Spring 3-5-2013

Genetic and Epigenetic Interactions in in vivo and in vitro Reprogramming

Margaret Ashley Young
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Immunology

Dissertation Examination Committee:
Timothy Ley, Chair
Kyunghee Choi
Tom Ellenberger
Mark Sands
Barry Sleckman
Matthew Walter

Genetic and Epigenetic Interactions in *in vivo* and *in vitro* Reprogramming

by

Margaret Ashley Young

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

May 2013

St. Louis, Missouri

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 2: Canoncial and non-canonical HOX expression patterns in acute myeloid leukemia

Chapter 3: Genetic heterogeneity of murine induced pluripotent stem (iPS) clones revealed by whole genome sequencing

LIST OF TABLES

ACKNOWLEDGEMENTS

There are many people who have helped me along the way to the completion of my thesis. Over the past four years, I have had the privilege to work with an incredible group of people in the Ley lab. First, I must thank my mentor, Tim Ley. There have been a few bumps along the way, but Tim has always had a positive attitude and an ability to turn a negative result into an unexpected finding. I have learned an incredible amount of science from Tim, but more importantly I have learned what it means to be a great mentor. To appreciate Tim's passion for training the next generation of scientists, one must just walk down the hallway of the 6th floor of the Southwest Tower to see his successful trainees. I look forward to following in their footsteps with Tim's continued support along the way.

My colleagues in the Ley lab, both past and present, have been an integral part of my graduate experience. In particular, I would like to thank Dan George and Tami Lamprecht for their help completing this work; it would not have been possible without them. In addition, I would like to thank Sheng Cai, Cynthia Li, David Germain, Lukas Wartman, John Welch, and Jeff Klco for their advice, support and most of all friendship. I also want to thank my extended lab family- the entire 6th floor as well as the ES core and the HSCS core. Working here has been a truly unique experience of collaboration and I thank all of them for their help over the years.

I would also like to thank the members of my thesis committee- Tom Ellenberger, Mark Sands, Matt Walter, Barry Sleckman and KC Choi. My initial thesis proposal is barely recognizable in the work presented here and they have been a great source of advice and

support along its evolution. I especially want to thank Mark Sands for lending his expertise in the MPSVII model, along with Marie Nunez from his lab. Their reagents and advice have been indispensible. I thank the MSTP for administrative and financial support of my work along with the Division of Hematology training grant. Most importantly, I would like to thank the patients who decided to make the most of a horrible situation and participate in the Washington University tumor-banking program. Although I will never know who they are, this work would not be possible without their contributions.

My graduate experience would not have been complete without the friends I have made along the way. I am extremely grateful to have a great group of friends that have been on the MSTP journey with me since I first came to St. Louis. Toni, Matt, Brian and Mark, you have all been a source of motivation and encouragement for me over the past 6 years, which I know will continue as we move along in our careers. In addition, St. Louis would not have been the same without Jess, Dena, LeRoy, Gabe, Diego and all of the amazing people I have met through medical and graduate school. I especially thank Steve for keeping me sane through the writing process.

I am lucky to be a part of a large family and they have always been my biggest supporters, especially my siblings, Jessie, Becky and Danny. Most importantly, I would like to acknowledge my mom, Lori, for the sacrifices she has made to get me here.

DEDICATION

As a tribute to my own genetics, I thank my grandpa, Walter Queen, for my work ethic, thirst for knowledge and somewhat obsessive-compulsive traits, which have made me the scientist I am today- I dedicate this thesis to his memory.

ABSTRACT OF THE DISSERTATION

Genetic and Epigenetic Interactions in *in vivo* and *in vitro* Reprogramming

by

Margaret Ashley Young

Doctor of Philosophy in Biology and Biomedical Sciences
Immunology

Washington University in St. Louis, 2013

Professor Timothy Ley, Chairperson

In cancer pathogenesis and induced pluripotent stem (iPS) cell production, an essential step for reprogramming is acquisition of self-renewal. In hematopoietic cells, HOX genes are partially responsible for self-renewal, and HOX gene dysregulation commonly occurs in acute myeloid leukemia (AML). HOX dysregulation is seen in AML with translocations involving HOX genes themselves (e.g. NUP98-HOXA9) and with other disease-initiating translocations (e.g. MLL translocations and inv(16)). However, HOX genes are also highly expressed in many AML samples without translocations; the mechanism that causes "dysregulation" in these cases is unknown. Whole genome sequencing of 45 *de novo* AML genomes showed that recurrent mutations in the HOX gene clusters are not responsible for the phenotype. Expression array data from 190 AML cases revealed that while translocations have unique HOX expression patterns, most AML cases predominantly express *HOXA5, A9, A10, B2, B3*, and *MEIS1* in a canonical, highly coordinated pattern that is virtually identical to that found in normal human CD34

cells. HOX gene "dysregulation" in these cases may therefore represent the persistence of a normal, stem cell-specific HOX gene expression pattern that is probably required for self-renewal, and "captured" by mutations that initiate leukemia in hematopoietic stem cells.

In vitro reprogramming induces self-renewal with overexpression of a cocktail of transcription factors, often in the form of integrating viruses, which have raised concerns about the genomic integrity of iPS lines. We performed whole genome sequencing of 10 murine iPS lines produced in 3 independent experiments. We found an average of 414 somatic nucleotide variants (SNVs) per iPS clone, with variant allele frequencies suggesting that the mutations occurred at or before reprogramming. In one experiment, four independent iPS clones contained 164 identical variants (6 protein-coding SNVs, 157 non-coding SNVs and 1 structural variant) that were also found in rare parental cells, suggesting that these rare cells were extraordinarily "fit" for reprogramming. Our data suggest that most of the mutations detected in iPS cells occurred prior to reprogramming and are simply "captured" by cloning; however, some preexisting mutations provide an advantage for reprogramming, and may provide novel insights into the genetic underpinnings of this process.

Chapter 1

Introduction

1.1 Development and reprogramming of somatic cells

In developmental biology, differentiation is the process of stem and progenitor cells evolving to generate all of the cell types of an organism. In this physiological setting, differentiation is generally a unidirectional process. The hematopoietic system can be used as an example of a well-studied differentiation pathway. All hematopoietic cells are generated from a single pool of hematopoietic stem cells (HSCs). Activating different transcription factors allows progenitors to become lymphoid (B, T and NK cells), myeloid (macrophages, neutrophils), erythrocytes, or megakaryocytes. These cell fate decisions are typically irreversible. This can be illustrated by development of T cells. The common lymphoid progenitor (CLP) is able to generate both B and T cells, and expresses the Notch receptor. When the CLP enters the thymus, it encounters its ligand, Jagged, activating Notch signaling [1]. Activated Notch not only promotes T cell development, but also blocks B cell development, ensuring that once a cell commits to being a T cell; it cannot convert to a B cell. Experimentally this can be seen as thymocyte deficiency in inactivated Notch models [2]. Conversely, B cell deficiency occurs in models of constitutively active Notch [3]. Cellular "reprogramming" involves manipulating cells to override the normal one-way process of development to generate stem cells from fully differentiated somatic cells.

In a living organism a process similar to reprogramming can sometimes lead to cancer. Oncogenic mutations are thought to allow a differentiated cell to take on the stem cell characteristic of self-renewal that can lead to the unchecked growth of a tumor. This reprogrammed cell is sometimes called a cancer stem cell (CSC). It is still unclear whether the CSC is a product of a differentiated cell being transformed to a stem cell, or

if an existing stem cell is altered to lose its normal growth restrictions to become transformed. Most likely, both scenarios occur.

In 2006, Yamanaka, *et al*. published the first report of *in vitro* reprogramming in mouse cells. By overexpressing a cocktail of factors (*Oct4, Sox2, Klf4 and c-Myc*) in fibroblasts, his laboratory was able to generate cells with an embryonic stem (ES) cell-like phenotype, which they named induced pluripotent stem (iPS) cells [4]. A year later, human iPS cells were reported using the same factors by Yamanaka's group; a second group, lead by James Thomson, generated human iPS cells with a slightly different cocktail (*OCT3/4, SOX 2, NANOG and LIN28*) [5, 6]. Highlighting the relationship between this *in vitro* reprogramming and the *in vivo* reprogramming of oncogenesis is the fact that iPS cells are tested for pluripotency by defining their ability to form teratomas in immune-deficient mice [4-6]. The mechanism behind both types of reprogramming is still not understood. To determine how a cell can be manipulated to become a new cell type, first one has to know what determines cell identity and fate.

Identity is defined at the organismal and cellular levels by genetics. Unique combinations of single nucleotide polymorphisms (SNPs) create diversity and distinguish one person from the next. But within a single person there are hundreds of different cell types. Each of these cells contains the same genetic information, yet they are able to perform distinct functions. These cell fate decisions are determined by epigenetic factors. The term epigenetics refers to any alteration of DNA that does not involve a permanent change in the actual nucleotide sequence. The two main layers of epigenetics are DNA methylation and histone modifications. The main consequences of these changes are alterations in gene expression. Epigenetic changes can be heritable, but unlike genomic changes, they

are also reversible (which can be seen in reprogramming). Since cell identity and fate is determined by a combination of both the genome and epigenome, it follows that alterations in either can have detrimental effects, such as oncogenic transformation.

1.2. Acute Myeloid Leukemia: A model of "*in vivo* reprogramming"

Acute myeloid leukemia (AML) is a heterogeneous group of diseases at the level of morphology, molecular lesions and prognosis. Strategies have been designed to stratify AML into groups that can help classify patients for therapy and prognosis. The French-American-British (FAB) classification system was developed over thirty years ago, and classifies AML (based on classic morphological characteristics) into 8 subtypes MO-M7 [7]. While FAB categories overlap some with cytogenetics (nearly all M3 or acute promyelocytic leukemia (APL) cases harbor the (15;17) translocation, for example), the World Health Organization (WHO) has created a more clinically useful classification scheme (AML with cytogenetic abnormalities, AML with a prior dysplasia, therapy related AML and "other AML" that does not fall into one of the previous categories)[8]. The variability of AML can be used to illustrate the different types of genetic alterations involved in cancer pathogenesis- point mutations, translocations and insertions/deletions.

1.2.1. Reprogramming translocations in AML

Recurrent translocations are currently the most useful marker for prognosis within AML. Translocations are large-scale chromosomal abnormalities, which can be visualized with low-resolution studies such as karyotyping and fluorescence *in situ* hybridization (FISH). Double-strand DNA breaks that occur throughout the genome are normally repaired by joining the adjacent regions back together with no loss of genomic integrity. However, if there are multiple double stranded breaks present in the genome at the same time, it is

possible to swap the chromosomes during the repair process, which leads to translocations.

As mentioned above, a translocation between chromosomes 15 and 17 is the defining factor of APL. This translocation creates a fusion protein involving the retinoic acid receptor alpha (RARα) and the promyelocytic leukemia gene (PML) to generate the novel PML-RARα protein [9]. Both PML and RARα contain DNA-binding motifs that are retained in the fusion protein and are thought to define a distinct set of target genes for PML-RARα [10]. Mouse models have been critical in defining PML-RARα as the initiating event in APL. The first mouse model of APL was created by expressing the PML-RARα fusion protein under the control of the human cathepsin G regulatory sequences (hCG-PR) to target expression to the promyelocyte compartment by the Ley laboratory and was subsequently confirmed by the Pandolfi group [11, 12]. While hCG-PR mice do develop leukemia, there is a latency of approximately 220 days with low penetrance (15 percent) [11]. The penetrance was increased to greater than 90 percent by knocking PML-RARα into the endogenous murine CG locus [13]. Co-expression of additional oncogenes (including activated FLT3, Bcl2 and activated K-ras) are able to increase the penetrance in the hCG- PML-RARα transgenic mouse [14-16]. This data, along with the finding of additional recurrent chromosomal abnormalities in APL mice, suggest that while PML-RARα is sufficient as an initiating event, progression hits are required for full leukemic transformation [17]. Another important effect of translocations causing fusion proteins is the reciprocal haploinsufficiency of the fusion partners which could also have pathogenic effects, although these are modest in the case of t(15,17) [18]. While the leukemogenic mechanism of these fusion proteins is not fully understood,

some have a favorable effect on prognosis [19]. In the case of patients with t(15;17), this is because all-trans retinoic acid is a highly effective, targeted therapy [20-24].

One of the most common translocations in AML involves chromosomes 8 and 21, causing a fusion protein of AML1 and ETO. AML1 is a subunit of the core binding factor (CBF) transcription factor; the β unit of CBF is involved in another common AML translocation inv(16) where it creates a fusion protein with the myosin heavy chain gene, MYH11. These novel CBF fusion proteins alter the activity of the transcription factor, but mechanisms of transformation are not fully understood [19, 25, 26]. Knock-in mouse models of AML1-ETO are embryonic lethal, [27, 28] so conditional models have been used to study the leukemogenic capacity of the fusion protein [29-31]. By expressing AML1-ETO via the Sca1 locus (targeting to the hematopoietic stem cell compartment), heterozygous HSCs showed increased survival *in vitro,* and the mice developed a myeloproliferative disorder [31]. Myeloid-lineage expression (including common myeloid progenitors) of AML1-ETO did not lead to malignancy on its own, but treatment with a DNA-alkylating mutagen was able to skew AML1-ETO expressing cells to myeloid leukemia, as opposed to lymphoblastic leukemia in their wildtype littermates [29]. Similar to PML-RARα, AML1-ETO acts as an initiating mutation, but requires additional progressive mutations to develop leukemia (such as *Flt3* and *c-Kit* mutations) [32, 33]. AML1-ETO translocations are also similar to t(15;17), in that they predict a favorable outcome [34].

In contrast, translocations involving the mixed lineage leukemia (*MLL*) gene on chromosome 11 confer a poor prognosis [35, 36]. There are more than fifty fusion partners involved in MLL translocations. In 1997, a group lead by Terrence Rabbits

showed that chimeric mice produced by injecting blastocysts with embryonic stem (ES) cells engineered with an MLL-AF9 fusion developed leukemia starting at 4 months, with most mice dying by one year. Interestingly, even though the animals showed widespread activity of the MLL promoter, tumors were restricted to the myeloid compartment [37]. Using mice heterozygous for the MLL-AF9 fusion (homozygous expression is embryonic lethal) the same group found that AML in these animals is preceded by myeloproliferation, indicating that, like the other reprogramming translocations, additional mutations are necessary to develop frank leukemia [38]. Using a retroviral system with bone marrow transplant, Armstrong's group proved that FLT3-ITD cooperates with MLL-AF9, decreasing latency from ~70 days to 30 days [39]. In addition to translocations, MLL can also have a "partial tandem duplication" mutation. While this is often associated with trisomy 11, there are also cases of normal karyotype MLL-PTD [40, 41]. The main molecular consequence of MLL fusions is thought to be overexpression of *HOX* genes [42, 43]. This association along with *HOX* expression in normal karyotype AML is explored in **Chapter 2.**

1.2.2. Recurrent mutations in AML

With advances in molecular techniques, there are now a handful of recurrent point mutations associated with AML. A point mutation refers to changing a single nucleotide; these mutations can be classified as missense, nonsense or silent. A silent mutation either falls within a non-coding region of DNA, or if within an exon, does not alter the amino acid sequence (e.g. codons AAA and AAG both code for lysine), and is generally considered to be innocuous. However, silent mutations that fall within regulatory sequences, such as splice site acceptor or donor sites, can have detrimental effects even

though they do not fall within a coding sequence. A missense mutation, however, changes the corresponding amino acid, which may or may not drastically alter protein function (e.g. the mutation that causes sickle cell anemia is an A to T substitution that changes a glutamic acid residue to valine). A nonsense mutation changes a coding triplet to a premature stop codon (e.g. C to T substitution changes the codon CGA (arginine) to TGA (STOP)), which can then create a truncated protein that may be non-functional, dominant negative, or cause nonsense mediated decay. One of the most common (4-7% of cases) point mutations seen in AML is in the protein kinase domain of the fms-related tyrosine kinase 3 (FLT3-TKD) [44, 45]. The mutation is at residue 835 within the activation loop of the kinase domain and causes the protein to be constitutively active even in the absence of ligand [46]. More recently, mutations in DNA methyltransferase 3A (*DNMT3A*) were shown to be present in 22% of AML patients [47, 48]; the most common mutation falls within the methyltransferase domain at residue 882. Both *FLT3* and *DNMT3A* point mutations predict a poor prognosis [47, 49-52].

Copy number variations (CNVs) are generated by the addition or deletion of nucleotides. Insertions and deletions cover the full spectrum from a single nucleotide to large blocks of sequence (megabases). Smaller CNVs are called “indels” and once again, FLT3 provides an example of this genetic aberration seen in AML. Even more common than point mutations within the activation loop of FLT3 are internal tandem duplications within the juxtamembrane domain of the protein (FLT3-ITD). FLT3-ITDs are a heterogeneous set of mutations with the length and position of the duplication being variable. FLT3-ITD mutations lead to a constitutively active kinase and a poor prognosis for the patient [53, 54]. *FLT3* and *DNMT3A* mutations are often found in conjunction

with mutations in the nucleophosmin (*NPM1*) gene [55]. Virtually all mutations in *NPM1* induce frame shifts that change the C-terminal amino acid sequence, causing a change in the distribution pattern of the protein (from nuclear to cytoplasmic) [55]. Mutations in the transcription factor CCAAT/enhancer binding protein alpha (C/EBP α) involved in myeloid differentiation and proliferation are seen in 7-10% of de novo AML cases and are comprised of all types of mutations- for example, 3-bp deletions and 3- or 15- bp duplications that cause in-frame alterations, as well as a nonsense point mutation at residue 284 [56]. These mutations have been shown to cause loss-of-function; on their own, *CEBPA* mutations are associated with favorable outcome [57, 58].

Larger CNVs allow for deletion or duplication of an entire gene or set of genes. Generally, tumor suppressor genes are involved in deletions and oncogenes are duplicated. A common chromosomal loss seen in AML (and myelodysplastic syndrome-MDS) is del(5q), where all or part of the long arm of chromosome 5 is deleted. While the borders of the deleted region are variable, there are commonly deleted regions (CDRs). There are multiple putative tumor-suppressor genes within these CDRs, including *RPS14* (a ribosomal subunit necessary for translation) and *α-catenin* (cytoskeletal remodeling protein) [59, 60]. Matt Walter's group analyzed knockdown of another gene that falls within del(5q3.2), *HSPA9*, in human primary hematopoietic cells and murine bone marrow transplant experiments. In the mouse model, there was a significant decrease in multiple lineages including hematopoietic progenitors; in human cells defects were limited to delayed maturation of erythroid cells and decreased cell growth. These results show that while HSPA9 haploinsufficiency is not leukemogenic on its own, it does

contribute to abnormal hematopoiesis [61]. Large deletions could be a less specific way to target multiple cooperative tumor-suppressors at once.

While these mutation types have been discussed as individual events, oncogenesis is thought to be a progressive process, and cancer cells harbor multiple genetic lesions of various mutation types. By gradually learning more about the mutational landscape of cancer, we can better understand the mechanisms of oncogenesis.

1.3. The epigenetic landscape

1.3.1. DNA methylation

In addition to genetic changes, alterations in epigenetic signatures are thought to alter gene expression and play a role in cancer progression. Epigenetic modification of the genome is normal; changing normal patterns is what can be pathogenic. The epigenetic landscape is much more complicated than the straightforward four base pair code that defines the genome. There are multiple layers of epigenetic modifications that are able to interact to regulate gene expression. Here, I will focus on the two main forms of epigenetic modifications: DNA methylation and histone modifications. DNA methylation is addition of a methyl side group to the 5 position of the pyrimidine ring of nucleotides (most commonly the cytosine residue of a CpG dinucleotide). DNA methyltransferases (DNMTs) are the proteins responsible for methylation. DNMT1 is a maintenance enzyme, while DNMT3A and DNMT3B are *de novo* methyltransferases. While underrepresented in the genome, CpG dinucleotides are concentrated in gene promoters as well as CpG islands. Methylation at promoters can act as a repressive signal for gene expression. Recent work has shown that gene expression is also affected by intragenic methylation [62]. As opposed to promoter methylation, intragenic methylation can be a

positive signal for expression. Many tumors show an overall hypomethylation phenotype, with focal hypermethylation at key promoters [63, 64].

The original methylation hypothesis suggested that hypomethylation caused loss of repression of oncogenes. However, it is now appreciated that there are at least three oncogenic mechanisms enabled by hypomethylation: 1) chromosomal instability, 2) loss of imprinting and 3) reactivation of transposable elements. A decrease in methylation creates chromosomal instability by favoring mitotic recombination, which can lead to translocations and loss of heterozygosity (important for loss of tumor suppressor expression in cancer cells). Hypomethylation in centromeric regions can lead to aneuploidy, another common feature of malignant cells. Imprinting is an inherited epigenetic mark that leads to mono-allelic expression of certain genes. Two imprinted genes on chromosome 11 are affected by global hypomethylation in cancer samples: *H19* and *IGF-2*. A decrease in methylation can lead to overexpression of the anti-apoptotic growth factor (IGF-2) and loss of a transformation-suppressing RNA, H19, in some childhood cancers [65]. Lastly, hypermethylation is a technique that human cells use to silence parasitic sequences that have integrated into the genome. These parasitic elements can act as transposons when expressed, inserting into novel sites and disrupting normal gene expression. Howard, *et al*. used a mouse model of hypomethylation to show that methylation status affected transposition of endogenous retroviral sequences. The hypomethylated genomes had transposons inserted into the *Notch1* locus, which led to oncogenic activation of the gene [66]. There are some examples of hypomethylation of an oncogene contributing to cancer by increasing expression. The multidrug resistance gene (*MDR1*) was shown to have an inverse relationship between its promoter's methylation

and its expression level [67]. While this alteration would not initiate cancer, it could have important affects on drug effectiveness for some affected patients.

Focal hypermethylation of promoters of tumor suppressors, however, is a recurrent observation in malignancies. The cyclin-dependent kinase inhibitor p15 (CDKN2B) is a tumor suppressor frequently altered in many cancers. While deletions of *p15* are common in acute lymphoblastic leukemia (ALL), deletions and mutations of *p15* are rare in AML. Instead, inactivation of p15 in AML is achieved by hypermethylation of the *p15* promoter [68-71]. In one study, 88% of adult AML patients showed *p15* promoter hypermethylation [70]. Another group showed that hypermethyation is associated with poor prognosis, but this has not been confirmed in additional studies [71].

DNA methylation can also indirectly cause genetic aberrations, revealing an interplay of genetics and epigenetics. Methylated cytosine residues undergo spontaneous deamination, which changes the cytosine to a thymine, a transition mutation. The APOBEC family of enzymes can also catalyze this deamination [72]. AID (activation-induced cytidine deaminase) is a member of the APOBEC family, which is responsible for generating antibody diversity by creating mutations within the lymphocyte receptor genes. Inappropriate expression of AID, however, has been shown to be oncogenic in some systems [73-77].

1.3.2. Histone Modifications

In cells, DNA is wrapped around histones to form chromatin. Histones are octomeric complexes composed of four proteins (H2A, H2B, H3 and H4: 2 subunits of each). The amino-terminal tail of each protein is exposed. Lysine residues on the tails of H3 and H4 can be acetylated or methylated to affect chromatin structure and gene expression. Lysine

methyltransferases (HMTs) are specific to a single lysine residue. For example, the enzyme EZH2 methylates H3K27, while MLL1 acts on H3K4. There are also histone demethylases (HDMs) that remove methyl groups, allowing histone modification to be a dynamic process [78]. Similar to DNA methylation, histone modifications can be positive or negative regulators of transcription, depending on where they are located within or near a gene. For example, H3 methylation within the coding region activates transcription, while methylation in the promoter has a negative effect [79]. Each lysine has 3 potential methylation sites allowing for mono-, di- and tri-methylated species. There are three lysine residues within histones thought to activate transcription- H3K4, H3K36 and H3K79. In contrast, H3K9, H3K27 and H4K20 have been implicated in repression [78, 80]. However, as mentioned above, H3K9me within the coding region can activate gene expression [79]. Lysines can also have an acetyl group added to them instead of a methyl group. Histone acetyltransferases (HATs) are less specific than their methyltransferase counterparts. One acetyltransferase can act on H3 and H4, and multiple lysine residues within each: CBP acetylates K14 and K18 of H3, and K5 and K8 of H4 [78]. Acetylation activates transcription, while histone deacetylases (HDACs) act to repress transcription. Acetylation is also thought to impact chromatin structure. Acetylation leads to loss of the positive charge of the lysine side group, which disrupts the charge interactions of DNA and the histones. This would lead to an opening of the chromatin, making the DNA more accessible for transcription [78]. Similar to the global hypomethylation in cancer cells, there is a global decrease in H4K20 trimethylation and H4K16 acetylation [81]. Muller-Tidow, *et al*. analyzed H3K9 methylation in primary AML samples and found a decrease in methylation in promoters of hundreds of genes.

This decrease often corresponded to an increase in gene expression [82]. Since epigenetic alterations are reversible, they are good candidates for treatment. Four epigenetic drugs have been approved by the US Food and Drug Administration: two DNMT1 inhibitors (5-azacitadine and decitabine) and two HDAC inhibitors (vorinostat and romidepsin) are now used clinically, and many more are in development [83].

The simplest illustrations of the interplay between genetics and epigenetics in cancer are mutations in genes responsible for epigenetic manipulations. There are many cases of chromosomal translocations involving HMTs, HDMS, HACs, and HDACs. For example, the *MLL* alterations mentioned above (translocations and MLL-PTD) are present in 5-10% of adult AML. MLL is the HMT that acts on H3K4 [83-85].

1.4. Recurrent mutations in AML highlight interactions of genetics and epigenetics

The impact of mutations affecting DNA methylation is still under investigation. In addition to the recurrent mutations discovered in the *de novo* methyltransferase DNMT3A [47, 48, 86] mutations in isocitrate dehydrogenase 1 and 2 (IDH1/2) and tet oncogene family member 2 (TET2) are also common in AML [87-93]. TET2 is thought to function as an indirect DNA de-methyltransferase. It catalyzes the conversion of 5-methylcytosine (5-mC) to 5-hydroxymethylcytosine (5-hmC). The 5-hmC is thought to be a target of a specific glycosylase that regenerates the original cytosine residue [94, 95]. IDH1/2 then plays an even more removed role in epigenetics. IDH1/2 are involved in citrate metabolism; activating mutations result in a neomorphic function of the enzymes to generate 2-hydroxyglutarate (2-HG) from alpha-ketoglutarate (αKG). While the novel 2-HG product may have some toxic effects on the cell, mutant IDH1/2 is thought to affect epigenetics by the subsequent decrease in αKG levels, inhibiting proteins (including

TET2) that are αKG -dependent. Loss-of-function TET2 mutations and gain-of-function IDH1/2 mutations would be predicted to have the same outcome, hypermethylation. This is supported by the fact that these mutations are mutually exclusive of each other [88]. Experimentally, there have been conflicting reports of the consequence of *IDH1/2* and *TET2* mutations. While Ari Melnick's group found the expected hypermethylation phenotype, another group lead by Anjana Rao reported hypomethylation in mutant *TET2* samples [88, 96]. Our own data, which will be discussed in **Chapter 2**, shows no change in global methylation levels with *IDH1/2* or *TET2* mutations.

The epigenetic effects of *DNMT3A* mutations are most likely varied. While the nonsense and frame shift mutations must cause loss-of-function, the effect of the missense mutations within the methyltransferase domain (including the most common R882 mutations) is still unknown. There is no difference in total 5-mC content between wildtype and *DNMT3A* mutated AML samples measured by liquid chromatography-tandem mass spectrometry [47]. While MeDIP-ChIP analysis showed some regions of differential methylation in DNMT3A mutant samples, they did not correlate with changes in gene expression [47].

Each of the reprogramming translocations described above also has a link to epigenetics. MLL is directly involved in epigenetics, with its role as a histone H3K4 methyltransferase. AML1-ETO and PML-RARα indirectly affect epigenetic modifications through their protein binding partners. AML1-ETO recruits both DNMT1 and HDAC1 to repressor complexes of AML1 target gene promoters [97]. The importance of these associations has been suggested by the use of inhibitors of both deacetylases and methyltransferases as anti-leukemic agents [97-99]. PML/RARα also

binds multiple epigenetic regulatory proteins, including DNMT3A, JMJD3 (an H3K27 demethylase), and SETDB1 (an H3K9 methylase) [100, 101].

1.5. Acquisition of self-renewal in AML

One of the main hallmarks of cancer is the acquisition of self-renewal in cancer initiating cells. There are 2 main pathways that are currently known to be associated with self-renewal in hematopoietic stem cells (HSCs) and AML: Notch and HOX. The Notch pathway's role in hematopoiesis is not fully understood, but multiple studies have shown its ability to impart self-renewal on hematopoietic progenitors. Notch signaling is initiated by binding of ligand (Jagged1 and 2 and Delta1, 3 and 4) to the extracellular domain of the Notch receptor. Upon binding, the receptor is cleaved to release the active intracellular domain (ICD) into the cell. The Notch ICD then transports to the nucleus where it can act on its targets through the transcription factor CSL and the co-activator, Mastermind-like (MAML) [102, 103]. Retroviral expression of constitutively active Notch ICD in hematopoietic stem cells rendered them immortalized (while still cytokine-dependent) without losing their ability to generate lymphoid and myeloid progeny [104]. While this study suggests that Notch is sufficient for inducing self-renewal in hematopoietic progenitors, other studies have shown that it is not required. Conditional knockouts of Notch1 and 2—and also Jagged1-- in the hematopoietic compartment did not demonstrate a defect in HSC function, nor did blocking Notch signaling downstream of the receptor with a dominant negative MAML or conditional knockout of CSL [105-107]. Taken together, these results suggest that HSCs have redundant pathways for self-renewal, such as HOX signaling.

Notch is involved in a translocation event, t(7;9), where its expression is controlled by the TCR locus in rare cases of T cell acute lymphoblastic leukemia (T-ALL) [108]; activating mutations of Notch1, however, are present in 50-60% of T-ALL cases [109]. In contrast, Notch mutations are very rarely found in cases of AML. Notch is expressed in all FAB subtypes, and several AML-associated fusion proteins are capable of activating Notch signaling in the absence of ligand [110-113].

HOX (homeobox containing) genes are a family of transcription factors highly conserved from *Drosophila* to humans. HOX genes were initially identified as regulators of positional identity along the anterior-posterior axis of animal models [114]. Their expression in hematopoietic progenitors suggests that they may also play a role in hematopoiesis, which has been studied extensively with overexpression and knockouts in mouse models [115-119]. Overexpression of Hoxb4 in murine HSCs stimulates *in vivo* and *ex vivo* expansion of HSCs without promoting leukemia, demonstrating its ability to promote self-renewal [120-122]. Based on the existence of HOX fusions with nucleoporin 98 (NUP98) in rare cases of AML, Keith Humphries' group has also tested multiple NUP98-HOX fusion proteins for their ability to expand HSCs *ex vivo* [123, 124]. Like HOXB4 alone, NUP98-HOXB4 is a potent stimulator of HSC expansion, as are NUP98 fusions with many other HOX genes (although to varying degrees) [125]. Interestingly, when only the homeobox domain of HOXA10 was fused with NUP98 (NUP98-HOXA10HD) it showed an even greater expansion capacity than NUP98-HOXB4 (>2000-fold vs. 300-fold) proving that the homeodomain alone is able to induce HSC self-renewal [126]. Similar to Notch, knockout models of HOX genes do not lead to severe impairments of hematopoiesis, illustrating not only compensation of other self-

renewal pathways such as Notch, but also the redundancy of the remaining HOX genes [127-129].

In addition to the NUP98 translocations mentioned above, HOX dysregulation is also thought to be a common feature of AML. MLL is an upstream regulator of HOX genes, and MLL translocations and MLL-PTD have been shown to cause dysregulation of HOX genes [84, 130]. Yan, *et al*. also reported a significant affect of *DNMT3A* R882H on expression of HOX genes [48]. One report showed that HOXA9 overexpression is the most predictive single factor for poor prognosis in AML [131]. On the other hand, *NPM1* mutations are associated with favorable outcome as well as HOX overexpression, demonstrating that the role HOX genes play in AML pathogenesis is still poorly understood [132-136]. In **Chapter 2**, we analyze a diverse set of 190 AML patients for HOX expression patterns, along with methylation and genetic studies, to better define the role of HOX gene dysregulation in AML.

1.6. Induced Pluripotent Stem cells: *in vitro* Reprogramming

1.6.1. The early years: Somatic cell nuclear transfer

While *in vitro* reprogramming has become a mainstream technique in the past four years, the concept has a much longer history. In the 1990's, somatic cell nuclear transfer (SCNT) was developed as a cloning technique. In this process, the nucleus of a somatic cell is removed and inserted into an oocyte, which has had its own nucleus removed. The oocyte is stimulated to divide and if successful, will form a blastocyst. While the most famous case of SCNT is the cloning of a sheep (Dolly) in 1996, the goal of SCNT was not reproductive cloning (which brings with it considerable ethical concerns) but therapeutic cloning to produce new embryonic stem (ES) cell lines [137]. SCNT proved

to have too many limitations to be practical. First and foremost, the technique requires significant manual, tedious work that cannot be moved to a large-scale system. The procedure also puts considerable stress on the oocyte and nucleus, so that only a very small proportion of transferred cells make it through to the blastocyst stage (from which an ES cell line can be made). The reagents themselves are limiting as well, since healthy female donors have to provide the oocytes. No ES cell lines have yet been successfully generated from SCNT. Since the introduction of transcription factor induced reprogramming, the SCNT field has largely been abandoned. However, very recently a group lead by Dieter Egli showed that some of the problems with SCNT can be overcome by leaving the oocyte nucleus intact, thereby creating triploid pluripotent cells [138]. Whether this finding will lead to a renewal of the SCNT field remains to be seen. Many of the limitations of SCNT have been addressed in transcription factor-mediated reprogramming. The only technical skills needed are basic tissue culture and viral transduction, which can easily be automated. However, the usefulness of iPS clones for generating safe cells for therapeutic purposes has not yet been established.

1.6.2. Reprogramming goes mainstream

Since their introduction over four years ago [4], iPS cells have allowed researchers a new way to model many biological processes and disease states [139-142]. Much of the "hype" surrounding iPS cells, however, is their potential as a therapeutic reagent [143-145]. This has been the driving force behind many of the modifications to iPS generation protocols. The initial 4-retrovirus method had large safety concerns due to retroviral-induced insertional mutagenesis, as well as the use of c-myc as a reprogramming factor,since it is a known oncogene. Reprogramming can now be accomplished without

c-myc [146, 147], with single lentiviruses [148-150], and by transiently expressing the reprogramming factors as DNA [151, 152], RNA [153] or protein [154]. While there is still considerable work being done to maximize efficiency and safety of reprogramming protocols, it is clear that we now need to decipher the mechanisms behind reprogramming and fully define the iPS phenotype.

1.6.3. Mechanisms of reprogramming

In 2009, Shinya Yamanaka published an editorial proposing 2 models to explain the low efficiency of iPS reprogramming: the "elite" vs. stochastic models [155]. According to the elite model, reprogramming has low efficiency because only a few cells within the donor pool are competent for reprogramming. For example, less-differentiated cells within a heterogeneous population are more readily transformed than their fully differentiated counterparts. The similarity of the frequency of multipotent stem cells within in the skin (0.067%) and initial reprogramming efficiency (~0.05%) support this model [156, 157]. However, with improved techniques, along with addition of small molecules to enhance reprogramming (such as valproic acid), efficiency is now reproducibly seen up to 10%, much higher than stem cell frequency [157, 158]. Yamanaka also postulates that since reprogramming efficiency is higher with insertional vectors, the positions of retroviral or lentiviral integrations may be important for activating or inactivating endogenous genes [155].

In contrast, the stochastic model states that the majority of differentiated cells are capable of reprogramming with alteration of their epigenetic landscape. In 1957 Conrad Waddington proposed a model of cell differentiation as a ball rolling down a series of slopes from a totipotent stem cell down to a lineage-committed cell. Along the way there

are valleys, which keep a cell from traveling in the reverse direction [159]. To maintain stem cells, there is an epigenetic block that keeps them from traveling down the path to differentiation. In the stochastic model, forced expression of the reprogramming transcription factor cocktail can move any cell back up to the totipotent level, but epigenetic changes must occur for full reprogramming to occur [155].

This "roadblock" hypothesis suggests that reprogramming is a purely epigenetic phenomenon, and is currently the dogma in the field. Successful reprogramming requires not only loss of methylation from pluripotency-related genes (such as *Nanog* and *Oct4*), but also new methylation to turn off expression of differentiation genes. Multiple groups have shown that iPS cells retain some "memory" of their parental cells at the epigenetic level [160]. A recent high-resolution analysis of the methylome of human iPSCs showed that although they are globally similar to ES cells, iPSCs have unique methylation patterns that could affect their therapeutic potential [161]. A report of successful reprogramming of mouse embryonic fibroblasts (MEFs) deficient for Dnmt3A and Dnmt3B (the *de novo* methyltransferases responsible for methylation in ES cells) shows that epigenetic changes produced by these genes are not required for reprogramming [162].

Until recently, karyotyping was the only measure of genomic stability performed on iPS cells [5, 6, 163, 164]. This low-resolution view of the genome could easily miss deleterious mutations (point mutations, small insertions and deletions, *etc*.). Earlier this year, high-resolution SNP genotyping was performed on human ES cells and iPSCs [165]. This study showed that human iPSCs have copy number alterations in oncogenes and tumor-suppressor genes not seen in ES cells. Another group used the same approach

to compare early- and late- passage human iPS lines. They found that late-passage iPS cells have fewer CNVs than early-passage lines, due to selection against highly mutated cells with passage in culture [166]. To test how reprogramming affects the genome at the nucleotide level, Gore, *et al*. performed exome sequencing on 22 human iPS lines. These lines were produced in multiple labs using five different transcription factor delivery protocols, including non-integrating DNA in episomes and messenger RNA. They found that each line had on average 5 single nucleotide variants (SNVs) in coding sequences. Although not discussed by the authors, there were no significant differences in SNV numbers that corresponded to the factor delivery method. They did report that the SNVs were enriched for in genes that have been shown to be involved in cancer [167]. They also found 3 examples where iPS lines derived in the same experiment from the same donor cells harbored identical mutations, but they did not explore this finding, or comment on its potential relevance for the "elite" model proposed above. **Chapter 3** details our whole genome sequencing results of ten mouse iPS lines, and introduces the idea of an elite model where reprogramming fitness mutations, as opposed to level of donor cell differentiation, can sometimes predetermine a cell's ability to be reprogrammed.

1.7. Summary

It is clear that we are just beginning to unravel the integration of genetics and epigenetics in iPS cells and cancer. With technological advancements in both fields, we now have the ability to detect subtle changes --such as single nucleotide mutations and methylation changes-- at individual CpG residues. As discussed above, these small alterations can have a big impact on cellular function. While translocations have identified groups of

favorable and poor risk patients with AML, the intermediate risk category is still relatively undefined. The discovery of *DNMT3A* mutations in 22% of these patients, and its impact on outcomes, illustrates the importance of high-resolution genomic studies in cancer [47-49]. Technical advancements are also changing the study of epigenetics; Illumina's first generation methylation bead chip covered 27,578 CpG sites across the genome. The sites were are all within the proximal promoters of over 14,000 genes including genes involved in cancer, imprinting and micro RNAs. The array is highly biased, so it could miss many important changes. Recently, Illumina released a new methylation bead chip that now covers 450,000 CpG sites. This new array covers 5' and 3' regions of genes with no bias for CpG islands, as was seen in its predecessor (Illumina, Inc . San Diego, CA). As shown in **Chapter 2** with our analysis of HOX expression in AML, by integrating high-resolution epigenetic and genetic techniques, we may be able to better understand the process of cellular reprogramming for both iPS cell generation, and for cancer pathogenesis.

References

1. Zuniga-Pflucker, J.C., *T-cell development made simple.* Nature reviews. Immunology, 2004. **4**(1): p. 67-72.
2. Radtke, F., et al., *Deficient T cell fate specification in mice with an induced inactivation of Notch1.* Immunity, 1999. **10**(5): p. 547-58.
3. Pui, J.C., et al., *Notch1 expression in early lymphopoiesis influences B versus T lineage determination.* Immunity, 1999. **11**(3): p. 299-308.
4. Takahashi, K. and S. Yamanaka, *Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors.* Cell, 2006. **126**(4): p. 663-76.
5. Takahashi, K., et al., *Induction of pluripotent stem cells from adult human fibroblasts by defined factors.* Cell, 2007. **131**(5): p. 861-72.
6. Yu, J., et al., *Induced pluripotent stem cell lines derived from human somatic cells.* Science, 2007. **318**(5858): p. 1917-20.
7. Bennett, J.M., et al., *Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group.* British journal of haematology, 1976. **33**(4): p. 451-8.
8. Vardiman, J.W., N.L. Harris, and R.D. Brunning, *The World Health Organization (WHO) classification of the myeloid neoplasms.* Blood, 2002. **100**(7): p. 2292-302.
9. Goddard, A.D., J. Borrow, and E. Solomon, *A previously uncharacterized gene, PML, is fused to the retinoic acid receptor alpha gene in acute promyelocytic leukaemia.* Leukemia, 1992. **6 Suppl 3**: p. 117S-119S.
10. Payton, J.E., et al., *High throughput digital quantification of mRNA abundance in primary human acute myeloid leukemia samples.* J Clin Invest, 2009. **119**(6): p. 1714-26.
11. Grisolano, J.L., et al., *Altered myeloid development and acute leukemia in transgenic mice expressing PML-RAR alpha under control of cathepsin G regulatory sequences.* Blood, 1997. **89**(2): p. 376-87.
12. He, L.Z., et al., *Acute leukemia with promyelocytic features in PML/RARalpha transgenic mice.* Proceedings of the National Academy of Sciences of the United States of America, 1997. **94**(10): p. 5302-7.
13. Westervelt, P., et al., *High-penetrance mouse model of acute promyelocytic leukemia with very low levels of PML-RARalpha expression.* Blood, 2003. **102**(5): p. 1857-65.
14. Chan, I.T., et al., *Oncogenic K-ras cooperates with PML-RAR alpha to induce an acute promyelocytic leukemia-like disease.* Blood, 2006. **108**(5): p. 1708-15.
15. Kelly, L.M., et al., *PML/RARalpha and FLT3-ITD induce an APL-like disease in a mouse model.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(12): p. 8283-8.
16. Kogan, S.C., et al., *BCL-2 cooperates with promyelocytic leukemia retinoic acid receptor alpha chimeric protein (PMLRARalpha) to block neutrophil differentiation and initiate acute leukemia.* The Journal of experimental medicine, 2001. **193**(4): p. 531-43.

17. Le Beau, M.M., et al., *Recurring chromosomal abnormalities in leukemia in PML-RARA transgenic mice identify cooperating events and genetic pathways to acute promyelocytic leukemia.* Blood, 2003. **102**(3): p. 1072-4.
18. Welch, J.S., et al., *Rara haploinsufficiency modestly influences the phenotype of acute promyelocytic leukemia in mice.* Blood. **117**(8): p. 2460-8.
19. Paschka, P., *Core binding factor acute myeloid leukemia.* Seminars in oncology, 2008. **35**(4): p. 410-7.
20. Huang, M.E., et al., *Use of all-trans retinoic acid in the treatment of acute promyelocytic leukemia.* Blood, 1988. **72**(2): p. 567-72.
21. Castaigne, S., et al., *All-trans retinoic acid as a differentiation therapy for acute promyelocytic leukemia. I. Clinical results.* Blood, 1990. **76**(9): p. 1704-9.
22. Tallman, M.S., et al., *All-trans-retinoic acid in acute promyelocytic leukemia.* The New England journal of medicine, 1997. **337**(15): p. 1021-8.
23. Warrell, R.P., Jr., et al., *Differentiation therapy of acute promyelocytic leukemia with tretinoin (all-trans-retinoic acid).* The New England journal of medicine, 1991. **324**(20): p. 1385-93.
24. Tallman, M.S., *Treatment of relapsed or refractory acute promyelocytic leukemia.* Best practice & research. Clinical haematology, 2007. **20**(1): p. 57-65.
25. Walker, H., F.J. Smith, and D.R. Betts, *Cytogenetics in acute myeloid leukaemia.* Blood reviews, 1994. **8**(1): p. 30-6.
26. Goyama, S. and J.C. Mulloy, *Molecular pathogenesis of core binding factor leukemia: current knowledge and future prospects.* International journal of hematology, 2011. **94**(2): p. 126-33.
27. Yergeau, D.A., et al., *Embryonic lethality and impairment of haematopoiesis in mice heterozygous for an AML1-ETO fusion gene.* Nature genetics, 1997. **15**(3): p. 303-6.
28. Okuda, T., et al., *Expression of a knocked-in AML1-ETO leukemia gene inhibits the establishment of normal definitive hematopoiesis and directly generates dysplastic hematopoietic progenitors.* Blood, 1998. **91**(9): p. 3134-43.
29. Yuan, Y., et al., *AML1-ETO expression is directly involved in the development of acute myeloid leukemia in the presence of additional mutations.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(18): p. 10398-403.
30. Higuchi, M., et al., *Expression of a conditional AML1-ETO oncogene bypasses embryonic lethality and establishes a murine model of human t(8;21) acute myeloid leukemia.* Cancer cell, 2002. **1**(1): p. 63-74.
31. Fenske, T.S., et al., *Stem cell expression of the AML1/ETO fusion protein induces a myeloproliferative disorder in mice.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(42): p. 15184-9.
32. Schessl, C., et al., *The AML1-ETO fusion gene and the FLT3 length mutation collaborate in inducing acute leukemia in mice.* The Journal of clinical investigation, 2005. **115**(8): p. 2159-68.
33. Wang, Y.Y., et al., *AML1-ETO and C-KIT mutation/overexpression in t(8;21) leukemia: implication in stepwise leukemogenesis and response to Gleevec.*

Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(4): p. 1104-9.
34. 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, 1998. **92**(7): p. 2322-33.
35. Schoch, C., et al., *AML with 11q23/MLL abnormalities as defined by the WHO classification: incidence, partner chromosomes, FAB subtype, age distribution, and prognostic impact in an unselected series of 1897 cytogenetically analyzed AML cases.* Blood, 2003. **102**(7): p. 2395-402.
36. Armstrong, S.A., et al., *MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia.* Nature genetics, 2002. **30**(1): p. 41-7.
37. Corral, J., et al., *An Mll-AF9 fusion gene made by homologous recombination causes acute leukemia in chimeric mice: a method to create fusion oncogenes.* Cell, 1996. **85**(6): p. 853-61.
38. Dobson, C.L., et al., *The mll-AF9 gene fusion in mice controls myeloproliferation and specifies acute myeloid leukaemogenesis.* The EMBO journal, 1999. **18**(13): p. 3564-74.
39. Stubbs, M.C., et al., *MLL-AF9 and FLT3 cooperation in acute myelogenous leukemia: development of a model for rapid therapeutic assessment.* Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K, 2008. **22**(1): p. 66-77.
40. Gaidzik, V. and K. Dohner, *Prognostic implications of gene mutations in acute myeloid leukemia with normal cytogenetics.* Seminars in oncology, 2008. **35**(4): p. 346-55.
41. Basecke, J., et al., *The MLL partial tandem duplication in acute myeloid leukaemia.* British journal of haematology, 2006. **135**(4): p. 438-49.
42. De Braekeleer, M., et al., *The MLL gene and translocations involving chromosomal band 11q23 in acute leukemia.* Anticancer research, 2005. **25**(3B): p. 1931-44.
43. Quentmeier, H., et al., *Expression of HOX genes in acute leukemia cell lines with and without MLL translocations.* Leukemia & lymphoma, 2004. **45**(3): p. 567-74.
44. Bacher, U., et al., *Prognostic relevance of FLT3-TKD mutations in AML: the combination matters--an analysis of 3082 patients.* Blood, 2008. **111**(5): p. 2527-37.
45. Yamamoto, Y., et al., *Activating mutation of D835 within the activation loop of FLT3 in human hematologic malignancies.* Blood, 2001. **97**(8): p. 2434-9.
46. Griffin, J.D., *Point mutations in the FLT3 gene in AML.* Blood, 2001. **97**(8): p. 2193A-2193.
47. Ley, T.J., et al., *DNMT3A mutations in acute myeloid leukemia.* The New England journal of medicine, 2010. **363**(25): p. 2424-33.
48. Yan, X.J., et al., *Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia.* Nature genetics, 2011. **43**(4): p. 309-15.

49. Markova, J., et al., *Prognostic impact of DNMT3A mutations in patients with intermediate cytogenetic risk profile acute myeloid leukemia.* European journal of haematology, 2011.
50. Govedarovic, N. and G. Marjanovic, *Frequency and prognostic impact of FLT3/ITD mutation in patients with acute myeloid leukaemia.* Journal of B.U.ON. : official journal of the Balkan Union of Oncology, 2011. **16**(1): p. 108-11.
51. Ravandi, F., et al., *Outcome of patients with FLT3-mutated acute myeloid leukemia in first relapse.* Leukemia research, 2010. **34**(6): p. 752-6.
52. Levis, M., et al., *Results from a randomized trial of salvage chemotherapy followed by lestaurtinib for patients with FLT3 mutant AML in first relapse.* Blood, 2011. **117**(12): p. 3294-301.
53. Sallmyr, A., et al., *Internal tandem duplication of FLT3 (FLT3/ITD) induces increased ROS production, DNA damage, and misrepair: implications for poor prognosis in AML.* Blood, 2008. **111**(6): p. 3173-82.
54. Schnittger, S., et al., *Analysis of FLT3 length mutations in 1003 patients with acute myeloid leukemia: correlation to cytogenetics, FAB subtype, and prognosis in the AMLCG study and usefulness as a marker for the detection of minimal residual disease.* Blood, 2002. **100**(1): p. 59-66.
55. Falini, B., et al., *Cytoplasmic nucleophosmin in acute myelogenous leukemia with a normal karyotype.* The New England journal of medicine, 2005. **352**(3): p. 254-66.
56. Shih, L.Y., et al., *AML patients with CEBPalpha mutations mostly retain identical mutant patterns but frequently change in allelic distribution at relapse: a comparative analysis on paired diagnosis and relapse samples.* Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K, 2006. **20**(4): p. 604-9.
57. Gombart, A.F., et al., *Mutations in the gene encoding the transcription factor CCAAT/enhancer binding protein alpha in myelodysplastic syndromes and acute myeloid leukemias.* Blood, 2002. **99**(4): p. 1332-40.
58. Leroy, H., et al., *CEBPA point mutations in hematological malignancies.* Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K, 2005. **19**(3): p. 329-34.
59. Eisenmann, K.M., et al., *5q- myelodysplastic syndromes: chromosome 5q genes direct a tumor-suppression network sensing actin dynamics.* Oncogene, 2009. **28**(39): p. 3429-41.
60. Liu, T.X., et al., *Chromosome 5q deletion and epigenetic suppression of the gene encoding alpha-catenin (CTNNA1) in myeloid cell transformation.* Nat Med, 2007. **13**(1): p. 78-83.
61. Chen, T.H., et al., *Knockdown of Hspa9, a del(5q31.2) gene, results in a decrease in hematopoietic progenitors in mice.* Blood, 2011. **117**(5): p. 1530-9.
62. Bauer, A.P., et al., *The impact of intragenic CpG content on gene expression.* Nucleic acids research, 2010. **38**(12): p. 3891-908.
63. Esteller, M. and J.G. Herman, *Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours.* The Journal of pathology, 2002. **196**(1): p. 1-7.

64. Esteller, M., *Profiling aberrant DNA methylation in hematologic neoplasms: a view from the tip of the iceberg.* Clinical immunology, 2003. **109**(1): p. 80-8.
65. Feinberg, A.P., *Imprinting of a genomic domain of 11p15 and loss of imprinting in cancer: an introduction.* Cancer research, 1999. **59**(7 Suppl): p. 1743s-1746s.
66. Howard, G., et al., *Activation and transposition of endogenous retroviral elements in hypomethylation induced tumors in mice.* Oncogene, 2008. **27**(3): p. 404-8.
67. Nakayama, M., et al., *Hypomethylation status of CpG sites at the promoter region and overexpression of the human MDR1 gene in acute myeloid leukemias.* Blood, 1998. **92**(11): p. 4296-307.
68. Herman, J.G., et al., *Hypermethylation-associated inactivation indicates a tumor suppressor role for p15INK4B.* Cancer Res, 1996. **56**(4): p. 722-7.
69. Aggerholm, A., et al., *Promoter hypermethylation of p15INK4B, HIC1, CDH1, and ER is frequent in myelodysplastic syndrome and predicts poor prognosis in early-stage patients.* Eur J Haematol, 2006. **76**(1): p. 23-32.
70. Herman, J.G., et al., *Distinct patterns of inactivation of p15INK4B and p16INK4A characterize the major types of hematological malignancies.* Cancer Res, 1997. **57**(5): p. 837-41.
71. Wong, I.H., et al., *Aberrant p15 promoter methylation in adult and childhood acute leukemias of nearly all morphologic subtypes: potential prognostic implications.* Blood, 2000. **95**(6): p. 1942-9.
72. Chahwan, R., S.N. Wontakal, and S. Roa, *Crosstalk between genetic and epigenetic information through cytosine deamination.* Trends in genetics : TIG, 2010. **26**(10): p. 443-8.
73. Endo, Y., et al., *Expression of activation-induced cytidine deaminase in human hepatocytes via NF-kappaB signaling.* Oncogene, 2007. **26**(38): p. 5587-95.
74. Lin, C., et al., *Nuclear receptor-induced chromosomal proximity and DNA breaks underlie specific translocations in cancer.* Cell, 2009. **139**(6): p. 1069-83.
75. Matsumoto, Y., et al., *Helicobacter pylori infection triggers aberrant expression of activation-induced cytidine deaminase in gastric epithelium.* Nature medicine, 2007. **13**(4): p. 470-6.
76. Morisawa, T., et al., *Organ-specific profiles of genetic changes in cancers caused by activation-induced cytidine deaminase expression.* International journal of cancer. Journal international du cancer, 2008. **123**(12): p. 2735-40.
77. Okazaki, I.M., et al., *Constitutive expression of AID leads to tumorigenesis.* The Journal of experimental medicine, 2003. **197**(9): p. 1173-81.
78. Kouzarides, T., *Chromatin modifications and their function.* Cell, 2007. **128**(4): p. 693-705.
79. Vakoc, C.R., et al., *Histone H3 lysine 9 methylation and HP1gamma are associated with transcription elongation through mammalian chromatin.* Mol Cell, 2005. **19**(3): p. 381-91.
80. Varier, R.A. and H.T. Timmers, *Histone lysine methylation and demethylation pathways in cancer.* Biochim Biophys Acta. **1815**(1): p. 75-89.
81. Fraga, M.F., et al., *Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer.* Nature genetics, 2005. **37**(4): p. 391-400.

82. Muller-Tidow, C., et al., *Profiling of histone H3 lysine 9 trimethylation levels predicts transcription factor activity and survival in acute myeloid leukemia.* Blood. **116**(18): p. 3564-71.
83. Rodriguez-Paredes, M. and M. Esteller, *Cancer epigenetics reaches mainstream oncology.* Nat Med. **17**(3): p. 330-9.
84. Dorrance, A.M., et al., *Mll partial tandem duplication induces aberrant Hox expression in vivo via specific epigenetic alterations.* J Clin Invest, 2006. **116**(10): p. 2707-16.
85. Thirman, M.J., et al., *Rearrangement of the MLL gene in acute lymphoblastic and acute myeloid leukemias with 11q23 chromosomal translocations.* N Engl J Med, 1993. **329**(13): p. 909-14.
86. Shen, Y., et al., *Gene mutation patterns and their prognostic impact in a cohort of 1,185 patients with acute myeloid leukemia.* Blood, 2011.
87. Delhommeau, F., et al., *Mutation in TET2 in myeloid cancers.* The New England journal of medicine, 2009. **360**(22): p. 2289-301.
88. Figueroa, M.E., et al., *Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation.* Cancer cell, 2010. **18**(6): p. 553-67.
89. Marcucci, G., et al., *IDH1 and IDH2 gene mutations identify novel molecular subsets within de novo cytogenetically normal acute myeloid leukemia: a Cancer and Leukemia Group B study.* Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 2010. **28**(14): p. 2348-55.
90. Mardis, E.R., et al., *Recurring mutations found by sequencing an acute myeloid leukemia genome.* The New England journal of medicine, 2009. **361**(11): p. 1058-66.
91. Paschka, P., et al., *IDH1 and IDH2 mutations are frequent genetic alterations in acute myeloid leukemia and confer adverse prognosis in cytogenetically normal acute myeloid leukemia with NPM1 mutation without FLT3 internal tandem duplication.* Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 2010. **28**(22): p. 3636-43.
92. Wagner, K., et al., *Impact of IDH1 R132 mutations and an IDH1 single nucleotide polymorphism in cytogenetically normal acute myeloid leukemia: SNP rs11554137 is an adverse prognostic factor.* Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 2010. **28**(14): p. 2356-64.
93. Ward, P.S., et al., *The common feature of leukemia-associated IDH1 and IDH2 mutations is a neomorphic enzyme activity converting alpha-ketoglutarate to 2-hydroxyglutarate.* Cancer cell, 2010. **17**(3): p. 225-34.
94. Ito, S., et al., *Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification.* Nature, 2010. **466**(7310): p. 1129-33.
95. Cannon, S.V., A. Cummings, and G.W. Teebor, *5-Hydroxymethylcytosine DNA glycosylase activity in mammalian tissue.* Biochemical and biophysical research communications, 1988. **151**(3): p. 1173-9.
96. Ko, M., et al., *Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2.* Nature, 2010. **468**(7325): p. 839-43.

97. Liu, S., et al., *Targeting AML1/ETO-histone deacetylase repressor complex: a novel mechanism for valproic acid-mediated gene expression and cellular differentiation in AML1/ETO-positive acute myeloid leukemia cells.* The Journal of pharmacology and experimental therapeutics, 2007. **321**(3): p. 953-60.
98. Klisovic, M.I., et al., *Depsipeptide (FR 901228) promotes histone acetylation, gene transcription, apoptosis and its activity is enhanced by DNA methyltransferase inhibitors in AML1/ETO-positive leukemic cells.* Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K, 2003. **17**(2): p. 350-8.
99. Barbetti, V., et al., *Selective anti-leukaemic activity of low-dose histone deacetylase inhibitor ITF2357 on AML1/ETO-positive cells.* Oncogene, 2008. **27**(12): p. 1767-78.
100. Martens, J.H., et al., *PML-RARalpha/RXR Alters the Epigenetic Landscape in Acute Promyelocytic Leukemia.* Cancer cell, 2010. **17**(2): p. 173-85.
101. Wang, K., et al., *PML/RARalpha targets promoter regions containing PU.1 consensus and RARE half sites in acute promyelocytic leukemia.* Cancer cell, 2010. **17**(2): p. 186-97.
102. Dahlberg, A., C. Delaney, and I.D. Bernstein, *Ex vivo expansion of human hematopoietic stem and progenitor cells.* Blood, 2011. **117**(23): p. 6083-90.
103. Fortini, M.E., *Notch signaling: the core pathway and its posttranslational regulation.* Developmental cell, 2009. **16**(5): p. 633-47.
104. Varnum-Finney, B., et al., *Pluripotent, cytokine-dependent, hematopoietic stem cells are immortalized by constitutive Notch1 signaling.* Nature medicine, 2000. **6**(11): p. 1278-81.
105. Maillard, I., et al., *Canonical notch signaling is dispensable for the maintenance of adult hematopoietic stem cells.* Cell stem cell, 2008. **2**(4): p. 356-66.
106. Mancini, S.J., et al., *Jagged1-dependent Notch signaling is dispensable for hematopoietic stem cell self-renewal and differentiation.* Blood, 2005. **105**(6): p. 2340-2.
107. Saito T, C.S., Hirai H., *Analysis of Notch2 conditional knockout mice: Notch2 deficient bone marrow cells can reconstitute both to lymphoid and myeloid lineages [abstract]*. Blood, 2001. **98**(68a).
108. Ellisen, L.W., et al., *TAN-1, the human homolog of the Drosophila notch gene, is broken by chromosomal translocations in T lymphoblastic neoplasms.* Cell, 1991. **66**(4): p. 649-61.
109. Weng, A.P., et al., *Activating mutations of NOTCH1 in human T cell acute lymphoblastic leukemia.* Science, 2004. **306**(5694): p. 269-71.
110. Mercher, T., et al., *The OTT-MAL fusion oncogene activates RBPJ-mediated transcription and induces acute megakaryoblastic leukemia in a knockin mouse model.* The Journal of clinical investigation, 2009. **119**(4): p. 852-64.
111. Salat, D., et al., *ETO, but not leukemogenic fusion protein AML1/ETO, augments RBP-Jkappa/SHARP-mediated repression of notch target genes.* Molecular and cellular biology, 2008. **28**(10): p. 3502-12.

112. Alcalay, M., et al., *Acute myeloid leukemia fusion proteins deregulate genes involved in stem cell maintenance and DNA repair.* The Journal of clinical investigation, 2003. **112**(11): p. 1751-61.
113. Payton, J.E., et al., *High throughput digital quantification of mRNA abundance in primary human acute myeloid leukemia samples.* The Journal of clinical investigation, 2009. **119**(6): p. 1714-26.
114. Krumlauf, R., *Hox genes in vertebrate development.* Cell, 1994. **78**(2): p. 191-201.
115. Giampaolo, A., et al., *Key functional role and lineage-specific expression of selected HOXB genes in purified hematopoietic progenitor differentiation.* Blood, 1994. **84**(11): p. 3637-47.
116. Moretti, P., et al., *Identification of homeobox genes expressed in human haemopoietic progenitor cells.* Gene, 1994. **144**(2): p. 213-9.
117. Kawagoe, H., et al., *Expression of HOX genes, HOX cofactors, and MLL in phenotypically and functionally defined subpopulations of leukemic and normal human hematopoietic cells.* Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K, 1999. **13**(5): p. 687-98.
118. Pineault, N., et al., *Differential expression of Hox, Meis1, and Pbx1 genes in primitive cells throughout murine hematopoietic ontogeny.* Experimental hematology, 2002. **30**(1): p. 49-57.
119. Sauvageau, G., et al., *Differential expression of homeobox genes in functionally distinct CD34+ subpopulations of human bone marrow cells.* Proceedings of the National Academy of Sciences of the United States of America, 1994. **91**(25): p. 12223-7.
120. Sauvageau, G., et al., *Overexpression of HOXB4 in hematopoietic cells causes the selective expansion of more primitive populations in vitro and in vivo.* Genes & development, 1995. **9**(14): p. 1753-65.
121. Antonchuk, J., G. Sauvageau, and R.K. Humphries, *HOXB4 overexpression mediates very rapid stem cell regeneration and competitive hematopoietic repopulation.* Experimental hematology, 2001. **29**(9): p. 1125-34.
122. Amsellem, S., et al., *Ex vivo expansion of human hematopoietic stem cells by direct delivery of the HOXB4 homeoprotein.* Nature medicine, 2003. **9**(11): p. 1423-7.
123. Slape, C. and P.D. Aplan, *The role of NUP98 gene fusions in hematologic malignancy.* Leukemia & lymphoma, 2004. **45**(7): p. 1341-50.
124. Nakamura, T., *NUP98 fusion in human leukemia: dysregulation of the nuclear pore and homeodomain proteins.* International journal of hematology, 2005. **82**(1): p. 21-7.
125. Argiropoulos, B. and R.K. Humphries, *Hox genes in hematopoiesis and leukemogenesis.* Oncogene, 2007. **26**(47): p. 6766-76.
126. Ohta, H., et al., *Near-maximal expansions of hematopoietic stem cells in culture using NUP98-HOX fusions.* Experimental hematology, 2007. **35**(5): p. 817-30.
127. Bijl, J., et al., *Analysis of HSC activity and compensatory Hox gene expression profile in Hoxb cluster mutant fetal liver cells.* Blood, 2006. **108**(1): p. 116-22.

128. Brun, A.C., et al., *Hoxb4-deficient mice undergo normal hematopoietic development but exhibit a mild proliferation defect in hematopoietic stem cells.* Blood, 2004. **103**(11): p. 4126-33.
129. Bjornsson, J.M., et al., *Reduced proliferative capacity of hematopoietic stem cells deficient in Hoxb3 and Hoxb4.* Molecular and cellular biology, 2003. **23**(11): p. 3872-83.
130. Ono, R., et al., *Mixed-lineage-leukemia (MLL) fusion protein collaborates with Ras to induce acute leukemia through aberrant Hox expression and Raf activation.* Leukemia, 2009. **23**(12): p. 2197-209.
131. Golub, T.R., et al., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 1999. **286**(5439): p. 531-7.
132. Becker, H., et al., *Favorable prognostic impact of NPM1 mutations in older patients with cytogenetically normal de novo acute myeloid leukemia and associated gene- and microRNA-expression signatures: a Cancer and Leukemia Group B study.* Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 2010. **28**(4): p. 596-604.
133. Vassiliou, G.S., et al., *Mutant nucleophosmin and cooperating pathways drive leukemia initiation and progression in mice.* Nature genetics, 2011. **43**(5): p. 470-5.
134. Alcalay, M., et al., *Acute myeloid leukemia bearing cytoplasmic nucleophosmin (NPMc+ AML) shows a distinct gene expression profile characterized by up-regulation of genes involved in stem-cell maintenance.* Blood, 2005. **106**(3): p. 899-902.
135. Verhaak, R.G., et al., *Mutations in nucleophosmin (NPM1) in acute myeloid leukemia (AML): association with other gene abnormalities and previously established gene expression signatures and their favorable prognostic significance.* Blood, 2005. **106**(12): p. 3747-54.
136. Haferlach, C., et al., *AML with mutated NPM1 carrying a normal or aberrant karyotype show overlapping biologic, pathologic, immunophenotypic, and prognostic features.* Blood, 2009. **114**(14): p. 3024-32.
137. Campbell, K.H., et al., *Sheep cloned by nuclear transfer from a cultured cell line.* Nature, 1996. **380**(6569): p. 64-6.
138. Noggle, S., et al., *Human oocytes reprogram somatic cells to a pluripotent state.* Nature, 2011. **478**(7367): p. 70-5.
139. Park, I.H., et al., *Disease-specific induced pluripotent stem cells.* Cell, 2008. **134**(5): p. 877-86.
140. Hanna, J., et al., *Treatment of sickle cell anemia mouse model with iPS cells generated from autologous skin.* Science, 2007. **318**(5858): p. 1920-3.
141. Moretti, A., et al., *Patient-specific induced pluripotent stem-cell models for long-QT syndrome.* N Engl J Med. **363**(15): p. 1397-409.
142. Maehr, R., et al., *Generation of pluripotent stem cells from patients with type 1 diabetes.* Proc Natl Acad Sci U S A, 2009. **106**(37): p. 15768-73.
143. Robbins, R.D., et al., *Inducible pluripotent stem cells: not quite ready for prime time?* Curr Opin Organ Transplant. **15**(1): p. 61-7.

144. Csete, M., *Translational prospects for human induced pluripotent stem cells.* Regen Med. **5**(4): p. 509-19.
145. Amabile, G. and A. Meissner, *Induced pluripotent stem cells: current progress and potential for regenerative medicine.* Trends Mol Med, 2009. **15**(2): p. 59-68.
146. Wernig, M., et al., *c-Myc is dispensable for direct reprogramming of mouse fibroblasts.* Cell Stem Cell, 2008. **2**(1): p. 10-2.
147. Huangfu, D., et al., *Induction of pluripotent stem cells from primary human fibroblasts with only Oct4 and Sox2.* Nat Biotechnol, 2008. **26**(11): p. 1269-75.
148. Chang, C.W., et al., *Polycistronic lentiviral vector for "hit and run" reprogramming of adult skin fibroblasts to induced pluripotent stem cells.* Stem cells, 2009. **27**(5): p. 1042-9.
149. Carey, B.W., et al., *Reprogramming of murine and human somatic cells using a single polycistronic vector.* Proc Natl Acad Sci U S A, 2009. **106**(1): p. 157-62.
150. Shao, L., et al., *Generation of iPS cells using defined factors linked via the self-cleaving 2A sequences in a single open reading frame.* Cell Res, 2009. **19**(3): p. 296-306.
151. Stadtfeld, M., et al., *Induced pluripotent stem cells generated without viral integration.* Science, 2008. **322**(5903): p. 945-9.
152. Okita, K., et al., *Generation of mouse induced pluripotent stem cells without viral vectors.* Science, 2008. **322**(5903): p. 949-53.
153. Warren, L., et al., *Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA.* Cell Stem Cell. **7**(5): p. 618-30.
154. Zhou, H., et al., *Generation of induced pluripotent stem cells using recombinant proteins.* Cell Stem Cell, 2009. **4**(5): p. 381-4.
155. Yamanaka, S., *Elite and stochastic models for induced pluripotent stem cell generation.* Nature, 2009. **460**(7251): p. 49-52.
156. Manini, I., et al., *Multi-potent progenitors in freshly isolated and cultured human mesenchymal stem cells: a comparison between adipose and dermal tissue.* Cell and tissue research, 2011. **344**(1): p. 85-95.
157. Nakagawa, M., et al., *Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts.* Nature biotechnology, 2008. **26**(1): p. 101-6.
158. Huangfu, D., et al., *Induction of pluripotent stem cells from primary human fibroblasts with only Oct4 and Sox2.* Nature biotechnology, 2008. **26**(11): p. 1269-75.
159. Waddington, C., *The Strategy of the genes, a discussion of some aspects of theoretical biology*1957, London: Allen and Unwin.
160. Kim, K., et al., *Epigenetic memory in induced pluripotent stem cells.* Nature. **467**(7313): p. 285-90.
161. Lister, R., et al., *Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells.* Nature.
162. Pawlak, M. and R. Jaenisch, *De novo DNA methylation by Dnmt3a and Dnmt3b is dispensable for nuclear reprogramming of somatic cells to a pluripotent state.* Genes Dev. **25**(10): p. 1035-40.

163. Wernig, M., et al., *In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state.* Nature, 2007. **448**(7151): p. 318-24.
164. Park, I.H., et al., *Generation of human-induced pluripotent stem cells.* Nat Protoc, 2008. **3**(7): p. 1180-6.
165. Laurent, L.C., et al., *Dynamic changes in the copy number of pluripotency and cell proliferation genes in human ESCs and iPSCs during reprogramming and time in culture.* Cell Stem Cell. **8**(1): p. 106-18.
166. Hussein, S.M., et al., *Copy number variation and selection during reprogramming to pluripotency.* Nature. **471**(7336): p. 58-62.
167. Gore, A., et al., *Somatic coding mutations in human induced pluripotent stem cells.* Nature. **471**(7336): p. 63-7.

Chapter 2

Canoncial and non-canonical HOX expression patterns in acute myeloid leukemia

Abstract

HOX gene dysregulation occurs in acute myeloid leukemia (AML) samples with translocations involving HOX genes themselves (e.g. NUP98-HOXA9), and in samples with other disease-initiating translocations (e.g. MLL translocations and inv(16)). However, HOX genes are also highly expressed in many AML samples without translocations; the mechanism that causes "dysregulation" in these cases is unknown. We first analyzed the data from 50 sequenced *de novo* AML genomes and found no recurrent mutations within or near the HOX genes that were likely to cause HOX gene dysregulation. Expression array data from 190 AML cases revealed that all AML cases with MLL translocations express HOXA cluster genes only, that inv(16) cases express HOXB genes only, and that t(15;17) and t(8;21) cases express none of the HOX genes. However, most AML cases predominantly express *HOXA5, A9, A10, B2, B3*, and *MEIS1* in a canonical, highly coordinated fashion that is similar to that found in normal human CD34 cells (which contain the hematopoietic stem cell population). This pattern is strongly associated with *NPM1* and *DNMT3A* mutations, but it can be found in AML samples with neither mutation. HOX gene expression patterns are not correlated with DNA methylation patterns in the HOX gene clusters. HOX gene "dysregulation" in many cases of AML may therefore represent the persistence of a normal, stem cell-specific HOX gene expression pattern that is probably required for self-renewal, "captured" by mutations that initiate disease in the hematopoietic stem cell that was transformed.

Introduction

While there are well known chromosomal translocations associated with favorable outcomes (inv(16), t(18;21) and t(15;17)) and poor outcomes (11q23, t(6;9) and 7(q)) [1-6], most patients have no translocations, and fall into the intermediate risk category. Outcomes within the intermediate risk category are highly variable. Many recent studies have focused on improving the classification of this large subset of patients, since it could improve early treatment decisions; one of the genomic strategies used has been gene expression profiling with array-based approaches, in an attempt to find gene expression signatures that segregate patients by outcome [6-8]. A recurrent theme seen in these gene-expression studies is homeobox (HOX) gene dysregulation in many AML patients [6, 9-13].

HOX genes are a highly conserved family of transcription factors originally identified for their role in anterior-posterior axis positioning in animal models [14]. Their expression in hematopoietic progenitors suggested that they may also play a role in hematopoiesis, which has been studied extensively with overexpression and knockout models in the mouse [15-19]. Overexpression of *Hoxb4* in murine hematopoietic stem cells (HSCs) stimulates *in vivo* and *ex vivo* expansion of HSCs without promoting leukemia, demonstrating its ability to promote self-renewal [20-22]. Copeland and colleagues showed that *HoxA7* and *A9* are able to cooperate with the Hox co-factor *Meis1* to generate myeloid leukemia in mice [23]. Further, Nakamura and colleagues showed that the HOX genes are down-regulated during myeloid differentiation, implying that their role in leukemogenesis may be to block differentiation [24].

The leukemogenic potential of HOX genes is most clearly seen in rare translocations of nucleoporin 98 (*NUP98*) and HOX genes (such as *HOXA9*, *A10*, *C10* and *D13*) [25-28]. *NUP98-HOX* fusions replace the regulatory region of the HOX gene with the N-terminus of *NUP98*. Humphries et. al. have tested multiple NUP98-HOX fusion proteins for their ability to expand HSCs *ex vivo* [29, 30]. Like HoxB4 alone, NUP98-HoxB4 is a potent stimulator of HSC expansion, as are NUP98 fusions with many other HOX genes (although to varying degrees) [31]. A fusion of the homeobox domain alone with NUP98 (NUP98-HOXA10HD) showed an even greater ability to expand HSPCs than NUP98-HOXB4 (>2000-fold vs. 300-fold), suggesting that the homeodomain alone is able to induce self-renewal [32].

MLL translocations are also associated with a HOX expression phenotype. Although MLL has over 60 fusion partners, HOX genes are never directly involved in MLL translocations. MLL is a histone methyltransferase responsible for methylation of H3K4, and is regulates HOX gene expression during development [31, 33]. The dependence of MLL and HOX genes in hematopoiesis was illustrated by experiments showing impaired hematopoietic differentiation by *Mll* deficient embryoid bodies, which could be rescued with overexpression of HOX genes [34]. The HOX overexpression phenotype seen in MLL translocations is most likely an effect of altered MLL function in these novel fusions [9, 35].

Mutations in nucleophosmin (*NPM1*) are very common in AML, and are also associated with high levels of HOX gene expression, even though NPM1 is neither a transcription factor nor an epigenetic regulator. Knockin mice expressing a mutant human *NPM1* gene (*NPM1c*) exhibited overexpression of mouse HOX genes in hematopoietic cells, showing that the *NPM1* mutation can somehow lead to dysregulated HOX expression [36]. More recently, mutations in the *de novo* DNA methyltransferase *DNMT3A* were shown to be highly recurrent in AML [37, 38]. Yan, *et al*. reported that these mutations are associated with hypomethylation and overexpression of *HOXB2* and *HOXB3* [38]. While mutations in both genes have been reported to affect HOX gene expression patterns, the *NPM1* mutation is associated with favorable outcomes, while *DNMT3A* mutations are associated with poor prognosis [37-41]. These studies suggest that HOX overexpression is not governed by a simple mutational pattern, and it does predict outcomes *per se*.

In this study, we analyzed a cohort of 190 carefully genotyped patients to better understand/define the HOX expression phenotype in AML. We found that the majority of patients express a subset of HOX genes (*HOXA5, A9, A10, B2, B3*, and *MEIS1*) at high levels. Cases with t(8;21) and t(15;17) minimally express HOX genes; patients with MLL translocations only express HOXA family members, and inv(16) cases express only HOXB family members . HOX gene expression levels do not correlate directly with DNA methylation patterns in the HOX gene clusters, nor are they supervised by any of the common recurrent AML mutations. Surprisingly, we found that the canonical HOX expression pattern seen in most AML cases is essentially identical to that of human CD34 cells; this data suggests that the HOX expression pattern in most cases of AML without

translocations is not actually dysregulated. Instead, HOX expression may represent the 'capture' of a normal pattern of HOX gene regulation in HSCs that is responsible for self-renewal, and subverted by mutations that initiate AML within this cellular compartment.

Materials and Methods

Human AML and normal sorted bone marrow samples

190 de novo adult AML bone marrow aspirates were analyzed. Patient selection has been previously described [42]; patient characteristics are listed (including recurrent mutations) in Table 2-1. Bone marrow aspirates were also obtained from healthy adult donors. This study was approved by the Human Research Protection Office at Washington University School of Medicine after patients and donors provided informed consent in accordance with the Declaration of Helsinki. Isolation of normal promyelocytes and neutrophils were performed as previously described [43]. Briefly, CD9-, CD14-, CD15hi, and CD16lo promeylocytes and CD9-, CD14-, CD15hi, and CD16hi neutrophils were isolated by high-speed cell sorting. CD34+ cells were isolated by MACS sorting according to the manufacturer's instructions (Miltenyi Biotec, Bergisch Gladbach, Germany).

DNA and RNA Isolation and Purification

DNA was extracted from AML bone marrow aspirates using the QIAamp DNA Mini Kit according to the manufacturer's instructions (Qiagen, Valencia, CA). DNA from all samples was quantified using UV spectroscopy (Nanodrop Technologies, Wilmington, DE). RNA from AML bone marrow aspirates was prepared from unfractionated snap-frozen cell pellets using Trizol reagent. For the sorted healthy bone marrow samples, sufficient cells were collected to perform the standard 1-cycle in vitro transcription protocol; avoiding the bias introduced by linear amplification (2-cycle) required for smaller amounts of RNA. Sorted cells were lysed in Trizol reagent (Invitrogen, Carlsbad,

CA) and stored at -80^{o}C until RNA purification. RNA from all samples was quantified using UV spectroscopy (Nanodrop Technologies, Wilmington, DE) and qualitatively assessed using a BioAnalyzer 2100 and RNA NanoChip assay (Agilent Technologies, Santa Clara, CA).

Analysis of RNA from AML and healthy cell populations

Gene chip analysis of all samples was done as previously described [44]. Samples were labeled and hybridized to Affymetrix Human Genome U133 Plus 2 Array GeneChip Microarrays (Affymetrix, Santa Clara, CA) using standard protocols. For 70 of the 190 patients, the Affymetrix data was orthogonally validated on the NanoString nCounter platform (see Table 2-1 for patient identifiers). Details of the nCounter Analysis System (NanoString Technologies) were reported previously [45]. Briefly, two sequence-specific probes were generated for each gene of interest (Table 2-4). The probes were complementary to a 100-base region of the target mRNA. One probe was covalently linked to an oligonucleotide containing biotin (the capture probe), and the other was linked to a color-coded molecular tag (the reporter probe). The nCounter CodeSet for these studies contained probe pairs for 46 test and control genes (data not shown for control genes). Each sample was hybridized in duplicate with 100 ng of total RNA in each reaction. All 46 genes were assayed simultaneously in multiplexed reactions. To account for slight differences in hybridization and purification efficiency, the raw data were normalized to OAZ1 (one of the control genes). Spearman correlations were calculated using SAS 9.1 statistical software to determine the correlation of NanoString and Affymetrix data, values are reported in Table 2-4.

Analysis of Affymetrix U133 Plus 2 array data

Spotfire software was used to generate heatmaps from the raw expression values for the best probe set of each HOX gene (defined as the probe set with the highest average expression level across all 190 samples- listed in Table 2-4). Samples were organized by translocations, mutations, and HOX expression levels as indicated. For heatmaps, maximum expression was set to a raw expression value 10,000 (red) and minimum was set to 0 (green), with a midpoint of 5000 (black). To correlate expression of genes with HOXA9, in an attempt to identify a potential upstream regulator, expression levels of all 54,613 probesets were compared to HOXA9 expression in all 190 samples.

Measuring 5-methylcytosine content by mass spectrometry.

Analysis was done as previously described [37]. Briefly, Samples were analyzed by liquid chromatography-tandem mass spectrometry (LC-MS/MS) using a Shimadzu SCL-10A VP HPLC system (Columbia, MD) coupled with a 4000 Q-Trap (MS/MS) mass spectrometer (Applied Biosystems, Foster City, CA). The Analyst software package (Version 1.4.2, Applied Biosystems, Foster City, CA) was used for instrument operation, data acquisition and analysis. Standard curves were generated using commercially available deoxyribonucleoside compounds (5-methyl-2'-deoxycytidine monophosphate disodium salt (mdCMP), 2'-deoxyguanosine-5'-monophosphate disodium salt (dGMP), 2'-deoxyadenosine-5'-monophosphate free acid (dAMP), 2'-deoxycytidine monophosphate sodium salt (dCMP), thymidine (T), 2'-deoxyadenosine monohydrate

(dA), 2'-deoxycyidine (dC) (USB Corporation, Cleveland, OH and Sigma Aldrich, St. Louis, MO). The LC-MS-MS analysis for patient samples was performed using 3.5-5 ng/μL of DNA hydrolysis products. The percentage of methylation was determined by either dividing the measured amount of mdCMP (in fmols) by the amount of dGMP in each sample or by dividing the sum of mdCMP and dCMP detected in each sample.

Bisulfite conversion of genomic DNA for methylation analysis

Cytosine residues in DNA samples (1 μg) were converted to thymidine with bisulfite using the Zymo Research EZ96 DNA methylation kit, according to the manufacturer's instructions (Zymo Research, Irvine, CA, USA). The completeness of bisulfite conversion and the amount of bisulfite-converted DNA for each sample was determined using a panel of four MethyLight-based quality control (QC) reactions as described previously [46]. All samples passed these QC tests and subsequently were entered into the Illumina Infinium DNA methylation data production pipeline.

Infinium DNA methylation profiling

The Illumina Infinium HumanMethylation450 DNA methylation assay was performed for all 190 samples according to the experimental protocol outlined by the manufacturer (Illumina, San Diego, CA, USA; User Guide part #15019519 A). The Illumina Infinium DNA methylation assay interrogates the DNA methylation status of 485,577 CpG dinucleotides as described [47]. The genomic characteristics of each probe are available for download via Illumina (ftp.illumina.com); our annotations are based upon version 1.2 of the HumanMethylation450 manifest. BeadChips were coated and imaged on an

Illumina iScan station. Binary IDAT files produced by the scanner were then processed directly using a development version of the Bioconductor *methylumi* package (Triche, 2011).

Data extraction and analysis. Analytic and control probe average intensities, as well as the number of beads registered for each probe, were extracted directly from IDAT files produced by the scanner. Background correction was performed via normal-exponential convolution on out-of-band intensities from 135,501 of the 485,577 analytic probes, taken to estimate within-array background [48]. An offset of +15 was employed as recommended by Shi, *et al.* [48]. Following background correction, beta values for each probe were computed as the fraction of the total intensity contributed by the methylated probe allele (M/(M+U)) for each of the 485,577 probes on the array. Non-detection probabilities (p-values) were computed using the empirical cumulative density function of the negative control probes in the appropriate channel for each allele [48]. Probes with fewer than three registered beads, or greater than a 0.01 (> 1%) non-detection probability for both alleles, were flagged as N/A and omitted from further analysis. Probes interrogating a CpG locus within 10 base pairs of a known single nucleotide polymorphism (SNP) were also omitted from analysis.

Data submission. Raw IDAT files and processed data for 190 patients with matching DNA methylation and gene expression data were deposited with the Data Coordination Center for the Cancer Genome Atlas project and are available from the TCGA Data Portal: http://tcgadata.nci.nih.gov/tcga/dataAccessMatrix.htm?mode=ApplyFilter&show

Matrix=true&diseaseType=LAML&availability=A&tumorNormal=TN&tumorNormal=T&platformType=2.

Results

Somatic mutations within the HOX gene clusters do not account for HOX gene overexpression in AML

To determine if HOX overexpression in AML is due to somatic mutations within or near the HOX genes, we analyzed the HOX gene cluster region from the whole genome sequencing results of 45 *de novo* AML samples with either normal karyotypes (n=35), MLL translocations (n=2) or t(15;17) (n=8). These patients represented a range of FAB subtypes, patient ages, and survivals (full patient details are in **Table 1**) as well as a wide range of HOX gene expression levels on the AffyMetrix U133+2 array (**Figure 1**). We defined somatic mutations in each of the four HOX gene clusters by comparing the sequence data from the bone marrow DNA to normal skin DNA, in a region that spanned from 100 kb upstream from the 5'- most gene to 100 kb downstream from the 3'- most gene. Ten of 45 samples contained somatic mutations within the HOX clusters (**Table 2**). The 10 patients with mutations are indicated in **Figure 1**, which shows that the mutations are not associated with the HOX gene expression phenotype. One sample had two mutations within 70 kb of each other in the HOXC cluster; both were in non-coding regions. The mutation closest to a gene was found 1,586 bp upstream from *HOXA13*. Although it is possible that some of these mutations could affect previously undefined regulatory regions, none were recurrent, and none were directly associated with patterns of HOX gene expression. None of the mutations were found within coding regions, and therefore none had translational consequences.

A canonical pattern of HOX gene expression in AML

As opposed to mouse models of AML, and NUP98-HOX fusion cases of human AML where single HOX genes are overexpressed, the data shown in **Figure 1** suggests that there is a specific set of HOXA and HOXB cluster genes that are highly expressed in some AML cases (predominantly *HOXA5, A9, A10, B2, B3*, and *MEIS1)*. Notably, HOX C and D cluster genes are rarely expressed in AML samples. To determine whether these patterns were restricted to normal karyotype cases, we increased our analysis to a total of 190 AML patients with a wide variety of cytogenetic and clinical features (**Table 1**). **Figure 2** and **Table 3** reveal the same canonical pattern of HOX gene expression in many of the samples in this larger set, with notable exceptions described below.

Due to the homology of homeobox domains in HOX genes, probe cross-hybridization could potentially explain the highly correlated expression patterns.

The HOX genes are most highly related among the paralog groups (i.e. *HOXA9*, *HOXB9*, *HOXC9* and *HOXD9*) and **Figure 1** shows that the Affymetrix probes can clearly distinguish among the expression levels of these genes. Regardless, we measured the expression levels of all HOX genes (and *MEIS1*) on an orthogonal platform (the NanoString nCounter platform) for 70 of the 190 AML samples; these values were highly correlated between the platforms for all HOX genes (**Table 4**). These data suggest that the Affymetrix U133 Plus 2 platform accurately measures patterns of HOX gene expression.

Unique patterns of HOX gene expression associated with AML translocations

Since AML-associated translocations have been linked with HOX gene overexpression, we compared patterns of HOX gene expression to cytogenetic and mutational data from the 190 AML cases. **Figure 4A** contains a heatmap that shows that samples with several well-defined AML translocations have unique HOX gene expression patterns; the actual expression values for each gene and patient are plotted in **Figure 4B.** Remarkably, none of the cases with t(15;17) or t(8;21) express any of the HOX genes. In contrast, cases with inv(16) express relatively low levels of only *HOXB2*, *HOXB3,* and *MEIS1*. As previously reported, MLL translocation cases uniformly overexpress *HOXA5*, *HOXA9*, and *HOXA10*, along with *MEIS1*; *HOX B* genes are not expressed in these cases, however. Patients with the MLL-PTD mutation express both HOXA and HOXB cluster genes in the canonical pattern seen with normal karyotype cases, suggesting that these two mutation types may have different operational mechanisms.

A variety of mutations associated with AML have also been reported to have dysregulated HOX gene expression. We therefore correlated HOX gene expression patterns with several common AML mutations in **Figure 5**. As previously reported, all AML cases with the NPM1 mutation exhibit the canonical HOX expression phenotype [39, 41, 49]. *DNMT3A* mutations, on the other hand, do not fully correlate with the phenotype; only 11/25 (44%) of patients with *DNMT3A* mutations alone have the canonical HOX expression pattern (for this analysis, the canonical HOX expression pattern was defined as raw values for *HOXA9* and *HOXB3* greater than 5,000 on the Affymetrix U133 Plus 2 array platform). *DNMT3A* mutations are positively correlated

with *NPM1* mutations [37]; if the doubly mutant cases are taken into account, the canonical HOX phenotype is detected in 33 of 47 cases (70%). *FLT3*, *IDH1/2* and *TET2* mutations all exhibit ~65% of cases with the canonical HOX phenotype (including patients with mutations in *NPM1* and/or *DNMT3A* mutations). While this data reveals positive correlations of the canonical HOX phenotype with some mutations, it also shows that some cases have the phenotype despite having none of the correlated mutations.

DNA methylation patterns near the HOX genes do not correlate with expression

DNMT3A, IDH1/2 and TET2 mutations have all been suggested to have effects on DNA methylation [38, 50, 51]; since there have been conflicting reports on the affect of these mutations on global DNA methylation, we first measured total 5-methylcytosine content by mass spectrometry in 70 selected samples [37]. Our data shows no significant association of 5-methylcytosine levels with any of the methylation-associated mutations (**Figure 6**). To determine whether the methylation status of the HOX gene clusters is associated with HOX expression levels, as suggested by Yan, *et al.*, we examined data from the Infinium Human Methylation 450 BeadChip (Illumina, San Diego, CA) for the same 190 patients for which expression data was available. **Figure 7** shows methylation data across the entire HOXA gene cluster (there are 542 CpG probes on the array that fall within this locus). The map was organized by the expression level of *HOXA9,* with lowest levels at the top, and highest at the bottom; there is no corresponding methylation gradient at the *HOXA9* promoter, or anywhere else in the *HOXA* gene cluster. The methylation array contained 450 CpG probes within the HOXB gene cluster; a minor hypomethylation phenotype was detected upstream from the *HOXB3* gene in some

samples with high levels of *HOXB3* expression (**Figure 8**; hypomethylation was defined as an average beta-value of less than 40% across the region of interest. This hypomethylation pattern is seen in 56% (36/64) of the top third of samples ranked on *HOXB3* expression. However, the hypomethylated samples do not cluster apart from the methylated samples, suggesting that hypomethylation is not the direct cause of *HOXB3* gene overexpression. By organizing the methylation data based on the common AML somatic mutations, there is no evidence of a phenotype in any of the 4 HOX gene clusters that is supervised by any of the common AML mutations (**Figure 9**).

The canonical HOX expression phenotype is also detected in normal human CD34 cells

Since neither mutations nor methylation patterns explain the canonical HOX phenotype, we next analyzed HOX gene expression patterns in normal hematopoietic cells. We analyzed expression array data from CD34+ cells, promyelocytes, and neutrophils purified from the bone marrow cells of five healthy volunteers [44]. Normal human CD34 cells express the same HOX genes as AML cells with the canonical phenotype (**Figure 10**); HOXA and B gene expression was downregulated in promyelocytes and neutrophils. To compare relative levels of expression in CD34 cells and AML samples, we normalized the expression of each gene to 100% for *HOXA9* (which is always the maximally expressed gene) in all samples (**Figure 11**). Remarkably, the relative expression of all HOX genes with respect to *HOXA9* is essentially the same in CD34 cells and AML cells.

Potential mechanisms underlying the canonical HOX expression phenotype

Since the HOXA and HOXB gene clusters are on different chromosomes, their coexpression in AML samples suggests that one or more *trans*-acting factors coregulate their expression. To identify potential *trans*-acting regulators *in silico*, we searched for common transcription factor motifs unique to the promoter regions of the expressed HOX genes. No common motifs were identified within 2,000 bp of *HOXA5*, *HOXA9*, *HOXA10*, *HOXB2*, *HOXB3* and *MEIS1* that were absent from the other HOX gene promoters (data not shown). We then analyzed all probesets on the AffyMetrix U133+2 array to identify genes with expression levels that were significantly correlated *HOXA9*, using a Spearman's correlation test. Of the 54,000+ probesets on the array, the only genes that were significantly coregulated were *HOXA9*, *HOXA5*, *HOXA10*, *HOXB2*, *HOXB3* and *MEIS1* (data not shown).

To determine whether clues regarding HOX gene regulation could be gleaned from mouse hematopoietic cells (which are easier to experimentally manipulate), we generated expression array data using the Affymetrix mouse Exon 1.0 array platform. We purified SLAM cells, KLS cells, promyelocytes, and neutrophils from at least three different young C57Bl/6 mice, and examined the *HOX* gene expression patterns from the arrays (**Figure 12**). Although the *Meis1* gene is developmentally regulated in a pattern that mimics human hematopoietic development, the *Hoxa* and *Hoxb* genes are not downregulated during terminal myeloid differentiation, and the pattern of expression is very different from that of human CD34 or AML cells. Unlike patients with AML FAB M3 and the t(15;17) translocation (which express none of the HOX genes), our mCG-

PML-RARA mouse model of APL expresses Hox genes in a pattern that is similar to that of normal progenitors (**Figure 13**).

Discussion

In this study, we explored HOX gene expression levels in a diverse group of AML patients, and found that four patterns exist based on cytogenetic findings: 1) Translocations that have no detectable effects on HOX gene expression (i.e. t(8;21) and t(15;17)) 2) Translocations associated with exclusive expression of either *HOXA* genes (MLL translocations) or *HOXB* genes (inv(16)); 3) Normal karyotype with a canonical pattern of expression of *HOXA5, A9, A10, B2, B3*, and the HOX co-factor *MEIS1*; 4) Normal karyotype with HOX gene expression below the level of detection. None of the common AML mutations supervise the canonical pattern of HOX gene expression, and DNA methylation patterns are not consistently correlated with HOX gene expression patterns. While many groups have reported HOX gene "dysregulation" in AML [9, 11, 12, 17, 23, 39, 40, 52-56], our studies clearly show that AML cases with the canonical pattern of HOX gene expression have the same pattern as that of normal human CD34 cells. These findings suggest that "dysregulation" in these AML cases more likely represents the "capture" of a normal pattern of HOX gene expression that is associated with the self-renewal properties of HSCs. The preservation of the HOX gene program in these cells may serve a mechanism to preserve self-renewal, an essential property of leukemia initiating cells.

Acquisition of self-renewal is thought to be a necessary step for the initiation of leukemia. Multiple pathways have been implicated in self-renewal in HSCs, including the HOX genes, along with the Notch, Hedgehog, Wnt, and FOXO signaling pathways [57]. Leukemia initiating fusion genes created by translocations, such as t(8;21) and t(15;17)

have been shown to directly induce self-renewal in mouse models [58, 59]; clearly, these translocations cause self-renewal via Hox-independent mechanisms. In support of this notion, our group has now shown that *PML-RARA* may initiate self-renewal program through the Notch pathway, since dominant negative MAML can abrogate the serial replating phenotype induced by *PML-RARA* (N. Grieselhuber, *et. al*. in preparation). We further analyzed the gene expression patterns of the four non-HOX self-renewal pathways to see if they are differentially expressed in normal karyotype patients with and without HOX expression; the majority of genes in these pathways are not expressed in AML samples. The genes that are expressed show no difference between the HOX positive and negative normal karyotype samples, suggesting that these pathways are not responsible for self-renewal in the HOX negative samples. A complete understanding of the genetic and epigenetic pathways in a large number of AML cases will be needed to identify all of the self-renewal pathways that are activated in this disease.

As opposed to the highly conserved t(15;17) translocation, MLL mutations represent a heterogeneous group of AML-associated mutations with variable HOX expression levels. While the MLL-PTD mutation is associated with the canonical expression pattern of *HOXA* and *HOXB* genes seen in the majority of cases of AML, MLL translocations are associated with overexpression of only the *HOXA5, A9,* and *A10* genes. This implies that the MLL fusion proteins have a restricted set of HOX target genes, compared to constitutively active MLL. Since MLL is a known upstream regulator of HOX genes, the link between MLL mutations and HOX gene overexpression is thought to be direct. However, most AML samples with HOX gene overexpression have no mutations

affecting MLL. Although other genes that have been reported to regulate HOX expression (e.g. *MOZ*, *CDX2*, and polycomb gene complex members), no mutations have been found in these genes in our set of sequenced AML samples (TJ Ley, *et al*., unpublished) and their expression does not correlate with HOX gene overexpression patterns in our samples [26, 60-63].

Finding the canonical HOX expression phenotype in a large subset of patients harboring *DNMT3A, IDH1/2 and/or TET2* mutations suggested a potential role for DNA methylation in controlling HOX gene expression patterns. However, global methylcytosine levels are not altered in patients with these mutations, and the methylation array data did not reveal a correlation between these mutations and CpG methylation in the HOX gene clusters. Our data conflicts with the report from Yan *et. al*., who demonstrated hypomethylation at 5 CpG dinucleotides in the *HOXB2* promoter of 17 *DNMT3A* mutated samples. In this study, we used the Infinium array to examine 992 CpG residues in the *HOXA* and *HOXB* loci of 47 AML samples with *DNMT3A* mutations and 143 samples without [38]. Although we did detect a small group of samples with hypomethylated regions in the *HOXB2* promoter, this finding was not restricted to *DNMT3A* mutant cases, and it was not linked directly to *HOXB2* overexpression. Although it is very unlikely that DNA methylation plays a major role in the regulation of HOX gene expression patterns, we cannot rule out a role for histone modifications; indeed, two of the known regulators of HOX genes, MLL and MOZ, modify histones and may directly affect HOX expression patterns [61]. A careful study of histone modifications in the HOX loci of primary AML cells may help to identify the key

upstream modifiers, if adequate numbers of viable cells can be obtained for CHiP-Seq analysis.

The coordinated, canonical expression of HOXA and HOXB genes, which lie on separate chromosomes, suggests the presence of an upstream transcriptional regulator. We performed an *in silico* analysis to identify a common transcription factor binding motif for the expressed HOX genes, but none was identified. We searched for AML genes that were coregulated with *HOXA9* in an attempt to find an upstream transcription factor, but found only the expressed HOX genes themselves (*A9, A5, A10, B2, B3, and MEIS1*). These data suggest that an autoregulatory loop may govern the expression of this subset of HOX genes. Although Hox gene autoregulation has previously been demonstrated in Drosophila, it has not yet been proven to occur in higher organisms [64]. Unfortunately, mice and humans express different subsets of HOX genes in hematopoietic cells, and regulate HOX genes differently during hematopoietic development; modeling an autoregulatory loop that mimics the human situation may therefore be difficult to achieve in murine systems.

Although HOX gene overexpression in AML has been appreciated for years, we have used genomic data to establish the differences between expression patterns in cases with and without common AML translocations; we have also defined a canonical HOX gene expression pattern in AML that is also found in normal human CD34 cells. Considering how common the "HOX phenotype" is in NK-AML cases, these results suggest that the

leukemia-initiating cell may often be a transformed HSC that retains its original HOX expression pattern. Since murine hematopoietic cells express different HOX genes in different patterns, modeling this phenomenon experimentally will be difficult. Although the mechanism underlying the HOX hematopoietic phenotype is still unclear, these data clarify the circumstances and patterns of HOX gene expression in many cases of AML. These data also show that it may be very difficult to differentially target the HOX pathway in AML, since it reflects the normal pattern of HOX gene expression in HSCs, where elimination of self-renewal would be expected to have disastrous consequences. Much additional work will be needed to identify all the genetic and epigenetic events that cause self-renewal in AML, and whether some will represent bona fide targets for novel therapeutic approaches.

References

1. 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, 2002. **100**(13): p. 4325-36.
2. Grimwade, D., et al., *The predictive value of hierarchical cytogenetic classification in older adults with acute myeloid leukemia (AML): analysis of 1065 patients entered into the United Kingdom Medical Research Council AML11 trial.* Blood, 2001. **98**(5): p. 1312-20.
3. 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, 1998. **92**(7): p. 2322-33.
4. Lowenberg, B., J.R. Downing, and A. Burnett, *Acute myeloid leukemia.* The New England journal of medicine, 1999. **341**(14): p. 1051-62.
5. Slovak, M.L., et al., *Karyotypic analysis predicts outcome of preremission and postremission therapy in adult acute myeloid leukemia: a Southwest Oncology Group/Eastern Cooperative Oncology Group Study.* Blood, 2000. **96**(13): p. 4075-83.
6. Valk, P.J., et al., *Prognostically useful gene-expression profiles in acute myeloid leukemia.* The New England journal of medicine, 2004. **350**(16): p. 1617-28.
7. Haferlach, T., et al., *Global approach to the diagnosis of leukemia using gene expression profiling.* Blood, 2005. **106**(4): p. 1189-98.
8. Olesen, L.H., et al., *Molecular typing of adult acute myeloid leukaemia: significance of translocations, tandem duplications, methylation, and selective gene expression profiling.* British journal of haematology, 2005. **131**(4): p. 457-67.
9. Armstrong, S.A., et al., *MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia.* Nature genetics, 2002. **30**(1): p. 41-7.
10. Bullinger, L., et al., *Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia.* The New England journal of medicine, 2004. **350**(16): p. 1605-16.
11. Drabkin, H.A., et al., *Quantitative HOX expression in chromosomally defined subsets of acute myelogenous leukemia.* Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K, 2002. **16**(2): p. 186-95.
12. Golub, T.R., et al., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 1999. **286**(5439): p. 531-7.
13. Yeoh, E.J., et al., *Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling.* Cancer cell, 2002. **1**(2): p. 133-43.
14. Krumlauf, R., *Hox genes in vertebrate development.* Cell, 1994. **78**(2): p. 191-201.

15. Giampaolo, A., et al., *Key functional role and lineage-specific expression of selected HOXB genes in purified hematopoietic progenitor differentiation.* Blood, 1994. **84**(11): p. 3637-47.
16. Moretti, P., et al., *Identification of homeobox genes expressed in human haemopoietic progenitor cells.* Gene, 1994. **144**(2): p. 213-9.
17. Kawagoe, H., et al., *Expression of HOX genes, HOX cofactors, and MLL in phenotypically and functionally defined subpopulations of leukemic and normal human hematopoietic cells.* Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K, 1999. **13**(5): p. 687-98.
18. Pineault, N., et al., *Differential expression of Hox, Meis1, and Pbx1 genes in primitive cells throughout murine hematopoietic ontogeny.* Experimental hematology, 2002. **30**(1): p. 49-57.
19. Sauvageau, G., et al., *Differential expression of homeobox genes in functionally distinct CD34+ subpopulations of human bone marrow cells.* Proceedings of the National Academy of Sciences of the United States of America, 1994. **91**(25): p. 12223-7.
20. Sauvageau, G., et al., *Overexpression of HOXB4 in hematopoietic cells causes the selective expansion of more primitive populations in vitro and in vivo.* Genes & development, 1995. **9**(14): p. 1753-65.
21. Antonchuk, J., G. Sauvageau, and R.K. Humphries, *HOXB4 overexpression mediates very rapid stem cell regeneration and competitive hematopoietic repopulation.* Experimental hematology, 2001. **29**(9): p. 1125-34.
22. Amsellem, S., et al., *Ex vivo expansion of human hematopoietic stem cells by direct delivery of the HOXB4 homeoprotein.* Nature medicine, 2003. **9**(11): p. 1423-7.
23. Nakamura, T., et al., *Cooperative activation of Hoxa and Pbx1-related genes in murine myeloid leukaemias.* Nature genetics, 1996. **12**(2): p. 149-53.
24. Fujino, T., et al., *Inhibition of myeloid differentiation by Hoxa9, Hoxb8, and Meis homeobox genes.* Experimental hematology, 2001. **29**(7): p. 856-63.
25. Borrow, J., et al., *The t(7;11)(p15;p15) translocation in acute myeloid leukaemia fuses the genes for nucleoporin NUP98 and class I homeoprotein HOXA9.* Nature genetics, 1996. **12**(2): p. 159-67.
26. Frohling, S., et al., *HOX gene regulation in acute myeloid leukemia: CDX marks the spot?* Cell cycle, 2007. **6**(18): p. 2241-5.
27. Pineault, N., et al., *Differential and common leukemogenic potentials of multiple NUP98-Hox fusion proteins alone or with Meis1.* Molecular and cellular biology, 2004. **24**(5): p. 1907-17.
28. Raza-Egilmez, S.Z., et al., *NUP98-HOXD13 gene fusion in therapy-related acute myelogenous leukemia.* Cancer research, 1998. **58**(19): p. 4269-73.
29. Slape, C. and P.D. Aplan, *The role of NUP98 gene fusions in hematologic malignancy.* Leukemia & lymphoma, 2004. **45**(7): p. 1341-50.
30. Nakamura, T., *NUP98 fusion in human leukemia: dysregulation of the nuclear pore and homeodomain proteins.* International journal of hematology, 2005. **82**(1): p. 21-7.

31. Argiropoulos, B. and R.K. Humphries, *Hox genes in hematopoiesis and leukemogenesis.* Oncogene, 2007. **26**(47): p. 6766-76.
32. Ohta, H., et al., *Near-maximal expansions of hematopoietic stem cells in culture using NUP98-HOX fusions.* Experimental hematology, 2007. **35**(5): p. 817-30.
33. Kouzarides, T., *Chromatin modifications and their function.* Cell, 2007. **128**(4): p. 693-705.
34. Ernst, P., et al., *An Mll-dependent Hox program drives hematopoietic progenitor expansion.* Current biology : CB, 2004. **14**(22): p. 2063-9.
35. Perkins, A., et al., *Homeobox gene expression plus autocrine growth factor production elicits myeloid leukemia.* Proceedings of the National Academy of Sciences of the United States of America, 1990. **87**(21): p. 8398-402.
36. Vassiliou, G.S., et al., *Mutant nucleophosmin and cooperating pathways drive leukemia initiation and progression in mice.* Nature genetics, 2011. **43**(5): p. 470-5.
37. Ley, T.J., et al., *DNMT3A mutations in acute myeloid leukemia.* The New England journal of medicine, 2010. **363**(25): p. 2424-33.
38. Yan, X.J., et al., *Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia.* Nature genetics, 2011. **43**(4): p. 309-15.
39. Becker, H., et al., *Favorable prognostic impact of NPM1 mutations in older patients with cytogenetically normal de novo acute myeloid leukemia and associated gene- and microRNA-expression signatures: a Cancer and Leukemia Group B study.* Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 2010. **28**(4): p. 596-604.
40. Haferlach, C., et al., *AML with mutated NPM1 carrying a normal or aberrant karyotype show overlapping biologic, pathologic, immunophenotypic, and prognostic features.* Blood, 2009. **114**(14): p. 3024-32.
41. Verhaak, R.G., et al., *Mutations in nucleophosmin (NPM1) in acute myeloid leukemia (AML): association with other gene abnormalities and previously established gene expression signatures and their favorable prognostic significance.* Blood, 2005. **106**(12): p. 3747-54.
42. Link, D.C., et al., *Distinct patterns of mutations occurring in de novo AML versus AML arising in the setting of severe congenital neutropenia.* Blood, 2007. **110**(5): p. 1648-55.
43. Grenda, D.S., et al., *Mutations of the ELA2 gene found in patients with severe congenital neutropenia induce the unfolded protein response and cellular apoptosis.* Blood, 2007. **110**(13): p. 4179-87.
44. Payton, J.E., et al., *High throughput digital quantification of mRNA abundance in primary human acute myeloid leukemia samples.* The Journal of clinical investigation, 2009. **119**(6): p. 1714-26.
45. Geiss, G.K., et al., *Direct multiplexed measurement of gene expression with color-coded probe pairs.* Nature biotechnology, 2008. **26**(3): p. 317-25.
46. Campan, M., et al., *MethyLight.* Methods in molecular biology, 2009. **507**: p. 325-37.

47. Bibikova, M., et al., *High density DNA methylation array with single CpG site resolution.* Genomics, 2011. **98**(4): p. 288-95.
48. Shi, W., A. Oshlack, and G.K. Smyth, *Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips.* Nucleic acids research, 2010. **38**(22): p. e204.
49. Alcalay, M., et al., *Acute myeloid leukemia bearing cytoplasmic nucleophosmin (NPMc+ AML) shows a distinct gene expression profile characterized by up-regulation of genes involved in stem-cell maintenance.* Blood, 2005. **106**(3): p. 899-902.
50. Figueroa, M.E., et al., *Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation.* Cancer cell, 2010. **18**(6): p. 553-67.
51. Ko, M., et al., *Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2.* Nature, 2010. **468**(7325): p. 839-43.
52. Andreeff, M., et al., *HOX expression patterns identify a common signature for favorable AML.* Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K, 2008. **22**(11): p. 2041-7.
53. Casas, S., et al., *Aberrant expression of HOXA9, DEK, CBL and CSF1R in acute myeloid leukemia.* Leukemia & lymphoma, 2003. **44**(11): p. 1935-41.
54. Lawrence, H.J., et al., *Frequent co-expression of the HOXA9 and MEIS1 homeobox genes in human myeloid leukemias.* Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K, 1999. **13**(12): p. 1993-9.
55. Rice, K.L. and J.D. Licht, *HOX deregulation in acute myeloid leukemia.* The Journal of clinical investigation, 2007. **117**(4): p. 865-8.
56. Roche, J., et al., *Hox expression in AML identifies a distinct subset of patients with intermediate cytogenetics.* Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K, 2004. **18**(6): p. 1059-63.
57. Heidel, F.H., B.G. Mar, and S.A. Armstrong, *Self-renewal related signaling in myeloid leukemia stem cells.* International journal of hematology, 2011. **94**(2): p. 109-17.
58. Tonks, A., et al., *The AML1-ETO fusion gene promotes extensive self-renewal of human primary erythroid cells.* Blood, 2003. **101**(2): p. 624-32.
59. Welch, J.S., W. Yuan, and T.J. Ley, *PML-RARA can increase hematopoietic self-renewal without causing a myeloproliferative disease in mice.* The Journal of clinical investigation, 2011. **121**(4): p. 1636-45.
60. Hanson, R.D., et al., *Mammalian Trithorax and polycomb-group homologues are antagonistic regulators of homeotic development.* Proceedings of the National Academy of Sciences of the United States of America, 1999. **96**(25): p. 14372-7.
61. Paggetti, J., et al., *Crosstalk between leukemia-associated proteins MOZ and MLL regulates HOX gene expression in human cord blood CD34+ cells.* Oncogene, 2010. **29**(36): p. 5019-31.

62. Raaphorst, F.M., *Deregulated expression of Polycomb-group oncogenes in human malignant lymphomas and epithelial tumors.* Human molecular genetics, 2005. **14 Spec No 1**: p. R93-R100.
63. Scholl, C., et al., *The homeobox gene CDX2 is aberrantly expressed in most cases of acute myeloid leukemia and promotes leukemogenesis.* The Journal of clinical investigation, 2007. **117**(4): p. 1037-48.
64. Bienz, M., *Homeotic genes and positional signalling in the Drosophila viscera.* Trends in genetics : TIG, 1994. **10**(1): p. 22-6.

Figure Legends

Figure 2-1. Heat map of expression data for MEIS1 and the HOX cluster genes from 45 *de novo* AML patient samples for which there is whole genome sequencing data.
RNA expression is plotted according to the normalized array mean, using a linear scale from 0 (green) to 5000 (black) to 10,000 or greater (red). Each row represents a patient and each column represents a single probeset from the Affymetrix U133 Plus 2 array. Translocations and mutations for each patient are indicated to the right, along with availability of RNASeq data. "Other" refers to patients with any cytogenetic abnormality that is not otherwise annotated. Patients are arranged according to cytogenetics. Numbers 1-10 to the right of the heatmap indicate the patients harboring the mutations listed in Table 2.

Figure 2-2. Raw data from Affymetrix U133 Plus 2 array for 190 patients *de novo* AML patient samples.
Each of the overexpressed HOX genes is plotted against each other to determine intra-HOX correlations. r^2 values are indicated, a complete list of r^2 and p-values for all of the combinations is in Table 3.

Figure 2-3. Correlation of patient characteristics and HOX Expression.
Expression values of *HOXA* (left) and *HOXB* (right) plotted against sex (A), age (B), %BM blast (C), EFS (D) and OS (E). p- and r^2- values are indicated for each comparison. None of the patient characteristics show a significant correlation with HOX expression.

Figure 2-4. A. Heat map of expression data for MEIS1 and the HOX cluster genes from 190 *de novo* AML patient samples by cytogenetics.
RNA expression is plotted according to the normalized array mean, using a linear scale from 0 (green) to 5000 (black) to 10,000 or greater (red). Each row represents a patient and each column represents a single probeset from the Affymetrix U133 Plus 2 array. Translocations and mutations for each patient are indicated to the right, along with availability of RNASeq data. "Other" refers to patients with any cytogenetic abnormality that is not otherwise annotated. Patients are arranged according to cytogenetics. B. Raw data from Affymetrix U133 Plus 2 array for 190 patients *de novo* AML patient samples shown in Panel A. Samples are categorized by translocations and mutations. "Other" refers to patients with any cytogenetic abnormality that is not otherwise annotated.

Figure 2-5. A. Heat map of expression data for MEIS1 and the HOX cluster genes from 190 *de novo* AML patient samples by recurrent mutations.
RNA expression is plotted according to the normalized array mean, using a linear scale from 0 (green) to 5000 (black) to 10,000 or greater (red). Each row represents a patient and each column represents a single probeset from the Affymetrix U133 Plus 2 array. Translocations and mutations for each patient are indicated to the right, along with availability of RNASeq data. "Other" refers to patients with any cytogenetic abnormality that is not otherwise annotated. Patients are arranged according to mutation status. B. Raw data from Affymetrix U133 Plus 2 array for 190 patients *de novo* AML patient

samples shown in Panel A. Samples are categorized by mutations and translocations. Patients with *DNMT3A* mutations are split into *DNMT3A* R882 mutations and "other" for non-R882 mutated residues.

Figure 2-6. LC-MS data of total methylcytosine content for 70 *de novo* AML patients.
Samples are categorized by mutations and translocations. Patients with *DNMT3A* mutations are split into *DNMT3A* R882 mutations and "other" for non-R882 mutated residues.

Figure 2-7. Methylation array data of the HOXA locus from 190 *de novo* AML patient samples by HOXA9 expression.
Estimated proportion of cytosine methylation in 190 *de novo* AML patient samples at loci within the HOXA cluster of genes, ordered left-to-right in the direction of transcription (reverse strand orientation). Methylation is estimated as the proportion of overall fluorescence intensity contributed by the methylated probe allele. Each row represents one patient and each column represents one interrogated cytosine probed by the Illumina HumanMethylation450 array. Translocations and mutations for each patient are indicated at right, along with availability of RNASeq data. 'Other' refers to patients with any cytogenetic abnormality that is not otherwise annotated. Patients are arranged bottom-to-top in descending order of HOXA9 expression. HOXA9 and B3 expression are shown shown in the left-most columns with a scale of 0 (white) to 10,000 (red).

Figure 2-8. Methylation array data of the HOXA locus from 190 *de novo* AML patient samples by HOXB3 expression.
Estimated proportion of cytosine methylation in 190 *de novo* AML patient samples at loci within the HOXB cluster of genes, ordered left-to-right in the direction of transcription (reverse strand orientation). Methylation is estimated as the proportion of overall fluorescence intensity contributed by the methylated probe allele. Each row represents one patient and each column represents one interrogated cytosine probed by the Illumina HumanMethylation450 array. Translocations and mutations for each patient are indicated at right, along with availability of RNASeq data. 'Other' refers to patients with any cytogenetic abnormality that is not otherwise annotated. Patients are arranged bottom-to-top in descending order of HOXB3 expression. HOXA9 and B3 expression are shown shown in the left-most columns with a scale of 0 (white) to 10,000 (red).

Figure 2-9. Methylation array data of the HOXA and HOXB loci from 190 *de novo* AML patient samples by recurrent mutations.
Estimated proportion of cytosine methylation in 190 *de novo* AML patient samples at loci within the HOXA (Panel A) and HOXB (Panel B) cluster of genes, ordered by left-to-right by direction of transcription (reverse strand orientation). Methylation is estimated as the proportion of overall fluorescence intensity contributed by the methylated probe allele. Each row represents one patient and each column represents one interrogated cytosine probed by the Illumina HumanMethylation450 array. Translocations and

mutations for each patient are indicated at right, along with availability of RNASeq data. 'Other' refers to patients with any cytogenetic abnormality that is not otherwise annotated. Patients are arranged according to mutation status; patients without major recurrent mutations are then arranged according to cytogenetic status.

Figure 2-10. HOX expression levels in healthy hematopoietic cells.
A. Heat map of expression data for MEIS1 and the HOX cluster genes from peripheral blood of healthy patient samples. Cells were sorted into different lineages- CD34+ cells (enriches for HSCs), committed progenitors (promyelocytes) and fully differentiated neutrophils (PMNs) (n=5 each). RNA expression is plotted according to the normalized array mean, using a linear scale from 0 (green) to 5000 (black) to 10,000 or greater (red). Each row represents a patient and each column represents a single probeset from the Affymetrix U133 Plus 2 array. B. Raw data from Affymetrix U133 Plus 2 array for healthy patient samples shown in Panel A.

Figure 2-11. Comparison of HOX expression pattern in AML vs. healthy hematopoietic cells.
For each cell type, expression level for each gene was averaged across all samples. For the AML samples only the 89 normal karyotype samples were analyzed since the translocations have unique expression patterns. Since the CD34+ cells are not purely HSCs, the HOX raw expression data is lower than that of the pure AML samples. Therefore, the healthy hematopoietic cells were then normalized with CD34+ *HOXA9* expression set to 100% and the AML samples were normalized separately to AML *HOXA9* expression set to 100%.

Figure 2-12. HOX expression pattern in mouse hematopoietic cells.
Raw data across all 4 HOX loci and *Meis1* from Affymetrix Mouse 430 2.0 array for cell populations enriched for HSCs (SLAM and KLS) as well as committed progenitors (promyelocytes) and full differentiated neutrophils.

Figure 2-13. HOX expression pattern in mouse acute promyelocytic leukemia samples.
Raw data across all 4 HOX loci and *Meis1* from Affymetrix Mouse 430 2.0 array for cell populations enriched for HSCs (SLAM and KLS) in acute promyelocytic leukemia (APL) mouse models (mCG-PR and CTSG KO) as well as APL tumor.

Table 2-1. Clinical characteristics of 190 *de novo* AML patients

UPN	NanoString data	WGS	Age	Sex	Race	FAB	Cytogenetics	%BM Blast	PB WBC@ presentation	%PB Blast	Vital Status (4.29.11)	EFS months (4.29.11)	OS months (4.29.11)	MLL status	DNMT3A	DNMT3A mutation consequences	FLT3	FLT3 other	IDH1	IDH2	NPM1
103342	Yes	Yes	61	F	W	M2	normal	43	9.8	32	Alive	35.6	36.1								
113971	Yes	Yes	57	F	W	M2	normal	43	3.5	35	Dead	8.6	28.4		G590fs/ 1.5Mb deletion	adds 28 aa before stop/del, chr2:24,460,563-25,807,506				R140Q	
142074	Yes	Yes	60	M	W	M4	normal	89	79.9	56	Dead	9.9	11.3		R882H	missense	ITD				W288fs
224143	Yes	Yes	67	F	W	M1	normal	76	45.6	49	Dead	0.8	0.8		F909C	missense					W288fs
225373	Yes	Yes	71	F	W	M2	normal	70	35.9	67	Dead	0.8	0.8						R132C		
246634	Yes	Yes	79	M	W	M4	normal	58	18.5	0	Dead	14.6	14.6		A741V,E477*	missense/nonsense					
273919	Yes	Yes	25	M	W	M2	normal	56	126.1	71	Alive	51.5	51.5								
335640	Yes	Yes	67	F	W	M5	normal	85	93.3	3	Dead	2.5	2.5		R882C	missense		D835Y			W288fs
426980	Yes	Yes	68	M	W	M2	normal	64	5.5	48	Dead	26.2	46.8							R140L	
445045	Yes	Yes	75	M	W	M2	normal	63	2.5	2	Alive	62.1	62.1								
452198	Yes	Yes	55	M	W	M5	normal	97	72.6	8	Alive	16.6	53.8		R882H	missense		D835H	R132H		W288fs
456892	Yes	Yes	58	M	W	M4	normal	58	19.9	10	Alive	45.0	45.0		E505*	nonsense					W288fs
509754	Yes	Yes	21	F	W	M1	normal	91	8.2	83	Alive	73.9	73.9						R132S		W288fs
545259	Yes	Yes	30	F	W	M1	normal	86	22.9	88	Alive	33.7	33.7								
548327	Yes	Yes	51	M	W	M1	normal	85	63.7	85	Alive	31.9	31.9						R132G		W288fs
573988	Yes	Yes	67	F	W	M4	normal	75	15.2	10	Dead	12.0	20.5							R140Q	W288fs
700717	Yes	Yes	45	M	W	M0	normal	75	2.3	0	Dead	3.8	11.2	MLL.PTD exon3-exon11							
775109	Yes	Yes	45	M	W	M5	normal	81	1.6	0	Alive	81.4	81.4	MLL/AF10 (RNA)							
807970	Yes	Yes	38	M	W	M1	normal	86	39.8	97	Alive	54.9	54.9						R132C		W288fs
808642	Yes	Yes	61	M	W	M1	normal	49	2.9	14	Dead	12.0	15.5				ITD				
816067	Yes	Yes	35	F	W	M5	normal	87	156.1	8	Alive	33.9	33.9		R882C	missense	ITD		R132C		W288fs
817156	Yes	Yes	54	M	B	M2	normal	67	45.0	83	Dead	8.5	34.0	MLL.PTD exon3-exon11	R882H	missense					
831711	Yes	Yes	57	F	W	M1	normal	64	2.5	61	Alive	53.4	53.4	MLL.PTD intron2-exon9, exon3-exon9							
869586	Yes	Yes	23	M	W	M4	normal	51	27.1	63	Dead	7.7	19.0								
906708	Yes	Yes	76	F	W	M4	normal	91	5.0	16	Dead	17.0	25.8								W288fs
933124	Yes	Yes	57	F	W	M1	normal	100	105.2	91	Dead	11.1	24.6		L723fs	adds 56 aa before stop	ITD				W288fs
991612	Yes	Yes	63	M	W	M2	normal	35	0.6	0	Alive	62.0	62.0								
123172	No	Yes	56	M	W	M1	normal	90	30.5	45	Dead	53.9	53.9					D835Y			W288fs
179223	No	Yes	82	F	W	M2	normal	53	1.4	25	Dead	1.5	1.5								W288fs
202127	No	Yes	68	F	W	M3	t(15;17)	85	44.2	32	Dead	1.0	1.0				ITD				
254137	No	Yes	31	F	W	M2	normal	63	12.6	45	Dead	6.6	6.6					598bp in frame insertion AGSSDNEYFYVDFREY			W288fs
321258	No	Yes	31	F	W	M3	t(15;17)	40	1.6	0	Alive	64.3	64.3								
327733	No	Yes	32	F	W	M1	normal	94	5.1	70	Dead	11.9	56.3						R132C		W288fs
400220	No	Yes	34	F	W	M4	normal	71	19.6	58	Dead	8.3	10.7	MLL.PTD exon4-exon7			ITD				
455499	No	Yes	29	F	H	M3	t(15;17)	85	1.2	1	Alive	37.1	37.1								
501944	No	Yes	40	F	W	M3	t(15;17),+8	90	0.4	0	Alive	73.6	73.6								
673778	No	Yes	53	M	W	M3	t(15;17)	42	1.0	29	Alive	76.1	76.1								
702808	No	Yes	75	F	W	M5	normal	41	78.1	0	Dead	5.3	6.3		R882H	missense					W288fs
753374	No	Yes	29	M	W	M2	normal	45	2.2	9	Dead	8.9	11.9								W288fs
758168	No	Yes	25	F	W	M3	t(15;17)	93	16.3	40	Dead	31.6	49.8								
804168	No	Yes	53	M	W	M1	normal	86	99.8	52	Dead	7.7	30.0				ITD				W288fs
849660	No	Yes	22	M	W	M1	normal	71	167.5	90	Dead	7.7	27.0								
863018	No	Yes	62	M	W	M3	t(15;17)	82	0.9	11	Alive	69.4	69.4								
907786	No	Yes	81	F	W	M5	normal	53	16.8	7	Dead	9.2	9.2								
943309	No	Yes	35	M	W	M3	del(7),+8, t(15;17)	90	1.5	7	Alive	63.1	63.1				ITD				
203	Yes	No	77	M	W	M0	normal	67	92.0	56	Dead	11.6	18.5							R140Q	
245	Yes	No	68	F	W	M1	+21 only	91	3.1	55	Dead	4.9	4.9							R140Q	
287	Yes	No	64	F	W	M3	normal	60	116.5	N/A	Dead	1.7	2.2		R882H	missense	ITD				W288fs
295	Yes	No	76	F	B	M2	not done	48	61.6	88	Dead	1.3	1.3							R140Q	W288fs
100232	Yes	No	81	M	W	M2	complex	46	1.5	0	Dead	1.4	1.4		R882C	missense					
150288	Yes	No	51	M	B	M1	normal	62	1.5	10	Alive	13.9	21.3								
186481	Yes	No	57	F	W	M4	normal	90	110.2	84	Dead	5.1	7.5		R882C	missense		D835	R132H		W288fs
208027	Yes	No	62	F	Asian	M2	1 cell +7	46	75.2	53	Dead	6.9	10.2		R803S	missense		D835			W288fs
237983	Yes	No	59	M	W	M2	-7 only	70	11.5	48	Alive	23.1	23.1							R140Q	
245450	Yes	No	63	M	W	M4	+8,-17	62	33.3	9	Dead	8.4	8.4		R729W	missense					
274429	Yes	No	64	M	W	M1	+8	60	1.2	5	Dead	27.4	27.4						R132C	R172K	
290344	Yes	No	49	F	W	M4	+11	51	12.4	5	Alive	39.8	39.8		R882C	missense					
303642	Yes	No	54	F	W	M1	normal	72	17.8	90	Alive	16.2	16.2						R132C		W288fs
311636	Yes	No	64	F	W	M4	normal	72	131.5	90	Dead	12.1	24.1							R140Q	W288fs
319955	Yes	No	45	F	W	M2	1~50dmin, der(6)	56	29.4	37	Alive	45.7	45.7								
322110	Yes	No	60	F	W	M1	9qh+	86	11.5	43	Alive	32.9	32.9				ITD			R140Q	W288fs
329614	Yes	No	68	F	W	M1	normal	90	202.2	91	Alive	4.3	22.0				ITD				W288fs
348685	Yes	No	55	F	W	M4	normal	69	38.4	57	Dead	6.3	8.1								
369065	Yes	No	59	M	W	M0	normal	61	1.4	2	Dead	17.2	20.2		splice_site del					R140Q	
375182	Yes	No	57	M	W	M5	normal	52	99.2	4	Dead	7.8	7.9		R882H	missense	ITD				W288fs
387919	Yes	No	57	F	W	M1	+8	91	19.8	91	Dead	13.3	15.1					D835Y	R132C		
431799	Yes	No	44	F	W	M1	normal	86	57.0	17	Dead	13.8	24.8		R882H	missense					W288fs
433325	Yes	No	51	F	W	M2	normal	64	34.5	79	Dead	10.2	16.3		Q515*	nonsense					W288fs
499862	Yes	No	78	M	W	M4	normal	70	16.8	28	Dead	0.6	0.6								
514066	Yes	No	35	F	W	M1	t(8;21)	81	56.6	86	Dead	4.6	8.0								
530962	Yes	No	76	M	W	M7	-Y	33	3.6	6	Dead	7.2	9.9		W314fs	creates in frame stop at 314					
578179	Yes	No	88	F	W	M1	normal	33	2.9	0	Dead	2.4	2.4								
594368	Yes	No	42	M	B	M1	complex [no favorable]	68	10.7	51	Dead	14.9	21.5		R729Q/D618fs (e15+2)	missense/adds RVLGLGRGLELPWLSLGCGA before stop				R140Q	
605322	Yes	No	43	F	B	M1	-7 only	87	183.7	91	Dead	4.0	4.0						R132C		
606061	Yes	No	37	F	W	M2	normal	52	18.7	33	Dead	8.5	9.3								
632729	Yes	No	75	M	W	M2	9qh+	76	98.2	80	Dead	0.1	0.1		R882C	missense	ITD				
708512	Yes	No	65	F	W	M4	complex [no favorable]	90	48.5	76	Dead	1.9	7.8		R882H	missense		D835E	R132H		W288fs
721214	Yes	No	41	F	W	M1	failed	92	151.8	90	Dead	5.7	5.7		R882H	missense	ITD				W288fs
730817	Yes	No	66	F	W	M5	normal	32	78.5	11	Dead	5.9	7.7		Q615fs			D835			
740266	Yes	No	58	M	W/Nat Amer	M0	complex [no favorable]	59	56.5	72	Alive	32.6	32.6		R882H	missense			R132H		
766126	Yes	No	60	F	W	M0	normal	50	2.1	14	Dead	13.8	13.8								
851929	Yes	No	69	F	W	M1	+8	42	10.1	32	Alive	24.8	24.8		R792H/R803fs	missense/fusion of e20:e22, fs adds SSAK before stop				R172K	
868442	Yes	No	52	M	W	M4	normal	75	329.2	51	Alive	16.7	16.7		R882H	missense		D835Y			W288fs
869922	Yes	No	56	F	W	M2	normal	60	202.7	96	Dead	9.6	16.4		R882H	missense					
914247	Yes	No	67	M	W	M4	normal	30	34.2	6	Dead	18.1	18.1	MLL.PTD exon3-exon11						R140Q	
957664	Yes	No	76	M	W	M5	+8	78	4.0	0	Dead	0.6	0.6		Q615*	nonsense					W288fs
972783	Yes	No	71	M	W	M0	+13	90	1.0	13	Dead	13.4	15.4							R140Q	
987523	Yes	No	27	M	B	M1	complex [no favorable]	77	277.7	88	Dead	3.0	10.5						R132C		
104851	No	No	25	F	B	M2	normal	53	140.1	87	Dead	7.3	11.5				ITD				
195182	No	No	63	M	W	M2	-7,t(9;22)	57	77.3	70	Dead	9.3	9.3								
242129	No	No	51	F	B	M4	complex [no favorable]	39	61.4	39	Alive	3.7	51.6	t(11;19)(q23;p13.1): FISH+							
255108	No	No	61	M	W	M0	+8	95	61.6	86	Alive	6.1	52.6				ITD				W288fs
255421	No	No	47	F	W	M1	+8 only	77	105.3	79	Dead	6.0	7.4								
296361	No	No	31	F	W	M5	normal	83	249.7	6	Dead	6.4	7.5		R882H	missense	ITD				W288fs

UPN	NanoString data	WGS	Age	Sex	Race	FAB	Cytogenetics	%BM Blast	PB WBC@ presentation	%PB Blast	Vital Status (4.29.11)	EFS months (4.29.11)	OS months (4.29.11)	MLL status	DNMT3A	DNMT3A mutation consequences	FLT3	FLT3 other	IDH1	IDH2	NPM1
303818	No	No	41	F	W	M0	complex [no favorable]	42	47.6	74	Alive	39.0	52.6								
312340	No	No	33	M	W	M2	de(9),del(20)	33	2.8	0	Dead	5.7	8.8								
325028	No	No	44	M	W	M2	complex [no favorable]	39	34.8	83	Dead	5.2	11.8								
326772	No	No	47	F	W	M2	complex [no favorable]	37	12.0	0	Dead	12.2	12.2								
332131	No	No	24	M	W	M1	normal	77	223.8	76	Dead	9.8	20.5		G8S		ITD				
345701	No	No	48	F	W	M1	inv(16)	75	72.9	41	Alive	8.8	26.6								
380949	No	No	59	F	W	M5	normal	95	12.1	0	Dead	11.4	11.8								
407992	No	No	33	M	W	M4	t(11;19), inv(12)	71	25.9	14	Dead	7.7	7.7	t(11;19)(q23;p13): FISH+				D835			
432398	No	No	67	M	W	M2	complex [no favorable]	50	2.1	17	Dead	10.3	12.4	add(11q23)							
434640	No	No	71	M	W	M0	complex [no favorable]	97	5.4	54	Dead	4.8	7.0								
470690	No	No	58	M	W	M4	i(17),+13	79	110.3	64	Dead	4.6	4.6								
498463	No	No	60	F	W	M1	del(7)(q11.2) only	87	46.5	65	Dead	4.9	17.4		P718L	missense					
537782	No	No	51	F	W	M3	t(15;17)	93	3.6	40	Dead	5.4	46.5								
569053	No	No	77	M	W	M6	del(20)	75	2.6	8	Dead	3.7	6.3								
619751	No	No	57	F	W	M4	inv(16)	61	13.6	17	Dead	10.8	45.8								
671473	No	No	42	F	W	M4	inv(16)	40	37.9	16	Alive	9.8	53.5								
723101	No	No	65	M	W	M1	normal only 15 metaphases	79	143.2	84	Dead	2.7	6.6	MLL.PTD exon3-exon11							
737451	No	No	21	F	W	M5	normal	90	103.7	72	Dead	7.5	7.5		R882H	missense					W288fs
806794	No	No	70	F	W	M1	pending	32	0.7	6	Dead	4.2	4.2								
809653	No	No	60	F	W	M1	complex [no favorable]	62	16.0	29	Dead	1.9	1.9								
852559	No	No	18	M	W	M1	complex [no favorable]	86	144.2	90	Dead	6.7	7.2								
861663	No	No	48	M	W	M0	pending	92	72.6	87	Dead	2.4	2.4								
868231	No	No	63	M	W	M4	normal	82	76.7	39	Dead	5.3	5.3				ITD				
884262	No	No	54	M	W	M6	complex [no favorable]	30	2.3	0	Dead	6.5	10.0								
923966	No	No	61	M	W	M5	t(9;11)	89	286.0	0	Dead	3.2	5.7	t(9;11)(p22;q23)							
992966	No	No	67	M	W	M1	normal	89	30.8	61	Dead	4.5	5.5	MLL.PTD exon4-exon7						R140Q	
104	No	No	50	M	W	M4	normal	88	16.9	52	Dead	8.7	12.0		R882C	missense			R132H		
110	No	No	61	F	W	M3	t(15;17)	44	1.0	N/A	Dead	24.1	25.7								
201	No	No	30	M	W	M3	normal	82	5.7	N/A	Alive	57.2	113.1								
237	No	No	46	M	W	M1	t(8;7;21)	90	47.7	81	Dead	16.1	30.6								
269	No	No	23	M	W	M2	+y	59	32.4	N/A	Alive	10.0	103.9								
296	No	No	81	M	W	M4	normal	98	90.6	37	Dead	8.2	8.2		R882H	missense		D835			W288fs
115225	No	No	22	M	Asian	nc	normal	80	174.8	94	Alive	15.0	15.0					597bp in frame ins			
123804	No	No	61	F	W	M5	+5(5)t(2;5), t(8;16)	71	29.6	55	Dead	0.0	0.0					D835H			
128392	No	No	65	M	W	M0	complex [no favorable]	83	0.9	0	Dead	17.1	17.1								
141273	No	No	64	F	B	M2	normal	80	2.6	46	Alive	15.6	19.2				No coverage	No coverage	No coverage	No coverage	
146218	No	No	76	M	W	M1	inv(16)	73	98.2	55	Dead	4.5	4.5								
147796	No	No	78	M	W	M4	complex [no favorable]	40	9.9	6	Dead	1.3	1.3								
150951	No	No	39	F	W	M0	del(5) only	75	2.3	4	Dead	22.2	26.3								
156704	No	No	49	M	W	M4	inv(16)	81	105.6	48	Dead	14.2	26.3								
179016	No	No	42	M	B	M3	complex [incl: t(15;17)]	88	2.1	2	Alive	32.2	32.2								
180168	No	No	51	F	W	M5	normal	57	177.8	0	Alive	20.4	20.4		R882H	missense					
193139	No	No	73	F	W	M2	complex	32	15.1	11	Dead	0.5	0.5								
258135	No	No	22	F	W	M2	inv(16), t(3;3)	55	51.8	70	Alive	41.1	41.1								
263578	No	No	52	F	W	M2	t(8;21)	67	4.1	18	Alive	95.2	95.2								
269542	No	No	76	M	B	M0	complex [no favorable]	74	14.5	82	Dead	0.3	0.3								
275382	No	No	65	M	W	M0	+8	99	2.9	34	Dead	32.3	32.3		R882H	missense					
275786	No	No	43	M	W	M2	t(8;21)	40	4.3	39	Alive	19.1	19.1				ITD				
291696	No	No	61	F	W	M3	t(15;17),+mar	73	86.4	68	Dead	0.2	0.2				ITD				
294154	No	No	44	M	W	M5	normal	74	8.8	5	Dead	0.3	0.3		R736H	missense					W288fs
318433	No	No	55	M	W	M4	t(2;4),inv(16)	35	22.2	8	Alive	87.6	87.6								
327929	No	No	50	M	B	M6	failed	70	5.5	0	Dead	3.7	8.9								
346190	No	No	64	F	W	M4	+i(11)(q10)	85	2.9	22	Dead	8.1	8.1	MLL.PTD exon3-exon11	R882P	missense					
399253	No	No	77	M	W	M4	normal	80	40.4	2	Dead	9.3	15.8		R882H	missense					
400830	No	No	73	F	W	M2	-7	37	1.7	16	Dead	11.0	11.0								
405655	No	No	83	M	W	M0	normal	75	93.9	74	Dead	0.1	0.1				ITD				
410324	No	No	60	M	W	M5	complex [no favorable]	94	6.7	30	Alive	36.4	36.4	del(11)(q23); FISH+							
412761	No	No	48	M	W	M5	normal on PB	83	116.2	3	Alive	64.4	64.4								W288fs
418499	No	No	35	M	W	M4	normal	90	52.9	18	Dead	17.0	17.0								W288fs
427366	No	No	66	F	W	M5	normal	90	120.0	7	Dead	0.7	0.7								W288fs
447649	No	No	74	M	W	M3	t(15;17)	51	128.3	Promy	Dead	0.7	0.7				ITD				
463352	No	No	25	M	W	M2	-Y,t(8;21),del(9)	72	12.6	50	Alive	17.7	17.7								
499294	No	No	57	M	W	M4	inv(16)	40	4.9	0	Alive	21.4	21.4								
507202	No	No	59	M	W	M4	normal	75	29.7	17	Alive	16.3	16.3					D835Y			W288fs
507696	No	No	55	F	W	M4	inv(16)	52	39.9	59	Alive	22.4	22.4								
553863	No	No	69	M	W	M0	complex [no favorable]	60	6.1	22	Dead	0.5	0.5								
558395	No	No	64	M	W	M4	del(11) only	95	2.3	0	Dead	0.3	0.3								
570755	No	No	63	M	W	M2	t(11;19)	80	17.0	71	Dead	3.9	3.9	t(11;19)(q23;p13.1): FISH+			ITD				
593890	No	No	35	M	W	M2	+21	34	38.0	40	Dead	14.6	55.4								
595704	No	No	67	F	W	M4	inv(16)	69	37.0	61	Alive	31.7	31.7								
617776	No	No	62	F	W	M2	normal	72	27.7	49	Dead	4.8	7.1		NULL	deleted					
633734	No	No	53	M	W	M1	del(7)(q21q36)	72	2.7	41	Dead	19.3	52.7								
635258	No	No	62	M	W	M5	normal	85	101.8	22	Alive	30.4	30.4								W288fs
670224	No	No	62	M	W	M4	normal	65	104.1	65	Dead	13.4	14.5								W288fs
690397	No	No	45	M	W	M5	t(6;11)	95	5.1	6	Dead	6.6	6.6	t(6;11)(q27;q23); FISH+							
692900	No	No	39	M	W	M0	+del(5)q, t(10;11)	83	5.0	53	Alive	86.1	86.1								
717456	No	No	50	F	W	M5	-7,t(9;11)	81	6.7	10	Dead	5.6	5.6	t(9;11)(p22;q23); FISH+							
729805	No	No	66	M	W	M4	complex	81	2.2	12	Dead	1.6	1.6								
750152	No	No	34	F	W	M2	t(8;21)	57	14.3	59	Alive	34.8	34.8								
757199	No	No	50	F	W	M3	t(15;17), t(16;19)	90	3.4	0	Alive	19.8	19.8								
767969	No	No	72	M	W	M1	complex	88	37.5	18	Alive	27.0	27.0								
794178	No	No	73	M	W	nc	complex	90	58.9	0	Dead	0.2	0.2								
807615	No	No	51	M	W	M1	normal	85	103.6	85	Dead	3.1	3.1		R882H	missense	ITD				
826984	No	No	51	F	W	M4	t(5;21)	73	2.9	2	Dead	8.0	11.1								
862507	No	No	55	F	Unk	M4	del(16),+22	54	114.5	24	Alive	61.3	61.3								
866660	No	No	41	M	W	M2	pending	47	148.4	62	Alive	13.4	13.4								
905179	No	No	75	F	W	M2	t(8;21), del(9)	48	8.3	54	Dead	18.2	19.2								
938150	No	No	60	F	W	M7	complex [no favorable]	58	0.8	0	Dead	12.0	13.6								
944281	No	No	31	M	W	M4	inv(16)	82	336.8	53	Alive	20.1	20.1					D835Y			
962561	No	No	76	F	W	M4	+13,-21	37	2.7	0	Dead	17.3	22.3								
964886	No	No	33	F	B	M3	normal	100	0.9	0	Alive	69.0	69.0				ITD				
974749	No	No	63	M	W	M1	+8, +8	81	46.4	97	Dead	1.2	1.2		R882H	missense			R132C		
982009	No	No	70	M	B	M1	normal	84	282.4	49	Dead	4.9	6.3					D835Y			W288fs
986000	No	No	77	M	W	M7	complex [no favorable]	47	1.9	0	Dead	4.5	4.5								
989176	No	No	76	F	W	M1	normal	99	352.3	98	Dead	0.4	0.4								W288fs
997292	No	No	66	F	W	M2	normal	70	3.4	6	Dead	5.2	5.2		K468R,K829R	missense/missense			R132C		

Table 2-2. Somatic mutations within HOX clusters identified by whole genome sequencing of 45 *de novo* AML patients.

Patient	831711	849660	863018	179223	246634	246634	775109	869586	501944	943309
Chromosome Position	7:27207807	12:52519096	17:43955804	7:27296211	12:52577061	12:52646401	12:52588242	17:44156725	17:44157110	17:44143238
Reference	A	C	A	G	C	T	C	G	T	A
Variant	G	T	G	T	T	G	T	T	C	C
Nearest HOX gene	HOXA13	HOXC13	HOXB1	HOXA13	HOXC13	HOXC11	HOXC13	HOXB13	HOXB13	HOXB13
Distance (-) up or (+) down-stream of TSS	-1,586	-99,862	6,104	-89,990	-41,897	-6,892	-30,716	2,426	2,041	15,913

Table 2-3. Expression correlation values for pairs of HOX genes.

Gene Pair	r^2 Value	p value
A5 v A9	0.6908	< 0.0001
A5 v A10	0.7369	< 0.0001
A5 v B2	0.3584	< 0.0001
A5 v B3	0.5828	< 0.0001
A5 v MEIS1	0.4878	< 0.0001
A9 v A10	0.7173	< 0.0001
A9 v B2	0.4877	< 0.0001
A9 v B3	0.5194	< 0.0001
A9 v MEIS1	0.5165	< 0.0001
A10 v B2	0.5079	< 0.0001
A10 v B3	0.5671	< 0.0001
A10 v MEIS1	0.4121	< 0.0001
B2 v B3	0.7165	< 0.0001
B2 v MEIS1	0.5167	< 0.0001
B3 v MEIS1	0.6758	< 0.0001

Table 2-4. Orthogonal validation of AffyMetrix U133 Plus 2 Array data with custom NanoString codeset

Gene	Affymetrix U133+2			NanoString			Spearman Correlation	
	Probe	Mean	Std Dev	Target Sequence	Mean	Std Dev	Irl	p value
HOXA1	214639_s_at	552.90	468.23	ATAATTCTGGACCAGAGACTTGGTGCGGGGTTAACACCTTCATCCAGATTGG GTGCCAGCATACATTTTCTGGTGGGCCTTAACATCCCTCCTGCTTTTA	29.98	27.55	0.78	<0.0001
HOXA2	214457_at	355.40	267.12	CCCAAAGTTTCCCAGTCTCGCCTTTAACCAGCAATGAGAAAAATCTGAAACA TTTTCAGCACCAGTCACCCACTGTTCCCAACTGCTTGTCAACAATGGG	41.98	40.36	0.70	<0.0001
HOXA3	235521_at	2989.00	4510.00	CCTGGTCTCCTCCGGGGATGCTTCGCGGTCTGTTATCGCGTCAGGAGGAAA GAATTGCTCCAAAAATCTGCACGCGGAGCGAAACAGTTTGAAAGGGACT	24.02	24.06	0.38	<0.0001
HOXA4	206289_at	1335.00	1204.00	TGCACTTCACAAATTAATGACCATGAGCTCGTTTTTGATAAACTCCAACTACA TCGAGCCCAAGTTCCCTCCCTTCGAGGAGTACGCGCAGCACAGCGGC	101.57	109.74	0.76	<0.0001
HOXA5	213844_at	10813.00	10452.00	CATAGTTCCGTGAGCGAGCAATTCAGGGACTCGGCGAGCATGCACTCCGG CAGGTACGGCTACGGCTACAATGGCATGGATCTCAGCGTCGGCCGCTCGG	525.47	770.29	0.83	<0.0001
HOXA6	208557_at	916.12	848.96	CGCGCTGAGGCCCTTCCCGGCCTCGTACGGGGCGTCGAGTCTCCCGGAC AAGACGTACACCTCACCTTGTTTCTACCAACAGTCCAACTCGGTCCTGGCC	462.36	648.99	0.67	<0.0001
HOXA7	235753_at	2229.00	1894.00	CGTATTATGTGAACGCGCTTTTTAGCAAATATACGGCGGGGGCTTCTCTGTT CCAAAATGCCGAGCCGACTTCTTGCTCCTTTGCTCCCAACTCACAGAG	198.35	183.51	0.66	<0.0001
HOXA9	209905_at	18241.00	15137.00	AACCGCCATTGGGCTACTGTAGATTTGTATCCTTGATGAATCTGGGGTTTCCA TCAGACTGAACTTACACTGTATATTTTGCAATAGTTACCTCAAGGCC	629.17	615.53	0.80	<0.0001
HOXA10	213150_at	4581.00	4823.00	GGGGTAAGCGGAATAAACTAGAGAAGGGAGACATTGTTTGGATTTCCTTTAT ACTGTGAAGTTACATGCATAAAAGGGTCAAACCTGTAGATGCAGAAAA	342.71	304.35	0.69	<0.0001
HOXA11	213823_at	311.42	667.80	TGGAGGTAGCTTTGAGGTGGAAGAGGGCTGCAAATCCTTGTGGGAAAAGA AATCTATGATTCCAGGTGGCATCAGTGTCTTTCCACTCCTCCTAGCCACC	20.11	40.06	0.39	<0.0001
HOXA13	231786_at	69.40	138.30	GAGTCGCGCCACGAACCCTTGGGTCTTCCCATGGAAAGCTACCAGCCCTG GGCGCTGCCCAACGGCTGGAACGGCCAAATGTACTGCCCCAAAGAGCAGG	7.92	13.62	0.25	0.0388
HOXB1	208224_at	225.36	194.16	GGGAACGAGCAGACCGCGAGCTTTGCACCGGCCTATGCTGATCTCCTCTC CGAGGACAAGGAAACACCCTGCCCTTCAGAACCTAACACCCCCACGGCCC	6.53	9.36	0.15	0.2155
HOXB2	205453_at	4628.00	3038.00	GAATTCCACTTTAATAAGTACCTGTGCCGGCCACGCCGCGTCGAGATCGCG GCCTTGCTGGACCTCACCGAAAGGCAGGTCAAAGTCTGGTTTCAGAACC	236.30	215.92	0.78	<0.0001
HOXB3	228904_at	8111.00	12274.00	TGTCCGTTTAAATGCTGCTGGGAGACTCGTAAAAAAATCATCGTGGACCTGG AGGATGAGAGGGGCGAGCTTTATTTCGGTCGGATTGCGGTGTGGTGGT	225.18	242.70	0.57	<0.0001
HOXB4	231767_at	1810.00	1464.00	CCTTTCTTTGTCCCCCACTCCCGATACCCAGCGAAAGCACCCTCTGACTGC CAGATAGTGCAGTGTTTTGGTCACGGTAACACACACACACTCTCCCTCA	551.54	551.18	0.85	<0.0001
HOXB5	205600_x_at	1516.00	1017.00	GCTTGTAAATATGGGGATAGTCTGGGTCAGACCCATCTCTCCCTTACCCATC TTGCTTCCAAGACCATTTGTAGTGAGCGAGTGGATGCTGTGCTACGTG	81.30	83.36	0.81	<0.0001
HOXB6	205366_s_at	1293.00	1968.00	CACCCATTCCTTTAAATCCGGAGGGGGAAAAAATCCCAAGGTCTGCAAAGG CGCGGCGCTCGGACTATAAAACACAACAAATCATAAACCCGGCGGAGCA	6.73	12.53	0.04	0.6165
HOXB7	216973_s_at	964.03	2011.00	AAATCTGGACTCTAATTCTGTAATATATCAAGGAATCTCGTAAAACCGACACTA AAACGTCCCTGCCTACAAATCATCCGGCCAAATTATGAGTTCATTG	24.04	38.16	0.71	<0.0001
HOXB8	229667_s_at	425.76	1488.00	GTGGTAGTATCTCGTAATAGCTTCTGTGTGTGAGCTACCGTGGATCTCCTTC CCTTCTCTTGGGGGCCGGGGGGAAAGAAAAGGATTTAAGCAAAGGCTC	125.02	367.07	0.28	<0.0001
HOXB9	216417_x_at	711.26	1010.00	GGTGGCTGTCGTGAAATTGTGCTTGTGTTTCGTGATTTCTTTGGGGGTGATT GTCTCGCTTGTTTTCAGTTGTCGATTATATGGGAGGGTTCTGGGTGGG	61.43	110.61	0.40	<0.0001
HOXB13	209844_at	44.28	36.86	CCACCAGGGTTCCCAAAGAACCTGGCCCAGTCATAATCATTCATCCTGACA GTGGCAATAATCACGATAACCAGTACTAGCTGCCATGATCGTTAGCCTC	9.33	27.23	0.04	0.674
HOXC4	206194_at	456.66	273.31	AGCGCCGCCAGCAAGCAACCCATAGTCTACCCATGGATGAAAAAAATTCAC GTTAGCACGGTGAACCCCAATTATAACGGAGGGGAACCCAAGCGCTCGA	4.75	8.61	0.38	0.0012
HOXC5	206739_at	180.67	147.88	CAGACTCTGGAACTCGAGAAAGAATTCCACTTTAACCGCTACCTCACTCGCC GCAGGCGCATAGAGATCGCCAACAACTTGTGTCTCAATGAGAGACAGA	6.65	11.63	0.03	0.8162
HOXC6	206858_s_at	424.56	1628.00	ACGTCGCCCTCAATTCCACCGCCTATGATCCAGTGAGGCATTTCTCGACCTA TGGAGCGGCCGTTGCCCAGAACCGGATCTACTCGACTCCCTTTTATTC	9.24	14.97	0.14	0.2589
HOXC8	221350_at	285.96	173.97	ACAGCCGGTATCAGACCTTGGAACTAGAAAAGGAGTTTCTCTTTAATCCTTAT TTGACACGAAAACGTCGGATTGAAGTCTCTCATGCCCTGGGACTGAC	7.59	13.13	0.17	0.1584
HOXC9	231936_at	102.22	205.04	AATAATCTTATGTATGTAAAACCCCGTTACGATGTCGGCGACGGGGCCCATC AGTAACTATTACGTGGACTCGCTCATCTCTCACGACAATGAAGACCTC	5.57	7.70	0.13	0.2898
HOXC10	218959_at	267.38	280.67	GGAAAGTTCGGCTAGTGTTCGTGTGTTTGTCGTAGCACCCAGAGCCTCCAC CAAACCCTCTCCATGTCTTTACCTCCCAGTCGCTCTAAGAATCTGCTTG	10.03	35.61	0.02	0.8763
HOXC11	206745_at	370.68	213.39	GGCGACAGTAGTGAGCGCCTGAGCCGAACAATCCTCGAACTAAAAGCCTTC CCTTGCCCATGTGAAAAGATCCGCTAAGACAGCATGTCTGCCAGCGGAA	7.30	4.38	0.08	0.5301
HOXC12	1553512_at	165.49	126.57	AGGGAACTCTCAGACCGCTTGAATCTTAGTGACCAGCAGGTCAAGATCTGG TTTCAGAACCGGAGAATGAAAAAGAAAAGACTTCTGTTGAGGGAGCAAG	17.50	10.88	0.13	0.2889
HOXC13	219832_s_at	69.82	77.52	ACTCCTCACACTTTCACCTTTACTGATTTCCAGAGGAAAGCTAGAGGATCTA GTTCAAGAGGCAAGAAGATCTGGCCCTCAATTAGCTAGATGTAGATGC	11.44	7.79	0.01	0.904
HOXD1	205975_s_at	245.71	213.15	CTCCCGCCTCCGGCCTCCCTGCCGCCTTCAGCACGTTCGAGTGGATGAAA GTGAAGAGGAATGCCTCTAAGAAAGGTAAACTCGCCGAGTATGGGGCCGC	2.65	7.43	0.05	0.5479
HOXD3	206602_s_at	84.47	100.38	TAAAGGATTAAGACCGCAAATTGTCCTTCATGGGTAGAGTCAGGAAGCCCG GTGGCGTGGCACAACACACTTTGGTCATTTCTCAAAAACCACAGTCCTC	4.82	5.23	0.04	0.5976
HOXD4	205522_at	132.27	117.52	TGTGAAAATTATGAGATGCTCACCAACCCGGTGATAAACTTGCTCCCTCGCC ATTGGCTGGCCTGGTCACATGGCTGCCCAACTTTATTCAGTTGACAGC	9.23	7.18	0.20	0.017
HOXD8	231906_at	294.16	145.46	TTACGGATACGATAACTTACAGAGACAGCCGATTTTTACGACCCAGCAAGAG GCCGAGCTGGTACAATATCCTGACTGTAAATCGTCCAGTGGTAATATT	6.44	7.82	0.14	0.2509
HOXD9	205605_at	158.68	137.55	TTTGGGGTTTCGCCCTATCCCACTCCCTCTCTTTCCTGCTCCATTGGTTCCT TAAGAAATGCTATATTTTGTGAGTGCAAGCTGGCTTGGGGAGCCCTCT	6.34	18.96	0.02	0.8303
HOXD10	229400_at	110.40	131.04	TCCCGTCTCTGGCCAAAAGAGAAGTGAACCACCAAAATATGGGTATGAATGT GCATCCTTATATACCTCAAGTAGACAGTTGGACAGATCCGAACAGATC	8.24	21.62	0.05	0.5245
HOXD11	214604_at	128.39	118.90	CAGCCTGCTCTCCGCAGGCCCACTGTCCTTGGGTTTAATGACGTCTCTTCT CTGTGGAACTTCACGATTCCTTCCCACGGTCAACTCGGGACCTCCCAGC	2.49	2.07	0.13	0.2868
HOXD13	207397_s_at	70.71	90.44	GGTAATTGAATCATTAGCTCTCAGCAGTTGCCCTGAGGCAAGTGGAAAGGC AGGCAGTGCTCTGGGGTCACCGAGAAAGTCTAAAAACAGGAGGCTGAAG	3.14	5.55	0.02	0.7586

Figure 2-1

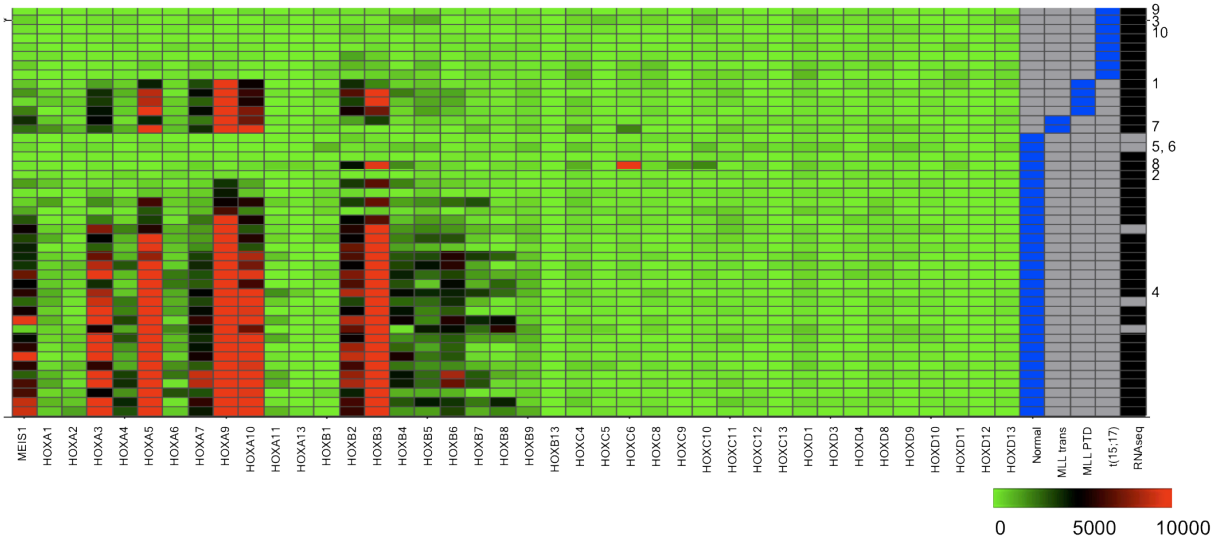

MEIS1
HOXA1
HOXA2
HOXA3
HOXA4
HOXA5
HOXA6
HOXA7
HOXA9
HOXA10
HOXA11
HOXA13
HOXB1
HOXB2
HOXB3
HOXB4
HOXB5
HOXB6
HOXB7
HOXB8
HOXB9
HOXB13
HOXC4
HOXC5
HOXC6
HOXC8
HOXC9
HOXC10
HOXC11
HOXC12
HOXC13
HOXD1
HOXD3
HOXD4
HOXD8
HOXD9
HOXD10
HOXD11
HOXD12
HOXD13
Normal
MLL trans
MLL PTD
t(15;17)
RNAseq
9
3
10
1
7
5, 6
8
2
4
0
5000
10000

Figure 2-2

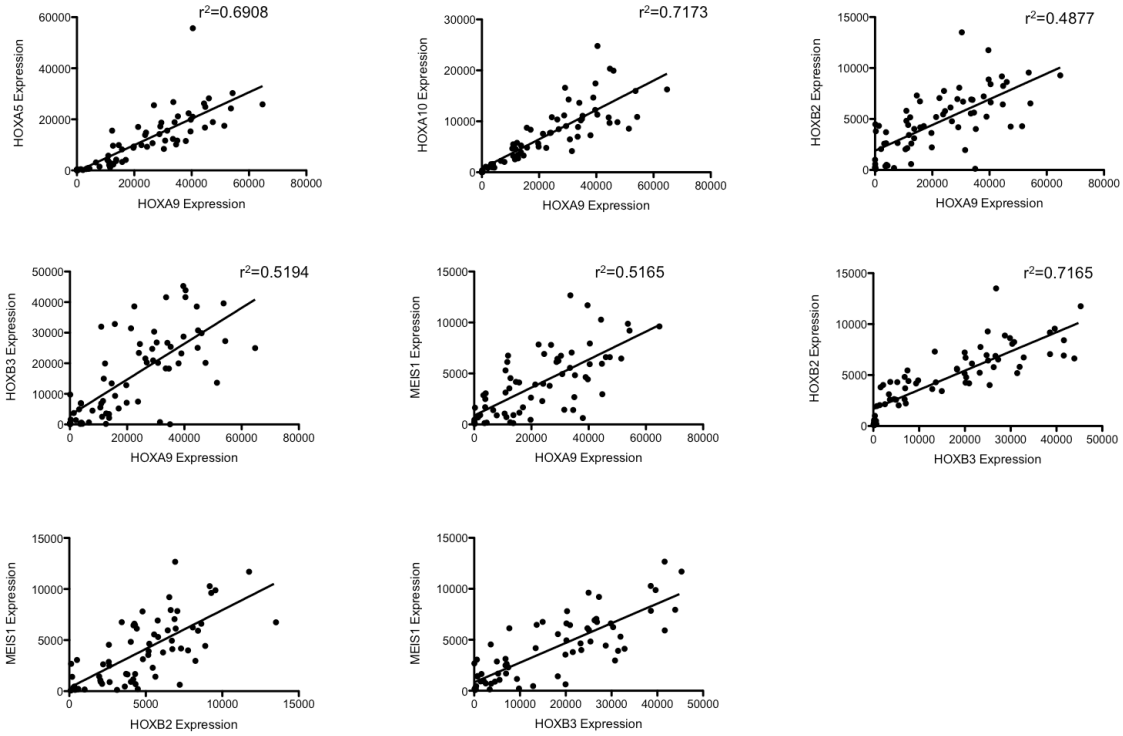

r²=0.6908
HOXA5 Expression
HOXA9 Expression
r²=0.7173
HOXA10 Expression
HOXA9 Expression
r²=0.4877
HOXB2 Expression
HOXA9 Expression
r²=0.5194
HOXB3 Expression
HOXA9 Expression
r²=0.5165
MEIS1 Expression
HOXA9 Expression
r²=0.7165
HOXB2 Expression
HOXB3 Expression
MEIS1 Expression
HOXB2 Expression
MEIS1 Expression
HOXB3 Expression

Figure 2-3

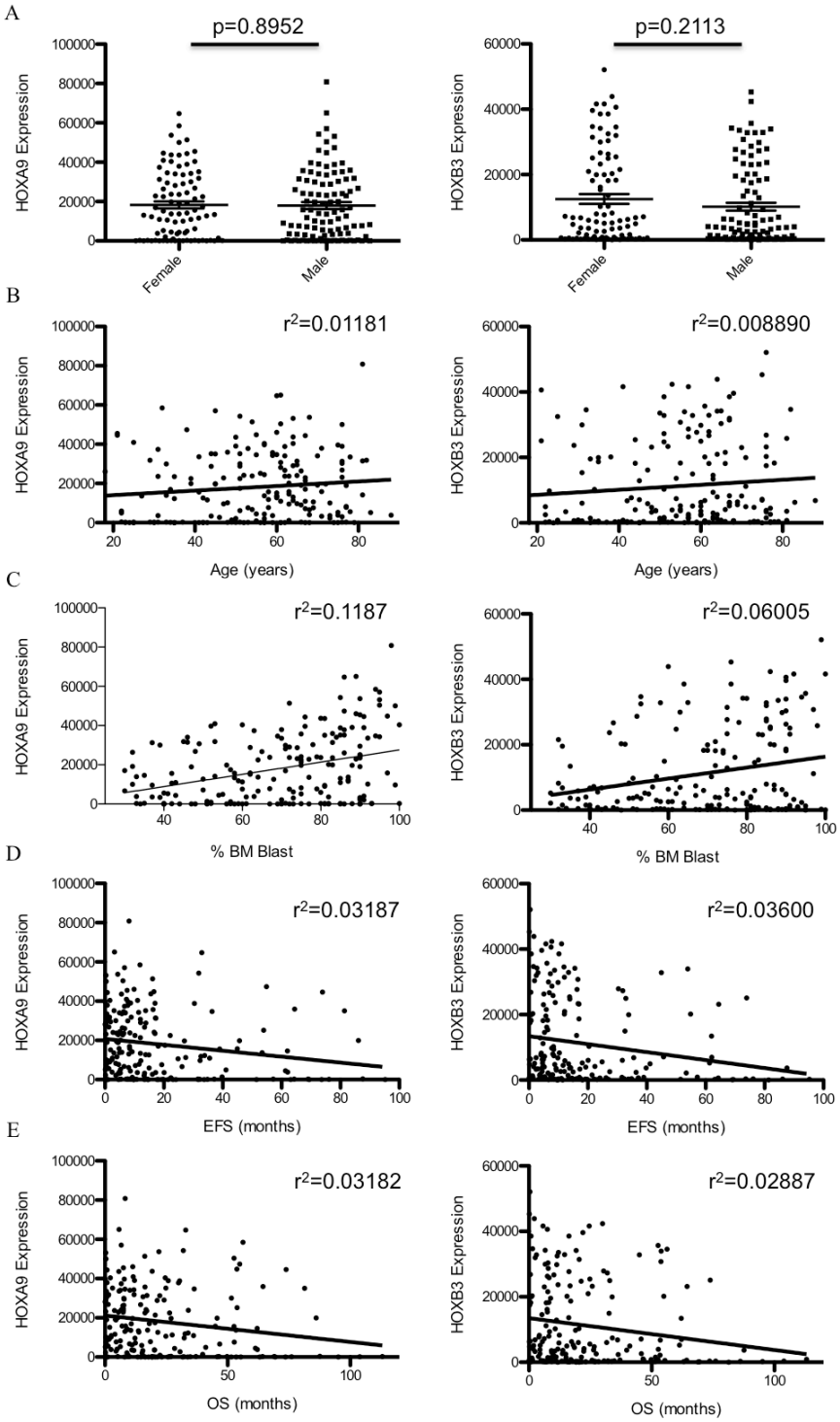
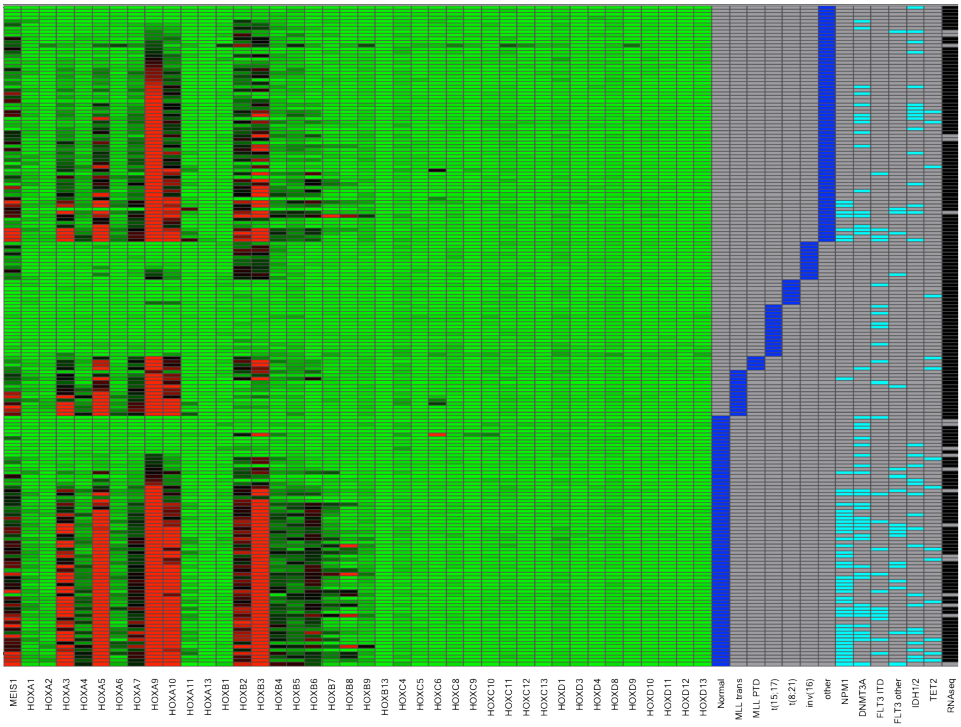

A
p=0.8952
p=0.2113
HOXA9 Expression
HOXB3 Expression
Female
Male
B
r2=0.01181
r2=0.008890
Age (years)
C
r2=0.1187
r2=0.06005
% BM Blast
D
r2=0.03187
r2=0.03600
EFS (months)
E
r2=0.03182
r2=0.02887
OS (months)

Figure 2-4

A.


MEIS1
HOXA1
HOXA2
HOXA3
HOXA4
HOXA5
HOXA6
HOXA7
HOXA9
HOXA10
HOXA11
HOXA13
HOXB1
HOXB2
HOXB3
HOXB4
HOXB5
HOXB6
HOXB7
HOXB8
HOXB9
HOXB13
HOXC4
HOXC5
HOXC6
HOXC8
HOXC9
HOXC10
HOXC11
HOXC12
HOXC13
HOXD1
HOXD3
HOXD4
HOXD8
HOXD9
HOXD10
HOXD11
HOXD12
HOXD13
Normal
MLL trans
MLL PTD
t(15;17)
t(8;21)
inv(16)
other
NPM1
DNMT3A
FLT3 ITD
FLT3 other
IDH1/2
TET2
RNAseq

B.

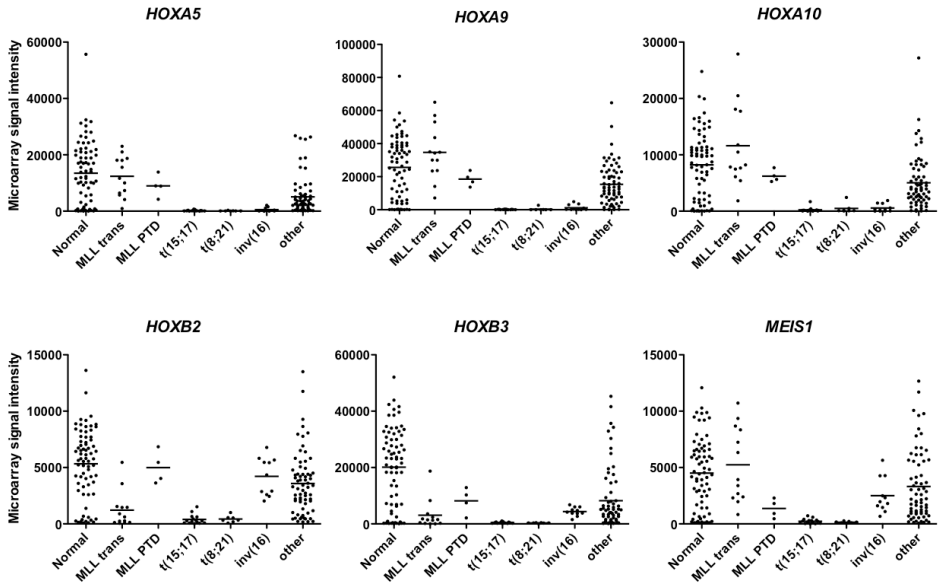
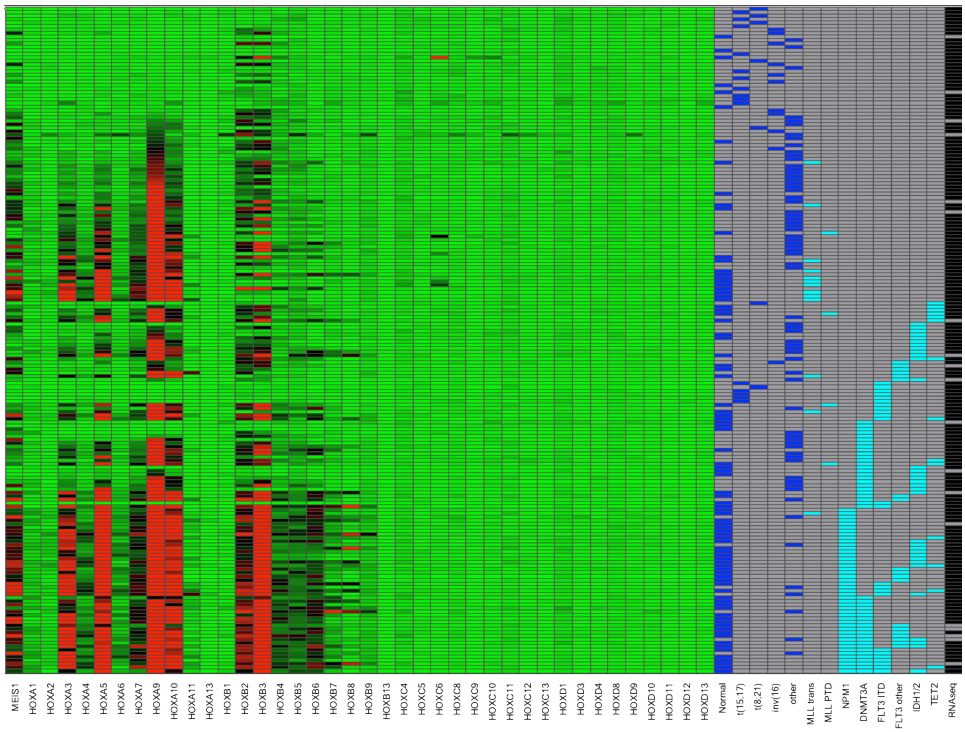

HOXA5
HOXA9
HOXA10
HOXB2
HOXB3
MEIS1
Microarray signal intensity
Normal
MLL trans
MLL PTD
t(15;17)
t(8;21)
inv(16)
other

Figure 2-5

A.


MEIS1
HOXA1
HOXA2
HOXA3
HOXA4
HOXA5
HOXA6
HOXA7
HOXA9
HOXA10
HOXA11
HOXA13
HOXB1
HOXB2
HOXB3
HOXB4
HOXB5
HOXB6
HOXB7
HOXB8
HOXB9
HOXB13
HOXC4
HOXC5
HOXC6
HOXC9
HOXC10
HOXC11
HOXC12
HOXC13
HOXD1
HOXD3
HOXD4
HOXD8
HOXD9
HOXD10
HOXD11
HOXD12
HOXD13
Normal
t(15;17)
t(8;21)
inv(16)
other
MLL trans
MLL PTD
NPM1
DNMT3A
FLT3 ITD
FLT3 other
IDH1/2
TET2
RNAseq

B.

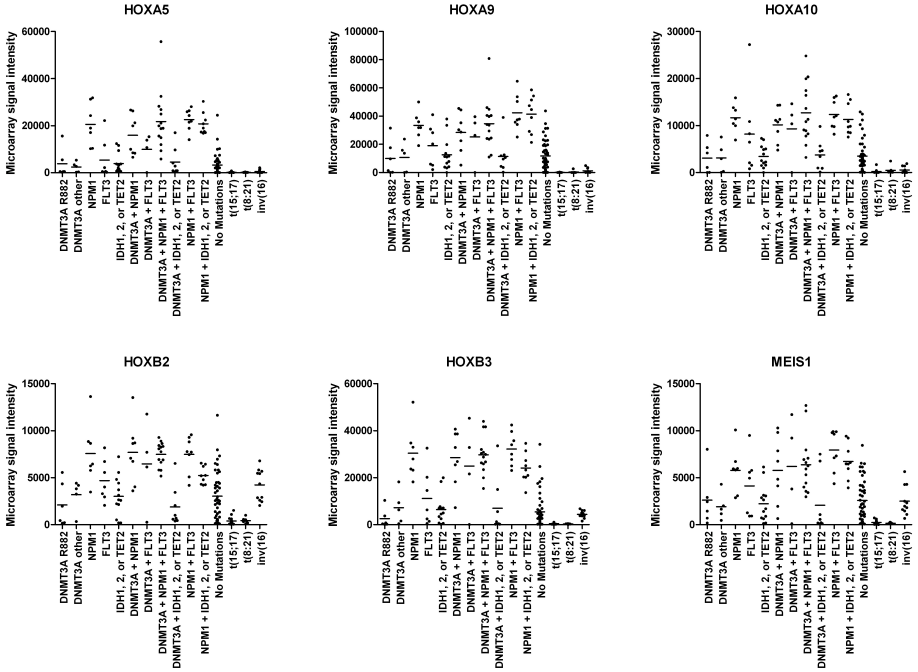

HOXA5
HOXA9
HOXA10
HOXB2
HOXB3
MEIS1
Microarray signal intensity
DNMT3A R882
DNMT3A other
NPM1
FLT3
IDH1, 2, or TET2
DNMT3A + NPM1
DNMT3A + FLT3
DNMT3A + NPM1 + FLT3
DNMT3A + IDH1, 2, or TET2
NPM1 + FLT3
NPM1 + IDH1, 2, or TET2
No Mutations
t(15;17)
t(8;21)
inv(16)

Figure 2-6

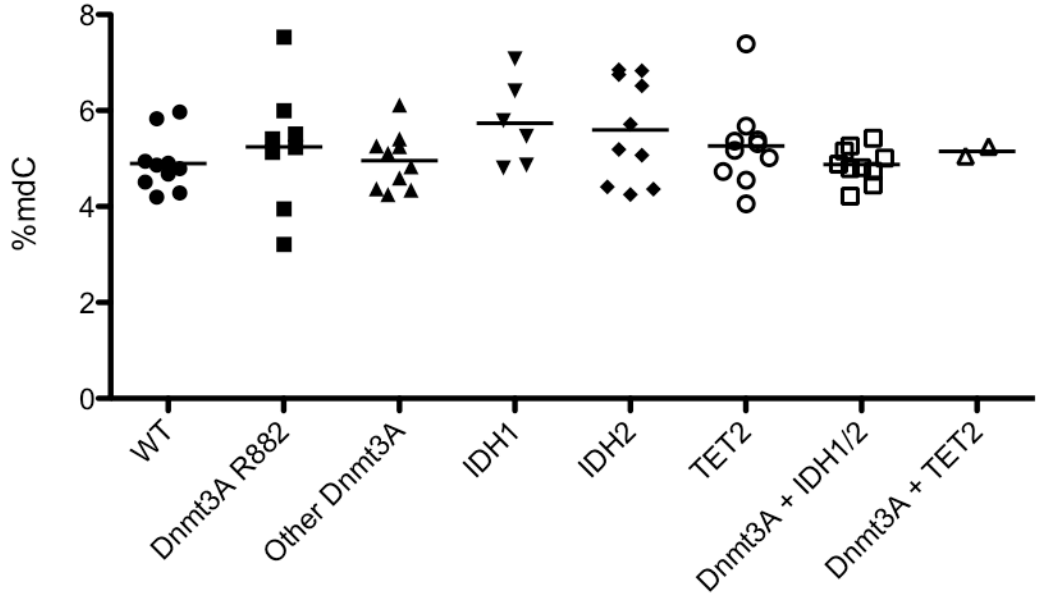

%mdC
8
6
4
2
0
WT
Dnmt3A R882
Other Dnmt3A
IDH1
IDH2
TET2
Dnmt3A + IDH1/2
Dnmt3A + TET2

Figure 2-7

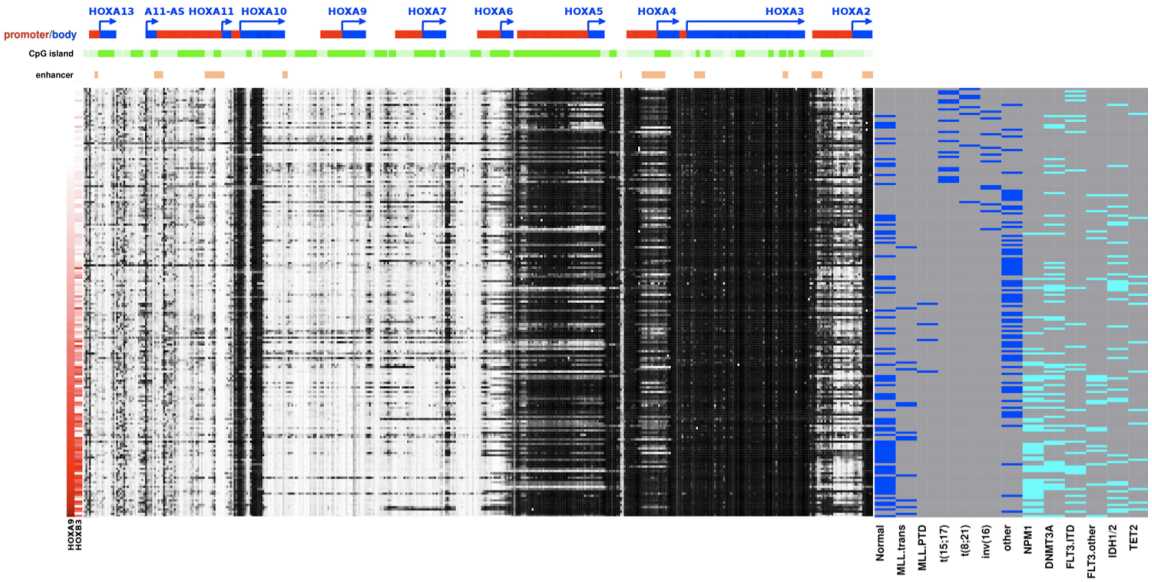

HOXA13
A11-AS
HOXA11
HOXA10
HOXA9
HOXA7
HOXA6
HOXA5
HOXA4
HOXA3
HOXA2
promoter/body
CpG island
enhancer
HOXA9
HOXB3
Normal
MLL_trans
MLL_PTD
t(15;17)
t(8;21)
inv(16)
other
NPM1
DNMT3A
FLT3.ITD
FLT3.other
IDH1/2
TET2

Figure 2-8

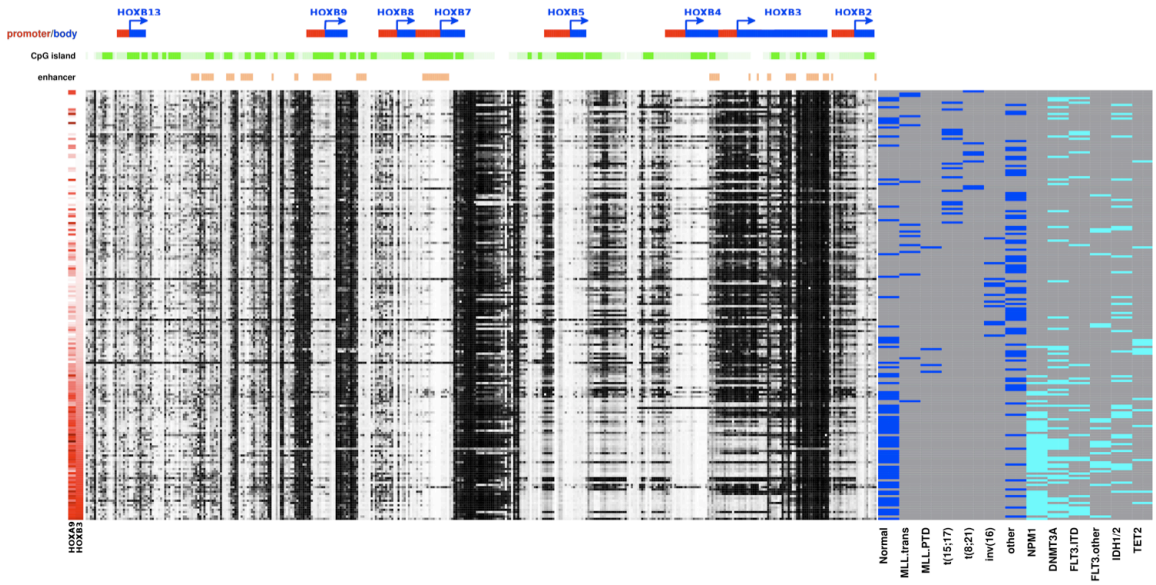
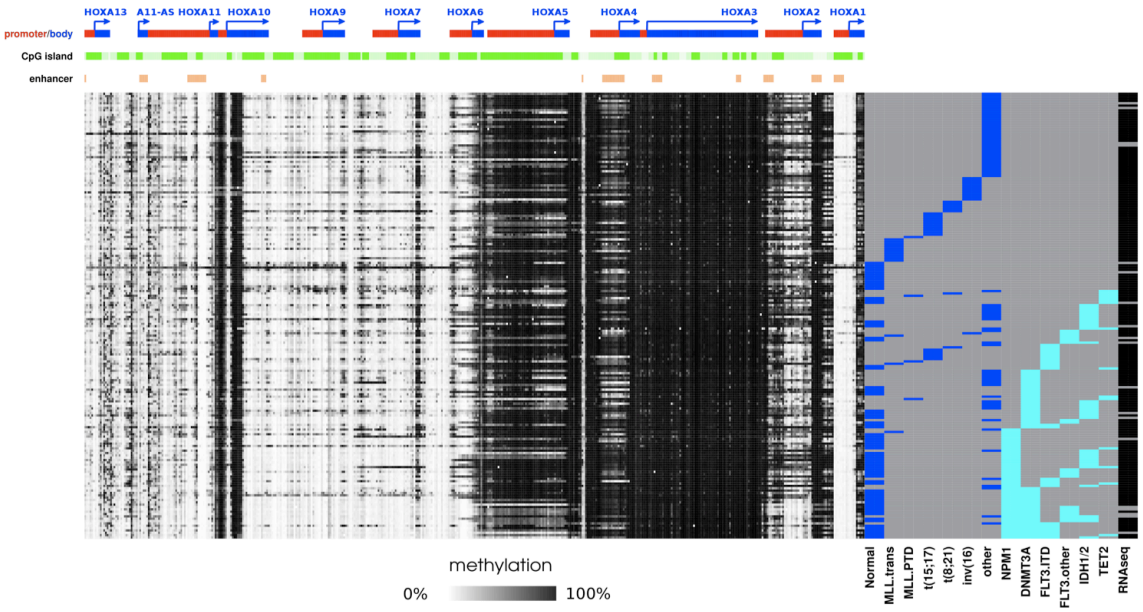

HOXB13
HOXB9
HOXB8
HOXB7
HOXB5
HOXB4
HOXB3
HOXB2
promoter/body
CpG island
enhancer
HOXA9
HOXB3
Normal
MLL.trans
MLL.PTD
t(15;17)
t(8;21)
inv(16)
other
NPM1
DNMT3A
FLT3.ITD
FLT3.other
IDH1/2
TET2

Figure 2-9

A.


HOXA13
A11-AS
HOXA11
HOXA10
HOXA9
HOXA7
HOXA6
HOXA5
HOXA4
HOXA3
HOXA2
HOXA1
promoter/body
CpG island
enhancer
methylation
0%
100%
Normal
MLL.trans
MLL.PTD
t(15;17)
t(8;21)
inv(16)
other
NPM1
DNMT3A
FLT3.ITD
FLT3.other
IDH1/2
TET2
RNAseq

B.

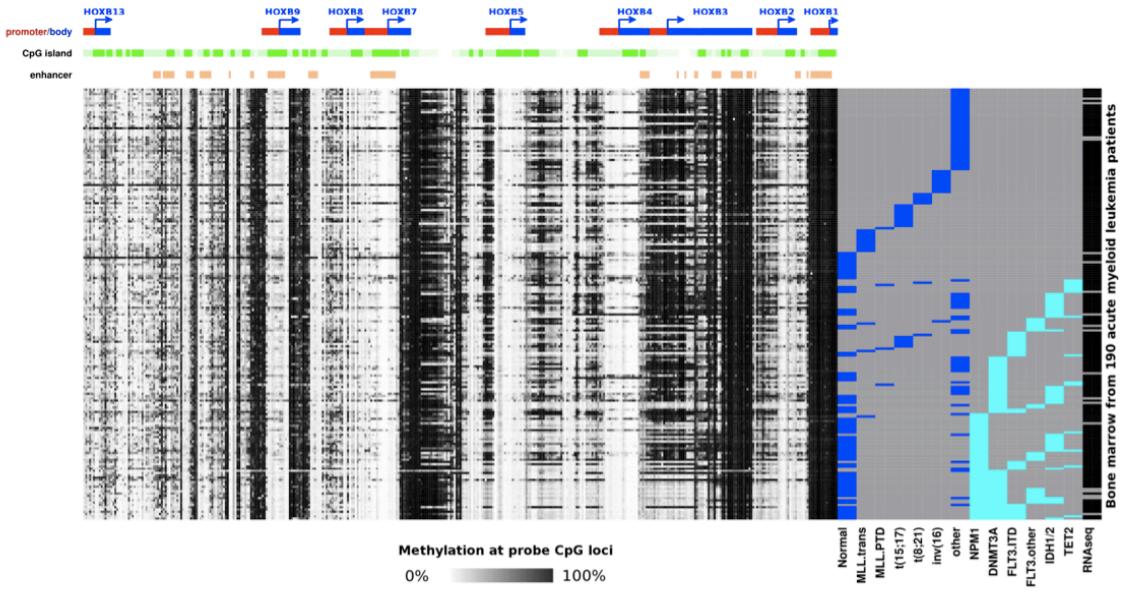
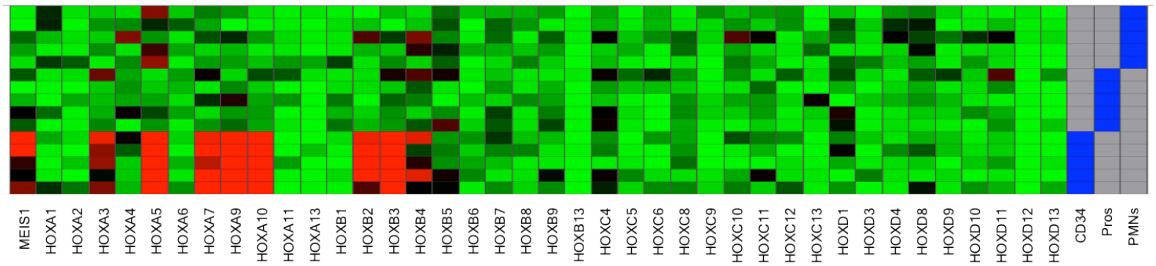

HOXB13
HOXB9
HOXB8
HOXB7
HOXB5
HOXB4
HOXB3
HOXB2
HOXB1
promoter/body
CpG island
enhancer
Methylation at probe CpG loci
0%
100%
Bone marrow from 190 acute myeloid leukemia patients
Normal
MLL.trans
MLL.PTD
t(15;17)
t(8;21)
inv(16)
other
NPM1
DNMT3A
FLT3.ITD
FLT3.other
IDH1/2
TET2
RNAseq

Figure 2-10

A.



B.

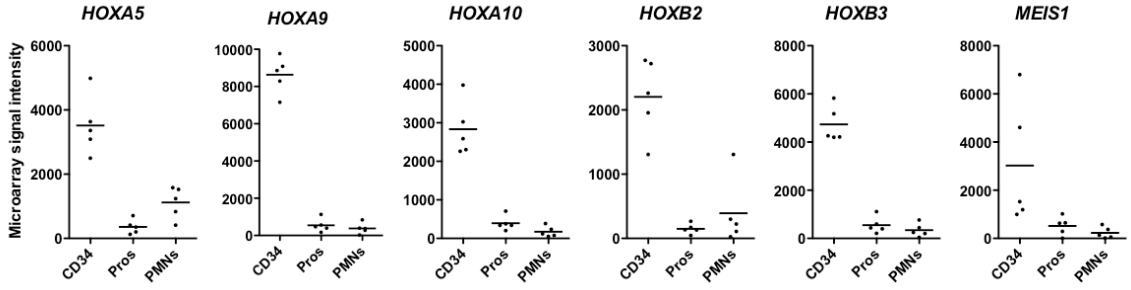


Figure 2-11

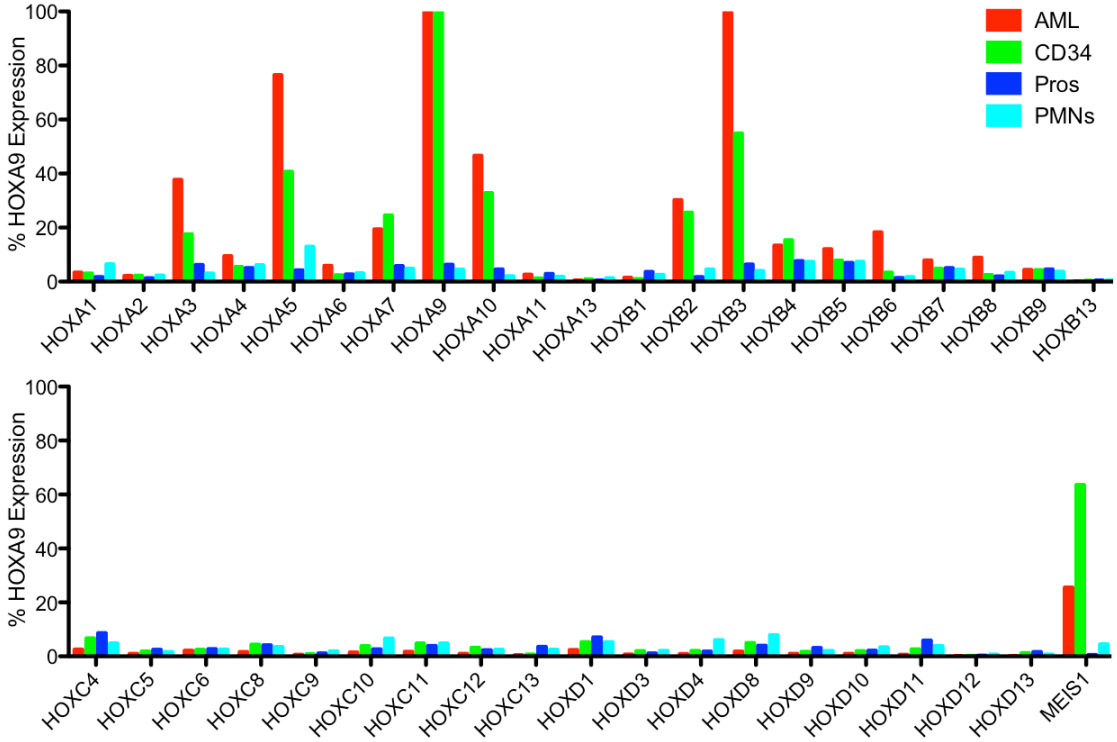

% HOXA9 Expression
100
80
60
40
20
0
AML
CD34
Pros
PMNs
HOXA1
HOXA2
HOXA3
HOXA4
HOXA5
HOXA6
HOXA7
HOXA9
HOXA10
HOXA11
HOXA13
HOXB1
HOXB2
HOXB3
HOXB4
HOXB5
HOXB6
HOXB7
HOXB8
HOXB9
HOXB13
% HOXA9 Expression
100
80
60
40
20
0
HOXC4
HOXC5
HOXC6
HOXC8
HOXC9
HOXC10
HOXC11
HOXC12
HOXC13
HOXD1
HOXD3
HOXD4
HOXD8
HOXD9
HOXD10
HOXD11
HOXD12
HOXD13
MEIS1

Figure 2-12

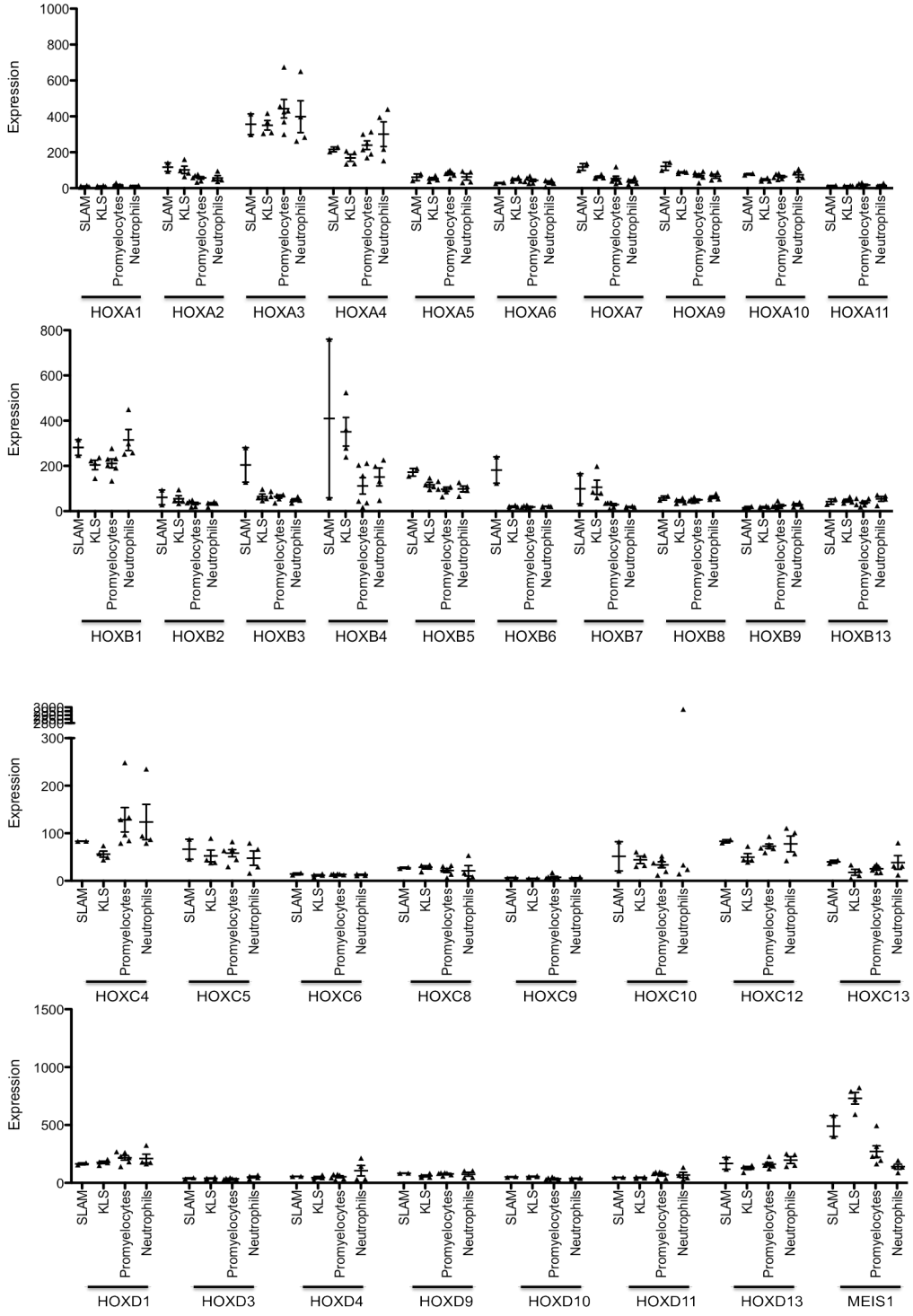
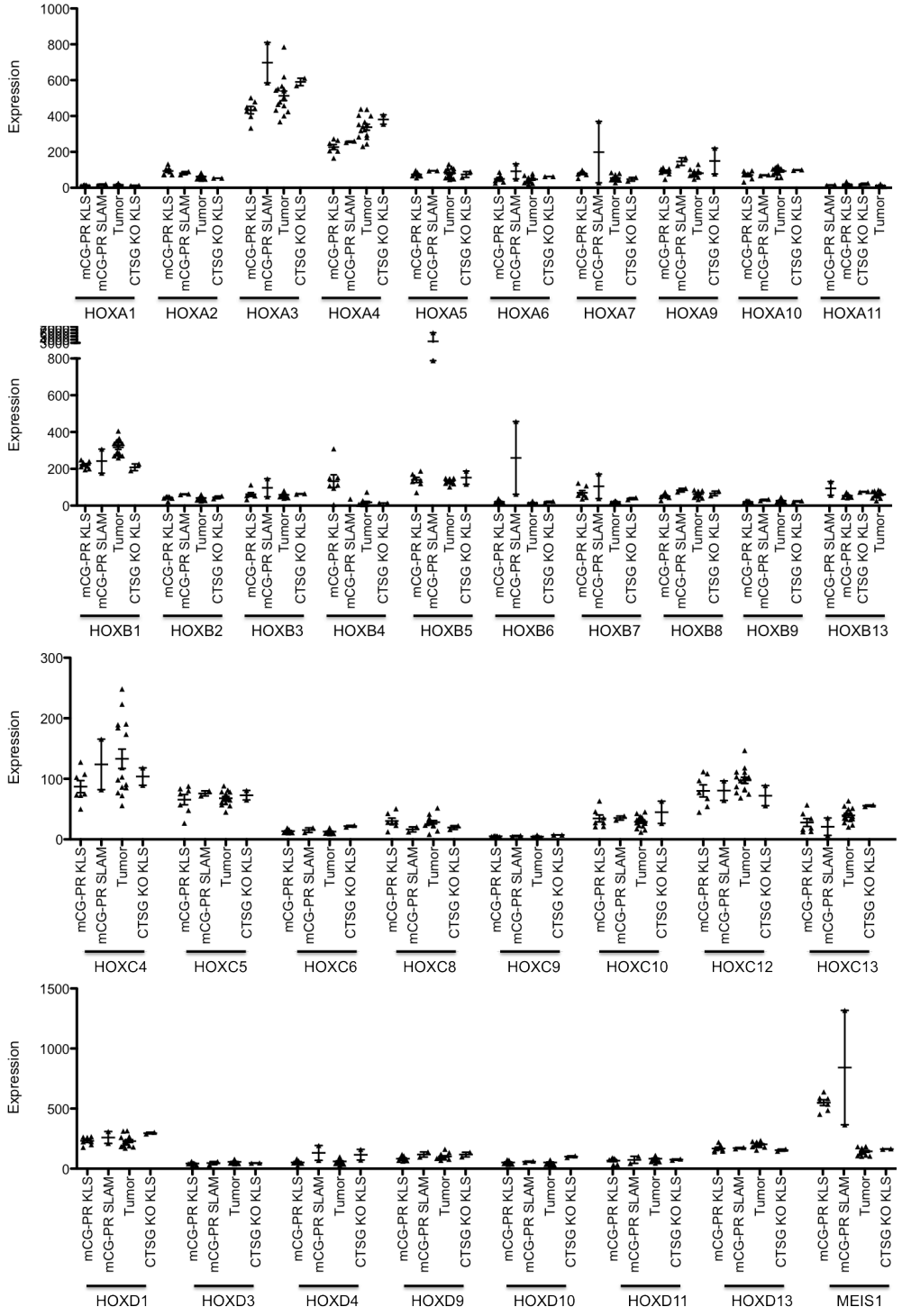

Expression
SLAM
KLS
Promyelocytes
Neutrophils
HOXA1
HOXA2
HOXA3
HOXA4
HOXA5
HOXA6
HOXA7
HOXA9
HOXA10
HOXA11
HOXB1
HOXB2
HOXB3
HOXB4
HOXB5
HOXB6
HOXB7
HOXB8
HOXB9
HOXB13
HOXC4
HOXC5
HOXC6
HOXC8
HOXC9
HOXC10
HOXC12
HOXC13
HOXD1
HOXD3
HOXD4
HOXD9
HOXD10
HOXD11
HOXD13
MEIS1

Figure 2-13


Expression
mCG-PR KLS
mCG-PR SLAM
Tumor
CTSG KO KLS
HOXA1
HOXA2
HOXA3
HOXA4
HOXA5
HOXA6
HOXA7
HOXA9
HOXA10
HOXA11
HOXB1
HOXB2
HOXB3
HOXB4
HOXB5
HOXB6
HOXB7
HOXB8
HOXB9
HOXB13
HOXC4
HOXC5
HOXC6
HOXC8
HOXC9
HOXC10
HOXC12
HOXC13
HOXD1
HOXD3
HOXD4
HOXD9
HOXD10
HOXD11
HOXD13
MEIS1

Chapter 3

Genetic heterogeneity of murine induced pluripotent stem (iPS) clones revealed by whole genome sequencing

Abstract

Use of integrative techniques for reprogramming of somatic cells into induced pluripotent stem (iPS) cells has led to concerns for their genomic integrity. Here we performed whole genome sequencing of 10 murine iPS lines produced in 3 independent experiments. Several hundred somatic nucleotide variants (SNVs) were identified in each clone, an average of 11 in protein-coding regions. The variant allele frequency of the first 2 experiments shows that the majority of SNVs are present in 50% of the reads, suggesting three possibilities for when these mutations could have arisen: 1) prior to reprogramming, 2) in a single burst at the time of reprogramming, or 3) in a burst after reprogramming leading to a selective advantage during expansion. In the third experiment, all four iPS lines contained 164 identical variants (6 protein-coding SNVs, 157 non-coding SNVs and 1 structural variant), as well as an average of less than 100 "private" SNVs unique to each clone. The common mutations were found in a subpopulation of the parental cells, proving that they were present before reprogramming. This suggests that one or more of these mutations may have been relevant for reprogramming fitness. For the private mutations, in addition to the majority of SNVs showing a 50% variant allele frequency, 3 of the clones have a subclone with variant allele frequencies of 20-30%. These subclones may represent acquisition of variants that have given a selective growth advantage to individual cells during post-reprogramming expansion. Our data suggest that most of the mutations detected in iPS cells occurred prior to reprogramming and are simply "captured" by cloning; however, some preexisting mutations provide an advantage for reprogramming, and may provide novel insights into the genetic underpinnings of this process.

Introduction

The discovery of methods to produce induced pluripotent stem (iPS) cells in 2006 has revolutionized the field of regenerative medicine. Yamanaka, *et al*. and Thomson, *et al*. showed that expression of a small set of transcription factors in mouse or human somatic cells is capable of reprogramming them to a pluripotent state [1-3]. Since then, there have been successful modifications of the reprogramming protocol that have decreased safety concerns, and increased the efficiency of the process. Reprogramming is now possible without the use of c-myc [4-6], and can be performed with single lentiviruses to limit integration dependent mutagenesis [4, 7, 8]; reprogramming has also been performed by transiently expressing specific cDNAs, RNA molecules, or proteins themselves [9-12]. Addition of demethylating agents and histone deacetylase inhibitors has also been shown to enhance iPS cell generation [5, 13, 14].

iPS cells provide novel models for the study of human diseases, and gene-corrected iPS cells may provide novel therapeutic reagents [15-17]. Daley and colleagues were the first to generate a panel of iPS lines from a variety of patients with complex diseases with genetic components, including Huntington's Disease (HD), Parkinson's Disease (PD) and juvenile-onset type I diabetes mellitus (JDM) [18]. Ellerby, *et al*. have been able to generate striatal neurons from the HD-iPS lines, which have the potential to be used for drug screening [19]. Since the Daley report, iPS lines have been generated from patients with a wide range of disorders [16, 20-29]. Jaenisch and Townes provided a proof of principle study in which they rescued a humanized sickle cell anemia mouse model by transplanting gene-corrected iPS cells that had been induced to undergo hematopoietic

differentiation [15]. Most recently, Cantz et.al. generated mice from tetraploid embryo aggregations of gene-corrected iPS cells, proving that genetic manipulation does not diminish the pluripotent phenotype of iPS cells [30].

Although advancements have been made in the generation of iPS cells, the mechanism behind reprogramming is not completely understood. Yamanaka proposed 2 potential models of reprogramming: "elite" vs. stochastic [17]. According to the elite model, reprogramming is inefficient because only a tiny fraction of cells within the donor pool are competent (or "fit") for reprogramming; perhaps less-differentiated cells within a heterogeneous population are more readily transformed than their fully differentiated counterparts. In contrast, the stochastic model suggested that all cells are equally capable of being reprogrammed with the correct balance of reprogramming factors. The stochastic model suggests that reprogramming is a purely epigenetic phenomenon. Reprogramming has also been shown to require new methylation patterns that presumably reflect the activation of endogenous pluripotency genes, and repression of differentiation genes from the parental cells. Ecker and colleagues defined the methylomes of iPS lines at single base resolution, and found that although they are globally similar to ES cells, iPS cells have regions of unique methylation [31]. Perhaps surprisingly, then, reprogramming of fibroblasts deficient in Dnmt3A and Dnmt3B (the *de novo* methyltransferases responsible for methylation in ES cells) revealed that neither of these genes are required for reprogramming [32]. This information, coupled with the fact that iPS cells have "memory" of the parental cells from which they were derived [33], suggests that there may be additional factors at play in iPS cell generation.

The genetics of iPS clones has been less well studied. Early, low-resolution studies showed that iPS lines generally have normal karyotypes [1, 3, 34, 35]. However, more recent analyses of iPS genomes suggest that there may be more subtle genetic consequences of reprogramming. Hall's group used whole genome sequencing data to detect structural variants (SVs) in 3 iPS lines derived from a single reprogramming experiment. They found a very small number of new SVs in the iPS lines, suggesting that reprogramming is not detrimental to genome stability [36]. Gore, *et al*. performed whole exome sequencing of 22 human iPS lines generated with multiple transcription factor delivery protocols in many different labs, using a variety of donor cell types. They detected an average of 5 non-synonymous point mutations per genome analyzed, which were enriched in genes found in the COSMIC database of cancer-associated genes [37, 38]; the authors concluded that iPS cell genomes often contain mutations that may be related to cancer pathogenesis, raising an important safety issue if these cells are to be used therapeutically.

In this study, we performed comprehensive whole genome sequencing to define all genetic events associated with reprogramming. We analyzed 10 murine iPS lines generated in 3 separate experiments. A polycistronic lentivirus was used for reprogramming, so that we could track integration events and genetically define each clone as unique. We found that all reprogrammed iPS lines have a set of several hundred mutations that are unique for each clone, with variant allele frequencies of ~50% for nearly all mutations. In one of the three experiments, however, each clone tested was

found to have a set of common mutations, along with a larger set of unique mutations; all of the mutations had variant allele frequencies of ~50%. These data strongly suggests that the MEF pool from which these cells were derived contained a tiny set of cells that were extraordinarily "fit" for reprogramming, probably due to the effects of one or more of the common mutations. The mutations found in the 10 iPS clones were not enriched in the COSMIC database. In sum, our data suggests that most mutations in iPS cells are random and benign, occurring during the growth and development of the fibroblasts that gave rise to each iPS clone. The mutational "footprint" of the fibroblasts that are reprogrammed are "captured" by the cloning event itself. Preexisting mutations in somatic cells should be captured by any reprogramming strategy during the cloning process; clearly, some may be relevant for reprogramming fitness. By sequencing large number of iPS clones, pathways that faciliate fitness may become evident, and exploitable to improve reprogramming efficiency and safety.

Materials and Methods

Production of iPS clones

We generated iPS clones from 3 parental cell lines: 1) wildtype mouse embryonic fibroblasts (MEFs), 2) wildtype tail tip fibroblasts (TTFs), 3) GusB -/- MEFs. iPS clones were generated from each line as previously described [4]. Briefly, $3x10^5$ of each cell type were seeded on 6 well plates and allowed to grow overnight. The next day, the cells were transduced with the OSK lentivirus (kindly provided by Tim Townes) at an MOI of 1-5. The cells were incubated with virus for 48 hours then trypsinized and transferred to a 100-mm dish without a feeder MEF layer. Cells were grown for 2-3 weeks with daily media changes. After 17-26 days (see Table 3-1) individual colonies were picked and expanded on MEF feeder layers.

Pluripotency characterization of iPS clones

All 10 iPS clones were assessed for ES cell-like morphology and stained for alkaline phosphatase according to the manufacturer's instructions (Millipore?). Cells were analyzed by flow cytometry for ES cell markers, SSEA-1, Nanog and Oct4. For intracellular Nanog and Oct4 detection cells were fixed with 4% paraformaldehyde and permeabilized with 1% saponin. In addition to these basic assays, the 4 clones from the third experiment underwent further analysis.

Total RNA comparison of ES cells vs. iPS cells. RNA was analyzed on the Mouse Exon 1.0 ST array (Affymetrix, Santa Clara, CA) as previously described [39]. Briefly, 100 individual colonies of each clone and ES cell controls were picked and RNA was purified

using TRIzol (Invitrogen, Carlsbad, CA) and amplified using the whole transcript WT-Ovation RNA Amplification System and biotin-labeled (NuGen Technologies, San Carlos, CA) (each samples prepared in duplicate). Amplified RNA was then applied to the Mouse Exon 1.0 ST array according to standard protocols from the Siteman Cancer Center, Molecular and Genomic Analysis Core Facility (http://pathology.wustl.edu/research/cores/lcg/index.php). Affymetrix Expression Console software was used to process array images, export signal data, and evaluate image and data quality relative to standard Affymetrix quality control metrics. Spotfire analysis software (TIBCO, Somerville, MA) was used for unsupervised hierarchical clustering of ES and iPS global RNA levels. For lentivirus insertion site analysis of experiment 3 clones, the best probe (defined as the probe with the highest average expression level) for each nearest neighbor gene was plotted.

Oct4 and Nanog promoter methylation status. DNA was prepared from 100 individually picked clones (to minimize contaminating feeder layer MEFs) using the QIAamp DNA Mini Kit according to the manufacturer's instructions (Qiagen, Valencia, CA). DNA samples (1 ug) were then converted with bisulfite using the Zymo Research EZ DNA Methylation Kit according to the manufacturer's instructions (Zymo Resarch, Irvine, CA). Bisulfite specific primers were used to amplify the promoter regions of Oct4 (5'-TCCAACCCTACTAACCCATCACC -3' and 5'- GGTTTTTTAGAGGATGGTTGAGTG -3') and Nanog (5'- ACCAAAAAAACCACACTCATATCAATATA -3' and 5'- GATTTTGTAGGTGGGAT-TAATTGTGAATTT -3') followed by deep digital sequencing on the 454-Flx platform. Each CpG dinucleotide was covered by >1000 reads, and the percentage of methylated C

residues were determined for each position.

Cystic teratoma formation. NOG mice were injected in the hindflank with 1 million iPS cells from each of the four iPS lines. Tumors were harvested after 5-6 weeks of growth and sectioned. Sections were H&E stained and analyzed. In all tumors, there were tissues of ectodermal, mesodermal, and endodermal origin noted. Examples of tissues of endodermal origin included ciliated pseudostratefied epithelium or columnar epithelium. Examples of tissues of mesodermal origin included bone, cardiac muscle, or cartilage. Fibrous connective tissue was also noted in some that did not appear to represent native tissue. Examples of tissues of ectodermal origin were components of the nervous system, particularly the MPS samples appeared to have abundant tissue representative of embryonal brain. The ES sample contained brain tissue that appeared better differentiated. Some tissues also had components of epidermis.

llumiIna whole genome shotgun library construction.

We followed the standard library construction procedure as previously described [40]. Briefly, we started with 100ng of DNA from each clone and the MEFs. After the DNA was fragmented, end repaired, and ligated with adaptors, we amplified the products and ran them on a polyacrylamide gel. We then excised gel slices between 300-350 bp and 450-500 bp for each sample, and eluted the DNA from each gel slice as the source for independent libraries.

Illumina exome library construction and capture

Illumina sequencing libraries were constructed according to the manufacturer's protocol (Illumina Inc, San Diego, CA) with the following modifications: 1) DNA was fragmented using a Covaris S2 DNA Sonicator (Covaris, Inc. Woburn, MA) to range in size between 100 and 500bp. 2) Illumina adapter-ligated library fragments were amplified in four 50μl PCR reactions for eight cycles. 3) Solid Phase Reversible Immobilization (SPRI) bead cleanup was used to remove primers from the PCR and select for 300-500bp fragments. Sequencing libraries were hybridized with the SureSelect^{XT} Mouse All Exon Kit (Agilent), which captures 49.6 Mb of the coding sequence from ~24,000 genes. KAPA qPCR was used to determine the quantity of library necessary to produce cluster counts appropriate for the Illumina HiSeq 2000.

Illumina sequencing.

After diluting the libraries to a 10pM concentration, we utilized the paired-end flow cell and cluster generation kits to produce flow cells with an average cluster density ranging between 1.9 - 3 million clusters per tile. We employed the standard sequencing kits (Illumina HiSeq Sequencing Kit) and performed 100 cycles of nucleotide incorporation. Following this first round of end sequencing, the flow cell was treated to remove the synthesized fragments, clusters were re-amplified, and the resulting fragments were linearized. We then annealed the second sequencing primer, and initiated another round of sequencing by synthesis to complete the read pairs. The read length for the second sequencing round matched that of the first. Following round two, we utilized an Illumina HiSeq 2000 machine and the Illumina sequencing pipeline, version 1.7 or 1.8, to analyze

the data and produced files containing high quality ("passed filter") reads with associated quality values.

Clonality analysis

Sub-clonality estimates were determined using the mutation allele frequencies from whole genome sequencing. To minimize the effect of coverage outliers from likely false positives, we pre-filtered each site to ensure that the coverage fell within ±2 median absolute deviations from the median coverage of all non-repetitive predictions within each clone. We drew a kernel density estimate (KDE) plot for variant allele frequencies using the density() function in R. A customized R function evaluated each KDE plot to determine the number of significant peaks, which served as an estimation of the number and relative composition of different sub-clones present within each iPS clone.

Significantly mutated gene analysis and pathway analysis

Since the mutational process in clones is analogous to the mutational process of a tumor, we used components of the unpublished Mutational Significance in Cancer (MuSiC) package to determine significantly mutated genes (SMG) and pathways. The SMG component of MuSiC assigns mutations to various categories, such as transition or transversion, and then uses methods including convolution, Fisher's test, and a likelihood test to combine the category-specific binomials to obtain an overall P-value. The result is appreciably more accurate than if these attributes were disregarded. The pathway analysis component of MuSiC is an implementation of PathScan[41].

Sequence Analysis

Reads were aligned using BWA 0.5.5[42] with quality trimming set at 5 to the NCBI build 37 of *Mus musculus* reference sequence augmented with an additional contig representing the complete OSK sequence. The resulting alignments were de-duplicated using Picard (http://picard.sourceforge.net) and quality recalibrated using the GATK[43] base quality recalibrator (v1.0.3471).

We called potential somatic single nucleotide variants using a modified version of SomaticSniper[44] to account for perfectly pure samples. Variants with a somatic score greater than 10 were filtered on a number of sequence features to remove false positives as described elsewhere[45]. Sites passing these filters were additionally filtered to remove variants where the difference between the clone and control variant allele frequency was 30% or less and the read depth was 10 reads or less. Non-genic variants in repetitive regions as identified from the UCSC Genome Browser[46] for mouse build37 were also excluded from further analysis. Variants passing all filters were annotated as previously described[47].

Sites chosen for validation were manually reviewed and validated using 454 sequencing as previously described[48]. Read counts were calculated by excluding reads where the called base was below a Phred quality of 15 for WGS data, with a mapping quality of 10 or a base quality less than 20 for the high depth Illumina data, and for a base quality less than 20 for the 454 data.

Indels were called using the GATK[43]. OSK insertion sites were predicted using BreakDancer[49] and SquareDancer, an in-house implementation of the CREST algorithm[50]. Copy number predictions were generated as previously described[47] and manually reviewed to determine their veracity.

Detection of common variants in parental MPSVII MEFs.

Rare variants were detected in the parental MEFs by novel StuI sites are generated by the *Apaf1* G16A and *Sbno2* A3783G variants. Regions surrounding each variant were amplified (Apaf1: 884 bp amplicon, 5'-CCCAAACACTTTGATGAACGA-3' and 5;-CTATAAGGACCTTGCTGCGC-3' ; Sbno2: 806 bp amplicon, 5'-GCTGCAGACTGACACA-GGAG-3' and 5'-AGCAGAGGCTCCCATGACTA-3'). PCR products were cloned into the pCR2.1 vector according to the manufacturer's instructions (Invitrogen, Carlsbad, CA). The number of clones in individual transformations was counted and each plate was pooled as a single sample (see Table 3-5 for pool sizes), creating mini-libraries of the parental MEFs. These mini-libraries were then digested with EcoRI to extract the amplicon from the pCR2.1 vector and StuI to detect the presence of the variant allele (Apaf1: 391 and 493 bp products; Sbno2: 454 and 352 bp products). Digestion products were run on 5% acrylamide gels and transferred to nitrocellulose. Nitrocellulose blots were probed with ^{32}P-labeled probes (Apaf1: the 493 bp product; Sbno2: the 352 bp product).

Additional MPSVII TTF iPS production and analysis.

iPS cells were produced from tail tip fibroblasts of an adult MPSVII mouse as described above for experiments 1-3. DNA was prepared as described above and the Apaf1 and Sbno2 regions were amplified as for the detection of variants in rare parental cells. PCR products were directly Sanger sequenced and analyzed for the variant allele.

Results

Experimental system

To investigate the genetic consequences of transcription factor-mediated reprogramming, we generated murine iPS lines using a established polycistronic lentivirus encoding three of the original four Yamanaka transcription factors, Oct4, Sox2 and Klf4 (OSK) [4]. We reprogrammed 3 independent fibroblast lines, all derived from C57Bl/6 mice; wildtype mouse embryonic fibroblasts (MEFs) and tail tip fibroblasts (TTF) in experiments 1 and 2, respectively, and MEFs derived from a disease model (murine mucopolysaccaridosis type VII- MPSVII) in the third experiment. Details of cell culture and transduction, followed by clone picking and expansion of individual clones, are provided in the Materials and Methods, and listed in **Table 1**.

All clones were examined for morphology and alkaline phosphatase positivity, as well as surface expression of the pluripotency markers SSEA-1, Nanog and Oct4. All clones had characteristics of pluripotent cells (**Table 1**). Since experiment 3 utilized MEFs derived from a disease model known to have growth and developmental defects (MPSVII, caused by a loss-of-function mutation in the *GusB* gene), we characterized the pluripotency of these iPS lines further [51, 52]. The Affymetrix Mouse Exon 1.0ST array was used to compare expression patterns in iPS lines and ES cells. Unsupervised hierarchal clustering analysis showed that the iPS clones and ES cell lines clustered randomly, suggesting that their global expression patterns are highly similar (**Figure 1**). The methylation status of the Oct4 and Nanog promoters was analyzed by bisulfite modification of genomic DNA from each of the four lines, along with ES and MEF controls. The promoter region of

each gene was amplified with bisulfite-specific primers, followed by deep digital sequencing on the 454-Flx platform. Each CpG dinucleotide was covered by >1000 reads, and the percentage of methylated C residues were determined for each position. **Figure 2** shows that the promoters were extensively methylated in MEFs, but not in ES or iPS cells. Lastly, NOG mice were injected in the hindflank with 1 million iPS cells from each of the four iPS lines; and line formed cystic teratomas containing all 3 germ layers (**Figure 3**).

Genomic Architecture of iPS Clones

Genomic DNA from each iPS line and their parental cells were used to make libraries that were subsequently sequenced with a paired-end approach on the Illumina GAIIX platform. Each sample was sequenced to a depth of 16x haploid coverage; a lower level of coverage is required for mutation discovery for inbred organisms (**Table 2**) [53]. The polycistronic OSK lentivirus was used for reprogramming to create genetic marks that would define each clone as unique. Indeed, each iPS line had 1-5 unique lentiviral insertion sites (**Table 2** and **Figure 4A**). For experiment 3, we used expression array data from each clone to demonstrate that the lentiviral insertions did not detectably alter expression of their nearest neighbor genes in any clone or insertion site (**Figure 4B**).

We first identified homozygous single nucleotide variants (SNVs) in the three parental cell lines (with respect to the reference C57Bl/6 genome). Each line had more than 2,000 homozygous SNVs, many of them within coding regions (**Figure 5**), illustrating the genetic drift that exists among B6 mice over time. This result shows the critical

importance of sequencing a matched parental control sample to identify "private" variants in individual iPS clones. Compared to its own parental line, each iPS clone contained a range of 176-701 SNVs (**Figure 6A**), 3-19 of which were within gene coding regions (**Figure 6B, Table 3**). Plotting the location of each SNV within the genome suggested that most variants are randomly distributed (**Figure 7**); a chi-squared analysis of the spatial distribution of mutations confirmed that the distribution was random.

The digital readcounts provided by whole genome sequencing allows for calculation of variant allele frequency for all SNVs. Variant allele frequency plots of all SNVs for all 10 iPS lines revealed that the variant frequency is distributed around a mean of ~50% (**Figure 8**), and that the distribution is consistent with a binomial distribution. This data suggests that the SNVs are probably heterozygous, and present in virtually all the cells in the iPS sample tested. Further, the presence of a single dominant clone in each sample suggests that all of the mutations arose at or before the time of reprogramming.

Mutational consequences of iPS SNVs

All of the SNVs reported within coding regions were secondarily validated by exome sequencing for experiments 1 and 2, or by 454 sequencing of PCR amplified regions for experiment 3 (**Table 3**). Each iPS genome contained an average of 10 coding region SNVs per genome (including missense, nonsense, splice site and silent mutations). In addition to the coding mutations, each iPS clone contained 186-679 predicted SNVs in non-coding regions of the genome (**Figure 6**). For the three iPS clones in experiments 1

and 2, all of the SNVs were unique to each genome, and there was no overlap among the variants detected in each experiment.

Among the 336-409 total SNVs detected in each iPS genome in experiment 3, however, there were 162 validated SNVs (5 within coding regions) that were detected in all four iPS lines (**Figure 9**). We tested for the 5 common coding region SNVs in two additional iPS lines generated from the same experiment, and both were heterozygous for all 5 variants (data not shown). To determine whether this set of common variants was not a consequence of reprogramming fibroblasts derived from the *GusB* -/- strain, we did a second reprogramming experiment using *GusB* -/- tail tip fibroblasts (TTFs) from a different mouse. We screened 10 individual iPS clones from the second reprogramming experiment for two of the shared mutations (*Apaf1* G16A and *Sbno2* A3783G), and did not detect either SNV in any clone (**Figure 10**).

To further confirm that the four iPS clones from experiment 3 had a common set of genetic variants, we also searched for common indels and structural variants. Somatic indel prediction analysis identified 32 variants that were present in all four iPS clones (**Table 4**); none had a translational effect. We also identified a common amplified region of ~130 kilobases on chromosome X that was present in all four iPS clones, but not detected in the parental MEFs (**Figure 11**). Two genes fall within this region, *Mug2* (a protease inhibitor) and *Gm10319* (a predicted gene).

The identification of a large set of shared genetic variants strongly suggests that all four of these iPS lines arose from a set of "founder" MEFs containing the same variants. However, none of the common SNVs were detected in the parental MEFs even with deep readcounts (which is limited by an error rate of 1%), suggesting that the founding population represents fewer than 1% of the total cells in the MEF pool. In an attempt to detect these rare cells, we took advantage of novel StuI restriction sites created by two of the common coding region SNVs (i.e. *Apaf1* G16A and *Sbno2* A3783G). We amplified regions containing these variants with PCR, and then cloned pools of these amplicons into the pCR2.1 vector (Invitrogen, Carlsbad, CA). DNA derived from pools containing hundreds of unique amplicons were then digested with StuI, and the digestion products were detected by Southern blot analysis. **Figure 12** shows representative blots of the Apaf1 and Sbno2 analyses. DNA from a pool of clones from one of the iPS cell lines reveals that ~50% of the starting DNA is digested by StuI. However, most pools of clones derived from the MEFs reveal no digestion products; only occasional pools contain very small amounts of the digestions products, suggesting that no more than one or two plasmids within the pool contains the variant allele. **Table 5** shows the total number of MEF pools screened for each variant, and the total number of clones screened in this pool.

Although these results are semiquantitative, it is clear that less than 1 MEF cell in 1,000 contains the common variants detected in all of the iPS clones. Further, these results make it extremely unlikely that the common variants arose after reprogramming event.

In addition to the common variants in the iPS clones of experiment 3, there were also 299-535 unique SNVs in each of the clones (**Figure 9**). To assess whether the private variants were acquired before or after the reprogramming event, we compared the variant allele frequencies of common SNVs vs. private SNVs. The common variants show the same variant allele frequency distribution seen for the other experiments in **Figure 8**. While the majority of the private SNVs are also present in ~50% of reads, clones 1, 2 and 6 have a subset of variants present in only 20-30% of reads, suggesting presence of a subclone (**Figure 13**).

Other than the common variants in experiment 3, there was no overlap in the genes containing SNVs among the three experiments. We also compared all of the genes with coding region SNVs in our dataset to that of Gore, *et al*. where the exomes of 23 human iPS lines were sequenced. A single gene, *ATM/Atm*, contained an SNV in both datasets: in our study, clone 5 from experiment 1 contains a splice site mutation at the 5' end of intron 23; one of the 22 human iPS clones contained a nonsynonymous SNV in *ATM* (X444Y) [37].

Gore, *et al*. also reported their human iPS clones displayed a significant enrichment of mutations in genes found in the COSMIC database [37]. We therefore examined the 89 unique genes with coding region SNVs in our 10 iPS clones, and found that that 36 (40.4%) were in the COSMIC database (**Table 3**). We also searched the COSMIC database for the 247 genes with homozygous variants in the three parental mice used in this study (compared to the B6 reference genome); 36.1% of these genes are likewise in

COSMIC, a value that is not significantly different from that of the iPS clones (p=0.5733).

We also performed a pathway analysis using the MuSiC suite to test for significantly mutated genes (SMGs), and common pathways that might be affected by mutations among the 10 clones. The only SMGs were the 4 common missense variants identified in experiment 3. A total of 50 pathways contained genes that were mutated (**Table 6**). Only 13 of these had a P-value of less than 0.05; all included one of the common variants from experiment 3. There were no pathways with hits in all 10 clones. This analysis suggests that in these 10 iPS lines, there is no common pathway that contains mutations that faciliate reprogramming.

Discussion

In this study, we have defined the genetic landscape of 10 independent murine iPS cell lines by whole genome sequencing. A polycistronic lentivirus was used for reprogramming so that virus integration sites could provide definitive marks for each clone; the absence of overlapping insertion sites shows that insertional mutagenesis does not target specific endogenous loci for successful reprogramming. In the first two experiments, we identified several hundred SNVs in each individual clone (with an average of 11 in coding sequences), but found no overlap among the mutations in any of the clones. In the third experiment, however, we identified a subset of common variants in all four iPS clones, as well as a set of "private" variants in each. Using genomic DNA from the parental MEFs used to create these clones, we were able to detect two of the common mutations at a very low frequency, demonstrating that rare "founder" cells within the total MEF population contained these mutations (i.e. they did not arise during reprogramming or expansion of the cells); this suggests that one or more the common mutations may have contributed to the extraordinary "fitness" of these cells for reprogramming. Using the mutational profiles and variant allele frequencies from all 10 clones, we were able estimate the timing of SNVs associated with reprogramming. Our data suggests that many SNVs may occur at or before the time of reprogramming, and that some may occur during expansion of the iPS clones in culture. Regardless, all iPS cell lines contain hundreds of mutations. Although most are probably functionally irrelevant (reflecting random genetic events in the starting cells that are "captured" by the process of cloning individual cells), some may contribute directly to reprogramming fitness.

The use of whole genome sequencing, with assessment of the variant allele frequencies of hundreds of mutations, is a very powerful tool that has allowed us to begin to assess whether the mutations detected in iPS clones are caused by reprogramming (i.e. the mutations are new), or whether they simply represent the genetic "signature" of the cell that was reprogrammed (i.e. the mutations were present before reprogramming, and were "captured" by virtue of the cloning event). In the first two experiments, this issue could not be resolved. All SNVs in all six clones were unique and distributed randomly in the genome, with variant allele frequencies of ~50%, suggesting that all the SNVs were heterozygous, and present in all the iPS cells. Three scenarios could explain these results: 1) the SNVs preexisted in the cell that was reprogrammed, and reflect the background mutations that were present in that cell, or 2) the variants all arose in a "burst" of mutational activity at the time of reprogramming, but they were not relevant for positive selection (which was provided by the reprogramming factors), or 3) the SNVs all arose in a burst of mutational activity after reprogramming; one or more mutations was important for selection, and the entire group of mutations was "captured" as a set in the iPS clone. Based on our knowledge of the background rate of mutations in somatic cells [54, Ley, *et al*., manuscript in preparation], and the large numbers of mutations detected in each clone, we suspect that the latter two scenarios are unlikely. Regardless, they cannot be ruled out with the data in the first two experiments.

However, in the third experiment, we detected a very large set of common mutations in all four iPS clones that were sequenced, including SNVs, indels, and a structural variant;

we detected 5/5 coding SNVs tested in two other iPS clones generated in the same experiment (importantly, these common variants were not detected in 10 iPS clones derived from another reprogramming experiment, using fibroblasts from an independent GusB deficient mouse, i.e. they are not related to the mouse strain itself). Using techniques that could detect point mutations in rare cells, we were able to detect two of the common mutations at a very low frequency in the starting MEF pool, suggesting that fewer than 1 MEF in 1000 contained the common mutation set. Since all six iPS lines obtained from the starting pool of 300,000 cells contained the same set of mutations, it is clear that these cells were much more likely to undergo reprogramming than the 'average' MEF in the pool. Although is not yet clear which of the mutations might have been responsible, some candidates are attractive (e.g. the Apaf1 mutation may alter a threshold for apoptosis during reprogramming stress). Sequencing the genomes of large numbers of iPS clones from independent pools of starting cells may allow for the identification of recurring mutations (or pathways) in iPS clones; these data may help to elucidate the genetic barriers to reprogramming, and provide novel approaches for improving the efficiency of the process.

Some data pertinent to this idea is already available. Gore, *et al*. performed whole exome sequencing on 22 human iPS lines that were created with a variety of reprogramming approaches (including 4-factor retroviral and lentiviral, 3-factor retroviral, episomal vector and messenger RNA). Among this set, 7 pairs of lines were generated from the same parental cells; in 3/7 pairs, the authors detected common SNVs (1-3 common mutations were detected, along with a few private mutations in each of the members of

the pair). None of these mutations was detected in other human iPS clones, and none were detected in our common set of mutations. Even though exome sequencing of iPS clone pairs should be less sensitive than whole genome sequencing of iPS clone trios, it is striking that 4 out of 10 sets of iPS clones derived from the same parental cells have shown common variants; this is not a rare event. Although the significance of these mutations is unknown, they are clearly not providing an overall selective advantage for cells before they are reprogrammed (since they are rare in the starting population). However, one or more of the common mutations would be expected to be capable of cooperating with reprogramming factors, raising a concern that cells derived from these iPS clones may be different from those derived from somatic cells without cooperating mutations.

As noted above, we evaluated variant allele frequencies in the iPS clones to help understand clonality and the timing of mutational events. The variant allele frequencies of all 10 clones averaged 50%, and fit a binomial distribution (**Figure 8**), suggesting that most variants are heterozygous, and present in virtually all cells in the sample. In experiment 3, we further examined the variant allele frequencies of the common vs. private mutations in each clone (**Figure 13**). The common variants occurred prior to reprogramming, and have an average variant allele frequency of 50%, as expected. The private variants also have a major peak of variant allele frequencies at 50%, but three clones (1, 2, and 6) also had a minor, secondary peak at 20-30%. These data suggest that the private mutations in each clone must have arisen after the common mutations, but still occurred at or before the time of reprogramming (since most variants are present in all

cells). However, the minor peak of variant allele frequency in three of the clones suggests that a third set of mutations arose in a subset of cells after reprogramming. Some of the mutations present in these subclones must allow for positive selection of these cells, allowing for their detection in the total pool of cells. Since we sampled clones at a single time in their existence, we do not know whether these late subclones are becoming dominant, are failing, or are stable; serial sampling would be required to resolve this question. However, these results do suggest that mutations in iPS clones may evolve with serial passaging, as previously suggested by Laurent, *et al.* [55].

Based on this information, we propose a model of how mutations are acquired during iPS cell development (**Figure 14**). Most of the time, individual iPS clones arise from unique cells within the transduced parental cell population. Each cell has a unique set of pre-existing background mutations, which are "captured" by the expansion and cloning of single cells with reprogramming (Panel A). However, there are some instances where pre-existing mutations are relevant for a cell's reprogramming susceptibility. These "super fit" cells are much more likely to be reprogrammed and detected as iPS clones, since their background mutations must cooperate with reprogramming factors (Panel B). Between the time when the fitness mutations are acquired and the reprogramming event occurs, additional private mutations are acquired (which may or may not further increase reprogramming efficiency). Finally, additional mutations can be acquired after reprogramming that provide additional selective pressure for the outgrowth of subclones.

Although one or more of the common mutations found in some iPS clones are very likely to be relevant for reprogramming, it is less clear whether the private mutations are. The private mutations are randomly distributed in genome space, which explains why so few fall in coding sequences (which comprise about 1% of the genome). The number of mutations detected in each iPS exome is about the same, regardless of whether the starting cells are from disease models, normal mice, or human volunteers. The method of reprogramming (even with non-integrating methods) did not affect total numbers of exonic mutations detected in the Gore, *et al.* study, and the numbers detected in our murine iPS clones were similar to theirs. Although Gore, *et al.* suggested that iPS mutations were enriched in cancer-associated genes in their human iPS clones, our study revealed that the percentage of exonic mutations found in the COSMIC database is the same for our iPS clones and the random background mutations that were detected in the three different B6 mice used to generate the clones (which arise because of genetic drift within inbred mice; $p=0.5733$). All in all, these additional observations support the model proposed above, where most of the mutations detected in iPS clones probably occurred before reprogramming, and are benign, irrelevant events that reflect the mutational history of the cell that was cloned in the reprogramming process.

Although there is a range of reprogramming fitness within a given population of somatic cells, these data show that it may be in part based on random genetic variants that occur within individual somatic cells, suggesting several important caveats for the field of somatic cell reprogramming. In the act of cloning individual somatic cells, background genetic variants will invariably be 'captured' by the cloning act itself, regardless of the

strategy used to induce reprogramming. If fitness mutations are present in rare somatic cells, they could potentially cooperate with reprogramming factors, regardless of how the factors are delivered (stably integrated viruses, transiently expressed plasmids, mRNAs, or proteins). Although this may have important consequences for the use of iPS clones in therapeutic settings, the sequencing of large numbers of iPS clones may also provide a new strategy for identifying genes that represent barriers for reprogramming, and suggest approaches to safely overcome them. Perhaps most importantly, these results strongly suggest that reprogramming is not a mutagenic event *per se,* and that knowledge gained from sequencing iPS cell genomes may help to refine the process and make it safer for therapeutic purposes.

References

1. Takahashi, K., et al., *Induction of pluripotent stem cells from adult human fibroblasts by defined factors.* Cell, 2007. **131**(5): p. 861-72.
2. Takahashi, K. and S. Yamanaka, *Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors.* Cell, 2006. **126**(4): p. 663-76.
3. Yu, J., et al., *Induced pluripotent stem cell lines derived from human somatic cells.* Science, 2007. **318**(5858): p. 1917-20.
4. Chang, C.W., et al., *Polycistronic lentiviral vector for "hit and run" reprogramming of adult skin fibroblasts to induced pluripotent stem cells.* Stem cells, 2009. **27**(5): p. 1042-9.
5. Huangfu, D., et al., *Induction of pluripotent stem cells from primary human fibroblasts with only Oct4 and Sox2.* Nature biotechnology, 2008. **26**(11): p. 1269-75.
6. Wernig, M., et al., *c-Myc is dispensable for direct reprogramming of mouse fibroblasts.* Cell stem cell, 2008. **2**(1): p. 10-2.
7. Carey, B.W., et al., *Reprogramming of murine and human somatic cells using a single polycistronic vector.* Proceedings of the National Academy of Sciences of the United States of America, 2009. **106**(1): p. 157-62.
8. Shao, L., et al., *Generation of iPS cells using defined factors linked via the self-cleaving 2A sequences in a single open reading frame.* Cell research, 2009. **19**(3): p. 296-306.
9. Stadtfeld, M., et al., *Induced pluripotent stem cells generated without viral integration.* Science, 2008. **322**(5903): p. 945-9.
10. Okita, K., et al., *Generation of mouse induced pluripotent stem cells without viral vectors.* Science, 2008. **322**(5903): p. 949-53.
11. Warren, L., et al., *Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA.* Cell stem cell, 2010. **7**(5): p. 618-30.
12. Zhou, H., et al., *Generation of induced pluripotent stem cells using recombinant proteins.* Cell stem cell, 2009. **4**(5): p. 381-4.
13. Mikkelsen, T.S., et al., *Dissecting direct reprogramming through integrative genomic analysis.* Nature, 2008. **454**(7200): p. 49-55.
14. Huangfu, D., et al., *Induction of pluripotent stem cells by defined factors is greatly improved by small-molecule compounds.* Nature biotechnology, 2008. **26**(7): p. 795-7.
15. Hanna, J., et al., *Treatment of sickle cell anemia mouse model with iPS cells generated from autologous skin.* Science, 2007. **318**(5858): p. 1920-3.
16. Raya, A., et al., *Disease-corrected haematopoietic progenitors from Fanconi anaemia induced pluripotent stem cells.* Nature, 2009. **460**(7251): p. 53-9.
17. Yamanaka, S., *Elite and stochastic models for induced pluripotent stem cell generation.* Nature, 2009. **460**(7251): p. 49-52.
18. Park, I.H., et al., *Disease-specific induced pluripotent stem cells.* Cell, 2008. **134**(5): p. 877-86.

19. Zhang, N., et al., *Characterization of Human Huntington's Disease Cell Model from Induced Pluripotent Stem Cells.* PLoS currents, 2010. **2**: p. RRN1193.
20. Ye, Z., et al., *Human-induced pluripotent stem cells from blood cells of healthy donors and patients with acquired blood disorders.* Blood, 2009. **114**(27): p. 5473-80.
21. Soldner, F., et al., *Parkinson's disease patient-derived induced pluripotent stem cells free of viral reprogramming factors.* Cell, 2009. **136**(5): p. 964-77.
22. Wang, Y., et al., *Generation of induced pluripotent stem cells from human beta-thalassemia fibroblast cells.* Cell research, 2009. **19**(9): p. 1120-3.
23. Carvajal-Vergara, X., et al., *Patient-specific induced pluripotent stem-cell-derived models of LEOPARD syndrome.* Nature, 2010. **465**(7299): p. 808-12.
24. Ebert, A.D., et al., *Induced pluripotent stem cells from a spinal muscular atrophy patient.* Nature, 2009. **457**(7227): p. 277-80.
25. Lee, G., et al., *Modelling pathogenesis and treatment of familial dysautonomia using patient-specific iPSCs.* Nature, 2009. **461**(7262): p. 402-6.
26. Moretti, A., et al., *Patient-specific induced pluripotent stem-cell models for long-QT syndrome.* The New England journal of medicine, 2010. **363**(15): p. 1397-409.
27. Jin, Z.B., et al., *Modeling retinal degeneration using patient-specific induced pluripotent stem cells.* PloS one, 2011. **6**(2): p. e17084.
28. Maehr, R., et al., *Generation of pluripotent stem cells from patients with type 1 diabetes.* Proceedings of the National Academy of Sciences of the United States of America, 2009. **106**(37): p. 15768-73.
29. Dimos, J.T., et al., *Induced pluripotent stem cells generated from patients with ALS can be differentiated into motor neurons.* Science, 2008. **321**(5893): p. 1218-21.
30. Wu, G., et al., *Generation of healthy mice from gene-corrected disease-specific induced pluripotent stem cells.* PLoS biology, 2011. **9**(7): p. e1001099.
31. Lister, R., et al., *Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells.* Nature.
32. Pawlak, M. and R. Jaenisch, *De novo DNA methylation by Dnmt3a and Dnmt3b is dispensable for nuclear reprogramming of somatic cells to a pluripotent state.* Genes Dev. **25**(10): p. 1035-40.
33. Kim, K., et al., *Epigenetic memory in induced pluripotent stem cells.* Nature, 2010. **467**(7313): p. 285-90.
34. Wernig, M., et al., *In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state.* Nature, 2007. **448**(7151): p. 318-24.
35. Park, I.H., et al., *Generation of human-induced pluripotent stem cells.* Nature protocols, 2008. **3**(7): p. 1180-6.
36. Quinlan, A.R., et al., *Genome Sequencing of Mouse Induced Pluripotent Stem Cells Reveals Retroelement Stability and Infrequent DNA Rearrangement during Reprogramming.* Cell stem cell, 2011. **9**(4): p. 366-73.
37. Gore, A., et al., *Somatic coding mutations in human induced pluripotent stem cells.* Nature, 2011. **471**(7336): p. 63-7.

38. Forbes, S.A., et al., *The Catalogue of Somatic Mutations in Cancer (COSMIC).* Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.], 2008. **Chapter 10**: p. Unit 10 11.
39. Wartman, L.D., et al., *Sequencing a mouse acute promyelocytic leukemia genome reveals genetic events relevant for disease progression.* The Journal of clinical investigation, 2011. **121**(4): p. 1445-55.
40. Mardis, E.R., et al., *Recurring mutations found by sequencing an acute myeloid leukemia genome.* The New England journal of medicine, 2009. **361**(11): p. 1058-66.
41. Wendl, M.C., et al., *PathScan: a tool for discerning mutational significance in groups of putative cancer genes.* Bioinformatics, 2011. **27**(12): p. 1595-602.
42. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
43. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nat Genet, 2011. **43**(5): p. 491-8.
44. Larson, D.E., et al., *Identification of Somatic Single Nucleotide Variants in Whole Genome Resequencing Data* 2011.
45. Koboldt, D.C., et al., *VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing.* 2011.
46. Karolchik, D., A.S. Hinrichs, and W.J. Kent, *The UCSC Genome Browser.* Curr Protoc Hum Genet, 2011. **Chapter 18**: p. Unit18 6.
47. Wartman, L.D., et al., *Sequencing a mouse acute promyelocytic leukemia genome reveals genetic events relevant for disease progression.* J Clin Invest, 2011. **121**(4): p. 1445-55.
48. Ding, L., et al., *Genome remodelling in a basal-like breast cancer metastasis and xenograft.* Nature, 2010. **464**(7291): p. 999-1005.
49. Chen, K., et al., *BreakDancer: an algorithm for high-resolution mapping of genomic structural variation.* Nat Methods, 2009. **6**(9): p. 677-81.
50. Wang, J., et al., *CREST maps somatic structural variation in cancer genomes with base-pair resolution.* Nat Methods, 2011. **8**(8): p. 652-4.
51. Meng, X.L., et al., *Induced pluripotent stem cells derived from mouse models of lysosomal storage disorders.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(17): p. 7886-91.
52. Sands, M.S. and E.H. Birkenmeier, *A single-base-pair deletion in the beta-glucuronidase gene accounts for the phenotype of murine mucopolysaccharidosis type VII.* Proceedings of the National Academy of Sciences of the United States of America, 1993. **90**(14): p. 6567-71.
53. Sarin, S., et al., *Caenorhabditis elegans mutant allele identification by whole-genome sequencing.* Nature methods, 2008. **5**(10): p. 865-7.
54. Warren, S.T., et al., *Elevated spontaneous mutation rate in Bloom syndrome fibroblasts.* Proceedings of the National Academy of Sciences of the United States of America, 1981. **78**(5): p. 3133-7.
55. Laurent, L.C., et al., *Dynamic changes in the copy number of pluripotency and cell proliferation genes in human ESCs and iPSCs during reprogramming and time in culture.* Cell stem cell, 2011. **8**(1): p. 106-18.

Figure Legends

Figure 3-1. Unsupervised cluster analysis of Mouse Exon 1.0 ST array data.
Each sample had two independent RNA samples (the second B6 Blu ES cell RNA sample was too low quality to put on the array). Samples are indicated below the heatmap. The MEFs segregate together and the iPS clones segregate with the ES cells.

Figure 3-2. Bisulfite sequencing of the Oct4 and Nanog promoters.
Genomic DNA from two separate preparations of MPS VII MEFs, B6 neo/hygro resistant (N/H) MEFs, MPS VII ES cells, B6 Blu ES cells and all 4 iPS lines was bisulfite treated with the Zymo EZ Methylation-Direct kit (D5020). The promoter regions of Oct4 (Panel A) and Nanog (Panel B) were then amplified using primers specific for bisulfite converted DNA (to select for fully converted DNA). Purified PCR products were directly analyzed by 454 sequencing. Using this method we were able to get sequence for >1000 PCR products of each promoter. The conversion rate was calculated as (# non CpG 'C' nucleotides converted to 'T'/ total # non CpG 'C' nucleotides) and was > 99% for each locus in each sample. Methylation at each position was then calculated as the "% Methylated Reads" (# reads with 'C' at CpG position/total # reads).

Figure 3-3. Cystic teratoma histology.
1 million ES or iPS cells were injected into the right hind flank of NOD/SCID/γ chain -/- mice. Teratomas were harvested 4-6 weeks later. All 3 germ layers were identified in each of the teratomas. Images are H&E stained slides with magnification noted.

Figure 3-4. OSK lentivirus insertion sites prove iPS clonality
A. Insertion sites of OSK lentivirus for each clone indicated on a graphic representation of the 20 mouse chromosomes. Each iPS line had 1-5 insertion sites. Experiment 1 in green, experiment 2 in red and experiment 3 in blue. B. Mouse Exon 1.0ST expression data (same samples from **Figure 1**) for the nearest neighbor genes of each of the insertion sites in the experiment 3 clones.

Figure 3-5. Homozygous SNVs in fibroblast lines compared to B6 reference genome.
A. Total SNVs B. SNVs within coding regions

Figure 3-6. SNVs in iPS clones compared to parental fibroblast lines.
A. Total SNVs B. SNVs within coding regions.

Figure 3-7. Circos plots illustrating genomic distribution of SNVs in each iPS clone
The genome of each iPS clone is illustrated as a circular diagram of the 20 chromosomes. Each red tick mark represents an individual SNV.

Figure 3-8. Variant allele frequency plots of iPS clones.
For each iPS clone, the variant allele frequency and read count depth of every SNV is plotted, the majority of all SNVs are present in half of the reads. A cutoff of 30% variant

allele frequency was used to eliminate reads from any potential contaminating feeder MEFs (which would have been derived from a different mouse).

Figure 3-9. Common and private mutations in the 4 iPS clones from experiment 3.
A. Coding region SNVs B. Non-coding region SNVs

Figure 3-10. Analysis for common SNVs in an additional MPSVII iPS clones from an independent reprogramming event.
MPSVII-derived TTFs were transduced with the OSK lentivirus and gDNA was prepared from 10 iPS clones. Apaf1 (Panel A) and Sbno2 (Panel B) regions were amplified from each clone along with iPS clone #1 from experiment 3 and iPS clone #1 from experiment 2 as positive and negative controls, respectively. PCR products underwent Sanger sequencing to determine presence of the variant allele. Each plot shows the variant bp +/- 3 bp on either side. All 10 clones were WT for both loci.

Figure 3-11. Common structural variant in the 4 iPS clones from experiment 3.
A single structural variant was identified in all 4 iPS clones from experiment 3 spanning ~130,000 kb on chromosome 6.

Figure 3-12: Detection of common variants in rare proportion of parental MEF population.
The Apaf1 and Sbno2 loci were amplified from the MPSVII MEF population from experiment 3 followed by cloning into the pCR2.1 vector. Clones were counted and pooled and gDNA was prepared from these clone pools. Each pool was then digested with StuI and EcoRI to detect the novel StuI sites generated by the Apaf1 G16A and Sbno2 A3783G variants. Digestion products were run on 5% polyacrylamide gels, transferred to nitrocellulose and analyzed by Southern blot for digested bands. Expected band lengths are indicated below the blot images. * indicates the variant-specific digestion products.
Figure 3-14. Variant allele frequency of private SNVs compared to common SNVs. p-values are listed at the top of the figure.

Figure 13. Comparison of common vs private variant allele frequencies
Variant allele frequencies for the common and private SNVs from the clones of experiment 3 were plotted and compared. The common SNVs are present in ~50% of reads, suggesting heterozygosity. The second, smaller peak in clones 1, 2 and 6 represents a subclone that arose after reprogramming.

Figure 14. Model of selection in iPS reprogramming.
A. In most caes, each cell has a unique set of pre-existing background mutations, which are “captured” by the expansion and cloning of single cells with reprogramming. B. In some cases, pre-existing mutations render a cell “super fit” for reprogramming. Private mutations can be acquired throughout the culture and reprogramming process: 1) between the time when the fitness mutations are acquired and the reprogramming event occurs, 2) at the time of reprogramming 3) after reprogramming. Mutations arising post-reprogramming can provide additional selective pressure for the outgrowth of subclones.

Table 3-1. Generation and characterization of iPS clones

Experiment (Parental Cells)	Clone	Age of mouse	Days in Culture Pre-transduction	Day post-transduction colony was picked	Days in culture from clone picking to gDNA prep (# passages)	ES-like morphology	Alkaline phosphatase staining positive	Flow cytometric staining		
								SSEA1 +	Nanog +	Oct 3/4+
1 (WT B6 MEF)	Clone 4	e13.5	6	17	29 (9)	Yes	Yes	Yes	Yes	Yes
	Clone 5	e13.5	6	17	29 (9)	Yes	Yes	Yes	Yes	Yes
	Clone 9	e13.5	6	17	29 (9)	Yes	Yes	Yes	Yes	Yes
2 (WT B6 TTF)	Clone 1	65d	20	26	16 (4)	Yes	Yes	Yes	Yes	Yes
	Clone 2	65d	20	26	16 (4)	Yes	Yes	Yes	Yes	Yes
	Clone 3	65d	20	26	16 (4)	Yes	Yes	Yes	Yes	Yes
3 (GusB-/- B6 MEF)	Clone 1	e13.5	6	17	24 (8)	Yes	Yes	Yes	Yes	Yes
	Clone 2	e13.5	6	17	24 (8)	Yes	Yes	Yes	Yes	Yes
	Clone 5	e13.5	6	17	24 (8)	Yes	Yes	Yes	Yes	Yes
	Clone 6	e13.5	6	17	24 (8)	Yes	Yes	Yes	Yes	Yes

Table 3-2. Whole genome sequencing coverage and lentivirus insertion sites of all 10 iPS clones

Experiment:Clone	WGS Haploid Coverage	OSK Coverage	# of Unique OSK integrations	Integration Positions
1:WT MEF	55.178	n/a	n/a	n/a
1:4	45.689	105.175	3	6:116062660; 11:54261983; 17:29179146
1:5	55.191	176.645	4	7:20201630; 9:34109963; 15:8307932; 16:36841079
1:9	57.668	104.485	2	12:56888952; X:82219170
1:WT TTF	22.727	n/a	n/a	n/a
2:1	26.56	148.404	1	X:23294264
2:2	24.377	66.531	2	7:102410141; 9:41255493
2:3	30.255	75.629	5	3:65184739; 7:66559629; 10:30896454; 12:70688422; 14:97803433
3:GusB-/- MEF	17.877	n/a	n/a	n/a
3:1	28.896	20.387	2	5:16530376; 9:72205724
3:2	23.958	12.279	2	4:139018701; 14:84892429
3:5	19.591	12.603	1	14:6229265
3:6	22.005	13.74	2	12:99479785; 16:38231164

Table 3-3. Validated coding region SNVs

Experiment: Clone	Position	Mutation	Translational Effect	Gene	Mutated in COSMIC?	Found in OMIM?	Variant Validation		
							Reference Reads	Variant Reads	Variant Frequency
1:4	1:52720233	G1697C	P566R	Mfsd6	YES	NO	96	76	0.44186047
1:4	13:11886861	A2419G	F807L	Ryr2	YES	YES	87	67	0.43506494
1:4	14:50760000	G313A	H105Y	Olfr728	NO	NO	90	95	0.51351351
1:4	7:104810281	A1707C	K569N	Rsf1	YES	NO	129	129	0.5
1:5	11:65885466	C5524T	V1842M	Dnahc9	YES	NO	207	105	0.33653846
1:5	11:71580404	C532T	R178W	Wscd1	NO	NO	82	55	0.40145985
1:5	19:6531496	G4460A	G1487D	Nrxn2	YES	NO	55	66	0.54545455
1:5	2:86195991	G547A	P183S	Olfr1056	NO	NO	171	129	0.43
1:5	7:27457861	T1177C	S393G	Cyp2b23	NO	NO	333	252	0.43076923
1:5	9:53303955	A3585(+2)G	e23+2	Atm	YES	YES	69	67	0.49264706
1:9	18:66158326	C287A	S96I	Lman1	NO	YES	287	154	0.34920635
1:9	19:10330864	C1135T	E379K	Dagla	YES	NO	255	144	0.36090226
1:9	7:4960568	T552C	G184	Fiz1	NO	NO	292	139	0.3225058
2:1	14:51135247	G367A	A123T	Olfr742	NO	NO	387	407	0.51259446
2:1	17:46662058	A3610T	S1204T	Cul9	YES	NO	34	23	0.40350877
2:1	18:4349082	T233G	Y78S	Map3k8	YES	YES	62	68	0.52307692
2:1	2:119346067	G1448C	A483G	Exd1	YES	NO	151	132	0.4664311
2:1	2:66522305	C3181T	V1061M	Scn7a	YES	NO	40	35	0.46666667
2:1	2:86829071	T261G	K87N	Olfr1101	NO	NO	68	74	0.52112676
2:1	4:115164785	G682A	R228C	Cyp4a14	NO	NO	45	31	0.40789474
2:1	6:119987197	T454C	T152A	Wnk1	YES	YES	93	1314	0.93323864
2:1	6:119986949	T702C	T234	Wnk1	NO	YES	183	1421	0.88591022
2:1	6:138314487	T414A	E138D	Lmo3	NO	NO	13	17	0.56666667
2:1	7:132816231	T2029C	T677A	Gtf3c1	YES	NO	275	270	0.49541284
2:1	4:43646587	C858T	L286	Npr2	NO	YES	56	45	0.44554455
2:1	9:56746380	T6667C	L2223	Cspg4	NO	NO	67	67	0.5
2:1	4:15828065	A507(-2)G	e8-2	Calb1	YES	NO	86	92	0.51685393
2:2	1:16450468	C307A	A103S	Stau2	NO	NO	48	40	0.45454545
2:2	11:73525754	C2021T	T674I	Zfp735	NO	NO	ND	ND	ND
2:2	13:66953683	A324C	C108W	Zfp640	NO	NO	5	3	0.375
2:2	3:113266509	G598A	H200Y	Amy1	NO	NO	91	80	0.46783626
2:2	5:137589316	C1215A	K405N	Trim56	YES	NO	137	127	0.48106061
2:2	6:30504261	C879G	S293R	Cpa2	YES	NO	45	42	0.48275862
2:2	7:112276751	T169C	I57V	ENSMUSG00000059768	NO	NO	30	18	0.375
2:2	7:12342475	C870G	I290M	V1rg2	NO	NO	138	113	0.4501992
2:2	7:148129017	C389A	T130K	Athl1	YES	NO	128	148	0.53623188
2:2	10:85342131	G735A	G245	Pwp1	NO	NO	26	21	0.44680851
2:2	15:75696200	C696A	L232	Gsdmd	NO	NO	94	74	0.44047619
2:2	2:163384795	C597A	V199	Hnf4a	NO	YES	63	27	0.3
2:3	1:34876755	T251A	Y84F	Fam168b	NO	NO	36	27	0.42857143
2:3	10:62780068	G3006T	M1002I	Herc4	YES	NO	43	47	0.52222222
2:3	11:117891399	C190G	P64A	ENSMUSG00000061395	NO	NO	16	17	0.51515152
2:3	11:94420228	C1679G	S560T	Acsf2	NO	NO	47	32	0.40506329
2:3	13:81030887	C1172A	P391Q	Arrdc3	YES	NO	30	25	0.45454545
2:3	15:44377040	T7376A	V2459E	Pkhd1l1	YES	NO	12	19	0.61290323
2:3	16:18399813	G1651A	A551T	Arvcf	YES	NO	13	10	0.43478261
2:3	2:76647962	T41358A	Q13786H	Ttn	YES	YES	39	28	0.41791045
2:3	5:20759811	C1455A	D485E	Pion	YES	NO	52	42	0.44680851
2:3	5:26445213	G375T	D125E	ENSMUSG00000067698	NO	NO	1034	252	0.1958042
2:3	7:51884733	G2957C	A986G	Myh14	NO	YES	13	8	0.38095238
2:3	7:99565750	G1895A	R632H	Dlg2	YES	NO	65	71	0.52205882
2:3	7:10574295	C400A	E134*	Vmn2r49	NO	NO	318	285	0.47263682
2:3	10:88446510	T1728A	L576	Ano4	NO	NO	37	42	0.53164557
2:3	16:88774114	G282T	I94	2310057N15Rik	NO	NO	25	27	0.51923077
2:3	17:72056421	C1623A	V541	Fam179a	NO	NO	47	58	0.55238095
2:3	2:85448358	T708C	K236	Olfr1000	NO	NO	76	66	0.46478873
2:3	4:151744896	C2640T	P880	Chd5	NO	NO	23	30	0.56603774
2:3	5:63121879	C1362T	T454	Arap2	NO	NO	35	35	0.5
2:3	X:121244726	C1516G	A506P	Vmn2r121	NO	NO	127	76	0.37438424
2:3	X:88880865	G1083T	K361N	LOC629542	NO	NO	47	37	0.44047619
2:3	X:137606686	G4303C	L1435V	Col4a6	YES	YES	16	11	0.40740741

Experiment: Clone	Position	Mutation	Translational Effect	Gene	Mutated in COSMIC?	Found in OMIM?	Variant Validation		
							Reference Reads	Variant Reads	Variant Frequency
3:1	1:99079629	G625A	A209T	EG634331	NO	NO	1093	385	0.26048714
3:1	6:119218294	C944T	A315V	Cacna2d4	YES	YES	305	904	0.74772539
3:1	10:90542754	G16A	R6C	Apaf1	YES	NO	947	533	0.36013514
3:1	14:27488308	C731T	R244Q	Pde12	NO	NO	2822	1818	0.39181034
3:1	10:79520428	A3783G	A1261	Sbno2	NO	NO	1438	1640	0.53281352
3:1	2:69374051	C760T	D254N	Lrp2	YES	YES	141	475	0.7711039
3:1	5:72630517	G666T	Q222H	Atp10d	YES	NO	594	2171	0.78517179
3:1	7:97388761	T598A	T200S	Ccdc83	NO	NO	1370	1048	0.43341605
3:1	13:64369144	C5T	T2M	1810034E14Rik	NO	NO	1087	1100	0.50297211
3:1	13:89747651	C968A	P323Q	Hapln1	YES	NO	226	441	0.66116942
3:1	2:76716141	C23955A	G7985	Ttn	YES	YES	774	732	0.48605578
3:2	1:99079629	G625A	A209T	EG634331	NO	NO	305	278	0.47684391
3:2	6:119218294	C944T	A315V	Cacna2d4	YES	YES	616	358	0.36755647
3:2	10:90542754	G16A	R6C	Apaf1	YES	NO	800	308	0.27797834
3:2	14:27488308	C731T	R244Q	Pde12	NO	NO	2030	1800	0.46997389
3:2	10:79520428	A3783G	A1261	Sbno2	NO	NO	1990	2106	0.51416016
3:2	2:164887834	G100A	D34N	Cd40	YES	YES	2006	2064	0.50712531
3:2	5:122919663	A791G	I264T	Atp2a2	YES	YES	1475	1531	0.5093147
3:2	7:47332084	T789G	E263D	Mrgpra5	NO	NO	432	399	0.4801444
3:2	8:106854405	C496A	L166I	Cmtm2b	NO	NO	987	1116	0.53067047
3:2	17:37436493	C873T	L291	Olfr101	NO	NO	423	212	0.33385827
3:5	1:99079629	G625A	A209T	EG634331	NO	NO	938	1004	0.51699279
3:5	6:119218294	C944T	A315V	Cacna2d4	YES	YES	603	487	0.44678899
3:5	10:90542754	G16A	R6C	Apaf1	YES	NO	485	291	0.375
3:5	14:27488308	C731T	R244Q	Pde12	NO	NO	2046	2266	0.5255102
3:5	10:79520428	A3783G	A1261	Sbno2	NO	NO	1894	2310	0.54947669
3:5	6:64013952	A883C	S295R	Grid2	YES	NO	663	975	0.5952381
3:5	17:32494082	G2342C	A781G	Wiz	NO	NO	1643	1387	0.45775578
3:5	17:34174154	G1352A	R451Q	Rxrb	YES	NO	1341	910	0.40426477
3:5	5:97636472	A215C	*72S	LOC100040598	NO	NO	805	805	0.5
3:5	17:33140786	C1293T	R431	Zfp952	NO	NO	1246	814	0.39514563
3:5	X:65963791	A1305G	Q435	Fmr1	NO	YES	324	278	0.46179402
3:6	1:99079629	G625A	A209T	EG634331	NO	NO	1324	249	0.15829625
3:6	6:119218294	C944T	A315V	Cacna2d4	YES	YES	751	367	0.32826476
3:6	10:90542754	G16A	R6C	Apaf1	YES	NO	802	451	0.35993615
3:6	14:27488308	C731T	R244Q	Pde12	NO	NO	2054	1825	0.47048208
3:6	10:79520428	A3783G	A1261	Sbno2	NO	NO	2017	1940	0.49027041
3:6	1:108948259	G89A	P30L	Serpinb3a	NO	NO	1084	1408	0.56500803
3:6	15:101612057	T755G	K252T	Krt72	YES	NO	707	1261	0.64075203
3:6	8:97894729	C1425T	F475	Mmp15	NO	NO	1101	271	0.19752187
3:6	13:73819506	G1638A	F546	Slc6a19	NO	YES	2032	491	0.19460959
3:6	17:25608777	G111A	T37	Tekt4	NO	NO	2798	2314	0.45266041

Table 3-4. Common indels in all 4 Experiment 3 iPS clones

Start Position	Indel Type	Reference	Variant	Gene	Amino Acid Change
1:151199053	DEL	A	0	None Annotated	n/a
1:186196114	DEL	T	0	None Annotated	n/a
3:48132682	INS	0	AT	None Annotated	n/a
10:108602704	DEL	TTT	0	None Annotated	n/a
10:113653029	INS	0	A	None Annotated	n/a
14:62764104	DEL	TGCTTCTGTAGACG	0	None Annotated	n/a
14:38669259	INS	0	C	None Annotated	n/a
14:83683846	INS	0	A	None Annotated	n/a
15:50209877	DEL	TCT	0	None Annotated	n/a
18:3689944	INS	0	T	None Annotated	n/a
18:48797211	INS	0	T	None Annotated	n/a
X:38193608	DEL	G	0	None Annotated	n/a
X:65453524	DEL	A	0	None Annotated	n/a
X:94905486	DEL	A	0	None Annotated	n/a
3:101347390	DEL	A	0	Atp1a1	RNA gene
10:35418517	DEL	A	0	ENSMUSG00000063953	RNA gene
18:51744954	DEL	G	0	ENSMUSG00000077317	RNA gene
12:34726624	INS	0	T	Hdac9	RNA gene
7:132434793	INS	0	A	4930533L02Rik	RNA gene
17:73797519	DEL	C	0	Capn13	RNA gene
4:43506005	INS	0	TC	Car9	RNA gene
3:89278370	DEL	C	0	Kcnn3	RNA gene
14:58753183	DEL	CTC	0	LOC100043554	RNA gene
16:18686784	INS	0	C	LOC622795	RNA gene
2:148805805	DEL	A	0	OTTMUSG00000015744	RNA gene
16:34605270	INS	0	AC	Ropn1	RNA gene
4:133201025	DEL	G	0	Zdhhc18	RNA gene
7:36203558	INS	0	AC	Ccdc123	Intronic: e9+178
X:72352644	DEL	GCGAAT	0	Dkc1	Intronic: e13+55
4:102178779	DEL	C	0	Pde4b	Intronic: e5+132
3:13814071	INS	0	A	Ralyl	Intronic: e1+36968
X:98328440	DEL	A	0	Snx12	Intronic: e5-30666

Table 3-5. Pools tested for presence of Apaf1 and Sbno2 variants

	# Colonies in Pool	StuI site detected by Southern blot?
Apaf1	648	NO
	698	NO
	662	YES
	230	NO
	205	NO
	149	NO
	144	NO
	448	NO
	192	NO
	164	NO
	117	NO
	98	NO
	183	NO
	100	NO
	431	NO
	143	NO
	79	NO
	100	NO
	129	NO
	91	NO
	157	NO
	391	NO
	39	NO
	49	NO
	60	NO
	138	NO
	192	NO
	177	NO
Sbno2	1548	YES
	1014	YES
	1686	YES
	976	YES
	141	YES
	207	NO
	203	NO
	169	NO

Table 3-6. Pathways identified by MUSIC suite as being enriched for in genes with SNVs

Pathway	Experiment 1: WT MEF			Experiment 2: WT TTF			Experiment 3: GusB-/- MEF				p-value	FDR
	4	5	9	1	2	3	1	2	5	6		
Cardiac muscle contraction	Ryr2						Cacna2d4	Cacna2d4	Cacna2d4	Cacna2d4	>0.001	>0.001
p53 signaling pathway		Atm					Apaf1	Apaf1	Apaf1	Apaf1	>0.001	>0.001
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	Ryr2						Cacna2d4	Cacna2d4, Atp2a2	Cacna2d4	Cacna2d4	>0.001	0.027
Apoptosis		Atm					Apaf1	Apaf1	Apaf1	Apaf1	>0.001	0.001
Hypertrophic cardiomyopathy (HCM)	Ryr2					Ttn	Cacna2d4	Cacna2d4, Atp2a2	Cacna2d4	Cacna2d4	0.001	0.028
Dilated cardiomyopathy	Ryr2					Ttn	Cacna2d4	Cacna2d4, Atp2a2	Cacna2d4	Cacna2d4	0.001	0.086
Amyotrophic lateral sclerosis (ALS)							Apaf1	Apaf1	Apaf1	Apaf1	0.001	0.005
Parkinson's disease							Apaf1	Apaf1	Apaf1	Apaf1	0.001	0.007
Small cell lung cancer							Apaf1	Apaf1	Apaf1, Rxrb	Apaf1	0.002	0.033
Alzheimer's disease							Apaf1	Apaf1, Atp2a2	Apaf1	Apaf1	0.003	0.017
Huntington's disease							Apaf1	Apaf1	Apaf1	Apaf1	0.020	0.124
MAPK signaling pathway				Map3k8			Cacna2d4	Cacna2d4	Cacna2d4	Cacna2d4	0.043	0.080
Pancreatic secretion	Ryr2				Cpa2			Atp2a2			0.065	0.116
PPAR signaling pathway				Cyp4a14					Rxrb		0.088	0.249
Toll-like receptor signaling pathway				Map3k8				Cd40			0.141	1.000
Viral myocarditis						Myh14		Cd40			0.196	0.508
Cell adhesion molecules (CAMs)		Nrxn2						Cd40			0.390	0.742
Calcium signaling pathway	Ryr2							Atp2a2			0.513	1.000
Olfactory transduction	Olfr1056			Olfr74, Olfr1101							0.058	0.304
Asthma								Cd40			0.08015	1
Intestinal immune network for IgA production								Cd40			0.13049	0.43836
Thyroid cancer									Rxrb		0.14428	1
Allograft rejection								Cd40			0.14926	1
Primary immunodeficiency								Cd40			0.1543	1
Autoimmune thyroid disease								Cd40			0.18528	1
Fatty acid metabolism				Cyp4a14							0.18542	0.19352
Malaria								Cd40			0.20754	1
Starch and sucrose metabolism					Amy1						0.21678	0.26482
Hedgehog signaling pathway							Lrp2				0.22494	0.54339
Carbohydrate digestion and absorption					Amy1						0.22762	1
Systemic lupus erythematosus								Cd40			0.23269	1
Non-small cell lung cancer									Rxrb		0.23429	1
Retinol metabolism				Cyp4a14							0.23612	0.36261
Arachidonic acid metabolism				Cyp4a14							0.24884	0.33438
Adipocytokine signaling pathway									Rxrb		0.27037	1
Long-term depression									Grid2		0.32792	1
T cell receptor signaling pathway				Map3k8							0.34932	1
Protein digestion and absorption					Cpa2						0.36819	1
Cell cycle									Atm		0.42245	0.84099
Toxoplasmosis								Cd40			0.42665	1
Protein processing in endoplasmic reticulum			Lman1								0.44594	0.94928
Vascular smooth muscle contraction				Cyp4a14							0.46016	1
Cytokine-cytokine receptor interaction								Cd40			0.46994	0.90212
Ubiquitin mediated proteolysis						Herc4					0.48474	0.99256
Amoebiasis										Serpinb3a	0.49434	1
Tight junction						Myh14					0.49889	1
Neuroactive ligand-receptor interaction									Grid2		0.59769	1
Regulation of actin cytoskeleton						Myh14					0.64666	1
Pathways in cancer									Rxrb		0.7795	1

Figure 3-1

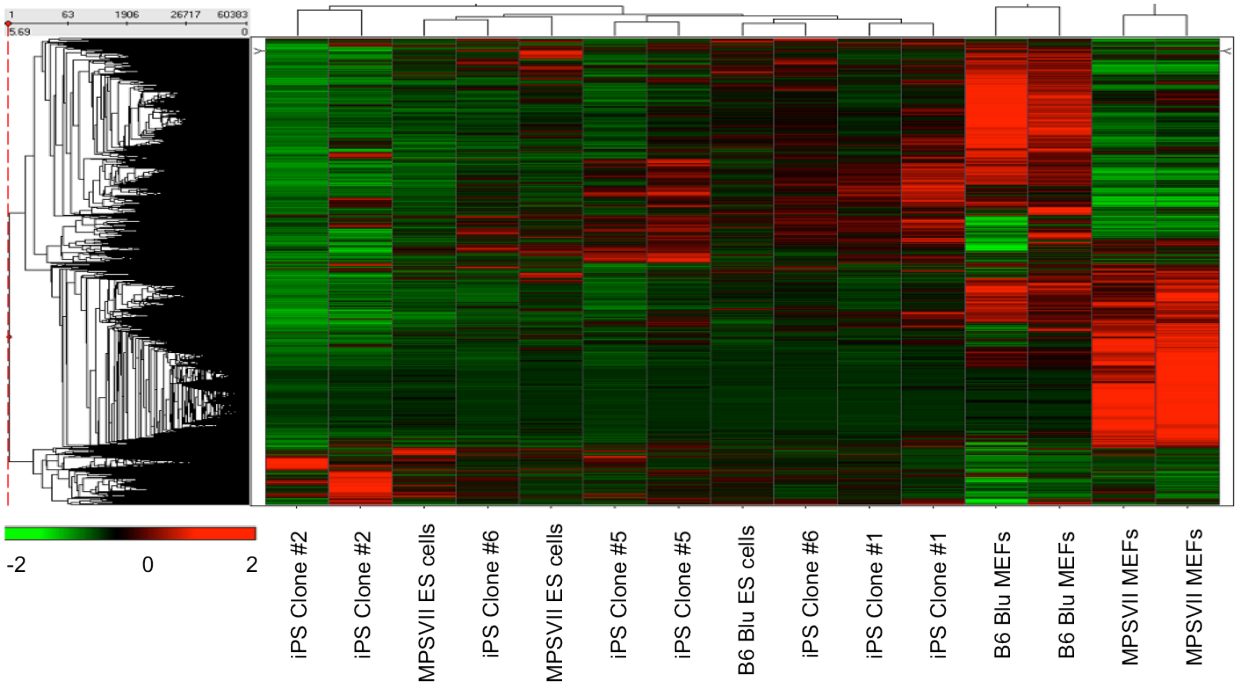
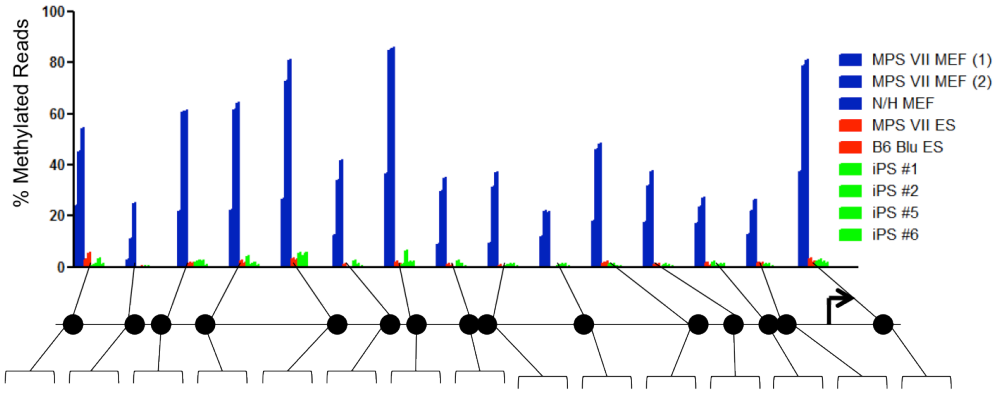

1
63
1906
26717
60383
5.69
0
-2
0
2
iPS Clone #2
iPS Clone #2
MPSVII ES cells
iPS Clone #6
MPSVII ES cells
iPS Clone #5
iPS Clone #5
B6 Blu ES cells
iPS Clone #6
iPS Clone #1
iPS Clone #1
B6 Blu MEFs
B6 Blu MEFs
MPSVII MEFs
MPSVII MEFs

Figure 3-2

A.

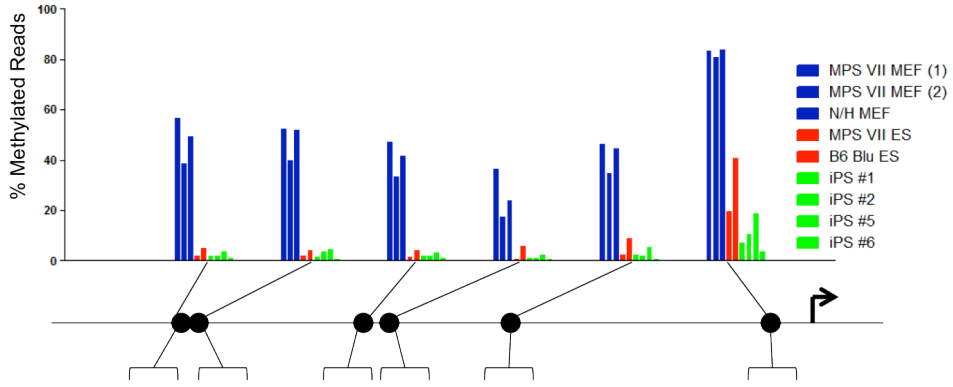
Oct 4 Promoter



24.02	2.92	21.76	22.37	26.69	12.62	36.56	8.79	9.27	11.95	18.06	17.41	16.96	12.76	37.28
45.22	11.11	60.80	61.51	72.61	33.83	84.82	29.50	31.33	21.93	45.97	31.71	23.50	22.11	78.86
54.21	24.78	60.95	64.16	80.91	41.75	85.45	34.64	36.94	21.57	48.29	37.53	27.17	26.41	81.03
3.10	0.11	1.53	2.59	3.24	1.36	1.92	1.15	0.68	0.00	1.65	1.28	1.83	1.83	3.28
5.42	0.28	1.63	1.46	3.08	0.13	1.18	0.78	0.33	0.00	2.05	1.35	0.82	1.53	1.99
0.86	0.32	2.20	4.20	5.44	0.00	1.06	0.00	0.70	0.60	1.37	0.17	2.21	1.14	2.25
1.39	0.32	2.68	1.08	4.38	2.35	6.25	2.61	1.30	1.12	1.31	1.36	0.85	1.17	2.99
3.20	0.14	2.55	1.86	5.70	1.05	2.13	1.39	1.21	1.36	0.53	0.90	1.09	0.37	1.91
1.04	0.07	0.93	0.75	0.00	0.31	1.87	0.22	0.22	0.45	0.48	0.34	1.26	0.11	1.56

B.

Nanog Promoter



56.75	52.35	47.08	36.56	46.54	83.39
38.70	40.03	33.57	17.43	34.75	80.76
49.31	51.74	41.41	23.96	44.71	83.94
2.05	1.77	1.62	0.74	2.38	19.54
5.13	3.96	3.89	5.91	8.83	40.55
1.93	1.37	1.73	1.06	2.17	7.27
1.86	3.84	1.92	1.15	2.09	10.42
3.45	4.33	3.06	2.45	5.30	18.65
1.05	0.66	0.93	0.61	0.74	3.49

Figure 3-3

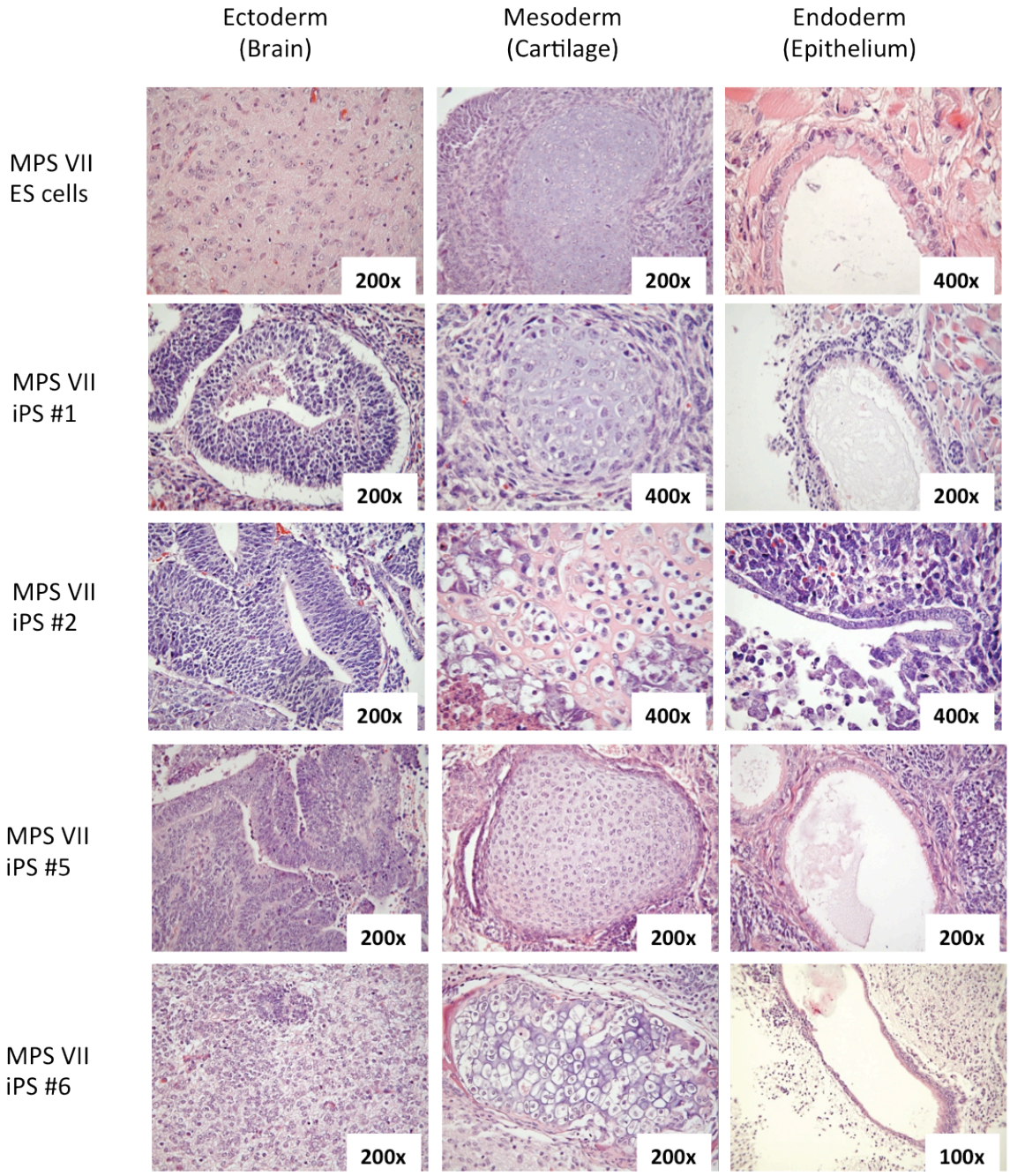
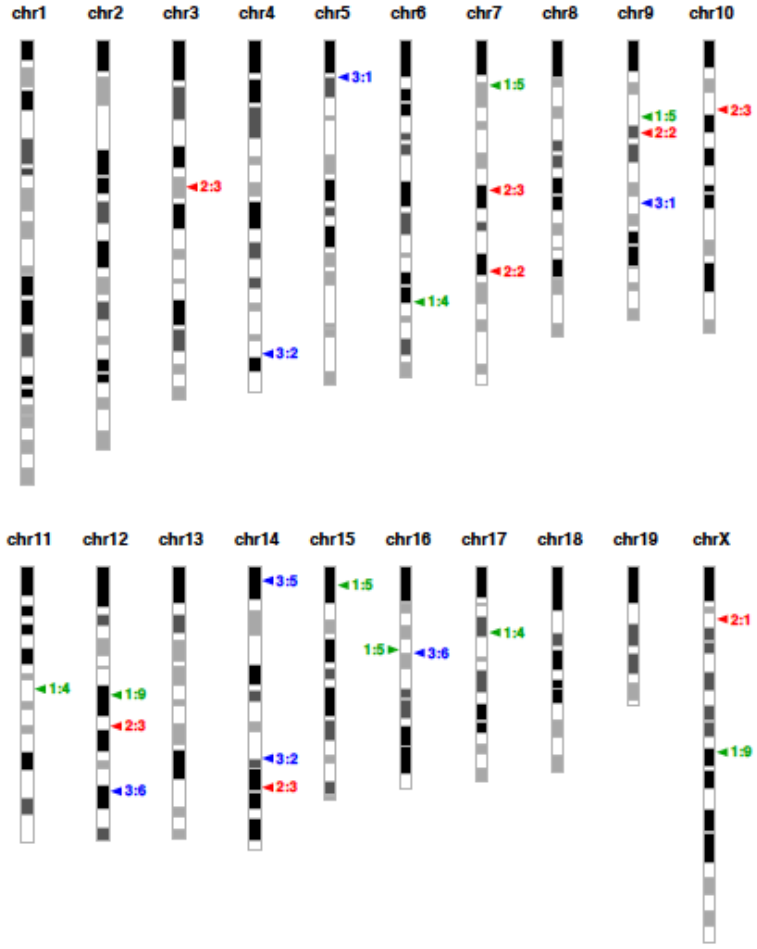

Ectoderm
(Brain)
Mesoderm
(Cartilage)
Endoderm
(Epithelium)
MPS VII
ES cells
200x
200x
400x
MPS VII
iPS #1
200x
400x
200x
MPS VII
iPS #2
200x
400x
400x
MPS VII
iPS #5
200x
200x
200x
MPS VII
iPS #6
200x
200x
100x

Figure 3-4

A.


chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chrX

B.

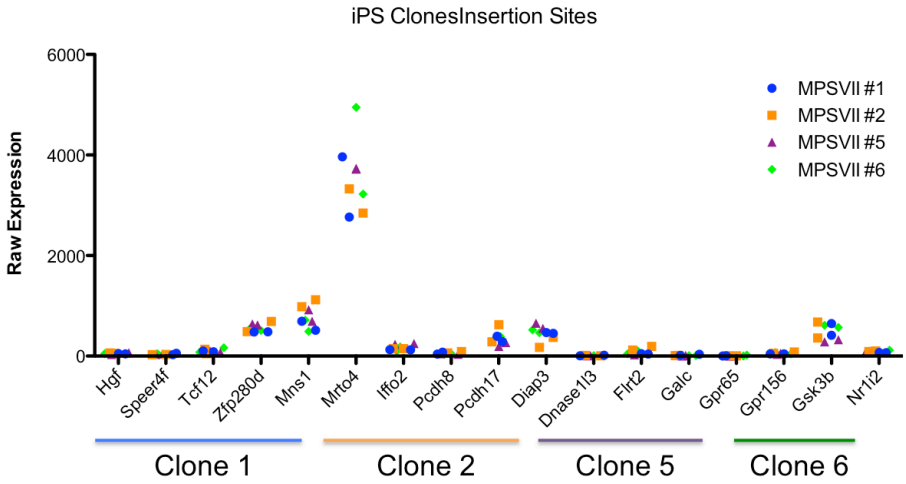
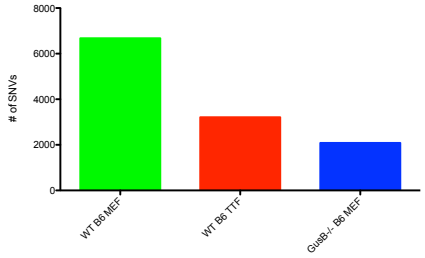

iPS ClonesInsertion Sites
Raw Expression
6000
4000
2000
0
MPSVII #1
MPSVII #2
MPSVII #5
MPSVII #6
Hgf
Speer4f
Tcf12
Zfp280d
Mns1
Mrto4
Iffo2
Pcdh8
Pcdh17
Diap3
Dnase1l3
Flrt2
Galc
Gpr65
Gpr156
Gsk3b
Nr1i2
Clone 1
Clone 2
Clone 5
Clone 6

Figure 3-5

A.


# of SNVs
8000
6000
4000
2000
0
WT B6 MEF
WT B6 TTF
GusBr/- B6 MEF

B.

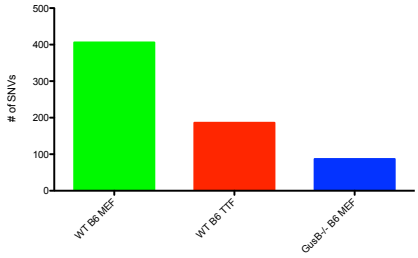
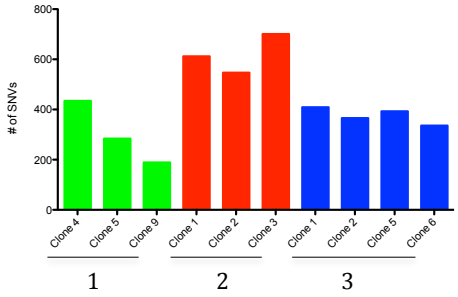

# of SNVs
500
400
300
200
100
0
WT B6 MEF
WT B6 TTF
GusBr/- B6 MEF

Figure 3-6

A.


800
600
400
200
0
# of SNVs
Clone 4
Clone 5
Clone 9
Clone 1
Clone 2
Clone 3
Clone 1
Clone 2
Clone 5
Clone 6
1
2
3

B.

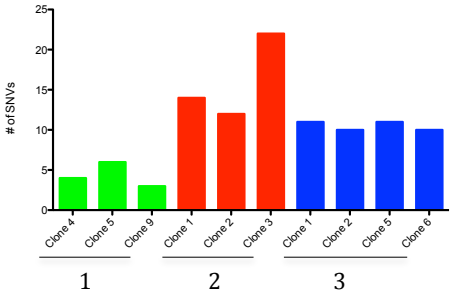

25
20
15
10
5
0
# of SNVs
Clone 4
Clone 5
Clone 9
Clone 1
Clone 2
Clone 3
Clone 1
Clone 2
Clone 5
Clone 6
1
2
3

Figure 3-7

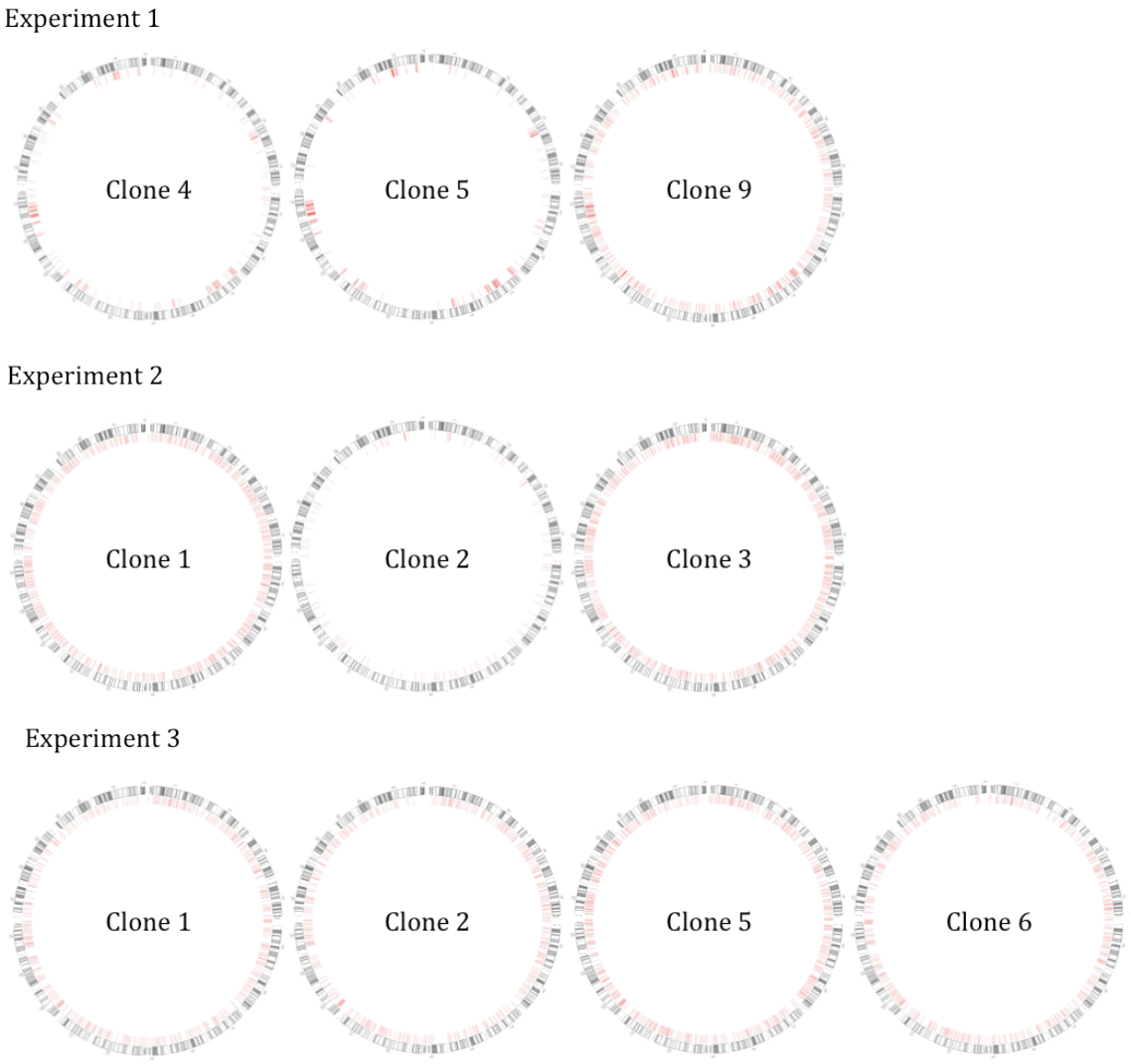

Experiment 1
Clone 4
Clone 5
Clone 9
Experiment 2
Clone 1
Clone 2
Clone 3
Experiment 3
Clone 1
Clone 2
Clone 5
Clone 6

Figure 3-8

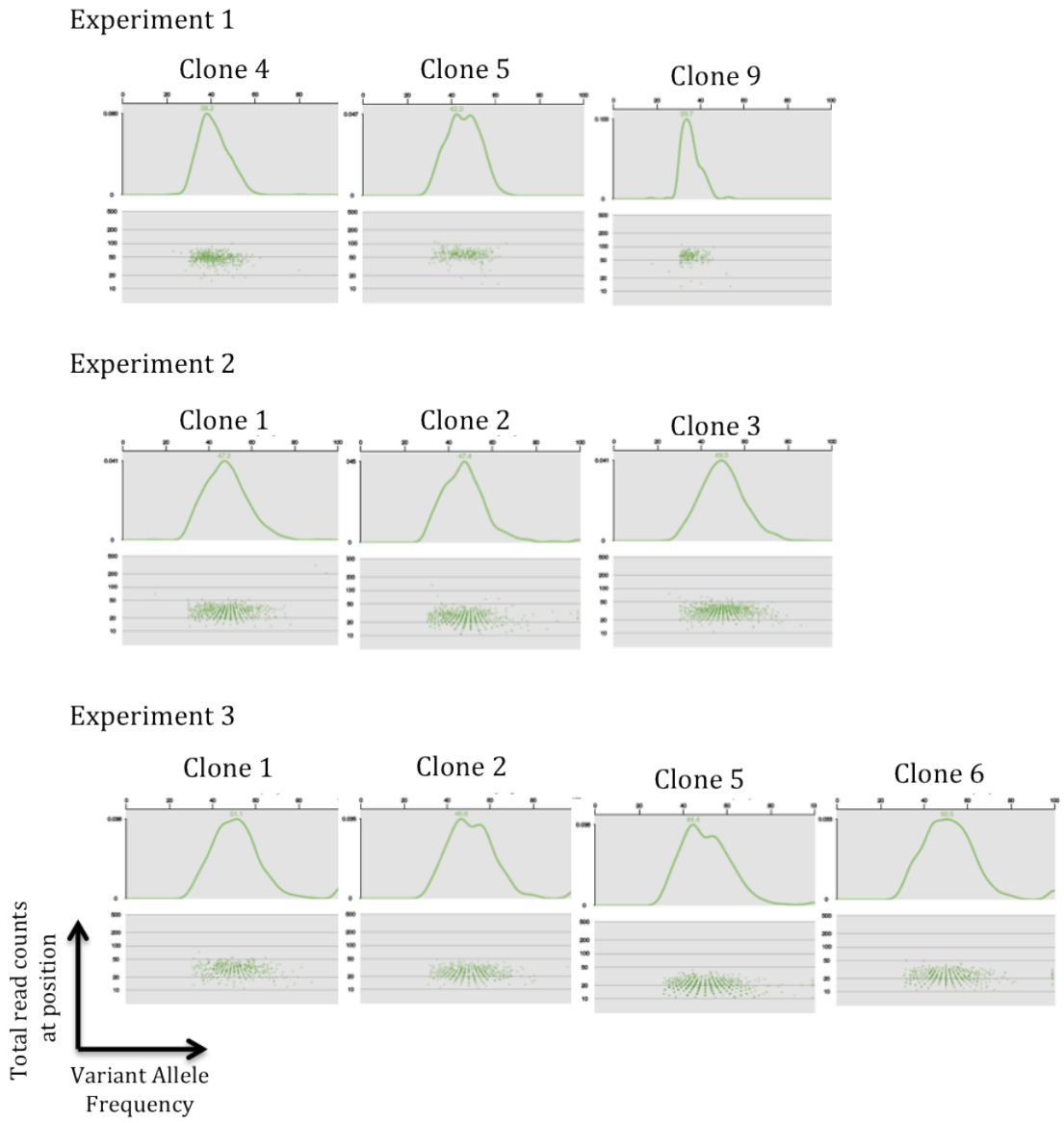
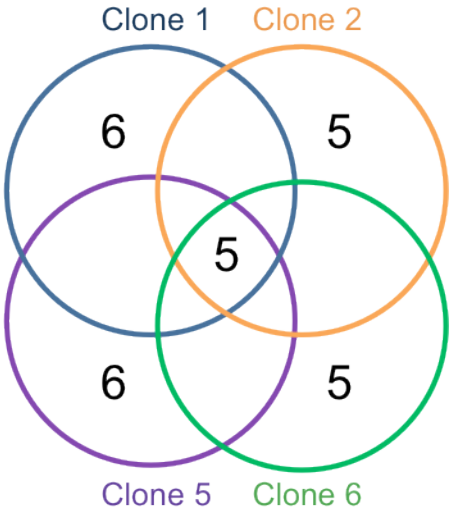

Experiment 1
Clone 4
Clone 5
Clone 9
Experiment 2
Clone 1
Clone 2
Clone 3
Experiment 3
Clone 1
Clone 2
Clone 5
Clone 6
Total read counts at position
Variant Allele Frequency

Figure 3-9

A.


Clone 1
Clone 2
6
5
5
6
5
Clone 5
Clone 6

B.


Clone 1
Clone 2
293
310
157
529
1
333
Clone 5
Clone 6

Figure 3-10

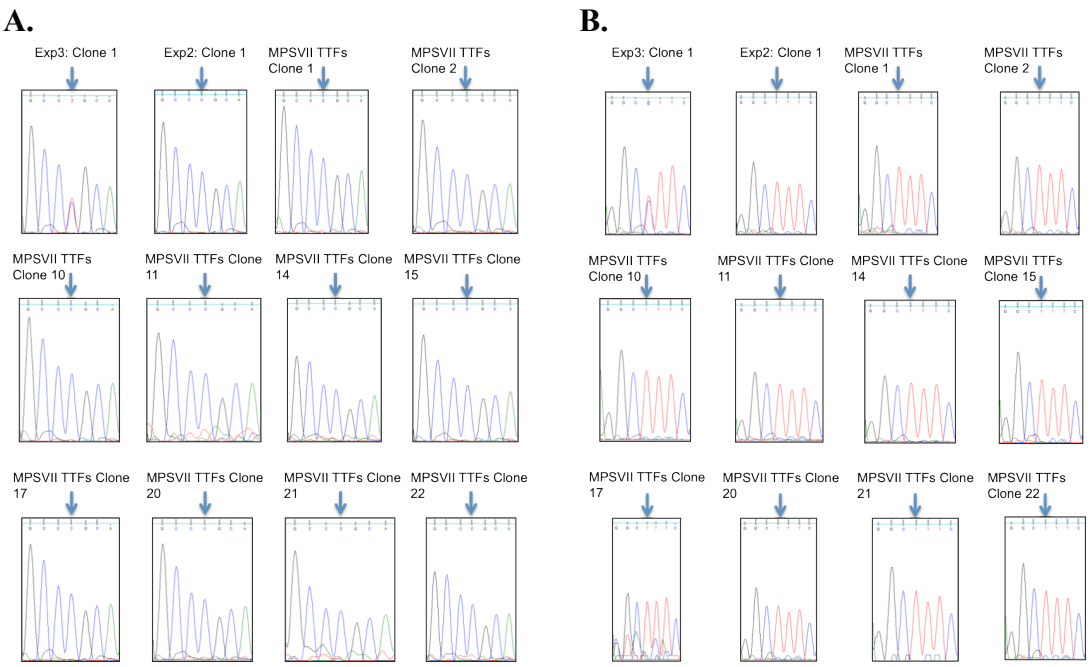

A.
Exp3: Clone 1
Exp2: Clone 1
MPSVII TTFs Clone 1
MPSVII TTFs Clone 2
MPSVII TTFs Clone 10
MPSVII TTFs Clone 11
MPSVII TTFs Clone 14
MPSVII TTFs Clone 15
MPSVII TTFs Clone 17
MPSVII TTFs Clone 20
MPSVII TTFs Clone 21
MPSVII TTFs Clone 22
B.
Exp3: Clone 1
Exp2: Clone 1
MPSVII TTFs Clone 1
MPSVII TTFs Clone 2
MPSVII TTFs Clone 10
MPSVII TTFs Clone 11
MPSVII TTFs Clone 14
MPSVII TTFs Clone 15
MPSVII TTFs Clone 17
MPSVII TTFs Clone 20
MPSVII TTFs Clone 21
MPSVII TTFs Clone 22

Figure 3-11

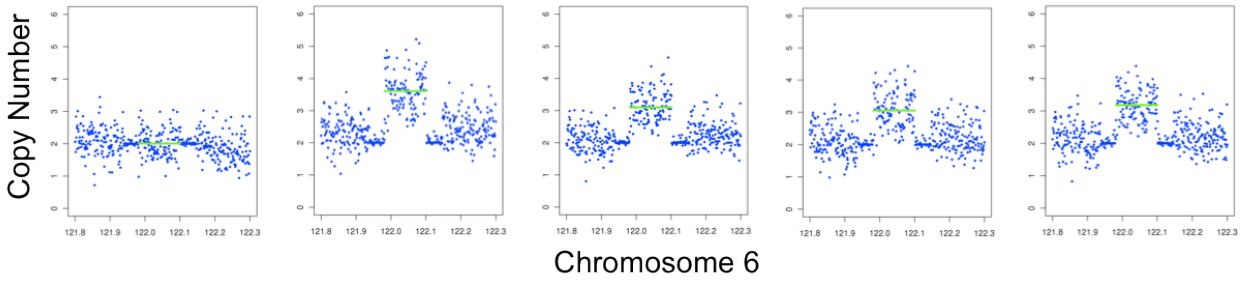


Figure 3-12

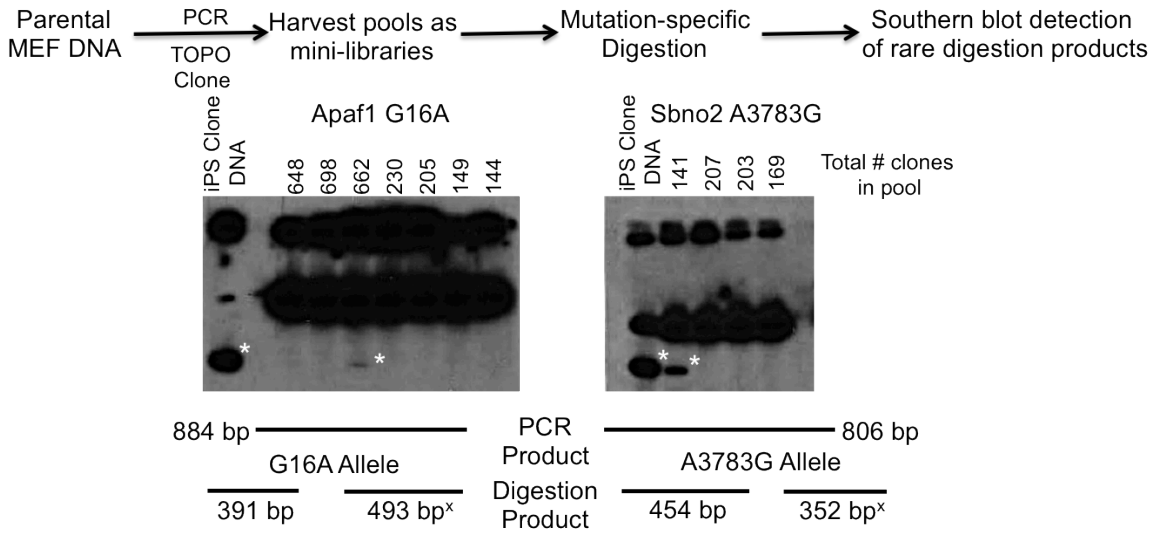

Parental MEF DNA
PCR
TOPO Clone
Harvest pools as mini-libraries
Mutation-specific Digestion
Southern blot detection of rare digestion products
Apaf1 G16A
iPS Clone DNA
648
698
662
230
205
149
144
Sbno2 A3783G
iPS Clone DNA
141
207
203
169
Total # clones in pool
884 bp
G16A Allele
391 bp
493 bpx
PCR Product
Digestion Product
806 bp
A3783G Allele
454 bp
352 bpx

Figure 3-13

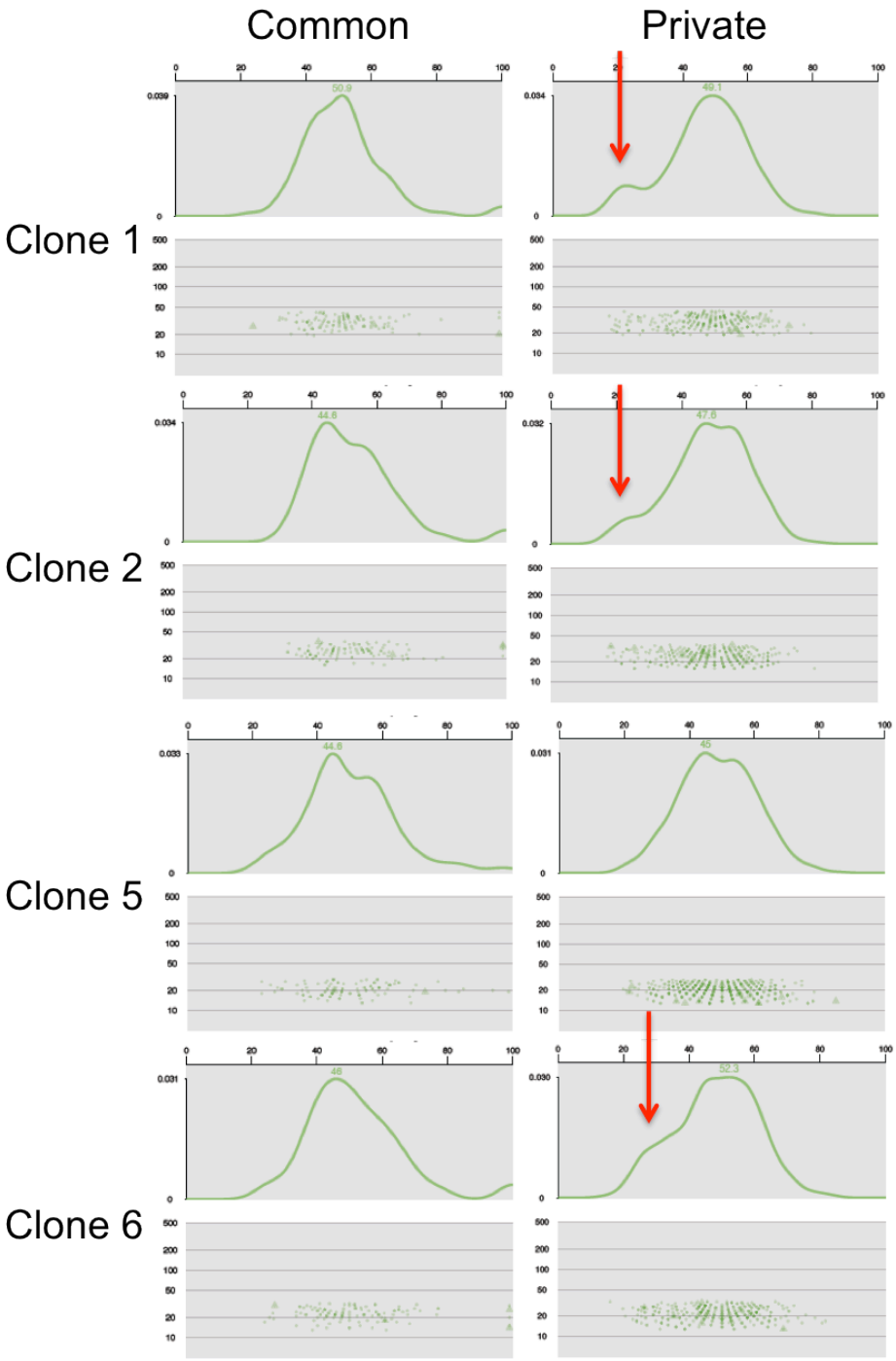
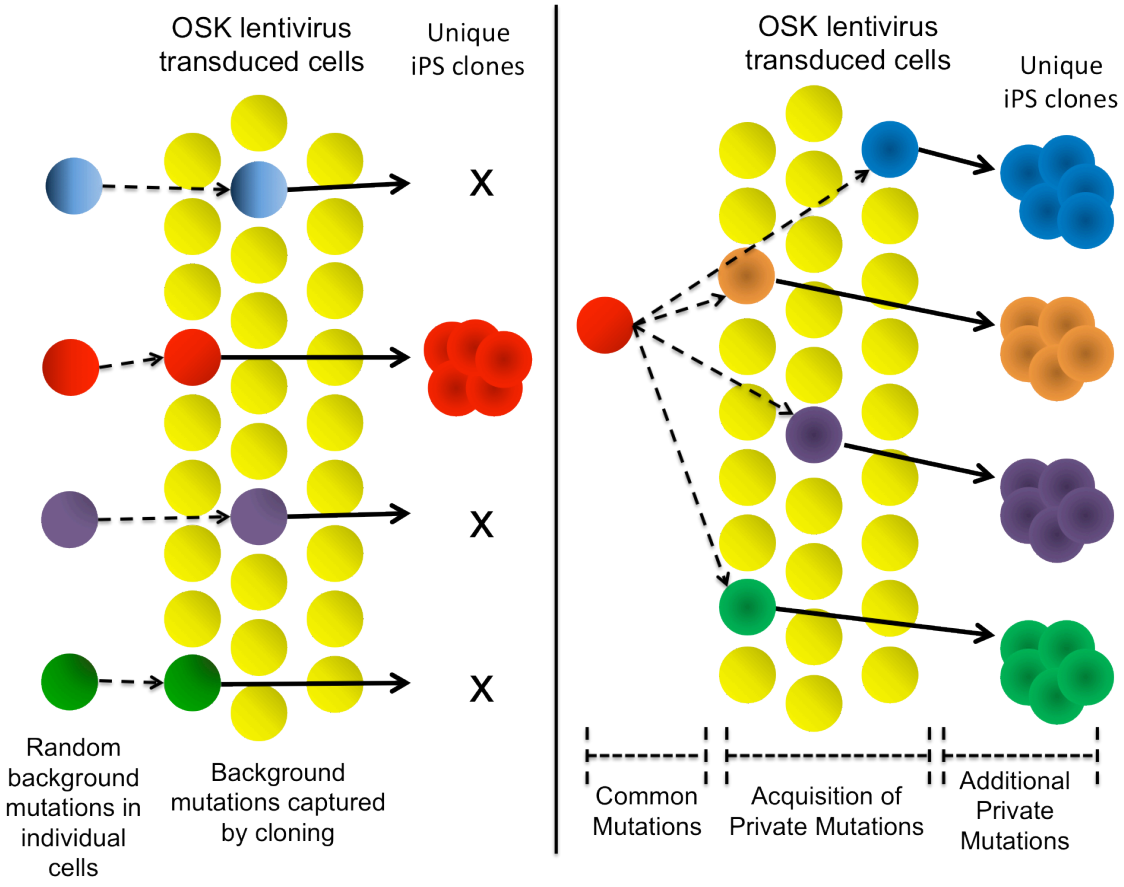


Figure 3-14


OSK lentivirus transduced cells
Unique iPS clones
X
X
X
Random background mutations in individual cells
Background mutations captured by cloning
OSK lentivirus transduced cells
Unique iPS clones
Common Mutations
Acquisition of Private Mutations
Additional Private Mutations

Chapter 4

Summary and Future Directions

Regulation of HOX expression in AML

In Chapter 2, we explored HOX expression in a diverse group of AML patients and found that patients fall into four categories based on HOX expression levels: 1) reprogramming translocations with HOX expression below the level of detection – t(8;21) and t(17;15); 2) reprogramming translocations that express either HOXA (MLL translocations) or HOXB (inv(16)) genes, but not both; 3) normal karyotype with high expression of *HOXA5, A9, A10, B2, B3*, and the HOX co-factor *MEIS1*; 4) normal karyotype with HOX expression below the level of detection. While multiple groups have reported HOX dysregulation in AML [1-12], we found that this "dysregulation" is actually a canonical pattern seen in healthy HSCs. The role of HOX genes in hematopoietic stem cell self-renewal has been shown in mouse models [13-15]. We hypothesize that capturing the normal HOX pattern gives leukemic cells self-renewal capabilities.

Coordinated expression of genes on separate chromosomes suggests the presence of an upstream regulator responsible for the HOX pattern seen in AML and healthy HSCs. This is supported by the HOX phenotype of MLL (a known regulator of HOX expression) mutated AML samples [16-19]. CDX2 is another regulator of HOX expression, which has recently been reported to be commonly dysregulated in AML [20]. However, in our set of 190 patients, there are 82 with no identified alterations in MLL or CDX2 that express HOX genes at high levels, suggesting that a novel upstream regulator(s) is affected in these patients. We performed *in silico* analyses to identify a common transcription factor binding motif or an upstream gene whose expression correlated with *HOXA9* gene expression and found only the subset of expressed HOX genes themselves.

This strongly suggests that an autoregulatory loop is responsible for the canonical pattern of HOX gene expression in AML.

To determine if HOX genes are capable of autoregulation, an *in vitro* reporter system could be developed using an YFP cassette under the regulation of the promoter regions of various HOX genes. If HOX genes are able to activate their own expression through promoter binding, then co-transfection of HOX cDNAs with the correct promoter-YFP constructs should induce YFP expression. The higher expression of *HOXA9* and *HOXB3* suggests that one or both of them could be the driver of the proposed autoregulatory loop. To test this, each cDNA would be co-transfected with promoter-YFP constructs of each of the expressed genes (*HOXA5, A9, A10, B2, B3* and *MEIS1*) to see if they can activate any of them. Any of the non-expressed HOX genes (the entire HOXC and HOXD cluster genes) would serve as important negative controls for this assay. A caveat for this experiment is including the appropriate regulatory sequences (which are not known) in the cloned promoter regions.

A more physiologic interrogation of HOX function in AML would involve knocking down individual HOX genes in primary AML samples. We now have a protocol to successfully culture AML samples on a feeder cell layer in the presence of a set of cytokines that allows the AML to expand without differentiating. This protocol could be used to assess the proliferative capacity of treated samples. Using siRNA targeted to each of the expressed HOX genes, single HOX genes could be knocked down to see the effect on the growth of HOX expressing AML cells. It could also require knock down of a

combination of HOX genes to see a self-renewal defect. These siRNAs should have no affect on the self-renewal phenotype of the AML samples that do not express HOX genes (t(8;21), t(15;17) and normal karyotype, HOX negative samples).

While these two experiments address the question of HOX autoregulation and its role in AML self-renewal, they would not answer how the autoregulatory loop is started. Since murine hematopoietic cells do not model the HOX regulation seen in human cells, this would require use of an *in vitro* system. We hypothesize that since the HOX pattern is present in HSCs, the initial activator of HOX gene expression would have to be expressed at some point before HSCs begin to develop. This HSC progenitor could potentially be derived experimentally by growing human ES or iPS cells on mouse OP9 bone marrow stromal cells. After 7 to 8 days of co-culture with OP9s, iPS cells differentiate into hematopoietic progenitors [21]. By harvesting cells on each day of co-culture and analyzing them on expression arrays, we may be able to determine at what stage the HSC HOX genes are first expressed, and search for an upstream regulator which is also turned on at that stage or just prior to it. To determine whether any identified upstream regulators are sufficient and/or necessary for the HOX expression in AML samples, we could return to the AML primary tumor stromal co-culture system. Putative regulatory genes could be overexpressed in the HOX negative samples to see whether they induce the canonical HOX expression pattern, and/or knocked down in the HOX positive samples to see whether they are necessary to maintain the HOX expression pattern.

In addition to determining the mechanisms underlying HOX expression in AML, our results also introduce the question of how the normal karyotype- HOX negative samples acquire self-renewal, a necessary step in tumor pathogenesis. HOX is one of many pathways that have been indicated in hematopoietic stem cell self-renewal. Mouse models have also shown that the Notch signaling pathway is sufficient to induce self-renewal in HSCs [22]. In unpublished work from our laboratory, we have seen that activation of Notch is responsible for self-renewal in murine bone marrow cells that express *PML-RARA*. Hedgehog, Wnt, and FOXO signaling have also been implicated in leukemic self-renewal, although all of them have stronger effects in chronic myeloid leukemia cells (reviewed in [23]). In data not shown, we analyzed proteins involved in the four non-HOX self-renewal pathways to determine whether they are differentially expressed in normal karyotype patients with and without HOX expression. The majority of genes in these pathways are not expressed in tumors. Those genes that are expressed show no difference between the HOX positive and negative normal karyotype samples, suggesting that their dysregulation is not responsible for self-renewal in the HOX negative samples. It is possible that these samples acquire self-renewal via a pathway that is not present in normal hematopoietic cells, but used for self-renewal in the stem cells of a different tissue type.

Functional consequences of SNVs in iPS reprogramming

In Chapter 3, we defined the genetic landscape of 10 iPS clones. In two experiments, we found 189 to 701 SNVs in each of six iPS clones. The majority of these variants were present in 50% of reads, suggesting that they are heterozygous, and present in all cells in

the sample. There are three possibilities for when these SNVs were acquired during the reprogramming process: 1) before reprogramming, 2) at the time of reprogramming, 3) after reprogramming, where one or more of the mutations would be required for selection of the affected cell (the other mutations, although irrelevant for selection, would be passengers in the selected cell). Distinguishing between these possibilities is currently not experimentally tractable. Currently, we pick individual clones from the lentiviral OSK transduction, and expand them in culture to get enough genomic DNA for whole genome sequencing. Laurent, *et al*. showed that continued passaging of human iPS lines leads to increases in copy number variants [24]. We saw a similar result in the private mutations in experiment 3, where variant allele frequencies indicated the presence of a subclone in 3 of 4 iPS lines. If individual cells could be sequenced, we could avoid expansion-related variants since they are clearly not required for the reprogramming process. While this would allow us to distinguish between the last 2 possibilities for the timing of variant acquisition, we would still have to attempt to detect the variants found in reprogrammed cells in the parental cells with highly sensitive methods, to distinguish these from variants caused by the reprogramming event.

In experiment 3, we obtained a unique and highly informative result, that of common variants among all of the iPS lines. Taking our results together with the report by Gore, *et al*. common variants have now been detected in 4 of 10 experiments where more than one iPS line has been examined from a single reprogramming event [25]. However, no common pathways were identified in the iPS clones from the four experiments with common variants. In order to identify reprogramming fitness pathways, many more iPS

lines will need to be sequenced. Since a common founder cell was detected in 40% of these initial studies, at least 20 additional sets of iPS lines should be studied to verify the frequency of this event (with a predicted 8 sets containing a common founder cell). To be confident that common variants are truly the consequence of a shared parental cell (and not the rare chance of 2 cells randomly having the same mutation), at least 4 individual iPS clones should be included in each set of iPS lines, leading to the sequencing of 100 genomes. The next decision to be made would be whether to perform whole genome sequencing or exome sequencing. Since any pathway identification would predominantly utilize variants within coding regions, it is tempting to say that exome sequencing would be sufficient. However, alterations in regulatory regions could also affect pathways; these variants would be missed by exome sequencing. By performing whole genome sequencing now, variants in currently unidentified regulatory regions and unannotated genes could be more relevant in the future, when a fully annotated genome is available.

Which iPS clones should be sequenced? Gore, *et al*. found common variants in iPS lines derived with 3- and 4- factor retrovirus and mRNA-mediated reprogramming, proving that the delivery method does not affect the selection of a "super fit" cell. In our study we used an integrating polycistronic lentivirus so that its integration sites could provide a definitive genetic mark for each clone; however, we now know that each iPS clone has a large group of "private" mutations, which can also be used to define clonality. Since lentiviral transduction with a polycistronic OSK cassette is efficient and clearly has a limited number viral integration sites, it would still be a good choice for iPS

reprogramming. Since mouse iPS clones have the same basic findings as human clones, since mouse cells are more readily available than human cells, since mouse genome sequencing requires less coverage (inbred strains), and mice do not require informed consent for whole genome sequencing, a mouse study might be a better choice for the large study proposed here.

Once whole genome sequencing has been performed for all 20 sets of iPS genomes, those with common variants could be taken forward to identify pathways enriched for in the variants. This pathway analysis could provide candidate genes for functional studies. These studies would include pathway manipulation (both overexpression and knock down) to determine effects on reprogramming efficiency. For example, in experiment 3, one of the common variants was Apaf1 R6C; *Apaf1* is a key regulator of the apoptotic response. If additional common variants in the apoptotic pathway were identified, then cells could be transfected with a plasmid encoding Apaf1 R6C (to see if this is a gain-of-function mutation), and transfected with siRNAs against Apaf1 (to see if it is a loss-of-function) prior to reprogramming with the OSK lentivirus. If transient manipulation of pathways could provide an increase in reprogramming fitness, these results could be used to improve the efficiency and safety of iPS reprogramming for therapeutic uses.

Integrating genetics and epigenetics- the transcriptome

In addition to whole genome sequencing, there are also multiple platforms available for assesing the epigenome in iPS cells. As described in Chapter 2, Illumina recently released their second generation methylation chip, which can assay 485,577 CpG dinucleotides at

single base pair resolution [26]. Methyl-Seq and MeDIP-seq are two additional methods to define methylation at single residues, by performing next-generation sequencing of fragments of DNA isolated by digestion with methylation-sensitive restriction enzymes, or immunoprecipitation with a 5-methylcytosine-specific antibody, respectively [27, 28]. Whole genome sequencing of bisulfite-treated DNA (all unmethylated cytosines are converted to thymine residues, while methylated cytosines are protected from the reaction) allows for a complete, unbiased view of methylation status at all cytosine residues in the genome. Although the mapping of a 3-bp genome presents a difficult computational issue, this problem has been addressed by the development of mapping programs such as BSMAP, but improvements are still required [29].

While these platforms allow for the identification of methylated regions, the biologic consequences of methylation are still not completely understood. For example, it has been reported that 80% of CpG islands in the genome are not located near genes, and most likely have no effect on gene expression [30]. While the dogma of the field has been that promoter methylation is associated with promoter silencing [30], many groups have now shown the importance of intragenic methylation as a positive regulator of gene expression [31-36]. In addition to cytosine methylation, histone modifications also play a large role in gene expression [37-39]. Chromatin immunoprecipitation followed by next-generation sequencing (ChIP-seq) is a technique similar to MeDIP-seq, except that antibodies specific to different histone modifications are used instead of antibodies specific for 5-methylcytosine residues [40, 41]. Deciphering the meaning of epigenetic

differences identified in these comprehensive techniques will be difficult, since the linkage of gene expression patterns to epigenetic modifications is far from complete.

One way to help define the consequences of genetic and epigenetic alterations is to analyze the transcriptome in parallel with the genome. Next-generation sequencing of cDNA (RNA-seq) allows for quantification of expression levels, and identification of post-transcriptional modifications in RNA [42-49]. For whole genome sequencing, parallel transcriptome analysis allows one to determine whether a mutant allele is expressed at equal levels as the wildtype allele, or whether it is associated with RNA degradation. Also, if a mutation is identified in a non-expressed gene, it is most likely not functionally relevant. For epigenetic studies, having paired RNAseq data will help define the transcriptional effects of different epigenetic modifications, clarifying their role in gene expression. Comprehensive analysis of the genome, epigenome, and transcriptome of cancer cells and iPS clones will allow for a greater understanding of the genetic and epigenetic interactions of reprogramming.

***in vivo* vs. *in vitro* reprogramming**

Comprehensive genomic analysis of AML samples and iPS lines has allowed us to identify similarities between these two forms of reprogramming. In both, we have determined that though there is a large mutational burden; it is clear that the majority of SNVs are passenger mutations that antedated reprogramming. Acquisition of mutations occurs over time in all organisms (genetic drift); in Chapter 3, this is illustrated by the homozygous variants detected in each of the three mice used to generate fibroblasts. In

other experiments from our group, we have sequenced individual HSCs from three healthy volunteers and seen the accumulation of benign mutations as a function of time [Ley, *et al*. manuscript in preparation]; similarly, there is a strong correlation between total mutations and age in the AML samples we have sequenced [Welch, *et al*. in preparation]. In AML, some recurrent mutations are able to act as "drivers" to initiate leukemic transformation, such as *DNMT3A* [50, 51]. The results of our third iPS experiment suggest that a similar phenomenon can happen in reprogramming, where one or more mutations (generated randomly during the life of the cell) creates a change in a cell that makes it "super fit" for reprogramming [25]. In a cancer cell, this kind of mutation would be known as a driver, since it can cooperate with progression mutations that give the cell a survival and/or proliferative advantage. Sequencing of more iPS clones may elucidate recurrent mutations involved in *in vitro* reprogramming, similar to what has been found in AML and other cancers.

Acquisition of self-renewal is a key step in both *in vivo* and *in vitro* reprogramming. In iPS production, overexpression of Oct4 and Sox2 activates pluripotency pathways in somatic cells (Oct4, Sox2 and Nanog. which then act in positive feedback loops to maintain self-renewal). The initiation of self-renewal is not as clear-cut in AML. While reprogramming translocations have been shown to induce self-renewal in hematopoietic cells [52, 53], in Chapter 2 we show that the majority of normal karyotype AML overexpress HOX genes in a canonical fashion that mimics that of normal CD34 cells, suggesting that these samples acquire self-renewal by "capturing" the normal program in the leukemia initiating cell. However, in the AML cases without HOX expression, other

known signaling pathways that affect self-renewal are not expressed; the mechanism for initiating the self-renewal in these samples remains an unknown.

Both *in vivo* and *in vitro* reprogramming lead to generation of cells with limitless growth potential. A complete understanding of the similarities between these two processes is important for the future therapeutic use of iPS cells. By definition, iPS cells form tumors when introduced into immunodeficient mice. In order to use iPS cells for regenerative medicine, they will almost certainly need to be differentiated into the desired cell type to avoid tumor growth. By understanding the pathways involved in tumorigenesis, it may be possible to more safely engineer iPS cells for therapeutic use. The key to using iPS cells clinically will be to make them less like tumors, and more like normal, differentiated tissue cells.

References

1. Andreeff, M., et al., *HOX expression patterns identify a common signature for favorable AML.* Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K, 2008. **22**(11): p. 2041-7.
2. Armstrong, S.A., et al., *MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia.* Nature genetics, 2002. **30**(1): p. 41-7.
3. Becker, H., et al., *Favorable prognostic impact of NPM1 mutations in older patients with cytogenetically normal de novo acute myeloid leukemia and associated gene- and microRNA-expression signatures: a Cancer and Leukemia Group B study.* Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 2010. **28**(4): p. 596-604.
4. Casas, S., et al., *Aberrant expression of HOXA9, DEK, CBL and CSF1R in acute myeloid leukemia.* Leukemia & lymphoma, 2003. **44**(11): p. 1935-41.
5. Drabkin, H.A., et al., *Quantitative HOX expression in chromosomally defined subsets of acute myelogenous leukemia.* Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K, 2002. **16**(2): p. 186-95.
6. Haferlach, C., et al., *AML with mutated NPM1 carrying a normal or aberrant karyotype show overlapping biologic, pathologic, immunophenotypic, and prognostic features.* Blood, 2009. **114**(14): p. 3024-32.
7. Kawagoe, H., et al., *Expression of HOX genes, HOX cofactors, and MLL in phenotypically and functionally defined subpopulations of leukemic and normal human hematopoietic cells.* Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K, 1999. **13**(5): p. 687-98.
8. Lawrence, H.J., et al., *Frequent co-expression of the HOXA9 and MEIS1 homeobox genes in human myeloid leukemias.* Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K, 1999. **13**(12): p. 1993-9.
9. Nakamura, T., et al., *Cooperative activation of Hoxa and Pbx1-related genes in murine myeloid leukaemias.* Nature genetics, 1996. **12**(2): p. 149-53.
10. Rice, K.L. and J.D. Licht, *HOX deregulation in acute myeloid leukemia.* The Journal of clinical investigation, 2007. **117**(4): p. 865-8.
11. Roche, J., et al., *Hox expression in AML identifies a distinct subset of patients with intermediate cytogenetics.* Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K, 2004. **18**(6): p. 1059-63.
12. Golub, T.R., et al., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 1999. **286**(5439): p. 531-7.
13. Amsellem, S., et al., *Ex vivo expansion of human hematopoietic stem cells by direct delivery of the HOXB4 homeoprotein.* Nature medicine, 2003. **9**(11): p. 1423-7.
14. Antonchuk, J., G. Sauvageau, and R.K. Humphries, *HOXB4 overexpression mediates very rapid stem cell regeneration and competitive hematopoietic repopulation.* Experimental hematology, 2001. **29**(9): p. 1125-34.

15. Sauvageau, G., et al., *Overexpression of HOXB4 in hematopoietic cells causes the selective expansion of more primitive populations in vitro and in vivo.* Genes & development, 1995. **9**(14): p. 1753-65.
16. Ono, R., et al., *Mixed-lineage-leukemia (MLL) fusion protein collaborates with Ras to induce acute leukemia through aberrant Hox expression and Raf activation.* Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K, 2009. **23**(12): p. 2197-209.
17. Paggetti, J., et al., *Crosstalk between leukemia-associated proteins MOZ and MLL regulates HOX gene expression in human cord blood CD34+ cells.* Oncogene, 2010. **29**(36): p. 5019-31.
18. Quentmeier, H., et al., *Expression of HOX genes in acute leukemia cell lines with and without MLL translocations.* Leukemia & lymphoma, 2004. **45**(3): p. 567-74.
19. Thirman, M.J., et al., *Rearrangement of the MLL gene in acute lymphoblastic and acute myeloid leukemias with 11q23 chromosomal translocations.* The New England journal of medicine, 1993. **329**(13): p. 909-14.
20. Scholl, C., et al., *The homeobox gene CDX2 is aberrantly expressed in most cases of acute myeloid leukemia and promotes leukemogenesis.* The Journal of clinical investigation, 2007. **117**(4): p. 1037-48.
21. Vodyanik, M.A. and Slukvin, II, *Hematoendothelial differentiation of human embryonic stem cells.* Current protocols in cell biology / editorial board, Juan S. Bonifacino ... [et al.], 2007. **Chapter 23**: p. Unit 23 6.
22. Varnum-Finney, B., et al., *Pluripotent, cytokine-dependent, hematopoietic stem cells are immortalized by constitutive Notch1 signaling.* Nature medicine, 2000. **6**(11): p. 1278-81.
23. Heidel, F.H., B.G. Mar, and S.A. Armstrong, *Self-renewal related signaling in myeloid leukemia stem cells.* International journal of hematology, 2011. **94**(2): p. 109-17.
24. Laurent, L.C., et al., *Dynamic changes in the copy number of pluripotency and cell proliferation genes in human ESCs and iPSCs during reprogramming and time in culture.* Cell stem cell, 2011. **8**(1): p. 106-18.
25. Gore, A., et al., *Somatic coding mutations in human induced pluripotent stem cells.* Nature, 2011. **471**(7336): p. 63-7.
26. Bibikova, M., et al., *High density DNA methylation array with single CpG site resolution.* Genomics, 2011. **98**(4): p. 288-95.
27. Brunner, A.L., et al., *Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver.* Genome research, 2009. **19**(6): p. 1044-56.
28. Down, T.A., et al., *A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis.* Nature biotechnology, 2008. **26**(7): p. 779-85.
29. Xi, Y. and W. Li, *BSMAP: whole genome bisulfite sequence MAPping program.* BMC bioinformatics, 2009. **10**: p. 232.
30. Takai, D. and P.A. Jones, *Comprehensive analysis of CpG islands in human chromosomes 21 and 22.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(6): p. 3740-5.

31. Lorincz, M.C., et al., *Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells.* Nature structural & molecular biology, 2004. **11**(11): p. 1068-75.
32. Maunakea, A.K., et al., *Conserved role of intragenic DNA methylation in regulating alternative promoters.* Nature, 2010. **466**(7303): p. 253-7.
33. Flanagan, J.M. and L. Wild, *An epigenetic role for noncoding RNAs and intragenic DNA methylation.* Genome biology, 2007. **8**(6): p. 307.
34. Ball, M.P., et al., *Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells.* Nature biotechnology, 2009. **27**(4): p. 361-8.
35. Rauch, T.A., et al., *A human B cell methylome at 100-base pair resolution.* Proceedings of the National Academy of Sciences of the United States of America, 2009. **106**(3): p. 671-8.
36. Eckhardt, F., et al., *DNA methylation profiling of human chromosomes 6, 20 and 22.* Nature genetics, 2006. **38**(12): p. 1378-85.
37. Peterson, C.L. and M.A. Laniel, *Histones and histone modifications.* Current biology : CB, 2004. **14**(14): p. R546-51.
38. Berger, S.L., *Histone modifications in transcriptional regulation.* Current opinion in genetics & development, 2002. **12**(2): p. 142-8.
39. Cheung, P. and P. Lau, *Epigenetic regulation by histone methylation and histone variants.* Molecular endocrinology, 2005. **19**(3): p. 563-73.
40. Lennartsson, A. and K. Ekwall, *Histone modification patterns and epigenetic codes.* Biochimica et biophysica acta, 2009. **1790**(9): p. 863-8.
41. Wang, Z., et al., *Combinatorial patterns of histone acetylations and methylations in the human genome.* Nature genetics, 2008. **40**(7): p. 897-903.
42. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nature reviews. Genetics, 2009. **10**(1): p. 57-63.
43. Nagalakshmi, U., et al., *The transcriptional landscape of the yeast genome defined by RNA sequencing.* Science, 2008. **320**(5881): p. 1344-9.
44. Wilhelm, B.T., et al., *Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution.* Nature, 2008. **453**(7199): p. 1239-43.
45. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nature methods, 2008. **5**(7): p. 621-8.
46. Lister, R., et al., *Highly integrated single-base resolution maps of the epigenome in Arabidopsis.* Cell, 2008. **133**(3): p. 523-36.
47. Cloonan, N., et al., *Stem cell transcriptome profiling via massive-scale mRNA sequencing.* Nature methods, 2008. **5**(7): p. 613-9.
48. Marioni, J.C., et al., *RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.* Genome research, 2008. **18**(9): p. 1509-17.
49. Morin, R., et al., *Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing.* BioTechniques, 2008. **45**(1): p. 81-94.

50. Ley, T.J., et al., *DNMT3A mutations in acute myeloid leukemia.* The New England journal of medicine, 2010. **363**(25): p. 2424-33.
51. Yan, X.J., et al., *Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia.* Nature genetics, 2011. **43**(4): p. 309-15.
52. Fenske, T.S., et al., *Stem cell expression of the AML1/ETO fusion protein induces a myeloproliferative disorder in mice.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(42): p. 15184-9.
53. Tonks, A., et al., *The AML1-ETO fusion gene promotes extensive self-renewal of human primary erythroid cells.* Blood, 2003. **101**(2): p. 624-32.

Margaret Ashley Young

EDUCATION

M.D./Ph.D. Student, Immunology, Washington University Medical Scientist Training Program, St. Louis, MO, **2005-2013**

- Thesis Advisor: Timothy J. Ley, M.D., Department of Medicine

B.S., Biological Sciences, Carnegie Mellon University, Pittsburgh, PA, **2001-2005**

- Concentration in Developmental and Cell Biology
- University Honors
- Mellon College of Science Research Honors
- Honors Thesis: "Proteomic Analysis of Cell Shape Changes during Gastrulation in Drosophila"

AWARDS, DISTINCTIONS, AND FELLOWSHIPS

Beckman Scholar, Carnegie Mellon University, Pittsburgh, PA, 2003-2004

- Laboratory of Jonathan S. Minden, Ph.D, Department of Biological Sciences

High Honor Dean's List, Carnegie Mellon University, Pittsburgh, PA, **2001-2005**

Science and Humanities Scholar, Carnegie Mellon University, Pittsburgh, PA, **2001-2005**

Howard Hughes Medical Institute, National Institutes of Health, Montgomery County Public Schools Student Teacher Internship Program, National Institutes of Health, Bethesda, MD, **2000-2001**

- Laboratory of George P. Chrousos, MD, Chief, Pediatric and Reproductive Endocrinology Branch, National Institutes of Health

PUBLICATIONS AND PRESENTATIONS

Young M, Larson D, Sun CW, George D, Ding L, Miller C, Lin L, Pawlik K, Chen K, Fan X, Schmidt H, Kalicki-Veizer J, Cook L, Swift G, Demeter R, Wendl M, Sands M, Mardis E, Wilson R, Townes T, Ley T. Background mutations in parental cells account for most of the genetic heterogeneity of induced pluripotent stem cells. Cell Stem Cell. 2012 May 4; 10(5):570-82.

Young M, Spencer D, Lamprecht T, Triche T, Germain D, Gilmore P, Larson D, Townsend R, Mardis E, Wilson R, Laird P, Ley T. Canonical and non-canonical HOX expression patterns in AML. *In preparation*.

Ding L, Ley T, Larson D, Miller C, Koboldt D, Welch J, Ritchey J, **Young M**, Lamprecht T, McLellan M, McMichael J, Wallis J, Lu C, Shen D, Harris C, Dooling D, Fulton R, Fulton L, Chen K, Schmidt H, Kalicki-Veizer J, Magrini V, Cook L, McGrath S, Vickery T, Wendl M, Heath S, Watson M, Link D, Tomasson M, Shannon W, Payton J, Kulkarni S, Westervelt P, Walter M, Graubert T, Mardis E, Wilson R, and DiPersio J. Patterns of clonal evolution in relapsed acute myeloid leukemia revealed by whole genome sequencing. Nature. 2012 Jan 11; 481(7382):506-10.

Gong L, Puri M, Ünlü M, **Young M**, Robertson K, Viswanathan S, Krishnaswamy A, Dowd SR and Minden JS. Drosophila Ventral Furrow Morphogenesis: A Proteomic Analysis. Development, 2004; 131:643-656.

Hiroi N, Wong ML, Licinio J, Park C, **Young M**, Gold PW, Chrousos GP, Bornstein SR. Expression of corticotropin releasing hormone receptors type I and type II mRNA in suicide victims and controls. Molecular Psychiatry, 2001; 6(5): 540-6.

Presentation: "Correcting Murine Mucopolysaccaridosis Type VII using homology directed repair induced by meganucleases" American Society of Clincal Investigation/ American Physicians Association Joint Meeting, Chicago, IL, **2010**

Presentation: "Meganuclease-Induced Correction of Murine Mucopolysaccaridosis Type VII (MPSVII)" Northwest Genome Engineering Consortium Workshop, Seattle, WA, **2009**

Presentation: "Proteomic Analysis of Cell Shape Changes during Gastrulation in Drosophila," Beckman Scholars Conference, Irvine, CA, **2004**

Presentation: "The Corticotropin Releasing Hormone (CRH) System in Human Sebocytes" Student Teacher Program Symposium, Chevy Chase, MD, **2001**

RESEARCH EXPERIENCE

Laboratory of Timothy J. Ley, M.D., Washington University School of Medicine, Department of Medicine, Division of Oncology, Section of Stem Cell Biology, St. Louis, MO, **2007-2011**

Laboratory of Jonathan S. Minden, Ph.D., Carnegie Mellon University, Department of Biological Sciences, Pittsburgh, PA, **2003-2005**

Laboratory of George P. Chrousos, M.D., National Institutes of Health, National Institute of Child Health and Development, Bethesda, MD, **2000-2001**