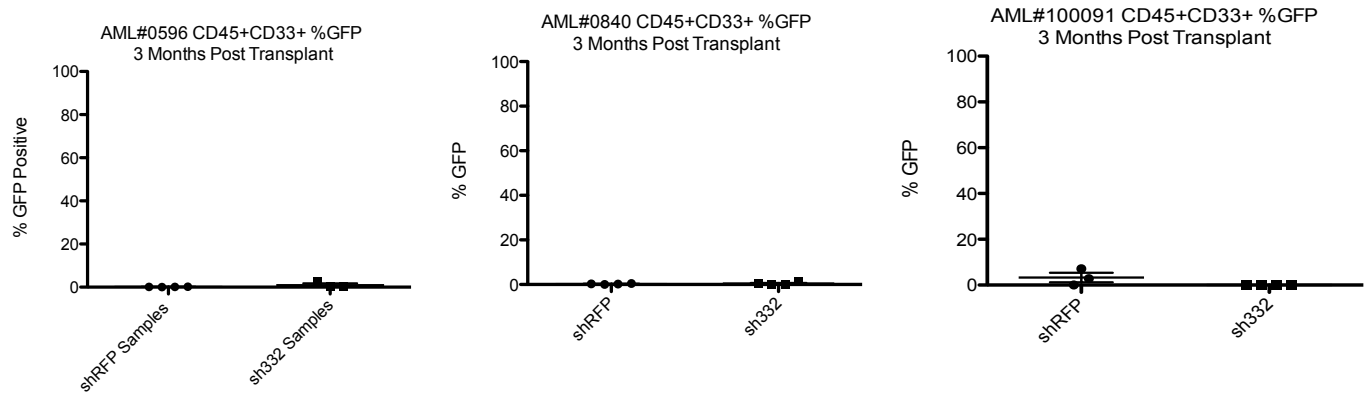


Figure 8. Analysis of AML #090191 transplant. AML #090191 was infected with either sh332 (a hairpin targeting MSI2), sh541(a second hairpin targeting MSI2) or shRFP lentiviral particles. AML samples infected with one of the hairpins were transplanted across 5 sub-lethally irradiated NSG mice each (for a total of 15 transplanted mice). 3 months post transplant, bone marrow was harvested and the human graft was analyzed for GFP expression. Notably, this leukemia sample had a large proportion of cells that were CD45- so we gated solely on CD33+ expression in order to analyze the human graft (A). Impressively, when GFP levels were analyzed in the shRFP transplanted mice, we saw no change in GFP levels compared to input levels. When GFP levels were examined in the sh332 or sh541 transplanted mice, we saw a significant decrease in GFP levels in the human graft compared to input (B).

A.



B.

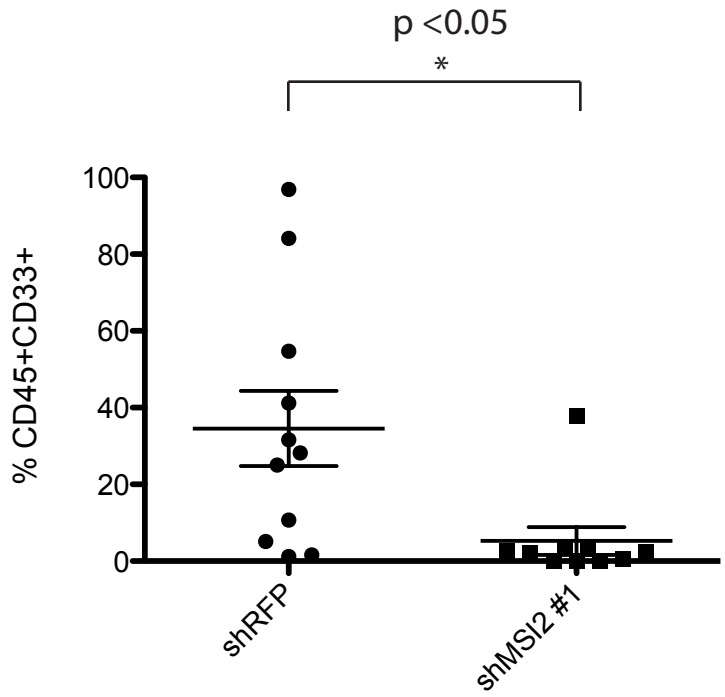


Figure 9. Analysis of AML samples 3-months post transplant. Three AML samples (#0596, #0840, and #100091) were treated with shMSI2 or shRFP lentiviral particles and transplanted in sub-lethally irradiated NSG mice (5 per hairpin). Mice were sacrificed 3-months post transplant and GFP levels were analyzed within the CD45+CD33+ human grafts. (A) No GFP expression was detected in any of the transduced AML samples. (B) Despite this, significant differences were seen in the engraftment levels between shRFP- and shMSI2-infected samples. Samples infected with an shRFP hairpin showed significantly higher engraftment levels compared to those samples infected with an shMSI2 hairpin. An unpaired, two-tailed t-test was performed

A)

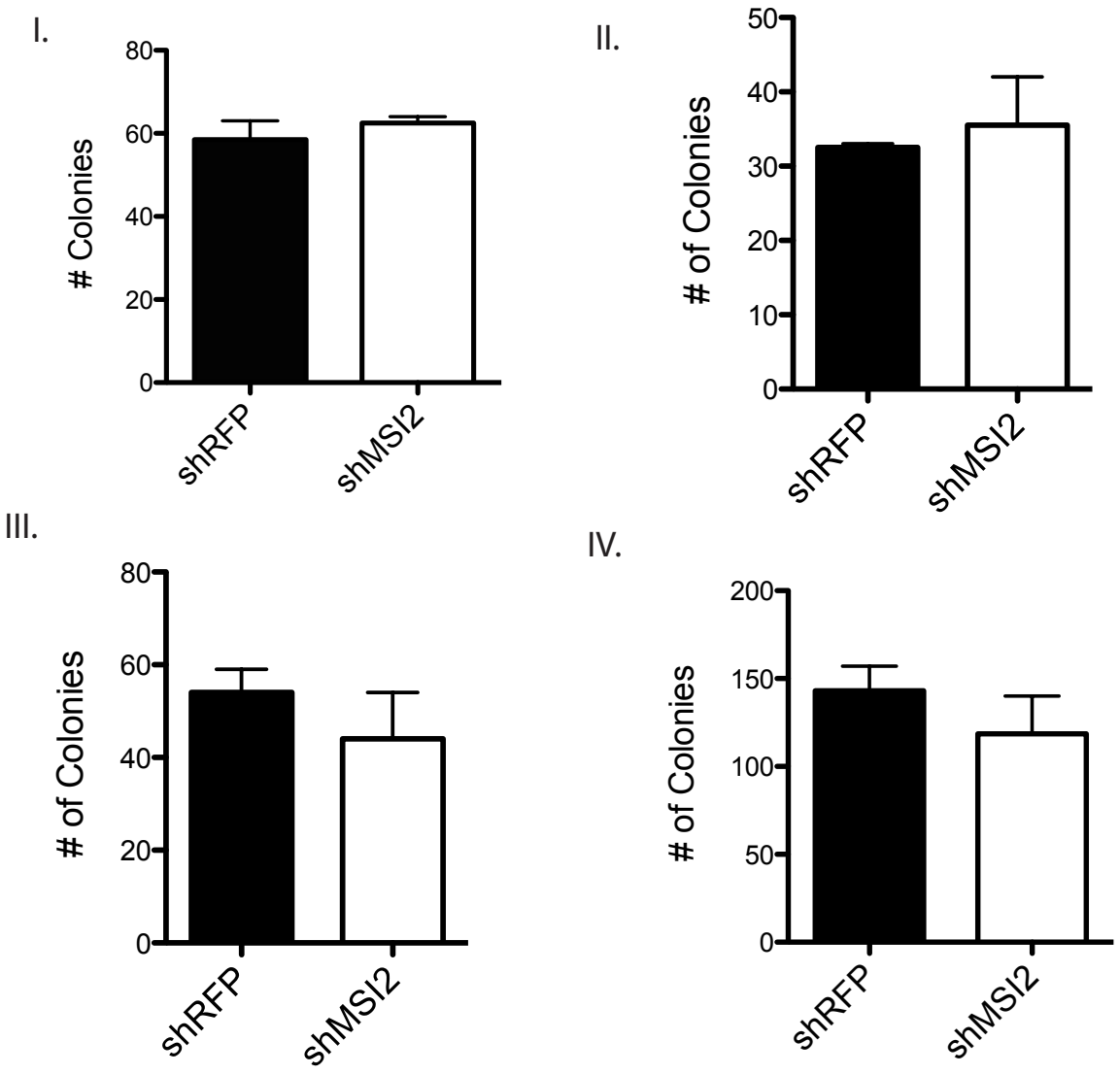


Figure 10. AML CFU Counts. 4 AML samples that were capable of producing colonies in methycellulose were infected with hairpins targeting MSI2 or an RFP control. Cells were sorted and 10 000 GFP+ cells were plated in duplicate wells of a 24-well plate in human colony gel. No significant differences in colony count were observed 14-days post plating (i) AML #090191 (ii) AML #100753 (iii) AML #110751 (iv) AML #0596.

Chapter 3:
Investigating the RNA-binding activity of MSI2

Abstract

The Musashi-2 (MSI2) RNA-binding protein (RBP) is a critical regulator of murine and human hematopoietic stem cells (HSCs). Despite this, its mechanism of action remains poorly understood. Our knowledge of how MSI2 functions is largely inferred from Musashi-1 (MSI1) studies. The function and RNA-binding properties of the MSI1 protein have been extensively characterized. MSI1 binds to a specific consensus motif in the 3' untranslated region (UTR) of target mRNAs and functions predominantly as a translational repressor. However, reports indicate that MSI2 can act to either increase or decrease the expression of target mRNAs. Therefore, it is important to understand the RNA binding properties of MSI2 and mechanism of action. Furthermore, although many studies have suggested that MSI2 and MSI1 act redundantly in certain tissues, others have indicated that the two proteins can play unique roles, even when expressed in the same cell population. Additionally, given the potent role MSI2 plays in the control of HSC self–renewal and differentiation, the identification of MSI2 target RNAs within the hematopoietic context is likely to reveal critical regulators of HSCs. Using cross-linking immunoprecipitation followed by sequencing (CLIP-Seq), we were able to exhaustively characterize the RNA-binding properties of MSI2. We discovered that MSI2 bound to the (G/U)UAGU motif that was localized predominantly to the 3'UTR of mRNAs. A ranked list of MSI2 RNA targets was generated and these targets were further validated by RNA immunoprecipitation followed by quantitative PCR (RIP-qPCR). Interestingly,

follow-up studies indicated that MSI2 could function to increase or decrease the expression of target mRNAs. Our data identify numerous MSI2 RNA targets that may represent potent regulators of hematopoietic stem cell self-renewal. They further suggest that MSI2 may have a complex mechanism of action that is dependent on the mRNA target to which it is bound.

Introduction

The hematopoietic system relies on intricate gene regulatory networks in order to balance self-renewal and differentiation. Disruptions in these regulatory networks can result in diseases such as leukemia and myelodysplasia. By better understanding the processes that control hematopoietic stem and progenitor cell (HSPC) self-renewal and differentiation, we can better understand the pathogenesis of such diseases in order to better develop curative therapies. At the same time, therapies such as hematopoietic stem cell transplantation (HSCT) rely on HSPCs in order to overcome patient morbidities[1]. HSCT is commonly used to repopulate the bone marrow of leukemia patients and to help target residual leukemic cells through a graft vs. leukemia effect[2]. A better understanding of the pathways active in HSPCs can lead to more effective HSCT regimes. Despite this, our understanding of the molecular pathways that underlie stem cell fate decisions remains limited. Complex transcriptional programs have been identified within populations of HSPCs, but the post-transcriptional regulation of gene expression further adds to the complexity of this system[3-5]. Notably, very few studies have addressed post-transcriptional mechanisms that control protein expression in HSPCs.

The complex process of gene expression begins with transcription in the nucleus. The transcribed mRNA then undergoes post-transcriptional events by interacting with numerous regulators such as RNA binding proteins (RBP) and non-coding RNAs before being translated into protein. Importantly, many studies report a weak correlation between mRNA transcript and protein levels, likely due to complex post-transcriptional regulatory networks present within cells[6]. RBPs play critical roles in the post-transcriptional control of gene expression by controlling events such as RNA splicing, transport, stability, and translation[7]. Furthermore, the post-transcriptional events that occur in eukaryotic cells are thought to be highly complex and coordinated. The "RNA Regulon" theory proposes that numerous RNAs may be regulated by multiple RBPs to co-regulate the expression of sets of related proteins[8,9]. This allows for the creation of plastic gene expression networks that can be assembled and disassembled rapidly.

Numerous RBPs have been shown to play critical roles in self-renewal, differentiation, and reprogramming[10,11], and it is likely that these proteins function to regulate patterns of gene expression at the post-transcriptional level. The MSI1 and MSI2 proteins are two well-known regulators of stem cell self-renewal[12]. *Msi2* is uniquely expressed in the hematopoietic system where it plays a critical role in the maintenance of HSC populations in both mice and humans[13,14]. Despite this, very little is known about the function of MSI2. Our current understanding of how the MSI2 protein functions in the hematopoietic system is largely inferred from studies of its homolog, MSI1. MSI1 is thought to mediate the translational repression of mRNA targets in vertebrates[15,16]. The enrichment of MSI2 in HSPCs along with its

functional roles in stem cell maintenance suggests that MSI2 functions to regulate the expression of critical genes involved in HSC quiescence, self-renewal, and differentiation. Despite this, an exhaustive characterization of the MSI2 RNA targets in the hematopoietic system was lacking. We thus set out to identify the RNA targets to which MSI2 binds and to further characterize the RNA-binding properties of MSI2.

To characterize the RNA-binding properties of MSI2 in an unbiased manner, we employed CLIP-Seq (Figure 1, p.141)[17]. This protocol allows for the identification of small RNA sequences attached to a protein of interest through next generation sequencing (NGS). Subsequent bioinformatic analysis of these RNA sequences provides an exhaustive profile of the RNA-binding characteristics of any given RBP.

Materials and Methods

CLIP-Sequencing

CLIP-seq was performed as previously described[17]. Briefly, 25 million NB4 cells were washed in PBS and UV-cross-linked at 400mJ/cm^2 on ice. Cells were pelleted, lysed in wash buffer (PBS, 0.1% SDS, 0.5% Na-Deoxycholate, 0.5% NP-40), DNase treated, and supernatants from lysates were collected for immunoprecipitation (IP). MSI2 was immunoprecipitated overnight using 5µg of anti-MSI2 antibody (EP1305Y, Abcam) and 50uL of Protein A Dynabeads (Invitrogen). Beads containing protein-RNA complexes were washed twice with wash buffer, high-salt wash buffer (5X PBS, 0.1% SDS, 0.5% Na-Deoxycholate, 0.5% NP-40), and PNK buffer (50 mM Tris-Cl pH 7.4, 10 mM $MgCl_2$, 0.5% NP-40). Samples were then treated with 0.2 U

micrococcal nuclease (MNase) for 5 minutes at 37 degrees with shaking to trim RNA. MNase inactivation was then carried out with PNK + EGTA buffer (50 mM Tris-Cl pH 7.4, 20 mM EGTA, 0.5% NP-40). RNA was dephosphorylated using alkaline phosphatase (CIP, NEB) at 37 degrees for 10 minutes followed by washing with PNK+EGTA, PNK buffer, and then 0.1 mg/mL BSA in nuclease free water. 3' RNA linker ligation was performed at 16 degrees overnight with the following adapter: 5' P-UGGAAUUCUCGGGUGCCAAGG-puromycin. Samples were then washed with PNK buffer, radiolabelled using P32-y-ATP (Perkin Elmer), run on a 4-12% Bis-Tris gel and transferred to a nitrocellulose membrane. The nitrocellulose membrane was developed *via* autoradiography and RNA-protein complexes 15-20 kDa above the molecular weight of Msi2 was extracted with Proteinase K followed by RNA extraction with acid phenol-chloroform. A 5' RNA linker 5' HO-GUUCAGAGUUCUACAGUCCGACGAUC-OH was ligated to the extracted RNA using T4 RNA ligase (Fermentas) for two hours at 37 degrees Celsius and the RNA was again purified using acid phenol-chloroform. Adapter ligated RNA was re-suspended in nuclease free water and reverse transcribed using Superscript III reverse transcriptase (Invitrogen). 15 cycles of PCR were performed using NEB Phusion Polymerase using a 3' PCR primer that contained a unique Illumina barcode sequence. PCR products were run on an 8% TBE gel. Products ranging between 150-200bp were extracted using the QIAquick gel extraction kit (Qiagen) and re-suspended in nuclease free water. Two separate libraries were prepared and sent for single-end 50bp Illumina sequencing at the Institute for Genomic Medicine at the University of California, San Diego. 47,098,127 reads from the first library passed

quality filtering of which 73.83% mapped uniquely to the human genome. 57,970,220 reads from the second library passed quality filtering of which 69.53% mapped uniquely to the human genome.

CLIP-seq mapping and cluster identification

Before sequence alignment of CLIP-seq reads to the human genome was performed, sequencing reads from libraries were trimmed of polyA tails, adapters, and low quality ends using Cutadapt[18] with parameters --match-read-wildcards --times 2 -e 0 -O 5 --quality-cutoff' 6 -m 18 -b TCGTATGCCGTCTTCTGCTTG -b ATCTCGTATGCCGTCTTCTGCTTG -b CGACAGGTTCAGAGTTCTACAGTCCGACGATC -b TGGAATTCTCGGGTGCCAAGG -b AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA -b TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT. Reads were then mapped against a database of repetitive elements derived from RepBase (version 18.05)[19]. Bowtie (version 1.0.0) with parameters -S -q -p 16 -e 100 -l 20 was used to align reads against an index generated from Repbase sequences[20]. Reads not mapped to Repbase sequences were aligned to the hg19 human genome (UCSC assembly) using STAR (version 2.3.0e) with parameters --outSAMunmapped Within –outFilterMultimapNmax 1 –outFilterMultimapScoreRange 1[21]. To identify clusters of significantly enriched CLIP-seq reads, PCR replicates were first removed using a custom script of the same method as *Darnell R., Cold Spring Harbor Protocols, 2012*[22]. Otherwise reads were kept at each nucleotide position when more than one read's 5' end was mapped. Clusters were then assigned using the CLIPper software

with parameters --bonferroni --superlocal --threshold-23. The ranked list of significant targets was calculated assuming a Poisson distribution, where the observed value is the number of reads in the cluster, and the background is the number of reads across the entire transcript and/ or across a window of 1000 bp +/- the predicted cluster (Figure 2, p.142).

RNA-Immunoprecipitation (RIP) followed by qPCR

10 million NB4 cells were pelleted, washed in ice-cold PBS and lysed in 300µL Polysome Lysis Buffer (10mM HEPES pH 7, 100mM potassium chloride, 5mM Magnesium chloride, 25mM EDTA, 0.5% NP-40, 2mM DTT, 50U/mL RNAse OUT). Lysate was spun at 14000g for 10 minutes at 4 degrees and the supernatant was removed. Protein G Dynabeads were washed with Citrate-Phosphate buffer (pH 5) and incubated for 2 hours at room temperature with 5µg of anti-MSI2 or 5µg of anti-IgG. Antibody-bead complexes were washed 3X in polysome lysis buffer, re-suspended in 750µL of polysome lysis buffer, and combined with 250µL cell lysate supernatant (1mL total). 100µL of this solution was withdrawn and set aside for a 10% input control. The lysate was incubated with the antibody bead complexes overnight at 4 degrees and washed 4X with ice-cold NT2 buffer (50mM TRIS-HCl pH 7.5, 150mM sodium chloride, 1mM magnesium chloride, 0.05% NP-40). Antibody bead complexes were suspended in Trizol solution and RNA was isolated.

Reverse transcription was performed using the qScript cDNA Synthesis Kit (Quanta) and qPCR was performed using PerfecTa qPCR Supermix (Quanta) and Taqman Universal probe library/ Primers: STMN1 F-gaagctaataaagagaaccgagagg R-

ttccgcacttcttcaatgtg Probe #46, GAPDH F- agccacatcgctcagacac R- gcccaatacgaccaaatcc Probe #60, Rac1 F- ctgatgcaggccatcaagt R- caggaaatgcattggttgtg probe #77, HSP90AB1 F- aaccgcatctatcgcatga R- catcaggaactgcagcattg probe #6, KIAA0101 F- cccagaaaggtgcttggtt R- gggttccctcctgcatattt probe #53, TSPAN3 F- ggacttgccacgtttgtcat R- atgctgcgatcaacctcatt probe #15, HDGF F- ggagagcaggggacttgc R- cctccttctcctctccttcag probe #16, TBX1 TVC F- gtgccggtggacgataag R- cgagtccgggtggtagtg probe #17, Cux1 F- gactctgccaggtggatgtt R- gttcgccaataccgttgc probe#36, MYB F- agcaaggtgcatgatcgtc R- gatcacaccatgatgaagaatcag probe #37, WDR77 F- gcatcaaggtttgggacct R- gaggcagcaacacaagtgac probe#62, LRPPRC F- gaagatgccttgaacttgaaaga R- gcctacatacttgccggtgt probe #56, HSP90AA1 F- gggcaacacctctacaagga R- cttgggtctgggtttcctc probe #46, HMGB2 F- tgaacagaaagcagctaagctaaa R- cttcttctttgagcctgttgg probe #81, U2AF1 F- aacatttaccgtaaccctcaaaac R- gcatctccacatcgctcac probe #48, CFL1 F- gtgccctctccttttcgttt R- ttgaacaccttgatgacaccat probe #5, Slain1 F- gaagcaaaattgcacaacctg R- ccctcagagtgctgagtgaa probe #75. Expression of target mRNA in the pulldown samples was normalized to expression of the same mRNA in the input samples.

Knockdown/ Overexpression and Western Blotting

MSI2 shRNAs were designed with the Dharmacon algorithm (http://www.dharmacon.com). Predicted sequences were synthesized as complimentary oligonucleotides, annealed and cloned downstream of the H1 promoter of the modfied cppt-PGK-EGFP-IRES-PAC-WPRE lentiviral expression

vector[24]. Sequences for the MSI2 targeting and control RFP targeting shRNAs were as follows: shMSI2, 5'-GAGAGATCCCACTACGAAA-3'; shRFP, 5'-GTGGGAGCGCGTGATGAAC-3'. Human MSI2 cDNA (BC001526; Open Biosystems) was subcloned into the MA-1 bi-directional lentiviral expression vector[25]. All lentivirus was prepared by transient transfection of 293FT cells with pMD2.G and psPAX2 packaging plasmids (Addgene) to create VSV-G pseudotyped lentiviral particles. All viral preparations were titrated on HeLa cells before use. Standard SDS-PAGE procedures were performed to validate the effect of MSI2 knockdown and overexpression in transduced and GFP+ sorted NB4 cells. Immunoblotting was performed with anti-MSI2 rabbit monoclonal IgG (EP1305Y, Epitomics), anti STMN1 (TA323949, Origene) and β-actin mouse monoclonal IgG (ACTBD11B7, Santa Cruz Biotechnology) antibodies. Secondary antibodies used were IRDye 680 goat anti-rabbit IgG and IRDye 800 goat anti-mouse IgG (LI-COR).

Results and Discussion

Standardization of the CLIP-Seq protocol

To begin our CLIP-Seq experiments, we first needed to standardize numerous conditions such as crosslinking time and intensity, MSI2 pull downs, and input cell numbers. In order to perform CLIP-Seq, a high quality antibody is required that is highly specific for the protein of interest. The antibody must be able to bind to the denatured protein and should be able to do so under high stringency conditions (1% NP-40, 0.1% SDS, and 0.5% deoxycholate). We identified a MSI2 antibody that was highly specific and bound to the MSI2 protein with high affinity

(Figure 3, p.143). The high affinity of this antibody was crucial since we were aiming to construct CLIP-Seq libraries from sorted populations of primary cells, which would greatly limit the number of input cells we could use. After identifying a high quality CLIP-grade antibody, we next sought to identify whether MSI2 could be efficiently cross-linked to RNAs. We sought to employ UV-B radiation in order to crosslink MSI2 with its RNA targets. Unlike formaldehyde and other crosslinking chemicals, UV-B radiation is specific for protein-nucleic acid complexes; it does not induce protein-protein crosslinking[26]. Crosslinking is believed to occur due to UV absorption by nucleic acid sequences located within angstroms of proteins. Importantly, previous groups have suggested that UV-B crosslinking efficiencies are quite low (on the order of 1-5%)[27]. Furthermore, it was suggested that the capacity for efficient UV-crosslinking might be protein specific since crosslinking requires the presence of aromatic amino acids at RNA-binding sites. To our surprise, the MSI2 protein-RNA complexes could be readily cross-linked at numerous different intensities of UV-B. Furthermore, cross-linked MSI2-RNA complexes could be detected with varying input cell numbers (Figure 4, p.144). We decided to crosslink protein-RNA complexes with 400 millijoules/cm^2 (mj/cm^2) since it provided a maximal cross-linking efficiency and other CLIP-Seq studies had successfully constructed libraries using this intensity and it did not lead to a drastic abundance of UV-induced mutations[28,29]. Importantly, one objective in determining the appropriate intensity of UV-B radiation is to limit the number of UV-induced mutations at the site of crosslinking. Previous studies have performed CLIP-Seq in the mouse brain and revealed UV-induced deletions in ~8-20% of mRNA tags after

crosslinking at 3 X 400 mJ/cm^2 [28]. This frequency of mutations still allowed for the accurate alignment and identification of CLIP reads and even allowed for the precise mapping of RNA-protein interactions due to the analysis of clustered mutation sites.

Once immunoprecipitation and crosslinking conditions were standardized, we next sought to standardize parameters that would allow for the successful construction of next generation sequencing (NGS) libraries. We opted to perform sequencing on the Illumina platform. One contributing factor was the availability of primer and adapter sequences that allowed us to design custom oligonucleotides for our CLIP-Seq libraries. To obtain CLIP tags of appropriate size, cross-linked RNA was partially digested. We employed an MNase digestion protocol and standardized the concentration of MNase and the digestion time in order to obtain RNA tags of optimal length. After proper MNase digestion, we performed radioactive western blotting, which allowed for the detection of MSI2-RNA complexes. Importantly, cross-linked RNA will slow the movement of protein-RNA complexes through an SDS-PAGE gel in a manner that is dependent on the size of the cross-linked mRNA. Previous studies identified that protein-RNA complexes that contain RNA of 50-70 nucleotides will run 15-20kDa above the non-cross-linked protein[30]. Therefore, the optimal region for the isolation of CLIP tags is typically 15-20kDa above the molecular weight of the protein of interest. As an additional control, cross-linked MSI2-RNA complexes were digested with a very high amount of MNase. This control is important to demonstrate the specificity of the radioactive signal (Figure 5, p.145). RNA-protein complexes that are digested with high concentrations of MNase should run very close to the molecular weight of the protein of interest since most of

the cross-linked RNA is destroyed except for the small regions that are directly protected by the protein. Incubation with dilute concentrations of MNase should result in a smearing pattern extending upwards from the molecular weight of the protein of interest due to the generation of variably sized protein-RNA complexes. It is important to note that cross-linked Msi2-RNA complexes were run on Novex NuPage gels since these gels are buffered at a pH very close to 7. The pH of traditional Laemmli SDS-PAGE gels can approach 9.5 during a run, which can result in the alkaline hydrolysis of RNA[30]. Furthermore, proteins were transferred onto a nitrocellulose membrane due to their inability to retain free RNA molecules. MSI2-RNA complexes were ultimately digested with 0.2 U MNase for 5 minutes at 37 degrees with shaking, and this generated MSI2-RNA complexes of variable length. An IgG control pull-down demonstrated that the signal detected upon radioactive western blotting was specific for the MSI2 protein (Figure 5, p.145). A high MNase control further revealed that the signal was specific for the RNA attached to the MSI2 protein. 3' and 5' RNA adapters were then ligated to the cDNA of interest. These RNA adapters contained nucleic acid sequences that were compatible with illumina NGS. Unique steps were taken to prevent the ligation of adapters to one another (Figure 6, p.146).

Numerous steps in the CLIP-Seq protocol prevent the formation of unwanted ligation products. To prevent the circularization of CLIP-derived RNA tags, the MNase trimmed RNA molecules are treated with a phosphatase on-bead. MNase trims RNA yielding 5'OH and 3'phosphate groups that could ligate to one another causing the circularization of the RNA molecule, preventing downstream analysis.

The phosphatase treatment prevents the circularization of MNase-digested RNA. Furthermore, to prevent the ligation of 5' and 3' adapters to each other, the 3'RNA adapter (P-UGGAAUUCUCGGGUGCCAAGG-puromycin) is first ligated to MNase-digested and phosphatase-treated RNA on-bead. The 3'RNA adapter is designed to contain a phosphate group on its 5' end allowing for ligation to the phosphatase treatment RNA. Several purification steps occur after the ligation of the 3' adapter and prior to the 5' adapter ligation. The 5' adapter (HO- GUUCAGAGUUCUACAGUCCGACGAUC-OH) is ligated after the extraction of protein-RNA complexes from a nitrocellulose membrane. The numerous washing steps, western blotting, and nitrocellulose transfer all ensure the efficient removal of free 3'adapter RNA molecules. Furthermore, the 3' RNA adapter was uniquely generated to contain a 3'puromycin moiety to ensure that the 3' adapter would only ligate to RNA tags at its 5'end. After the 5' adapter is ligated to the CLIP RNA, reverse transcription was performed using a RT primer specific for the 3' adapter sequence (5'GCCTTGGCACCCGAGAATTCCA). PCR was performed using NEB Phusion polymerase. A standard forward PCR primer (5' AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA) was designed that bound to the 5' adapter sequence and introduced additional sequences to the 5'end that allowed for the binding of cDNA to the Illumina flow cell, and generated a common site for a sequencing primers to adhere and sequence into the CLIP-derived RNA. To allow for multiplexing of CLIP-Seq libraries, 3' primer sequences were designed that contained unique index sequences. The following primer was used to bind to the common 3' adapter and sequence the index of a

barcoded RNA: 5' GATCGGAAGAGCACACGTCTGAACTCCAGTCAC. Independently derived CLIP-Seq libraries were PCR amplified with unique 3' reverse primers allowing each to be uniquely barcoded.

PCR-amplified CLIP-Seq libraries were run on an 8% Tris/Borate/EDTA (TBE) gel and products between 150 and 200 base pairs (bp) were isolated, extracted using the QIAquick gel extraction kit, and sent for single-end 50bp sequencing. The combination of PCR primers and adapter sequences added an extra 113bp onto the CLIP-derived RNA. Consequently, PCR products in the 150-200bp region should represent CLIP tags between ~ 40 and 85 nucleotides. Notably, when RNA-input was low, a potent primer-dimer approximately 120 nucleotides would appear; it was critical to exclude this product prior to NGS.

Overview of CLIP Bioinformatics

The bioinformatic workup of CLIP-Seq libraries can be simplified as follows: (1) Filter reads to remove introduced sequences, low quality reads, and repetitive elements (2) Align reads to the human genome and remove PCR duplicates (3) Identify significant 'clusters' of reads and (4) Analyze the properties of these 'clusters'[31]. The libraries produced after Illumina sequencing were filtered using Cutadapt software. This software was written in Python and is freely available from the Python Package Index (PyPI). The FASTQ reads obtained from Illumina sequencing are passed to the cudadapt software that removes adapter sequences and low quality reads. Filtered reads are then aligned to Repbase, a database of repetitive DNA elements[19]. Repetitive DNA sequence motifs in eukaryotic genomes

can repeat hundreds or thousands of times. These repetitive elements make up a large proportion of nuclear DNA yet their function remains largely unknown[32]. The presence of repetitive elements in DNA and RNA libraries is difficult to interpret and complicates analysis due to the inability to map such reads uniquely to the genome. Consequently, bioinformatic analysis of DNA and RNA libraries typically exclude any repetitive elements from analysis and instead focus on the analysis of unique genomic regions[32]. In order to remove repetitive elements, the Bowtie sequence aligner was used to align filtered CLIP-Seq reads against the Repbase human repository[20]. Reads that did not align to this database were used for downstream analysis.

Bowtie, an open-source software program written in C++, is an ultrafast and memory efficient alignment software that is specifically designed to align short reads to large genomes. It is an open-source software program written in C++. The Bowtie software uses a unique algorithm known as a Bhurrows-Wheeler Transform (BWT) in order to compress a large reference genome into a small 'index' that can be queried much more efficiently and requires less memory to operate when compared to more complex alignment software[20]. Bowtie is 35-200 times faster than traditional alignment software and can be performed on a simple PC workstation with less than 2 GB of memory. Bowtie's speed of alignment comes with trade-offs in the quality of alignment and thus is used for the quick alignment and removal of repetitive elements. Importantly, Bowtie does not report gapped alignments and therefore will fail to report reads that sit on intron-exon junctions.

Reads that do not map to the Repbase database are aligned to the human genome using the Spliced Transcript Alignment to a Reference (STAR) aligner[21]. The accurate alignment of NGS reads must overcome two major hurdles: (1) The alignment of reads that contain mismatches and (2) mapping sequences derived from non-contiguous regions of the genome. Mismatches can occur due to genomic variations, sequencing errors, or through mutations introduced through experimental protocols (i.e. UV-induced mutations introduced in CLIP-Seq reads). The STAR aligner uses novel strategies for spliced alignment at a high mapping speed. This comes at the trade-off of a large memory footprint that is required for STAR alignment- roughly 27GB of RAM. Resultantly; therefore this software is typically run off of a server platform[21].

Once reads are aligned to the human genome, PCR duplicates must be collapsed in order to remove any bias that may have been introduced during the PCR amplification of adapter-ligated CLIP RNA. Due to low yields of adapter-ligated RNA and the need to add additional sequences onto the CLIP tags for sequencing purposes, PCR amplification is required. PCR amplification biases occur due to differences in DNA sequence content and length that affect the kinetics of DNA annealing and denaturation[33]. Differences in sequence content and length result in different amplification efficiencies. This is reflected upon the analysis of PCR-amplified libraries, which often reveal a large proportion of identical reads[34]. Most bioinformatic pipelines collapse identical reads in order to prevent PCR bias from affecting downstream analysis[33]. Importantly, libraries generated through techniques such as CLIP-Seq often have a high proportion of repetitive reads and it

is often difficult to discern whether these are a result of PCR bias or result from unique RNA molecules[34]. As opposed to techniques such as exome sequencing, CLIP-Seq preferentially isolates very specific populations of RNA. Furthermore, the processing of these RNAs in order to isolate protein-RNA interactions sites can further enhance the proportion of identical reads. Importantly, these identical reads may originate from unique mRNA molecules and not from PCR duplication events. For example, MNase displays a notable bias in sequence specificity with a majority of cut sites being centered on A/T-containing dinucleotides[35]. This cutting bias in combination with a size-selection bias can result in the detection of numerous identical reads that do not originate from PCR duplication events. Thus, collapsing identical reads during CLIP analysis has the benefit of removing the effects of PCR bias but can result in a drastic reduction in the number of usable reads partially due to the removal of biologically significant reads. In order to properly discern between PCR duplicates and identical reads originating from unique protein-RNA interactions, many studies have included randomer sequences in their 3'adapters[36]. Randomers are typically a sequence of 5-10 random nucleotides that are incorporated into the 3' adapter. When a 3' adapter includes a randomer sequence, it is possible to more accurately identify PCR duplicates. Reads with identical sequences but unique randomers originate from unique RNA molecules while those reads with identical sequences and identical barcodes are likely PCR duplicates.

Clusters (also known as 'peaks') were identified using the Clipper software[17]. This custom python script was developed by our collaborators in the Yeo lab and is a significant piece of software. Unlike other peak calling programs, Clipper was

specifically designed for CLIP-Seq. Traditionally, peak finding programs had been designed for chromatin immunoprecipitaion (ChIP)-Seq applications and were concerned with the identification of DNA and not RNA clusters. ChIP-Seq peak finding algorithms use genome wide cutoffs to define significance since all genes are found at equal levels[23]. Due to differences in gene expression, a CLIP-Seq peak finding algorithm cannot use a genome wide cut-off and instead must use other means to determine what peaks are significant[23]. In order to take into account the level of gene expression, Clipper makes use of an algorithm that calculates a false discovery rate (FDR) threshold on a gene-by-gene basis. Peaks are defined as those areas that contain read heights greater than the FDR threshold. Additional algorithms are then used to determine the shape of the peak. The peak boundaries are those areas that fall below the FDR cutoff or are local minima above the FDR[23]. These algorithms allows for the identification of multiple peaks within a close proximity. The ranked list of significant targets was calculated assuming a Poisson distribution, where the observed value is the number of reads in the cluster, and the background is the number of reads across the entire transcript. Motif analysis was performed using the HOMER algorithm[37].

CLIP- Seq reveals the RNA binding properties of MSI2

Two independent MSI2 CLIP-Seq libraries were prepared using the human promyeloctic leukemia cell line, NB4. Libraries were sequenced at the Institute for Genomic Medicine at the University of California, San Diego. 47,098,127 reads from the first library passed quality filtering of which 73.83% mapped uniquely to the

human genome. 57,970,220 reads from the second library passed quality filtering of which 69.53% mapped uniquely to the human genome. Clusters bound 5,552 common genes and 9,246 clusters overlapped significantly between the two CLIP-Seq libraries (Figure 7, p.147). The top 40% of clusters were analyzed and were predominantly found in the 3' UTR of mRNA and were associated significantly with the consensus sequence: (G/U)UAGU (Figure 8, p.148). The MSI2 consensus motif we identified shows striking similarity to the MSI1 consensus motif, (G/A)U1-3AGU that was previously identified using an *in vitro* selection assay[38]. The significant overlap in consensus motifs is not surprising given that the RNA binding domain of MSI1 and MSI2 shares an 87% sequence identity. A recent study examining the binding specificities of the Musashi family members lends further support to these consensus motifs. This study identified that the UAG motif is critical for RNA binding and that it makes the largest contribution to RNA-binding affinity[39]. Furthermore, nuclear magnetic resonance (NMR) spectroscopy experiments involving the Msi1 protein show that Msi1 binds to the GUAGU oligomer with high affinity[40]. Overall, numerous pieces of evidence suggest that the Msi1 and Msi2 consensus motifs require a conserved UAG motif in order to bind RNA targets. Furthermore, given the similarities between Msi1 and Msi2 consensus motifs, there is a possibility of significant overlap in the RNA-binding targets of these two proteins. However, there is also the possibility that minor differences in the consensus motifs may diversify the RNA-binding properties of MSI1 and MSI2, allowing for the recognition of unique RNAs as well.

To further confirm that the mRNAs identified by CLIP-Seq are in fact MSI2

targets, RNA immunoprecipitation followed by qPCR was performed. Of 16 highly significant mRNA targets identified by CLIP-Seq, 6 showed a greater than 10-fold enrichment when Msi2 was immunoprecipitated vs. an IgG control (Figure 9, p.150). These were: Rac1 (24.6-fold), KIAA0101 (32.22-fold), TSPAN3 (18.18-fold), Cux1 (26-fold), STMN1 (10-fold), and HMGB2 (27-fold). Mass spectrometry analysis of the most immature fraction of the mouse hematopoietic system reveals that all of these genes are expressed at the protein level[4]. Importantly, Tetraspanin-3 (TSPAN3) was recently described as a Msi2 target in the mouse hematopoietic system and has previously been identified as a Msi1 target mRNA in HEK293 cells. In the mouse hematopoietic system, RIP-qPCR revealed a significant enrichment of TSPAN3 mRNA upon MSI2 immunoprecipitation vs. control[41]. Furthermore, overexpression of Msi2 in murine hematopoietic cells increased levels of Tspan3 while inhibition of Msi2 resulted in a 6-fold reduction in Tspan3 levels. Additional experiments indicated that Tspan3 is most highly expressed in HSPCs and that it may play a critical role in the propagation of *de novo* AML.

STMN1 is a MSI2 target that is post-transcriptionally enhanced by MSI2 overexpression

Stathmin-1 (STMN1), also known as leukemia-associated phosphoprotein p18, is a microtubule destabilizing protein that sequesters alpha/beta-tubulin heterodimers preventing microtubule formation[42]. Phosphorylation of Stmn1 at conserved serine residues reduces the affinity for alpha/beta-tubulin heterodimers resulting in a potent inactivation of the Stmn1 protein[43]. Stmn1 is overexpressed in

many cancers and is thought to be a marker of poor prognosis[44]. Interestingly, studies using the erythroid leukemia cell line K562 report a down regulation of STMN1 levels upon chemically induced megakaryocytic differentiation[45]. Additionally, treatment of the promyelocytic leukemia cell line HL-60 with the chemical phorbol myristate acetate (PMA), which results in monocytic differentiation of these cells, is associated with a significant increase in STMN1 phosphorylation levels[46]. HL-60 cells can be differentiated to monocytes using vitamin D3, and differentiated to neutrophils using a combination of all-trans retinoic acid (ATRA) and granulocyte colony stimulating factor (G-CSF). Treatment of HL-60 cells with vitamin D3 and ATRA/G-CSF result in the down-regulation of Stmn1 levels[46]. Interestingly, *Stmn1* is expressed highly in both murine and human HSPC populations thus we decided to investigate the functional impact that changes in MSI2 levels have on *Stmn1*[4,47]. Interestingly, knockdown and overexpression studies indicated that MSI2 acts to enhance protein levels of STMN1. Knockdown of MSI2 in NB4 cells decreased STMN1 levels while overexpression of *Msi2* resulted in increased levels of Stmn1 expression (Figure 9, p.149). Given its close sequence identity to Msi1, we hypothesized that Msi2 would function in a similar manner as a translational repressor. This is clearly not the case for *Stmn1* mRNA; instead, MSI2 seems to be functioning in an opposite role by enhancing the protein levels of its target mRNA. As previously mentioned, MSI2 has been shown to enhance the protein levels of another one of its target mRNAs, Tspan3[41]. Still, other studies indicate that MSI2 acts to decrease the expression of target mRNAs[12]. Studies from our own lab investigating the relationship between MSI2 and another one of its

target mRNAs, Cyp1b1, have clearly demonstrated that MSI2 acts to decrease expression levels of this target[14]. Consequently, it appears that the MSI2 protein may act to increase or decrease expression levels of its bound mRNAs in a context-dependent manner.

MSI2 attenuates AHR signaling to expand human HSCs

Interestingly, amongst the top 2% of MSI2 CLIP-Seq targets was *CYP1B1* ($p<10^{-18}$) a component of the Aryl Hydrocarbon Receptor (AHR) signaling complex that is up regulated upon AHR activation[14]. Of note, inhibition of the AHR pathway with the small molecule StemRegenin1 (SR1) promotes the *ex vivo* expansion of human HSCs. When treated with SR1, human cord blood cells show a greater than 50-fold increase in CD34+ cells[14]. The *CYP1B1* mRNA had a significant MSI2 binding site in the distal end of its 3'UTR (Figure 10, p.151). Further studies by a fellow graduate student in our lab demonstrated an uncoupling of CYP1B1 protein and mRNA levels following *Msi2* overexpression in primary human Lin-CD34+ cord blood cells[14]. Seven days post-transduction with a lentivirus overexpressing *MSI2* resulted in a 1.7-fold up-regulation of *CYP1B1* mRNA despite a 2-fold down-regulation in protein levels. Additional luciferase assays showed that MSI2 binding the 3'UTR of *CYP1B1* was able to repress translation[14]. Further experiments involving the inhibition of *CYP1B1* in Lin-CD34+ human cord blood cells with (E)-2,3',4,5'-Tetramethoxystilbene (TMS) resulted in a 2-fold expansion of total CD34+ cells. This data suggests that CYP1B1, a molecule that is normally used to report on the activity of the AHR pathway is itself a regulator of HSC differentiation that can

be post-transcriptionally down-regulated by Msi2.

The mechanism through which Msi2 functions is currently unknown and thus it is difficult to postulate how Msi2 regulates its target mRNAs. Our CLIP-Seq analysis has clearly demonstrated that Msi2 binds specific mRNAs at conserved motifs and indicates that Msi2 may act to either enhance or reduce protein levels of target mRNAs. Importantly, the biological effects of the Msi2 protein may be target specific. The Msi2 consensus motif is quite simple yet it is able to bind to precise locations on specific mRNA molecules. Resultantly, other factors are likely to contribute to MSI2-RNA binding. Importantly, the binding of RNA recognition motifs (RRM) to consensus motifs is known to cause structural changes in RBPs that can result in the recruitment of other proteins that are critical for stable interactions with the mRNA. It is likely that the MSI2 protein functions in complexes with other proteins in order to mediate high affinity interactions with specific mRNAs. It is likely that the protein composition of MSI2 RBP complexes determines the specificity and affinity of RNA binding and may also determine the ultimate fate of the bound mRNAs.

References

1 Copelan, E. A. Hematopoietic stem-cell transplantation. *N Engl J Med* **354**, 1813-1826, doi:10.1056/NEJMra052638 (2006).

2 Atkins, H. L., Muraro, P. A., van Laar, J. M. & Pavletic, S. Z. Autologous hematopoietic stem cell transplantation for autoimmune disease--is it now ready for prime time? *Biol Blood Marrow Transplant* **18**, S177-183, doi:10.1016/j.bbmt.2011.11.020 (2012).

3 Klimmeck, D. *et al.* Transcriptome-wide profiling and posttranscriptional analysis of hematopoietic stem/progenitor cell differentiation toward

myeloid commitment. *Stem Cell Reports* **3**, 858-875, doi:10.1016/j.stemcr.2014.08.012 (2014).

4 Cabezas-Wallscheid, N. *et al.* Identification of regulatory networks in HSCs and their immediate progeny via integrated proteome, transcriptome, and DNA methylome analysis. *Cell Stem Cell* **15**, 507-522, doi:10.1016/j.stem.2014.07.005 (2014).

5 Kim, Y. C. *et al.* The transcriptome of human CD34+ hematopoietic stem-progenitor cells. *Proc Natl Acad Sci U S A* **106**, 8278-8283, doi:10.1073/pnas.0903390106 (2009).

6 Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet* **13**, 227-232, doi:10.1038/nrg3185 (2012).

7 Glisovic, T., Bachorik, J. L., Yong, J. & Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett* **582**, 1977-1986, doi:10.1016/j.febslet.2008.03.004 (2008).

8 Keene, J. D. RNA regulons: coordination of post-transcriptional events. *Nat Rev Genet* **8**, 533-543, doi:10.1038/nrg2111 (2007).

9 Cosker, K. E., Fenstermacher, S. J., Pazyra-Murphy, M. F., Elliott, H. L. & Segal, R. A. The RNA-binding protein SFPQ orchestrates an RNA regulon to promote axon viability. *Nat Neurosci* **19**, 690-696, doi:10.1038/nn.4280 (2016).

10 Hattori, A., Buac, K. & Ito, T. Regulation of Stem Cell Self-Renewal and Oncogenesis by RNA-Binding Proteins. *Adv Exp Med Biol* **907**, 153-188, doi:10.1007/978-3-319-29073-7_7 (2016).

11 Ye, J. & Blelloch, R. Regulation of pluripotency by RNA binding proteins. *Cell Stem Cell* **15**, 271-280, doi:10.1016/j.stem.2014.08.010 (2014).

12 Fox, R. G., Park, F. D., Koechlein, C. S., Kritzik, M. & Reya, T. Musashi signaling in stem cells and cancer. *Annu Rev Cell Dev Biol* **31**, 249-267, doi:10.1146/annurev-cellbio-100814-125446 (2015).

13 Hope, K. J. *et al.* An RNAi screen identifies Msi2 and Prox1 as having opposite roles in the regulation of hematopoietic stem cell activity. *Cell Stem Cell* **7**, 101-113, doi:10.1016/j.stem.2010.06.007 (2010).

14 Rentas, S. *et al.* Musashi-2 attenuates AHR signalling to expand human haematopoietic stem cells. *Nature* **532**, 508-511, doi:10.1038/nature17665 (2016).

15 Kawahara, H. *et al.* Neural RNA-binding protein Musashi1 inhibits translation initiation by competing with eIF4G for PABP. *J Cell Biol* **181**, 639-653, doi:10.1083/jcb.200708004 (2008).

16 MacNicol, M. C., Cragle, C. E. & MacNicol, A. M. Context-dependent regulation of Musashi-mediated mRNA translation and cell cycle regulation. *Cell Cycle* **10**, 39-44, doi:10.4161/cc.10.1.14388 (2011).

17 Yeo, G. W. *et al.* An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nat Struct Mol Biol* **16**, 130-137, doi:10.1038/nsmb.1545 (2009).

18 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10-12, doi::http://dx.doi.org/10.14806/ej.17.1.200 (2011).
19 Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**, 11, doi:10.1186/s13100-015-0041-9 (2015).
20 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25, doi:10.1186/gb-2009-10-3-r25 (2009).
21 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).
22 Darnell, R. CLIP (cross-linking and immunoprecipitation) identification of RNAs bound by a specific protein. *Cold Spring Harb Protoc* **2012**, 1146-1160, doi:10.1101/pdb.prot072132 (2012).
23 Lovci, M. T. *et al.* Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat Struct Mol Biol* **20**, 1434-1442, doi:10.1038/nsmb.2699 (2013).
24 Doulatov, S. *et al.* PLZF is a regulator of homeostatic and cytokine-induced myeloid development. *Genes Dev* **23**, 2076-2087, doi:10.1101/gad.1788109 (2009).
25 van Galen, P. *et al.* The unfolded protein response governs integrity of the haematopoietic stem-cell pool during stress. *Nature* **510**, 268-272, doi:10.1038/nature13228 (2014).
26 Chodosh, L. A. UV crosslinking of proteins to nucleic acids. *Curr Protoc Mol Biol* **Chapter 12**, Unit 12 15, doi:10.1002/0471142727.mb1205s36 (2001).
27 Darnell, R. B. HITS-CLIP: panoramic views of protein-RNA regulation in living cells. *Wiley Interdiscip Rev RNA* **1**, 266-286, doi:10.1002/wrna.31 (2010).
28 Zhang, C. & Darnell, R. B. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat Biotechnol* **29**, 607-614, doi:10.1038/nbt.1873 (2011).
29 Masuda, A. *et al.* Position-specific binding of FUS to nascent RNA regulates mRNA length. *Genes Dev* **29**, 1045-1057, doi:10.1101/gad.255737.114 (2015).
30 Ule, J., Jensen, K., Mele, A. & Darnell, R. B. CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods* **37**, 376-386, doi:10.1016/j.ymeth.2005.07.018 (2005).
31 Wang, T. *et al.* Design and bioinformatics analysis of genome-wide CLIP experiments. *Nucleic Acids Res* **43**, 5263-5274, doi:10.1093/nar/gkv439 (2015).
32 Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* **13**, 36-46, doi:10.1038/nrg3117 (2012).
33 Meyer, C. A. & Liu, X. S. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat Rev Genet* **15**, 709-721, doi:10.1038/nrg3788 (2014).

34 Van Nostrand, E. L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* **13**, 508-514, doi:10.1038/nmeth.3810 (2016).
35 Allan, J., Fraser, R. M., Owen-Hughes, T. & Keszenman-Pereyra, D. Micrococcal nuclease does not substantially bias nucleosome mapping. *J Mol Biol* **417**, 152-164, doi:10.1016/j.jmb.2012.01.043 (2012).
36 Konig, J., Zarnack, K., Luscombe, N. M. & Ule, J. Protein-RNA interactions: new genomic technologies and perspectives. *Nat Rev Genet* **13**, 77-83, doi:10.1038/nrg3141 (2011).
37 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010).
38 Imai, T. *et al.* The neural RNA-binding protein Musashi1 translationally regulates mammalian numb gene expression by interacting with its mRNA. *Mol Cell Biol* **21**, 3888-3900, doi:10.1128/MCB.21.12.3888-3900.2001 (2001).
39 Zearfoss, N. R. *et al.* A conserved three-nucleotide core motif defines Musashi RNA binding specificity. *J Biol Chem* **289**, 35530-35541, doi:10.1074/jbc.M114.597112 (2014).
40 Ohyama, T. *et al.* Structure of Musashi1 in a complex with target RNA: the role of aromatic stacking interactions. *Nucleic Acids Res* **40**, 3218-3231, doi:10.1093/nar/gkr1139 (2012).
41 Kwon, H. Y. *et al.* Tetraspanin 3 Is Required for the Development and Propagation of Acute Myelogenous Leukemia. *Cell Stem Cell* **17**, 152-164, doi:10.1016/j.stem.2015.06.006 (2015).
42 Sellin, M. E., Holmfeldt, P., Stenmark, S. & Gullberg, M. Op18/Stathmin counteracts the activity of overexpressed tubulin-disrupting proteins in a human leukemia cell line. *Exp Cell Res* **314**, 1367-1377, doi:10.1016/j.yexcr.2007.12.018 (2008).
43 Amayed, P., Pantaloni, D. & Carlier, M. F. The effect of stathmin phosphorylation on microtubule assembly depends on tubulin critical concentration. *J Biol Chem* **277**, 22718-22724, doi:10.1074/jbc.M111605200 (2002).
44 Kouzu, Y. *et al.* Overexpression of stathmin in oral squamous-cell carcinoma: correlation with tumour progression and poor prognosis. *Br J Cancer* **94**, 717-723, doi:10.1038/sj.bjc.6602991 (2006).
45 Rubin, C. I., French, D. L. & Atweh, G. F. Stathmin expression and megakaryocyte differentiation: a potential role in polyploidy. *Exp Hematol* **31**, 389-397 (2003).
46 Machado-Neto, J. A., Saad, S. T. & Traina, F. Stathmin 1 in normal and malignant hematopoiesis. *BMB Rep* **47**, 660-665 (2014).
47 Bagger, F. O. *et al.* BloodSpot: a database of gene expression profiles and transcriptional programs for healthy and malignant haematopoiesis. *Nucleic Acids Res* **44**, D917-924, doi:10.1093/nar/gkv1101 (2016).

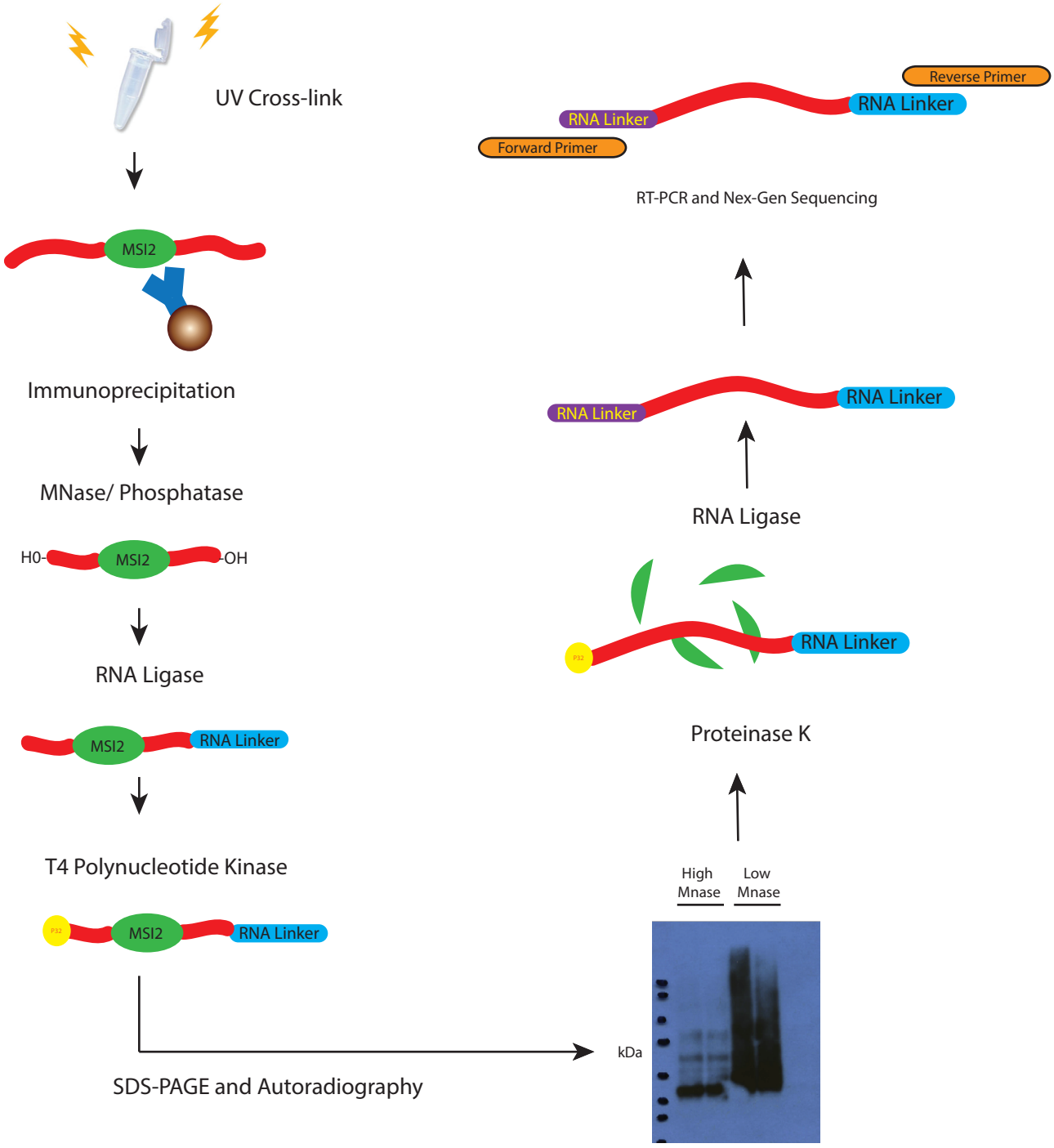


Figure 1. Overview of the CLIP-Seq protocol. Cells are harvested, cross-linked with 400mj/cm2 UV-B radiation and protein-RNA complexes are isolated *via* immunoprecipitation. Cross-linked RNAs are trimmed using micrococcal nuclease (Mnase) and dephosphorylated. A 3'RNA adapter is ligated to the RNA molecules. Bead complexes are extensively washed to remove excess adapter and RNA is 5' labelled with P32. RNA-protein complexes are run on an SDS-PAGE gel and transferred to a nitrocellulose membrane. The region of the membrane 15-25kDa above the molecular weight of the protein is extracted and incubated with proteinase K. RNA molecules are collected *via* acid-phenol-chloroform extraction and a 5'RNA linker is ligated. Adapter ligated RNA is reverse transcribed and PCR-amplified.

	CLIP-Seq Bioinformatic Workflow	
Step	Program	Purpose
1	Cutadapt	Remove adapter sequences and low-quality reads
2	Bowtie	Align reads to RepBase (database of repetitive elements). Reads that align are discarded.
3	STAR	Align reads to the reference genome. Allows for gapped alignment
4	SAMTools	Collapse PCR duplicates
5	CLIPPER	Analyze all of the reads that were aligned to the human genome and determine regions where significant 'clustering' occurs
6	HOMER	Identify enriched motifs in the genomic regions identified by CLIPPER

Figure 2. CLIP Bioinformatic Steps. This table presents a chronological list of software programs that are used in the analysis of CLIP-Seq data.

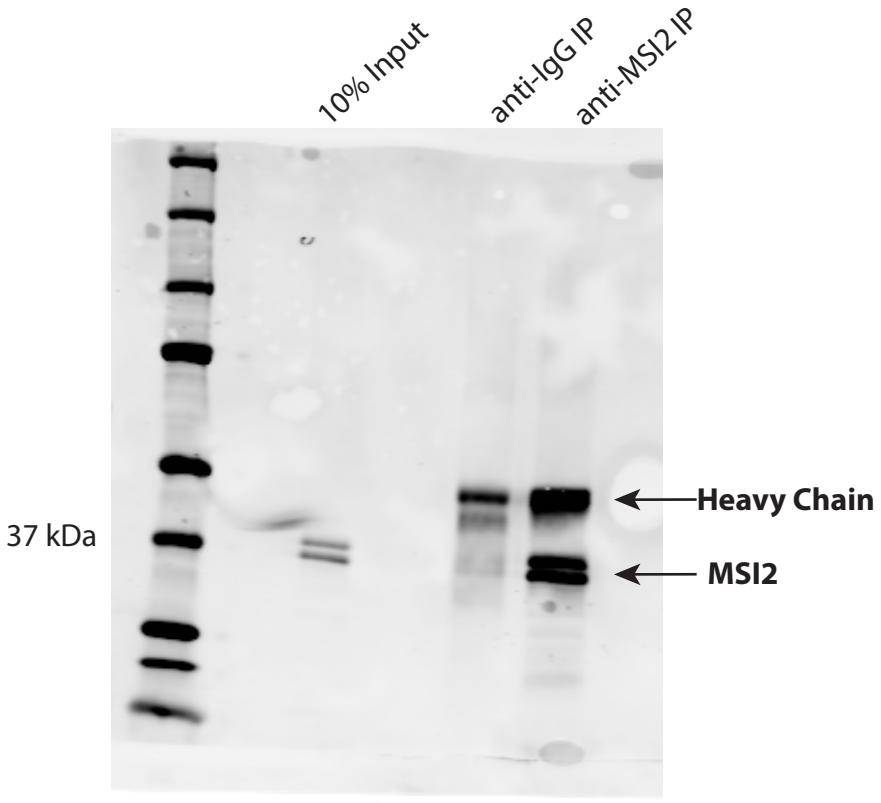


Figure 3. MSI2 Pulldown. 1 million NB4 cells were lysed in 0.1% SDS, 0.5% Deoxycholate, and 0.5% NP-40 in PBS. 10% of the lysate was set aside and used as an input control. 5µg of anti-MSI2 or 5µg of rabbit IgG were incubated with 50µL of washed protein A dynabeads for 2 hours. Antibody-bead complexes were washed and incubated with equal amounts of the remaining lysate. Immunoprecipitation was performed in a 1mL volume (topped up with lysis buffer) overnight. Antibody-beads complexes were boiled in Laemmli buffer and subjected to SDS-PAGE. Anti-MSI2 was used to detect MSI2 protein in the pulldown.

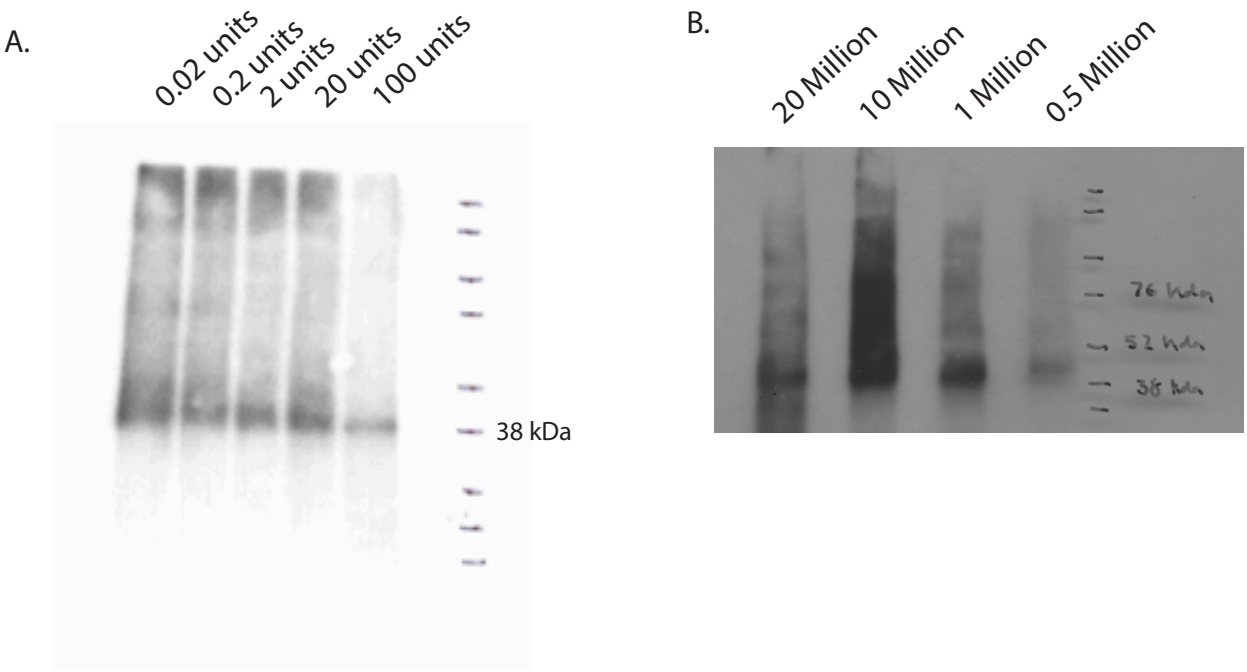


Figure 4. MSI2 UV crosslinking optimization. (A) NB4 cells were treated with a varying amount of MNase and subjected to radiactive western blotting. Blots were detected *via* autoradiography. 100 units of MNase resulted in a dramatic loss of larger protein-RNA products. Incubation with 0.2 and 0.02 units of Mnase appeared to produce the most uniform smearing patterns. (B) Radiolabelled protein-RNA complexes could be detected when as little as 0.5 million input NB4 cells were used.

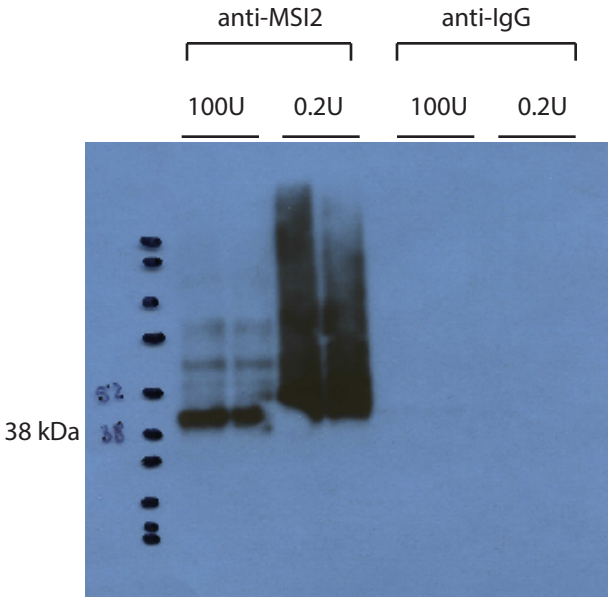


Figure 5. High and low Mnase and IgG controls. 25 million NB4 cells were immunoprecipitated with either anti-MSI2 or an IgG control. Immunoprecipitated cells were incubated with either 100U or 0.2U of Mnase. RNA-protein complexes were radiolabelled, radioactive western blotting was performed and complexes were detected *via* autoradiography. When MSI2 was immunoprecipitated, significant radioactive signals were detected. When an IgG immunoprecipitation was performed, no radioactive signal was detected. When MSI2-RNA complexes were treated with 100U of Mnase, one predominant band was detected near the molecular weight of MSI2. When MSI2-RNA complexes were treated with 0.2 units of MNase, a large smearing pattern was detected extending upwards from the molecular weight of MSI2

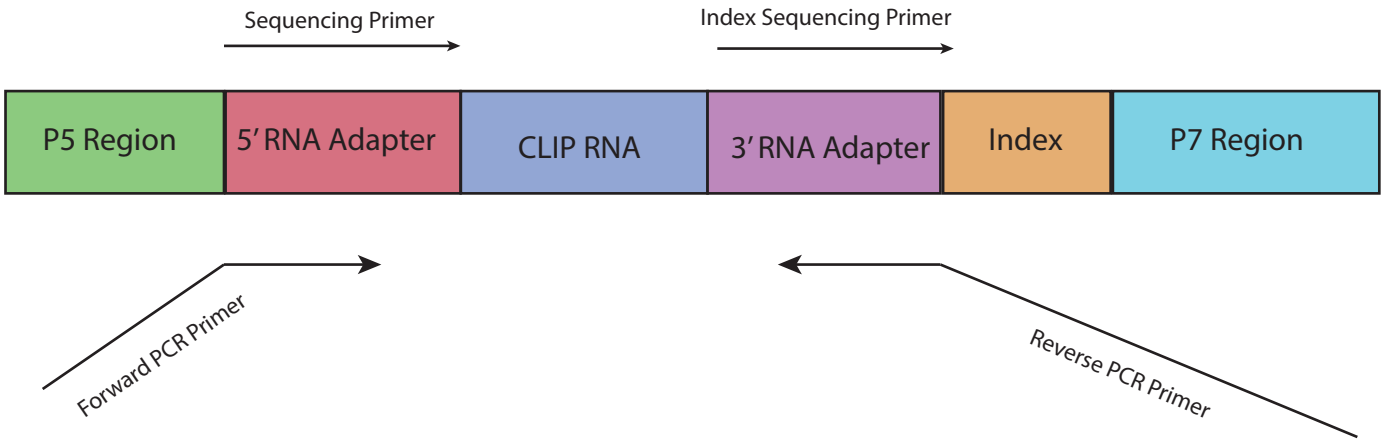
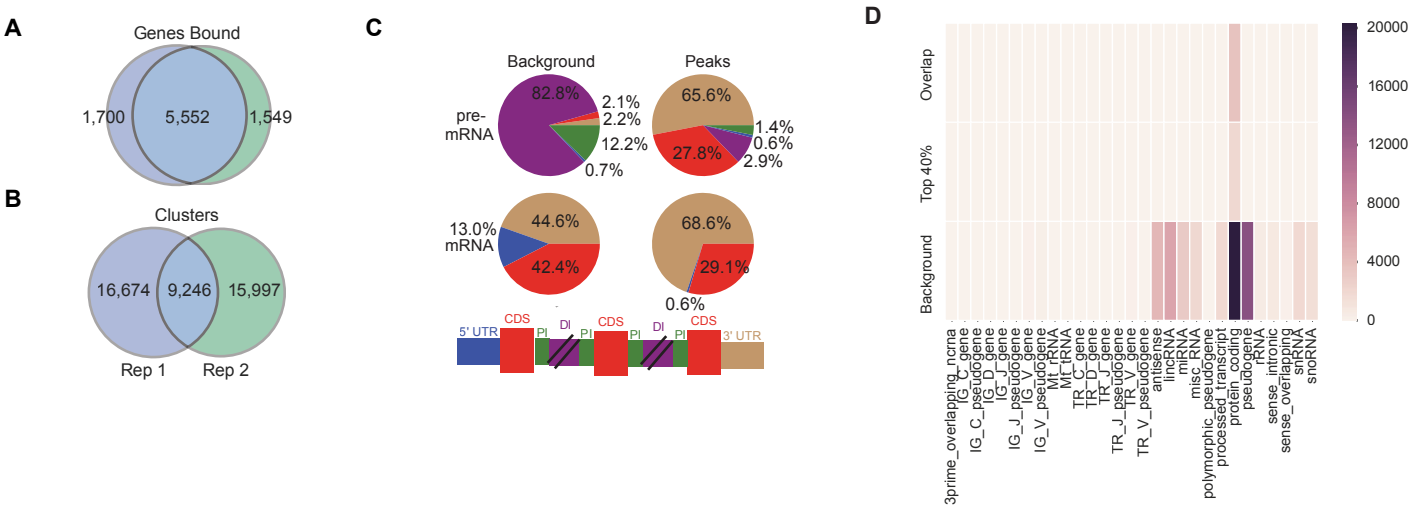
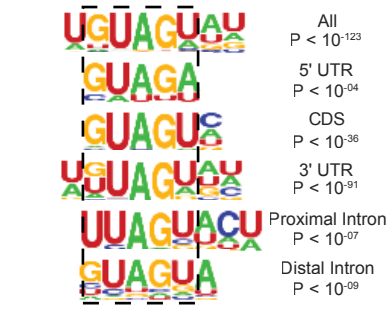


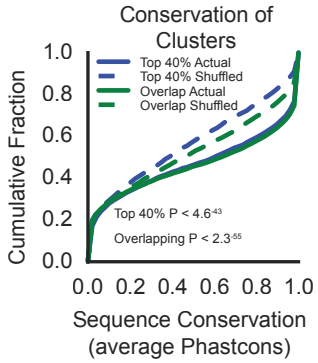
Figure 6. CLIP PCR Strategy. The 5'RNA adapter and the 3'RNA adapter contained sequences for the illumina read sequencing primer and index sequencing primers respectively. The forward PCR primer used to amplify the CLIP library annealed partially to the 5'adapter sequence. The forward primer contained the P5 DNA region that allows for binding to the illumina flow cell. The reverse primer used to amplify the CLIP library annealed partially to the 3'RNA adapter. The reverse primer contained a unique index sequence for each library and the P7 DNA sequence that allows for binding to the illumina flow cell

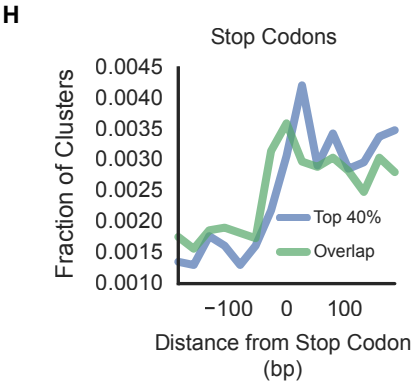
Rank	Chromosome	Start	End	Gene	P-Value
1	chr12	20704378	20704482	PDE3A	3.74E-91
2	chr1	91852786	91852929	HFM1	2.66E-90
3	chr11	18429004	18429132	LDHA	1.60E-72
4	chr11	65622297	65622473	CFL1	3.97E-64
5	chr15	77338371	77338510	TSPAN3	1.57E-61
6	chr1	26227334	26227453	STMN1	1.29E-55
7	chr11	62609107	62609254	WDR74	6.43E-55
8	chr16	83845127	83845243	HSBP1	2.05E-53
9	chr1	228682333	228682531	RNF187	7.75E-52
10	chr2	47387673	47387827	CALM2	2.75E-51
11	chr7	5646159	5646281	FSCN1	5.92E-46
12	chr1	26227089	26227208	STMN1	7.06E-46
13	chr8	98863800	98863932	LAPTM4B	1.63E-45
14	chr4	83274981	83275117	HNRNPD	8.68E-40
15	chr21	33040855	33040981	SOD1	1.33E-38
16	chr17	37006614	37006729	RPL23	1.52E-38
17	chr4	140211172	140211248	NDUFC1	1.17E-36
18	chr6	135539756	135539862	MYB	3.06E-36
19	chr2	44115318	44115415	LRPPRC	4.65E-36
20	chr6	163992585	163992684	QKI	6.28E-36
21	chr2	29023357	29023433	PPP1CB	6.27E-35
22	chr6	30692473	30692596	TUBB	2.60E-34
23	chr2	133012533	133012611	ANKRD30BL	1.44E-33
24	chr7	134127340	134127436	AKR1B1	1.96E-33
25	chr7	6442998	6443071	RAC1	1.05E-32
26	chr1	156711908	156712021	HDGF	1.43E-32
27	chr14	102547599	102547713	HSP90AA1	1.94E-32
28	chr2	38294846	38294935	CYP1B1	6.12E-32
29	chr3	32496142	32496220	CMTM7	1.08E-31
30	chr3	195777194	195777294	TFRC	1.28E-31
31	chr20	30781075	30781165	PLAGL2	2.72E-31
32	chr5	314760	314880	PDCD6	3.11E-31
33	chr13	46700688	46700853	LCP1	6.01E-31
34	chr2	201736074	201736153	PPIL3	8.24E-31
35	chr4	174253166	174253284	HMGB2	1.58E-30
36	chr15	64658017	64658120	KIAA0101	1.62E-30
37	chr6	31797755	31797868	HSPA1B	3.63E-30
38	chr1	245016832	245016912	HNRNPU	4.18E-30
39	chr1	151239762	151239874	PSMD4	5.81E-30
40	chr10	102123184	102123289	SCD	6.55E-30
41	chr4	174252988	174253069	HMGB2	1.09E-29
42	chr10	98281015	98281103	TM9SF3	1.20E-29

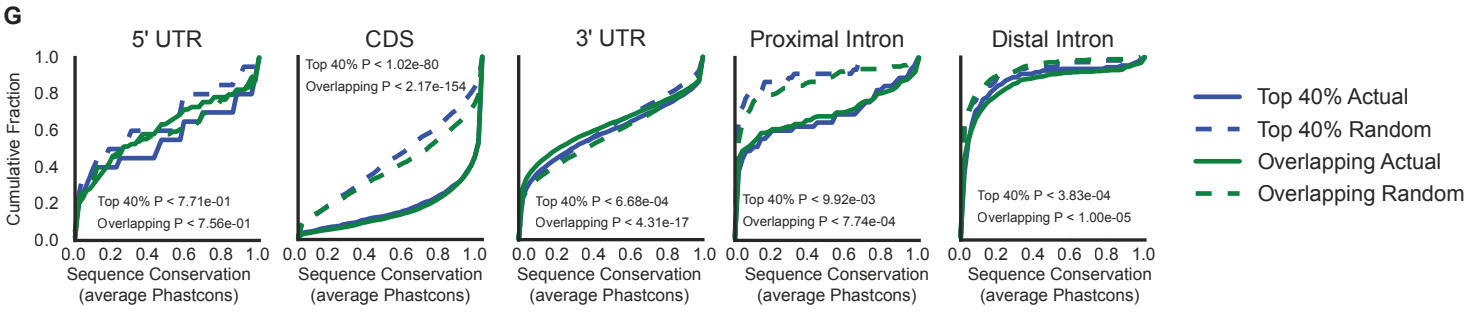
Figure 7. MSI2 CLIP-Seq ranked targets. Analysis of MSI2 CLIP-Seq reads by CLIPPER software revealed the following list of significantly enriched clusters. 'Start' indicates the starting genomic location of the cluster and 'End' indicates where the cluster terminates.


A
Genes Bound
1,700
5,552
1,549
B
Clusters
16,674
9,246
15,997
Rep 1
Rep 2
C
Background
Peaks
pre-mRNA
82.8%
2.1%
2.2%
12.2%
0.7%
65.6%
1.4%
0.6%
27.8%
2.9%
13.0% mRNA
44.6%
42.4%
68.6%
29.1%
0.6%
5' UTR
CDS
PI
DI
CDS
PI
DI
CDS
3' UTR
D
Overlap
Top 40%
Background
20000
16000
12000
8000
4000
0
3prime_overlapping_ncrna
IG_C_gene
IG_C_pseudogene
IG_D_gene
IG_J_gene
IG_J_pseudogene
IG_V_gene
IG_V_pseudogene
MT_tRNA
TR_C_gene
TR_D_gene
TR_J_gene
TR_J_pseudogene
TR_V_gene
TR_V_pseudogene
antisense
lincRNA
miRNA
misc_RNA
polymorphic_pseudogene
processed_transcript
protein_coding
pseudogene
rRNA
sense_intronic
sense_overlapping
snRNA
snoRNA


E
All
P < 10-123
5' UTR
P < 10-04
CDS
P < 10-36
3' UTR
P < 10-91
Proximal Intron
P < 10-07
Distal Intron
P < 10-09


F
Conservation of Clusters
Top 40% Actual
Top 40% Shuffled
Overlap Actual
Overlap Shuffled
Cumulative Fraction
Top 40% P < 4.6-43
Overlapping P < 2.3-55
Sequence Conservation (average Phastcons)


H
Stop Codons
Fraction of Clusters
Top 40%
Overlap
Distance from Stop Codon (bp)


G
5' UTR
CDS
3' UTR
Proximal Intron
Distal Intron
Cumulative Fraction
Top 40% P < 7.71e-01
Overlapping P < 7.56e-01
Top 40% P < 1.02e-80
Overlapping P < 2.17e-154
Top 40% P < 6.68e-04
Overlapping P < 4.31e-17
Top 40% P < 9.92e-03
Overlapping P < 7.74e-04
Top 40% P < 3.83e-04
Overlapping P < 1.00e-05
Sequence Conservation (average Phastcons)
Top 40% Actual
Top 40% Random
Overlapping Actual
Overlapping Random

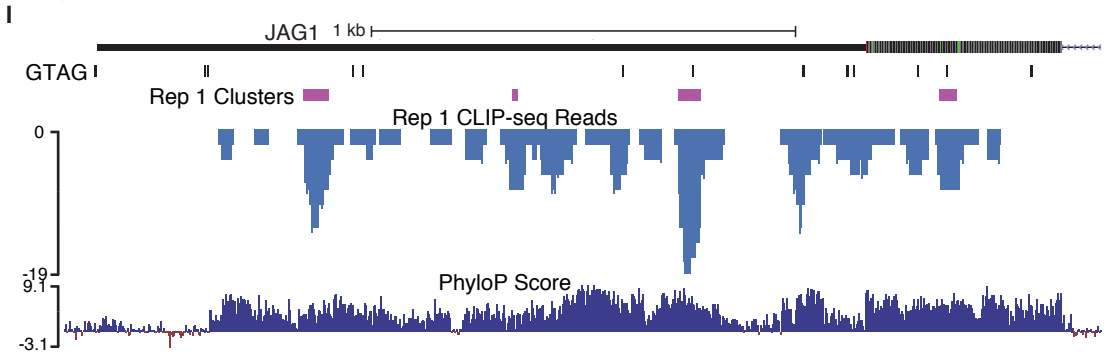

I
JAG1
1 kb
GTAG
Rep 1 Clusters
Rep 1 CLIP-seq Reads
PhyloP Score

Figure 8. MSI2 preferentially binds mature mRNA within the 3'UTR
(A) Venn diagram of overlap between MSI2 target genes in two replicate experiments. (B) Venn diagram showing that a statistically significant overlap ($p<0.0001$, by hypergeometric test) of the clusters between the two replicates. (C) Pie charts showing the percent of CLIP-seq clusters in different genic regions, namely 5'UTR, coding exons (CDS), proximal (<500 bases from an exon) and distal (>500 bases) intronic regions and 3'UTRs. (D) Heatmap indicating the number of different classes of Gencode annotated genes that contain at least one predicted MSI2 binding site. (E) Consensus motifs within MSI2 clusters in the different genic regions. P-values for the most statistically significant enriched motif is presented for the top 40% of overlapping clusters between replicates. (F) Cumulative distribution function of mean conservation score (Phastcons) of MSI2 clusters, compared to a shuffled background control, computed for all overlapping clusters and the top 40% of overlapping clusters. P-values were obtained by a Kolgomorov-Smirnov two-tailed test comparing the distributions from actual and shuffled locations. (G) Cumulative distribution function of mean conservation score (Phastcons) of MSI2 clusters, compared to a shuffled background control, computed for overlapping clusters between the replicates and the top 40% of overlapping clusters found the different genic regions. P-values were obtained by a Kolgomorov-Smirnov two-tailed test comparing the distributions from actual and shuffled locations. (H) Number of clusters within 200 bases of the annotated stop codon in known mRNA transcripts for all overlapping clusters between replicates and the top 40% of overlapping clusters. (I) Genome browser views displaying CLIP-seq mapped reads from replicate 1 (blue), predicted clusters (purple), exact matches for the GUAG sequence (black) and mammal conservation scores (PhyloP) in the 3' UTRs for a previously predicted MSI1 target.

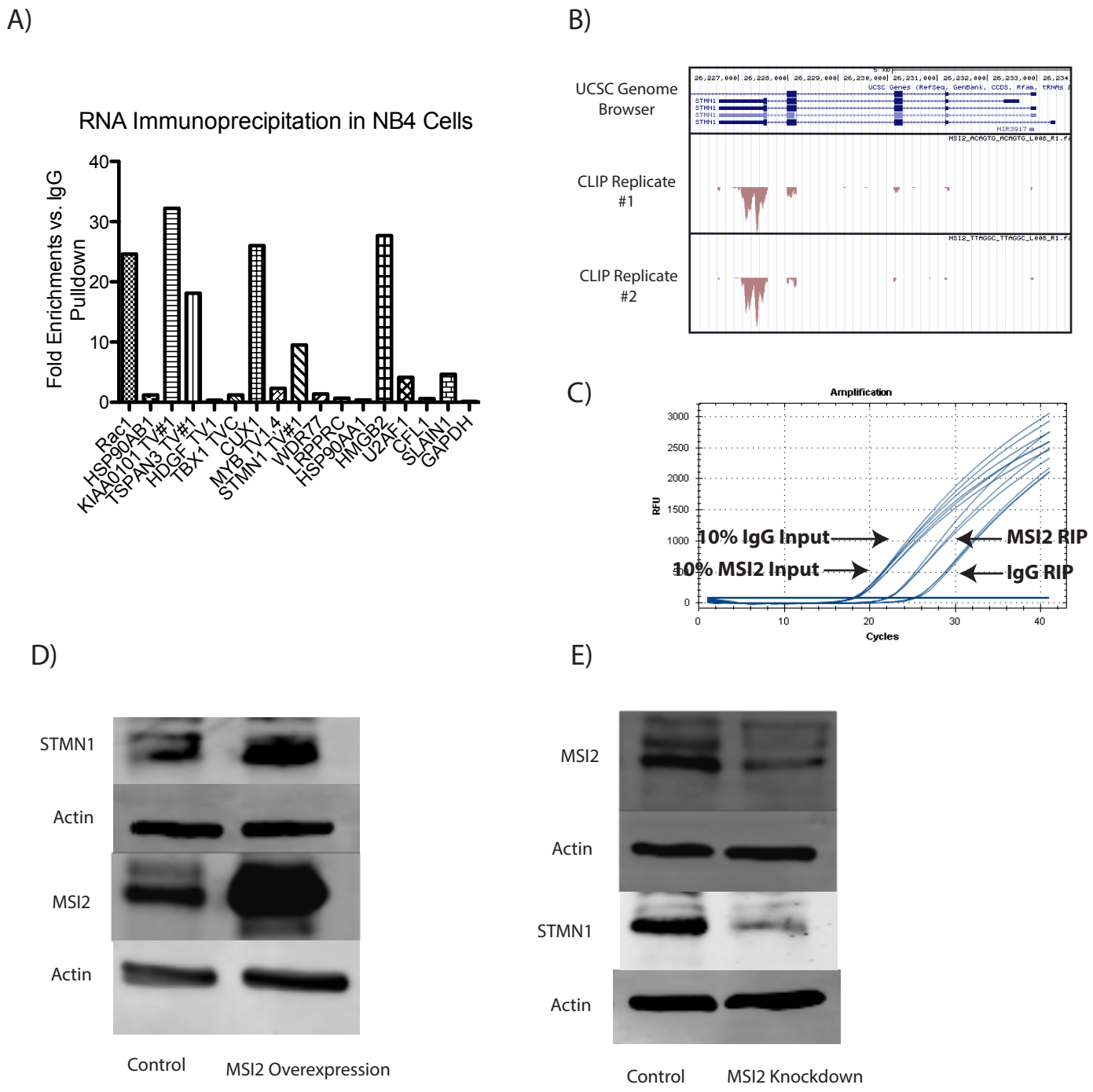


Figure 9. MSI2 RNA immunoprecipitation and the valdiation of STMN1 as an MSI2 target. (A) 16 CLIP seq targets were analyzed by qPCR after MSI2 RNA-immunoprecipitation. 6 targets (Rac1, KIAA0101, TSPAN3, Cux1, STMN1, and HMGB2 were enriched >10-fold vs an IgG pulldown. (B) Duplicate CLIP-Seq experiments show significant clusters on the STMN1 3'UTR. Clusters are identical between CLIP replciates. (C) qPCR was performed in triplicate on RNA isolated from MSI2 and IgG immunoprecipitations. qPCR was also performed in 10% input samples taken prior to IgG and MSI2 immunoprecipitation. qPCR analysis reveals a significant enrichment of STMN1 RNA in MSI2 IPs vs. IgG despite identical levels in the input RNA samples. (D) Overexpression of MSI2 results in an upregulation of STMN1 protein levels. (E) Knockdown of MSI2 results in the down-regualtion of STMN1 protein levels

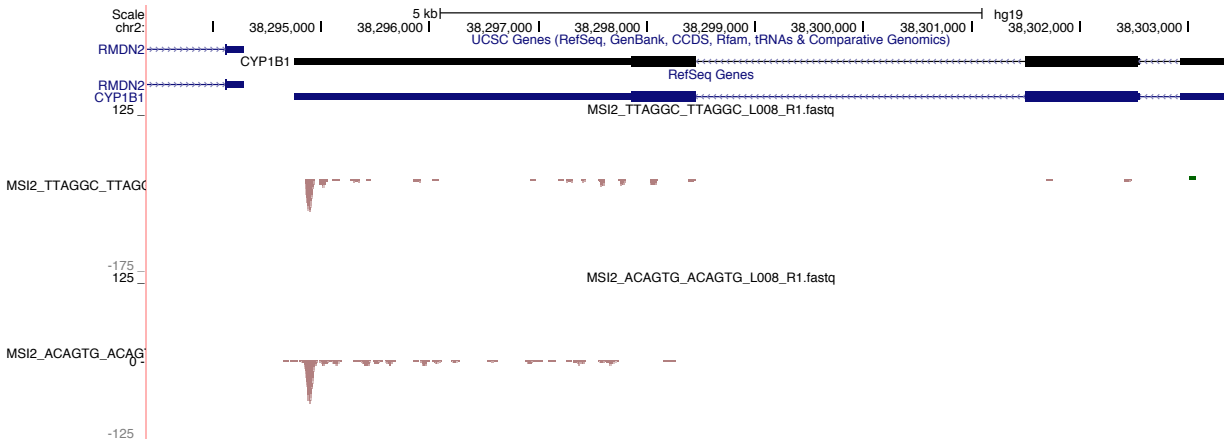


Figure 10. MSI2 binds to the 3'UTR of Cyp1b1. MSI2 CLIP-Seq reads were aligned to the hg19 human reference genome. Duplicate CLIP-Seq experiments identify a highly significant cluster located on the terminal end of the 3' UTR

Chapter 4:
Identification of MSI2 Protein Interactors

Abstract

The molecular circuitry occurring within Hematopoietic Stem and Progenitor Cells (HSPCs) is largely unknown. Though our understanding of the transcriptional networks occurring in HSPCs is growing due to advances in sequencing technology, our knowledge of protein circuitries lags far behind. This is predominantly due to difficulties in identifying proteins in such small cell populations. Current data suggests that Musashi-2 (MSI2) is a critical regulator of hematopoietic stem cell self-renewal. Despite this, our knowledge of the precise function of the MSI2 protein remains limited. The precise binding of MSI2 to specific mRNAs, despite its relatively simple consensus motif, in combination with its enigmatic role in translational control, suggests that this protein does not function in isolation. Importantly, protein complexes that include RNA-binding protein (RBPs) are known to affect RBP target specificity and affinity and control the fate of the mRNA targets. In order to identify MSI2 protein interactors, we employed clustered-regularly interspaced short palindromic sequence (CRISPR)-mediated tagging of the *Msi2* locus to generate a cell line that expressed an endogenous MSI2-BirA* protein. BioID analysis of this cell line revealed numerous proteins that were in close proximity to the MSI2 protein *in vivo*. We identified Insulin-like Growth Factor 2 mRNA Binding Protein 2 (IGF2BP2) as an MSI2 binding protein that can be detected *via* Co-IP assays. Impressively, analysis of mass spectrometry data indicates that Igf2bp2 is the most significantly down regulated protein when murine long-term

hematopoietic stem cells (LT-HSCs) first commit to short term-HSCs (ST-HSCs). Transplantation assays reveal a critical role for the IGF2BP2 protein in murine HSC function. We hypothesize that IGF2BP2/MSI2 protein complexes play a critical role in the regulation of target mRNAs in LT-HSCs and that the downregualtion of *Igf2bp2* results in altered regulation of target mRNAs in more committed cells.

Introduction

HSPCs consist of multipotent cells that are able to maintain the production of billions of mature blood cells everyday. The molecular pathways active in these cells control self-renewal and differentiation in order to maintain hematopoietic function throughout the lifetime of an organism and to enhance levels of hematopoiesis during times of stress. Having a thorough understanding of the molecular mechanisms that control hematopoiesis can help enhance their clinical application and can aid in our understanding of aberrant self-renewal events such as those that are thought to occur in leukemia. Despite this, our understanding of the molecular circuitry that is present within these cells remains quite poor. Transcript profiling of HSPC populations has allowed for the identification of complex gene regulatory patterns that are critical for the maintenance of multipotency and commitment to differentiation[1-3]. However, very little is known about the proteomic composition and even less is known about protein-protein interaction networks that are active within these immature cells[2]. Until recently, large-scale protein data from fractionated hematopoietic populations have been lacking due to the scarcity of these cells. However, advances in technology have allowed for the identification of

hematopoietic stem and progenitor cell proteomes allowing for a more in-depth analysis of protein regulatory networks active in these cells[2,4]. Importantly, these studies have supported the idea that specific post-transcriptional mechanisms may be active in HSCs[2].

The MSI2 protein has previously been identified as a potent regulator of murine and human hematopoietic stem cells[5,6]. Identification of the RNA-networks regulated by MSI2 has shed light on its role in HSPC maintenance and has helped in the identification of potentially novel regulators of HSPCs. However, no studies to date have extensively studied the MSI2 protein interactome. RNA-binding proteins are commonly found in complexes with other proteins[7,8]. Their incorporation into larger protein complexes has the potential to alter their RNA-binding profiles and can impact the fate of RNA targets[7]. *Msi2* is expressed in both long-term and short-term HSCs but also in multipotent progenitors (MPPs) as well and in lymphoid- and myeloid-biased progenitors[2]. All of these cell populations have unique functional properties and unique molecular programs are thought to underlie either the maintenance of self-renewal or the push to differentiation[9]. Importantly, though a loss of MSI2 can greatly impact the self-renewal function of long-term HSCs, the protein is also highly expressed in progenitor cells with limited to no abilities for self-renewal[2]. It is likely that unique circuitries are active in these different cell populations and that the MSI2 circuitry existing in LT-HSCs is ultimately responsible for its functional role in self-renewal. Differences in MSI2 function may exist due to differences in protein binding partners that affect the RNA-binding profile or the fate of bound MSI2 targets.

In the current body of work, we elucidate novel MSI2 protein interactors and identify the IGF2BP2 protein as a potent MSI2-binding partner that is highly expressed at the protein level in LT-HSCs and whose protein expression drops significantly upon transition to ST-HSCs and MPPs. Preliminary experiments show that the IGF2BP2 protein is critical for the maintenance of murine HSCs and we postulate that the MSI2-IGF2BP2 interaction is critical for the proper function of murine HSCs.

Materials and Methods

Generation of MSI2-BirA mESC cell lines*

Nickase guides targeting the genomic region downstream of the MSI2 stop codon were designed using MIT CRISPR design software at http://crispr.mit.edu. The guides: TGTGTACACGTATGAGCGTA and CTCGCTGAACCCCCTTTCAA were cloned into the pSpCas9n(BB)-2A-Puro (pX462) as previously described[10]. MSI2-BirA* and MSI2-P2A-BirA* repair templates were generated by cloning MSI2 genomic regions out of BAC RP23-19J8 into the pL452 vector (Figure 1, p.178). The 800bp region upstream of the MSI2 stop codon was amplified using primers F-AGGTTACATCCTGCATCAGATTC and R- GTGGTATCCATTTGTAAAGGCCG. The BirA-Flag and P2A-BirA-Flag sequences were cloned using the primers F-AAGGACAACACCGTGCCCCT R- TTACTTGTCATCGTCATCCTTG P2A Rev-TTACTTGTCATCGTCATCCTTGTAATC. The genomic region immediately downstream of the MSI2 stop codon up to 300 bases into the first intron was amplified using the primers F- GCAGGCGCTTCCATTGCC and R-

CACTCAAGAAGGAAAAAGATGAGGG. These products were infused into the pL452 vector between the KpnI and SalI sites. The 800bp sequence continuing downstream from the MSI2 intron was amplified using the primers F- GGAAACTGGAAGGTGCTCTC and R- CATATAGAGGGTTTACTTGGTTAG and inserted into the pL452 vector between the BamHI and NotI cut sites. To generate MSI2-BirA* and MSI2-P2A-BirA* cell lines, 3 million mouse embryonic stem cells (mESCs) were transfected with 3ug of forward and reverse pX462 nickase vectors and 1ug of the repair template. Cells were incubated with 250ug/mL of G418 at 48 hours post transfection for a 10-day period. Individual colonies that grew out under G418 selection were picked into duplicate 96-well plates and PCR screens were performed to identify clones that contained the MSI2-BirA* fusion. One of the 96-well plates was aspirated, washed with PBS, and incubated at -80 degrees for 1 hour. Cells were then allowed to warm to room temperature and lysed in 30uL of 1X digestion buffer (1X Taq PCR buffer, 0.5% NP40, 0.5% Tween-20, 1mg/mL proteinase K). The plate was then incubated at 60 degrees for 2 hours followed by 95 degrees for 15 minutes. 3uL of the crude extract was used as input in a 50uL PCR reaction. One PCR screen targeting the 5' end of the repair template was performed using the primers F- GTA AGG CGC AAG CCC TGT CTG and R- CTT GGA AAA TGC TGG TGT CTC CAC TC. Another PCR screen targeting the 3' end of the repair template was performed using primers F- CTG CAG CTC CCC AAA GGC C and R- GCT CTT GGG GTG AGG AGC AGA TAC. Msi2-P2A-BirA* clones and Msi2-BirA* clones that were positive for 5' and 3' screens were further validated by western blotting. Neo excision was performed on positive clones by transfecting 1 million cells with pCX-NLS-Cre-Puro.

24 hours post-transfection, cells were treated with 1.5ug/mL puromycin for 2 days. Cells were then plated at a density of 5000 cells per 10cm plate and allowed to grow until colonies were visible to the naked eye. Individual colonies were screened for Neo excision (Figure 2, p.179).

MSI2-BirA BioID*

2 MSI2-BirA* and 2 MSI2-P2A-BirA* mESC cell lines were each grown in 5 X 15cm plates in standard mESC media (DMEM, 0.1mM NEAA, 6mM L-glutamine, 0.1mM B-mercaptoethanol, 15% FBS, 6ng/mL LIF) on 0.1% gelatin coated plates. Cells were grown until 60% confluency and then incubated with standard mESC media + 50uM biotin for 24 hours. Cells were then scraped and pelleted in 15-mL tubes and incubated at 4 degrees for 1 hour on an end-over-end rotator. Samples were centrifuged for 30 minutes at 16000rpm and the supernatant transferred to a new 15mL tube. 36uL of streptavidin-sepharose beads were washed twice in RIPA and incubated with cleared supernatants for 6 hours at 4 degrees. Samples were centrifuged at 2000rpm for 2 minutes and washed 4X with 50mM ammonium bicarbonate. Beads were snap frozen and sent to Institute for Research in Immunology and Cancer (IRIC) at the university of Montreal for mass spectrometry analysis.

Co-immunoprecipitations

3 million mESCs were washed in PBS and pelleted. Cells were lysed in 700 **µ**L 1X Gentle Soft Lysis Buffer (10mM NaCl, 0.5% NP40, 0.5% 2-mercaptoethanol, 5mM

EDTA, 20mM PIPES pH 7.4) and incubated on ice for 20 minutes. Tubes were centrifuged at 15 000 rpm for 15 minutes at 4 degrees and the supernatant was transferred to a new tube. 5**µ**g of anti-MSI2 (EP1305Y) or 5**µ**g of rabbit IgG (sc-2027) was added to the supernatant and the solution was allowed to rotate overnight at 4 degrees. 20**µ**L of protein G Dynabeads were washed with gentle soft lysis buffer and added to the antibody-lysate solution for 2 hours at 4 degrees. Unbound supernatant was removed and bead complexes were washed 4X with gentle soft buffer. Proteins were eluted off beads by boiling in 40**µ**L of 1X Laemmlli buffer. 20**µ**L of protein lysate were run on a 10% Bis-Tris polyacrymlaide gel, transferred to a PVDF membrane, blocked with 5% BSA and probed with anti-MSI2 (EP1305Y)and anti-Igf2bp2 (D4R2F).

Igf2bp2 Knockdown, CFU-Assay and Transplantation

MicroRNA-embedded shRNAs targeting control (shLuciferase-CCGATATGGGCTGAATACAAAT) and mouse *Igf2bp2* (5-shIGF2BP2-CTCCCTTAGAGATTTTGTAAAA) were designed from the latest sensor-based rules or sequence scores and cloned into the pZIP-mEF1a-miR-E lentiviral backbone (TransOMIC Technologies; further adapted in lab)[11,12] (Figure 3, p.180). Clones were sequenced and tested in mouse embryonic fibroblasts (MEFs); validated shRNAs were virally packaged alongside pMD2.G (Addgene #12259) and psPAX2 (Addgene #12260) vectors *via* transfection of HEK293FT cells using Lipofectamine LTX (ThermoFisher Scientific). Concentrated viral supernatant was titrated on HeLa cells and a multiplicity of infection of 100 was used to infect CD45.1+Lin-

CD150+CD48- flow-sorted populations from fresh mouse bone marrow. Sorted cells were pre-plated at 2,500 cells per well of a 96-well ultra-low binding plate in StemSpan SFEM media (Stem Cell Technologies) supplemented with mSCF (100 ng/mL), mTPO (100 ng/mL), mIL-3 (10 ng/mL) and mIL-6 (10 ng/mL) cytokines approximately 20 hours prior to the addition of viral supernatant. Cells were kept in culture for another 72 hours before gene transfer levels were measured *via* flow cytometry (Day 0) and half-well cell culture equivalents were intravenously transplanted into lethally-irradiated C57Bl/6 mice along with 1x10^5 whole bone marrow competitor cells per recipient mouse (Figure 3, p.180). Peripheral blood grafts were monitored at 4-week intervals *via* flow cytometry up to16 weeks post-transplant, after which animals were sacrificed to further analyze grafts within the bone marrow, spleen and thymus. 300 GFP+ CD45.1CD150+CD48- cells were plated into 1.1mL of murine colony gel (reachbio #1202) and plated in duplicate into 35mm plates. Plates were incubated for 10 days and BFU-E, CFU-G, CFU-M, CFU-GM, and CFU-GEMM were counted.

Construction of Igf2bp2-P2A-ZsGreen Lentiviral Particles

Murine *Igf2bp2* cDNAs were ordered from the GE Dharmacon mammalian genome collection. *Igf2bp2* transcript variant #1 was amplified from the MMM1013-202798440 plasmid using the primers F- ATGATGAACAAGCTGTACATTGGGA and R-CTTGCTGCGCTGTGGGG. IGF2BP2 transcript variant #2 was amplified from the MMM1013-202709335 plasmid using the primers F-ATGGAAGTTGACTACTCAGTCTCT and R- CTTGCTGCGCTGTGGGG. The P2A sequence

was amplified using the primers F- GGAAGCGGAGCTACTAACTT and R- AGGTCCAGGGTTCTCCTC and the ZsGreen sequence was amplified using the F- ATGGCCCAGTCCAAGCAC and R- TCAGGGCAAGGCGGA primers. The IGF2BP2, P2A, and ZsGreen sequences were infused into the pZIP-mEF1a vector between the AgeI and MluI sites.

Results and Discussion

BioID reveals MSI2 protein interactors

BioID analysis of endogenously tagged MSI2 in mESCs revealed 8 proteins that showed differential patterns of biotinylation between MSI2-BirA* and MSI2-P2A-BirA* controls[13] (Figure 4, p.181). These included insulin-like growth factor 2 mRNA-binding protein 1 (IGF2BP1), Musashi-2 (MSI2), Insulin-like growth factor 2 mRNA-binding protein 2 (IGF2BP2), Far upstream element binding protein 3 (FUBP3), Elongation factor 1 alpha 1 (EF1A1), Ataxin-2-like protein (ATXN2L), Proline-rich coiled-coil 2C (PRRC2C), and Polyadenylate binding protein 1 (PABPC1). All of these proteins are thought to act, at least in part, as RNA binding proteins[14]. Importantly, mass spectrometry data revealed the presence of MSI2 in the list of preferentially biotinylated proteins. Since the BirA* enzyme is located immediately downstream of the MSI2 protein in experimental lines but not in P2A control lines, we expected to see an enrichment of biotinylated MSI2 and this discovery suggests that our CRISPR-BioID technique is acting to preferentially guide BirA* to the normal physiological locations where MSI2 functions.

Of immediate interest to us, was the detection of both IGF2BP1 and IGF2BP2 as proteins that are preferentially biotinylated when BirA* is covalently attached to the MSI2 protein. IGF2BP1 and IGF2BP2 are members of the insulin-like growth factor 2 mRNA binding protein family[15]. The detection of both family members in our BioID screen suggested to us that MSI2 might form unique protein-protein interactions with these family members. The insulin-like growth factor 2 binding proteins (IGF2BPs) are a family of three cytoplasmic RNA-binding proteins: IGF2BP1, IGF2BP2, and IGF2BP3. IGF2BP1 and IGF2BP3 are referred to as 'oncofetal' proteins since they are only detected during embryogenesis but are commonly up regulated in tumor samples[15]. In contrast, the IGF2BP2 protein is expressed in numerous adult tissues. Not much is known about their RNA targets or mechanism of action but these RNA binding proteins are generally thought to bind target mRNAs in order to control their stability, transport, or translation[15]. IGF2BP1 and IGF2BP3 are thought to act as post-transcriptional drivers of cancer progression[15]. IGF2BP3 is commonly detected in a variety of malignant neoplasms but is absent from adjacent benign tissues. High expression of *Igf2bp3* is associated with poor outcomes in ovarian cancer, cervical carcinoma, hepatic carcinoma, and renal carcinoma[16-19]. Strikingly, a retrospective study involving 501 renal cell tumours revealed that patients with IGF2BP3 positive tumours had dramatically reduced rates of 5-year overall survival[19]. Patients with stage I tumours that were IGF2BP3 positive had a 32% overall survival rate while those with IGF2BP3 negative tumours had an 89% survival rate. This significant difference in overall survival was seen regardless of tumour grade. Notably, patients with stage III

IGF2BP3 positive tumours had an overall survival of 14% while those patients with IGF2BP3 negative tumours had a 58% overall survival rate. In a similar manner, overexpression of *Igf2bp1* associates with advanced clinical stage and poor outcome in a variety of tumour samples as well. *Igf2bp1* expression correlates with poor patient outcomes in ovarian cancer, colorectal cancer, and neuroblastoma[15,20,21]. In the leukemic setting, *Igf2bp1* is highly expressed in t(12;21)(p13;q22)-positive ALL[22]. Its down regulation in these samples impairs leukemic growth by attenuating cell cycle progression and increasing rates of cell death. Though numerous studies implicate IGF2BP1 and IGF2BP3 as potent oncogenes whose expression is detrimental to patient outcome, not many studies implicate IGF2BP2 in aggressive tumour formation. Instead IGF2BP2, is implicated in a variety of normal physiological processes such metabolism and muscle cell motility[23,24].

Igf2bp2 is uniquely expressed in dormant HSCs

Importantly, our MSI2-BirA* screen was performed in mESCs and thus we decided to examine whether *Igf2bp1* or *Igf2bp2* were expressed in the hematopoietic context to determine whether a MSI2 and IGF2BP2 interaction could occur within hematopoietic stem or progenitor cells. To do this, we consulted a recently published hematopoietic stem and progenitor cell proteome and transcriptome database[2]. Importantly, hematopoietic stem and progenitor cells are thought to have unique gene expression patterns that control self-renewal and differentiation[2,9]. This database provides a comprehensive insight into the molecular profiles of these early HSPCs.

Five populations of stem and progenitor cells were transcriptionally profiled in the aforementioned database; these five populations are referred to as 'HSC', 'MPP1', 'MPP2', 'MPP3', and 'MPP4'[2] (Figure 5, p.182). Additionally, the 'HSC' and 'MPP1' proteomes were analyzed *via* mass spectrometry. These populations are hierarchically organized based on reconstitution capabilities and cell-cycle status with 'HSC' at the apex followed by 'MPP1', 'MPP2', 'MPP3', and 'MPP4'[2,9]. The cell surface marker profiles for the various populations were as follows: HSC: LSK CD150+CD48-CD135-CD34-, MPP1: LSK CD150+CD48-CD135-CD34+, MPP2: LSK CD150+CD48+CD135-CD34+, MPP3: LSK CD150-CD48+ CD135- CD34+, and MPP4: LSK CD150-CD48+ CD135+CD34+. The LSK CD150+CD48- immunophenotype identifies a highly purified population of hematopoietic stem cells that give rise to long-term multilineage engraftment[25]. Both 'HSC' and 'MPP1' populations represent functional hematopoietic stem cells yet the HSC population outlined above represents a much more quiescent population of cells when compared to the 'MPP1' population[9]. The acquisition of CD34 is thought to be an early event that corresponds to the activation of HSCs from their dormant state[9]. CD48+ cells are only capable of short-term engraftment[2]. The 'MPP2' population represents an early multipotent progenitor population. 'MPP3' is a myeloid biased MPP population while 'MPP4' is a lymphoid biased MPP population[2].

In analyzing the transcriptomes of the 'HSC', 'MPP1', 'MPP2', 'MPP3', and 'MPP4' populations, we observed that *Igf2bp2* transcript levels were highly expressed in the HSC populations and dropped significantly as cells became more differentiated. In accordance with its 'oncofetal' expression pattern, *Igf2bp1* was not

expressed in any of the hematopoietic cell populations. Strikingly, *Igf2bp2* transcript levels dropped 3-fold between 'HSC' and 'MPP1' populations and continued to drop as the proliferation status of the various cell populations increased and the cells became more differentiated. *Igf2bp2* mRNA levels were 1.74-fold *Gapdh* in 'HSC' populations, in which 70% of cells are in G0. *Igf2bp2* transcript levels dropped to 0.58-fold *Gapdh* in 'MPP1' populations, in which 40% of cells are in G0. In the most proliferative MPP population, 'MPP4', where only 16% of cells are in G0, *Igf2bp2* mRNA levels dropped to 0.06-fold *Gapdh*- a 29-fold decrease from 'HSC' *Igf2bp2* transcript levels[2]. Strikingly, when 'HSC' and 'MPP1' proteomes were compared, IGF2BP2 protein levels dropped 8.2-fold, making it the most potently down-regulated protein upon the switch from CD34 negative dormant HSCs to CD34 positive cycling HSCs[2] (Figure 6, p.183). *Msi2* transcript expression was less dynamic between these two stem cell populations. *Msi2* transcript levels were nearly 10-fold *Gapdh* in the 'HSC' population and dropped to 7.4-fold *Gapdh* in the 'MPP1' population, representing a 33% decrease in transcript levels. The proteomic comparison between 'HSC' and 'MPP1' populations revealed a relatively constant expression of MSI2 protein levels. *Msi2* transcript dropped between 'MPP1' and 'MPP2' populations. In the 'MPP2' population, *Msi2* transcript level was 3-fold that of *Gapdh* representing a 250% decrease from the 'MPP1' population. Importantly, the 'MPP2' population not only has a greater proportion of cycling cells when compared to 'HSC' and 'MPP1', but it shows a dramatically different reconstitution profile. Unlike 'HSC' and 'MPP1' populations, 'MPP2' cells are unable to support long-term

engraftment in a mouse[2,9]. The significant drop in *Msi2* transcript seems to correlate with the transition to a population of cells with limited self-renewal capabilities.

Co-immunoprecipitation identifies a direct interaction between Igf2bp2 and MSI2

Based on this data, we set forth with the hypothesis that MSI2 and IGF2BP2 can directly interact in order to control hematopoietic stem cell quiescence and self-renewal. A critical first step was to prove that the MSI2 and IGF2BP2 proteins were in fact direct interactors. Importantly, the BirA* protein can biotinylate proteins that are direct binding partners with MSI2 but it can also biotinylate near-neighbors that are not direct interactors[13]. Co-IP analysis determined that MSI2 and IGF2BP2 are in fact direct protein-binding partners (Figure 7, p.184). With this in mind, we formulated a hypothesis where MSI2 and IGF2BP2 can form a complex in order to regulate a distinct set of mRNAs. Importantly, RBP complexes have been postulated to influence RNA-binding specificities and target fates and studies have indicated that RBP complexes can have distinct RNA binding profiles that are different from either of the constituent RBPs[7,8]. We postulated that MSI2 and IGF2BP2 RBP complexes might bind to and regulate unique RNAs that play a critical role in the maintenance of HSC quiescence. In such a situation, a significant decrease in IGF2BP2 protein would result in a significant decrease in IGF2BP2-MSI2 protein complexes resulting in the altered regulation of IGF2BP2-MSI2 RNA targets and a possible increase in HSC cycling.

Knockdown of Igf2bp2 impairs HSC function

Our first step in describing the role of the MSI2-IGF2BP2 complex was to first elucidate the role of the IGF2BP2 protein in the murine hematopoietic system. The role of MSI2 in murine hematopoiesis has been described extensively[5,6]. It is has been identified as a critical regulator of hematopoietic stem cell self-renewal whose loss results in severe defects in hematopoietic reconstitution and a loss of self-renewal capacity. Little is known about the role of IGF2BP2 in murine hematopoiesis. Correlative data implicates IGF2BP2 as a potential regulator of murine HSCs. As mentioned previously, IGF2BP2 is the most significantly down-regulated protein upon the transition from dormant to cycling murine HSCs[2]. Furthermore, an analysis of 142 microarrays covering over 40 different hematopoietic cell types reveals that the expression of *Igf2bp2* mRNA is restricted to murine HSCs[26]. Other studies indicate that a conserved *Igf2bp2* pathway is a key regulator of stem cell function in numerous organ systems across a variety of model organisms[2,27-29].

Various components of the LIN28-let-7-HMGA2-IGF2BP2 pathway have been broadly implicated in stem cell self-renewal. The LIN28 protein was first identified in *C. elegans* where loss-of-function mutations resulted in the enhanced differentiation of hypodermal and vulval stem cells[27]. Since then, LIN28 has been identified in numerous stem cell populations and is considered to be a common protein that defines "stemness"[30]. Two orthologues of the *C. elegans LIN28* protein are found in vertebrates, *LIN28A* and *LIN28B*[31]. Both are expressed at high levels in mouse embryos and embryonic stem cells and decrease as these cells differentiate[27].

Furthermore, LIN28A, along with OCT4, SOX2, and NANOG, can promote the formation of induced pluripotent stem cells from human fibroblasts[27]. In mouse embryonic stem cells, Lin28 proteins are direct inhibitors of let-7 miRNA biogenesis[27]. Importantly, members of the let-7 miRNA family are potent inhibitors of mESC self-renewal. The transcription factor High-Mobility Group AT-Hook 2 (*Hmga2*) has been identified as a let-7 target; overexpression of let-7 miRNAs results in a potent down-regulation of *Hmga2*[32]. Interestingly, Hmga2 has been shown to bind to a regulatory region within the first intron of *Igf2bp2*, enhancing its transcription[33]. Hmga2 does not transcriptionally up-regulate Igf2bp1 or Igf2bp3[33]. Interestingly, Hmga2 is one of only 22 proteins (including Igf2bp2) that is preferentially up-regulated in murine HSCs and functional studies indicate that the *Lin28*-let-7-*Hmga2* pathway contributes to the enhanced self-renewal of fetal liver HSCs when compared to adult HSCs[29]. Studies indicate that HSCs derived from fetal liver have a much higher capacity for self-renewal when compared to identical populations isolated from adult bone marrow[28,34]. Interestingly, fetal-liver derived LSK and CD150+CD48- HSCs have a much higher expression of *Lin28b*, *Igf2bp2*, and *Hmga2* when compared to identical populations derived from adult cells. The transduction of adult LSK and CD150+CD48- HSCs with *Lin28b* results in a significant down-regulation of let-7a levels, an elevation in *Hmga2* transcript and protein levels, and an increase in *Igf2bp2* transcript levels[29]. Ectopic expression of both *Lin28* and *Hmga2* can enhance the self-renewal of adult bone marrow-derived HSCs. Furthermore, fetal liver HSCs from *Hmga2* knockout mice have a reduced self-renewal capacity when compared to wild-type controls. Interestingly, comparative

affymetrix array analysis between *Hmga2* -/- and wild-type cells identifies *Igf2bp2* as one of seven transcripts (including *Hmga2*) that is downregulated upon a loss of *Hmga2*[29].

To test the impact that a loss of *Igf2bp2* has on murine HSCs, we infected donor CD45.1+CD150+CD48- murine HSCs with lentiviral particles that contained short-hairpins targeting *Igf2bp2*. This population represents an extremely pure collection of murine cells that are able to maintain long-term multipotent engraftment when transplanted into recipient mice[25]. As mentioned previously, this population can be further fractionated into CD34 negative and CD34 positive populations that differ remarkably in their cell cycle status[2]. *Igf2bp2* expression is highest in the dormant CD34-CD150+CD48- HSC population and drops significantly in CD34+CD150+CD48- HSCs. Donor cells were derived from B6.SJL mice and infected with lentiviral particles expressing either a short-hairpin targeting luciferase (shLuc) or a short hairpin targeting *Igf2bp2* (5-*shIgf2bp2*). The 5-sh*Igf2bp2* hairpin was previously validated to knockdown *Igf2bp2* transcript levels 45% in mouse embryonic fibroblasts (MEFs). 1250 D0 equivalent lentiviral-infected CD150+CD48- cells (representing at least 500 functional HSCs) were transplanted into lethally irradiated C57/Bl6 mice along with 100 000 C57BL/6 whole bone marrow competitor cells. B6.SJL mice are congeneic with C57BL/6 mice at the CD45 locus; B6.SJL mice express CD45.1 and C57BL/6 mice express CD45.2 Importantly, CD45 is a pan-hematopoietic marker and the expression of different CD45 molecules allows for the discrimination of donor vs. host cells and thus allows us to specifically analyze the behavior of the short-hairpin infected graft. Importantly,

lentiviral particles that deliver a short-hairpin into infected cells, also deliver a ZsGreen construct that allows for the identification of virally infected cells. When CD45.1+CD150+CD48- cells are incubated with lentiviral particles, this population becomes a mosaic of infected and non-infected cells. By analyzing the ratio of ZsGreen positive to ZsGreen negative cells within the CD45.1 graft, we can determine whether the expression of the hairpin is having any positive or deleterious effect on HSC function. Cells infected with lentiviral particles containing hairpins that have no effect on HSC biology would be expected to behave in an identical manner to non-infected cells[5]. Thus the ratio of ZsGreen positive to ZsGreen negative cells should be maintained when compared to initial transduction levels. However, if a hairpin is deleterious to the function of HSCs, then the repopulation of these HSCs would be impaired and we would expect a relative increase in the percentage of non-infected CD45.1 cells. In this assay, the CD45.2 competitor cells serve to maintain hematopoiesis throughout the initial stages of the transplant. Importantly, the initial repopulation by HSCs is slow since this population does not contain progenitor cells. The competitor bone marrow cells provide the necessary cells required to maintain hematopoiesis immediately after lethal irradiation.

Peripheral blood from C57BL/6 mice was analyzed after 1 month. Remarkably, sh*Igf2bp2* infected HSCs displayed a severe impairment in repopulation compared to shLuc infected HSCs. A significant drop in the proportion of ZsGreen positive cells was noted in grafts derived from sh*Igf2bp2*-infected cells in comparison to those grafts derived from shLuc-infected cells (Figure 8, p.185).

Despite this, no significant difference were seen in the total engraftment of CD45.1+ ZsGreen negative cells suggesting that a loss of Igf2bp2 greatly impairs the repopulation of murine HSCs. We initially hypothesized that a loss of Igf2bp2 in murine HSCs would promote the transition of dormant CD34 negative CD150+CD48- HSCs to cycling CD34 positive CD150+CD48- HSCs. We hypothesized that this would result in the enhanced repopulation of primary mice during the initial stages of transplantation with an eventual decrease in repopulation at later time points that correspond to repopulation from HSCs. This does not appear to be the case. Further analysis of sh*Igf2bp2*-infected cells will be required to fully elucidate the functional consequence that a loss of Igf2bp2 has in the murine HSC compartment. Even if a loss of Igf2bp2 resulted in the rapid differentiation of HSCs into MPPs, we would not expect a rapid decline in engraftment after 1 month since MPPs are capable of short-term engraftment over a period of 4-6 weeks. Based on this preliminary data, it is likely that a loss of Igf2bp2 results in HSC dysfunction that greatly impairs the ability of HSCs to give rise to more differentiated progeny. It is likely that a loss of Igf2bp2 in HSCs impairs cell cycle status, results in apoptosis, or blocks differentiation. Interestingly, a loss of Igf2bp2 in CD150+CD48- murine HSCs significantly impaired the *in vitro* colony forming capabilities of these cells when compared to shLuc-infected controls. A knockdown of Igf2bp2 in murine HSCs allowed for the formation of CFU-E, CFU-M, CFU-GM, CFU-G, and CFU-GEMM colonies albeit at much lower levels compared to controls (Figure 8, p.185). We initially hypothesized that a loss of Igf2bp2 would result in greater HSC cycling and would result in the more rapid generation of progenitor cell populations. Instead we

saw an opposite effect. As mentioned previously, a knockdown of Igf2bp2 may result in a block in HSC differentiation or an increase in apoptosis. This could explain why a loss of Igf2bp2 impairs 1-month repopulation and why a loss of IGF2BP2 impairs HSC-derived colony formation *in vitro*. The mechanisms through which IGF2BP2 functions are likely to be complex and dependent on the presence of other critical regulators and binding partners.

IGF2 signaling in hematopoiesis

Insulin-like growth factor 2 (IGF2) signaling plays a complex role in the regulation of HSCs and IGF2BP2 is likely to play a part in this regulation. Previous studies have identified an imprinted gene network that contributes to the enhanced self-renewal of LT-HSCs through the regulation of Insulin-like growth factor 1 receptor (IGF1R) signaling[28]. A differentially methylated region (DMR) on mouse chromosome 7 located upstream of the non-coding RNA, *H19*, has been identified that controls the mono-allelic expression of *H19* and *Igf2*[35]. The methylation status of the DMR controls the transcriptional insulation of the maternal *Igf2* allele and the transcriptional silencing of the paternal *H19* allele. Interestingly, when inherited paternally, the DMR allows for the expression of *Igf2* but prevents the expression of *H19*. When inherited maternally, the DMR is un-methylated and *Igf2* expression is repressed but *H19* expression is permitted. A deletion in the maternal DMR results in the down-regulation of *H19* and the up regulation of *Igf2*[28]. Interestingly, *H19* is a source of miR-675, a miRNA that is highest in LT-HSCs where it plays an important role in the suppression of the IGF1R[28]. Deletion of the maternal DMR results in the

activation and proliferation of LT-HSCs and their eventual exhaustion[28]. Transplant studies indicate that LT-HSCs with a deletion in the maternal *H19* DMR show greatly impaired engraftment at the 3-month time point[28]. This is thought to be the result of the activation of the IGF2-IGF1R pathway. Binding of IGF2 to the IGF1R results in the activation of this receptor and the eventual phosphorylation of FOXO3, a transcription factor that arrests the cell cycle. Studies indicate that phosphorylation of FOXO3 impairs its function[36]. Deletion of the maternal DMR results in the enhanced phosphorylation of FOXO3 in murine LT-HSCs[28]. Remarkably, the deletion of the maternal DMR results in the expression of pFOXO3 in 75% of LT-HSCs vs. 15% in controls. Interestingly, the ectopic overexpression of miR-675 results in a dramatic increase in Igf1R protein levels in CD34- LSKs and enhances the percentage of quiescent CD34-LSKs[28]. It has thus been postulated that the IGF1R regulates the transition from LT- to ST-HSCs. Notably, the ectopic expression of IGF2 alone does not result in the activation of LT-HSCs. A concomitant increase in IGF1R is required as well. Interestingly, another insulin-like growth factor receptor, IGF2R, is also expressed in CD34-LSK[28]. The IGF family consists of two ligands, IGF1 and IGF2, and two receptors, IGF2R and IGF1R[37]. IGF2 can bind to both IGF1R and IGF2R. Binding of IGF2 to IGF1R results in the phosphorylation of intracellular tyrosine kinase domains and subsequent downstream signaling. The IGF2R does not contain an intracellular kinase domain and the binding of IGF2 to IGF2R does not activate signaling[37]. IGF2R is thought to be a dummy receptor that functions to dampen IGF signaling. The deletion in the maternal *H19* DMR results in a significant up regulation of *Igf1r* in LT-HSCs but does not effect the expression of *Igf2r*[28]. It is

thus possible that a balance between IGF2-IGF2R and IGF2-IGF1R signaling may play a critical role in controlling the activation state of LT-HSCs. Under such a mechanism, the deletion of the maternal DMR would result in preferential signaling through the IGF2-IGF1R pathway resulting in the activation of LT-HSCS. Interestingly, IGF2BP2 is known to bind to the 5'UTR of *Igf2* in order to enhance the expression of *Igf2*. It is possible that IGF2BP2 plays a critical role in priming HSCs for activation. In the process of dormant HSC activation, IGF2BP2 might play a critical role in maintaining IGF2 levels allowing for signaling through IGF1R when the activation of dormant HSCs is required. Importantly, IGF2 is a paracrine growth factor and thus the overexpression of IGF2BP2 might serve to enhance the local expression of IGF2 in the HSC microenvironment. HSCs would thus be surrounded by high levels of IGF2 and signaling through the IGF1R could occur rapidly upon the up-regulation of this receptor as CD34 negative dormant HSC populations transition to CD34 positive cycling HSCs.

To further analyze the effects of *Igf2bp2* knockdown in HSCs, the ratio of ZsGreen positive to ZsGreen negative cells within the CD45.1 donor population will be monitored in the peripheral blood up to 4-months post-transplant. At this point the mice will be sacrificed and the bone marrow harvested for a more in-depth analysis of the remaining graft. The analysis of cell cycle status, apoptosis, quiescence, and the frequency of progenitor and stem cell populations will allow for a more in-depth understanding of the role of Igf2bp2 in murine hematopoiesis.

To further elucidate the role of IGF2BP2 in murine hematopoietic stem cell function, we have generated separate lentiviral particles that overexpress *Igf2bp2*

transcript variant #1 and *Igf2bp2* transcript variant #2 in frame with a P2A-ZsGreen construct. Upon translation of the Igf2bp2-P2A-ZsGreen transcript, the P2A sequence causes the ribosome to skip an amide bond resulting in the formation of equal levels of the two independent proteins. Thus ZsGreen can be used as a reporter of *Igf2bp2* levels in infected cells. Future experiments will involve the infection of CD45.1+CD150+CD48- cells with lentiviral particles that overexpress *Igf2bp2* TV#1 or *Igf2bp2* TV#2. The analysis of these cells through a variety of *in vitro* and *in vivo* assays will aid in the understanding of IGF2BP2 function in hematopoiesis.

Overall, this work describes a novel interaction between the MSI2 and IGF2BP2 proteins. Given the unique expression pattern of both proteins in hematopoietic stem and progenitor cell populations along with functional data implicating the two in hematopoietic stem cell self-renewal, we believe that MSI2-IGF2BP2 protein complexes may play a critical role in the regulation of HSC function. We hypothesize that MSI2 and IGF2BP2 are able to form a protein complex that is able to post-transcriptionally regulate a network of mRNAs that are critical for proper HSC function. Importantly, the differential expression of *Igf2bp2* between dormant and cycling LT-HSCs may result in the altered regulation of critical mRNAs between these two populations. The identification of IGF2BP2-MSI2 regulated mRNAs is likely to reveal critical regulators of HSC quiescence.

References

1 Kim, Y. C. *et al.* The transcriptome of human CD34+ hematopoietic stem-progenitor cells. *Proc Natl Acad Sci U S A* **106**, 8278-8283, doi:10.1073/pnas.0903390106 (2009).

2 Cabezas-Wallscheid, N. *et al.* Identification of regulatory networks in HSCs and their immediate progeny via integrated proteome, transcriptome, and DNA methylome analysis. *Cell Stem Cell* **15**, 507-522, doi:10.1016/j.stem.2014.07.005 (2014).
3 Klimmeck, D. *et al.* Transcriptome-wide profiling and posttranscriptional analysis of hematopoietic stem/progenitor cell differentiation toward myeloid commitment. *Stem Cell Reports* **3**, 858-875, doi:10.1016/j.stemcr.2014.08.012 (2014).
4 Klimmeck, D. *et al.* Proteomic cornerstones of hematopoietic stem cell differentiation: distinct signatures of multipotent progenitors and myeloid committed cells. *Mol Cell Proteomics* **11**, 286-302, doi:10.1074/mcp.M111.016790 (2012).
5 Hope, K. J. *et al.* An RNAi screen identifies Msi2 and Prox1 as having opposite roles in the regulation of hematopoietic stem cell activity. *Cell Stem Cell* **7**, 101-113, doi:10.1016/j.stem.2010.06.007 (2010).
6 Rentas, S. *et al.* Musashi-2 attenuates AHR signalling to expand human haematopoietic stem cells. *Nature* **532**, 508-511, doi:10.1038/nature17665 (2016).
7 Lunde, B. M., Moore, C. & Varani, G. RNA-binding proteins: modular design for efficient function. *Nat Rev Mol Cell Biol* **8**, 479-490, doi:10.1038/nrm2178 (2007).
8 Glisovic, T., Bachorik, J. L., Yong, J. & Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett* **582**, 1977-1986, doi:10.1016/j.febslet.2008.03.004 (2008).
9 Wilson, A. *et al.* Hematopoietic stem cells reversibly switch from dormancy to self-renewal during homeostasis and repair. *Cell* **135**, 1118-1129, doi:10.1016/j.cell.2008.10.048 (2008).
10 Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* **8**, 2281-2308, doi:10.1038/nprot.2013.143 (2013).
11 Fellmann, C. *et al.* An optimized microRNA backbone for effective single-copy RNAi. *Cell Rep* **5**, 1704-1713, doi:10.1016/j.celrep.2013.11.020 (2013).
12 Kampmann, M. *et al.* Next-generation libraries for robust RNA interference-based genome-wide screens. *Proc Natl Acad Sci U S A* **112**, E3384-3391, doi:10.1073/pnas.1508821112 (2015).
13 Roux, K. J., Kim, D. I. & Burke, B. BioID: a screen for protein-protein interactions. *Curr Protoc Protein Sci* **74**, Unit 19 23, doi:10.1002/0471140864.ps1923s74 (2013).
14 Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat Rev Genet* **15**, 829-845, doi:10.1038/nrg3813 (2014).
15 Bell, J. L. *et al.* Insulin-like growth factor 2 mRNA-binding proteins (IGF2BPs): post-transcriptional drivers of cancer progression? *Cell Mol Life Sci* **70**, 2657-2675, doi:10.1007/s00018-012-1186-z (2013).
16 Hsu, K. F. *et al.* Overexpression of the RNA-binding proteins Lin28B and IGF2BP3 (IMP3) is associated with chemoresistance and poor disease

outcome in ovarian cancer. *Br J Cancer* **113**, 414-424, doi:10.1038/bjc.2015.254 (2015).
17 Lu, D. *et al.* IMP3, a new biomarker to predict progression of cervical intraepithelial neoplasia into invasive cancer. *Am J Surg Pathol* **35**, 1638-1645, doi:10.1097/PAS.0b013e31823272d4 (2011).
18 Gao, Y. *et al.* IMP3 expression is associated with poor outcome and epigenetic deregulation in intrahepatic cholangiocarcinoma. *Hum Pathol* **45**, 1184-1191, doi:10.1016/j.humpath.2014.01.016 (2014).
19 Park, J. Y., Choe, M., Kang, Y. & Lee, S. S. IMP3, a Promising Prognostic Marker in Clear Cell Renal Cell Carcinoma. *Korean J Pathol* **48**, 108-116, doi:10.4132/KoreanJPathol.2014.48.2.108 (2014).
20 Hamilton, K. E. *et al.* IMP1 promotes tumor growth, dissemination and a tumor-initiating cell phenotype in colorectal cancer cell xenografts. *Carcinogenesis* **34**, 2647-2654, doi:10.1093/carcin/bgt217 (2013).
21 Kobel, M. *et al.* Expression of the RNA-binding protein IMP1 correlates with poor prognosis in ovarian carcinoma. *Oncogene* **26**, 7584-7589, doi:10.1038/sj.onc.1210563 (2007).
22 Stoskus, M., Vaitkeviciene, G., Eidukaite, A. & Griskevicius, L. ETV6/RUNX1 transcript is a target of RNA-binding protein IGF2BP1 in t(12;21)(p13;q22)-positive acute lymphoblastic leukemia. *Blood Cells Mol Dis* **57**, 30-34, doi:10.1016/j.bcmd.2015.11.006 (2016).
23 Boudoukha, S., Cuvellier, S. & Polesskaya, A. Role of the RNA-binding protein IMP-2 in muscle cell motility. *Mol Cell Biol* **30**, 5710-5725, doi:10.1128/MCB.00665-10 (2010).
24 Dai, N. *et al.* IGF2BP2/IMP2-Deficient mice resist obesity through enhanced translation of Ucp1 mRNA and Other mRNAs encoding mitochondrial proteins. *Cell Metab* **21**, 609-621, doi:10.1016/j.cmet.2015.03.006 (2015).
25 Oguro, H., Ding, L. & Morrison, S. J. SLAM family markers resolve functionally distinct subpopulations of hematopoietic stem cells and multipotent progenitors. *Cell Stem Cell* **13**, 102-116, doi:10.1016/j.stem.2013.05.014 (2013).
26 Riddell, J. *et al.* Reprogramming committed murine blood cells to induced hematopoietic stem cells with defined factors. *Cell* **157**, 549-564, doi:10.1016/j.cell.2014.04.006 (2014).
27 Shyh-Chang, N. & Daley, G. Q. Lin28: primal regulator of growth and metabolism in stem cells. *Cell Stem Cell* **12**, 395-406, doi:10.1016/j.stem.2013.03.005 (2013).
28 Venkatraman, A. *et al.* Maternal imprinting at the H19-Igf2 locus maintains adult haematopoietic stem cell quiescence. *Nature* **500**, 345-349, doi:10.1038/nature12303 (2013).
29 Copley, M. R. *et al.* The Lin28b-let-7-Hmga2 axis determines the higher self-renewal potential of fetal haematopoietic stem cells. *Nat Cell Biol* **15**, 916-925, doi:10.1038/ncb2783 (2013).

30 Liu, Y. *et al.* Lin28 induces epithelial-to-mesenchymal transition and stemness via downregulation of let-7a in breast cancer cells. *PLoS One* **8**, e83083, doi:10.1371/journal.pone.0083083 (2013).
31 Stefani, G., Chen, X., Zhao, H. & Slack, F. J. A novel mechanism of LIN-28 regulation of let-7 microRNA expression revealed by in vivo HITS-CLIP in C. elegans. *RNA* **21**, 985-996, doi:10.1261/rna.045542.114 (2015).
32 Lee, Y. S. & Dutta, A. The tumor suppressor microRNA let-7 represses the HMGA2 oncogene. *Genes Dev* **21**, 1025-1030, doi:10.1101/gad.1540407 (2007).
33 Li, Z. *et al.* An HMGA2-IGF2BP2 axis regulates myoblast proliferation and myogenesis. *Dev Cell* **23**, 1176-1188, doi:10.1016/j.devcel.2012.10.019 (2012).
34 Kunimoto, H. *et al.* Tet2 disruption leads to enhanced self-renewal and altered differentiation of fetal liver hematopoietic stem cells. *Sci Rep* **2**, 273, doi:10.1038/srep00273 (2012).
35 Park, K. Y., Sellars, E. A., Grinberg, A., Huang, S. P. & Pfeifer, K. The H19 differentially methylated region marks the parental origin of a heterologous locus without gametic DNA methylation. *Mol Cell Biol* **24**, 3588-3595 (2004).
36 Boccitto, M. & Kalb, R. G. Regulation of Foxo-dependent transcription by post-translational modifications. *Curr Drug Targets* **12**, 1303-1310 (2011).
37 Boone, D. N. & Lee, A. V. Targeting the insulin-like growth factor receptor: developing biomarkers from gene expression profiling. *Crit Rev Oncog* **17**, 161-173 (2012).

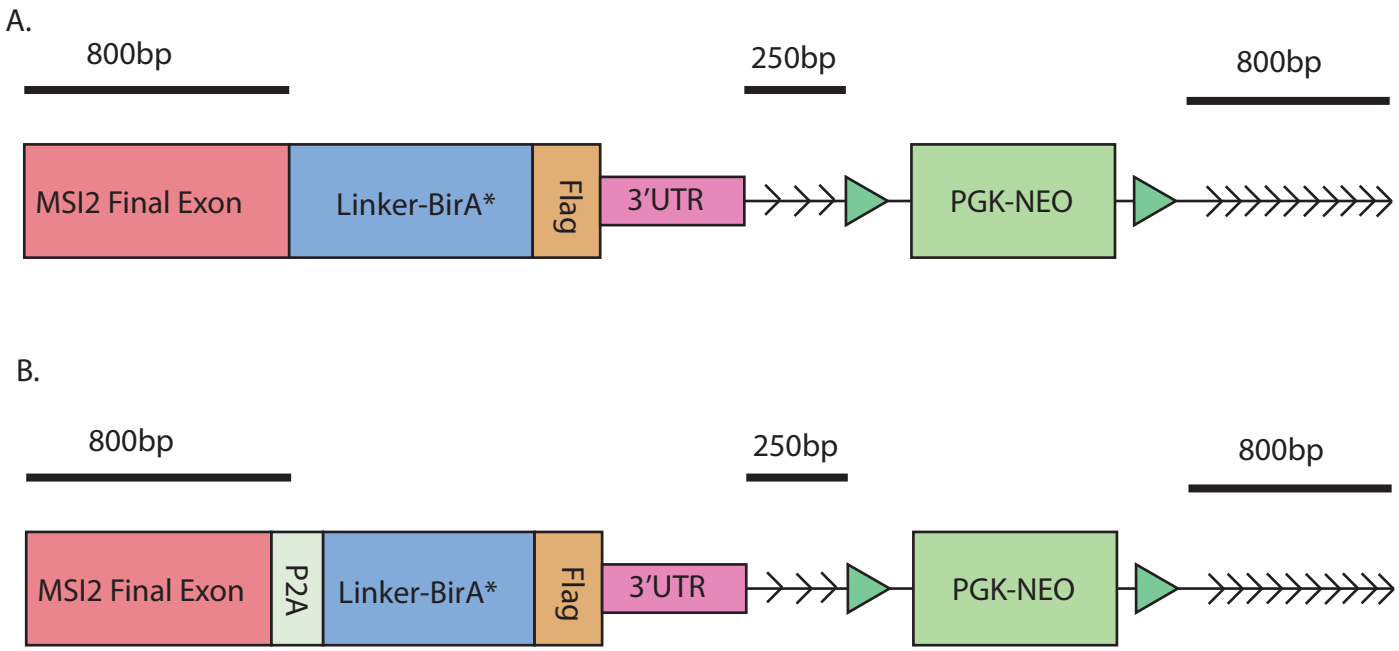


Figure 1. Structure of the MSI2-BirA* repair templates. (A) The final 800bp sequence upstream of the MSI2 stop codon were cloned in frame with a linker-BirA*-3X-Flag sequence. Importantly, the MSI2 stop codon was removed and a stop codon was placed immedately downstream of the flag sequence. The MSI2 3'UTR and the adjacent 250 bp genomic region was cloned immediately downstream of the Flag sequence followed by a lox-P (green triangles) flanked PGK-Neo cassette. An additional 800bp genomic sequence located immediately downstream from the previously cloned 3'UTR+250bp sequence was inserted downstream from the PGK-Neo Cassette. (B) The 'P2A' control repair template is identical to the experimental template with the addition of a P2A sequence located in frame between the MSI2 final exon and the linker-BirA*-3X-Flag sequence.

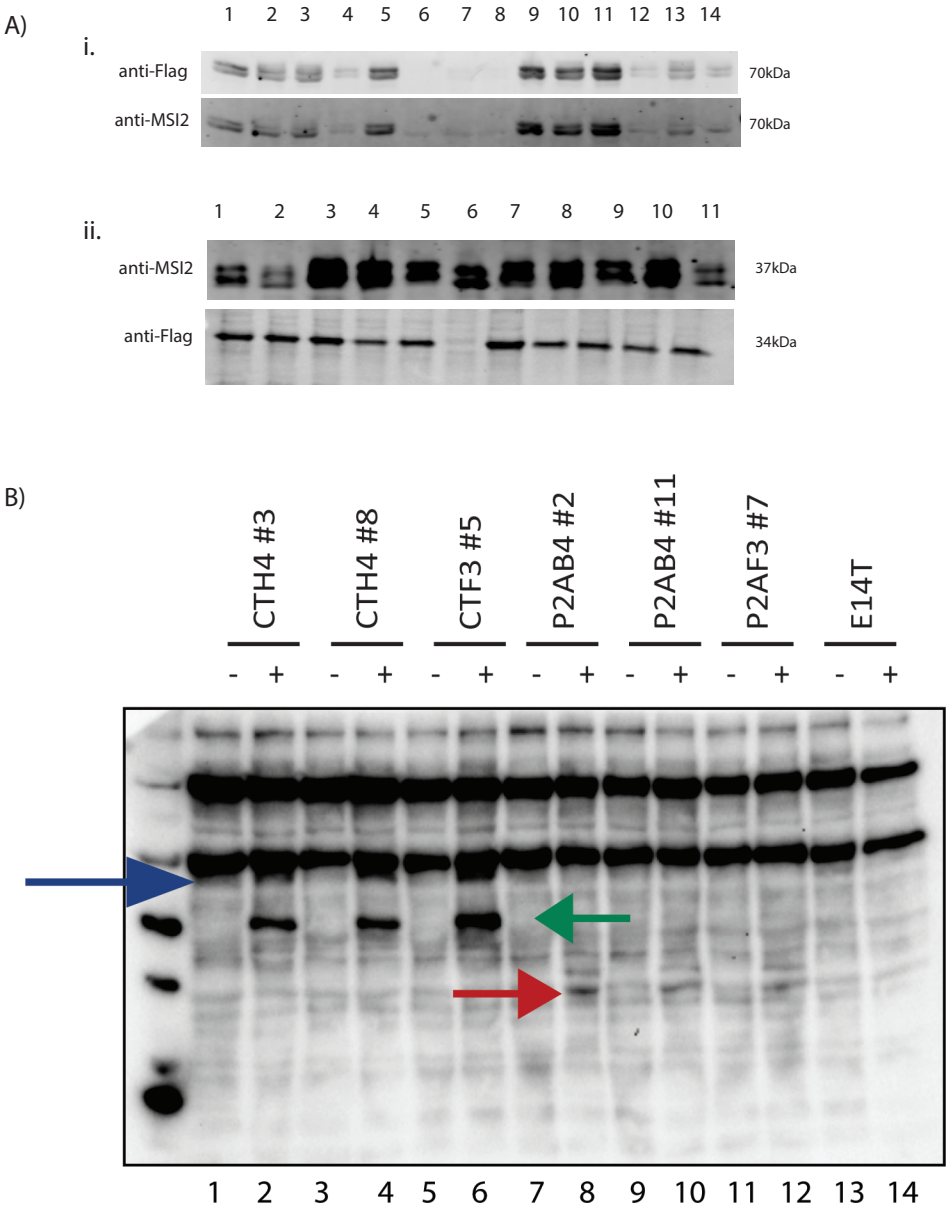


Figure 2. Validation of MSI2-BirA* clones. (A) Numerous MSI2-BirA* and MSI2-P2A-BirA* clones were subject to PCR screening.14 MSI2-BirA* and 11 MSI2-P2A-BirA* clones that were positive for PCR screens were analyzed via western blotting. (i) Western blotting of MSI2-BirA* clones identifies a 70kDa protein that is detected by anti-flag and anti-MSI2. (ii) Western blotting of MSI2-P2A-BirA* clones identifies MSI2 at 37kDa and a 34kDa band that is detected by anti-Flag (P2A-BirA*-3X-Flag). (B) Three MSI2-BirA* (CTH4 #3, CTH4 #8, and CTF3 #5) and three MSI2-P2A-BirA* (P2AB4 #5, P2AB4#11, and P2AF3 #7) cell lines were incubated with (+) or without (-) 50uM biotin. E14T wild-type cells were treated as well. Western blotting was performed and the blot was developed using a streptavidin-HRP. Specific patterns of biotinylation were detected only in MSI2-BirA* cell lines treated with 50uM biotin (red and blue arrows). A biotinyalted protein was also specifically detected only in P2A-BirA* cells lines treated with 50uM biotin(red arrow). This band corresponds to the molecular weight of the free P2A-BirA* protein and likely represents autobiotinylation of this product.

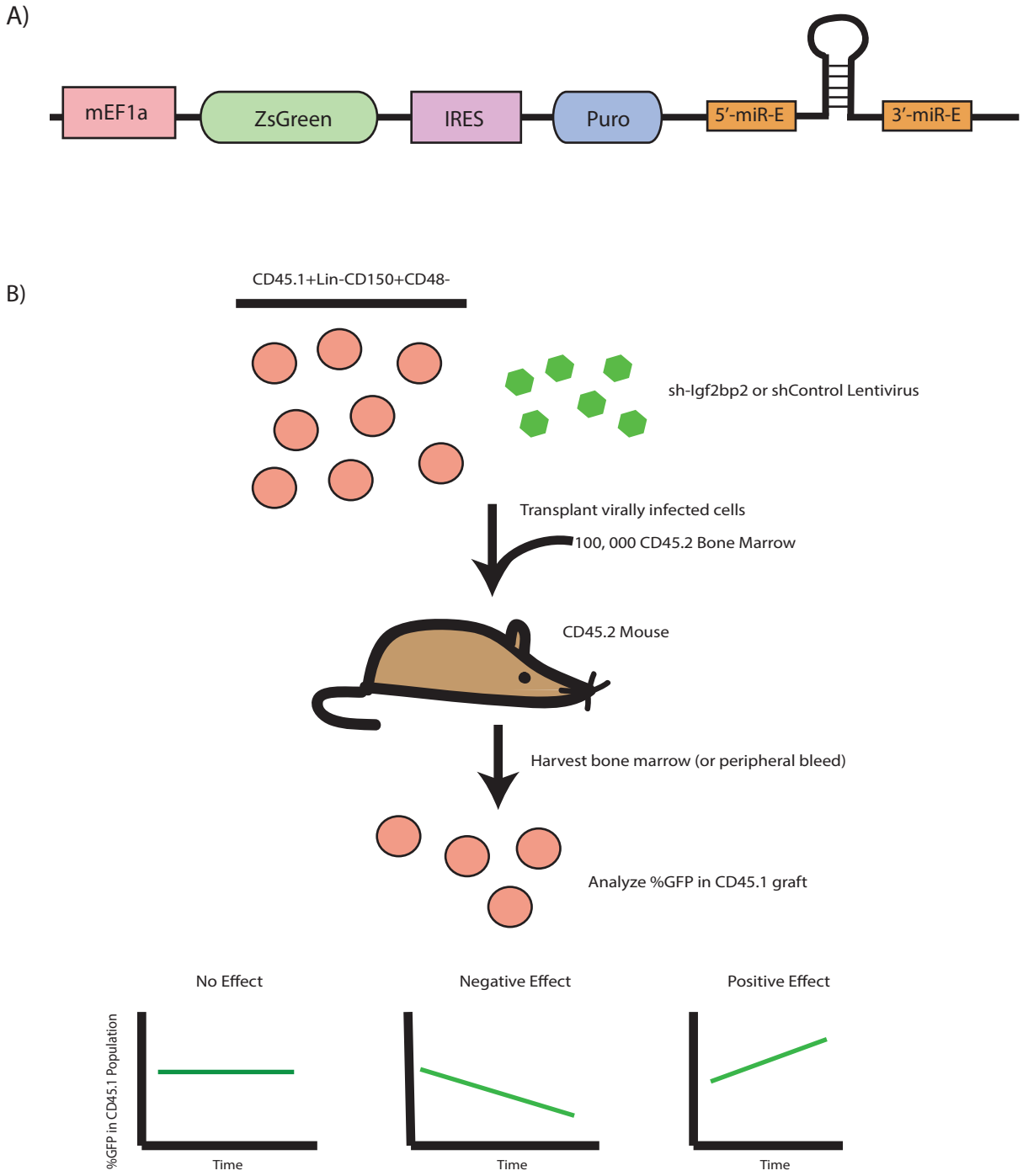


Figure 3. shIGF2BP2 vector design and outline of the HSC transplantation method. (A) The mEF1a promoter drove the transcription of a ZsGreen-IRES-Puro-miRE-Igf2BP2 transcript. The shIGF2BP2 was embedded within a miR-E backbone. (B) CD45.1 Lin-CD150+CD48- HSCs were infected with shIGF2BP2 or control lentiviral particles. These cells were transplanted with 100, 000 CD45.2 bone marrow helper cells to maintain the initial stages of engraftment into lethally iradiated mice. GFP levels in the CD45.1 graft will be analyzed at 4, 8, 12, and 16 weeks, to determine the effect of the shIGF2BP2 hairpin. After 16 weeks, the bone marrow will be harvested and further analyzed to determine the impact that a loss of IGF2BP2 has on hematopoetic reconstitution.

	Unique Peptide Counts			
Proteins	MSI2-BirA #1	MSI2-BirA #2	P2A #1	P2A #2
IGF2BP1	29	30	1	4
IGF2BP2	25	27	0	0
FUBP3	17	16	0	0
PRRC2C	11	8	0	0
EF1A1	11	11	1	1
MSI2	8	9	0	1
ATXN2L	6	10	0	0
PABP1	4	4	0	0

Figure 4. BioID identifies proximally interacting proteins. BioID analysis was performed in two MSI2-BirA* cell lines and two MSI2-P2A-BirA* cell lines. Mass spectrometry analsysis revealed 8 proteins that are preferentially biotinyalted in MSI2-BirA* but not MSI2-P2A-BirA* cell lines

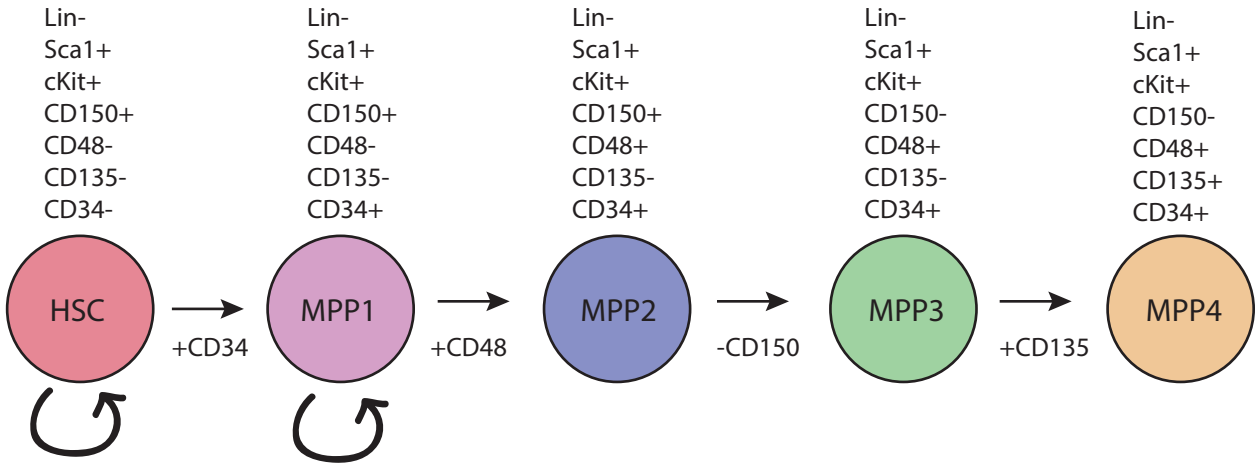


Figure 5. Schematic of murine HSC and MPP populations. LT-HSCs (HSC) are characterized as LSKCD150+CD48-CD135-CD34-. These cells show robust multilineage engraftment when transplanted in primary and secondary mice. The expression of CD34+ is associated with ST-HSCs (MPP1). These cells show robust multilineage reconstitution in primary mice (>3 months) but impaired repopulation in secondary mice. MPP2, MPP3, and MPP4 represent multipotent progenitor populations. These cells give rise to short-term (4-6 week) multilineage engraftment when transplanted. MPP3 cells show a myeloid-biased reconstitution while MPP4 cells show a lymphoid biased reconstitution.

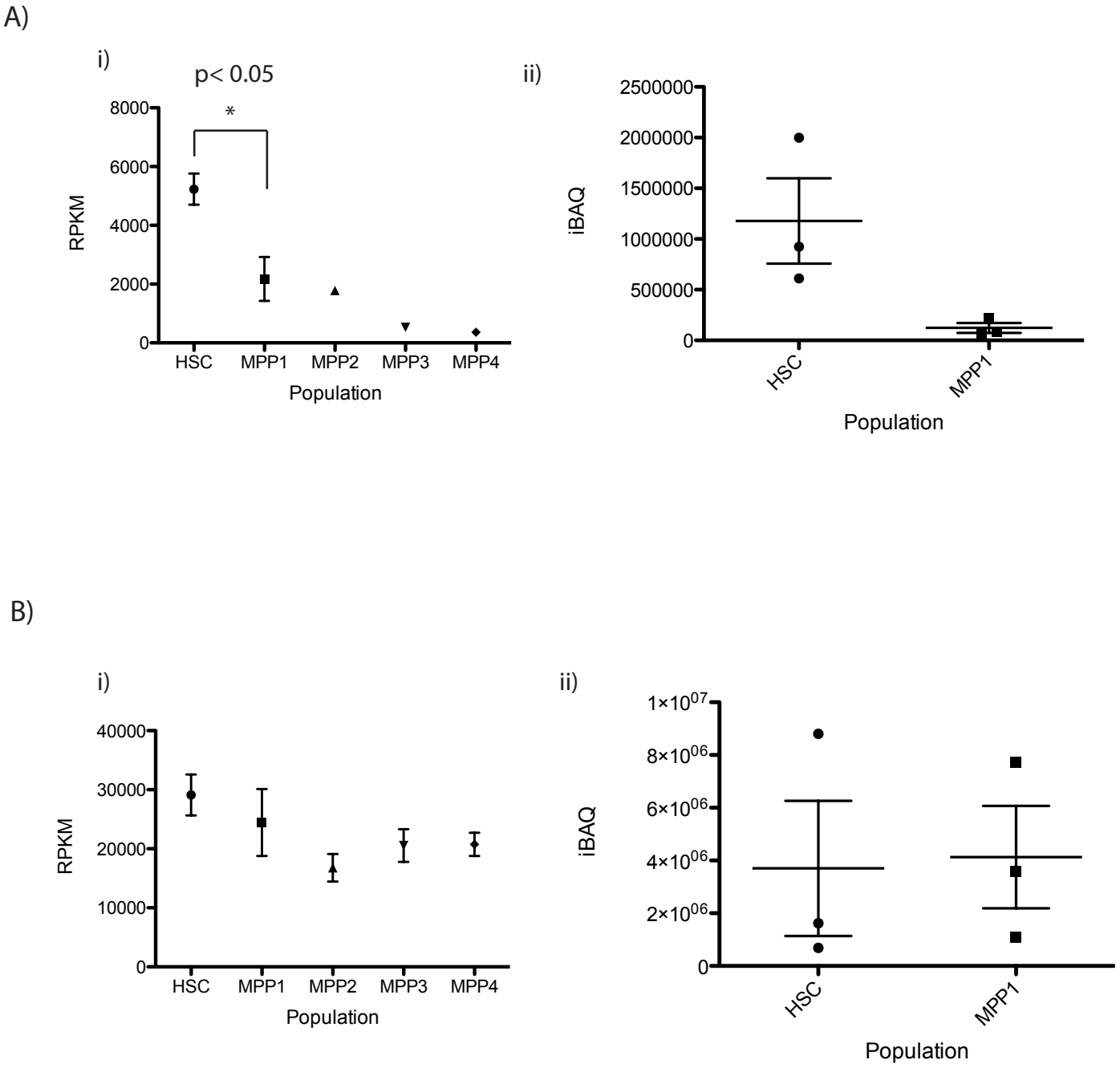


Figure 6. Levels of Msi2 and Igf2bp2 in mouse hematopoietic stem cells and multipotent progenitors. (A) (i) RNA-Seq identifies IGF2BP2 transcript levels across HSCs and more commited multipotent progenitor populations. (ii) Mass spectrometry identifies IGF2BP2 protein levels in LT-HSCs (HSC) and ST-HSCs (MPP1). (B) (i) RNA-Seq identifies MSI2 transcript levels across HSCs and more committed multipotent progenitor populations. (ii) Mass spectrometry identifies MSI2 protein levels in LT-HSCs (HSC) and ST-HSCs(MPP1).

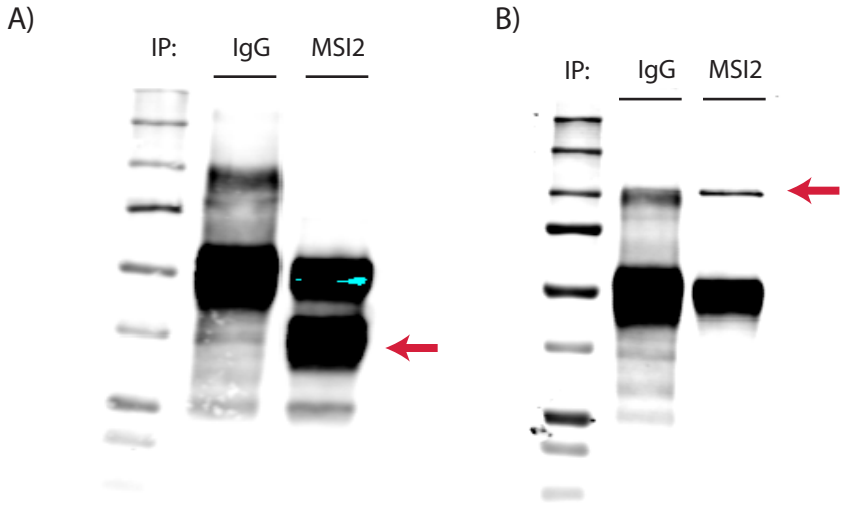
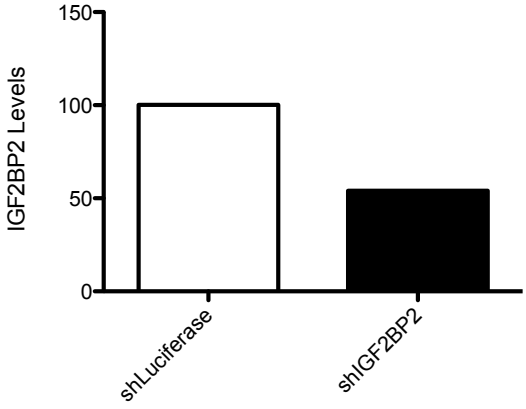


Figure 7. Co-IP of Msi2 in mESCs identifies Igf2bp2 as a direct Msi2-interacting protein. Co-immunoprecipitaiton was performed in Igf2BP2-BirA* cell lines using the gentle-soft lysis buffer. 3 million mESCs were lysed per pulldown with either a control rabbit IgG or a rabbit-anti-Msi2. Lysate-antibody complexes were incubated for 3 hours and isolated *via* incubation with protein G dynabeads. Proteins were eluted in Laemmli sample buffer and analyzed *via* western blotting. Lysates were probed with anti-Msi2 (A) and revealed a significant enrichment of Msi2 (red arrow) in the Msi2 pulldown vs control. Lysates were probed with anti-Igf2bp2 (B) and revealed the presence of the Igf2bp2-BirA* fusion protein (red arrow) in the Msi2 pulldown lysate that was absent from the control pulldown

A)



B)

Plate	BFU-E	CFU-G	CFU-M	CFU-GM	CFU-GEMM	Total
shLuc #1	18	52	17	21	1	109
shLuc #2	24	72	19	28	2	145
shIgf2bp2 #1	20	10	21	17	0	68
shIgf2bp2 #2	22	18	33	12	0	85

*

p< 0.05

C)

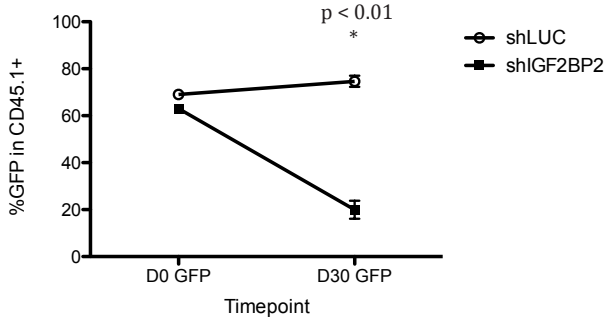


Figure 8. Igf2bp2 knockdown and its impact on CFU formation and 1-month repopualtion. (A) sh*Igf2bp2* knockdown was tested in murine embryonic fibroblasts (MEFs). Infection of cells with sh*Igf2bp2* lentivirus results in a 45% knockdown of Igf2bp2 levels. (B) CD150+CD48-Lin- cells were infected with sh*Igf2bp2* lentivirus or control and plated in duplicate for CFU-assays. sh*Igf2bp2*-infected cells produced significantly less CFU-G colonies after 10-days in methylcellulose culture when comapred to shLuc infected cells. An unpaired, two-tailed t-test was performed (C) Lin-CD150+CD48- mouse HSCs were infected with a lentivirus expressing sh*Igf2bp2* or shLuc and a ZsGreen marker. Cells were transplanted into lethally irradiated mice. Peripheral blood analysis 1-month post-transplant shows a significant drop off in GFP in the sh*Igf2bp2*-infected graph. GFP are maintained in the shLuc infected graft. An unpaired, two-tailed t-test was performed .

Chapter 5: Concluding Remarks

The studies outlined in this PhD thesis have identified, in an unbiased manner, the biochemical mechanisms through which MSI2 functions. Notably, this body of work has identified numerous proteins that are likely to play critical roles in controlling the balance between hematopoietic stem cell self-renewal *vs* differentiation. A critical analysis of these targets will likely reveal novel pathways and regulators of hematopoietic biology.

Importantly, MSI2 is a potent regulator of normal hematopoietic- and leukemic- stem cell function[1,2]. Our previous studies indicate that an overexpression of MSI2 in CD34+ cord blood cells results in a 17-fold increase in short-term repopulating cells and a 23-fold *ex vivo* expansion of long-term repopulating cells[2]. The *ex vivo* expansion of short-term and long-term repopulating cells in cord blood samples is clinically relevant due to their use in hematopoietic stem cell transplants (HSCT). Umbilical cord blood-derived hematopoietic cells are an important source for HSCT. Cord blood-derived cells are used in HSCT in pediatric patients where they are just as good as bone marrow derived cells[3]. Furthermore, the lower immunogenicity of cord blood samples results in much less severe graft-vs-host disease despite a maintenance in graft-vs-leukemia effects[4]. Also, cord blood transplants can be performed between unrelated donors with one or two HLA mismatches[4]. In adult patients however, the use of umbilical cord blood-derived HSCs typically results in delayed engraftment due to the low numbers of hematopoietic stem cells found in cord blood. On a per volume basis, cord blood

has a higher number of primitive progenitors and long-term repopulating cells compared to bone marrow derived samples. However, due to the small sample sizes, the absolute number of hematopoietic cells in a cord blood sample is much smaller[3]. Importantly, a critical determinant of successful HSCT is the absolute number of nucleated cells in the transplanted sample. Resultantly, CB transplants usually result in a significant delay in engraftment in adults (>60kg) and are more successful when limited to pediatric patients[3]. Cord blood numbers can be increased by transplanting multiple cord blood samples or by transplanting *ex vivo* expanded cells. The goal is to increase the number of HSCs as well as committed progenitors; this will allow for a rapid transient reconstitution of the bone marrow as well as a robust long-term engraftment. A clinically relevant expansion of cord blood-derived HSPCs will greatly improve their use in adult HSCT.

Our studies indicate that the MSI2 protein plays a critical role in the *ex vivo* expansion of cord-blood derived hematopoietic stem and progenitor cells (HSPCs)[2]. By better understanding the biochemical mechanisms through which MSI2 functions, we can better understand the processes that control HSPC expansion. Our previous study used lentiviral-mediated overexpression of MSI2 in order to produce cord blood expansion. This is not a clinically applicable technique due to the random nature of lentiviral integration and its associated insertional mutagenesis. Instead, a clinically relevant technique for the *ex vivo* expansion of cord-blood derived HSPCs would involve the treatment of these cells with a drug cocktail that can be removed prior to transplantation. Though our previous studies focus on the regulation of Cyp1b1 by MSI2, treatment of cord blood-derived HSPCs with the Cyp1b1 inhibitor

(E)-2,3',4,5'-tetramethoxystilbene (TMS) only results in a 1.5-fold expansion of CD34+ cells. In contrast, other studies targeting the Ahr signaling pathway have demonstrated that incubation of CD34+ HSPCs with the Ahr antagonist StemRegenin1 (SR1) and cytokines results in a 17-fold increase in immune-deficient mouse repopulating cells[5].

Notably, in our ranked list of MSI2 RNA targets, there are 27 additional targets other than Cyp1b1 that are more significantly bound by MSI2. Notably, the vast majority of these genes have no known role in hematopoiesis and it is very possible that these genes represent novel regulators of hematopoiesis. The analysis of hematopoietic gene expression databases reveals that many of these targets show distinct patterns of expression across the hematopoietic hierarchy where they are either highly enriched in HSPCs or severely downregulated[6,7]. One interesting experiment to perform (perhaps by a future graduate student) would be to overexpress those MSI2 targets whose protein is expressed at low levels in HSPCs and to knockdown those MSI2 targets whose protein is overexpressed in HSPCs. Such a screen in human cord blood samples will likely identify novel genes that play critical roles in the maintenance of HSPCs.

Furthermore by specifically altering the levels of MSI2 target RNA, we may be able to generate a more robust HSPC expansion. MSI2 is thought to regulate the post-transcriptional expression of target mRNAs. Importantly, many studies indicate that RNA-binding proteins (RBPs) often function in complexes with other RBPs and non-RBPs. The composition of RBP complexes can have a great affect on the affinity of RNA binding and the fate of bound mRNAs[8,9]. For example, if the

regulation of a target RNA by MSI2 is dependent on an interaction with 'Protein X', then we can imagine that by simply overexpressing MSI2, the effect on its target RNA will be much more attenuated compared to an overexpression of MSI2 and 'Protein X'. We can overcome this effect by directly altering levels of the downstream RNA target instead. Furthermore, to elucidate those proteins that are responsible for the MSI2 phenotype, I believe that an interesting experiment would be to overexpress MSI2 and subject the sample to analysis by mass spectrometry. By identifying proteins that are significantly altered upon changes in MSI2 expression and are known MSI2 targets, we can put together a composite picture of the impact that MSI2 is having on its RNA targets and we can begin to answer some very important questions: Does MSI2 enhance the expression of target mRNAs? Does MSI2 attenuate the expression of target mRNAs? Are the protein levels of some targets increased while others decrease? By examining the differences in protein expression following changes in MSI2, perhaps we can start to put together an idea of combinatorial changes that occur in order to enhance HSPC function. As mentioned previously, RBPs are thought to co-ordinate 'RNA regulons'. These 'RNA regulons' consist of functionally related mRNAs whose expression is co-ordinated post-transcriptionally by RBPs[10,11]. Perhaps MSI2 binds to a variety of mRNAs that play a critical role in the maintenance of HSPCs and their coordinated regulation can dramatically enhance HSPC function. Importantly, MSI2 may act on some of these mRNA targets to enhance their expression and it may act on other targets to attenuate their expression. This can be accomplished through the formation of unique protein complexes between MSI2 and other protein interactors.

The downstream analysis of MSI2 protein binding partners will further help in the elucidation of MSI2 function and mechanisms that are active in HSPCs. MSI2 is expressed in both HSCs and more committed progenitors. Differences in MSI2 protein binding partners may be responsible for its unique role in these different compartments. Our BioID analysis of MSI2 identified the IGF2BP2 protein as a potent MSI2 binding partner. Importantly, IGF2BP2 is preferentially expressed in the most dormant population of HSCs. I believe that the further analysis of the relationship between MSI2 and IGF2BP2 will elucidate novel mechanisms responsible for the maintenance of HSC quiescence and the transition from dormant to actively cycling HSCs. I hypothesize that MSI2 and IGF2BP2 form a complex that regulates the post-transcriptional expression of a distinct set of mRNAs that play a critical role in the proper function of HSCs. Specifically, I believe that our lab should focus some resources studying the regulation of IGF2 signaling by MSI2. As mentioned previously, the up-regulation of IGF2 and IGF1R has been shown to impair HSC quiescence *via* the activation of IGF1R signaling upon IGF2 binding[12]. Importantly, studies indicate that IGF2BP2 can bind to the 5'UTR of IGF2 in order to enhance its translation[13]. I would like to see someone investigate whether MSI2 can bind to the 3'UTR of IGF2 and interact with IGF2BP2 in order to circularize the IGF2 mRNA. Perhaps the strong affinity of IGF2BP2 for MSI2 will occlude the interaction between MSI2 and poly(A) binding protein (PABP). This circularization reaction could further serve to bring the PABP in close proximity to eIF4G. As mentioned previously, the interaction between PABP and eIF4G is critical for the initiation of translation[14]. This could be a possible mechanism through which MSI2 and IGF2BP2

enhance IGF2 expression. Notably, IGF2 is a secreted cytokine that can act in a paracrine manner. I would like to investigate whether the paracrine secretion of IGF2 by LT-HSCs primes these cells for activation. I imagine a situation where LT-HSCs reside in a niche surrounded by high levels of IGF2. The expression of IGF2R on LT-HSCs attenuates IGF2 signaling and studies indicate a down-regulation of IGF2R upon the transition to cycling HSC[6]. When the activation of HSCs is required, one of the first steps may be the down-regulation of IGF2R and a relative increase in the IGF1R. The abundance of IGF2 in the stem cell niche would result in enhanced IGF1R signaling and the activation of dormant HSCs.

To identify novel pathways that are active in the maintenance of AML, it will be interesting to examine the expression of MSI2 RNA targets in various AML samples. Though MSI2 correlates with CD34+ expression in human AML and patient data reveals a correlation with MSI2 and aggressive leukemias, not all AML samples express MSI2. Despite this, it is possible that pathways through which MSI2 acts are still active. If we think of MSI2 as a repressor of translation, it is possible to imagine a situation where mRNA targets that are typically repressed by MSI2 are repressed through other means (i.e. promoter methylation, down regulation of signaling protein, etc.). In an opposite manner, we can imagine a situation where mRNA targets that are typically up regulated by MSI2 have their expression enhanced through other mechanisms. Thus I think it is of the utmost important to examine the expression of MSI2 CLIP targets across a large number of AML samples in order to see whether the expression of any targets are conserved.

My initial studies with AML taught me that these samples are quite difficult to work with. Notably, when using lentiviral-transduction assays, it is often impossible to efficiently transduce AML samples, especially when a large amount of cells are required to appropriately engraft. For future experiments, it would be ideal to select AML samples that have a large leukemic stem cell (LSC) fraction and can engraft robustly with a relatively small number of cells. This would allow for the transduction of these cells at a high MOI, which could possibly help in the maintenance of transgene expression throughout the duration of the xenograft. It is interesting to note that the one AML sample that did not reveal any transgene silencing was the one sample that had the highest level of transduction (>95%). At such high transduction efficiency, it is likely that most cells received numerous copies of the transgene cassette. This may have contributed to the maintenance of GFP expression in these samples. To further prevent the issues with transgene silencing, I would consider switching lentivectors for future experiments. Notably, the pLB vector (Addgene #11619) is a lentivector that drives GFP expression off a CMV promoter and a short-hairpin off a U6 promoter. Importantly, genetic elements known to suppress epigenetic silencing have been incorporated upstream and downstream of the expression cassettes. This may help to maintain the expression of GFP and our hairpin throughout the duration of the transplant.

Overall, I hope that this body of work motivates future researchers to continue studying the biochemical pathways through which MSI2 functions. Studying the MSI2 protein was filled with many hardships. The techniques I used were complex and difficult to comprehend. They required months of

standardization and much trial and error before they yielded any useful results. Despite all of this, it was well worth the effort. I am happy to have contributed this knowledge to the scientific community. I hope that the elucidation of the biochemical pathways involved in hematopoietic stem cell biology will one day result in the development of novel therapeutics that will help alleviate the burden of human disease.

References

1 Kharas, M. G. *et al.* Musashi-2 regulates normal hematopoiesis and promotes aggressive myeloid leukemia. *Nat Med* **16**, 903-908, doi:10.1038/nm.2187 (2010).

2 Rentas, S. *et al.* Musashi-2 attenuates AHR signalling to expand human haematopoietic stem cells. *Nature* **532**, 508-511, doi:10.1038/nature17665 (2016).

3 Flores-Guzman, P., Fernandez-Sanchez, V. & Mayani, H. Concise review: ex vivo expansion of cord blood-derived hematopoietic stem and progenitor cells: basic principles, experimental approaches, and impact in regenerative medicine. *Stem Cells Transl Med* **2**, 830-838, doi:10.5966/sctm.2013-0071 (2013).

4 Kurtzberg, J. Update on umbilical cord blood transplantation. *Curr Opin Pediatr* **21**, 22-29 (2009).

5 Boitano, A. E. *et al.* Aryl hydrocarbon receptor antagonists promote the expansion of human hematopoietic stem cells. *Science* **329**, 1345-1348, doi:10.1126/science.1191536 (2010).

6 Cabezas-Wallscheid, N. *et al.* Identification of regulatory networks in HSCs and their immediate progeny via integrated proteome, transcriptome, and DNA methylome analysis. *Cell Stem Cell* **15**, 507-522, doi:10.1016/j.stem.2014.07.005 (2014).

7 Bagger, F. O. *et al.* BloodSpot: a database of gene expression profiles and transcriptional programs for healthy and malignant haematopoiesis. *Nucleic Acids Res* **44**, D917-924, doi:10.1093/nar/gkv1101 (2016).

8 Campbell, Z. T. *et al.* Cooperativity in RNA-protein interactions: global analysis of RNA binding specificity. *Cell Rep* **1**, 570-581, doi:10.1016/j.celrep.2012.04.003 (2012).

9 Jankowsky, E. & Harris, M. E. Specificity and nonspecificity in RNA-protein interactions. *Nat Rev Mol Cell Biol* **16**, 533-544, doi:10.1038/nrm4032 (2015).

10 Cosker, K. E., Fenstermacher, S. J., Pazyra-Murphy, M. F., Elliott, H. L. & Segal, R. A. The RNA-binding protein SFPQ orchestrates an RNA regulon to promote axon viability. *Nat Neurosci* **19**, 690-696, doi:10.1038/nn.4280 (2016).

11 Keene, J. D. RNA regulons: coordination of post-transcriptional events. *Nat Rev Genet* **8**, 533-543, doi:10.1038/nrg2111 (2007).

12 Venkatraman, A. *et al.* Maternal imprinting at the H19-Igf2 locus maintains adult haematopoietic stem cell quiescence. *Nature* **500**, 345-349, doi:10.1038/nature12303 (2013).

13 Gu, T. *et al.* IGF2BP2 and IGF2 genetic effects in diabetes and diabetic nephropathy. *J Diabetes Complications* **26**, 393-398, doi:10.1016/j.jdiacomp.2012.05.012 (2012).

14 Kahvejian, A., Svitkin, Y. V., Sukarieh, R., M'Boutchou, M. N. & Sonenberg, N. Mammalian poly(A)-binding protein is a eukaryotic translation initiation factor, which acts via multiple mechanisms. *Genes Dev* **19**, 104-113, doi:10.1101/gad.1262905 (2005).

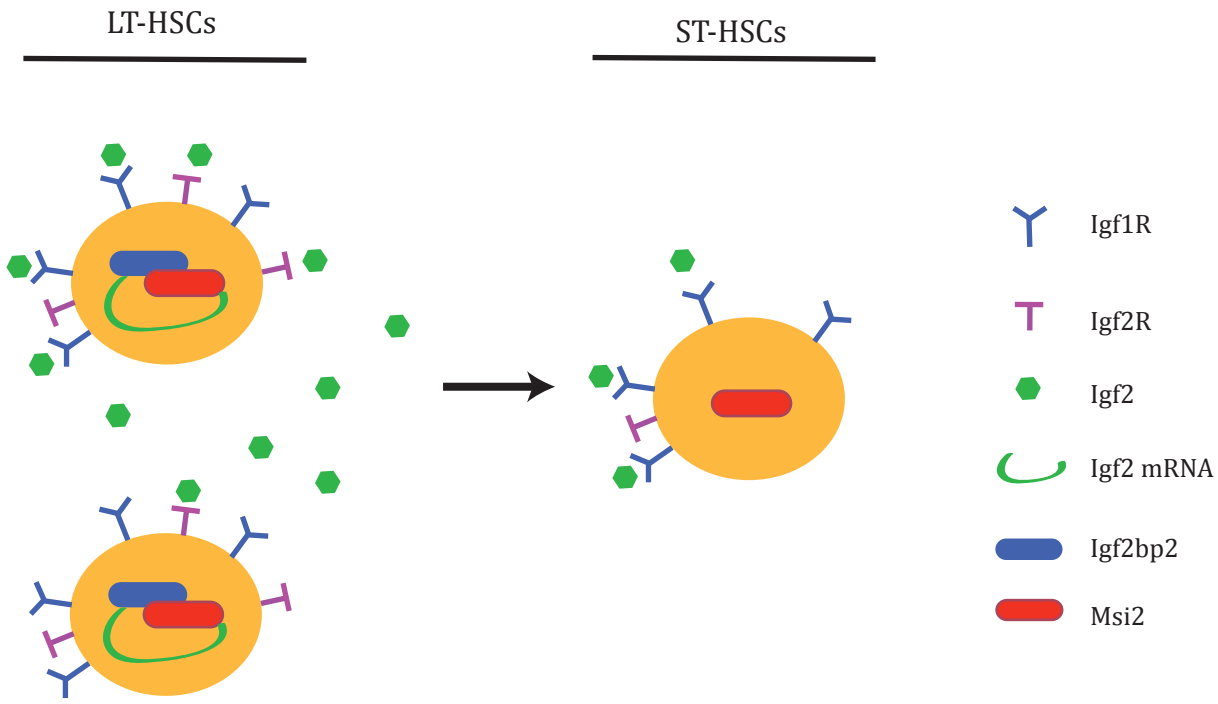


Figure 1. Model of Msi2 and Igf2bp2 control of HSC function. Msi2 and Igf2bp2 act in concert to facilitate the expression of the secreted protein Igf2. Resultantly, long-term HSCs reside in a niche that is rich in Igf2. Upon transition of dormant long-term HSCs to cycling-short-term HSCs, there is a loss of Igf2bp2 protein which is thought to contribute to the preferential expression of Igf1R. Preferential signaling of Igf2 through Igf1R results in enhanced cycling of these short-term HSCs