A Targeted CRISPR-Cas9 Screen in AML Cells

A Thesis Presented

by

Sam Blake Chiappone

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Master of Science

in

Biochemistry and Cell Biology

Stony Brook University

December 2015

Stonybrook University

The Graduate School

Sam Blake Chiappone

We, the thesis committee for the above candidate for the

Master of Science degree, hereby recommend

acceptance of this thesis

Lingbo Zhang

Group Leader and Fellow, CSHL

Arne Stenlund

Associate Professor, CSHL

This thesis is accepted by the graduate school

Charles Taber

Dean of the Graduate School

Abstract of the Thesis

A Targeted CRISPR-Cas9 Screen in AML Cells

by

Sam Blake Chiappone

Master of Science

in

Biochemistry and Cell Biology

Stony Brook University

2015

The genetic underpinnings of a cancer are key to understanding the etiology of the disease and thus to providing a cure, because any robust cure of the disease will need to target a therapeutic window, a gene or a group of genes that are different between the cancerous cells and the normal cells. The research shown in this thesis gives insight into the genetic causes of acute myelogenous leukemia in an attempt to find a therapeutic window for this disease. A genetic screen using CRISPR-Cas9 is designed and conducted in a cell line designated RN 2-5 that was determined to recapitulate the acute myelogenous leukemia phenotype in culture and in a mouse model of the disease. Following up on the most important hits of this screen will very likely lead to important therapeutic targets. Two bioinformatics methods, a machine-learning algorithm and a statistical approach, are also used to identify key features of the expression of genes targeted in the screen. A combination of genetic screening and bioinformatics will have great efficacy in the future in finding genes differentially expressed, and differentially required, by cancer cells versus normal cells.

Table of Contents

List of Figures and Tables

Introduction

Hematopoiesis is the process of generating new blood, in particular new blood cells. There are over a dozen different mature blood cell types, including two major types known as red blood cells (RBC), derived from erythroblasts, and platelets, derived from megakaryocytes. The largest diversity of blood cell types, however, is in the category of white blood cells, which include B-lymphocytes, T-lymphocytes, basophils, eosinophils, and monocytes, to name a few. All of these diverse cell types, from platelets to monocytes, are derived from a single stem cell type, the Hematopoietic Stem Cell (HSC). This stem cell population originates in the fetal liver, migrates to the blood compartment, and in mice briefly occupies the spleen,[1] before finally migrating to the bone marrow compartment where it remains for the lifetime of the organism in all mammals.

The HSC differentiates into its diverse progeny in a series of hierarchical steps, each of which represents a different progenitor cell population. There are, in total, dozens of such progenitor cell types, and these cell types, along with their ultimate stem cell the HSC, are known as Hematopoietic Stem and Progenitor Cells (HSPC). Disregulation of any of these stages can result in a vast array of different blood disorders. This can be due either to aberrant downregulation of differentiation from an HSPC stage or of the self-renewal of that particular HSPC population itself, which results in various cytopenias or, in the special case of RBC, in anemia – or upregulation of any of these stages can result in various different blood cancers, including lymphoid leukemias (T- and B- cell lines), myeloid leukemias (the eosin/baso/neutrophil and monocytic cell lines), erythroleukemias (the RBC cell line), and megakaryocytic leukemia (the megakaryocyte/platelet cell line).

Thus, hematopoiesis provides a map of all the different possible hematological disorders, each of which can result from either up- or down- regulation of each of the HSPC stages. Frequently, hematological disorders will affect a subset of mature cell types – for instance, all of the lymphoid white blood cells – because it stems from an upstream progenitor stage. The various hematological cytopenias and malignancies represent a huge portion of the causes of human mortality. Anemia alone affects 1.62 billion people worldwide[2] and although half of these cases are due to iron deficiency, especially in developing countries, the other half (810 million cases) are due to inborn or acquired defects in hematopoiesis, especially defects associated with aging. The financial burden of treating these cases is in the tens or even hundreds of billions of dollars. It is worth noting that this is the cost of treating just one hematological disorder, albeit a common one with many different manifestations.

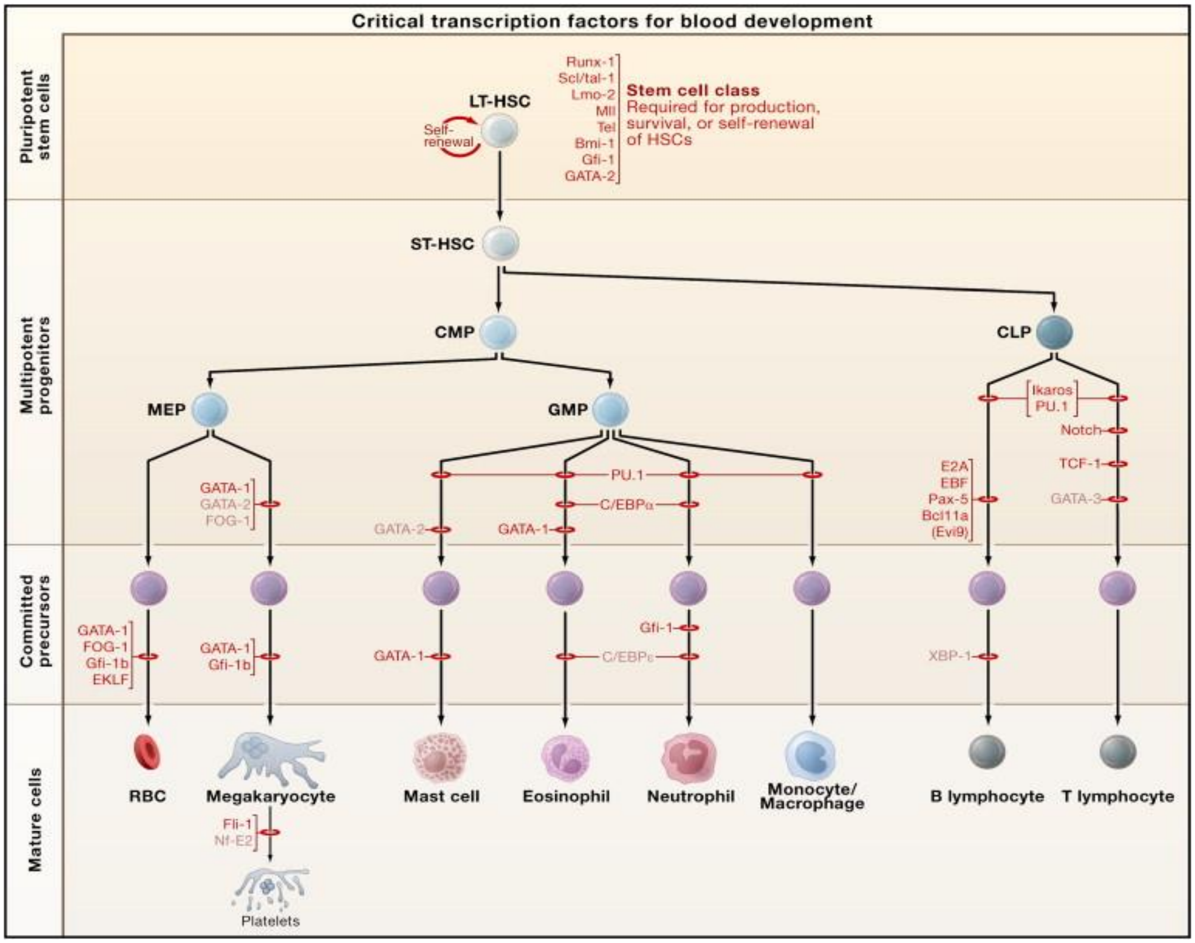


Figure 1. Schematic diagram of hematopoietic stages, along with some of the major transcription factors involved in differentiation at each stage. The differentiation pathway begins with an HSC, which becomes all of the blood cell types. The bottom row represents mature cell types, which have finished differentiation and stopped dividing. The bottom row also nicely orders the different main lineages in blood, from right to left; RBC, platelet, the myeloid lineage (although sometimes the term myeloid also refers to the RBC and megakaryocyte lineages), and the lymphoid lineage. Adapted from *Hematopoiesis: An Evolving Paradigm for Stem Cell Biology* S.H. Orkin and L.I. Zon 2008.[3]

This thesis revolves around understanding a hematological disorder involving upregulation of a hematopoietic lineage group, the myeloid lineage. This hematological malignancy is known as acute myelogenous leukemia (AML). An estimated 352,000 people are affected by AML worldwide.[4] Over half of these cases are lethal, according to patient outcomes in the United States. Cases in significantly less developed countries are likely to have even poorer outcomes. There is only one diagnostic subcategory of AML, called Acute Promyelocytic Leukemia (APL), which has a relatively good prognosis, and this is due to the efficacy of All-Trans Retinoic Acid (ATRA) in treating this particular subcategory of AML. The other forms of AML are refractory to this treatment, and ATRA is not included in the standard chemotherapy regime for non-APL AML. In particular, the Mixed-Lineage Leukemia (MLL) form of AML has a dismal prognosis. This subcategory of AML is extremely refractory to the current standard regime of chemotherapy. The research presented herein is part of a larger attempt to develop a novel therapy for the treatment of AML, and it was done under the guidance of, and where it is noted, with the direct assistance of, Lingbo Zhang (LZ) of Cold Spring Harbor Laboratory.

Although the genetic causes of some subcategories of AML are known – for instance, MLL AML is known to involve translocations at the 11q23 locus, which is the site of the *MLL* gene, also known as the Histone-lysine N-methyltransferase 2A gene, which attach aberrant enzymatic activities to the targeting domain of the MLL gene product – the genetic underpinnings of the majority of AML cases are not fully understood. The genetic aberrations involved in the majority of AML cases are likely to be diverse, meaning that there are many possible changes that can contribute, and combinatorial, meaning that more than one change needs to occur, and different changes can substitute for one another. Therefore, finding the different combinations of changes that can lead to AML is difficult, because there may be many different such combinations. In fact, it may be that the majority of AML patients each have a unique combination, in which case approaching the problem of treatment from a gene-specific or cytogenetic-specific perspective becomes particularly onerous. Interrogating the function of one gene at a time in the etiology of AML is, in that case, unlikely to be useful.

Instead, a high-throughput approach can be used to identify genes for which loss-of-function stimulates progression of the disease, and for which loss-of-function halts the progression of the disease and the proliferation of the diseased cells. Such high-throughput genetic methods using CRISPR-Cas9 technology have recently been pioneered in the Genome-wide CRISPR Knock-Out (GeCKO) experiments by the Feng Zhang lab at the Broad Institute.[5] These experiments use CRISPR-Cas9 technology, in a mammalian system, to knock out every gene in the mammalian genome, and recapitulate a phenotype. In the simplest case, this phenotype is cell proliferation or cell death. In their experiments, they were able to identify many genes known to be strongly required for cell survival and proliferation, in particular ribosomal RNA genes and other genes involved in gene transcription and protein synthesis. They were also able to demonstrate, in melanoma cells, which genes were required for sensitivity to the BRAF inhibitor Vemurafenib. This demonstrated the ability to use CRISPR-Cas9 in a genome-wide screen to identify genes responsible for a particular phenotype.

The CRISPR-Cas9 method itself is an adaptation of a bacterial defense mechanism against viruses. The genomes of bacteria, such as *Streptococcus pyogenes*, house a number of regions that each consist of short palindromic repeats separated by short ~20 bp variable regions called protospacers. The repeats are known as Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR). It turned out that each of the protospacers was complimentary to a sequence in a viral genome, and that when these CRISPR-containing regions are transcribed, each of the protospacers is cleaved into individual ~20 bp RNA's that act as CRISPR-targeting RNA's (crRNA). The way that the protospacers are processed in such a way, is that an RNA complimentary to the short palindromic repeat, known as the trans-activating crRNA (tracrRNA), hybridizes to the repeat, causing endonucleolytic processing by RNase III in the bacterium. Then both the crRNA (the original protospacer) and the tracrRNA binds to a CRISPR-Associated (Cas) enzyme, and can target the enzyme to a gene sequence. The Cas nuclease of choice for CRISPR technology is the Cas9 enzyme. The only requirement in the target region for Cas9 cutting, other than being complimentary to the crRNA, is that it must have next to the target sequence the Protospacer Adjacency Motif (PAM) sequence, which for Cas9 is NGG. One of the key advancements pioneered by the Feng Zhang laboratory is the discovery that when the crRNA and tracrRNA are synthesized as one long RNA transcript, they can still bind to Cas9 and direct it to genomic regions. Thus only one RNA is needed, and such an RNA is known as a single-guide RNA (sgRNA).

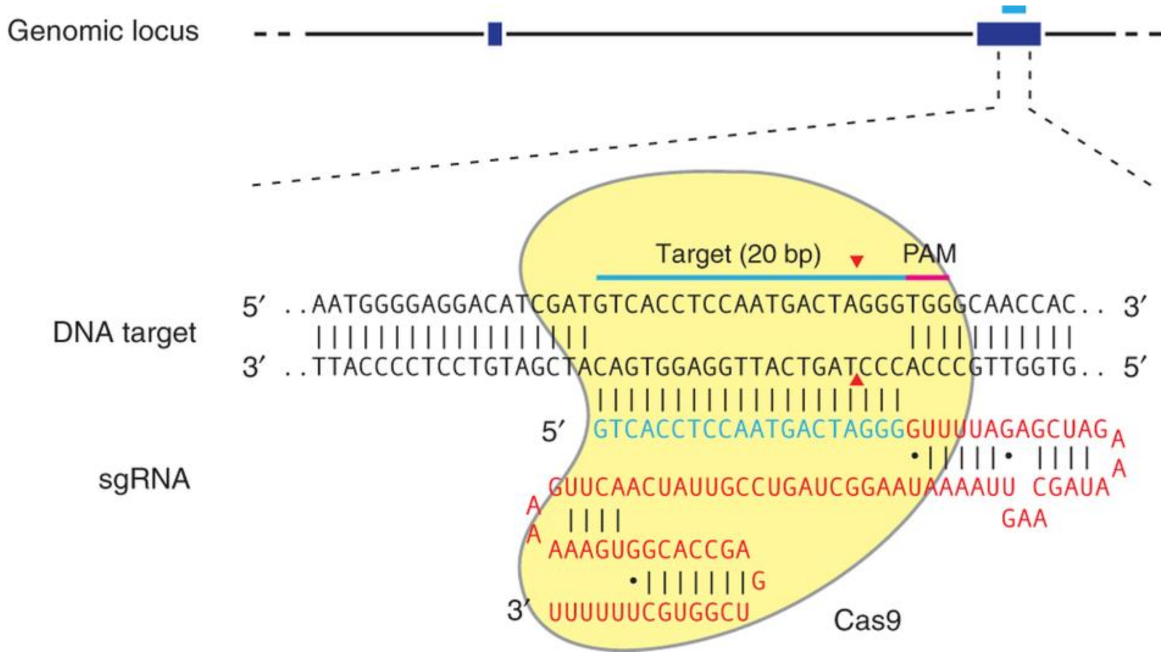

Genomic locus
Target (20 bp)
PAM
DNA target
5′ ..AATGGGGAGGACATCGATGTCACCTCCAATGACTAGGGTGGGCAACCAC.. 3′
3′ ..TTACCCCTCCTGTAGCTACAGTGGAGGTTACTGATCCCACCCGTTGGTG.. 5′
sgRNA
5′ GTCACCTCCAATGACTAGGGGUUUUAGAGCUAG
A
A
GUUCAACUAUUGCCUGAUCGGAAUAAAAUU CGAUA
GAA
A
A
AAAGUGGCACCGA
G
3′ UUUUUUCGUGGCU
Cas9

Figure 2. Outline of the CRISPR-Cas9 system. The sgRNA consists of the crRNA, highlighted in blue, and the tracrRNA, highlighted in red. This artificial, composite RNA fits into the RNA-binding pocket of the Cas9 enzyme. The crRNA forms base pairs with the strand of genomic DNA shown, and the enzyme make two endonucleolytic cleavages at the positions marked by reds triangles through the activity of two distinct endonuclease domains. The only other requirement is the PAM, shown next to the target sequence. Adapted from *Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells* O. Shalem *et al* 2013.[6]

The key feature of the Cas9 enzyme is that it is an RNA-dependant DNA endonucleases. This means that it cleaves DNA endonucleolytically based on an RNA template. The outcome of this is that, in contrast to previous genome-editing technology such as Transcription Activator-Like EndoNucleases

(TALEN's), only one enzyme can be used to edit virtually any genetic sequence in the genome (it is programmable), depending on which sgRNA's it is given as a template, and it can do this in duplex, triplex, or multiplex, depending on how many sgRNA's it is given. Although the native Cas9 enzyme introduces double strand breaks by cleaving the two strands separately with two different endonuclease domains, a Cas9 nickase has been engineered that only has one active endonuclease domain. This allows for genuine editing of the genome, including directed mutagenesis and inserting an entire gene into the genome, by introducing the Cas9 enzyme and two RNA's, one for each end of the region you want to alter, but with the opposite sense, such that an effective double-strand break is introduced. This area can potentially be very large. Then DNA complimentary to the sticky ends produced by such cleavage can be introduced, which will be incorporated in the way of traditional homologous recombination, albeit with much greater frequency. Note that both the normal CRISPR-Cas9 protocol and the CRISPR-Cas9 double-nickase protocol rely on DNA repair to achieve mutagenesis, in the first case because Non-Homologous End-Joining (NHEJ) is required to create a frame-shift, and in the second case because homologous repair is required to achieve homologous recombination. Other uses for CRISPR-Cas9 utilizing a nucloelytically dead Cas9 (both endonuclease domains inactivated) have been found, which utilize the CRISPR-Cas9 system purely for its targeting capabilities.[7]

Following the work of Shalem *et al* in the Feng Zhang lab, we undertook a smaller scale library knockout screen to identify metabolic genes important to the etiology of AML. The work of Feng Zhang relies on a Lentiviral vector to deliver the DNA required for CRISPR-Cas9 directly to the nucleus of target cells, where it integrates into the host genomic DNA. The mRNA for the *S. pyogenes* Cas9 enzyme as well as the sgRNA itself is then transcribed from the integrated DNA, completing the requirements for CRISPR-Cas9. This vector, called LentiCRISPR v2 is the other key contribution from the Feng Zhang lab, and it is shown in figure 3. For our library knockout screen in mice, we designed a library of sgDNA's covering a list of 372 genes provided by LZ. The guides were taken from the supplemental material of

Yusa *et al* [8] in which they used bioinformatics to identify all potential sgDNA's in gene sequences in mice, and then applied strict filtering to remove potential sgDNA sequences that had significant off-targets elsewhere in the genome. Each of the 372 genes was assigned five different sgDNA's unless fewer than five were available, in which case all the available sgDNA's were used. This resulted in a grand total of 1702 sgDNA's. These 1702 sgDNA's were then used in a targeted library knockout screen based on the LentiCRISPR v2 vector.

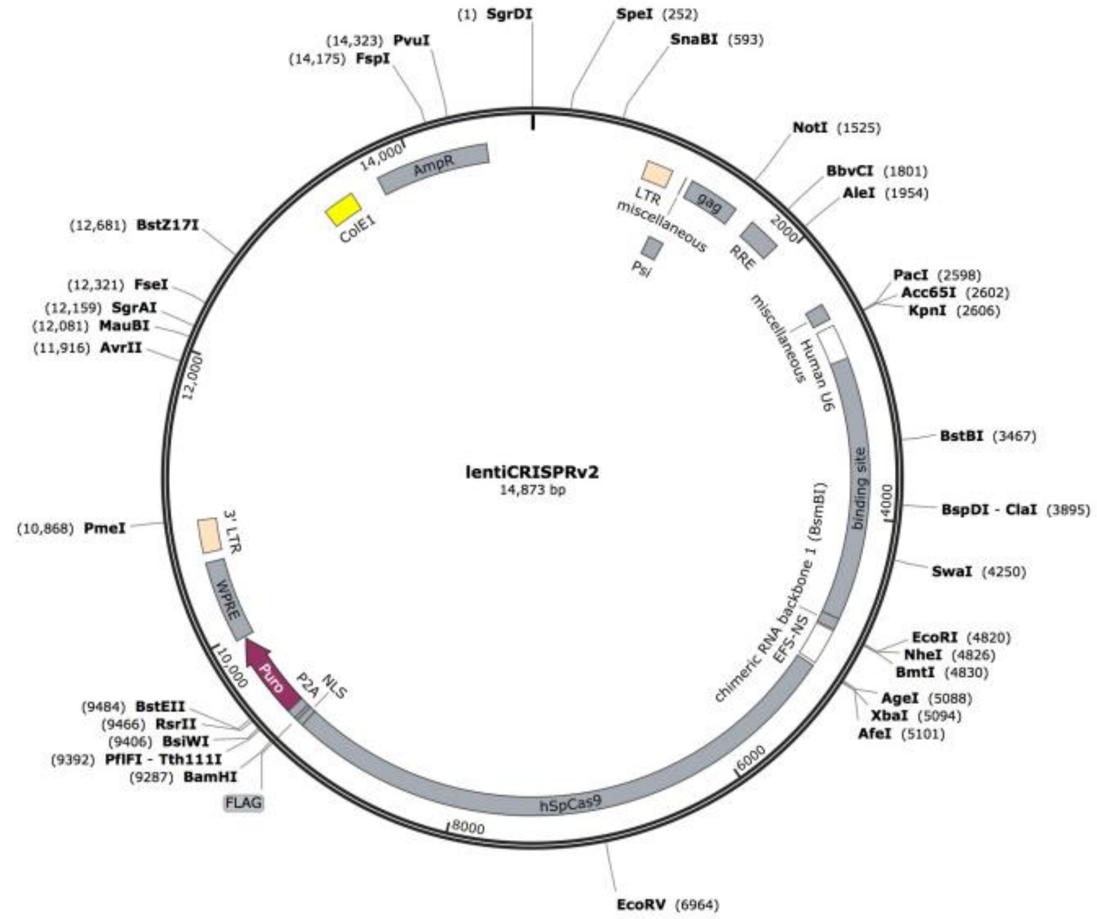


Figure 3. A diagram of the LentiCRISPR v2 plasmid from the Feng Zhang lab. The plasmid contains the *gag* gene from HIV-1 (lentivirus) and lentiviral integration sites (5'- and 3'- LTR) as well as the lentiviral packaging signal (Psi). Between the LTR's is a cloning site for the different sgRNA sequences, flanked by the human U6 promoter and the tracrDNA sequence, and an open reading frame containing the *S. pyogenes* Cas9 gene linked to the Puromycin resistance gene via a picornavirus P2A autocleavage site, both under control of the EFS promoter. This plasmid was used as the backbone of our targeted library knockout screen. Adapted from Addgene entry Plasmid #52961.

Results and Methods

A cloning scheme was developed such that the 372 genes could be split into two lists, each of which could be amplified separately, then cloned into the LentiCRISPR v2 plasmid using Gibson assembly, and finally used to transfect and infect in a pooled manner. The cloning scheme is outlined in figure 4. First, 1702 oligos were ordered from Invitrogen in a pooled sample, each one consisting of a fragment of the human U6 promoter, a G nucleotide appended directly afterward, a 19-nt guide sequence that varied between each of the oligos, followed by the tracrDNA sequence, and finally one of two different barcodes that served as primer sites for the next step of the cloning. The barcodes were chosen by using the Primer3 web-based tool and selecting for the same melting temperature as the forward primer, a random sequence, and no similar sequence in the first 4 bp's. The single G was appended before each of the sgDNA's to ensure that every sgDNA was transcribed from the U6 promoter at the same level.[9]

Table 1. Protocol for PCR amplification with Platinum *Pfx* DNA Polymerase.

Component	Volume (µL)
10X *Pfx* Amplification Buffer	5.5
10 mM dNTP Mixture	1.5
50 mM $MgSO_4$	1
Primers (10 µM each)	1.5 (each)
Platinum *Pfx* DNAP	0.4
Molecular-grade H_2O	37
Template	2 (1st PCR) or 5 (2nd PCR)

Once the pooled oligos were received, the two lists of genes of genes were PCR amplified separately in two sequential PCR reactions. The protocol used (per reaction) for both PCR reactions is shown in Table 1, and is generally the same as the manufacturer's protocol for Platinum *Pfx* DNA

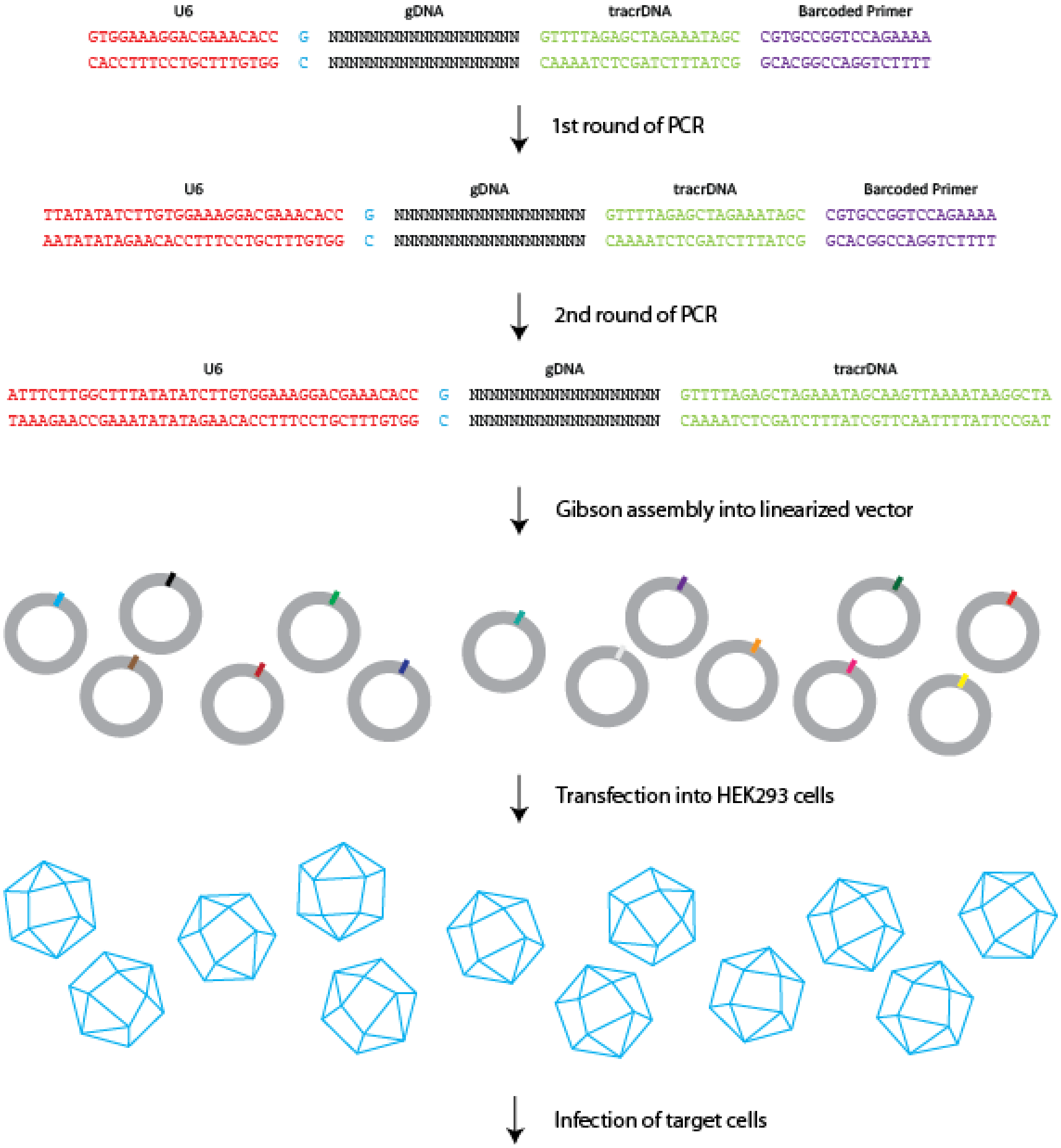

U6
gDNA
tracrDNA
Barcoded Primer
GTGGAAAGGACGAAACACC G NNNNNNNNNNNNNNNNNNNN GTTTTAGAGCTAGAAATAGC CGTGCCGGTCCAGAAAA
CACCTTTCCTGCTTTGTGG C NNNNNNNNNNNNNNNNNNNN CAAAATCTCGATCTTTATCG GCACGGCCAGGTCTTTT
1st round of PCR
U6
gDNA
tracrDNA
Barcoded Primer
TTATATATCTTGTGGAAAGGACGAAACACC G NNNNNNNNNNNNNNNNNNNN GTTTTAGAGCTAGAAATAGC CGTGCCGGTCCAGAAAA
AATATATAGAACACCTTTCCTGCTTTGTGG C NNNNNNNNNNNNNNNNNNNN CAAAATCTCGATCTTTATCG GCACGGCCAGGTCTTTT
2nd round of PCR
U6
gDNA
tracrDNA
ATTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACC G NNNNNNNNNNNNNNNNNNNN GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTA
TAAAGAACCGAAATATATAGAACACCTTTCCTGCTTTGTGG C NNNNNNNNNNNNNNNNNNNN CAAAATCTCGATCTTTATCGTTCAATTTTATTCCGAT
Gibson assembly into linearized vector
Transfection into HEK293 cells
Infection of target cells

Figure 4. Schematic outline of the library cloning and infection protocol. The protocol begins with 1702 oligos ordered from Invitrogen. The first round of PCR serves to selectively amplify the list of genes separately as well as to add a portion of the human U6 sequence. The second round of PCR replaces the barcode with an extended portion of the tracrDNA sequence, as well as further extending the U6 sequence. After linearizing the vector, the vector and the pool of amplified oligos are Gibson assembled into a pool of plasmids, resulting in a pool of plasmids containing many different inserts, symbolized by the tiny region of variable color in the circular plasmids. Finally, the pooled plasmids are used to transfect HEK 293 cells together with the plasmids required for lentiviral production, and the transfected cells produce a pool of viruses each containing a different plasmid, which is then ready to be used to infect any cells of interest.

Polymerase PCR from Invitrogen. The first reaction was done using two different reverse primers complimentary to the two different barcodes. The forward primer was common to both PCR reactions, and had a sequence complimentary to the fragment of the U6 promoter in the oligos, plus an additional stretch of the U6 seqeunce. The next reaction was performed using the product of the first reaction as template, and using a different forward primer with even more of the human U6 sequence in it, and a reverse primer that was complimentary to the tracrDNA sequence in the oligos, plus additional tracrDNA

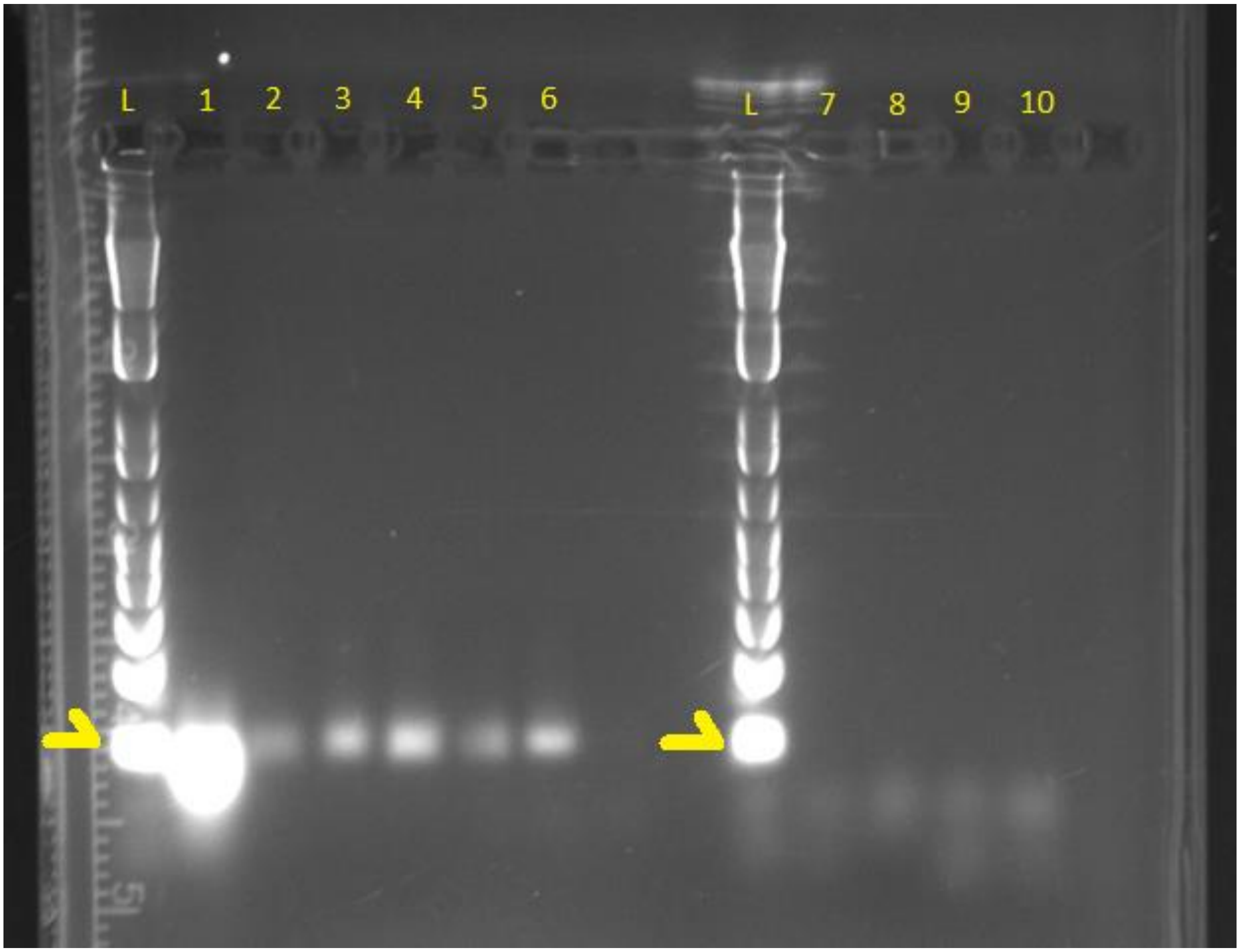


Figure 5. The results of PCR amplification after the second round of PCR. The lanes are as follows, from left to right: L, Invitrogen 1kb DNA Ladder, lane 1, 10X (30 µL) product from 1st PCR reaction, lane 2, 1X product of 1st PCR, then four lanes (lanes 3-6) with barcodes one, two, three, and four (each carrying a different list of genes) after the 2nd PCR reaction. These were followed by two empty lanes, then another lane of DNA Ladder, followed by four negative controls (lanes 7 -10) from the 1st PCR reaction for the four different barcodes. The fuzzy bands in the negative control lanes are primer-dimer. The lack of these bands in the product lanes indicates that PCR went to completion. The half-arrows mark the bottom of the 1 Kb ladder, a size of 100 bp's, from which it can be seen that the ~60 bp oligos increased in length to ~100 bp over the 2 rounds of PCR, which was desired.

sequence. Once the two rounds of PCR were finished, the two sets of oligos in the library had been amplified separately, and in addition, the length of the oligos had been extended, due to the use of primers with additional sequence, such that the oligos had sufficient sequence length at either end to perform Gibson assembly. The results after the second round of PCR are shown in figure 5.

The next step in the cloning scheme was to linearize the LentiCRISPR v2 vector. The restriction enzyme sites for insertion of the 20 bp sgDNA were two BsmBI sites with opposite sense, one immediately after the human U6 promoter, and one directly before the tracrDNA sequence. The BsmBI enzyme cuts outside of its recognition sequence, by 1 nucleotide on the 5'-3' strand and 5 nucleotides on the 3'-5' strand. The LentiCRISPR v2 plasmid was restriction digested by BsmBI from New England Biolabs (NEB) with 10 Units/1 µg DNA in NEBuffer 3.1 at 55˚ C for 2 hours. Figure 6 shows the results of digestion of a similar but smaller plasmid, the pL-CRISPR.EFS.GFP plasmid (AddGene entry #57818), which in addition to the eGFP gene has all of the same elements as the LentiCRISPR v2 plasmid, including the same restriction sites after the human U6 promoter and before the tracrDNA. This band was subsequently excised and purified with Qiagen's Gel Purification kit. After restriction at the two sites, a large (> 1.8 kb) insert is released, leaving a pair of non-complimentary sticky ends. This is useful to reduce the self-ligation background during cloning, but it means that the traditional cloning method of adding a restriction site to the oligos and subsequent digestion, will not allow the vector and oligo to be ligated. Furthermore, the tracrDNA sequence is invariant, and must be transcribed contiguously with the 20 bp guide sequence, which must also be immediately downstream of the human U6 promoter. For these reasons, Gibson assembly was used to clone the guide sequences into the vector plasmid.

Gibson assembly is a procedure developed by Daniel Gibson at the Venter Institute[10]. This method allows seamless integration of DNA's as long as those DNA's have homology of a moderate length, usually 30 to 50 bp's of homology. This method utilizes three enzymes to catalyze three steps.

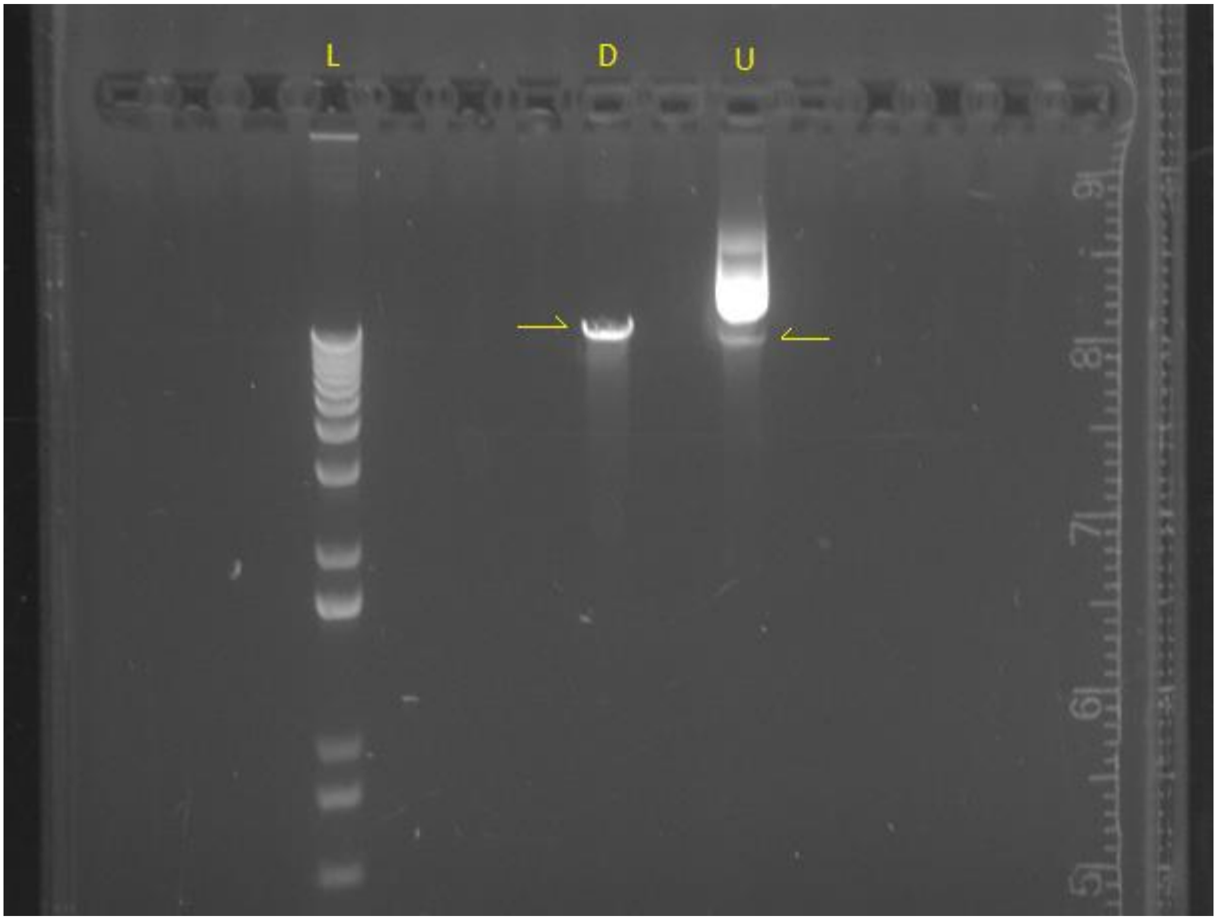


Figure 6. The results of BsmBI digestion of pL-CRISPR.EFS.GFP plasmid. The ladder, L, is the 1 kb ladder from Invitrogen, showing the approximate size of the bands is just >12 kb, which is the expected size (13.5 kb). The two lanes show the digested DNA, D, on the left and the undigested plasmid DNA, U, on the right. The undigested plasmid shows two bands, one for relaxed circular DNA (top band) and one for supercoiled circular DNA (bottom band). The digested DNA fits in directly in-between them, which is the expected position for linear DNA. The half-arrows mark the sizes of the linear and supercoiled DNA for comparison. There is also a faint shadow in the digested lane of undigested, supercoiled DNA. The linear DNA band was subsequently excised from the gel and purified using Qiagen's Gel Purification kit.

The first is an 5'-exonuclease that degrades the 5' end of both DNA's to be integrated. Once the 5' ends of both DNA's to be integrated are chewed back, if the two DNA's have homologous sequences at opposite ends (at the 3' end on one and the 5' end of the other), then the homologous regions will base-pair with one another. The next step in Gibson assembly uses DNA polymerase to fill in the single-stranded gap generated by the exonuclease, between the two base-paired DNA's. Finally, a ligase fixes the nick between the two DNA's, generating a seamlessly integrated DNA molecule. Up to several DNA fragments can be assembled simultaneously in one reaction using the Gibson method.

In this experiment, prior to Gibson assembly, the linearized vector was treated with Calf-Intestinal Phosphatase (CIP) from NEB with 10 Units/1 µg of DNA in NEBuffer 3 at 37˚ C for 1 hour, and the oligo insert was treated with T4 Polynucleotide Kinase from NEB with 10 Units/1 µg of DNA in T4 Ligase Buffer (NEB) at 37˚ C for 30 minutes. Both the vector and the insert were cleaned up with Qiagen's PCR Purification kit after enzyme treatment. Gibson assembly was carried out by using Gibson Assembly Master Mix from NEB, with 15 µL of Master Mix combined with 15 µL of a DNA mixture containing 50 ng of vector and 250 ng of insert, and incubated at 50˚ C for 15 minutes.

Before the use of the plasmid library to produce lentivirus and infect target cells, however, the target cells must be generated and characterized. Four cell lines were generously donated from another lab, designated as RN 2-5, NN, IDH 140, and IDH 170. These cell lines were suspected to recapitulate the AML phenotype. The first step in characterizing these cells was to assess their growth and get an estimate of their doubling time. Figure 7 shows the growth of each of the four cell lines in culture, as well as the growth of normal bone marrow and spleen cells of mice from Jackson lab. Several different culture media were tested. The final culture medium used was based on RPMI with 10% FBS, supplanted with 10 µg/mL of Stem Cell Factor (SCF), 2 µg/mL Interleukin 6 (IL6), 2 µg/mL IL3, 2 mL of a penicillin-streptomycin solution, and 2 mL of L-Glutamine. The cells were counted using a BrightLine Haemocytometer on an inverted brightfield microscope.

The bone marrow cells were acquired by dissecting mice, cleaning the femurs of muscle and connective tissue, making two cross-sectional cuts through the femurs, and then using a syringe to aspirate marrow with 20% Fetal-Bovine Serum (FBS)-containing Phosphate Buffered Saline (PBS). Spleen cells were acquired from the same mice, by dissecting out the spleen, crushing the spleen against a cell-strainer into 20% FBS/PBS, and washing the cells twice between centrifugations in 20% FBS/PBS. All four of the cell lines and the primary (bone marrow and spleen) cells were cultured in the same medium, and split whenever the cell concentration reached >1 million cells/mL. As can be seen from the graph, normal blood-lineage (bone marrow and spleen) cells grew hardly at all over the time course, whereas two of the cell lines, the RN 2-5 and NN cells, grew very aggressively. The IDH 140 cells grew very slowly over the time course, while the IDH 170 cells grew hardly at all or even died out over the time course.

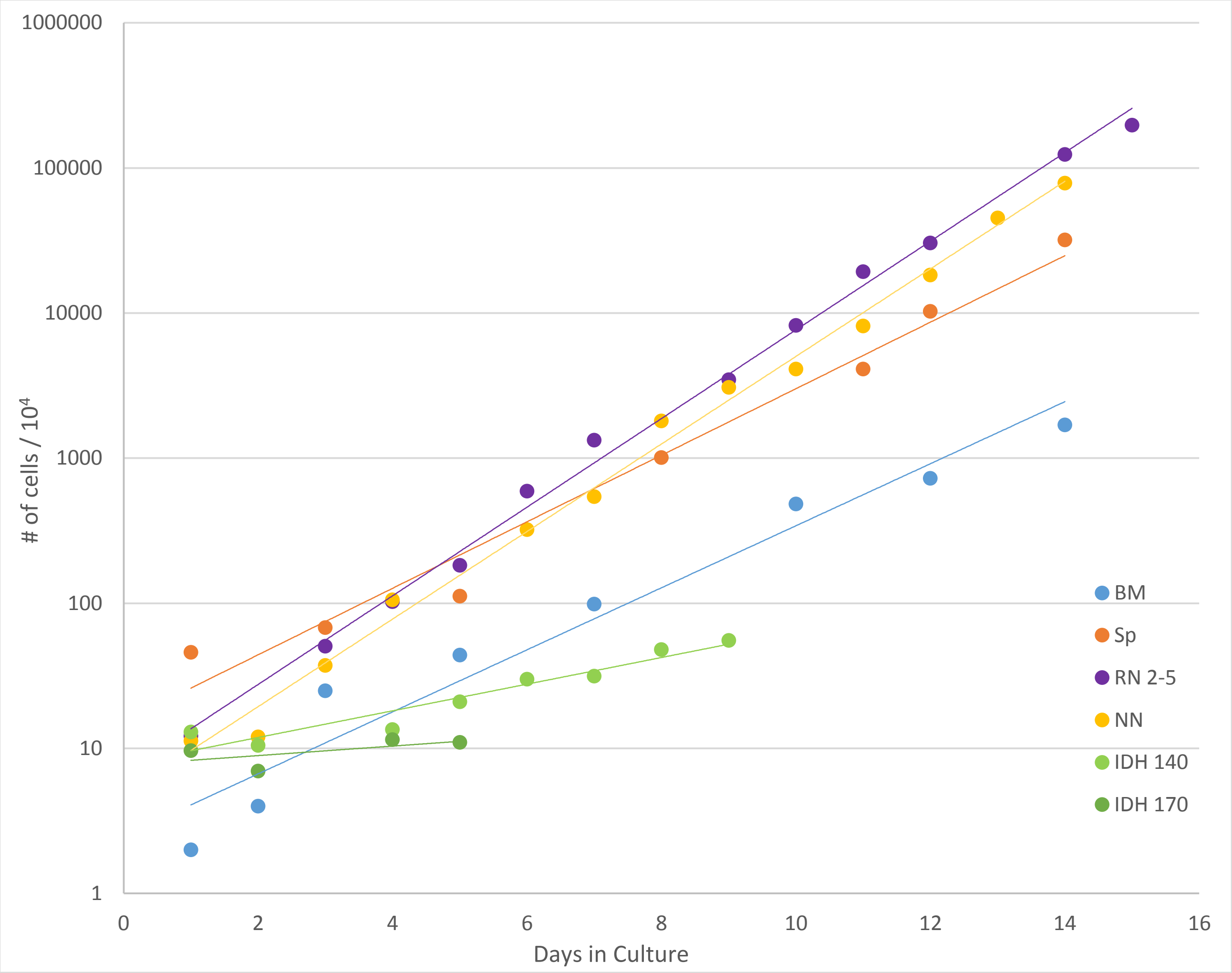


Figure 7. Four cell lines and two primary cell collections were cultured in RPMI/10% FBS/SCF/IL6/IL3/Penn-Strep/L-Gln medium and counted via haemocytometer. The growth curves were fit by an exponential plot – note the logarithmic scale on the y-axis. The y-axis value is the number of cells in culture, normalized by their dilution after splitting. All of the fits had an R^2 above 90%, except for IDH 170 which had an R^2 above 30% indicating that it may not be growing at all. The primary cells from bone marrow (BM) and spleen (Sp) had approximately parallel curves, indicating they had approximately the same rate of growth. Two of the cell lines, the RN 2-5 and NN cell lines, also had parallel curves which were steeper in slope than the primary cells, indicating that they had a higher rate of growth. The IDH 140 had a much slower rate of growth than any of the other cells, according to the low slope of its curve, while the IDH 170 had a nearly flat curve, showing that these cells grew barely or not at all.

In order to generate a mouse model of leukemia for subsequent investigation, a standard procedure for generating transgenic mice was utilized. In this method, recipient mice are lethally irradiated with 900 centiGrays of radiation (1 Gray is 1 Joule of radiant energy/Kilogram of body mass) from a Cesium-137 source. With this radiation dose, the hematopoietic system of the mice is severely compromised, which is thought to be due to the near-complete loss of the HSPC population in their bone marrow. Mice irradiated like this will die in approximately two weeks, the time that it takes for remaining HSPC to terminally differentiate. However, mice injected with HSPC's from a donor will survive, because the irradiated mice will allow these donor cells to engraft into the bone marrow niche previously occupied by their native HSPC's. Injected bone-marrow cells have long been known to be able to 'home' to the bone marrow niche through the blood stream. In this way, transgenic mice can be generated, by transplanting with transgene- or knock-out- containing HSPC.

For this experiment, Jackson lab mice were lethally irradiated at 900 cGy and fed an antibiotic-containing recovery chow. The next day, these mice were retro-orbitally (behind the back of the eyeball) injected with either cells harvested from the bone marrow of other mice, or one of the cell lines that had been characterized previously, along with a negative control that did not receive a transplant. The negative control mice started to weaken at approximately 12 days post-irradiation, and over half of the mice were dead at 15 days post-irradiation. All of the irradiated mice were dead by 18 days post-irradiation. The mice that received injections of normal bone-marrow cells were visibly healthier than the non-transplanted mice at day 12, and most of the mice survived perfectly healthily to day 18 and beyond, although a few of the transplanted mice (~25%) died. The mice transplanted with RN 2-5 or NN cells, in contrast, instead of dying at day 15 post-irradiation, started to weaken, develop patches of missing fur, and lose pigmentation at the tips of their tails, reminiscent of a leukemic phenotype. These mice all die at approximately three weeks post-irradiation.

These mice, when dissected alongside normally-transplanted mice, showed clear morphological signs of highly-progressed leukemic malignancy: they had an enlarged spleen with white, irregular inclusions, a scattering of white inclusions in viscera and other organs, and a hugely enlarged, soft liver showing complete loss of red-brown coloration, turning a completely pale-yellow color. This morphology highly corroborated the leukemic phenotype expected from these cells. The histology of cells harvested from both normally-transplanted and RN 2-5 transplanted mice, collected in the same way as described previously, is shown in Figure 8. The cells were slide-transferred by a Shandon Cytospin centrifuge and stained by May-Grunwald-Giemsa staining, a standard hematological stain that colors nuclei violet, and the cytoplasm dark blue, with the cytoplasm of highly-dividing (malignant) cells staining darker than terminally-differentiated cells. The histology of the cells recaptured from RN 2-5 and normally-transplanted mice exactly matches the histology of the respective cells in culture.

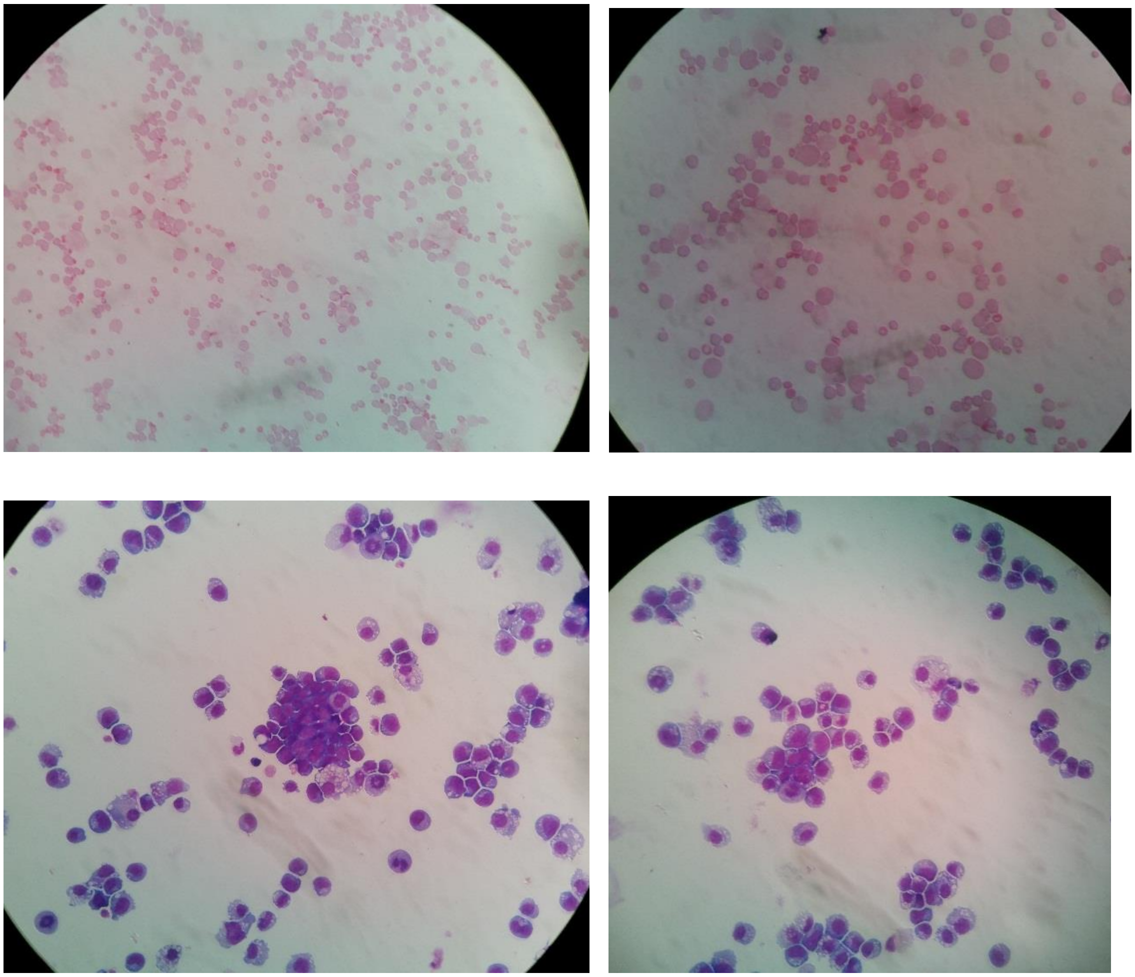


Figure 8. Images taken with an Zeiss AxioPlan upright brightfield microscope of cells harvested from spleen (top) and RN 2-5 cells (bottom) after May-Grunwald-Giemsa staining. The RN 2-5 cells show a characteristic myeloid-leukemic morphology, including deep blue staining of the cytoplasm. The clear vesicles dominating the cytoplasm of some of the RN 2-5 cells is indicative of partial differentiation to the myeloid lineage, and a number of cells can be seen to actively be undergoing mitosis by their condensed chromosomes.

Once the cells had been assayed for growth and the mouse model had been established, the CRISPR-based knockout screen could be conducted. In order to do this, the pool of cloned LentiCRISPR v2 plasmids, now containing a library of sgDNA inserts, were transfected into HEK 293 cells, along with the plasmids required for lentiviral production, in this case the second-generation lentivirus system plasmids pCMV-VSV-G (AddGene #8454) and psPAX2 (AddGene #12260). To transfect the HEK 293 cells, first a mixture of 125 µL of 1:10 Fugene 6 in Opti-MEM was made by placing 112.5 µL Opti-MEM in a Falcon tube and then adding 12.5 µL of Fugene 6 (do not add Fugene 6 reagent directly to the Falcon tube, or any other plastic container without prior dilution) and gentle agitation until thoroughly mixed. Then 5 ng of Pax2 and 3 ng of VSVG were combined with 7.5 ng of LentiCRISPR v2 cloned library, and this DNA mixture was combined with the Fugene 6/Opti-MEM mixture. This solution was incubated at room temperature for 20 minutes, before drop-wise addition to a 10-cm dish previously plated with HEK 293 cells, and which was >70% confluent. The HEK 293 cells were left in their incubator for at least 48 hours in order to produce virus.

The virus was harvested by drawing off the medium from the HEK 293 dish with a 5-mL syringe, and then attaching a 0.45 –micron Supor membrane syringe-filter. The target cells, in this case RN 2-5 cells, were spun down and the medium removed. Then, the virus-containing medium was ejected through the filter into the RN 2-5 cells, which were then resuspended in this virus-media. The cells were replated, and 5 µL of 1000x Polybrene was then added, and the plate was mixed gently. The cells were then spinfected, by centrifugation at 2000 rpm for 105 minutes in a centrifuge pre-warmed to 37° C. The cells were adhered to the bottom of the plate, despite being suspension cells. The media was then changed while the cells were still adhered to the plate. Puromycin was added 24 hours after spinfection to a concentration of 10 µg/mL (previously determined to be effective) in order to select for cells infected with LentiCRISPR v2, which has a Puromycin resistance gene (unlike the pL-CRISPR.EFS.GFP plasmid). The results of puromycin selection after infection are shown in figure 9, which indicates an infection efficiency of ~3%. Considering that 8 million cells were infected, this means that ~240,000 cells were transformed, providing coverage of at least 200-fold for each sgDNA (i.e. on average 200 cells were infected with each different sgDNA-containing plasmid for list 1 sgDNA's).

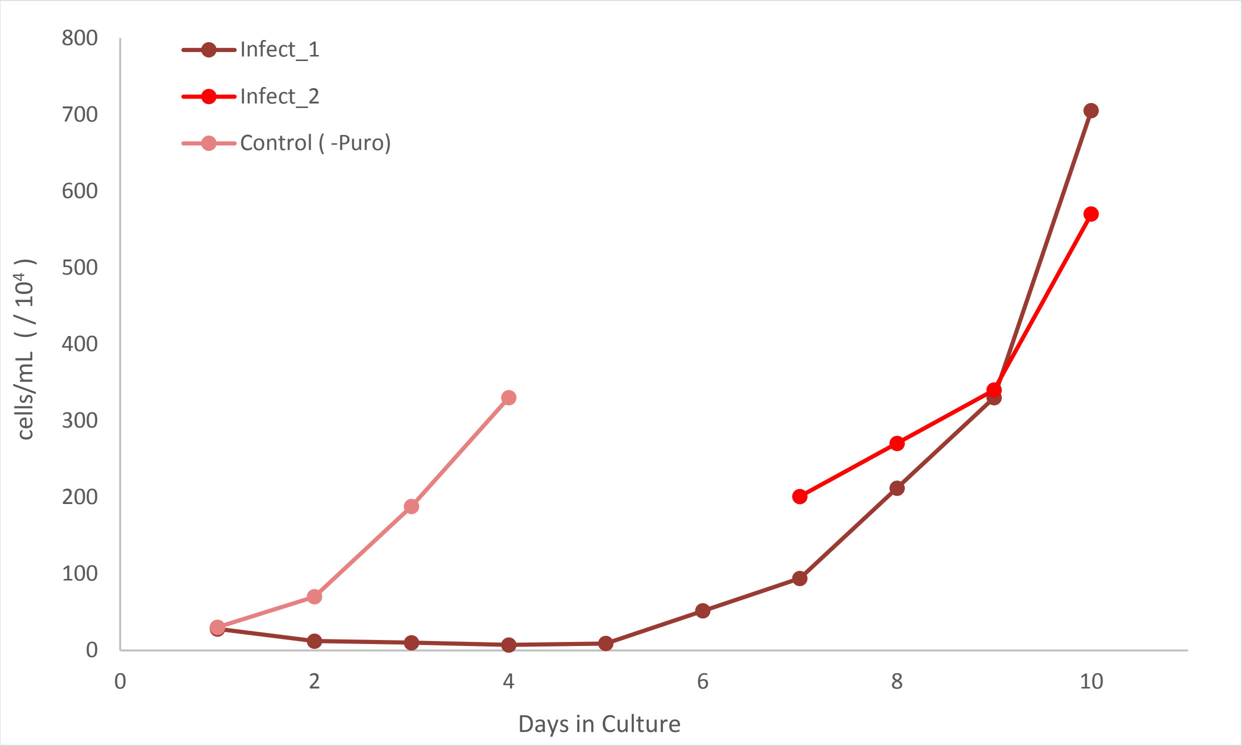


Figure 9. Spinfected RN 2-5 cells were divided into 3 wells, and two of these wells were treated with 10 µg/mL Puromycin. The third well served as a negative control (-Puro). Puromycin selection began right after counting on Day 1, and ended right before counting on Day 3. All counting was done with a BrightLine haemocytometer. Judging from the cell count of puromycin vs control on days 3 and 4, the infection efficiency was either 5% or 2%, from which it was judged that around 3% of cells had been transformed, giving at least 200-fold coverage for list 1 sgDNA's.

The final step in the CRISPR knockout screen is to determine which genes that, when knocked out, resulted in deproliferation of the leukemia cells, and which genes that, when knocked out, increased proliferation of the leukemia cells. This required quantification of the representation of the individual guides in the genomic DNA of the transformed cells. The most advanced method for doing this is to use 'Deep Sequencing' (DeepSeq), a next-generation sequencing method available on Illumina's HiSeq platform. This platform can give many millions (~150 million) reads of a target sequence defined by the experimenter. The technology of HiSeq is described in figure 10. This technology uses a flow-cell with a silicate panel microfabricated to have short DNA's of a defined sequence on its surface. These short DNA's are complimentary to an adaptor sequence, so that any DNA with that adapter sequence at the end, provided by the experimenter, will base-pair with them. The next step goes through a round of DNA synthesis using DNA Polymerase, so that the short DNA's attached to the flowcell now have the complement of the sequences that were attached to the adapters, which includes the sequence of interest and another adapter complimentary to a different set of DNA's attached to the flowcell. The DNA's with the target sequence provided by the experimenter are now flushed out of the flowcell. This now allows for cluster growth through further rounds of PCR, in which the different adapter-bound sequences amplify each other, leading to the formation of a 'PCR colony' (Polony). Once many millions of polonies have been seeded, the sequencing step can now begin.

The sequencing is based on a variant of Sanger-sequencing (sequencing-by-synthesis) in which the amplicon is built one nucleotide at a time. This is achieved by providing to the flowcell all of the components necessary for PCR, including a primer chosen by the experimenter that defines the position to be sequenced, just as in normal sequencing, and dNTP's labelled by a fluorescent moiety that also acts as a chain terminator. They key feature of this chemical moiety is that, in addition to being fluorescent and acting as a chain terminator, it is also a reversible chain terminator – that is, it can be removed. So the machine provides the components of Sanger-sequencing to the flowcell, the DNAP goes through one round of extension across the whole flowcell, and then it stalls. Each polony is now labelled with one kind of fluorescent moiety corresponding to the nucleotide at that position of its amplicon. The machine then flushes out the PCR components, and goes through a cycle where a series of advanced optics, including a microscope capable of detecting every polony, uses a laser to fluorescently visualize each of the positions on the flowcell. Ideally, every polony is detected separately by the optics. In reality, many of the polonys are overlapping or are just barely separated. A software component of CASAVA, the program package used to run the machine and do post-processing, is able to determine, based on the first four nucleotides of sequence, each individual polony during cluster identification. After the visualization cycle, the next cycle provided to the flowcell by the machine cleaves off the fluorescent chain terminator. Then, the cycles repeat themselves, with each round

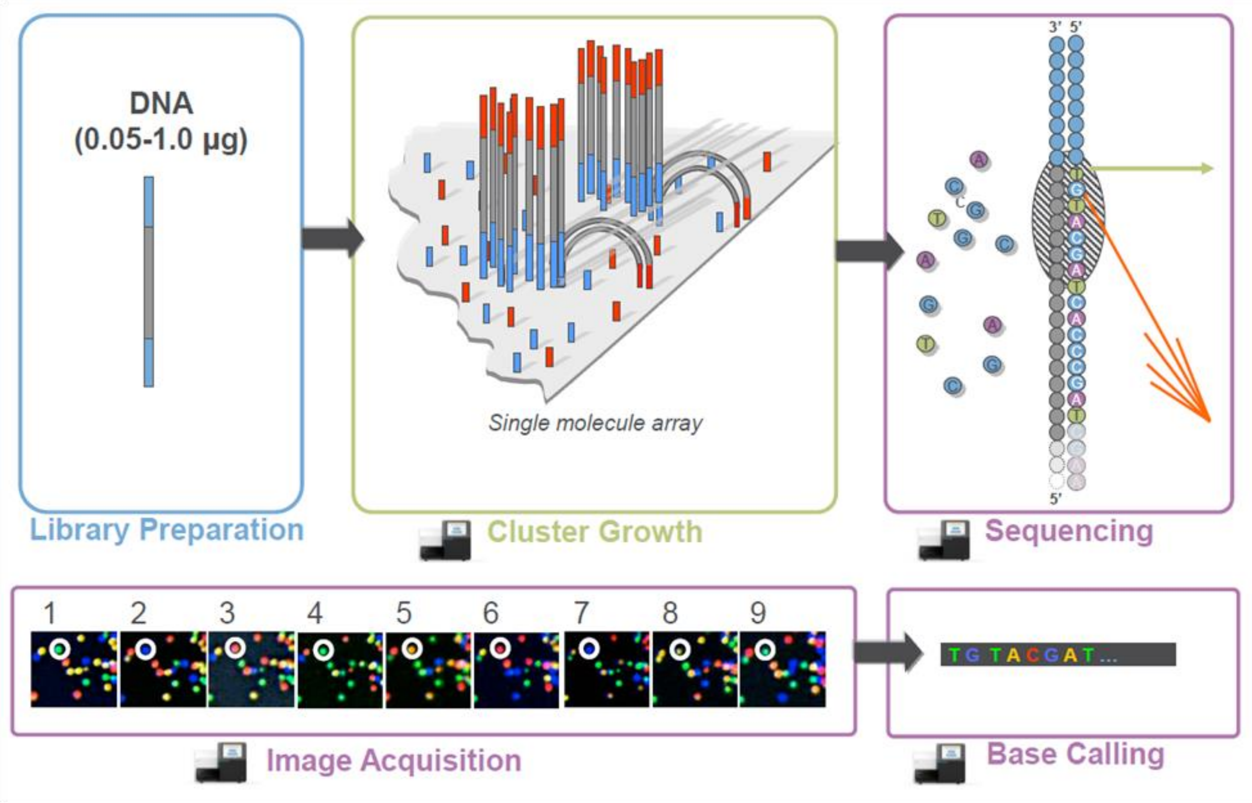


Figure 10. The description by Illumina of their MiSeq platform, a benchtop version of the same technology used by their HiSeq platform. This shows the steps of library preparation, which attaches adapter sequences to both ends of the sequence of interest (although misleading from the image, there should be two different adapters at either end), cluster growth resulting in polonies, and the actual sequencing step using fluorescent dNTP's. Image acquisition takes place between each cycle of single-nucleotide extension. From these images, a series of base calls can be made, from which the sequences of what are ultimately many individual DNA molecules can be determined. This image can be found on Illumina's website or at MIT's BioMicroCenter webpage http://openwetware.org/wiki/BioMicroCenter:Sequencing

sequencing one more nucleotide position. This data is eventually used by the CASAVA package to build the sequences of each of the polonies, resulting in a series of reads in fastq format. Each of these reads corresponds to a single molecule of DNA, and so then the number of reads can be compared, in a DeepSeq experiment, to determine the relative abundance of DNA sequences.

For this experiment genomic DNA was collected from infected cells every other day of culture using Qiagen's Blood and Tissue extraction kit. Sample preparation was done by two steps of PCR, one which amplified the region between the LTR (the site of genomic integration from lentiviral vectors), and a second that was done using primers that also contained the Illumina adapter sequences and a barcode. These primers amplified the region from just upstream of the end of the human U6 promoter to a sequence after the tracrDNA. The forward primer in this PCR reaction contained one of 15 different barcodes provided by Illumina[11] to label the samples collected at each timepoint, so that all of the samples could be sequenced together in one HiSeq run. These barcodes, which lay right after the forward Illumina adapter, also provided the heterogeneity required at the first four nucleotides for robust cluster identification. DNA collected and amplified by LZ from an infection done in RN 2-5 and normal HSPC was sent for sequencing to Eurofins Genomics, a company specializing in next-generation sequencing. The sequencing primer used was the Illumina Universal sequencing primer, which is complimentary to the forward Illumina adapter, so that the sequencing started just after the adapter at the site of the barcodes.

From Eurofins, over 150 million reads were collected across 15 samples. Approximately 140 million of these reads passed filter. These reads came in fastq format, meaning that they were text files with four lines per read; a descriptor line, a line with the actual sequence, a register line, and a line with the quality-score for each position in the read sequence. In order to demultiplex the data into individual samples, and then determine the number of reads for each sgDNA, custom computer code was developed and executed on CSHL's High-Performance Compute Cluster (HPCC). This cluster consists of 100 compute nodes, 2 development (head-) nodes, and 2 high-memory nodes. Each of the nodes has 32 2.7 GHz processors from IBM (with hyperthreading) and 128 Gigabytes of RAM, except for the high-memory nodes, which each have 1.5 Terrabytes of RAM. All computer code was written in the UNIX-based Red-Hat Enterprise Linux (RHEL) environment. For the first step of data processing, the reads were extracted from their fastq files (the data came in 38 fastq files with 16 million lines each) and demultiplexed into individual samples. For instance, for the first barcode, consisting of the sequence ATCACGAC, the following was executed:

```
grep -Ew ^ATCACGAC\|^..CACGAC\|^.T.ACGAC\|^.TC.CGAC\|^.TCA.GAC\|^.TCAC
.AC\|^.TCACG.C\|^.TCACGA.\|^A..ACGAC\|^A.C.CGAC\|^A.CA.GAC\|^A.CAC.AC\
|^A.CACG.C\|^A.CACGA.\|^AT..CGAC\|^AT.A.GAC\|^AT.AC.AC\|^AT.ACG.C\|^AT
.ACGA.\|^ATC..GAC\|^ATC.C.AC\|^ATC.CG.C\|^ATC.CGA.\|^ATCA..AC\|^ATCA.G
.C\|^ATCA.GA.\|^ATCAC..C\|^ATCAC.A.\|^ATCACG.. *.fastq >> sample1.txt
```

which searched through all fastq files for sequences starting with the desired barcode, allowing for two mismatches in the sequence, and put these lines into the sample1 text file. For each of the samples' sequence files, the number of reads for each sgDNA sequences were counted by using a C-shell script containing the following line:

```
agrep -1 –c NNNNNNNNNNNNNNNNNNN sample#.txt | grep –v tot >> out#.txt
```

for each sgDNA sequence in the library, where the series of N's represents the unique 19 bp guide sequence, and 'agrep' is an approximate search program developed at UCSF.[12] The result of sequence demultiplexing and counting is shown in figure 11. This figure shows the distribution of sgDNA's recovered at the end of culture (Day 14) minus the sgDNA's recovered at the start of culture (Day 2) from RN 2-5 cells. This distribution is, in its major features, similar to other DeepSeq data recovered from other genetic screens. The key outcome of this screen is the knowledge of which metabolic genes, when knocked out of the AML cells, have the strongest negative influence on cell proliferation. In addition to this, although most genes can be seen to inhibit cell growth when knocked out, certain genes can be seen to give a selective advantage to cell growth when knocked out.

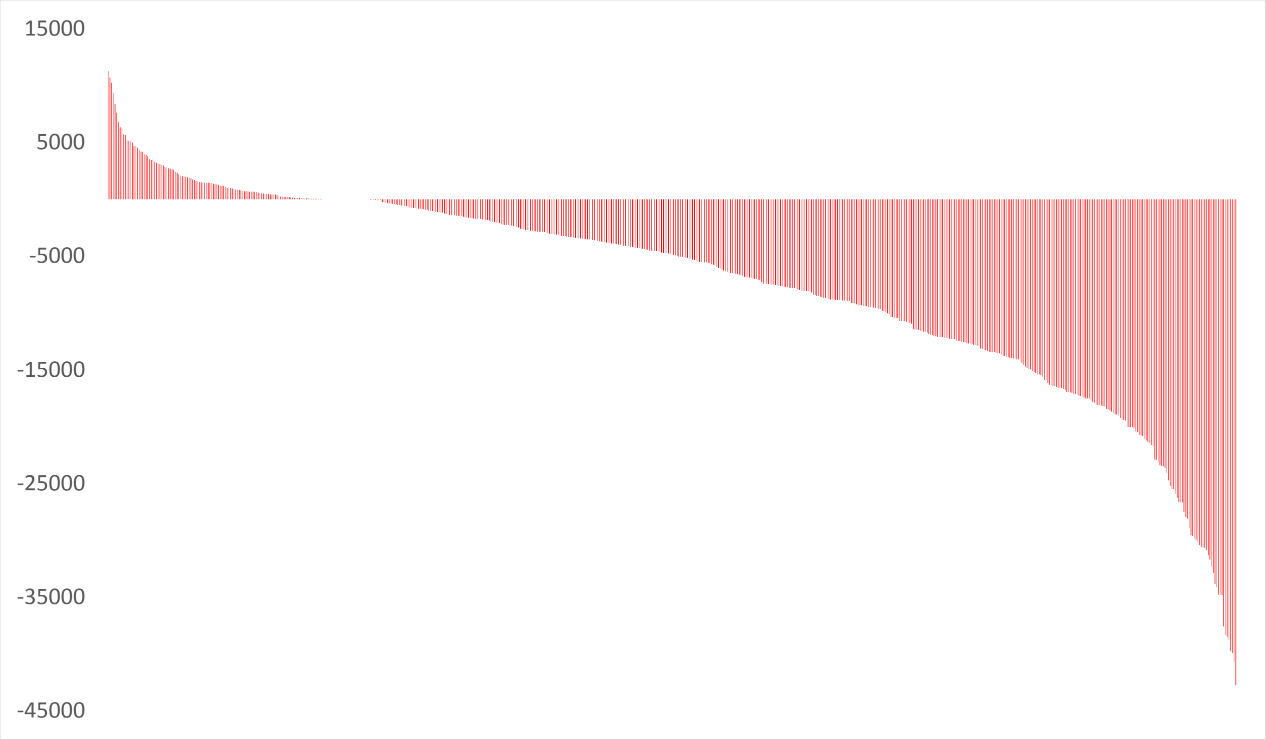


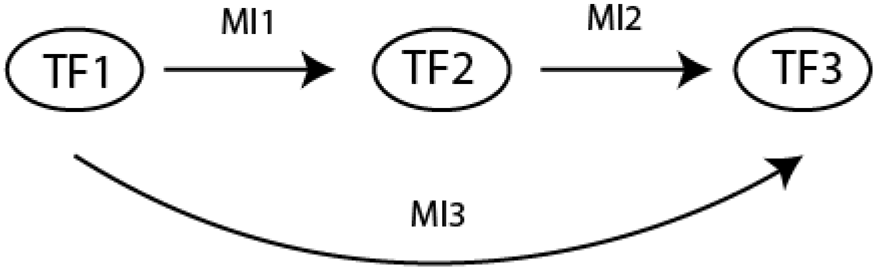
Figure 11. The results of a CRISPR-based knockout screen in RN 2-5 cells. The y-axis (unitless) represents the number of reads, normalized as the reads found at day 14 of culture minus the reads found at day 2 of culture. The x-axis is a bin for each of the sgDNA sequences (thus this is a histogram). As can be seen from the distribution of the normalized reads, there were many genes for which knockout hampered the ability of these cells to proliferate, which might be expected considering that all of the genes in this list were metabolic genes. There were, however, a number of genes that, when knocked out, conferred a modest proliferative advantage to these cells.

Table 2. Characteristics of distribution from DeepSeq-based sgDNA recovery and analysis

Total sgDNA's	1538	
Total number of Reads	150,734,483	100%
Reads matched to Barcodes	146,742,708	97.35%
Barcoded Reads matched to sgDNA's	89,612,649	59.45%

It would be beneficial, while investigating the function of metabolic genes, to know what the transcriptional control of the metabolic genes being studied is. For instance, if there is a master regulator of the genes being interrogated, and it is the case that these genes are necessary for the AML phenotype, then intervention at the level of the master regulator, either using a CRISPR-Cas9 knockout or with a pharmaceutical intervention, would likely be able to affect the progression of the AML cells. If the set of genes controlled by such a regulator are more necessary for the survival of the AML cells than the normal HSPC cells, then such an intervention would likely be targeting a therapeutic window. Deducing the activity of such a master regulator, however, requires a large amount of data, specifically data about the expression of many metabolic genes, as well as the expression of known or putative transcription factors.

While the transfection-infection was being conducted, and while waiting for sequencing results from Eurofins, a bioinformatics approach based on the ARACNE algorithm[13] was taken in order to attempt to find master regulators of the metabolic program in AML cells. The ARACNE algorithm calculates the mutual information (MI), which is an information-theoretical measure of the co-occurrence of two variables, in this case the co-expression of two genes (a high MI indicates a high correlation of co-expression), and then applies a further information-theoretic principle, the data-processing inequality,[14] which states that any MI between two variables that are related by virtue of an intervening third variable, cannot have a greater value than the MI between the co-related variables and the intervening variable. Another way of putting this is that, for the scheme below, which could be for instance a transcriptional program where TF1 causes the expression of TF2, which then causes the expression of TF3;



The variables TF1, TF2 and TF3 all have MI's between each other, but the indirect MI between TF1 and TF3 (MI3) is due to the direct interaction of TF1 with TF2 (MI1) and then TF2 and TF3 (MI2). This type of situation poses a ubiquitous problem in bioinformatic analysis, because from the co-expression data alone, it is impossible to reconstruct the pathway shown above. However, according to the principle of the data-processing inequality, the value MI3 must be no greater than the value of either MI1 or MI2 (in other words MI3 $\leq$ {MI1,MI2}).

Therefore, for linear systems of transcription factor activity, and even whole transcriptional programs without significant loops or redundancies in their network topology, the order of transcription factor activity can be reconstructed. ARACNE's key feature is, after calculation of all possible MI's, application of the data-processing inequality principle to eliminate the least values from every possible triple containing a given gene. This gives the putative order of transcriptional priority for that gene. The ARACNE2 program is given a matrix containing a set of expression values for each gene, calculates the MI between each gene, giving a symmetric matrix as the output (the adjacency matrix), and then eliminates least-triples. The ARACNE2 program allows for several levels of parameterization. The first is a threshold for the MI allowable in the calculation. Genes with very low MI with the given gene to reconstruct are then removed from subsequent calculation. Another parameter is the tolerance, which is a value that gives space for the data-processing inequality. This allows the inclusion of MI's that are the least in a triple by only a small margin (lower than the tolerance), which allows for biases and noise in the data. ARACNE2 can also be given a list of genes for which to reconstruct putative transcription factors from the larger matrix of all genes.

For the ARACNE2 calculation, RNA-Seq data from The Cancer Genome Atlas (TCGA) was used, which contains a data set from 200 AML patients.[15] This data set has data for array-based mRNA expression, miRNA expression, DNA methylation, copy-number variation, and RNA-Seq from AML patient samples. The RNA-Seq data contained aligned whole-genome measurements in RPKM (Reads Per Kilobase per Million mapped reads) count. The ARACNE2 program was given a list of genes annotated as metabolic genes in the KEGG database for which to reconstruct putative transcription factors. A bootstrapping method, which uses permutation of patient samples to increase statistical confidence at the expense of greatly increased computation time, was employed. Two different parameterizations, one with a relatively high MI threshold (MI corresponding to p = 1x10^{-20}) and a tolerance of 0.1, and one with a lower MI threshold (MI corresponding to p = 1x10^{-7}) and a tolerance of 0.05, were used. Both of these calculations were run on the HPCC at CSHL. Figure 12 shows the results

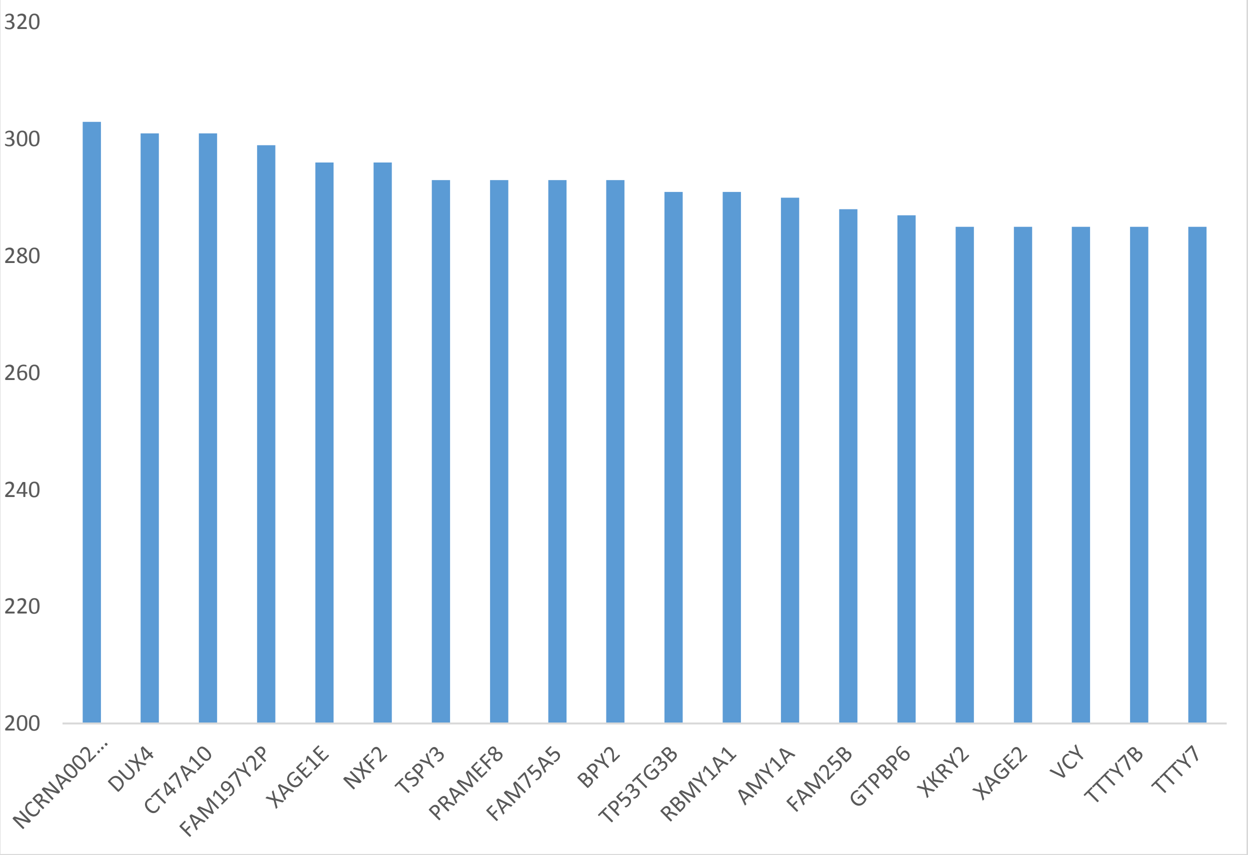


Figure 12. The results of an ARACNE calculation run with p=10^{-20} and tolerance=0.1 on RNA-Seq data from TCGA. The genetic loci identified as regulators are labelled on the x-axis, and the number of genes annotated in the KEGG as metabolic genes, for which that loci was identified as a regulator for, is shown on the y-axis. This is important because a master-regulator of the metabolic network would be expected to bind to a large subset of the genes in the network. The second highest hit in the putative regulators is DUX4, a known transcription factor.

of the p=1x10^{-20} calculation. This shows the gene, on the x-axis, and the score for that gene , which is the number of metabolic genes for which that gene was identified as controlling, on the y-axis. Only the 20 top-scoring loci are shown.

In addition to the TCGA data set on AML, there are also a number of other large microarray-based data sets taken from AML patients. One of these data sets is stored in the Gene Expression Omnibus (GEO) as GEO entry GSE1159. This data set, based on the Affymetrix HG-U133A gene expression microarray, was taken from 285 AML patients and 8 normal controls.[16] The ARACNE algorithm was also employed to reconstruct the same network of metabolic genes in GSE1159 as in the RNA-Seq data set from TCGA. A number of different parameterizations were also run on the GSE1159 data. Figure 13 shows the results of the same parameterizations and bootstrapping method used on

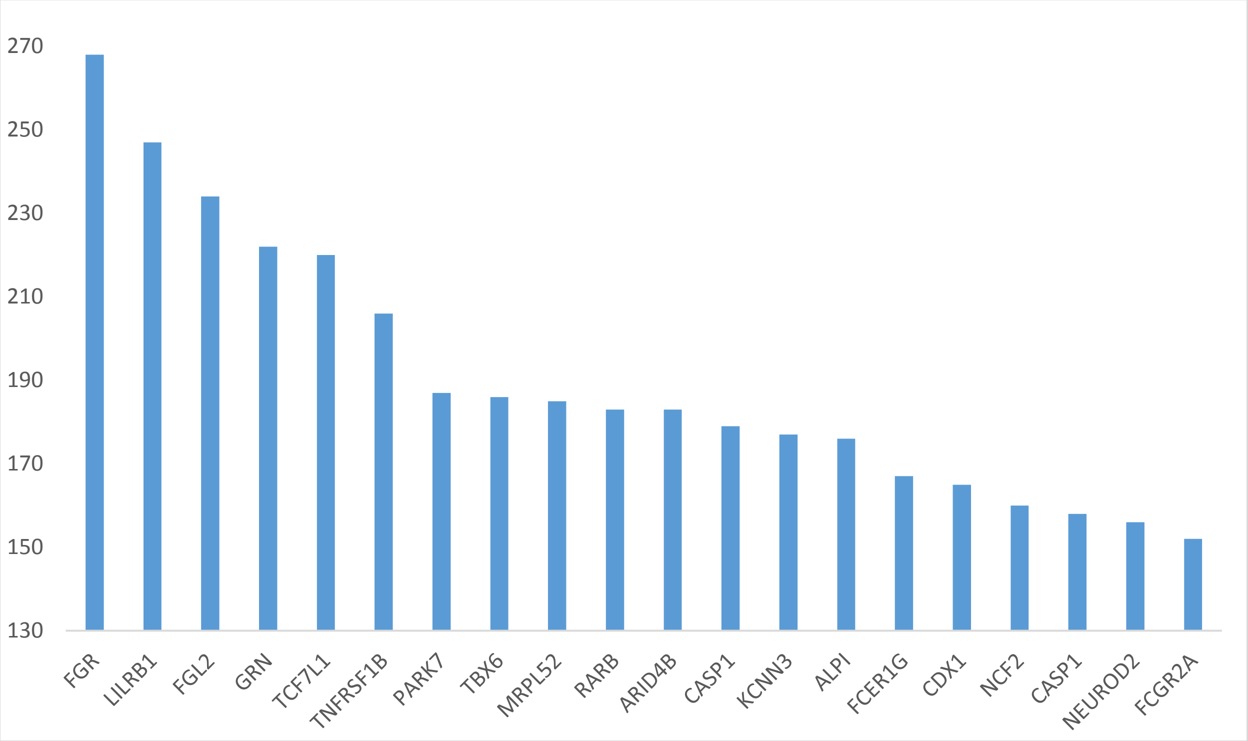


Figure 13. The results of supervised ARACNE analysis, run with p=10^{-20} and tolerance=0.1, on whole-genome microarray expression data in GSE1159. 6 out of the top-scoring 20 hits were known transcription factors; TCF7L1, TBX6, RARB, ARID4B, CDX1, and NEUROD2. Compared with the results in figure 12, the supervised analysis also gave better separation of hits – the very top, 1st scorer is separated from the 20th scorer by almost 2-fold (100%), whereas the separation in the previous reconstruction was only 6%. This may indicate that the supervised analysis is much better at scoring the hits.

GSE1159 as for the result in figure 12 on TCGA data. However, unlike in the previous analysis, an additional functionality of ARACNE was utilized, which allows ARACNE to be given a list of known transcription factors, which are then given preference in the data-processing inequality step, independent of the tolerance.

When removing the least MI in a triple, if one of the MI's represents the interaction of the gene with a locus in the list of transcription factors given to ARACNE, the ARACNE program will not remove that locus unless the other MI in the triple is also with a transcription factor. This encapsulates the logic that, if one of the genes in the triple is a transcription factor and the other is not, even if the MI for the transcription factor is lower, it is still impossible for the interaction with the non-transcription factor (for instance, another metabolic enzyme) to be more direct than with the transcription factor. As can be seen from figure 13, the network reconstructed from the supervised analysis (giving ARACNE a list of known transcription factors) in GSE1159 is much different than the unsupervised analysis in TCGA RNA-Seq data. The results from GSE1159 are generally better than the previous results, given that 6 of the top-scoring 20 hits in this network reconstruction are known transcription factors. It would be interesting, in the future, to run a supervised ARACNE reconstruction on the TCGA RNA-Seq data.

The ARACNE reconstruction wasn't the only analysis run on the GSE1159 data set, though. A statistical approach recently described in a Nature Biotechnology paper[17] was also employed on the GSE1159 data. In this paper's approach, first the root-mean squared deviation (RMSD) between the expression of all genes annotated in the KEGG as metabolic genes was calculated by

$$RMSD = \sqrt{\sum_{i=1}^{n} \frac{(\, Mean[\log_2 x_i] - Mean[\log_2 y_i]\,)^2}{n}}$$

where x_i is the expression of the i^{th} gene in the tumor array, y_i is the expression of the i^{th} gene in the control array, n is the total number of genes annotated as metabolic genes, and the operator *Mean[]* is understood to be the calculation of the arithmetic mean of the bracketed values. From a quick look at this equation, it is obvious that if x and y have no differences in expression, the RMSD between them will be 0, and if there are differences in their expression, the RMSD will be greater than 0, with how much greater than 0 depending on how great their difference in expression is. The same method of calculating the RMSD was employed using the expression arrays in GSE1159, except for the minor further consideration that each of the probe-sets for each of the genes were used as the i^{th} values (and thus n was the total number of probe-sets). Using this method of calculation, the RMSD between the AML patients and CD34+ controls was found to be 0.710, the RMSD between the AML patients and NBM controls was found to be 0.699, and the RMSD between the CD34+ and NBM controls was 1.030, implying that the metabolic regime of the cancer cells is somewhere between the metabolism of the CD34+ stem cells and the NBM differentiated cells.

In the original paper that GSE1159 came from, Valk *et al* 2004, common genetic and cytogenetic abnormalities in AML were also assayed for each of the 285 AML patients in the study, and these are annotated in each expression array. In order to find out if there were differences between the different subtypes of AML, each of the different subtypes were manually separated, and then the RMSD was calculated between each of them. In addition to this, the 8 control samples were also from two different cellular sources, Normal Bone Marrow aspirates (NBM) from whole bone marrow, and purified CD34+ cells which are more stem-like in nature, so in order to look at the metabolic gene expression differences between these two cellular subgroups, and between these subgroups and the different subtypes of AML, they were also separated and had RMSD's calculated separately. The calculated RMSD's between all the different genetic and cytogenetic subtypes of AML assayed and the control cell types is shown below in Table 3. As can be seen, the greatest difference in gene expression is between the NBM and CD34 categories, while the RMSD between the different subtypes of AML are significantly lower, implying that the metabolism of cancer cells are all somehow more similar to each other than to normal cell types. This is a result similar to what was found by Hu *et al* in the Nature analysis paper from which the RMSD calculation was taken. In addition, the single lowest RMSD on the table is between the two different types of FLT3 mutation, FLT3 ITD and FLT3 TKD, which is perhaps expected because they affect the same gene.

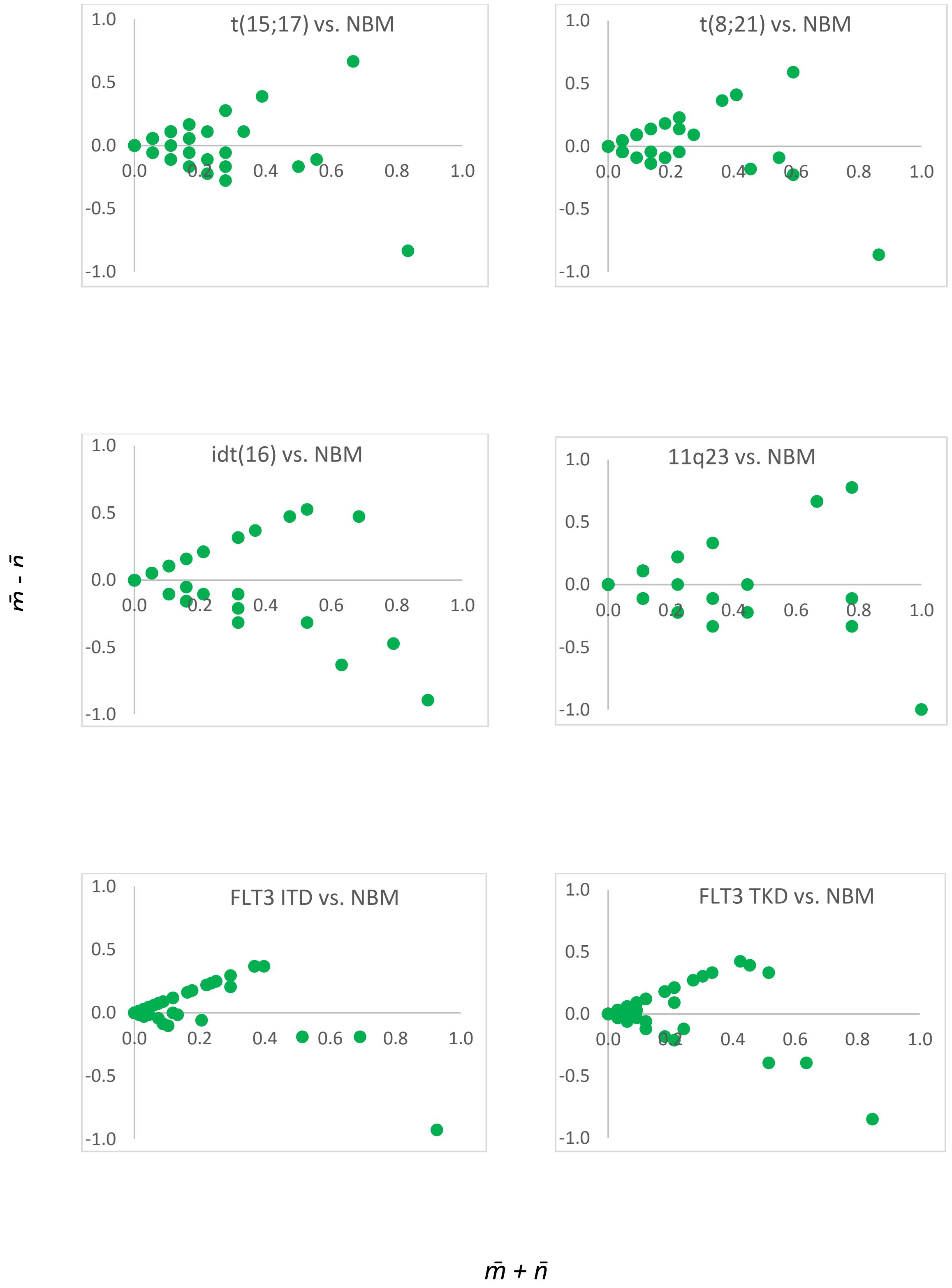
Table 3. The RMSD between different subtypes of AML and control cell types

	t(5;17)	t(8;21)	idt(16)	11q23	FLT3 ITD	FLT3 TKD	N-RAS	K-RAS	EVI1	CEBPA	CD34	NBM
t(5;17)												
t(8;21)	0.683											
idt(16)	0.759	0.611										
11q23	0.690	0.651	0.627									
FLT3 ITD	0.543	0.506	0.529	0.454								
FLT3 TKD	0.539	0.495	0.405	0.420	0.248							
N-RAS	0.628	0.456	0.331	0.496	0.357	0.274						
K-RAS	0.731	0.628	0.495	0.565	0.468	0.397	0.417					
EVI1	0.667	0.573	0.622	0.566	0.419	0.448	0.429	0.593				
CEBPA	0.672	0.538	0.618	0.542	0.433	0.443	0.466	0.632	0.506			
CD34	0.949	0.799	0.913	0.869	0.711	0.764	0.776	0.892	0.669	0.772		
NBM	0.856	0.892	0.827	0.821	0.798	0.748	0.693	0.791	0.742	0.830	1.030	

Hu *et al* also used devised a method employing a statistical test called the Wilcoxon signed-rank test, which a nonparametric test of the medians, not to be confused with the Wilcoxon rank-sum test, (sometimes also called the Mann-Whitney U-test) which is also nonparametric and uses ranking. The

Wilcoxon signed-rank test is a paired test, meaning that it takes as its test sample a set of differences between two related variables. However, according to the test that Hu *et al* performed, the paired differences are between the expression of a gene in the tumor sample and the expression of the same gene in the control sample. Furthermore, they broke up the list of all genes annotated as metabolic genes into the individual pathways that those genes are part of according to the KEGG, so that each pathway was its own sample population. Then a statistical test was performed on each set of genes represented by a pathway, for each patient in the data set. After performing these statistical tests, the results of the right-sided rejection of the null hypothesis (implying overexpression of the pathway in tumor tissue versus control) and the left-sided rejection of the null hypothesis (implying underexpression of the pathway in tumor tissue versus control) are tallied. Then the values $\bar{m}$ and $\bar{n}$ are calculated, where $\bar{m}$ is the average number of tumor samples with a given pathway underexpressed, and $\bar{n}$ is the average number of tumor samples with the given pathway overexpressed. This allows for estimates of how different the tumor samples are from normal ($\bar{m} + \bar{n}$) and whether that difference represents net overexpression, underexpression, or neither ($\bar{m} - \bar{n}$).

This method was applied to the data for metabolic gene expression in GSE1159 using custom code in the R programming environment on CSHL's HPCC. To perform the Wilcoxon signed-rank tests, the R package *exactRankTests* was used, and the results of each test were Bonferroni corrected for multiple hypothesis testing. The results of running and tallying the tests for each pathway in each of the 285 AML patients in GSE1159 are shown in figure 14. Each of the dots represents the $\bar{m} + \bar{n}$ and $\bar{m} - \bar{n}$ values for a single pathway annotated in the KEGG ontology for metabolic pathways. Patients were manually broken up according to subcategory, and the same thing was done for each genetic and cytogenetic subcategory of AML versus both the CD34+ and NBM controls. This allowed for identification of whether certain subtypes of AML relied on certain pathways more than other subtypes. Only the calculations between the NBM and AML subtypes are shown.


t(15;17) vs. NBM
t(8;21) vs. NBM
idt(16) vs. NBM
11q23 vs. NBM
FLT3 ITD vs. NBM
FLT3 TKD vs. NBM
1.0
0.5
0.0
-0.5
-1.0
0.0
0.2
0.4
0.6
0.8
1.0
m̄ - n̄
m̄ + n̄

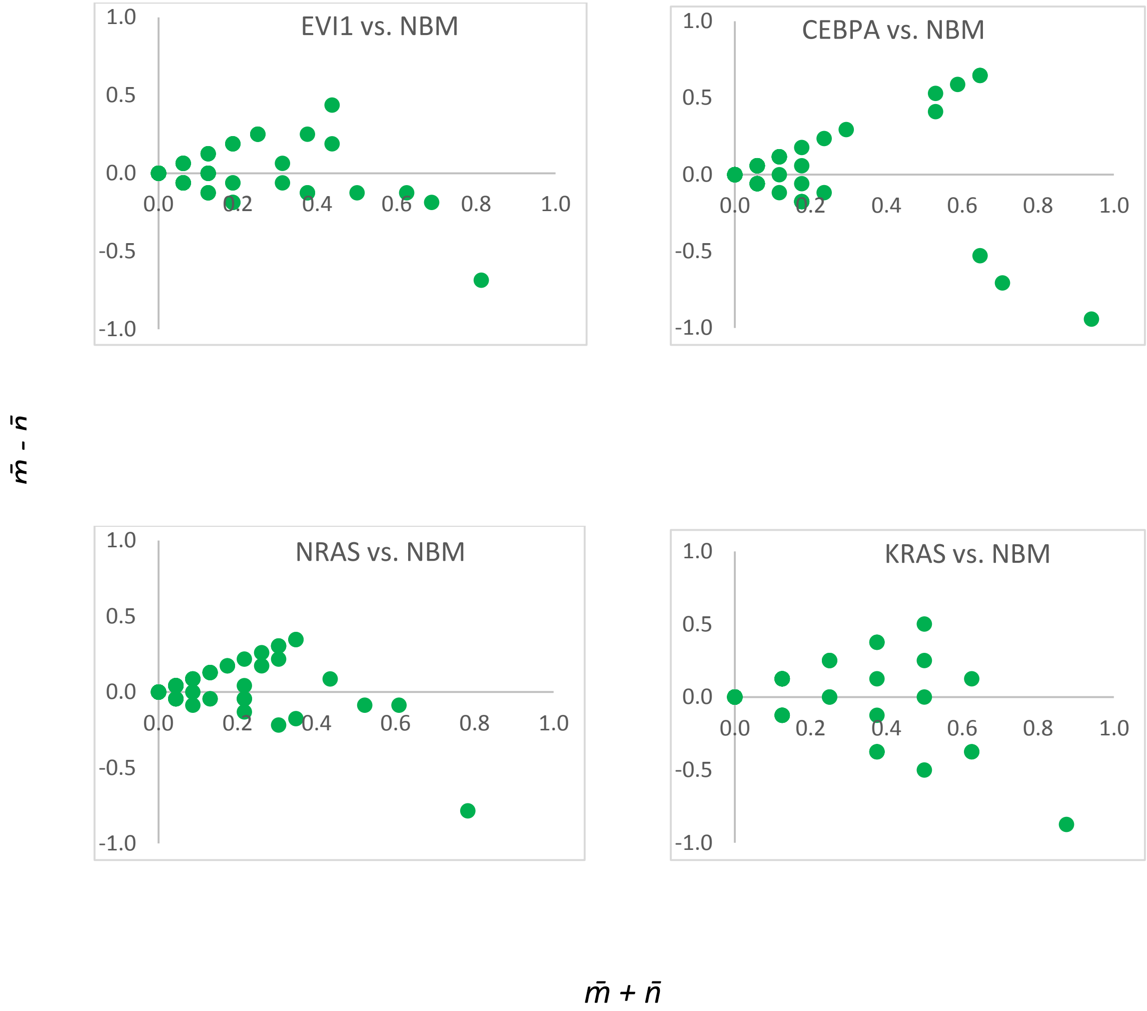


Figure 14. The results of Wilcoxon rank-sum based pathway expression analysis on different subtypes of AML versus NBM control. Each dot represents a particular metabolic pathway. The x-axis is the m̄ + n̄ value and the y-axis is the m̄ - n̄ value for each particular pathway. The distribution of the dots can be seen to be different between the different subtypes of AML. Only the values as calculated between the NBM control and the subtypes is shown. In general, there is a moderate difference in expression between the tumor and NBM control, although some subtypes can be seen to have more of a difference than others.

Discussion

In the work done for this thesis, a genetic screen using CRISPR-Cas9 technology was conducted, in which a set of genes never before screened for in AML cells were interrogated. There are, to date, only a handful of CRISPR-Cas9 genetic screens in mammalian cells published in the primary literature, and most of these focus on genome-wide studies, trying to recapture genes known to be absolutely essential and thus validate the CRISPR-Cas9 based screening method. This genetic screen employs a different, much smaller scope, with a much more directed aim – to find metabolic genes that are necessary for progression in AML-type malignancy. This is the first screen of its kind, to the best of our knowledge. We successfully designed, executed, and analyzed this screen following the work of the Feng Zhang laboratory in whole-genome screens. We also were able to characterize a number of new cell lines, and generate a mouse model of advanced AML in order to support further work. The most aggressive cell line, the RN 2-5 cells, were chosen to conduct the CRISPR-Cas9 knockout screen.

In recent years, there has been a renewed interest in metabolic genes as drug targets in chemotherapy. The metabolism of cancer cells has long been known to be vastly different from normal cells, providing a therapeutic window, but therapeutic interventions have lagged behind recognition of this difference. Recent developments in drug design have allowed targeting of specific metabolic pathways, such as the ones with a therapeutic window in cancer treatment, and renewed interest in cancer metabolism has paved the way for genetic screening of metabolic genes. This work has already identified a number of genes that halt proliferation of the highly aggressive RN 2-5 cells when knocked out, and so serve as potential drug targets for AML treatment. Even more interestingly, perhaps, is the finding of a smaller number of metabolic genes that, when knocked out, confer a proliferative advantage in these cells. These genes may be important for the progression of the AML phenotype.

The next phase of the research, already underway, is to follow up on some of the hits from the genetic screen. This can be done by a targeted CRISPR-Cas9 knockout of a single gene identified by the screen, and monitoring the phenotype of the knockout, both in culture and in the mouse model of malignancy. The biochemical effect of the knockout can be determined using mass spectrometry to quantify the change in metabolic flux through pathways known to be associated with the gene target. CSHL's Mass Spectrometry and Proteomics core facility is also outfitted to perform metabolomics analysis, and samples generated by our lab have been sent to the core facility for analysis. In this way, the underlying mechanism of knockout-mediated deproliferation (or enhanced proliferation) can be interrogated.

We also employed a machine-learning program, the ARACNE algorithm, to identify potential master-regulator targets in therapeutic interventions for AML. The results shown above for the TCGA RNA-Seq data set, along with results from calculations performed on other data sets, have shown promise in finding known transcription factors. This bioinformatics program may be able to find transcription factors important for the AML phenotype but not important for normal HSPC cells, and thus open a new therapeutic window, or even to find new genes with transcription factor activity. A statistical approach was also used in order to characterize the metabolic gene expression program of AML cells versus the gene expression program of normal cells. In the future, a combination of bioinformatics and genetic screening will help to identify the underlying themes of AML etiology.

Bibliography

[1] Mikkola, H.K. and S.H. Orkin, *The journey of developing hematopoietic stem cells.* Development, 2006. **133**(19): p. 3733-44.

[2] WHO Global Database on Anaemia, *Worldwide prevalence of anaemia 1993 – 2005.* WHO

[2] Orkin, S.H. and L.I. Zon, *Hematopoiesis: an evolving paradigm for stem cell biology.* Cell, 2008. **132**(4): p. 631-44.

[4] Union for International Cancer Control, *Acute Myelogenous Leukemia and Acute Promyelocytic Leukemia*. WHO

[5] Shalem, O., et al., *Genome-scale CRISPR-Cas9 knockout screening in human cells.* Science, 2014. **343**(6166): p. 84-7.

[6] Ran, F.A., et al., *Genome engineering using the CRISPR-Cas9 system.* Nat Protoc, 2013. **8**(11): p. 2281-308.

[7] Qi, L.S., et al., *Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression.* Cell, 2013. **152**(5): p. 1173-83.

[8] Koike-Yusa, H., et al., *Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library.* Nat Biotechnol, 2014. **32**(3): p. 267-73.

[9] Ma, H., et al., *Pol III Promoters to Express Small RNAs: Delineation of Transcription Initiation.* Mol Ther Nucleic Acids, 2014. **3**: p. e161.

[10] Gibson, D.G., et al., *Enzymatic assembly of DNA molecules up to several hundred kilobases.* Nat Methods, 2009. **6**(5): p. 343-5.

[11] https://support.illumina.com/downloads/illumina-customer-sequence-letter.html

[12] http://www.tgries.de/agrep/

[13] Margolin, A.A., et al., *Reverse engineering cellular networks.* Nat Protoc, 2006. **1**(2): p. 662-71.

[14] Margolin, A.A., et al., *ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.* BMC Bioinformatics, 2006. **7 Suppl 1**: p. S7.

[15] Cancer Genome Atlas Research, N., *Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia.* N Engl J Med, 2013. **368**(22): p. 2059-74.

[16] Valk, P.J., Verhaak, R.G., et al., *Prognostically useful gene-expression profiles in acute myeloid leukemia.* N Engl J Med, 2004. **350**(16): p. 1617-28

[17] Hu, J., et al., *Heterogeneity of tumor-induced gene expression changes in the human metabolic network.* Nat Biotechnol, 2013. **31**(6): p. 522-529