GUIDESCAN SOFTWARE FOR IMPROVED SINGLE AND PAIRED CRISPR GUIDE RNA DESIGN COUPLED WITH COMPUTATIONAL STUDIES IN LEUKEMIA

A Dissertation
Presented to the Faculty of the Weill Cornell Graduate School
of Medical Sciences
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

By
Alexendar Reinaldo Pérez
December 2017

GuideScan Software For Improved Single and Paired CRISPR Guide RNA Design Coupled with Computational Studies in Leukemia

Alexendar Reinaldo Pérez
Cornell University 2017

CRISPR technology has revolutionized the field of genome engineering. CRISPR allows for the easy and efficient manipulation of virtually any genetic locus through a two-component system: a CRISPR endonuclease and guide RNA (sgRNA). These components form a complex that enacts double strand breaks in target DNA. The repair of the double strand break is the main mechanism by which genetic editing of a locus takes place. While the endonuclease cleaves target DNA, it is the sgRNA that specifies targets through complementary binding to a target site. Determining the specificity of sgRNAs to their target site represented a crucial challenge to the genome-engineering field. To facilitate the design of CRISPR libraries, we developed Guidescan, a software package that allowed for the customizable production of sgRNA databases that were guaranteed to match user-defined requirements for sgRNA uniqueness.

Furthermore, several computational studies of leukemia are described in this thesis that illustrate different molecular actors and mechanisms through which a leukemic like disease, Myelodysplastic Syndrome, can progress towards leukemia, how leukemia hijacks a splicing protein to maintain its pathology, and finally, how a leukemia can develop resistance to a targeted therapy.

Biographical Sketch

Alexendar Reinaldo Vincent Pérez was born in Pasadena, California. He attended Cornell University in Ithaca, New York for his undergraduate degree where he majored in Computational Biology and minored in Science of Earth Systems. He matriculated to the Tri-Institutional MD-Ph.D program in the summer of 2012.

After he defends his thesis he will return to medical school to complete the final clinical years of the MD-Ph.D program. Once he graduates the MD-Ph.D program he will return home to California.

Dedication

This dissertation is dedicated to the following individuals whose significance they each understand in their own right:

- My father and mother, Dr. Reinaldo J. Pérez and Madeline K. Pérez, who sacrificed everything to let me achieve all that I have.
- Reinaldo I. Pérez -, Palmidia M. Pérez, Vince E. Reilly -, and Dean Reilly –
- Joana Vidigal
- My brothers and sisters, Richard Pérez, Madi Pérez, Rey Pérez, Laura Rajabi, and Cyrus Rajabi
- My uncles and aunts, Vince Reilly, Ginger Neubauer -, Maria Pérez, Hector Pérez, Teresa Pérez, Richard Reilly, Laura Kuhn, and Gary Kuhn
- John Foo and Dan Hickey
- Jimmy Castellanos, Raúl Martínez-McFaline, Lisa Noble, and Andrew Milewski
- Simin Zhang and Jimmy Chen
- Sara Rahman
- Sue Hayden
- Javier -, Isidoro -, Juan –
- Karla Vergara Pérez and Tomas Castellanos
- El Viejo -
- Pasadena, California

Acknowledgements

I wish to acknowledge my Ph.D mentors Drs. Andrea Ventura and Christina Leslie for their support and mentorship throughout my training. Furthermore, I wish to acknowledge Dr. Joana Alves Vidigal, Dr. Yuri Pritykin, and Sagar Chhanagawala for their invaluable help and friendship throughout my graduate tenure. Additionally, I wish to acknowledge the members of my thesis committee: Drs. Michael Kharas, James Hudspeth, Oliver Elemento, and Olaf Andersen for their guidance and mentorship. Finally, I would like to acknowledge the Tri-Institutional MD-Ph.D program, specifically Dr. Ruth Gotian, Renee Horton, and Dr. Olaf Andersen for providing me with the support and resources I needed to come to this point.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Foundations of Genome Engineering:

Gregor Mendel's study, in the garden of a Czech monastery, of the attributes of peas established the idea that traits were heritable from one generation to the next[1]. Initially ignored, his pioneering results were rediscovered approximately twenty years later by Hugo de Vries, Carl Correns, and Erich von Tschermak who replicated his results and postulated that heredity was carried as discrete units across generations[2–4]. In 1905 William Bateson gave the term 'genetics' to the field of study that investigated the patterns and mechanisms of inheritance[5]. In 1944, Oswald Avery, Colin MacLeod, and Maclyn McCarty of The Rockefeller University demonstrated that genetic information was encoded in the molecule deoxyribonucleic acid (DNA)[6]. With the identity of genetic material confirmed to be DNA, a flurry of studies followed which established a paradigm that phenotypes derive from the transcription of DNA into ribonucleic acid (RNA) and translation of RNA into protein[7] (Figure 1).

In reality this dogma is more nuanced with the understanding that phenotype can be influenced by the actions of non-coding RNA, the topology of chromatin superstructures, epigenetic modifications to DNA and chromatin, and the environment[8–12]. However, this understanding is a more modern interpretation of the central dogma, which was not appreciated until the late twentieth century.

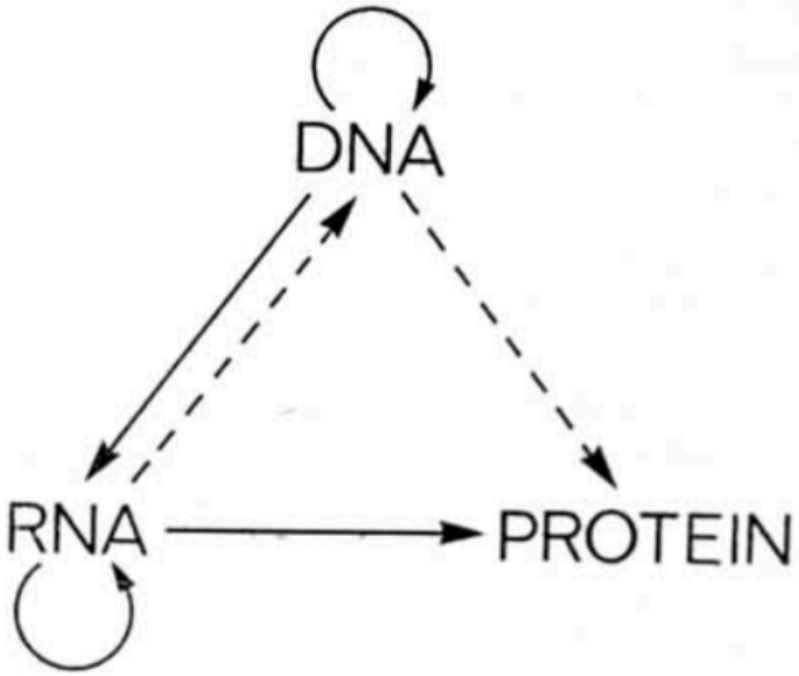


Figure 1: Central dogma of molecular biology as conceptualized by Francis Crick. Solid lines illustrated established transfers, while dotted lines show transfers observed in special cases or systems. (Crick F. 1970)

Figure 1: Central dogma of molecular biology

Throughout the development of the central dogma it was noted that phenotypes follow statistical distributions[13]. Implicit in this observation was the notion that within the central dogma, variance exists. This variance, from a nucleotide-centric standpoint, can derive from changes in DNA. These changes in DNA can occur in a population of individuals through a myriad of mechanisms including positional base changes, insertions and deletions of extra bases at a given locus, inversions of base sequences, and translocations of chromosomal arms[14,15]. While this variance exists naturally, the observation that alterations on the genetic level can yield changes on the phenotypic level created a paradigm in the life sciences. This paradigm stated that through altering the base sequence composition of DNA an investigator could potentially alter a phenotype. This concept would form the basis of what would eventually become the field of genome engineering.

Genome engineering can be broadly defined as the ability to precisely modify any arbitrary sequence, whether it be coding or non-coding, in any

arbitrary genome. While this definition of genome engineering fits well with the modern interpretation of the central dogma, the initial advances in genome engineering derived from a more protein centric viewpoint. The understanding that perturbations in genotype can deliver distinct phenotypes through the modification of effector proteins highlighted an early interest in the ability to change the base sequence of protein coding genes, which would thereby alter the protein itself and potentially replicate a phenotype. However, to achieve this modification investigators would need to discretely target and modify sequence-specific locations in the genome. One of the first attempts at sequence specific targeting and modification of a genetic locus occurred in the late 1950's when investigators used oligonucleotides cross-linked with bleomycin or psoralen to generate site-specific modifications in yeast and mammalian cells[16–21]. While this early attempt at genome engineering demonstrated an important proof-of-concept, that site-specific editing of a genomic target was possible, it failed to produce a robust method for gene editing.

In the late 1970s revolutionary experiments by Wigler and Axel illustrated that mammalian cells deficient in thymidine kinase (tk) could have their tk gene function restored by exposing these cells to a co-precipitate of calcium phosphate with DNA containing a herpes virus thymidine kinase (HSV-tk)[22,23]. The experiment demonstrated that co-precipitate DNA could accumulate on a cell's surface membrane and undergo endocytosis by the cell and correct the deficient enzymatic activity. Unfortunately, this initial method was rather inefficient in that the desired transformation would only occur in one out of a million cells exposed to the co-precipitate[24]. This efficiency problem was tackled and optimized in work done by Mario Capecchi and colleagues who advanced the method by microinjecting the HSV-tk gene, linked to viral nuclear homing sequences, directly into the nucleus of murine cells and observing a one million fold increase in transformation efficiency[25]. Capecchi's work revealed a key insight into the mechanism by which this transformation takes place. To appreciate this

insight, however, it is necessary to understand how a cell responds to a double strand break in the structure of its DNA.

Nuclear Repair Response to DNA Double Strand Breaks:

When DNA undergoes a double strand break (DSB), the cell will attempt to repair the break typically through either Non-Homologous End Joining (NHEJ) or Homologous Recombination[26,27] (Figure 2). While both pathways fix DSB they do so in distinct fashions. NHEJ pushes back together the blunt ends of the broken DNA to ligate the break[28–30]. This process is error-prone and can result in the introduction of insertions and deletions (indels) in the repaired DNA locus[31]. In contrast, homologous recombination mends DSB by using a sister template for homologous recombination and perfectly repairs the DSB with the template provided by the sister chromatid[32]. The presence of a sister template limits the stages in the cell cycle where homologous recombination repair can take place to the S and G2 phases, with greatest efficiency of repair occurring in the early S phase[33,34]. By contrast, NHEJ repair can occur at any point in the cell cycle[35]. If an engineered template is introduced into a cell by a researcher it can be treated as a sister template by the cell and used to enact a targeted modification via homologous recombination (also knows as homology directed repair [HDR])[36].

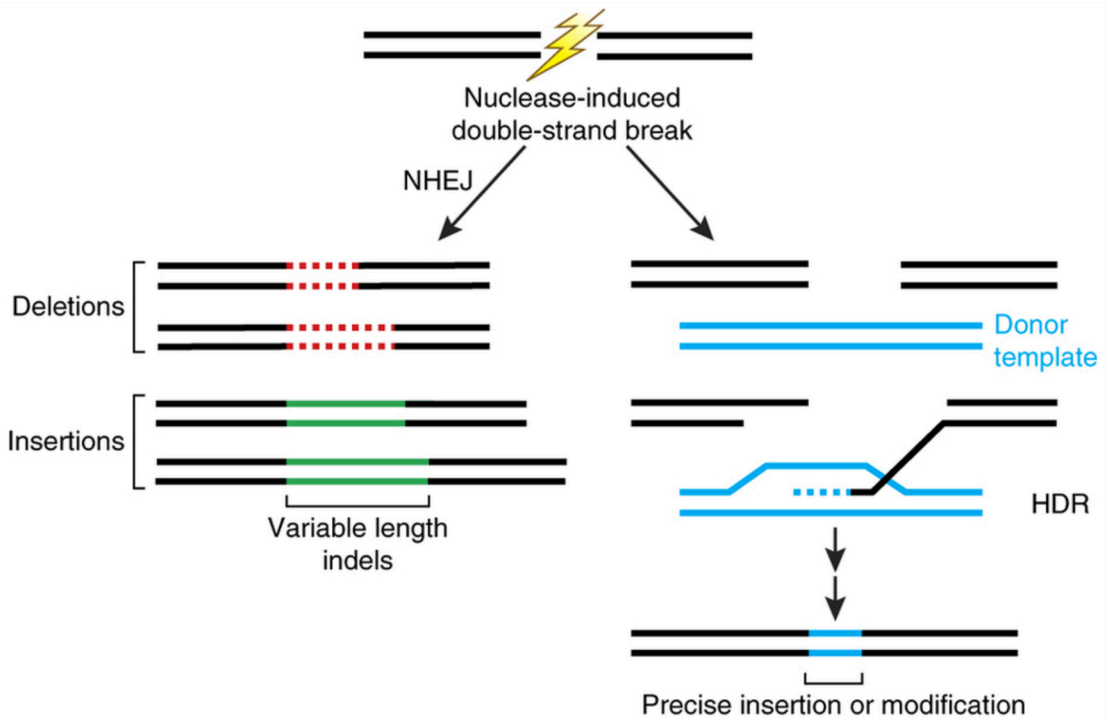


Figure 2: Double strand DNA repair through imprecise non-homologous end joining (NHEJ) introducing insertions and deletions of variable length. Alternatively, DNA can repair more precisely with homology directed repair (HDR) with donor template . (Sander et al. 2014)

Figure 2: DNA DSB repair

Genome Engineering and Homologous Recombination:

What Capecchi and colleagues discovered in the process of microinjecting HSV-tk into murine nuclei was that when many copies of the HSK-tk plasmid were introduced into a nucleus, they integrated into genomic loci in a highly ordered head-to-tail fashion[37]. The probability of such a head-to-tail structure occurring by chance at a single locus, x times, can be assessed with the following measure:

$$P(x) = \frac{1}{2^x}$$

where P(x) is the probability of observing x independent copies of a plasmid integrating into a single genetic locus in a continuous head-to-tail orientation. In Capecchi and colleague's original experiments they observed approximately 100 HSV-tk carrying plasmids all incorporated, in a structured

head-to-tail fashion, into a single genetic locus which represented an event with the probability of occurrence of:

$$P(100) = \frac{1}{2^{100}} = 7e^{-31}$$

Such a probability, they reasoned, was so unlikely to occur by chance that a basic biological mechanism had to be at play. Further studies by Capecchi's group proved that such a structure was achieved through homologous recombination of the cell's genome with the DNA located on the plasmid[38]. This work demonstrated, for the first time, that mammalian cells could undergo homologous recombination with experimentally engineered exogenous DNA molecules[38]. The importance of this discovery was not lost upon investigators as they quickly realized that, by harnessing the homologous repair machinery intrinsic to the cell, they could potentially modify any gene in a cell through the introduction of engineered DNA molecules[39]. Using this technology to generate transgenic mice carrying a desired genetic modification, researchers originally microinjected DNA into the nucleus of a mouse zygote. While this process was robust it was also tedious. In the mid-1980s the usage of mouse embryonic stems cells (mESC) to create germline transgenic mice began to take hold with investigators finding they could engineer genomic modifications through electroporating mESC with the desired modification DNA[40,41]. While this process allowed for a high throughput way to generate genomic alterations in a mouse, it was not modular nor did it easily allow for the custom targeting of genomic regions outside of murine system.

Meganucleases, Zinc-Finger Nucleases, and TALE Nucleases:

The first hint that researchers could develop a method that would allow for the arbitrary editing of a genetic locus in a wide array of model organisms and conditions came with the discovery of meganucleases (MGNs)[42]. These endonucleases derive from a large set of organisms including bacteriophage,

archaea, bacteria, fungi, yeast, algae, and even some plant species[42–45]. MGNs contain DNA binding domains that are characterized by having a large sequence recognition site typically on the order of twelve to forty bases long[46]. The size of this recognition site, coupled with its low tolerance for mismatches, makes MGNs highly specific to their target regions. Once the MGN binds a target site the catalytic domains on the endonuclease induce a DSB that allows for the HDR and NHEJ pathways to be activated, which like previous methods, provides a mechanism to custom engineer a genetic locus[45,47,48]. MGNs represented a tool that could potentially modify any cognate sequence in any organism. MGNs represented the first step towards establishing a generalized genome-engineering system[49,50]. However, the same specificity that made MGNs attractive also limited their use. While hundreds of MGNs have been discovered the existence of target sites within a studied genome remains restrictive. In fact, to discover a MGN eighteen base recognition site by chance in the human genome would require the human genome to be approximately twenty-three times larger than it presently is.

$$\frac{Possible\ 18mers}{Bases\ in\ Human\ genome} = \frac{4^{18}}{3 * 10^{9}} = 22.9$$

While efforts were made to modify MGNs sequence recognition sites so that they could home to desired targets within a genome, progress was slow coming. The problem was that the MGN sequence recognition site was not easily altered and thus not trivial to customize[51,52]. This changed with the introduction of zinc finger nucleases. (Figure 3a, 3b)

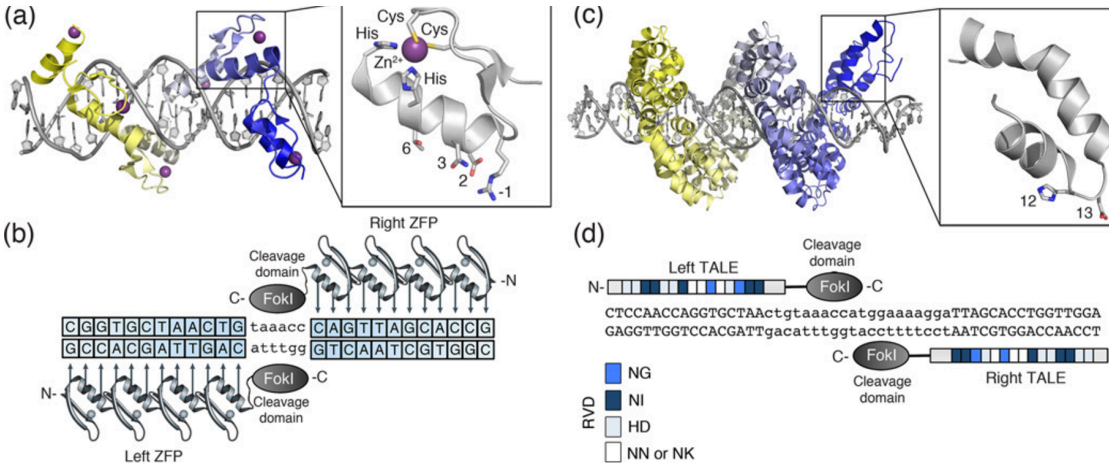


Figure 3: Zinc finger and TALE proteins **a.)** Zinc finger in complex with DNA. Zinc fingers are composed of approximately 30 amino acids and possess DNA contact domains which recognize 3-4 bp of DNA in the major groove. **b.)** Diagram of zinc finger nuclease dimer binding to a target site. The zinc finger target site is separated by a 5-7 bp spacer which is recognized by a FokI cleaving domain. The action of the zinc-finger dimer binds target sites while the cleavage of the target site is achieved by the ligated FokI domains. **c.)** TALE protein in complex with DNA. TALE proteins consist of 33-35 amino acids and are able to recognize a single bp of DNA through the use of dual hyper-variable residues (RVDs). **d.)** Diagram of TALE nuclease (TALEN) binding a target site. Like zinc finger nucleases, TALEN target sites consist of two binding sites separated by a spacer sequence 12-20bp in length. (Cathomen et al. 2008)

Figure 3: Zinc Finger and TALE proteins

Unlike MGNs zinc finger nucleases (ZFNs) were endonucleases that utilized modular domains to target and cut DNA[53–55]. The advent of ZFNs allowed investigators, for the first time, to robustly and accurately engineer site-specific modifications into a target locus by using a zinc-finger module that recognized a triplet of bases in the major groove of DNA[56,57]. These modules could be assembled together with a highly conserved linker sequences to yield a multi-modular ZFN that could recognize 9-18 continuous bases in length[53]. This modular array could then be joined to an effector domain, like a nuclease, and allow for site-specific modification at a locus[53]. Despite allowing for precise targeting of a sequence, ZFNs failed to gain widespread use. ZFNs suffered from a difficulty in design in that the linker sequence between zinc finger modules made their assembly difficult[58]. This design obstacle was partially overcome with the discovery of TALE proteins (Figure 3c,3d).

TALEs are bacterial derived proteins that, like MGNs and ZFNs, contain DNA-binding domains[59,60]. They are composed of a repeat sequence of 33-35 amino acids that form a domain that recognizes a single base pair[61,62]. Again, like ZFNs, these repeat sequence domains can be modularized to recognize a series of bases and joined to an effector domain, like a nuclease, to form TALENs[63,64]. Unlike ZFNs, the TALE modules do not require assembly with a linker sequence and the ability to recognize single bases as opposed to a triplet of bases required by ZFNs simultaneously simplified the assembly of these sequence-targeting proteins and increased their targeting resolution[65]. However, the difficulties inherent in protein design, synthesis, and validation, especially given the presence of repeat sequences in the TALE domains, limited the widespread adoption and use of TALENs[58]. This changed with the emergence of CRISPR systems.

Basic Biology of CRISPR:

Clustered-regularly-interspersed-short-palindromic-repeats (CRISPR) were first described in 1987 when Nakata and colleagues reported a series of 29-base repeats (termed direct repeats) interspersed by 32-base non-repetitive sequences (termed spacers) in the E. coli iap gene[66]. The presence of direct repeats and spacers was subsequently discovered to be a broadly conserved characteristic present in greater than 40% of bacteria and 90% of archaea[67,68]. However, the purpose of these alternating sequences remained elusive until 2005 when investigators determined a bacteriophage derived origin for the spacer sequences[69–71]. Compounding this finding was experimental evidence demonstrating that archaea with spacer sequences matching sequences found in a bacteriophage's genome, were immune to infection from that phage[70]. These findings suggested that CRISPR direct repeats and spacers constituted a component of the adaptive immune system of bacteria and archaea[70,71]. However, the mechanisms through which this immunity was established remained speculative.

Around the same time that the CRISPR direct repeats and spacers were realized to be widespread and broadly conserved across bacteria and

archaea, another important discovery regarding the CRISPR system came about: CRISPR -associated (Cas) genes[72,73]. CRISPR direct repeats were discovered to not be the only widely conserved sequences present in the CRISPR system with the realization that several protein coding gene clusters rested adjacent to the direct repeat and spacer regions. These gene clusters coded for a set of nuclease enzymes that facilitated the recognition and destruction of target nucleic acids and acquisition of new spacer sequences. From the diversity of these gene clusters and their products, CRISPR was divided into three distinct systems: Type I, Type II, and Type III[74,75] (Figure 4). Type I and Type III systems depend on multiple Cas proteins to form complexes, which target and degrade target double-stranded DNA[76]. By contrast Type II systems consist of notably fewer Cas proteins, oftentimes achieving their double-stranded DNA nuclease activity with just one enzyme[76].

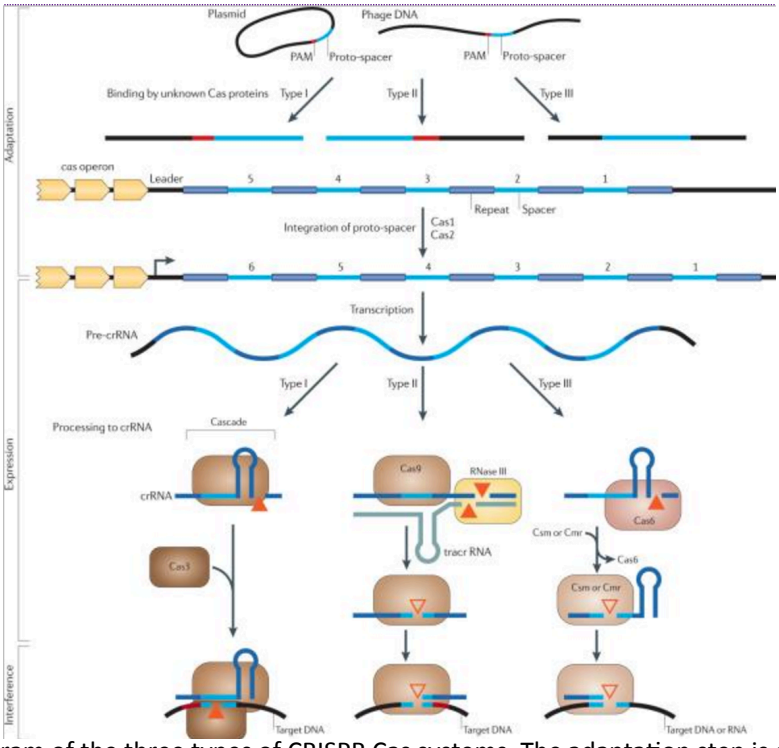


Figure 4: Diagram of the three types of CRISPR Cas systems. The adaptation step is conserved across all three systems, while the expression and interference steps in the systems are distinct (Doudna et al. 2014)

Figure 4: CRISPR Cas System

Once it was understood that spacer sequences derived from viral origin and Cas genes were nucleases, the field moved quickly towards understanding the mechanisms underlying CRISPR systems. In short order, it was demonstrated that CRISPR spacers represented a nucleotide-based memory of past infections and Cas enzymes were responsible for spacer acquisition upon novel infection as well as neutralizing phage[77,78]. Furthermore it was shown that CRISPR arrays were transcribed and cleaved into constituent components (termed crRNAs) that contained individual spacer sequences that guided Cas enzyme nuclease activity against DNA[76]. As study intensified on the Type II system, a critical piece of the CRISPR puzzle was elucidated; the protospacer adjacent motif (PAM) sequence. It was noticed that Type I and Type II Cas proteins did not target all sequences in the bacterial genomes where complementarity between the crRNA and the genome existed (notably, Cas enzymes never targeted the CRISPR array itself)[79]. Rather it was discovered that Cas enzymes appeared to target only those complementary regions that had a stereotyped sequence adjacent to the complementarity site. These sequences were termed protospacer adjacent motif (PAM) sequences and are recognized directly by the Cas enzyme[80,81]. Evidence that PAM sequences were required for targeting was bolstered by the fact that PAM sequences are completely absent from the CRISPR array direct repeats[79]. While Cas enzymes can recognize several PAM sequences, it was further demonstrated that they have a hierarchy of preferred PAMs. Likewise these PAM sequences vary across the different organisms that utilize CRISPR systems. As an example, a Type II Cas enzyme of intense study is spCas9 (derived from Streptococcus pyogenes) and has a preferred PAM sequence of NGG (N is a wildcard) with lesser degrees of tolerance for NAG, NGA, and NTG[82,83].

CRISPR as a Model for Bacterial and Archaea Adaptive Immunity:

The model depicting how a CRISPR system acts as a form of adaptive immunity in bacteria and archaea starts with an infection by a bacteriophage. Upon injection of the bacteriophage's DNA into the cell, the invading DNA is

recognized by two Cas enzymes, Cas1 and Cas2, which cleave the invading DNA into smaller components known as protospacers. These protospacers are brought into the CRISPR array as spacers between the first direct repeat in the array and the array leader sequence. A direct repeat sequence is then reestablished upstream of the new spacer and in this manner new spacers are continually prepended to the array thereby establishing nucleotide based memory of past viral infections. When the CRISPR array is called into action in the cell's adaptive immune response, the CRISPR array and Cas genes are transcribed and crRNAs are created. The processing of the CRISPR transcript and generation of the crRNAs vary across Type I, II, and III CRISPR systems[75,76].

In Type I systems the direct repeats of the CRISPR transcript form hairpin loops and cleavage of the crRNAs occurs at the junction of the single stranded RNA and double stranded RNA present at the boundary of the hairpin loops by Cas6e and Cas6f enzymes. The resulting crRNAs are then loaded into a multi-Cas enzyme complex (termed CASCADE complex) where target DNA interference is established through simultaneous PAM site recognition and complementary binding between crRNA and target sequence[76,84,85].

In Type II systems the direct repeats also form double stranded RNA, but instead of forming hairpin loop structures, the direct repeats bind to a noncoding piece of RNA (later discovered to be the tracrRNA). The double stranded RNA is then used by Cas9 and RNaseIII to cleave the CRISPR transcript into crRNAs. The resulting crRNAs undergo 5' trimming to form mature crRNAs that are then loaded into Cas9 where target DNA is cleaved by recognition of a PAM site by the Cas9 enzyme followed by the complementary binding between the target site and the crRNA[84–86].

In Type III systems the requirement for double stranded RNA is absent and the CRISPR transcript is cleaved by a Cas6 homolog. The resulting crRNAs undergo 3' trimming to form mature crRNAs that are then loaded into various Cas enzymes that are used to cleave target DNA. Interestingly, Type

III Cas enzymes do not require the recognition of the PAM sequence in order to deploy their nuclease activity[84].

Consequently, when a bacteria or archaea is infected by a bacteriophage, the phage DNA is cleaved and remembered as the spacer sequences in the CRISPR array and a host immune response is enabled through the direct cleavage of the DNA in the cytosol or extraction of integrated viral DNA in the bacterial or archaea genome through the activity of targeting crRNAs and Cas nucleases.

CRISPR Systems as a RNA-Guided Genome Editing Tool:

As the mechanisms by which CRISPR mediates bacterial and archaea adaptive immunity became established, the idea that CRISPR systems could be used as a genome editing tool started to gain traction. The understanding of this potential became clear when one of the key components in Type II crRNA biogenesis with the Cas9 enzyme was uncovered: the trans-activating crRNA (tracrRNA). The tracrRNA is a conserved non-coding RNA that binds with crRNA in the Type II system to form a crRNA-tracrRNA hybrid that is important in the processing of the CRISPR transcript[87]. The tracrRNA provides the secondary structure needed to allow the crRNA to be properly loaded into the Cas9 enzyme[86,88]. With the discovery that only the tracrRNA, crRNA, and Cas9 enzyme were needed to effect sequence target activity, investigators realized the potential of the CRISPR-Cas9 system to serve as a RNA-guided genome editing technology. This potential was further developed when the system was reduced from three to two components with the creation of a synthetic single guide RNA (sgRNA) that resulted from the fusion of a crRNA and tracrRNA sequences[89]. Researchers could now simply vary the approximately 20-nucleotide complementary region of the 5' end of the sgRNA (crRNA component of the sgRNA) to target any region of the genome they desired provided a PAM sequence existed 3' to the target site. The target site would undergo a double-strand break through Cas9's nuclease activity and the NHEJ or HDR repair pathways would be activated[90]. The CRISPR-Cas9 system was shown to function in a myriad of cell types,

including mammalian cells, at which point the modern era of genome engineering was born[91,92].

Applications of CRISPR-Systems in Genome Engineering:

Site-specific sequence targeting using CRISPR-Cas9 opened up a vast array of previously intractable problems in genome engineering (Figure 5a). Cas9 induced frameshift mutations in coding sequences of target genes allowed for the generation of gene knockout models through rapid and efficient means[93]. Furthermore, the introduction of multiple sgRNAs into a cell allowed investigators to simultaneously assess knockout mutations across a set of genes in one experiment[94–96]. This ability also allows CRISPR sgRNAs to be used in a paired way enabling large-scale positive and negative deletion screens in both the coding and non-coding regions of the genome[93,97]. However, the engineering of frameshift mutations represented only the beginning of CRISPR's potential applications.

The editing potential of CRISPR systems also allows for the recapitulation of the genetic variants present in disease. The CRISPR-Cas9 system, coupled with an oligonucleotide carrying a potential pathogenic alteration, allows investigators to directly interrogate the role of a genetic variant in disease pathogenesis[98]. This represents a substantial advance in allowing researchers to model disease by dissecting its genetic underpinnings instead of relying on animal models that may only phenocopy disease traits. This use of CRISPR systems is especially intriguing when applied to understanding complex human diseases such as diabetes, cancer, and schizophrenia. It also offers a lens to understand the results of genome-wide association studies where it is often difficult to determine which of several genetic variants, in linkage disequilibrium with a haplotype, are causing the phenotype[99]. Additionally, CRISPR-Cas9 systems can be used to model pathogenic genetic lesions in their native regulatory environment; such as with the modeling of the EML4-ALK fusion present in a rare set of lung cancers[100] (Figure 5b).

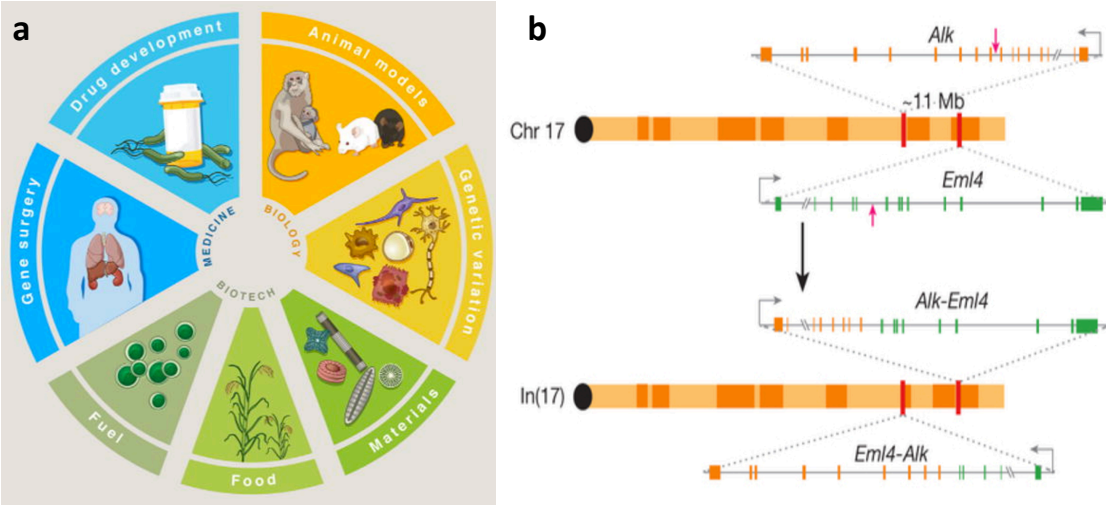


Figure 5: Applications of CRISPR systems. **a.)** The potential use of CRISPR systems is broad with applications being sought in medicine as an advance in gene therapy, optimize certain applications in synthetic biology, engineer crops that are resistant to pests and increases global food security, assist in the development of organic fuels based on ethanol or algae, streamline the production of drug precursors needed for pharmaceutical development, and finally in generating animal models of biology or disease that more precisely replicate a genetic variant in its native background environment such as in [117] **b.)** the Eml4-Alk gene fusion which has been observed in a subset of lung adenocarcinomas. Such modeling expresses the fusion protein under its natural regulatory environment. (Wright et al. 2016)

Figure 5: Applications of CRISPR

Beyond direct sequence modification, CRISPR-Cas9 systems can also be utilized to affect transcriptional regulation through the use of a version of Cas9 where the nuclease domains have been rendered inactive (dCas9)[101]. dCas9 can bind to target DNA and repress transcription by sterically inhibiting the activity of transcription machinery. This inhibition of transcription (termed CRISPRi) works well in bacteria and archaea and shows promise in eukaryotes[102]. Tethering transcriptional repressors to the dCas9 enzyme can augment the inhibitory activity of CRISPRi[103]. Conversely, tethering transcriptional activators to the dCas9 enzyme and directing the multiple dCas9s to a promoter sequence can promote transcription of a target gene[103].

Overall, CRISPR systems in general and the Cas9 system specifically, allow investigators to quickly and efficiently target and modify site sequence composition and expression. However, the uses of CRISPR-systems are not limited to just research.

Clinical Potential of CRISPR Systems:

Perhaps the most exciting potential application of CRISPR systems is their use as a therapeutic in human diseases. Genome engineering has long held the promise of serving as a treatment for genetic diseases. While early attempts at treating aliments such as cystic fibrosis made use of wild-type genes brought into diseased cells through adenoviruses, these approaches did not give sustained phenotypes because their effectors, over time, became diluted or inactivated[104,105]. With the advent of ZFN and TALENs further progress was made, even to the point where ZFNs were utilized in a clinical trail to engineer protective knockout mutation in the CCR5 receptor of human T-cells against the HIV virus[106]. However, these tools remained difficult to synthesis, validate, and ultimately use in many clinical settings[57]. The accessibly and versatility of CRISPR systems makes the task of engineering changes to a patient's genome achievable on a clinical time scale. A well-studied CRISPR system set to appear in upcoming clinical trails is the Cas9 system.

Cas9's ability to induce DSBs repaired by the NHEJ pathway makes it an exciting tool to combat diseases characterized by dominant-negative mutations. Illnesses such as transthyretin-related hereditary amyloidosis or the genetically dominant version of retinitis pigmentosum are characterized as having one mutant allele whose protein product makes the cell haploinsufficient for the gene[107,108]. A Cas9 sgRNA designed against the mutated allele can introduce frameshift mutations into the gene that could disrupt the protein structure thereby eliminating the pathogenic dominant negative action of the protein.

Paired-sgRNA approaches, where two sgRNAs are designed against a target, possess the potential to serve as a therapeutic in treating tri-nucleotide repeat disorders such as Huntington's Disease, Myotonic dystrophy, and Friedreich's ataxia[109–111]. SgRNAs can be designed flanking repeat expansion regions and used to delete these pathologic regions of the genome. This approach holds special potential for Friedreich's ataxia, a neurodegenerative

disease that primarily affects children. The tri-nucleotide repeats in Friedreich's ataxia occurs in a non-coding region, which makes frameshift mutations caused by the indels created by NHEJ less worrisome[111] (Figure 6).

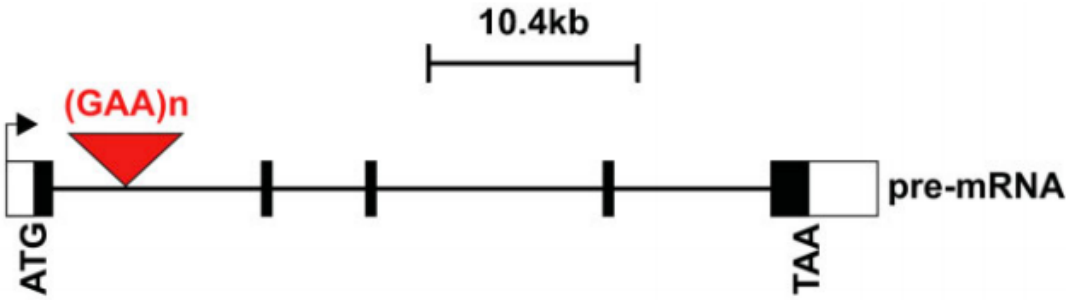


Figure 6: Gene diagram of frataxin with the pathogenic tri-nucleotide repeat (red) shown to occur in the intronic region of the gene. This tri-nucleotide repeat is believed, through a poorly understood mechanism, to be involved in the pathological mechanism underlying Friedreich's Ataxia. (Campuzano et al. 1996)

Figure 6: Frataxin gene

Cas9 induced DSBs may also be repaired through the HDR pathway, which allows for the precise editing of loss-of-functions genes characteristic of diseases such as cystic fibrosis, sickle-cell anemia, and Duchenne muscular dystrophy[112–114]. Wild-type oligonucleotide templates can be introduced with a Cas9 sgRNA targeted against a mutant site. The resulting DSB can undergo HDR and incorporate the wild-type template and in so eliminate the pathogenic mutation. While promising, this approach may still be several years down the line as the HDR repair frequency of DSB sites is low compared to NHEJ repair at the same site[24].

However, the application of CRISPR systems to the clinic is not just theoretical. Presently, the Cas9 system is being used in clinical trials designed to engineer chimeric antigen receptor (CAR) T cells to re-recognize cancer antigens present in patient's with prostate cancer[115]. Investigations are underway to use the Cas9 system to engineer the protective CCR5 mutation in human T cells to delete PCSK9 or angiopoietin to treat statin-resistant

hyperlipidemia and hypercholesterolemia[99]. CRISPR systems are already being used at the forefront of the precision medicine revolution.

Challenges Facing CRISPR Systems as Genome Engineering Tools:

The promise of CRISPR systems for the field of genome engineering is immense, however, it is not without its challenges[58]. Chief among these concerns is the notion of off-target cutting by the CRISPR endonucleases[58,90,99]. The specificity of CRISPR systems derives from the complementary sequence present in the sgRNA that recognizes a target site in a genome[116]. The target site is composed of a sequence of approximately twenty nucleotides that is, ideally, unique in a genome for a given PAM. Yet, even if a sequence is uniquely occurring for a specific PAM, it does not mean it is immune from off-target cleavage. The sequence could have perfect occurrences with alternative PAM sequences. These sites would be cleaved with reduced, but still notable, efficiency. For example, in the Cas9 system the NAG alternative PAM sequence is cleaved at approximately 20-25% the efficiency of the NGG PAM[117]. The occurrence of perfect match target sites with alternative PAMs represents a considerable off-target effect for CRISPR systems. However, alternative PAM perfect match off-targets are not the only concern regarding CRISPR off-targets. The complementary region of the sgRNA is capable of withstanding mismatches[116]. Similar to short-hairpin RNA (shRNA), sgRNAs can cleave off-target sequences that differ from a target site by only a few bases[116,118]. While the rules regarding mismatch tolerance in sgRNAs are an area of active investigation, the literature supports the idea that sgRNAs with a 20-nucleotide long complementary region can withstand upwards of 6 positional mismatches to a target[116]. The ability to design sgRNA that target precisely in a genome is essential for both research and clinical applications of CRISPR systems. Consequently, the ability to correctly enumerate all potential off-target sites for a given sgRNA is critical for the development of CRISPR as a genome engineering technology.

The understanding that sgRNA specificity derives from genome specific target site similarity raises an additional concern for using CRISPR

systems for genome editing: designing sgRNAs from a reference genome. In 2003 the Human Genome Project released the initial assembly of the human genome[119,120]. This assembly, along with assemblies for various other organisms, has been made more complete with advances in next-generation sequencing over the years allowing for the creation of organismal reference genomes[121,122]. Although these reference genomes are invaluable for bioinformatics, they cannot be used in all designs of sgRNAs. Although the sequence conservation among individuals in a non-clonal species is exquisitely high, it is never one hundred percent (excluding identical twins). Consequently, the genomes of cell lines derived from a specific individual will differ slightly from another. Furthermore, should CRISPR systems ever to be used in clinical medicine, sgRNAs would have to be designed with the knowledge that patient genomes have overlapping, but distinct, sets of target sites. Designing sgRNAs from a reference genome makes an incorrect assumption that a set of target sites is present in all individuals of a given species. An sgRNA that may be uniquely occurring in one individual's genome but may have multiple perfect or near-perfect occurrences in another individual's genome. These off-target effects could confound research findings or even induce disease if used clinically[123]. Ultimately, the use of CRISPR systems in research and medicine will depend on the ability to create individual-specific sets of sgRNAs.

SgRNA on-target cutting efficiency constitutes an additional major concern of researchers and clinicians when using CRISPR systems. An sgRNA's ability to uniquely cleave a target site is made void if the cleavage efficiency at the site is minimal. While several groups have been made efforts to understand the variables contributing to sgRNA cutting efficiency, the general rules governing cleavage efficiency remain relatively undefined[124–126]. Furthermore, present methods of assessing an individual sgRNAs' cleavage efficiency largely ignore the specificity of the sgRNA[124–126]. The cutting efficiency of sgRNAs is likely multifactorial in etiology and methods to predict such efficiency likely depend on a complete understanding of all these

factors. As a result a more complete understanding of the factors that contribute to sgRNA on-target cutting efficiency are needed if CRISPR systems are to be used more broadly.

Ultimately, researchers and clinicians using CRISPR systems for genome engineering tasks seek sgRNAs that are optimized for both target specificity and cutting efficiency. Various groups have developed tools that attempt to service this task, however, they suffer from certain limitations.

Limitations of Current sgRNA Selection Methods:

Many methods exist which attempt to select sgRNAs optimized to be specific to a given target site as well as quantify how efficiently that target site will be cut[126–129]. Unfortunately, these tools tend to suffer from a common set of limitations. For one, these tools will generate sgRNAs using the reference genome. With the exception of the method described in the forthcoming chapter, none of the currently available tools allow for the construction of individual specific sgRNA databases. Consequently, these tools will produce a subset of sgRNAs that may have non-existent or even heavily expanded target spaces in a particular individual of a given species due to genetic variation among individuals.

Secondly, most presently available sgRNA selection tools provide sgRNAs only for the CRISPR-Cas9 system. While this CRISPR system is one of the most widely used, it is not the only CRISPR system that could be used for genome engineering purposes. Numerous other Type II CRISPR systems hold the potential to serve as genome engineering systems. In particular the Type II CRISPR system involving the Cpf1 enzyme garners much interest among investigators due to its ability to induce staggered end cuts at a target site as opposed to the blunt end cutting induced by Cas9[130]. Few tools provide sgRNAs using non-Cas9 CRISPR systems and many methods appear to be specific to generating CRISPR-Cas9 sgRNA sets. As a result many of these tools are inflexible to advances in the CRISPR field should a more desirable endonuclease come about that replicates the popularity of Cas9.

Thirdly, many methods only consider sgRNAs that possess the canonical NGG PAM sequence for sgRNA selection. These methods will often consider the NAG alternative PAM in their off-target assessment, but few tools consider other potential alternative PAMs[126,127,129,131]. This represents a limitation given that multiple alternative PAM sites exist for the Cas9 CRISPR system and knowledge of all potential cutting sites will be essential if the technology is to ever be translated into the clinic. However, the limitation is not only relevant to off-target searches. Many groups are attempting to increase the specificity of the Cas9 CRISPR system by altering PAM site specificity[132]. This typically takes the form of expanding the PAM sequence that Cas9 recognizes from NGG to a stereotyped k-mer that is longer than NGG. The inability of many tools to easily consider expanded sets of alternative PAM sequences inhibits their broader utility.

Fourthly, virtually all current methods fail to correctly identify potential off-targets within a defined number of mismatches. Several groups have noted the failure of popular sgRNA selection tools to enumerate potential off-targets that are even one base different from the intended target site[116,124]. The inability of tools to determine the target universe for a sgRNA undermines the purpose of sgRNA selection tools. A correct and exhaustive enumeration of an sgRNA's target universe represents an essential requirement for CRISPR systems if they are to be used as genome editing tools. The failure of present methods to precisely determine the targetable universe for a sgRNA within a finite set of mismatches illustrates a critical need in the CRISPR field.

Lastly, all current sgRNA selection tools consider on-target cutting efficiency, either without considering or considering with incomplete knowledge, the specificity of a given sgRNA[124–126].

Overall these limitations of present sgRNA selection tools constitute a critical and unmet need in the CRISPR genome engineering field. The task of providing a potential solution to overcome these limitations is the essence of the following body of work.

Chapter 2: GuideScan

The sgRNA Specificity Problem:

Targeted alterations of a genomic locus using a CRISPR system depends on a.) the recognition of a PAM sequence by a CRISPR endonuclease and b.) complementary binding between the sgRNA and the target site. The matching of a sgRNA to a target site is accomplished through hydrogen bonding between nucleobases[133,134] (Figure 7a,b,c).

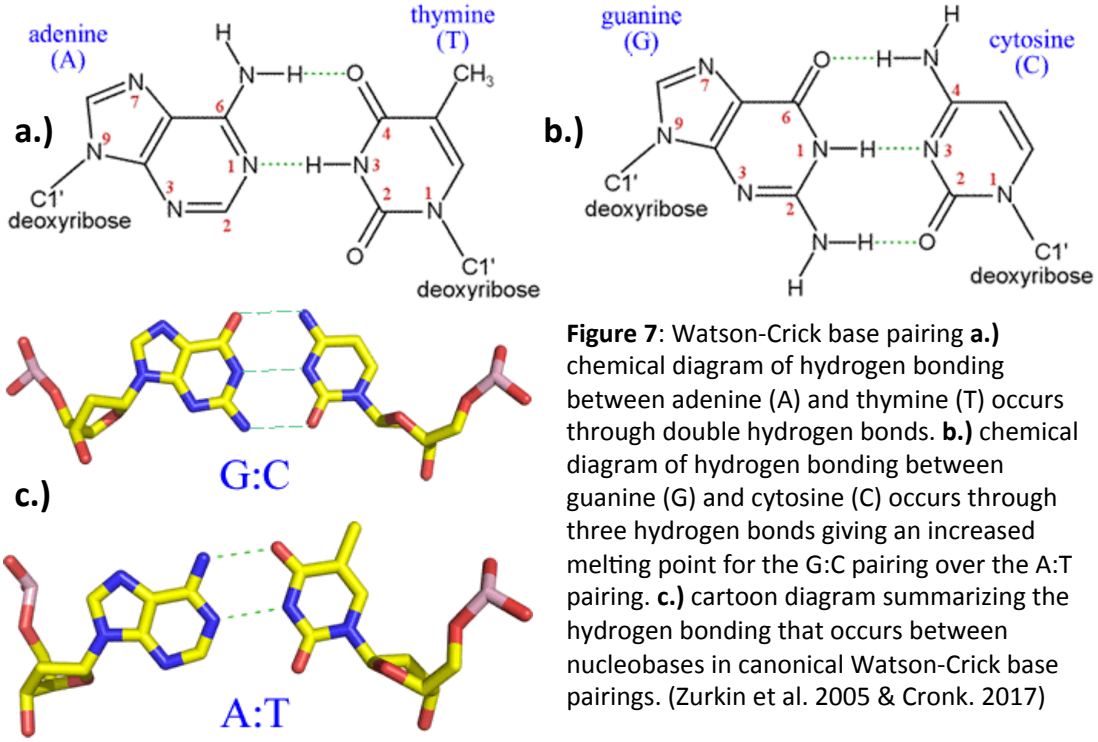


Figure 7: Watson-Crick base pairing **a.)** chemical diagram of hydrogen bonding between adenine (A) and thymine (T) occurs through double hydrogen bonds. **b.)** chemical diagram of hydrogen bonding between guanine (G) and cytosine (C) occurs through three hydrogen bonds giving an increased melting point for the G:C pairing over the A:T pairing. **c.)** cartoon diagram summarizing the hydrogen bonding that occurs between nucleobases in canonical Watson-Crick base pairings. (Zurkin et al. 2005 & Cronk. 2017)

Figure 7: Watson-Crick hydrogen bonding

As hydrogen bonding is a non-covalent interaction, the pairing between the CRISPR endonuclease and the target site is non-permanent. The strength of the interaction is influenced by the length of the complementary region, the GC content of the site, and the amount of mismatches between the complementary and target regions[133]. Ultimately, whether a sgRNA will bind to a target region reduces to a question of thermodynamics and whether the sgRNA is sufficiently similar to a target site for it to be energetically favorable

to bind. SgRNAs with perfect complementarity to a target site possess the greatest favorability of binding, but sgRNAs that differ from a target site by only a few nucleobases may also be able to bind. Stated plainly, sgRNAs can bind target sites with mismatches.

It is the binding of a sgRNA to a non-target loci in a genome that constitutes an off-target effect. Consequently, off-targets take on two forms a.) unknown loci in the genome that perfectly match a sgRNA and b.) known or unknown loci that are degenerate to the sgRNA complementary region, but which possess a free energy profile favorable to binding. For CRISPR systems these degenerate sites are distinct to a sgRNA by a finite number of positional mismatches[116,124]. An experimenter needs to know the target space of a sgRNA. The importance of this task derives from the fact that CRISPR systems induce a DSB at a target location that results in a mutation at that site. Furthermore, if a target site is present and the CRISPR endonuclease is active then theoretically the endonuclease will attempt to cut the region until the target site is disrupted and the endonuclease can no longer bind[135] (Figure 8a-e).

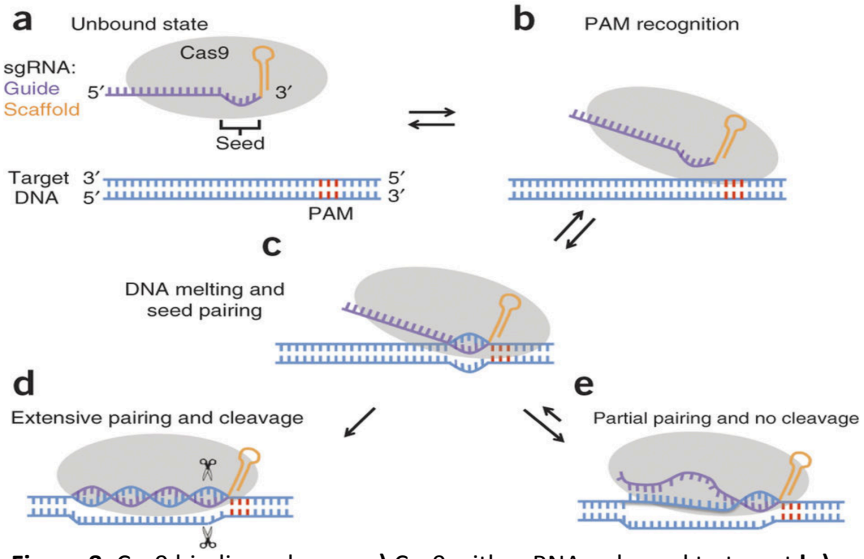


Figure 8: Cas9 binding scheme **a.)** Cas9 with sgRNA unbound to target **b.)** Cas9 enzyme recognizing PAM sequence **c.)** sgRNA and target site base pairing **d.)** sgRNA pairing and ultimate cleavage of target **e.)** Cas9 aborts cleavage if sgRNA and target site unable to base pair. (Wu et al. 2014)

Figure 8: Cas9 repeat target scheme

The reality that CRISPR systems effect permanent changes in a genome and will continuously attempt cleavage at target loci, highlights the need of experimenters to understand the target space of a given sgRNA to assess it's fidelity to it's target. If multiple genomic cuts occur without the experimenter's knowledge then the conclusions drawn from the experiment are in doubt. Understanding a sgRNA's target fidelity is the crux of the sgRNA specificity problem.

Naïve Genomic Off-Target Search:

Essential to determining a sgRNA's specificity is knowledge about its potential off-targets. Given the size of a sgRNA's complementary region and the amount of bases present in various model organisms, there is a high likelihood that a given sgRNA will be uniquely occurring in a given genome[136].

$$\frac{Possible\ 20mers}{Size\ of\ Genome} = \frac{4^{20}}{Size\ of\ Genome\ (bp)} = Magnitude\ of\ 20mer\ Excess$$

This likelihood is increased when the size of the genome is restricted to only the space of 20mers that are 3' flanked by a Cas9 PAM sequence, which in this case we will define to be NGG or NAG (Table 1).

$$\frac{Possible\ 20mers}{(PAM\ occurrence)} = \frac{4^{20}}{(PAM\ occurrence)} = Magnitude\ of\ Target\ Excess$$

Table 1: Magnitude of target excess

Organism (Assembly)	$\frac{Possible\ 20mers}{(PAM\ occurrence)}$	$Magnitude\ of\ Target\ Excess$
Human (hg38)	$\frac{4^{20}}{744,651,681}$	1,476
Mouse (mm10)	$\frac{4^{20}}{666,467,906}$	1,649
Zebrafish (danRer10)	$\frac{4^{20}}{251,921,183}$	4,264
Fly (dm6)	$\frac{4^{20}}{28,720,587}$	38,283
C. elegan (ce11)	$\frac{4^{20}}{16,867,034}$	65,187
Yeast (SacCerv3)	$\frac{4^{20}}{2,358,265}$	466,237

Despite this high potential for a sgRNA to be unique, it is known that sequence repeats are characteristic of many genomes[137–139]. Therefore it cannot be assumed that a given sgRNA has a unique target site.

Determining a sgRNA's target specificity therefore becomes a matter of accounting for all potential cleavage sites available to the sgRNA within a set number of mismatches. Accomplishing this task naively requires the scanning of the entire genome for PAM sequences and computing the target sequence adjacent to each PAM. Once all target sequences are enumerated then the genome must be scanned again, comparing each specific target sequence against all other target sequences in the genome. This comparison would take the form of computing a Hamming distance between a specific target sequence and all other target sequences in the genome. In this manner the entire space of degenerative neighbors (henceforth called a mismatch neighborhood) could be determined. Unfortunately, this process is quadratic in complexity (Figure 9).

```
for PAM in genome:
    compute complementary sequence from PAM
    append complementary sequence to a list

for sgRNA in list of sgRNAs:

    for PAM in genome:
        compute complementary sequence from PAM
        compute Hamming distance between sgRNA and complementary sequence
```

Figure 9: Pseudo code of naive off-target determination

$$l = length\ of\ genome$$

$$h = Hamming\ distance\ function$$

$$z = number\ of\ PAM\ sites$$

$$q = number\ of\ sgRNAs\ queried$$

$$l * z * h(l_1 + l_2 + \cdots + l_q) = Complexity\ of\ Function$$

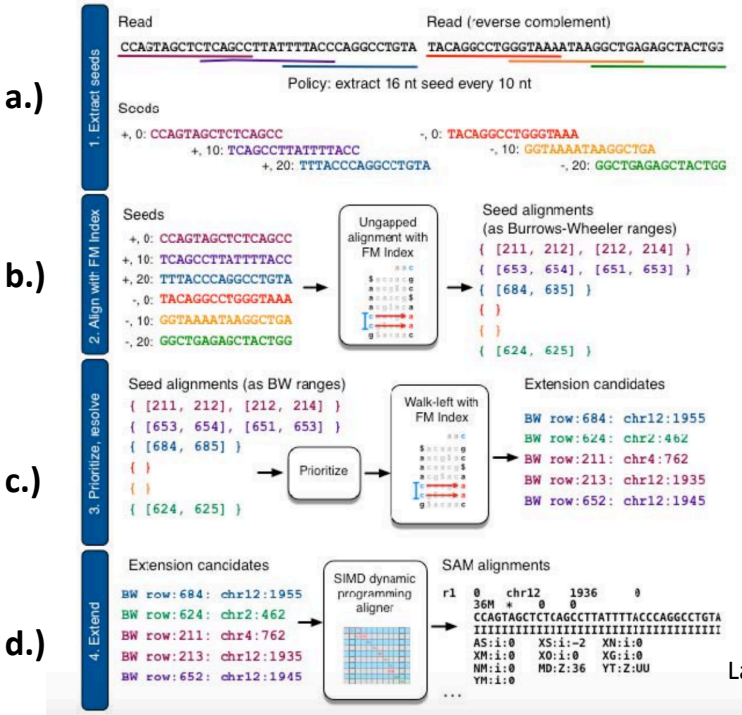
$$O(l^2) = Complexity\ of\ Function$$

This process is inefficient from the repeat scanning of an entire genome for each sgRNA. In reality, only the set of sequences adjacent to a 3' PAM sequence need to be considered, which would reduce the scanning space. However, it is the repetitive comparison of a sgRNA complementary sequence against all possible target sites that makes this process computationally expensive. Ultimately, this naïve approach for determining sgRNA specificity is unfeasible for determining the off-targets of sgRNAs. A faster and more efficient approach is needed.

Genome Aligners:

A natural solution to the off-target search would be to employ genomic sequence aligners (henceforth referred to as simply aligner). An aligner is an essential tool in bioinformatics that serves to rapidly and efficiently map sequence data to a reference genome. The precise methods through which

this mapping takes place varies across tools, but a general description can be made using the Bowtie2 aligner as a case example[140].



Langmead et al. 2012

Figure 10: Bowtie 2 workflow **a.)** Split sequence reads as substrings **b.)** Align substrings to FM-Index using Burrows-Wheeler transformation **c.)** randomly select Burrows-Wheeler range and resolve offset to reference genome **d.)** Final resolve of alignment to reference through SIMD

Figure 10: Bowtie2 workflow

Reads need to be accurately assigned to a discrete location(s) in a genome, which is the principle task of an aligner. The aligner processes the reads as a string data-type and splits these strings into substrings that are a defined length for the forward and reverse complement (Figure 10a). These substrings are called seed sequences. Next the seed sequences are aligned, in an ungapped manner, to a reference genome using an index (Figure 10b). In Bowtie2 this index takes the form of an FM-Index that uses a Burrows-Wheeler transformation to give a Burrows-Wheeler range (appendix). The resulting Burrows-Wheeler ranges are prioritized such that rows with smaller ranges are assigned a higher priority for mapping determination. Bowtie2 will select rows randomly with a probability proportional to the row's priority weight and resolve the offset of the row with the reference genome using the FM-Index's *walk-left* procedure (Figure 10c). Finally, the aligner performs

Single Instruction Multiple Data (SIMD) accelerated dynamic programming alignment on the vicinity of the resolved alignments until either a.) a sufficient number of alignments are examined b.) all seed hits are examined or c.) dynamic programming effort limit is achieved (Figure 10d).

The advantage of aligners is that they are optimized to be efficient and rapid in their alignment of reads to genomic loci. Additionally, they are capable of tolerating mismatches and/or indels between a read and an alignment location. Given that an sgRNA can be thought of as a twenty base pair read, the ability of an aligner to quickly assign a genetic coordinate(s), as well as determine degenerate alignment loci, makes it an appealing solution for determining a sgRNA's off-targets.

Suboptimal Behavior of Genome Aligners in Off-Target Search:

The intended purpose of aligners is to robustly and accurately map reads to a reference genome. However, to accomplish this task using the best alignment algorithms between two strings, requires an algorithm that is quadratic in complexity[141].

Consequently, this behavior is unfeasible for alignment purposes and heuristics must be employed to make alignments quicker while remaining robust and accurate. A common heuristic used by aligners is to break reads into substrings and optimize alignments for those substrings. While this and other heuristics speed up aligner behavior, they can allow for missed potential alignment sites that represent false negative events[141].

These false negative events can be empirically shown when one uses a standard aligner to map reads that have multiple known perfect sequence matches to the hg38 assembly of the Human reference genome[140] (Table 2).

Table 2: Aligner empirical error

At least X perfect sequence matches	Total Reads	Captured All Possible Read Alignments	Missed at Least One Possible Read Alignment	Missed Read Alignment Percentage
X = 2	100	48	52	52%
X = 10	100	37	63	63%
X = 100	100	5	95	95%
X = 1000	100	5	95	95%
X = 10000	100	0	100	100%
X = 100000	12	0	12	100%

When dealing with the CRISPR system the existence of these false negatives is worrisome since they indicate sites in the genome that have the potential of being cut and permanently altered. These false negative sites represent an underreporting of a sgRNA's target space and may compromise the results or analysis of a CRISPR experiment. Consequently, to determine the complete target space of a sgRNA and thereby determine it's specificity, a different approach is needed to determine off-targets.

Retrieval Tree:

To accurately determine the target space of any sgRNA requires the complete knowledge of all the potential target sites, within a set number of mismatches, in a genome. To do this assessment with recursive scanning of a genome is computationally unfeasible, however, with the utilization of a retrieval tree (henceforth referred to as trie) this determination becomes tractable[142,143].

A trie is a specific type of ordered tree where each node contains a character value and the root node represents an empty string (Figure 11). Children nodes have the same prefix, which is derived from the traversal of

their parent and ancestor nodes. A trie is traversed using the pre-order procedure and thus trie traversals are linear in complexity. Additionally, to build a trie from a collection of strings has simple complexity.

$$l = length\ of\ word$$

$$w = number\ of\ words\ in\ collection$$

$$O(lw) = \ Complexity\ of\ Trie\ Construction$$

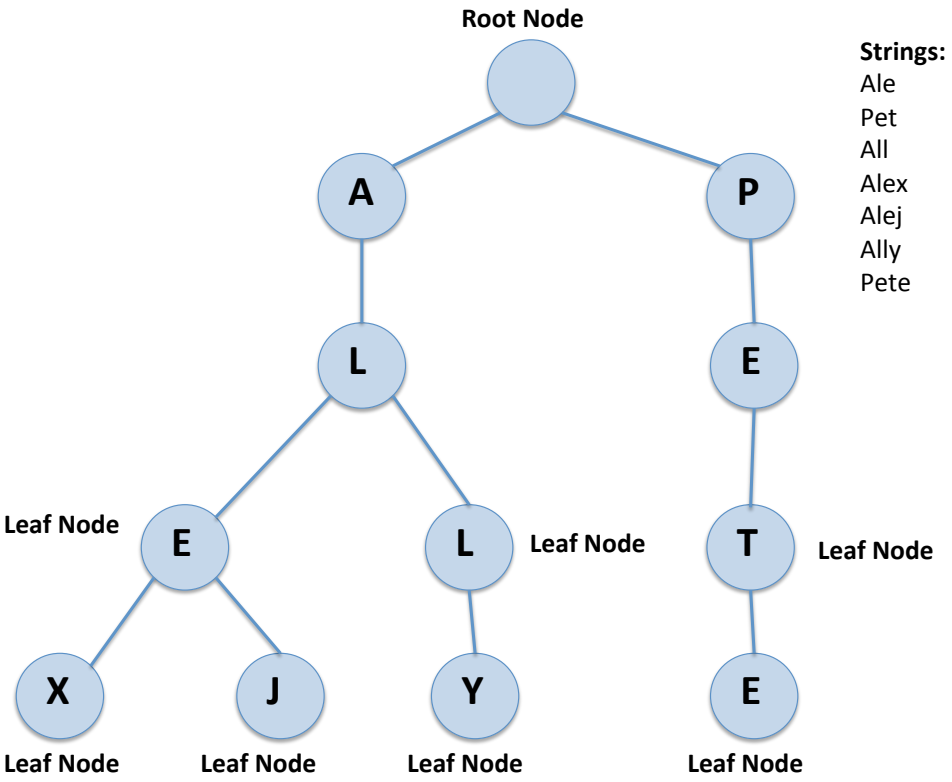


Figure 11: Trie data structure

Figure 11: Trie data structure

Furthermore, for a given traversal of a string through a trie, a mismatch neighborhood for the string can be computed. Stated another way a string and all strings in the trie that are similar to it, by at most a distance h, can be identified through a traversal. This distance can be either a Hamming distance or Levenshtein distance[144,145].

Trie Based Off-Target Search:

Given the simple complexity of construction and traversals, coupled with its ability to enumerate mismatch neighborhoods for strings, a trie represents a tractable solution for determining off-targets, and thereby specificity, of sgRNAs. The utilization of a trie for this purpose would start with the scanning of a genome for PAM sequences. This scanning is a O(n) process and simply scales with the size of the genome. When a PAM sequence is identified, a sgRNA complementary sequence is computed by taking a sequence of length *l* that is adjacent to the PAM site. As the sequences are extracted from the genome they are stored in a file that represents the universe of sgRNAs for a given genome.

This universe of sgRNA's is then used for the construction of the trie, which is a process that is O(*lw*) in complexity. Since *l* will be constant, however, the process reduces to O(w), which is simply the amount of sgRNAs in the genome. The branches in the trie will represent sgRNA complementary sequences and the leaf nodes associated with each branch will record the amount of times the sgRNA sequence occurs in the genome. Simply by constructing the trie, one is able to determine which sgRNA sequences are uniquely occurring in the genome. The ability to determine sgRNA uniqueness is essential in determining sgRNA specificity. This determination can be made exactly and through a process that is linear in complexity.

Furthermore, the complete mismatch neighborhood of a sgRNA can be enumerated through traversing a trie, which is also a process that is linear in complexity. SgRNAs have been shown to tolerate positional mismatches, but have low tolerance for insertions and deletion[124]. Consequently, computing the Hamming distance at a value h, for a given sgRNA as it traverses through the constituent sequences in the trie will exhaustively determine the degenerate sequences to which a sgRNA can potentially cleave. Trie construction and trie traversals, both linear complexity processes, can therefore determine the uniqueness and off-target space of a given sgRNA. The trie data structure represents a solution to the sgRNA specificity problem.

GuideScan Algorithm:

Prior to GuideScan, more methods suffered from several restrictions including a.) generating sgRNAs strictly from a reference genome b.) providing sgRNAs only for the Cas9 CRISPR system c.) accounting for multiple alternative PAM sites d.) incomplete identification of all potential off-target cut sites and e.) determining both cutting efficiency and cutting specificity. These restrictions ultimately limit the use of CRISPR systems in both research and clinical settings. The GuideScan algorithm was designed specifically to address these limitations[123] (Figure 12).

```
---GuideScan Algorithm---

for PAM in genome:
    compute complementary sequence from PAM
    write complementary sequence to a file A

for complementary sequence in file A:
    add complementary sequence to trie

for complementary sequence in file A:
    if complementary sequence has canonical PAM:
        if complementary sequence occurs > 1:
            continue
        else:
            traverse trie and assess Hamming distance h
            if complementary sequence has any mismatch neighbor <= Hamming distance:
                continue
            else:
                write to file B
    else:
        continue

for complementary sequence in file B:
    traverse trie and assess Hamming distance q
    write out data to SAM file

convert SAM file to BAM file
index BAM file
```

Figure 12: GuideScan algorithm

GuideScan is a software package that robustly and accurately determines the specificity of all sgRNAs present in a genome as well as constructs a database of sgRNAs that conform to prescribed uniqueness standards detailed by a user. GuideScan requires only a single input, which is a FASTA file. This FASTA file can come from any organism or disease condition. Furthermore, GuideScan allows for the customization of output

through the use of twenty-three parameters all of which possess default values for the generation of a Cas9 database.

GuideScan begins by reading in a FASTA file and scanning for PAM sequences. The user specifies the identity of canonical PAM sequences that represent intended targets. Additionally, the user may also specify the identity of alternative PAM sequences that represent potential off-target sites. The position of the PAM sequence, either 3' or 5' to the sgRNA sequence recognition site, can also be determined by the user (Figure 13a).

When GuideScan detects either a canonical or alternative PAM site it computes the target sequence of this site by selecting a series of k sequences either upstream or downstream of the PAM (depends on the position of the PAM sequence to the sgRNA sequence recognition site). The value of k is also a parameter that can be specified by the user. Once these target site sequences are computed they are written to a file (Figure 13b).

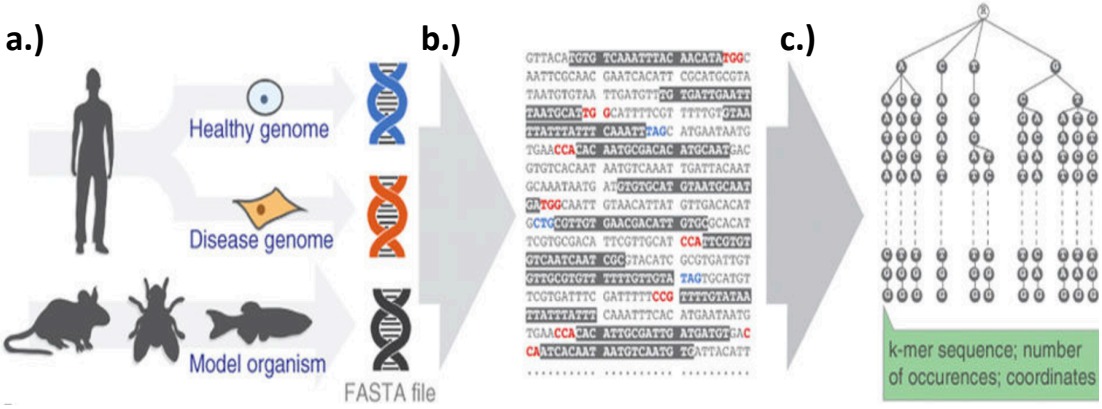


Figure 13: Workflow of GuideScan algorithm. **a.)** A user selects a genomic sequence file in the format of a FASTA file as input for GuideScan. This FASTA file is arbitrary in nature and can represent the genomes of various model organisms, genomes reflective of disease states, or wild-type genomes. **b.)** The FASTA file is scanned for canonical PAM sequences associated with desired sgRNAs in addition to searching for alternative PAM sequences that represent potential off-target cut sites. The target sites associated with each PAM sequence are extracted from the genome and are stored in an output file. The PAM sequence scanning enumerates the target space of the genome. **c.)** The target sites from the output file are used to construct a trie. If a target site occurs more than once, the amount of times it occurs in the genome is stored in the leaf node of the trie along with the genomic coordinate information associated with the target site. Traversing the trie for sequences similar, within h mismatches, to a query sequence determines the sequence's mismatch neighborhood.

Figure 13: GuideScan workflow

Once all possible target sites are determined and written to a file, the contents of the file are used in the construction of a trie. Each branch of the trie represents a potential target site and has a leaf node that enumerates how often the sequence occurs in the genome. As mentioned earlier, the construction of the trie from target sequences determines the uniqueness of every target site in the genome (Figure 13c).

However, the trie data structure allows for a more rigid definition of uniqueness. Trie traversals allow for the assessment of a Hamming distance between an input sequence and the constituent sequences of the trie. These traversals can determine if a sequence has any mismatch neighbors within a distance h. This ability of a trie allows a user to enforce a higher level of uniqueness on the sgRNAs by enforcing the usage of only those sequences that have no mismatches neighbors within a given distance h. In other words, trie traversals allow a user to determine which subset of sgRNA sequences are uniquely occurring in the genome up to h mismatches. The sgRNA sequences that meet this enhanced level of specificity are then written to a file. The distance h is another parameter determined by the user.

Once the set of sgRNAs unique up to h mismatches (S_h) is determined, the algorithm allows a user to enumerate the potential off-targets for each sgRNA in set S_h up to q mismatches (q > h). The rationale behind this is that a user may wish to automatically exclude those sgRNAs that have near off-targets ($\leq h\ mismatches$), but still desires to have sufficient sgRNAs to broadly and precisely target a genome. However, the higher the value of h, the fewer sgRNAs will be available. Reducing the size of set S_h functionally reduces the size of the targetable genome, which reduces the resolution on which the genome can be edited. Rather than arbitrarily reducing the size of S_h by increasing the value of h, a user can choose a value of h that makes S_h exclude those sgRNAs most likely to represent troublesome off-targets (off-targets with a Hamming distance of one for example). The remaining sgRNAs can then have their off-targets enumerated up to a distance q and this additional off-target information can be used in making the final sgRNA

selection. By construction the elements of S_h will have no off-targets at $\leq h$ mismatches and can only have off-targets at p mismatches where $h < p \leq q$. As when sgRNA uniqueness was assessed, the enumeration of off-targets up to q mismatches is done through taking the sgRNAs in S_h and traversing them through the trie, enumerating mismatch neighborhoods up to Hamming distance q. This process satisfies the need to have numerous precise sgRNAs since $S_h > S_q$ but a user has the knowledge of a sgRNA's target space up to q mismatches.

As the sgRNA uniqueness or off-target information is computed, it is written out to a Sequence Alignment Map (SAM) file. This file will contain the standard SAM fields in addition to three additional fields. Theses added fields detail the off-target distance assessment (q), the maximum amount of off-targets recorded in the file for a given sgRNA (parameter selected by the user), and the coordinates of potential off-target cut sites. To reduce the size of the SAM file, the off-target information is recorded in the SAM file as a hex-byte array. For Cas9 databases two further fields may be recorded to each sgRNA in the SAM file: a cutting efficiency score and a specificity score.

Once the SAM file is generated, GuideScan utilizes Samtools to a create a Binary Alignment Map (BAM) file and index file that both reduces the size of the file and accelerates the search of the database file by genomic coordinate[146]. The index increases the speed of database query from a linear process to a logarithmic process[143].

$$O(n) = \ Complexity\ of\ Linear\ Database\ Search$$

$$O(\log n) = Complexity\ of\ Index\ Based\ Database\ Search$$

Consequently, GuideScan creates CRISPR databases and determines sgRNA specificity in linear time and allows for the lookup of sgRNAs by coordinate in logarithmic time.

Cas9 sgRNA Cutting Efficiency Score:

If a GuideScan database is generated for the Cas9 CRISPR system, the BAM file can be further enriched with cutting efficiency scores for each sgRNA in the database. GuideScan utilizes the Rule Set 2 cutting efficiency score, which quantifies how likely a given sgRNA is to efficiency cut it's target site[124]. The score is developed from a learned boosted regression tree model that computes a cutting efficiency score from a myriad of factors including sequence features.

Rule Set 2 scores are defined only for Cas9 sgRNAs with a twenty base pair complementary sequence and the NGG PAM. Consequently, if a user desires to enrich a GuideScan database with Rule Set 2 scores, the algorithm must first ensure that the database contains sgRNAs that conform to the requirements of Rule Set 2. GuideScan achieves this by checking the header sequence in the GuideScan database BAM file and searches specifically for the values associated with canonical PAM sequence and length of the sgRNA's complementary sequence. If, and only if, both values conform to Rule Set 2 requirements will the score be computed for each sgRNA in the database.

Specifically, for each line in the BAM file, GuideScan will inspect the 5' coordinate and strand sequence present in the file and compute the coordinates for a thirty base sequence region (Figure 14).

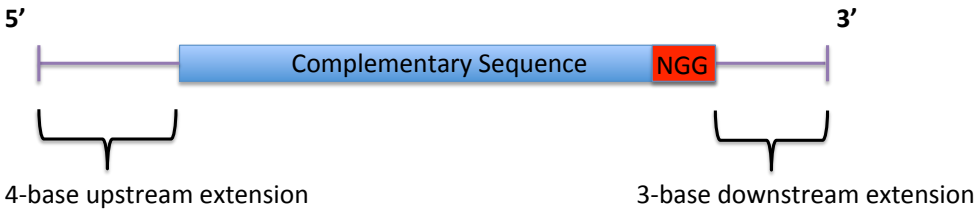


Figure 14: Rule Set 2 on-target cutting efficiency score sequence requirement.

Figure 14: Rule Set 2 sequence requirement

While sgRNAs in the GuideScan database have a complementary region of twenty bases, Rule Set 2 requires sequences of thirty bases to compute a cutting efficiency score, and hence why a thirty base lookup is done. Once the coordinates for the thirty bases are determined, querying an indexed version of the input FASTA file retrieves the sequence itself. The Rule Set 2 model then processes the sequence and the score is recorded as an additional field in the BAM file.

Cas9 sgRNA Specificity Score:

An additional metric available to GuideScan Cas9 databases is a specificity score. This score quantifies how likely a sgRNA is to cut only an intended target site given information about it's off-targets up to a Hamming distance of q. The specificity score uses the cutting frequency determination (CFD) mismatch matrix to compute the likelihood that a given off-target site, dissimilar to the sgRNA complementary sequence by $\leq$ q mismatches, will be cut[124] (Figure 15).

The CFD mismatch matrix is composed of empirically determined values that represent how deleterious a given mismatch, in the twenty base complementary region, is to sgRNA binding and cutting efficiency. Each position in a twenty base complementary region was assessed with every base sequence as a mismatch and the disruption of the mismatch to cutting efficiency was determined. If a degenerate target site is dissimilar to a sgRNA complementary sequence by one mismatch then the CFD score is the empirical value of the mismatch position in the matrix. If more than one mismatch exists, then the CFD score is the product of the values for each mismatch. The CFD score can thereby determine a value for any potential off-target site with any number of mismatches between the degenerate target site and sgRNA complementary sequence.

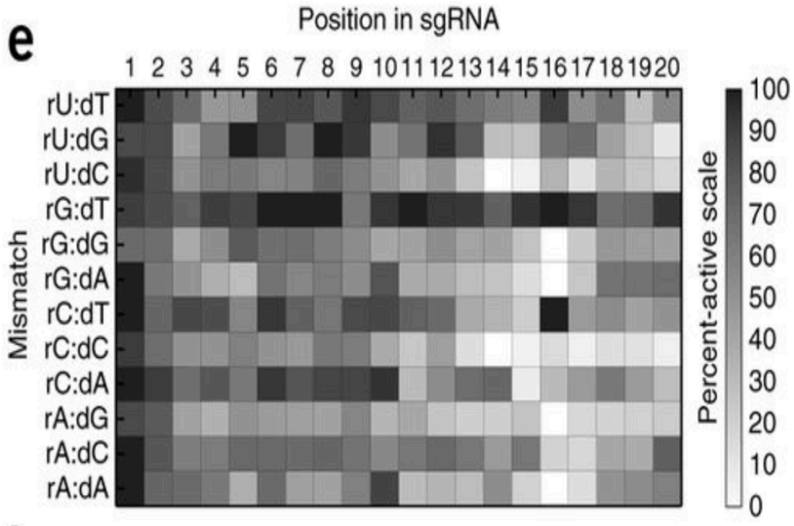


Figure 15: Cutting Frequency Determination (CFD) mismatch cutting activity matrix. This matrix quantifies how deleterious a positional mismatch is between a sgRNA and a target site. (Doench et al. 2016)

Figure 15: Cutting frequency determination matrix

As with the Rule Set 2 score, the CFD mismatch matrix is defined only for Cas9 sgRNAs with a twenty base pair complementary sequence and the NGG PAM. Consequently, before specificity scores can be computed, the GuideScan database BAM file will be assessed to verify it meets the requirements. If, and only if, the database conforms to CFD scoring requirements will GuideScan compute a specificity score for a sgRNA.

The CFD score computes the off-target likelihood of cutting for a single degenerate target site. GuideScan computes a specificity score by looking up the sequence associated with each off-target determined out to q mismatches and determining the CFD score by comparing the off-target sequence against the sgRNA complementary sequence. GuideScan then takes the CFD score and multiplies it by the amount of times the off-target sequence occurs in the genome. The resulting value is then aggregated as a denominator and a composite specificity score for a sgRNA is determined.

$$n = represents\ the\ unique\ targetable\ sites\ within\ up\ to\ z\ mismatches$$

$$z_i = number\ of\ times\ the\ i^{th}\ neighbor\ occurs\ in\ the\ genome$$

$$CFD_i = \ CFD\ score\ for\ the\ i^{th}\ unique\ targetable\ site$$

$$Specificity\ Score = \frac{1}{\sum_{i=1}^{n} CFD_i * z_i}$$

In this manner, GuideScan determines a specificity score for each sgRNA in the database by evaluating its entire mismatch neighborhood. The exhaustive enumeration of mismatch neighbors makes the specificity score reflective of a sgRNA's possible target space. The specificity score for each sgRNA is recorded as an additional field in the GuideScan database BAM file.

Genomic Feature Annotation:

When choosing sgRNAs, a researcher may be interested in selecting only those sgRNAs that cut within specific genomic features. For example an experimentalist may be using sgRNAs to disrupt the reading frame of a protein coding sequence in which case they would want to choose sgRNAs that target exons. Whether it is exons or another genomic feature, GuideScan allows for the annotation of sgRNAs that overlap any arbitrary feature through its use of interval trees (appendix).

Interval trees are data structures that allow for the efficient storage and query of a set of overlapping intervals. While the details of an interval tree are discussed in the appendix, for the purposes of understanding their usage in annotation it is sufficient to state that these structures allows for the discovery of all overlapping intervals at a single interval (Figure 16). Furthermore, they are efficient in both their construction and query[147].

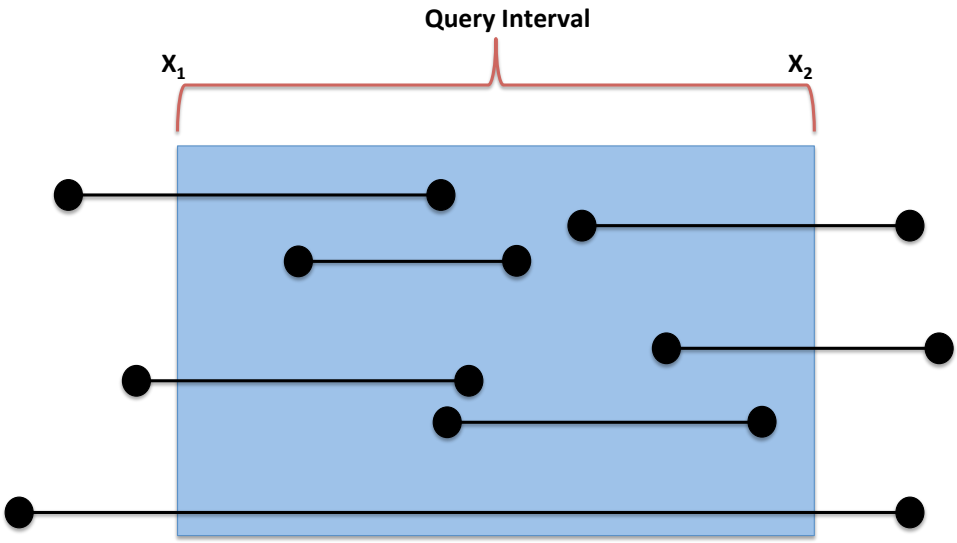


Figure 16: Overlapping segments within a query interval. The query of a single interval [X_1,X_2] will return all segments that intersect this interval. This basic principle is the underlying logic behind interval trees.

Figure 16: Interval tree conceptual diagram

$$n = nodes\ in\ tree, k = number\ of\ intervals\ output$$

$$O(n\ log\ n) = \ Complexity\ of\ Interval\ Tree\ Construction$$

$$O(\log n + k) = \ Complexity\ of\ Interval\ Tree\ Search$$

A practical example of the use of interval trees for annotation purposes is to return to the problem of selecting sgRNAs that overlap exons. Imagine a researcher is conducting a loss of function experiment with sgRNA's targeting a given protein. This protein has several isoforms that use different exons, however some exons are expressed across all isoforms. Querying an interval tree, constructed from the coordinates of all the exons in the genome, allows the investigator to see which sgRNAs overlap exons expressed across all isoforms. In this manner, a researcher can choose a single sgRNA that will effectively cut, and hopefully disrupt, all isoforms of the protein leading to an effective loss of function result.

GuideScan will automatically construct interval trees to annotate sgRNAs if it is delivered a Browser Extensible Data (BED) format file when a user queries a GuideScan database. The annotation is done at time of database query to allow a user the maximum flexibility in determining what annotations they would like to apply to a GuideScan database. The annotation is displayed as a field in the output that results when a GuideScan database is queried.

GuideScan Database Query Output Options:

GuideScan allows users to extract sgRNAs from a database according to several selection parameters. To start, a researcher can have GuideScan return sgRNAs from *within* a target site or from the regions *flanking* a target site. The user defines the size of these flanking regions. Additionally, if the user requests sgRNAs flanking a target region, and a pair of flanking sgRNAs exist, then GuideScan automatically will generate an oligonucleotide that has both sgRNA complementary regions cloned in, provided they meet the sequence requirements described in the Vidigal & Ventura dual sgRNA delivery system[148].

The selected sgRNAs can be sorted according to at least two, and at most four, parameters (depends if database is for Cas9). The number of enumerated off-targets up to distance q can sort all queries. In this process those sgRNAs with the least total number of off-targets appear first in the output while those with the most appear last. Additionally, sgRNAs can be sorted by their proximity to a target site. For *within* queries, sgRNAs are sorted with those appearing closest to the 5' end, as seen from the positive strand, appearing first and those nearest the 3' end appearing last. For *flanking* queries GuideScan output appears in two parts: sgRNAs upstream of the 5' coordinate and sgRNAs downstream of the 3' coordinate. For a flanking query sorted by proximity to target site, the sgRNAs upstream of the 5' coordinate are sorted with those sgRNAs closest to the 5' coordinate listed first and those furthest away last. Likewise for the sgRNAs downstream of the 3' coordinate the sgRNAs closest to the 3' coordinate are listed first and those

furthest away last. Additionally, if the database is composed of Cas9 sgRNAs, and cutting efficiency scores and specificity scores are included in the database, then sgRNAs can be sorted by either score. For both cutting efficiency and specificity, for either within or flanking queries, the sgRNAs are sorted with those sgRNAs with the highest score listed first and those with the lowest scores listed last.

Furthermore, GuideScan is capable of selecting the top m sgRNAs for a user, for either within or flanking queries, by using a double sort method utilizing two selection parameters. The user sets the value of m. As an example, should an investigator choose to have GuideScan select the top m sgRNAs prioritizing the off-target option then all the sgRNAs for a queried region will be sorted according to their off-targets values, after which the top m will be selected for a second sort. In this second sort, if a cutting efficiency score is present in the database then these m sgRNAs will be resorted by efficiency score, otherwise they will be resorted by proximity to query boundary. A full detailing of the double sort method for GuideScan sgRNA selection is detailed in the following table (Table 3).

Table 3: GuideScan double sort selection

Sort Option	First Sort	First Sort (Cas9)	Second Sort	Second Sort (Cas9)
Fewest Off-targets	Fewest Off-targets	NA	Coordinates Closest to Query Boundary	Cutting Efficiency Score
Coordinates Closest to Query Boundary	Coordinates Closest to Query Boundary	NA	Fewest Off-targets	NA
Cutting Efficiency Score	NA	Cutting Efficiency Score	Fewest Off-targets	NA
Specificity	NA	Specificity	Coordinates Closest to Query Boundary	Cutting Efficiency Score

In addition to the various ways in which GuideScan can display its output, it also is capable of handling distinct input formats. GuideScan databases are ultimately organized and queried by genomic coordinates; however, the software was designed to handle four distinct forms of query, which serves to maximize its utility to researchers.

First, GuideScan supports batch queries that allow a user to specify a set of genomic coordinates that GuideScan reads in and extracts sgRNAs for. Batch query output is displayed according to user determined output parameters. Specifically, batch queries take files as input. These files are either of the Gene Transfer Format/General Feature Format (GTF/GFF), text, or BED formats. If the file is of the text of GTF/GFF format then arbitrary unique identifiers for each sgRNA will be assigned to each sgRNA. However,

if the file is a BED file then the fourth field potentially can serve as a unique identifier for output sgRNAs. If the file is a BED file GuideScan determines if any multiplicity exists in the fourth column and if none exists, it uses the values of the fourth field for creating unique identifiers. However, if even one instance of multiplicity exists then arbitrary labeling will be enacted.

The second type of query GuideScan supports is the direct genomic coordinate query. A user can type in a genomic coordinate for their region of interest and set their output parameters to receive sgRNAs for their region of interest. These sgRNAs will have arbitrarily generated unique identifiers.

The third type of query allowed by GuideScan is query by genomic feature. In this format a user provides a BED file where the fourth field is composed of unique feature identifiers. For example, imagine a BED file in which the fourth field consists of gene names and the first three fields compose the chromosome, start, and end coordinates of the genes. GuideScan will create a dictionary data structure with the elements of the fourth field as keys and genomic coordinate as values. The user can then specify, as either a direct query or a batch query, the feature name and get sgRNAs for that feature with GuideScan processing the request as a coordinate query in the background. These sgRNAs will have arbitrarily generated unique identifiers.

The last query permitted by GuideScan is query by sequence. This query format requires a user to specify the indexed FASTA file that was used as input for the generation of the GuideScan database. Additionally, the user must give a FASTA file, composed of the sequences of interest, to GuideScan as query input. GuideScan will then use BLAT to locally align the sequences to the indexed FASTA file. If, and only if, a perfect sequence match is found then the output of BLAT will be converted into coordinates and used by GuideScan to extract sgRNAs for the determined region [149].

GuideScan Database sgRNA Density and Target Resolution:

The final output of the GuideScan algorithm is a sgRNA database in a BAM file format. The sgRNA's in this database are guaranteed to be unique up to h mismatches, which means that not all possible sgRNAs will be included in S_h. To ensure that the sgRNA database still has a density of sgRNAs that allow it to precisely edit any arbitrary genomic locus, a GuideScan database for the mm10 assembly of the mouse genome was generated with sgRNAs unique up to two mismatches. To determine sgRNA density, the mouse genome was binned into fifty kilobase regions and the quantity of sgRNAs per bin was determined using both the mm10 GuideScan database and the only other genome-wide database of sgRNAs available at the time (Hsu database)[129] (Figure 17).

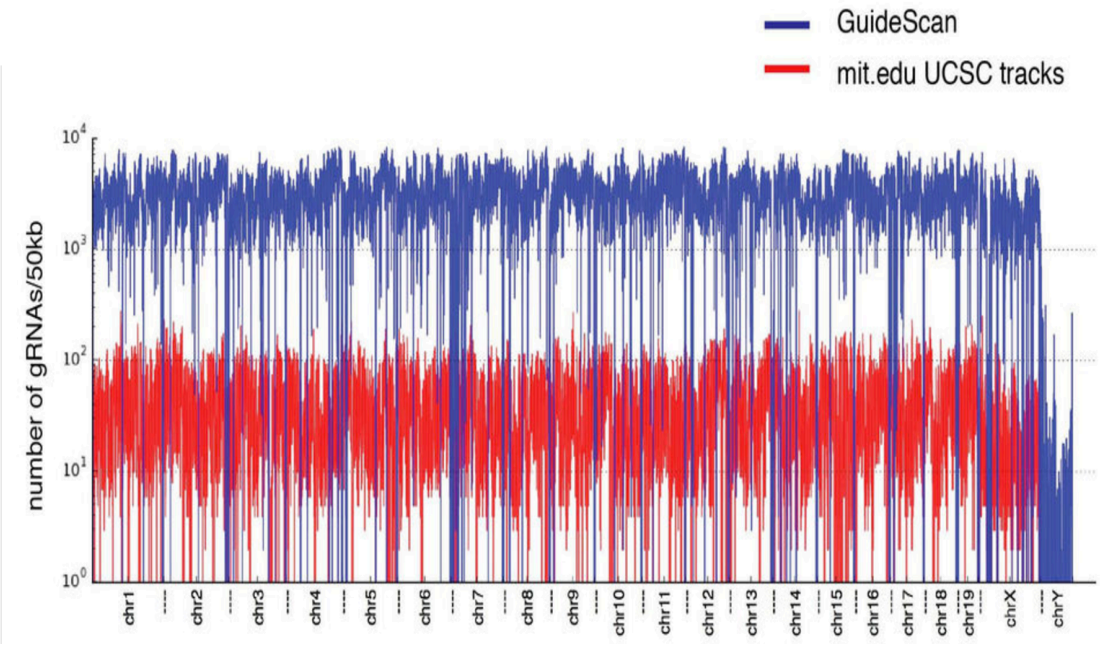


Figure 17: sgRNA density per 50 kilobase region in the mm10 assembly of the mouse genome. The GuideScan database was filtered so sgRNAs were guaranteed to be unique up to two mismatches. The mit.edu UCSC tracks were unfiltered and their density was computed using the entire contents of the track.

Figure 17: GuideScan mm10 sgRNA density

This comparison showed that the filtering of sgRNAs for a uniqueness of up to two mismatches did not profoundly impact the density of sgRNAs in the genome.

To demonstrate that GuideScan sgRNA density translated to a better cutting resolution, sgRNAs were designed against non-coding features including all enhancers, microRNAs (miRNA), long non-coding RNAs (lncRNAs), and CTCF sites in the mm10 assembly of the mouse genome[150–153]. Flanking sgRNAs were designed against each genomic feature and the distance between the nearest sgRNA and the 5' or 3' genomic coordinate defining the feature was determined. Once again the GuideScan and Hsu databases were used for this comparison (Figure 18).

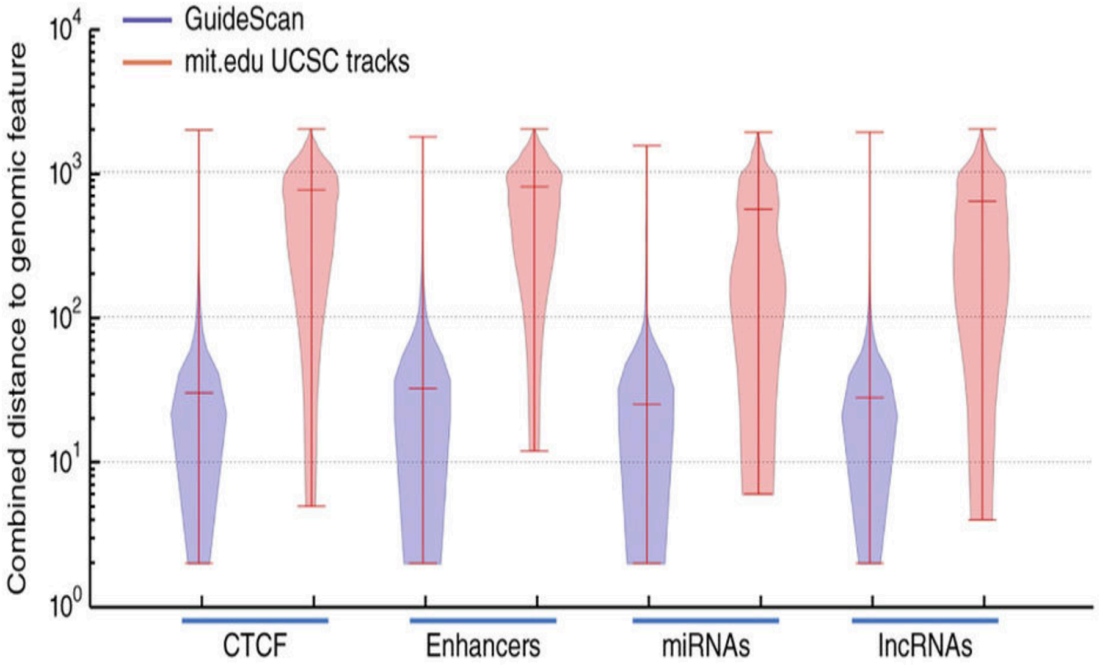


Figure 18: Combined flanking distance for all CTCF site, Enhancers, miRNAs, and lncRNAs in the mm10 assembly of the mouse genome were computed using a mm10 GuideScan database where all sgRNAs were unique up to two mismatches and the mit.edu UCSC tracks.

Figure 18: Target flanking distance for non-coding elements

The resulting analysis showed that sgRNAs could be designed against enhancers, miRNAs, lncRNAs and CTCF sites with a combined flanking distance of 31, 24, 27, and 29 base pairs respectively with the GuideScan database. For the Hsu database the combined flanking distance was 783,

716, 774, and 781 base pairs respectively. Consequently, this comparison demonstrated that GuideScan's more rigid definition of uniqueness not only does not functionally affect sgRNA density, but it also has modest effect on target resolution.

To ensure that sgRNAs designed against these target sites were functional, one miRNA cluster and one enhancer site in the mouse genome were chosen at random for deletion by paired sgRNAs. Both of these sites had paired sgRNAs designed against them and were sequenced to verify the presence of the deletion (Figure 19a,b). Overall, the requirement that GuideScan databases contain sgRNAs unique up to h (where h is reasonably low) mismatches produced databases with excellent density, resolution, and activity of sgRNAs.

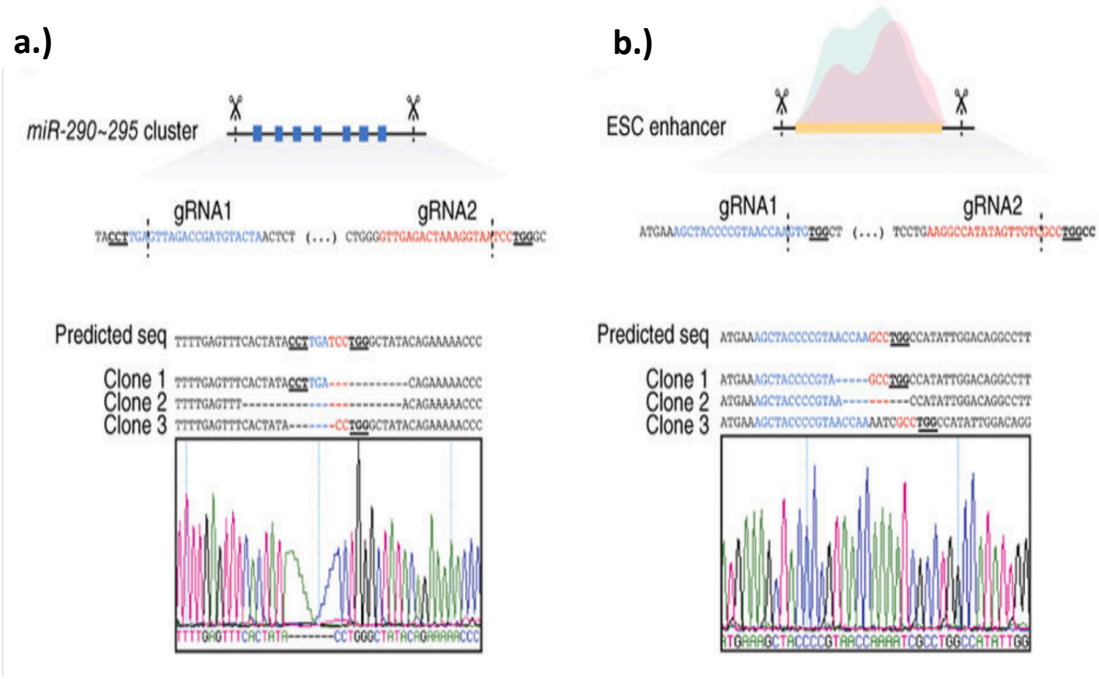


Figure 19: Example deletions of genomic regions containing RNA (**a)** and DNA **(b)** non-coding elements using pairs of sgRNAs designed by GuideScan. sgRNA sequences, blue and red; PAM sequences, bold underlined. The predicted sequence after deletion, the sequences after three edited alleles, and a representative chromatogram are shown for each target locus. [19]

Figure 19: Example deletions of miRNA cluster and enhancer

GuideScan Database sgRNA Specificity:

GuideScan databases have sgRNAs filtered for a user-defined level of uniqueness. However, it remained to be seen whether this augmented level of

uniqueness translates to higher sgRNA specificity. To make this determination genomic coordinates of fifty random mouse (mm10) protein coding genes, non-coding elements, and repeat masked regions were chosen for sgRNA design by three widely used tools: mit.edu, CRISPRScan, and E-CRISP in addition to GuideScan [126,127,129]. All three tools required sequence input for sgRNA design; therefore genomic coordinates were translated into sequences and delivered as input to the methods. The sgRNAs outputted by each method were then assessed for mismatch neighbors by traversing them through the mm10 trie and determining the identity of mismatch neighbors up to a Hamming distance of two. This assessment showed that all methods, except GuideScan, produced sgRNAs that either had multiple perfect sequence match target sites or target sites dissimilar by only one mismatch (Figure 20). These sites represented off-targets with extremely high cutting potential[116].

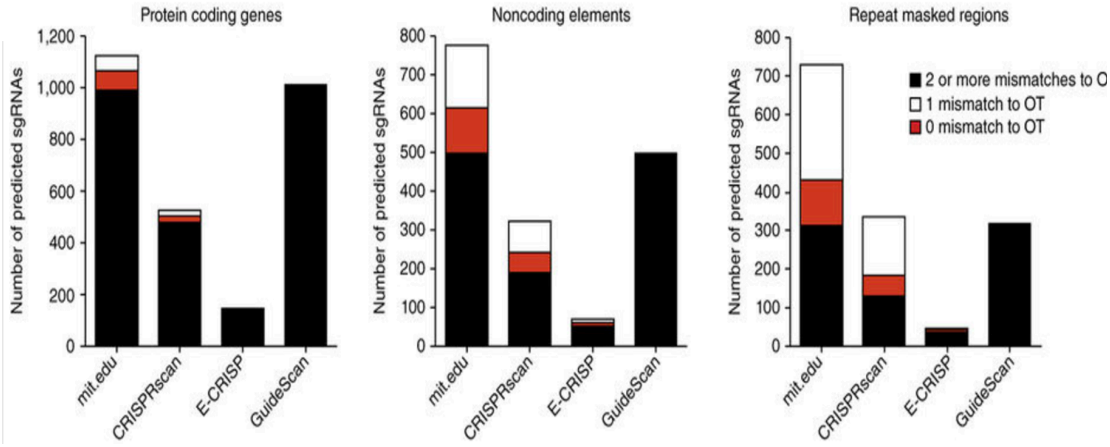


Figure 20: Number of mouse sgRNAs designed against fifty random protein coding genes, fifty random non-coding elements, and fifty random repeat masked regions. All tools but GuideScan delivered sgRNAs that contained multiple perfect sequence match target sites and degenerate target sites different from the intended target site by only one.

Figure 20: Tool comparison with single and perfect sequence match target sites

Multiple perfect target site matches represent a particularly troublesome off-target for CRISPR usage. To investigate how many times perfect target site matches occurred in the sgRNAs outputted by competing methods, a strict enumeration of the target site occurrence in the genome was computed for each duplicitous sgRNA. This showed that some sgRNAs

delivered by these methods occurred in the genome tens of thousands of times (Figure 21).

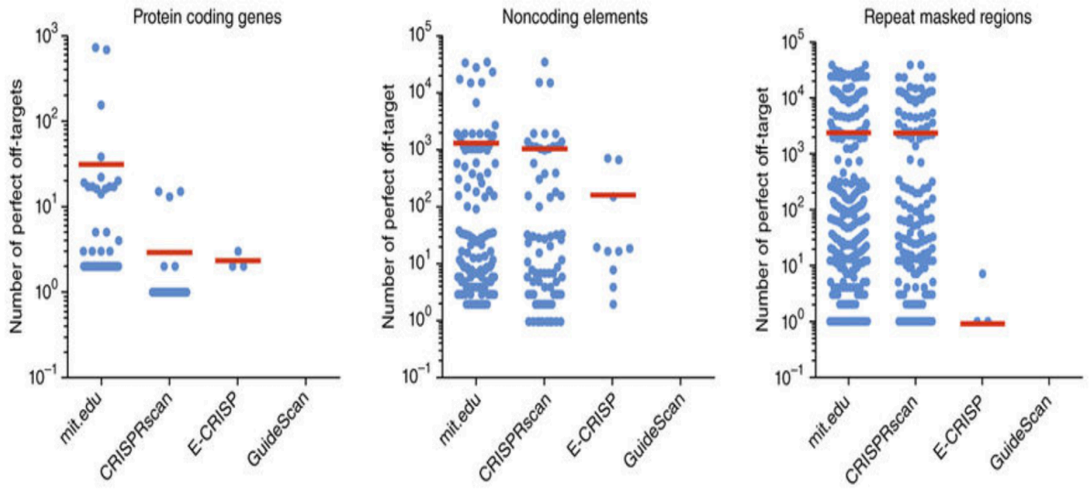


Figure 21: sgRNAs designed against fifty random protein coding genes, fifty random non-coding elements, and fifty random repeat masked regions with multiple perfect sequence match target sites. All tools but GuideScan returned sgRNAs that had sgRNAs that were not uniquely occurring in the mm10 assembly of the mouse genome.

Figure 21: Quantity of perfect sequence match target sites

Such a magnitude of perfect target site matches in the genome makes the usage of such sgRNAs worrisome given that the abundance of target sites would likely lead to prolific DSBs and potentially compromise the survival of the cell. This effect could be particularly worrisome when using these sgRNAs as part of a negative or positive selection screen.

Additionally, many selection methods assign a specificity score to their sgRNAs that takes into account off-target information. A particularly popular method was the web interface tool from MIT (henceforth termed mit.edu). This method assigned a specificity score based on a number ranging from one to one hundred with one indicating highly nonspecific and one hundred indicating highly specific. Additionally, these values took on a color score that illustrated the mit.edu's assessment for the specificity of the sgRNA. Green represented sgRNAs that were most specific to their target, yellow represented sgRNAs that were somewhat specific to their target, and red represented sgRNAs that were likely non-specific to their target. Ideally those sgRNAs with multiple perfect target site matches would be assigned a low

specificity score since they possess multiple exact target sites in a genome. However, when the mit.edu specificity scores were compared against the amount of perfect match target sites no correlation was observed (Figure 22).

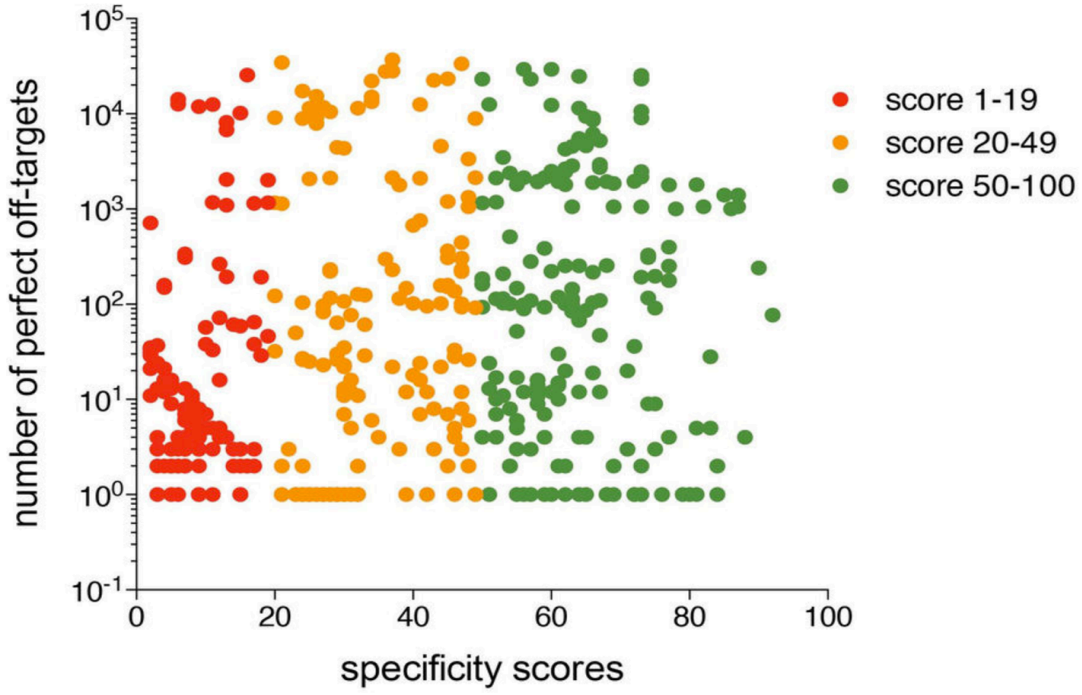


Figure 22: Dot plot showing specificity scores and number of perfect sequence match target sites for sgRNAs designed by mit.edu. Red dots indicate low specificity, yellow dots medium specificity, and green dots indicate high specificity.

Figure 22: mit.edu specificity score

Furthermore, upon closer analysis of the mit.edu tool's output it was noted that many sgRNAs with multiple perfect target site matches were being reported as uniquely occurring in the mouse genome. This result was worrisome because sgRNAs will target and cut perfect sequence matches with approximately equal efficiency (Figure 23).

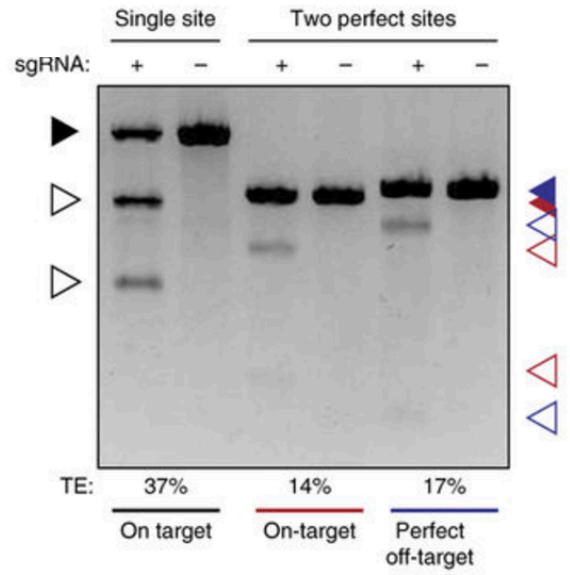


Figure 23: T7 cleavage assay for sgRNAs having single (black, on-target) or multiple (red, blue, on-target) perfect matches in genome. Position of cleavage substrates (filled triangles), position of cleavage product (open triangles).

Figure 23: Cleavage assay of perfect sequence match target sites

Furthermore, if a sgRNA is described as being uniquely occurring, and in reality is not, then multiple genetic lesions can occur unbeknownst to the experimenter that potentially compromises the interpretation of the experiment.

CRISPR systems induce DSB in DNA at a target site. If two DSB occur in a chromosome an inversion or a deletion can result. Furthermore, should two DSBs occur on separate chromosomes then translocations between the chromosomes can occur. In reality, however, when multiple DSB occur across or within multiple chromosomes then multiple translocations and inversions will occur. These alterations, if unintended, pose tremendous

difficulty in interpreting the result of a CRISPR experiment. This is especially true for chromosomal translocations that are known to be oncogenic in several well-described cases[154–156] (Figure 24). We investigated a sgRNA from the mit.edu tool that was termed to be highly specific and uniquely occurring in the genome, but which had three perfect sequence match target sites in the genome. When we transfected this sgRNA we observed cutting at all three target sites and the generation of translocations. Such a result is undesirable to an investigator.

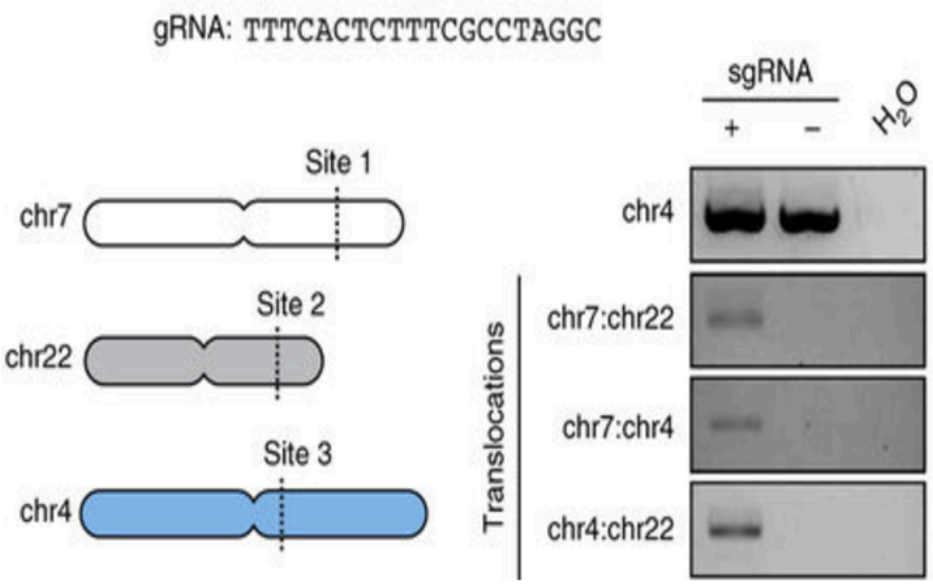


Figure 24: Left, schematic representation of chromosomal translocation. Right, PCR-based identification of chromosomal translocations between perfect sequence match target sites. + sgRNA, – empty vector. sgRNA is marked as highly specific by mit.edu (score = 78)

Figure 24: Undesired translocations

Multiple perfect sequence match target sites can also pose a problem even if the target sites are not located distantly from another. If CRISPR systems are being used to engineer a specific genomic alteration at a given locus with the HDR mechanism, then it is imperative that a target site be unique. However, if sgRNAs are designed against a region that has locally repetitive sequences then the genetic alteration will only occur in a minority of cells. Furthermore, the readout of the experiment will give non-descript bands that reveal the multi-cut nature of the sgRNA. Again, we investigated a

sgRNA from the mit.edu method that was stated to be specific and uniquely targeting in the genome, but in reality had several local perfect match target sites. In transfecting this sgRNA we observed the creation of non-descript bands revealing the multiple cutting tendency of the sgRNA (Figure 25). These repetitive local cut sites, again, can be undesirable to a researcher.

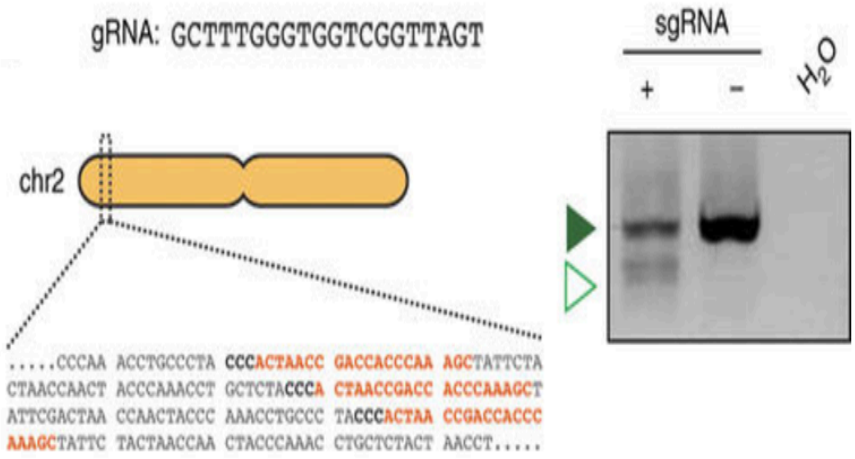


Figure 25: Left, schematic representation of the chromosomal locations of three perfect sequence match target sites , all within chromosome two, of a sgRNA labeled highly specific by mit.edu (score = 89). Genomic sequence: target sites, red; PAM sequence, bold. Right, PCR-based identification of chromosomal deletions between target sites. Positions of the wild-type amplicon, filled triangle; position of deletion amplicon, open triangle. +, sgRNA; -, empty plasmid.

Figure 25: Undesired local target sites

While specific examples demonstrated the unintended effects of non-unique sgRNAs, it remained to be determined if the higher standard of uniqueness GuideScan enforces on its sgRNAs translated globally into increased sgRNA specificity. To determine if GuideScan sgRNAs as a whole were more specific than the sgRNAs returned by mit.edu, CRISPRScan, and E-CRISP, specificity scores were computed for all the sgRNAs outputted by the tools. The distributions of specificity scores were then assessed and the differences between populations were statistically evaluated using the Kolmogorov-Smirnov test (Figure 26). As a population GuideScan sgRNAs were significantly more specific to their target sites than sgRNAs from either

mit.edu or CRISPRScan ($p < 2.2 \times 10^{-16}$). The sgRNAs from E-CRISP were not significantly more specific than the sgRNAs from GuideScan; however, the amount of sgRNAs delivered by E-CRISP was an order of magnitude less than the amount of sgRNAs returned by GuideScan. Furthermore, output by E-CRISP still contained sgRNAs that had multiple perfect sequence match target sites as well as target sites distinct from a sgRNA complementary region by only one mismatch, albeit less frequently than mit.edu or CRISPRScan. Consequently, GuideScan was the only tool that delivered completely unique sgRNAs, which were significantly more specific than mit.edu and CRISPRScan and an order of magnitude more numerous than E-CRISP.

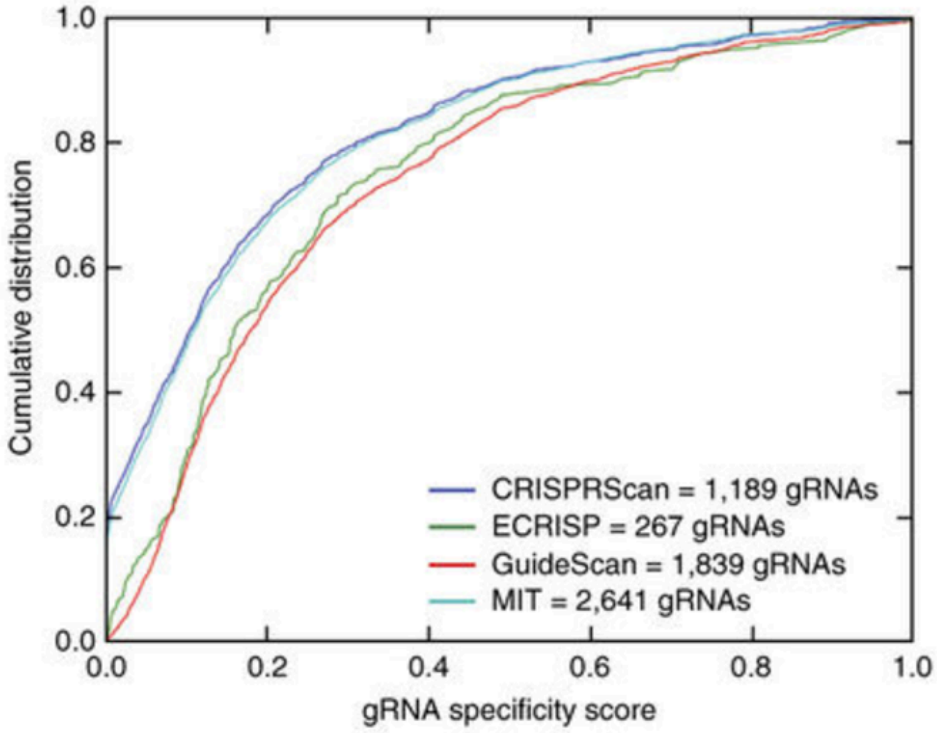


Figure 26: Cumulative distribution plot of specificity scores for each of the sgRNAs designed by each tool.

Figure 26: GuideScan specificity score

Competitor Methods Off-Target Search:

CRISPRScan, mit.edu, and E-CRISP are conceptually robust and ingeniously designed tools, yet they fail to relay the complete target space information about a given sgRNA within a set number of mismatches. As a

consequence, these methods will overestimate the specificity of sgRNAs and occasionally miss off-targets with high cutting potential. It appears the reason for this is that all three methods rely on genome aligners for their off-target search. As previously discussed, these tools are excellent for aligning high throughput sequencing data, but are not optimized for exhaustive alignment mapping. These tools utilize heuristics to determine mapping sites in an efficient and rapid manner. However, these heuristics are known to give false negatives. As a result, when genome aligners are used to determine the off-targets of a sgRNA, only occasionally will the complete target space be enumerated by the aligner. Unfortunately, in a system that enacts permanent changes in a genome such as CRISPR, the complete knowledge of a target space is essential. It is the enumeration of all genomic target sites coupled with the exhaustive determination of sgRNA mismatch neighborhoods by the trie data structure that gives GuideScan it's advantage over aligner based methods.

GuideScan Command Line Tool:

The GuideScan algorithm was designed as a python software package that can be installed system-wide on a machine. The package can generate and query databases from the command line and creates its own output directory with required intermediate and final output files. The software can be downloaded from a public repository at https://bitbucket.org/arp2012/guidescan_public/overview. Furthermore, the package is also available as a Docker container at https://hub.docker.com/r/xerez/guidescan/.

GuideScan Web Interface:

To better facilitate the usage of GuideScan, a web interface was created. This interface contains pre-computed Cas9 and Cpf1 GuideScan databases, for the human (hg38), mouse (mm10), zebrafish (danRer10), fruit fly (dm6), Caenorhabditis elegans (ce11), and yeast (SacCer3) genomes. These genomes can be queried by coordinate, gene symbol, sequence, text

file, BED file, or GTF file upload for all organisms with an easy to use front end. The back end of the site runs completely off the GuideScan software.

The web interface also allows for the direct download of the hosted GuideScan databases as well as directs a user to the software repository and Docker instance should an investigator want to utilize GuideScan to create their own custom CRISPR databases.

Since the launch of the GuideScan web interface on March 1, 2017 through July 11, 2017 the site has serviced 1,577 unique users with retention rate over fifty percent (Figure 27).

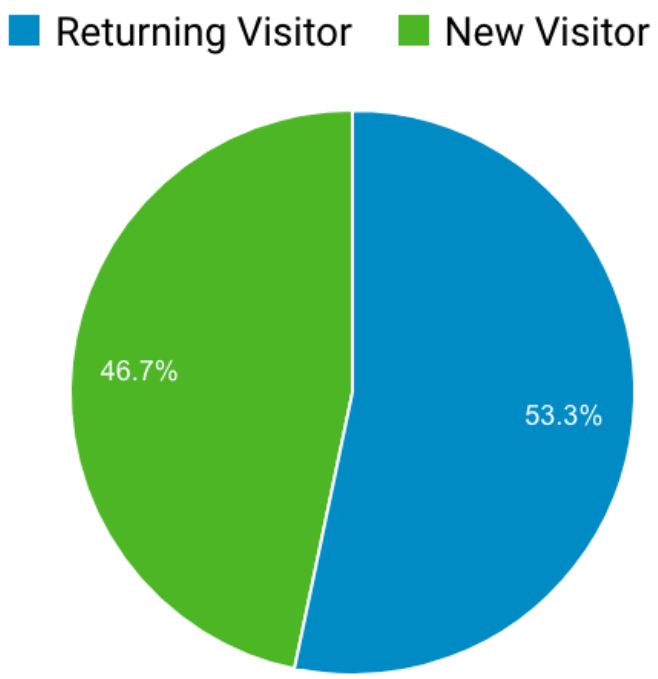


Figure 27: pie plot diagram showing site visits by familiar IP addresses and by novel IP addresses to GuideScan web interface.

Figure 27: GuideScan web interface retention

These users have generated 3,368 unique sessions and 11,187 page-views. User IP addresses originate from fifty-one distinct countries and 405 different cities (Figure 28). Site usage follows a weekly cyclic pattern with lowest activity occurring on the weekend.

The web interface can be found at www.guidescan.com. The site is constructed with the CherryPy, a python web framework. The website was co-

developed with Sagar Chhangawala and main algorithm was co-developed with Yuri Pritykin.

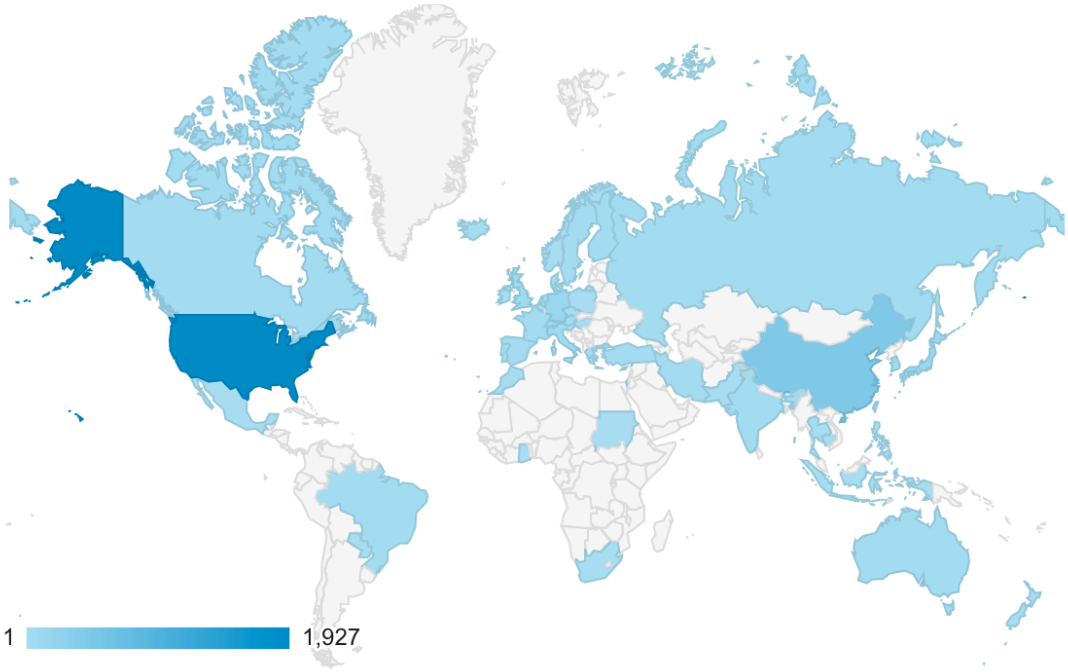


Figure 28: Map of country usage of GuideScan web interface. Darker blue indicates greater usage.

Figure 28: GuideScan web interface global access

Chapter 3: RBMX

Overview of Splicing:

The most recent assessment by Ensembl states that there are 20,412 genes in the human genome[136]. Not all of these genes encode proteins as their final gene product, with some genes encoding functional forms of RNA[157,158]. Consequently, only a subset of genes is needed to generate the 92,179 proteins that constitute the known human proteome[159]. This discrepancy between the quantity of genes and their protein products is achieved through a fundamental genetic process known as splicing.

Genes, in a broad sense, can be defined as segments of DNA that undergo transcription. Protein coding genes are those DNA sequences that, when transcribed, ultimately form messenger RNA (mRNA). These genes are composed of two distinct types of DNA sequences: those sequences that ultimately interact with a ribosome and those sequences that do not. The sections of DNA that interact with a ribosome are termed exons, while the non-interacting sections are called introns. When DNA is transcribed, introns are either concurrently or shortly thereafter removed from the resulting RNA transcript to form the mRNA[160]. The removal of introns from a RNA transcript is the general definition of splicing (Figure 29).

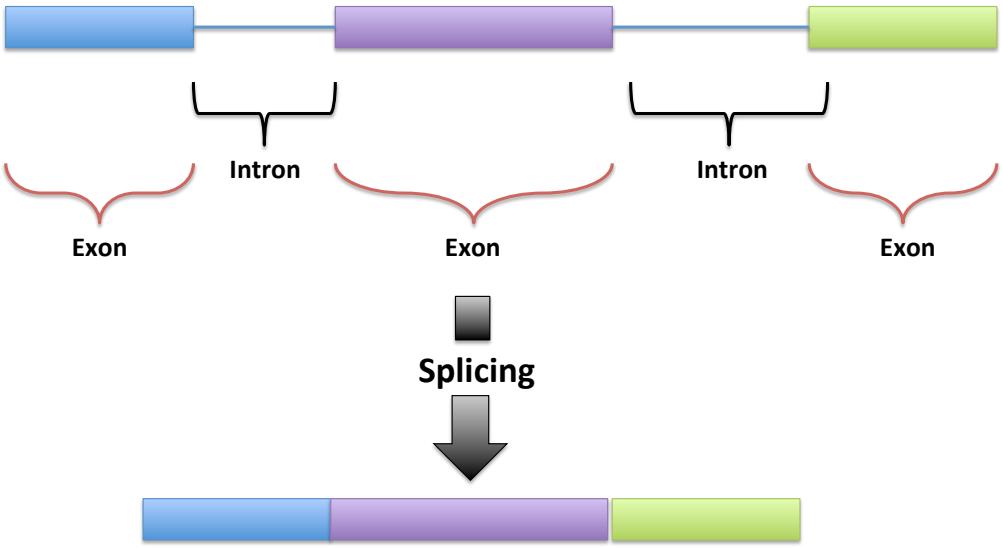


Figure 29: General diagram of splicing where all introns in a transcript are spliced out leaving only exons. While the cartoon illustrates splicing in isolation from transcription, the reality is that splicing occurs simultaneously with, or shortly after, transcription.

Figure 29: Overview of splicing

Splicing is an evolutionarily ancient process that is functionally present across all domains of life. Similar splicing machinery exists in both eukaryotes and prokaryotes, but the utilization of splicing varies between these kingdoms. While eukaryotes frequently splice protein coding RNA transcripts and to a lesser extent non-coding RNA transcripts, prokaryotes splice less often and commonly focus on non-coding RNA transcripts[161–163]. Splicing occurs in archaea as well, but it appears limited to tRNA splicing[164,165]. Interestingly, though prokaryotes and eukaryotes possess similar splicing machinery, the splicing mechanism present in archaea most closely resembles the mechanism present in eukaryotes[164]. The machinery available for splicing is most numerous in eukaryotes with this domain of life possessing at least three established pathways: spliceosomal complex, self-splicing introns, and tRNA splicing[166–169].

Among all the splicing pathways, the one that is conserved across all domains of life is tRNA splicing as it is crucial for the generation of tRNA molecules[170]. Self-splicing introns are found in prokaryotes and eukaryotes

and used in the creation of ribozymes[171,172]. Eukaryotes enact most of their splicing through the unique molecular machinery of the spliceosome. In fact the presence of the spliceosome is one of the characteristics defining the eukaryote domain of life[173,174].

The spliceosome is composed of five small nuclear RNAs (snRNA) that interact with various heterogeneous nuclear ribonucleoproteins (hnRNPs) that are present in the nucleus[175,176] (Figure 30).

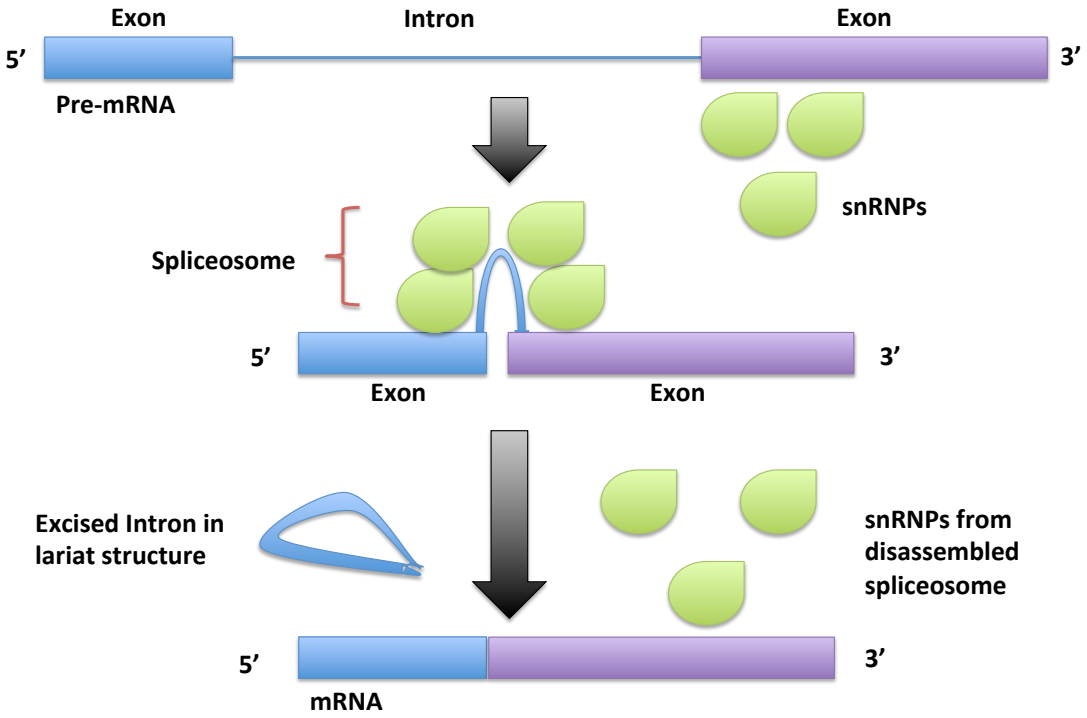


Figure 30: Spliceosome assembly, structure, and disassembly in the process of excising an intron from a pre-mRNA transcript.

Figure 30: Spliceosome

When DNA is transcribed into RNA, it exists first as heterogeneous nuclear RNA (hgRNA) that contains both introns and exons. The hgRNA is bound by hnRNPs, which assist in preventing self-binding of the transcript, the transport of mRNA out of the nucleus, and associating the hgRNA with splicing machinery. When the snRNAs and the hgRNPs come together they form small nuclear ribonucleo proteins (snRNPs), which constitute the main actors in the spliceosome[175,176]. The spliceosome removes introns from the hgRNA through a two-step biochemical reaction (Figure 31a-c).

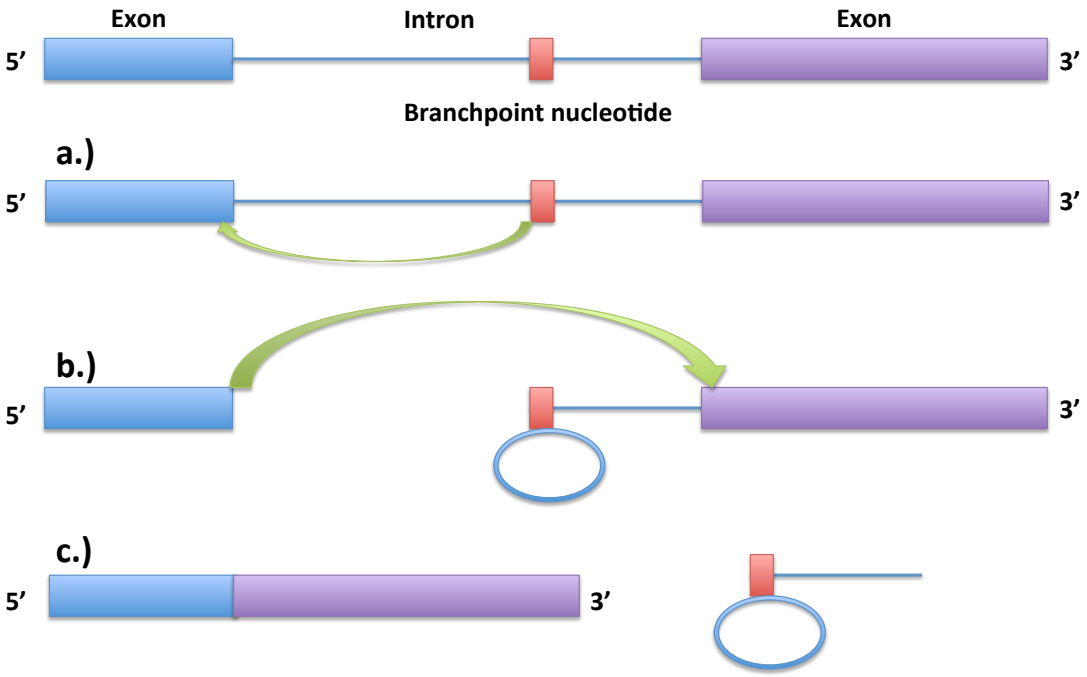


Figure 31: Diagram of splicing biochemical reaction. **a.)** branchpoint nucleotide nucleophilic attack on 5' exon **b.)** exposed exon nucleophilic attack on 3' exon **c.)** spliced exons and excised intron in lariat structure.

Figure 31: Chemical mechanism of splicing

In the first step, the splicing machinery defines a nucleotide in the target intron to be a branchpoint nucleotide[177]. The hydroxyl group on this branchpoint nucleotide then performs a nucleophilic attack on the base at the 5' splice site between the intron and the 5' exon[177] (Figure 31a). This nucleophilic attack results in the creation of a lariat structure intermediate. In the second step, the hydroxyl group on the released exon engages in a nucleophilic attack on the base at the 3' splice site, which releases the lariat

structure and joins the two exons[177] (Figure 31b,c). When the intron is released it has all the snRNPs bound to it and it is this final assembly that is called the spliceosome. Shortly, after the intron is released the snRNPs detach and the process repeats.

Alternative Splicing:

However, introns are not the only components of the hgRNA that can be spliced out. Exons are also capable of being excised as a consequence of splicing. This excision contributes to the diversity of protein products encoded by the genome[178,179]. Furthermore, introns do not always need to be excised and some persist as retained introns in the final transcript[178,179]. Utilizing splicing to express distinct versions of a processed hgRNA (termed isoforms) is the essence of alternative splicing[180].

In the human genome there is an average of 8.8 exons and 7.8 introns per gene[181]. Furthermore, it is believed that at least ninety-five percent of all multi-exon genes undergo some form of alternative spicing[182]. If one assumes that the average gene is spliced with the average values of exons and introns, one sees the tremendous genetic diversity available to the human genome through alternative splicing.

$$Possible\ Isoforms\ with\ Splicing\ of\ Exons = 2^{8.8} \approx 446$$

$$Possible\ Isoforms\ with\ Splicing\ of\ Exons\ and\ Introns = 2^{8.8+7.8} \approx 99{,}334$$

Adding to complexity of the splicing system is the fact that there are at least five forms of splicing that are known to occur in eukaryotic genomes: exon skipping, mutually exclusive exons, alternative 5' splice site, alternative 3' splice site, and retained introns[178–180] (Figure 32a-e).

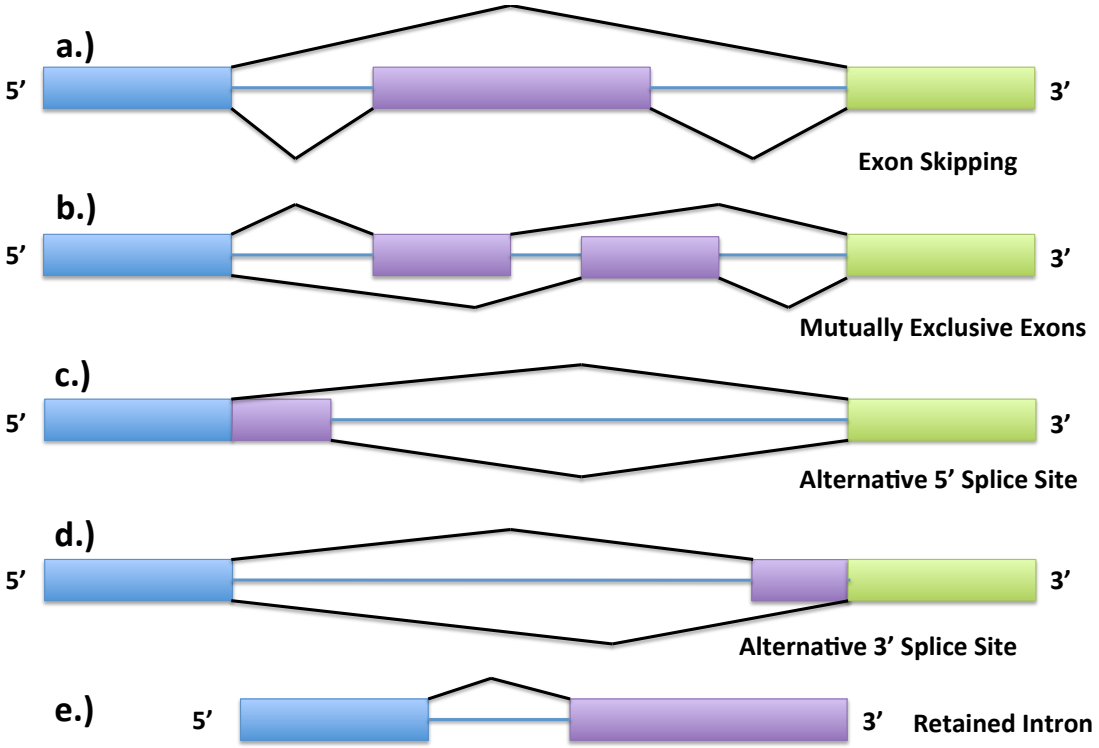


Figure 32: Alternative Splicing events **a.)** skipped exons **b.)** mutually exclusive exons **c.)** alternative 5' splice sites **d.)** alternative 3' splice site **e.)** retained intron

Figure 32: Alternative splicing events

Exon skipping, also known as cassette exon, is when an exon is either spliced out or retained during the processing of a hgRNA (Figure 32a). This form of splicing is the most common form of splicing in the human genome[178]. Mutually exclusive exons represent splicing events where, within a set of exons, one and only one is retained in a final transcript (Figure 32b). In other words mutually exclusive exons consists of exons that are never seen together in the same mRNA. Alternative 5' and 3' splice sites effectively change the size of a target exon by specifying distinct splice site boundaries in an exon that deviate from what is normally observed (Figure 32c,d). Intuitively, alternative 5' splice sites push the 5' splice site further into the 5' exon. Likewise, the alternative 3' splice sites push the 3' splice site further

into the 3' exon. The last, and rarest, form of alternative splicing is the retained intron[178] (Figure 32e). The retained intron is distinct from exon skipping in that with exon skipping the exons are flanked by intron sequences, while in retained introns the introns are flanked by exon sequences[178–180]. Furthermore, these retained introns must be in the same reading frame as the flanking exons and code for amino acids. If the retained intron is not in frame or does not code for the proper amino acids then the resulting polypeptide will be of altered, limited, or null functionality.

While generally not recognized as alternative splicing events, eukaryotic genomes are also capable of further expression diversity with the usage of multiple promoter or polyadenylation sites. These alternations affect transcriptional regulation at the 5' and 3' end points of a mRNA transcript respectively[179,180].

RBMX:

Previous studies investigating the role of MUSASHI2 in leukemia identified RBMX as a direct actor in MUSASHI2's riboproteomic network[183]. To further evaluate its significance in leukemia it was included in a mouse in-vivo pooled shRNA screen of 613 shRNAs against 128 genes in MLL-AF9 cells. Four shRNAs against RBMX in the bone marrow and two shRNAs against RBMX in the spleen were both highly depleted. Those cells with RBMX depletion also showed dramatic reduction of colony formation and increased differentiation. These results in aggregate suggested that RBMX could represent a putative oncogene.

RBMX, also known as hnRNP G, is a hgRNP whose function remained largely uncharacterized for nearly a decade after it's discovery[184,185]. Initial studies demonstrated that RBMX associated with the spliceosome and appeared to influence alternative splicing, which conformed to RBMX's expected function as a hgRNP[186]. However, the specific in vivo function of RBMX remained unknown[187]. Recent studies have implicated RBMX in various functions pertaining to genomic stability, suggesting a potential role of RBMX in genome maintenance.

RBMX is ubiquitously expressed across all organs in the human body and has been implicated in various cellular processes[188,189]. Such expression prevalence suggests a role of RBMX that may be of general importance to a cell. Recently, it was shown that RBMX acts as a critical regulator for sister chromatid cohesion during cell division. Investigators knocked down RBMX using small interfering RNA (siRNA) and observed that cells that underwent RBMX depletion had an increased accumulation of mitotic cells arrested in prometaphase. Additionally, researchers noticed that the RBMX depleted cells frequently had misaligned chromosomes on the metaphase plate. Pursuing this phenotype, researchers discovered that RBMX directly associated with chromatin in an RNA independent manner and was needed in the maintenance of cohesin, a protein complex that regulates the separation of sister chromatids during cell division[187]. Furthermore, they noted that RBMX was detected in the nucleus during the G2 and S phases of cell division, but was absent in the nucleus and cytoplasm in M phase[187]. These results suggested a role for RBMX in genome maintenance since misaligned chromosomes on the metaphase plate indicate the presence of unattached kinetochores. Unattached kinetochores activate the spindle checkpoint signal that disallows the cell to progress into anaphase [190]. This checkpoint acts to prevent aneuploidy in daughter cells and can be viewed as a means to ensure genomic stability.

The suggestion that RBMX contributed to genome maintenance was given further credence when investigators demonstrated that RBMX played a role in the homologous recombination pathway in response to DSBs in DNA[189]. In the process of conducting a screen to identify genes involved in the homologous recombination pathway, researchers found that one of their top hits was RBMX[189]. They observed that, in response to DNA damage, RBMX localized transiently to the site of damage in a PARP dependent manner[189]. Furthermore, when RBMX was knocked down by siRNA, investigators noted that the rate of homologous recombination dropped to just seven percent of the rate observed in control cells[189]. Additionally,

researchers observed that cells that underwent RBMX depletion were more sensitive to DNA damaging agents[189]. Interestingly, homologous recombination efficiency was not decreased when either RBMX was depleted or PARP was inhibited[189]. This result indicated that RBMX recruitment to the site of DNA damage was not essential, and suggested its role in mediating homologous recombination may be due to splicing effects on DNA repair proteins. In fact researchers noted that BRCA2 levels, which were decreased in RBMX depletion, could be rescued by reintroducing RBMX[189].

Role of RBMX in Disease:

RBMX function has been noted in diseases whose pathophysiology depends on DNA damage. Systemic lupus erythematous (SLE) is a devastating autoimmune disease whose pathophysiological mechanism remains undetermined[191]. A diagnosis of SLE can be made through the testing of whether certain autoantibodies are present in a patient's blood. Among the most specific of all SLE autoantibodies is an antibody to double stranded DNA (anti-dsDNA)[192,193]. Though the mechanism behind the production of antibodies against DNA remains speculative, one of the consequences of this immune reaction is genomic instability and ultimately cell death through apoptosis of a target cell[194]. Interestingly, RBMX has been implicated in playing a role in SLE[184].

Another disease whose pathology appears to be influenced by RBMX expression is endometrial cancer[195]. Specifically, RBMX function in endometrial cancer affected the expression of the estrogen receptor alpha isoform, D7, by regulating the splicing of exon seven[195]. This lead to increased levels of the isoform ERaD7 which correlated with better survival[195]. A similar role for RBMX was noted in human oral squamous cell carcinomas, where increased expression of RBMX reversed neoplastic phenotypes in mice[196]. RBMX also appears to be a downstream effector of the classic tumor suppressor protein p53, where it assists p53 in ensuring the fidelity of DNA end joining activity in response to DNA damage[197]. These findings suggested that RBMX can act as a tumor suppressor gene, acting to

ensure genomic stability through both splicing and DNA DSB repair mechanisms[189,196,197].

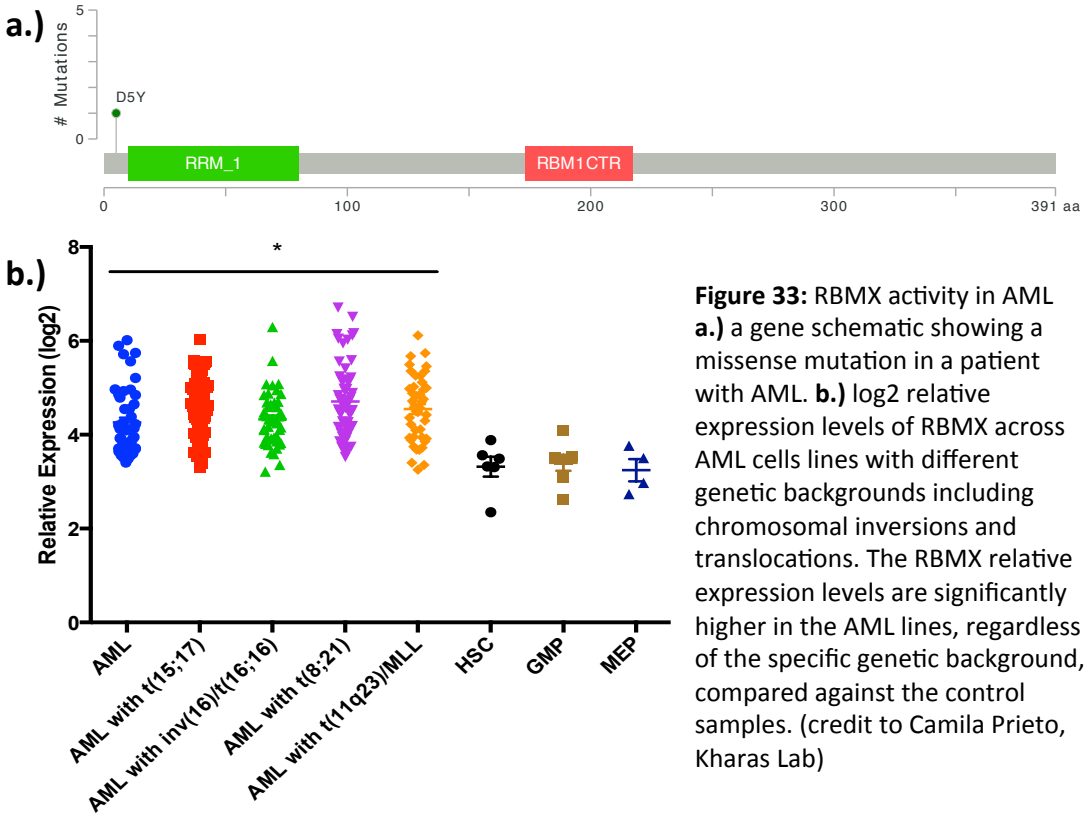


Figure 33: RBMX activity in AML **a.)** a gene schematic showing a missense mutation in a patient with AML. **b.)** log2 relative expression levels of RBMX across AML cells lines with different genetic backgrounds including chromosomal inversions and translocations. The RBMX relative expression levels are significantly higher in the AML lines, regardless of the specific genetic background, compared against the control samples. (credit to Camila Prieto, Kharas Lab)

Figure 33: RBMX activity in AML

While the role of RBMX has been noted in a few solid cancers, it remains broadly undescribed in liquid malignancies. Recently, a missense mutation in RBMX was reported in a patient with acute myeloid leukemia (AML)[198] (Figure 33a). To determine the role RBMX played in AML pathology, RBMX levels were assessed across AML cell lines with various genetic backgrounds[199] (Figure 33b). These results demonstrated that RBMX levels were significantly increased in AML cell lines compared to control cell lines, thus inferring a role for RBMX in AML pathology.

RBMX Knockdown Experiment in AML and Data Pre-Processing:

To dissect the role of RBMX in AML, short hairpin RNAs (shRNAs) were designed against RBMX and transfected into MOLM13 AML leukemia cells and control cells for twenty-four hours. Afterwards these cells underwent

puromycin selection for another twenty-four hours. Finally, after forty-eight hours post-transduction the cells were harvested for RNA sequencing and shRNA knockdown efficiency analysis (Figure 34a-b).

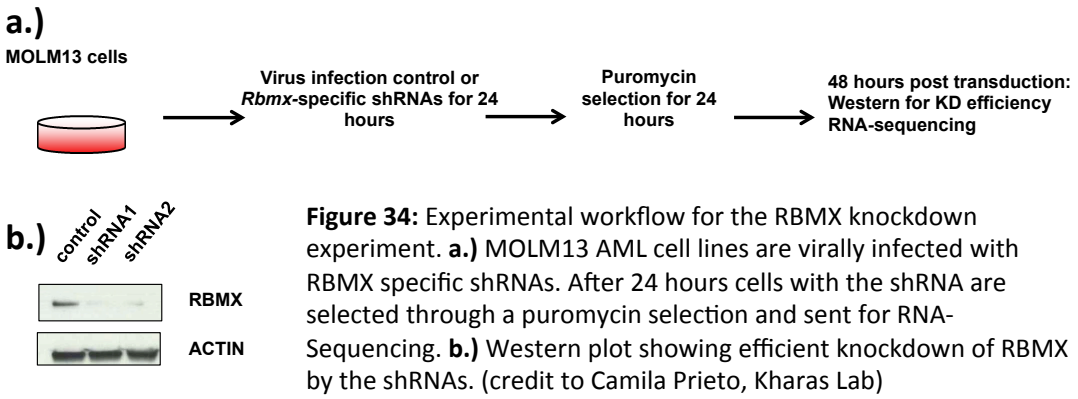


Figure 34: Experimental workflow for the RBMX knockdown experiment. **a.)** MOLM13 AML cell lines are virally infected with RBMX specific shRNAs. After 24 hours cells with the shRNA are selected through a puromycin selection and sent for RNA-Sequencing. **b.)** Western plot showing efficient knockdown of RBMX by the shRNAs. (credit to Camila Prieto, Kharas Lab)

Figure 34: RBMX experimental workflow

The RNA sequencing protocol returned FASTA files as output. These FASTA files were aligned to the human genome using the STAR aligner and were processed to BAM files using SAMtools[146,200]. Read counts were computed from the BAM files using the function summarizeOverlaps from the GenomicAlignments package[201]. Differential Expression analysis was done using the DESeq software package[202]. Statistically significant differential expression of genes between the RBMX knockdown and the control conditions was assessed using a pairwise negative binomial test. The nominal p-values that resulted from this analysis were corrected using the Benjamini-Hochberg procedure[203]. Differential splicing analysis was done using the rMATS software package[204]. Splicing events were deemed significant if they possessed a q-value less .05. Sashimi plots were generated from the MISO software package[205].

Differential Gene Expression Analysis:

Differential gene expression analysis was conducted between the RBMX knockdown samples and control samples. The result of this analysis demonstrated that fifty-eight genes were significant in their differential expression after false discovery rate (FDR) correction (Figure 35).

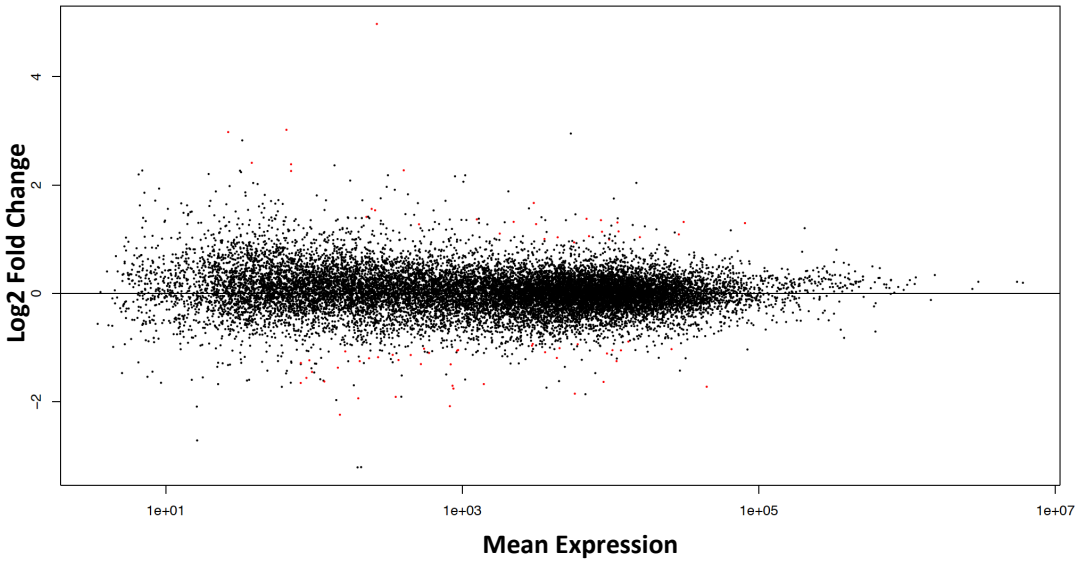


Figure 35: MA plot of the universe of assessed genes that underwent differential expression analysis. Genes that have a log2 fold change difference that is significantly different from the null hypothesis and maintains significance after false discovery rate correction are labeled in red. Genes that do not meet these requirements are labeled black.

Figure 35: RBMX MA plot

To determine how well this set of genes discriminated RBMX knockdown versus control conditions two unsupervised techniques were employed. The counts for each gene were first normalized across samples, with each gene taking the value of a Z-score. The resulting matrix of Z-scores was then hierarchically clustered and a heatmap was generated[206]. This heatmap demonstrated that the significant genes accurately partitioned RBMX knockdowns from controls (Figure 36).

A principle components analysis (PCA) was also conducted with the significant differentially expressed genes[207] (Figure 37). This analysis, in addition to illustrating the distinct clustering of RBMX depleted samples

compared to controls, also demonstrated the amount of variance that was accounted for by separating RBMX depletion from controls. The higher the variance accounted for by a principle component (PC), the more indicative of the importance of that PC in partitioning the space. In the PCA of RBMX depleted samples versus controls the first PC, which primarily separated the knockdown condition from the controls, accounted for approximately seventy-eight percent of the variance in the data.

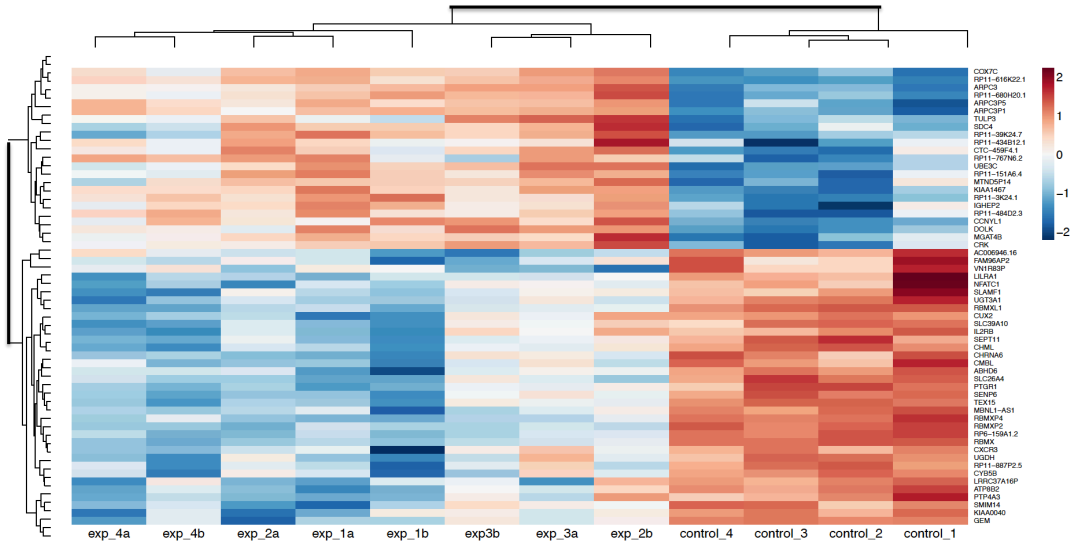


Figure 36: Genes that were deemed to be significant in their differential expression between the RBMX knockdown and control conditions had their read counts transformed into Z-scores, computed across samples, and were hierarchically clustered. A heatmap was created along with the clustering to illustrate relative expression levels for each gene across samples. The clustering places control samples in their own super-cluster and the RBMX knockdown samples in their own super-cluster demonstrating that the differential genes can effectively partition the sample space.

Figure 36: RBMX heatmap

This result illustrated that the genes that were deemed to be statistically significant in their differential expression constituted a set of data that could strongly differentiate RBMX depletion from controls. It is traditional in PCA analysis to display PCs with the largest eigenvalues, and therefore account for the largest variance in the data, in sequential order. In this tradition the first PC accounts for the most variance while the second PC accounts for the second most variance. In the PCA of the experimental samples, it is interesting to note that the second PC only accounts for six percent of the

variance. This second PC separates control and RBMX depletion samples from themselves. The value of the second PC also indicates that no other individual PC accounts for more than six percent of variance. Consequently, the PCA demonstrates that separation of RBMX depletion from controls is the most significant partitioning factor (PC1) followed by a less pronounced intra-class separation (where classes are RBMX depletion and control samples). Overall, this means that while the amount of differentially expressed genes are small, their ability to distinguish RBMX depletion from control conditions is profound.

However, the initial characterization of RBMX was as a splicing protein and consequently, its ability to affect differential expression in genes may be modest compared to its ability to affect differential splicing. Consequently, to determine the effect of RBMX depletion in AML cells a splicing analysis was performed.

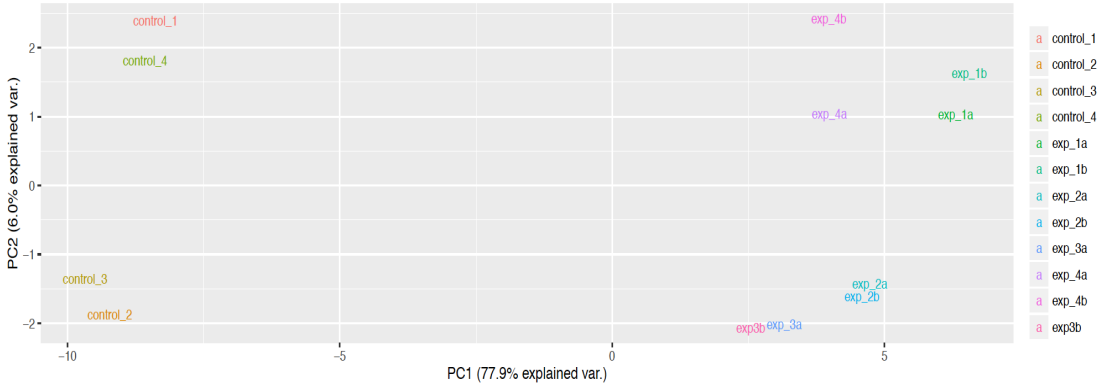


Figure 37: Principle components analysis (PCA) of experimental samples. Samples were partitioned based on genes that were deemed to be significant in their differential expression between the RBMX knockdown and control conditions. As with the heatmap the genes had their counts converted to Z-scores for each gene across samples. The PCA of these samples nicely separates control samples from RBMX knockdown samples with the first principle component, which separates controls from RBMX knockdowns, accounting for 77.9% of the variance present in the data. The second principle component separates the elements of each class from one another. In this case class is defined as the control and RBMX knockdown conditions. The second principle component accounts for 6.0% of the variance in the data, which notably smaller than the variance accounted for by the first principle component. This indicates that the data is nicely and primarily separated by class using the significant differentially expressed genes.

Figure 37: RBMX Principle Components Analysis

Splicing Analysis Methods:

The biological benefit of splicing is immense. Splicing allows for the tremendous diversity of gene products from a relatively small set of genes. Indeed for every additional element that can be spliced out of a transcript there is an exponential increase in the diversity of the transcript (Table 4).

Table 4: Splicing elements and possible isoforms

Amount of Splice Elements (n)	Possible Isoforms
2	4
5	32
8	256
16	65,536

$$Possible\ Isoforms = 2^n\ where\ n\ is\ amount\ of\ splice\ elements$$

However, the downside of this enormous diversity is that the analysis of splicing products can be quite complicated. Added to this is the fact that the depth of sequencing required to robustly detect splicing events is rather immense and hence the cost of sequencing these splicing experiments is substantial. The reads present in each sample for the RBMX knockdown experiment are shown below to illustrate the depths to which the samples were sequenced to robustly detect splicing events (Table 5).

Table 5: RBMX sample depth

Sample Name	Input Reads	Uniquely Mapped Reads
Rbmx73-04272016 (KD)	95,767,004	85,689,975
Rbmx74-05192016 (KD)	98,612,191	85,779,294
Rbmx74-04272016 (KD)	90,056,141	79,443,050
Rbmx73-05122016 (KD)	111,792,625	100,902,413
Rbmx74-05052016 (KD)	114,673,099	104,229,784
Rbmx73-05192016 (KD)	120,087,552	105,241,726
Rbmx73-05052016 (KD)	134,173,603	121,035,950
Rbmx74-05122016 (KD)	149,553,038	136,502,430
Scr-05052016 (Control)	109,911,802	101,450,085
Scr-05122016 (Control)	116,990,568	107,226,586
Scr-04272016 (Control)	114,656,786	103,434,769
Scr-05192016 (Control)	166,471,863	146,565,920

A consequence of this large read depth and cost per experiment was that splicing experiments were initially done without replicates. The analysis of non-replicate data can be non-trivial and will tend towards employing Bayesian methods[208]. As the field progressed, the cost of sequencing decreased and replicate data for splicing experiments became more common. The analysis of replicate data could be done with the non-replicate methods by combining replicates into a merged sample; however this diminished the value of replicates. Instead replicate-based tools were developed for splicing analysis. These tools tended towards being non-Bayesian in nature. Consequently, numerous splicing tools proliferated, with a sampling of popular methods detailed below[205,209–216] (Table 6).

Table 6: Differential splicing methods

Method	Description
DiffSplice	Low precision, poorly detect skipped exons, assembles transcriptome based on graph theory methods
Cufflinks	Performs well at medium read depths, calls alternative 5' and 3' splice sites well, computationally slow
DEX-Seq	Utilizes a negative binomial generalized linear model, with proper annotation file calls splicing events accurately. Incomplete annotations cause significant problems for calling splicing events.
MATS	Annotates simple splicing events well. Performs poorly on complex splicing events. Performs the best of all methods on real data.
SeqGSEA	Generally precise method that integrates differential expression with differential splicing. Computation time increases dramatically with increased permutations.
MISO	Bayesian method based calling that allows for the easy generation and visualization of plots. Computationally, scales with size of input files.
DSGseq	Medium precision compared to other methods, does not report p-values, good at detecting retained intron events
SplicingCompass	Medium precision compared to other methods, good at detecting skipped exon events
rDiff-param	Low precision, computationally fast

While these tools made admirable efforts in tackling the splicing problem the results across various splicing methods depended on the type of

analysis performed, with no single method performing the best across all conditions. A recent study investigating various tools ability to call splicing events did demonstrate, however, that the MATS splicing method bested competing methods in accurately calling simple alternative splicing events[216]. In this case simple splicing events are defined as alternative splicing events such as exon skipping, mutually exclusive exons, alternative 5' splice site, alternative 3' splice site, and retained introns. By contrast complex splicing events are defined as events where multiple simple splicing events are combined[216]. At the time of this study MATS was already augmented to handle replicate data. This adjustment to the MATS software was formalized with the introduction of rMATS, which was shown to outperform competing methods and demonstrated the added precision of using replicate data to call splicing events[204] (Figure 38). Consequently, given the accuracy and robustness of the rMATS software, it was the method chosen for the RBMX depletion splicing analysis.

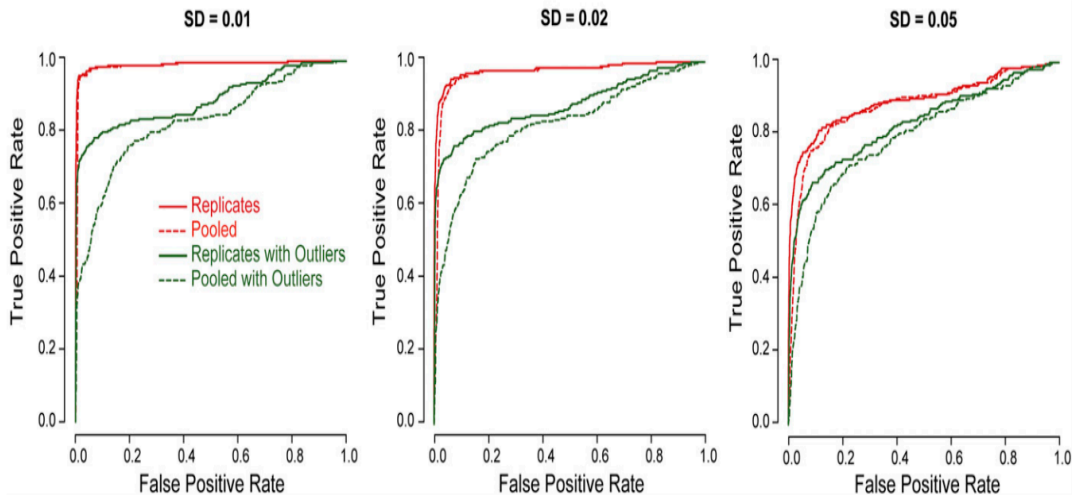


Figure 38: Receiver operator characteristic curves that demonstrate the benefit of using replicate data over pooling replicates into a single merged file. In this simulation study true positive and negatives are known and the variance in the dataset can be defined precisely. The analysis demonstrates that the use of replicates bested the pooled data method across increasing levels of variance. The advantage of using replicate data is pronounced when the data contains outliers. Overall, the use of replicate data appears to offer a significant advantage in precise calling of events. (Shihao et al. 2014)

Figure 38: rMATS pooling vs. replicate ROCs

rMATS:

Briefly, the rMATS software calls significant differentially spliced events through utilizing a generalized linear mixed model with a novel link function that serves to normalize read lengths[204]. Specifically, the rMATS model uses a hierarchical framework to model inclusion events (percent spliced in: PSI) that simultaneously accounts for estimation uncertainty in individual replicates and variability among replicates[204]. Furthermore, rMATS employs a flexible hypothesis-testing framework that allows a user to define a cutoff for the null hypothesis in assessing statistical significance[204]. Typically, the null hypothesis is set to zero (no difference between the conditions being tested), but rMATS allows the null hypothesis to be accepted if a difference is less than a value n, which is defined by the user. This allows rMATS to be more stringent in calling splicing events compared to other methods. Additionally, rMATS normalizes the lengths of individual splice variants, which allows it to call all major types of alternative splicing events and use reads mapped to both exons and splice junctions [204]. Finally, rMATS allows for the usage of paired replicates, which makes it attractive for numerous types of timed intervention studies[204].

Characteristics of Differential Splicing Events in RBMX Depletion Experiment:

The overwhelming majority of significant differentially spliced events in the RBMX depletion study were skipped exons (64.6% of all called events). This observation conforms with previous studies that indicated that skipped exons are the most common splicing event observed in mammalian genomes[178]. After skipped exons the second most commonly observed event was mutually exclusive exons (30.0% of all called events), followed by retained introns (2.7% of all called events), alternative 3' splice site (1.7% of all called events) and alternative 5' splice sites (1% of all called events) (Figure 39). Consequently, the knockdown of RBMX in MOLM13 cells leads

to differential splicing events that manifest primarily as skipped exons and mutually exclusive exons.

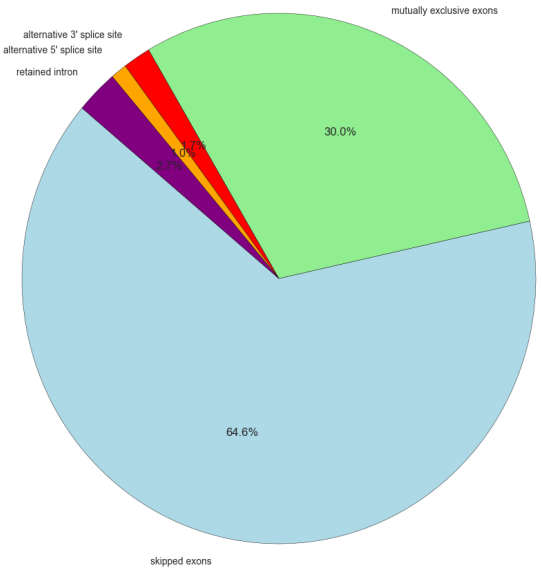


Figure 39: Pie plot of all significant differential splicing events in the RBMX knockdown experiment. For an event to be included it must be differentially spliced to a degree that it maintains statistical significance after false discovery rate correction.

Figure 39: Pie plot of significant differentially spliced events

In addition to classifying alternative splicing events, the directionality of each spliced event was determined. The directionality of a spliced event is simply the test condition that has the higher event inclusion level. This definition can be illustrated most clearly by way of an example. Imagine that a significant differentially spliced skipped exon event exists. This skipped exon is significant because the level of exon inclusion is sufficiently distinct between the RBMX depletion condition and control condition such that the probability of seeing this difference by chance falls below a threshold probability, even after correction for multiple hypothesis testing. If the exon inclusion is greater in the RBMX depletion condition then the directionality of the event is towards the RBMX condition. Conversely, if the exon inclusion is greater in the control condition, then the directionality of the event is towards the control condition. In the RBMX depletion experiment it was observed that

there was greater exon inclusion in the direction of the RBMX condition for skipped exons, retained introns, alternative 3' splice sites, and alternative 5' splice sites. For mutually exclusive exons there was no notable directionality between the RBMX depletion and control conditions (Figure 40). Therefore following RBMX knockdown, there appeared to be preferential exon inclusion, where normally there would have been exon exclusion, for the majority of alternative splicing events. Overall, RBMX knockdown appeared to increase exon retention in AML MOLM13 cell lines.

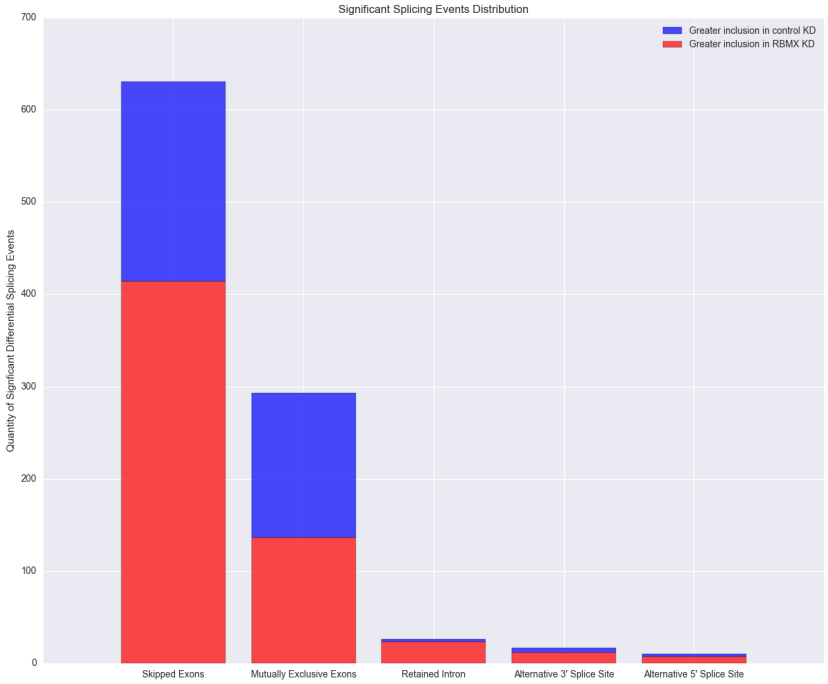


Figure 40: A bar plot of significant alternative splicing events. Events in red show greater inclusion in the RBMX knockdown condition, while blue shows great inclusion in the control condition

Figure 40: Bar plot of significant alternative splicing events

Gene Ontology Analysis of Significant Alternatively Spliced Genes:

To determine what pathways were effected by RBMX knockdown a gene ontology analysis, using the GOrilla software, was done on all significant differentially spliced genes[217] (Figure 41). This analysis revealed that five pathways were affected by the knockdown: negative regulation of intrinsic apoptotic signaling pathway in response to DNA damage, regulation of intrinsic apoptotic signaling pathway, regulation in response to DNA damage

stimulus, negative regulation in response to DNA damage stimulus, and intrinsic apoptotic signaling pathway in response to DNA damage. Of these pathways two remained significant after multiple hypothesis testing (MHT) correction: regulation of intrinsic apoptotic signaling pathway and the negative regulation of intrinsic apoptotic signaling pathway in response to DNA damage. The results of the gene ontology analysis supported the notion that RBMX plays a role in genome maintenance, since the depletion of RBMX primarily affects DNA damage response or apoptosis pathways. To understand if these gene ontology terms were the consequence of the global effects of RBMX knockdown or primarily the consequence of one type of splicing event, each alternative splicing event was individually investigated.

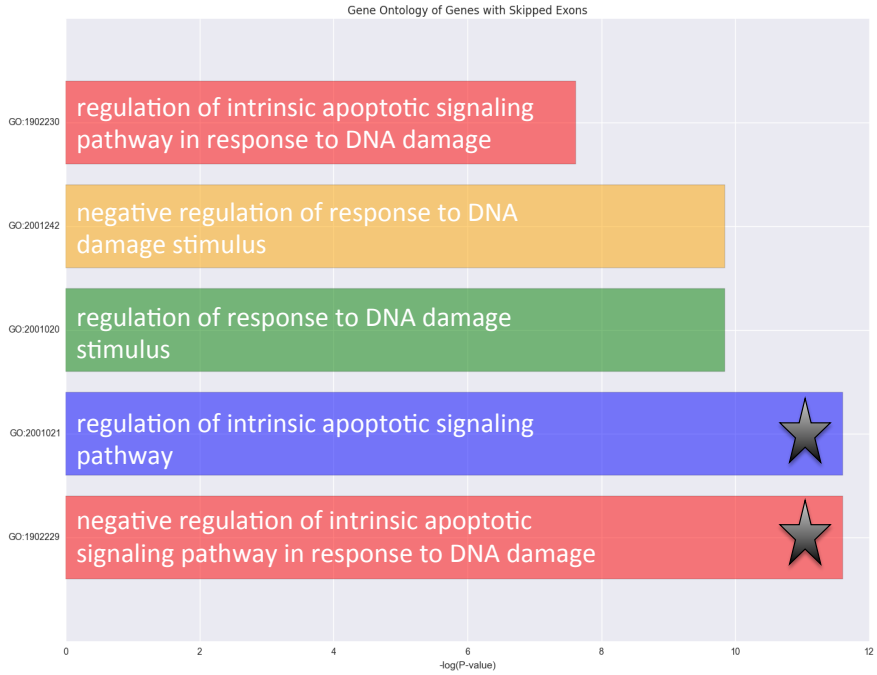


Figure 41: Gene ontology plot of terms that resulted from a GOrilla analysis of all significant differentially spliced genes. Only GO:2001021 and GO:1902229 remained significant after MHT correction

Figure 41: Gene ontology pathways using all significant alternative splicing events

Skipped Exon Events:

The significant differentially spliced skipped exon events that followed RBMX knockdown included 414 events with higher inclusion levels in RBMX condition compared to 217 events with higher inclusion levels in the control

condition (Figure 42). The skipped exon events between the RBMX and control conditions targeted rather independent sets of genes, with only five common genes between them including: SCRIB, ALG11, TBCE, SLC25A26, and CD44 (Figure 43). Of these common genes SCRIB and CD44 have been implicated in having roles in cancer[218–220]. Furthermore, when the genes that constituted these sets were further investigated it was realized that some genes were associated with multiple significant differential splicing events. Specifically, of the 414 events with higher inclusion in the RBMX condition, only 287 genes were affected. Likewise, of the 217 events with higher inclusion in the control condition only 175 genes were affected. Interestingly, there was little overlap between significant differentially spliced and expressed genes, with the intersection consisting of ALDH18A1, PABPC4, KLHL23, and MYL9.

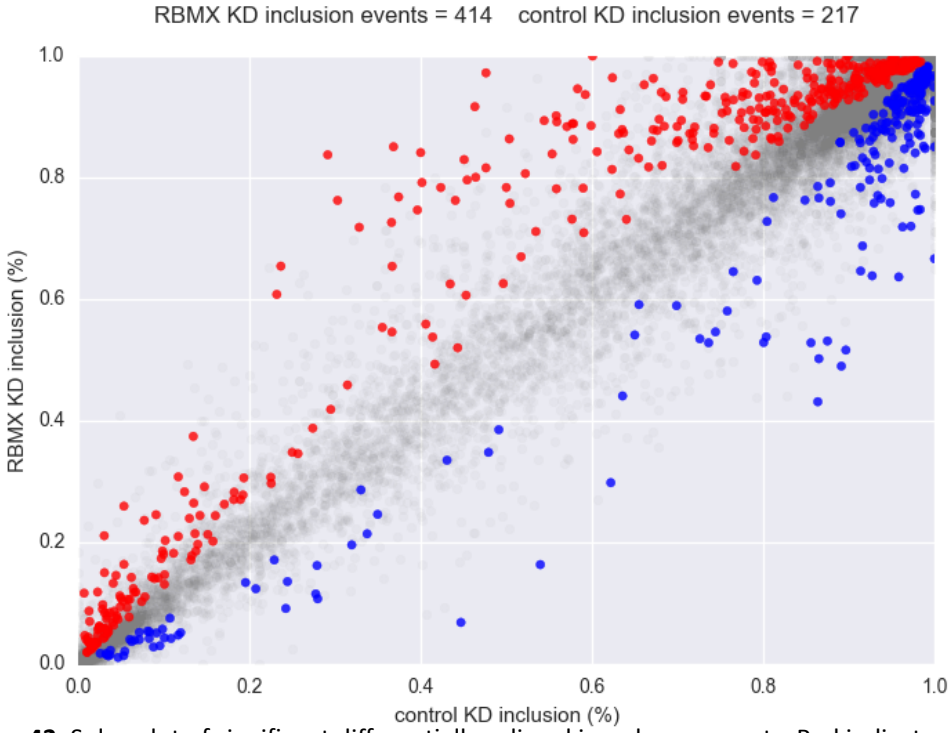


Figure 42: Splay plot of significant differentially splice skipped exon events. Red indicates significant events with higher inclusion in the RBMX knockdown condition, while blue indicates significant events with higher inclusion in the control condition.

Figure 42: Splay plot of skipped exon events

A gene ontology analysis was conducted on all genes with significant skipped exon events revealing four terms that remained significant after MHT

correction: regulation of intrinsic apoptotic signaling pathway, regulation of response to DNA damage stimulus, negative regulation of intrinsic apoptotic signaling pathway response to DNA damage, and negative regulation of response to DNA damage stimulus (Figure 44a). These terms closely mirrored the terms that were revealed as a result of gene ontology analysis done on all significant differentially spliced events. However, when the analysis was restricted to just skipped exons the p-values associated with the terms increased notably and several terms gained significance after MHT correction that were not significant in the global gene ontology study. These results prompted a second gene ontology study where the genes that had higher inclusion in the RBMX knockdown condition were studied alone. The results of this study gave identical gene ontology terms as those seen in the gene ontology of all skipped exons (Figure 44b). When the same analysis was done for those genes with higher inclusion in the control condition only two terms were returned: cation transport and ion transport. Consequently, the gene ontology terms for skipped exons come from the set of genes that have higher exon retention levels in the RBMX knockdown condition.

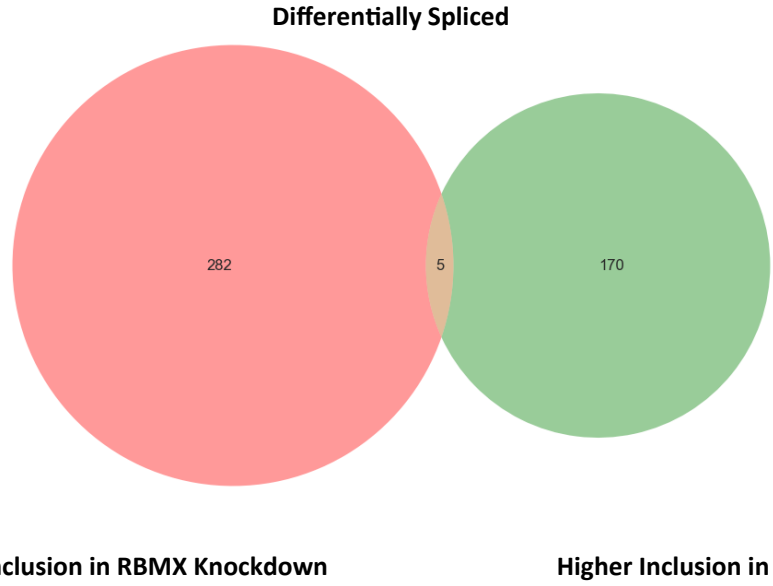


Figure 43: Venn diagram of genes with significant differentially spliced skipped exons between the RBMX knockdown condition and the control condition. The two sets of genes are largely independent of one another with an intersection of only five genes: SCRIB, ALG11, TBCE, SLC25A26, and CD44. There are notably more genes with skipped exons that retain an exon in response to RBMX knockdown than lose an exon.

Figure 43: Venn diagram of significant skipped exons between RBMX KD and control conditions

Many of the genes with higher exon retention in the RBMX knockdown had multiple significant skipped exon events associated with them. The gene with the most skipped exon events was the CD44 gene (Figure 45). This gene is a cell surface molecule that is involved in a myriad of functions including cell proliferation, cell differentiation, cell migration, angiogenesis, and cell survival signaling[221]. Furthermore, it also believed that alternative splicing of this gene generates tumor specific isoforms and high levels of it are needed to generate leukemic cells[222]. Additionally, splice variants of CD44 have been implicated in head and neck squamous carcinomas (i.e.: oral squamous cell carcinoma) and endometriosis, which can predispose women to uterine cancer[223–225]. Other genes that had multiple significant events associated with them were the DAXX and POLL genes. The DAXX gene, also known as the death-associated protein 6, is a protein involved in the apoptosis pathway[226]. Similarly, the POLL gene has a function in DNA double strand break repair[227,228].

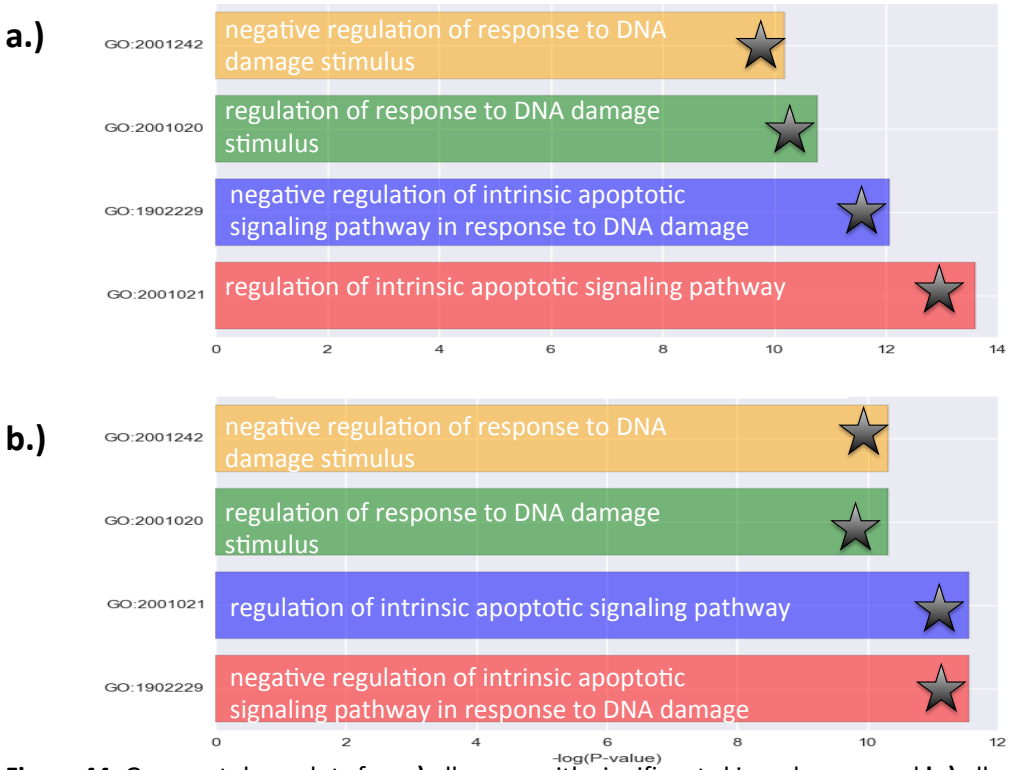


Figure 44: Gene ontology plots for **a.)** all genes with significant skipped exons and **b.)** all genes with significant skipped exons with higher inclusion in the RBMX knockdown condition

Figure 44: Gene ontology of skipped exon events

To better determine how these splicing events were affecting target genes a domain analysis was conducted using interval trees constructed with UniProt gene domain annotations[159]. Using these interval trees, domain annotations overlapping the coordinates of the splicing event could be determined. Domains were determined to be associated with skipped exons with higher exon retention in the RBMX knockdown, if they significantly associated with these skipped exons over all other non-significant splicing events called by rMATS using the Fisher Exact Test (Figure 46). The most significantly associated domains in this analysis both belonged exclusively to the CD44 gene: CD44-antigen and stem. Interestingly, the BRCT domain associated exclusively with the POLL gene. The BRCT domain is the C-terminal domain of a breast cancer susceptibility protein and is found in those proteins that play a role in cell cycle checkpoint responses to DNA damage[229,230]. Consequently, in response to RBMX depletion, the identity of the genes involved in skipped exon events, and the domains that are spliced, both reinforce the idea that RBMX contributes to genome maintenance.

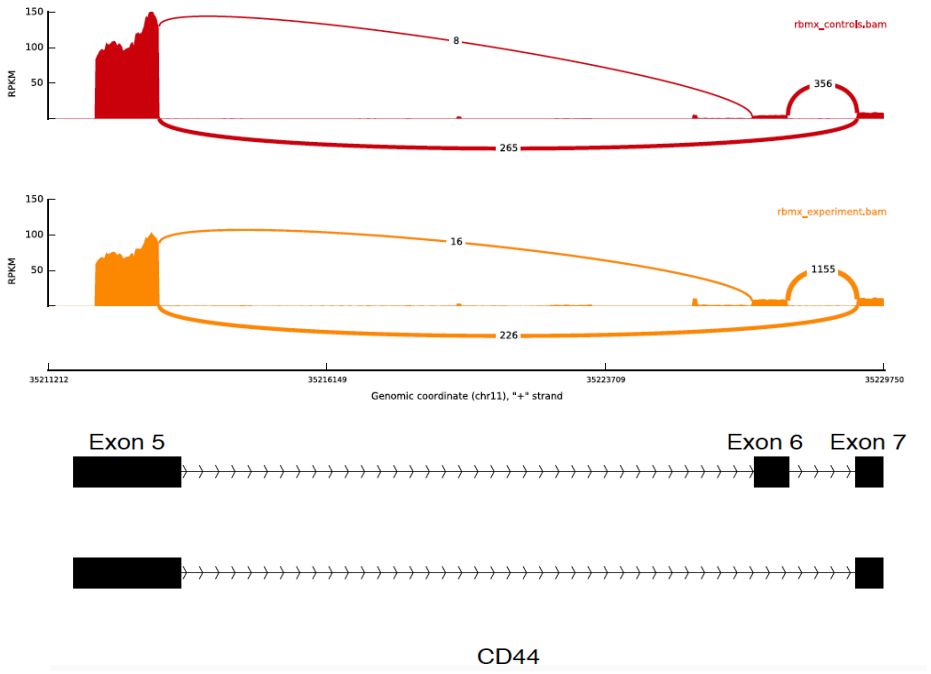


Figure 45: Sashimi plot showing an example of differential exon retention in the CD44 gene. Specifically, the retention is seen as the greater amount of reads spanning the junctions of exon 6 and exon 7 in the RBMX knockdown condition compared to control

Figure 45: Sashimi plot of CD44

Finally, RBMX depletion had certain global effects on the potential types of gene isoforms expressed. Using interval trees constructed of gene transcript annotations, the lengths and number of exons were assessed for significant differently spliced skipped exon events. Interestingly, RBMX depletion significantly shifted splicing events towards the exons at the end of transcripts (Figure 47b). Considering just those genes with significant skipped exons, RBMX knockdown significantly targeted those genes with fewer exons (Figure 47a).

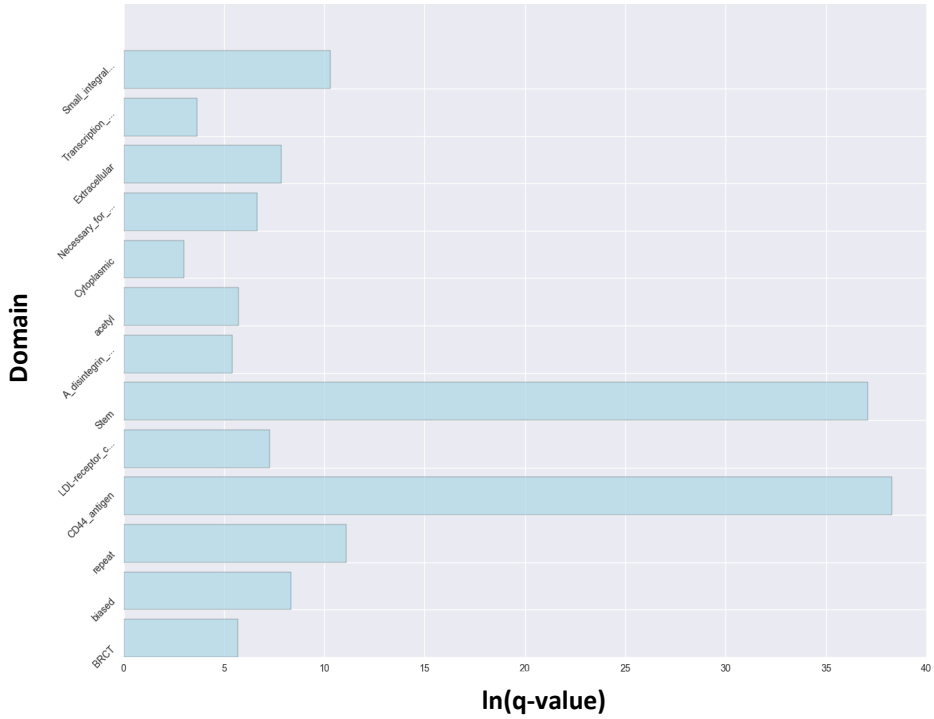


Figure 46: Bar plot showing domain annotations that are significantly associated with genes that have significant skipped exon events in which there is higher exon retention in the RBMX knockdown condition.

Figure 46: Bar plot of domains significantly associated with skipped exon events

Overall, RBMX knockdown generated a large amount of skipped exon events in genes that influence the apoptotic and DNA damage response pathways. The pathways implicated are virtually identical to the pathways tagged with the gene ontology analysis of all significant differentially spliced genes. Furthermore, it appears that only those skipped exon events with higher exon retention in the RBMX knockdown condition contribute to the generation of these gene ontology terms. Several of these genes have

multiple significant skipped exon events and the domains these events overlap, reinforce the idea that RBMX splicing affects genes that are essential to genome maintenance.

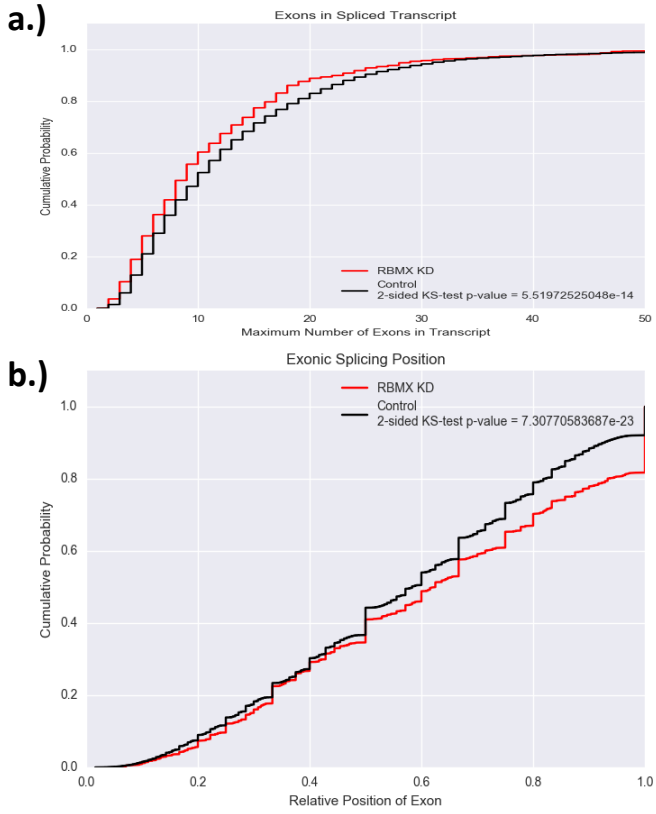


Figure 47: Descriptive global effects of RBMX knockdown. **a.)** RBMX knockdown preferentially affected transcripts with fewer exons. In this study only those genes, and their associated transcripts, that had significant skipped exon events were considered in the RBMX KD condition. **b.)** RBMX knockdown demonstrated a significant preference for splicing events that occurred towards the end of a transcript. In this cumulative distribution plot the relative position of the exon is defined as 1.0 if the exon in question is the last exon in the transcript. Conversely, an exon with a relative position of 0.0 is the first exon in the transcript. Any values between 0.0 and 1.0 are middle exons. The plot shows those genes with significant skipped exon events tended to have skipped exon events in their distal exons.

Figure 47: CDFs of skipped exon events

Mutually Exclusive Exons:

The significant differentially spliced mutually exclusive exon events that followed RBMX knockdown included 136 events with higher inclusion levels in RBMX condition compared to 157 events with higher inclusion levels in the control condition (Figure 48). The mutually exclusive exon events between the RBMX and control conditions targeted largely independent sets of genes, with only ten common genes between them including: FNTA, AK2, VAPA, CCDC14, N4BP2L2, CNPY3, TTC13, POLL, PHF3, and KIAA1191. Of these genes, POLL is notable for it's implication as a DNA damage repair protein. As with skipped exon events, when the genes that constituted these sets were further investigated it was realized that some genes had multiple significant differential splicing events. Specifically, of the 136 events with

higher inclusion in the RBMX condition, only 104 genes were affected. Likewise, of the 157 events with higher inclusion in the control condition only 123 genes were affected (Figure 49a). Interestingly, there was little overlap between significant differentially spliced and expressed genes, with the intersection consisting of PABPC4. However, a gene ontology analysis of the genes with significant mutually exclusive exons did not yield any specific gene ontology terms and individual investigation of several genes showed functions that were largely related to basic biochemical pathways.

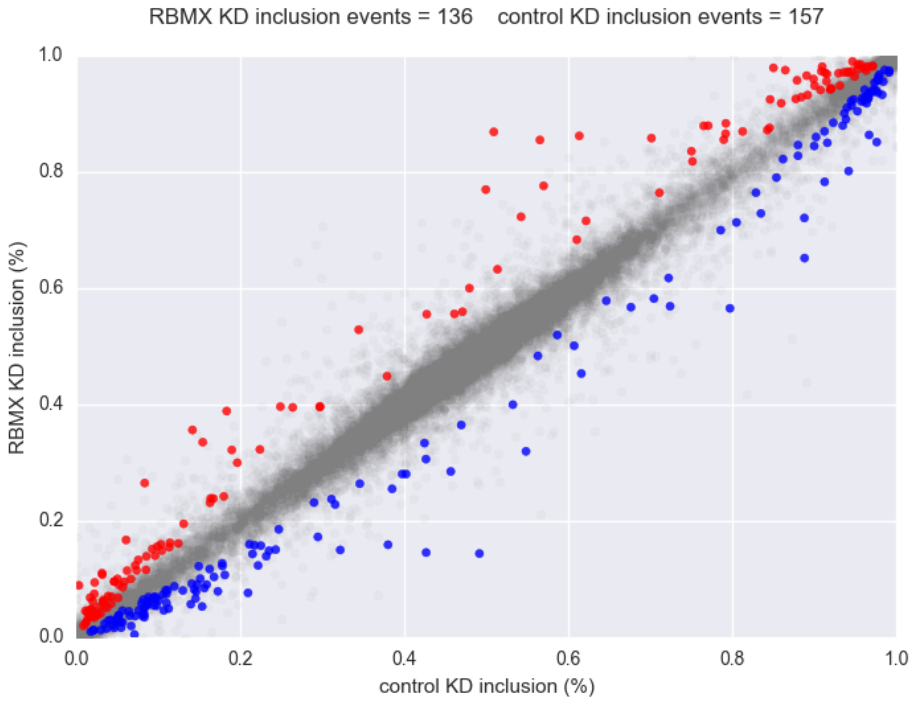


Figure 48: Splay plot of significant differentially splice mutually exclusive exon events. Red indicates significant events with higher inclusion in the RBMX knockdown condition, while blue indicates significant events with higher inclusion in the control condition.

Figure 48: Splay plot of mutually exclusive exon events

Unlike with skipped exons, RBMX depletion did not significantly shift mutually exclusive splicing events towards the exons at the end of a transcript. However, genes with significant mutually exclusive exon events had fewer exons than would be expected (Figure 49b). Overall, mutually exclusive exons signify the second most numerous alternative splicing event when RBMX is depleted from MOLM13 cells. However, the genes experiencing significant mutually exclusive exon events do not replicate, or

contribute to, the terms seen in the global gene ontology study. Similar to what was seen with skipped exon events, mutually exclusive exons that occur in the context of RBMX depletion, tend to be associated with transcripts that possess fewer exons.

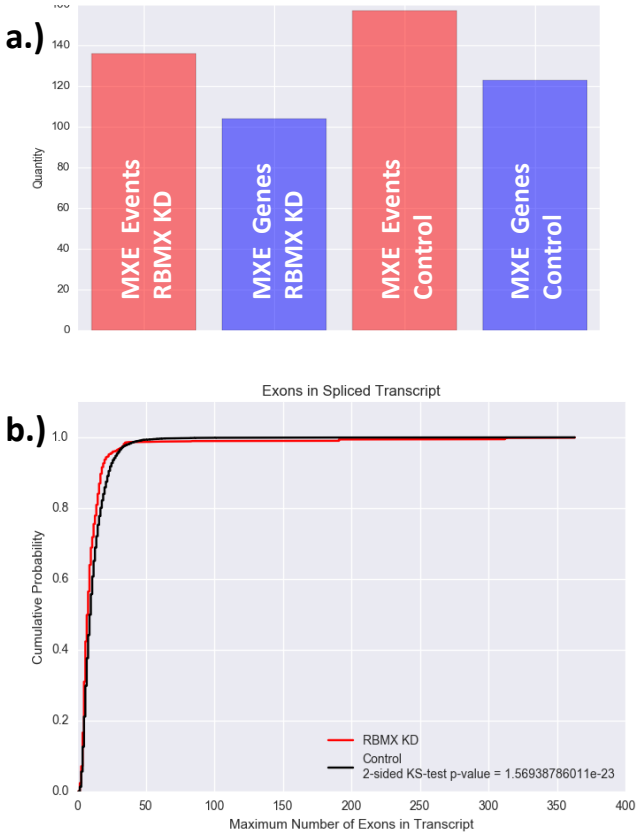


Figure 49: Mutually exclusive exon splicing events **a.)** a bar plot demonstrating the amount of significant mutually exclusive exon events and the genes they occur in for both the RBMX and control conditions. As illustrated, mutually exclusive exon events outnumber genes enumerated indicating that some genes possess multiple significant mutually exclusive exon events. **b.)** RBMX knockdown preferentially affected transcripts with fewer exons. In this study only those genes, and their associated transcripts, that had significant mutually exclusive exon events were considered in the RBMX KD condition.

Figure 49: Mutually exclusive exon events CDF and bar plot

Retained Introns:

The significant differentially spliced retained events that followed RBMX knockdown included twenty-three events with higher inclusion levels in RBMX condition compared to three events with higher inclusion levels in the control condition (Figure 50). The retained intron events between the RBMX and control conditions targeted independent sets of genes, with no intersecting genes between them. When the genes that constituted these sets were further investigated it was realized that only the genes with higher inclusion levels in the RBMX knockdown condition had any genes with multiple significant retained intron events. Specifically, of the twenty-three events with higher inclusion in the RBMX condition, only nineteen genes were

affected (Figure 51a). There was no overlap between significant differentially spliced and expressed genes. Additionally, a gene ontology analysis on significant differentially spliced retained introns did not yield any gene ontology terms, likely due to the small sample size of genes in this splicing category.

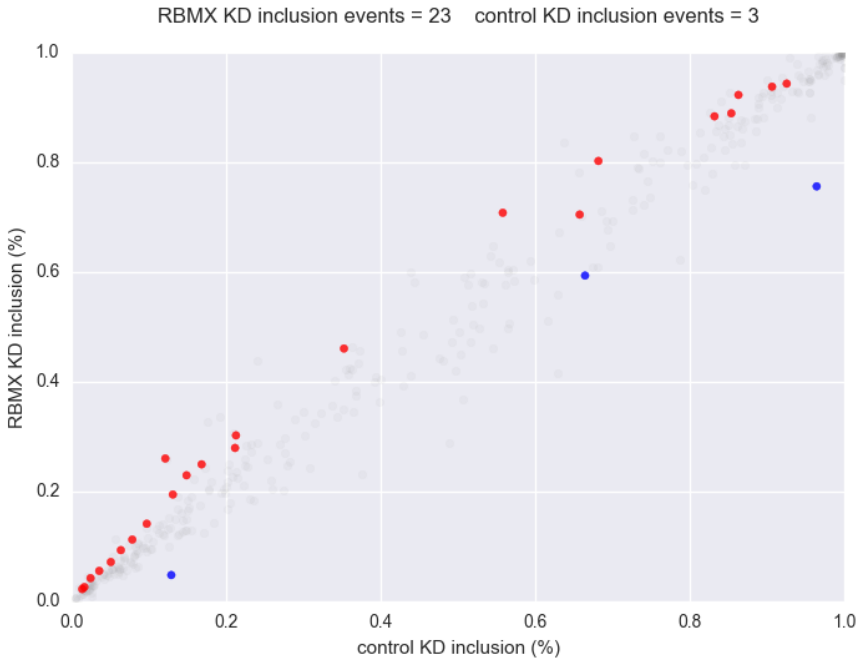


Figure 50: Splay plot of significant differentially splice retained intron events. Red indicates significant events with higher inclusion in the RBMX knockdown condition, while blue indicates significant events with higher inclusion in the control condition.

Figure 50: Splay plot of retained introns

The observation seen with skipped exons that during RBMX depletion splicing events tend to target exons towards the end of the transcript was not replicated with retained exons. Additionally, while the there was a significant difference in the amount of exons present in a transcript of a genes with significant retained intron events versus control, a clear pattern was not discernible indicating that the populations are distinct but a trend is not obvious (Figure 51b). Once again this is likely due to the small sample size of genes present in the retained intron category.

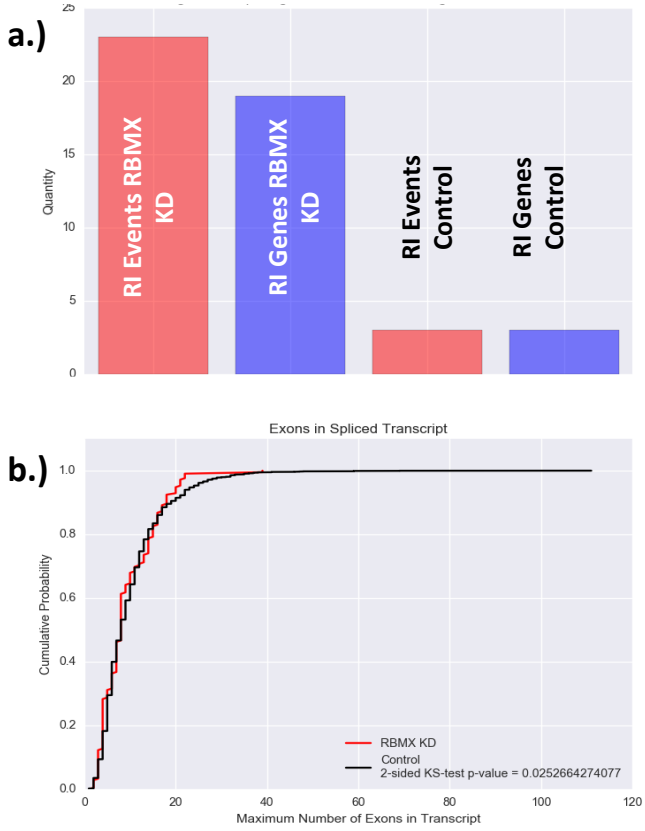


Figure 51: Retained intron splicing events **a.)** a bar plot demonstrating the amount of significant retained intron events and the genes they occur in for both the RBMX and control conditions. As illustrated, retained intron events outnumber genes for the RBMX knockdown condition only. This indicates that some RBMX knockdown genes possess multiple significant retained intron events. **b.)** RBMX knockdown appears to have an association with the number of exons in a transcript. However, the pattern is unclear for the retained intron case. In this study only those genes, and their associated transcripts, that had significant retained intron events were considered in the RBMX KD condition.

Figure 51: Retained intron CDF and bar plot

Alternative 3' Splice Site:

The significant differentially spliced alternative 3' splice site that followed RBMX knockdown included twelve events with higher inclusion levels in RBMX condition compared to five events with higher inclusion levels in the control condition (Figure 52). The alternative 3' splice site events between the RBMX and control conditions targeted independent sets of genes, with no intersecting genes between them. When the genes that

constituted these sets were further investigated it was realized that only the genes with higher inclusion levels in the RBMX knockdown condition had any genes with multiple significant alternative 3' splice sites events. Specifically, of the twelve events with higher inclusion in the RBMX condition, only nine genes were affected (Figure 53a). There was no overlap between significant differentially spliced and expressed genes. Additionally, a gene ontology analysis on genes with significant differentially spliced alternative 3' splice site events did not yield any gene ontology terms, likely due to the small sample size of genes in this splicing category.

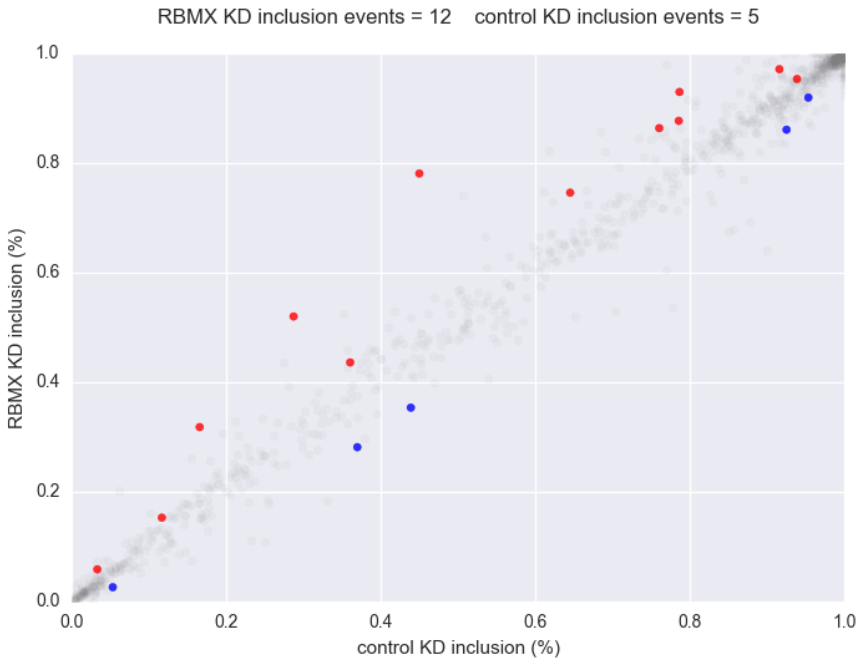


Figure 52: Splay plot of significant differentially splice alternative 3' splice site events. Red indicates significant events with higher inclusion in the RBMX knockdown condition, while blue indicates significant events with higher inclusion in the control condition.

Figure 52: Splay plot of alternative 3' splice site

The observation seen with skipped exons that during RBMX depletion splicing events tend to target exons towards the end of the transcript was reversed with alternative 3' splice sites (Figure 53b). Genes with significant alternative 3' splice sites had a significant shift towards having splicing occur at the beginning of their transcript. Furthermore, as with skipped exons and mutually exclusive exons, alternative 3' splice sites that occur in the context of

RBMX depletion, tend to be associated with transcripts that possess fewer exons (Figure 53c).

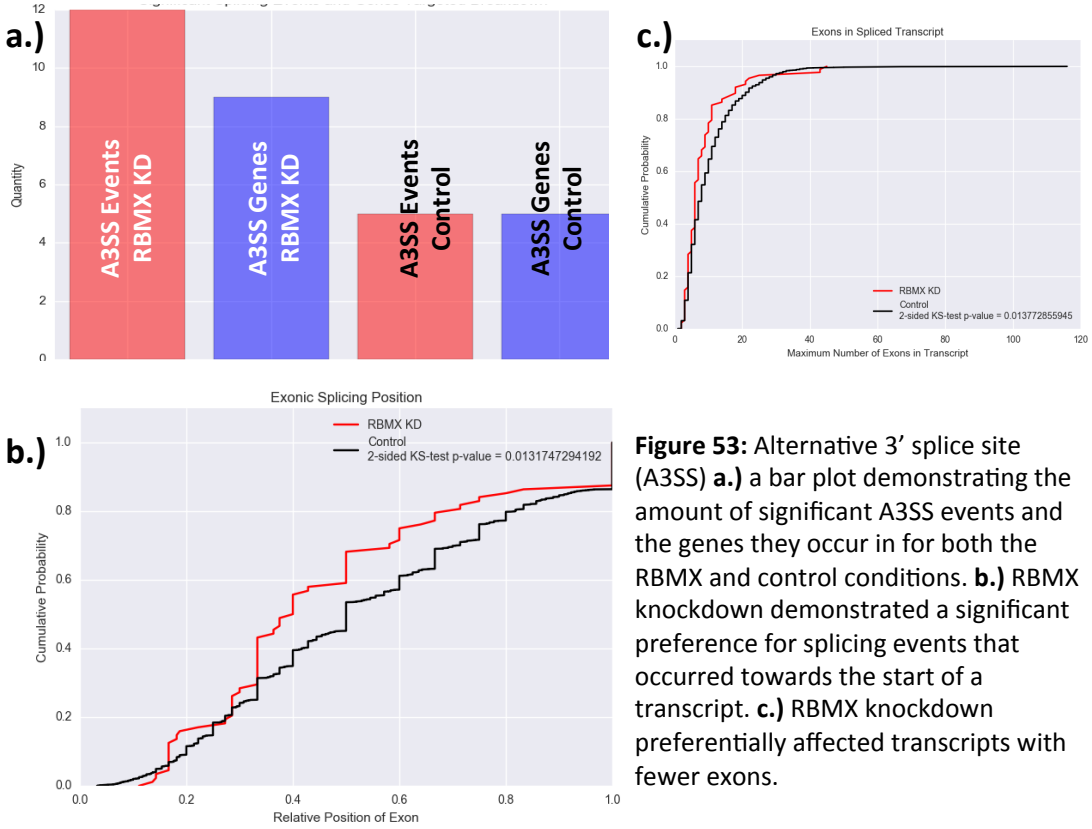


Figure 53: Alternative 3' splice site (A3SS) **a.)** a bar plot demonstrating the amount of significant A3SS events and the genes they occur in for both the RBMX and control conditions. **b.)** RBMX knockdown demonstrated a significant preference for splicing events that occurred towards the start of a transcript. **c.)** RBMX knockdown preferentially affected transcripts with fewer exons.

Figure 53: Alternative 3' splice site CDFs and bar plot

Alternative 5' Splice Sites:

The significant differentially spliced alternative 5' splice site events that followed RBMX knockdown included seven events with higher inclusion levels in RBMX condition compared to three events with higher inclusion levels in the control condition (Figure 54). The alternative 5' splice site events between the RBMX and control conditions targeted independent sets of genes, with no intersecting genes between them. When the genes that constituted these sets were further investigated it was realized that some genes were associated with multiple significant differential splicing events. Specifically, of the seven events with higher inclusion in the RBMX condition, only four genes were affected. Likewise, of the three events with higher inclusion in the control condition only two genes were affected. There was no overlap between significant differentially spliced and expressed genes. Additionally, a gene

ontology analysis on genes with significant differentially spliced alternative 5' splice site events did not yield any gene ontology terms, likely due to the small sample size of genes in this splicing category.

The repeated observation that RBMX depletion tends to shift splicing events towards the end of the transcript was not replicated with alternative 5' splice sites nor was the observation that RBMX depletion appears to cause splicing events on genes with fewer exons. However, this result may be due to the small sample size of the splicing events in this category.

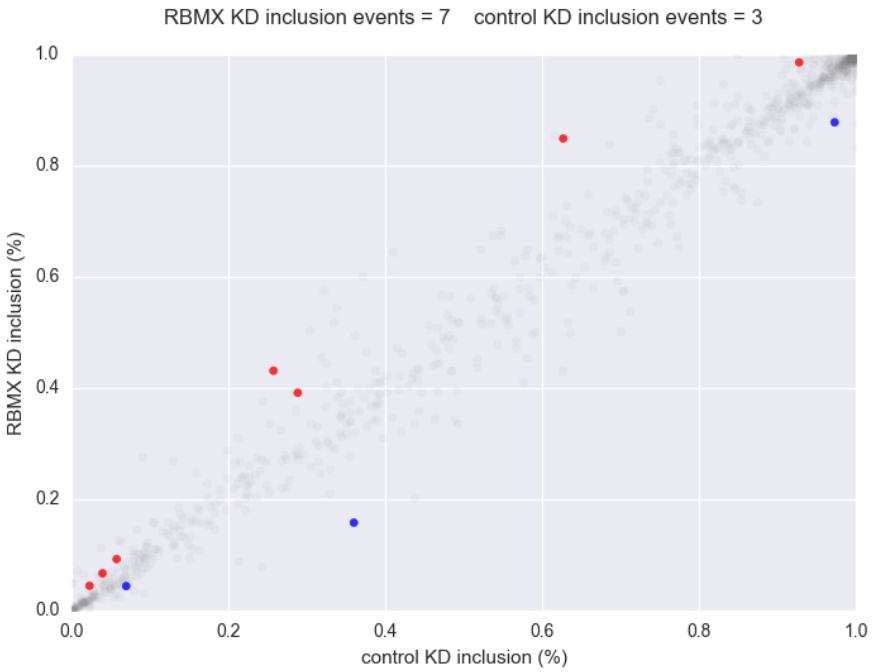


Figure 54: Splay plot of significant differentially splice alternative 5' splice site events. Red indicates significant events with higher inclusion in the RBMX knockdown condition, while blue indicates significant events with higher inclusion in the control condition.

Figure 54: Splay plot of alternative 5' splice site

Alternative Splicing Events in the Context of RBMX Depletion:

Gene ontology analysis of all significant differentially spliced genes indicated that RBMX depletion specifically affects cellular pathways responsible for DNA damage repair and apoptosis regulation. Investigating the roles each splicing event contributed to the generation of this annotation, it was noted that only the skipped exon events could replicate the gene

ontology result. Specifically, within skipped exons it was only those events where an exon was retained in response to RBMX depletion that the gene ontology terms could be generated. Further delving into those genes that retained exons in response to RBMX knockdown revealed that numerous genes had multiple significant skipped exons events. Among this subset of genes several such as DAXX and POLL supported the notion that RBMX plays a role in regulating the proteins involved in maintaining genomic stability. Interestingly, the gene with the most significant skipped exon events was CD44, which is a gene implicated in various cancers. Perhaps most interestingly CD44 is reported to be essential for creating leukemic cells in AML[222]. Additionally, the domains affected by significant differential splicing in response to RBMX depletion suggest targets that may contribute to cancer evolution. The most suggestive domain in this category is the BRCT domain that is the repeated target of skipped exon events in the POLL gene. Overall it appears that RBMX plays a role in genome maintenance through regulating the splicing of genomic maintenance proteins primarily through the splicing mechanism of skipped exons.

Reverse Phase Protein Array Data:

Reverse phase protein array (RPPA) is a tool that allows investigators to assess the expression levels of a set of a priori selected proteins across a large number of samples simultaneously (appendix). The details of the experiment are described in the appendix, but it is sufficient to understand that this procedure aims to quantify protein expression in a quantitative manner so differential expression of proteins across samples can be assessed. RPPA data was collected for six RBMX depletion experiments and three control experiments and protein expression levels were quantified. Significant differential expression for a protein between the two conditions was determined using the Wilcoxon ranked sum test on the normLog2 values of the data as reported by the facility that conducted the RPPA experiment. A total of 304 proteins were analyzed in this experiment, of which twenty-six had nominally significant p-values (Table 7).

Table 7: RPPA nominally significant genes

Nominally Significant Protein (gene name)
53BP1
Akt_pS473
ATR_pS428
b-Catenin_pT41_S45
Caveolin-1
cdc25C
Chk1
Chk2
Cyclin-B1
Cyclin-D3
Glutaminase
HES1
INPP4b
JNK2
MCT4
PAR
PARP
Paxillin
PLK1
Rb
Rictor
S6_pS240_S244
XPA

Unfortunately, after MHT correction all p-values became non-significant. Furthermore, there was no intersection between the RPPA genes and the genes that were significantly differentially spliced.

H3k9me3 Observation:

A few studies have suggested an association between RBMX and H3k9me3 methylation[187,231]. H3k9me3 is a histone mark that is associated with repressed gene expression given its high correlation with constitutive heterochromatin[232]. To investigate this potential association, H3k9me3 Chip-Seq data from k562 chronic myelogenous leukemia (CML) cell lines was downloaded from ENCODE[150]. Unfortunately, ENCODE did not have H3k9me3 Chip-Seq data for the MOLM13 cell line, but the two cells lines are sufficiently similar to allow for a preliminary, albeit not conclusive, analysis about the association of H3k9me3 marks in the context of RBMX knockdown. The 16,446 H3k9me3 peaks in the k562 file were converted into an interval tree and used to annotate whether significant differential splicing events overlapped H3k9me3 peaks.

Across all significant differentially spliced events only three events were noted to have any overlap with H3k9me3 peaks. Furthermore, these splicing events were all skipped exon where there was greater exon retention in the face of RBMX knockdown. The skipped exon events occurred in the genes PRKACA, ZNF792, and MBD1. ZNF792 is a zinc finger protein involved broadly with transcription and PRKACA is a catalytic subunit of protein kinase A[233–235]. Interestingly MBD1, also known as methyl-CpG-binding domain protein 1, is involved in binding methylated sequences to influence transcription[236]. Specifically, the protein is known to affect chromatin modification through its interaction with H3K9 methyltransferase, which is the enzyme responsible for laying down H3k9me3 methylation marks[237]. Additionally, the location of the skipped exon corresponds to the last exon in every MBD1 isoform (Table 8).

Table 8: MBD1 last exon isoforms

MBD1 Isoform Exon Splicing
MBD1_ENST00000382948.9_1_exon_16_of_16
MBD1_ENST00000353909.7_1_exon_15_of_15
MBD1_ENST00000591416.5_1_exon_16_of_16
MBD1_ENST00000269468.9_1_exon_16_of_16
MBD1_ENST00000347968.7_1_exon_15_of_15
MBD1_ENST00000585595.5_1_exon_17_of_17
MBD1_ENST00000589541.5_1_exon_7_of_7
MBD1_ENST00000398495.6_1_exon_15_of_15
MBD1_ENST00000457839.6_1_exon_17_of_17
MBD1_ENST00000398493.5_1_exon_15_of_15
MBD1_ENST00000398488.5_1_exon_13_of_13
MBD1_ENST00000591535.5_1_exon_13_of_13

H3k9me3 methylation peaks were also overlapped with differentially expressed genes and a total of seventeen out of 141 genes were found to overlap at least one of these peaks. Overall, it was interesting to note the intersection of H3k9me3 methylation marks with both differentially spliced and expressed genes. Of particular interest is that RBMX depletion appears to affect the splicing of the MBD1 gene, which may play a role in regulating the establishment of H3k9me3 marks (Table 9).

Table 9: H3k9me3 marks overlapping differentially expressed genes

Differentially Expressed Genes Overlapping H3K9me3 Methylation Peaks
JPH1
AFAP1-AS1
FAM46A
SLC38A1
SLC39A10
CUX2
GGTA1P
PRKAR1B
GALNT12
SLAMF1
NCS1
SLC16A7
NLRC3
SIPA1L2
IL18R1
UBE3C
CKAP4

Chapter 4: Musashi2

MUSASHI2:

MUSASHI2 is an RNA-binding protein that is expressed broadly across human tissues[238,183]. It was noted to have a role in regulating tissue stem cell processes as well as being able to act as a translational inhibitor[239]. MUSASHI2 appears to play a general role in tissue stem cell biology. It has been shown to affect the development of embryonic stem cells and influences control over the proliferation of neuronal progenitors in the developing central nervous system[240,241]. Crucially, MUSASHI2 has also been implicated in regulating hematopoietic stems cells[183].

The origin cells of the hematopoietic system consist of the long-term hematopoietic stem cells, the short-term hematopoietic stem cells and the multipotent progenitor cells[242,243]. These cells give rise to the myeloid and lymphoid linages and their controlled asymmetric differentiation is essential for establishing normal populations of myeloid and lymphoid cells[242,243]. Specifically, it is the stem cell self-renewal and subsequent differentiation processes that must be tightly regulated in ensure normal hematopoietic homoeostasis. MUSASHI2 appears to regulate the proliferation rate of hematopoietic stem cells[239]. In fact it was shown that if musashi2 levels are decreased a reduction in hematopoietic stem cell primitive progenitors cells follows[238].

MUSASHI2's role as a regulator of hematopoietic stem cell proliferation suggests that disruption of its activity could affect the normal biology of the hematopoietic cells. Indeed it has been noted that MUSASHI2 plays a critical role in the pathologies of both acute myeloid leukemia and chronic myeloid leukemia[183,244–246]. In both diseases, MUSASHI2 dysregulates stem cell proliferation and appears to inhibit the differentiation of myeloid cells thus contributing to the pathology of leukemia.

Myelodysplastic Syndromes:

Myelodysplastic syndromes (MDS) are a constellation of symptoms that are indicative of bone marrow failure[247,248]. The syndrome is generally accepted to be an acquired disorder. The pathology is thought to be linked to environmental or chemical exposures to compounds such as xylene, petroleum related hydrocarbons, agent orange, mercury, lead, and even tobacco smoke[249,250]. Furthermore, patients with prior exposure to chemotherapy, specifically alkylating chemotherapeutics, or heavy doses of ionizing radiation have a higher likelihood of developing MDS[251]. Despite numerous associations with various environmental toxins, the direct mechanism underlying MDS pathogenesis remains undetermined.

MDS is insidious in its onset, often developing initially without any symptoms. As the disease progresses a patient can acquire a pancytopenia resulting from bone marrow failure[252,253]. This pancytopenia gives a patient tremendous fatigue, dyspnea, acquired hemophilia, and frequent infection[252,254]. Disease staging and progression is traditionally done using the French-America-British (FAB) classification that classifies MDS by hematological characteristics[255] (Table 10).

Table 10: French-American-British MDS classifications

FAB Classification	Abbreviation	Description
Refractory anemia	RA	<5% myeloblasts in bone marrow, abnormalities seen in red cell precursors
Refractory anemia with ring sideroblasts	RARS	<5% myeloblasts in bone marrow, >15% ring sideroblasts in bone marrow
Refractory anemia with excess blasts	RAEB	5-19% myeloblasts in bone marrow
Chronic myelomonocytic leukemia	CMML	<20% myeloblasts in bone marrow, > $1x10^9$/L monocytes in peripheral circulation
Acute myelogenous leukemia	AML	>20% myeloblasts in bone marrow

A final stage of this classification is chronic myelomonocytic leukemia (CMML). Furthermore, some cases of MDS progress to particularly aggressive forms of acute myelogenous leukemia (AML) that tends to be resistant to chemotherapy[256,257].

The progression of MDS to chronic and acute forms of leukemia suggests that MDS represents a pre or pseudo leukemic state. Though the precise mechanism responsible for MDS pathology is unknown, it is thought to be tied to underlying genomic instability in hematopoietic stem cells[258]. Consequently, it's possible the molecular actors implicated in leukemia progression are shared with MDS. Specifically, the role of MUSASHI2 in MDS has been a topic of general investigation[259]. In the study below, it was shown that MUSASHI2 was required for maintaining activated MDS stem cells. This

finding sheds light on the molecular basis of MDS pathology and potentially even suggests musashi2 as a therapeutic target.

Contribution:

It was my privilege to contribute to the study of musashi2 in MDS pathology in the following ways[260]. Briefly, musashi2 was overexpressed in NHD13 background mice and hematopoietic stem/progenitor cell were extracted for transcriptome profiling analysis via RNA-sequencing. Using RNA-seq data from the laboratory of Dr. Michael Kharas along with publically available microarray data from MDS patients, I demonstrated that the most aggressive (non-leukemic) FAB MDS classification, RAEB, had significantly higher levels of MUSASHI2 expression than the other less aggressive FAB classifications[261].

Furthermore, I converted MUSASHI2 levels to a Z-score value and stratified the survival data in the following way: patients whose MUSASHI2 Z-score was less than negative one were deemed to be MUSASHI2 low, patients with MUSASHI2 Z-scores between and including the range of negative one to one where deemed to be MUSASHI2 normal, and finally those patients with MUSASHI2 levels above one were deemed to be MUSASHI2 high. When the patient survival data was stratified according to MUSASHI2 levels, I demonstrated that there was a significant survival difference between those patients that had high MUSASHI2 levels and low MUSASHI2 levels. Additionally, while the difference in survival between MUSASHI2 high and MUSASHI2 normal was not significant, a survival trend could clearly be seen in the data. This demonstrated that just by looking at MUSASHI2 levels one could hypothesize patient long-term survival.

Additionally, I performed the differential gene expression analysis of musashi2 overexpression in the NHD13 background. I determined the significant genes, accounting for the NHD13 background, by using a generalized linear model. Furthermore, I showed that the control condition, NHD13 condition, and NHD13 with musashi2 over-expression condition all

clustered distinctly from another. These significant genes represented a genetic signature of musashi2 overexpression in the NHD13 background.

The initial gene expression analysis was conducted in a mouse background, and it was desired to see if the genes that were significant in mouse could also be used to cluster patients into distinct cohorts. Using human microarray data I selected human probes that corresponded to the human orthologues of the mouse genes deemed significant due to musashi2 overexpression. Using these human genes I clustered the patient samples and showed that four primary cohorts resulted. One particular cluster was significantly enriched for patients with MUSASHI2 high levels. I further showed that this cluster had significantly worst survival compared to all other clusters.

Overall, my contribution to the study was in computationally correlating musashi2 levels with patient survival and showing that a musashi2 specific genetic signature present in mouse translated to informative clustering of human patients into cohorts. Specifically, the cohort with the highest mean MUSASHI2 level had worst survival compared to all other cohorts.

nature COMMUNICATIONS

ARTICLE

Received 29 Jul 2015 | Accepted 14 Jan 2016 | Published 22 Feb 2016

DOI: 10.1038/ncomms10739 OPEN

MSI2 is required for maintaining activated myelodysplastic syndrome stem cells

James Taggart[1,*], Tzu-Chieh Ho[1,*], Elianna Amin[1,*], Haiming Xu[2,*], Trevor S. Barlowe[1], Alexendar R. Perez[3], Benjamin H. Durham[4], Patrick Tivnan[1], Rachel Okabe[1], Arthur Chow[1], Ly Vu[1], Sun Mi Park[1], Camila Prieto[1], Christopher Famulare[5], Minal Patel[5], Christopher J. Lengner[6], Amit Verma[7], Gail Roboz[8], Monica Guzman[9], Virginia M. Klimek[5], Omar Abdel-Wahab[4], Christina Leslie[3], Stephen D. Nimer[10] & Michael G. Kharas[1]

Myelodysplastic syndromes (MDS) are driven by complex genetic and epigenetic alterations. The MSI2 RNA-binding protein has been demonstrated to have a role in acute myeloid leukaemia and stem cell function, but its role in MDS is unknown. Here, we demonstrate that elevated MSI2 expression correlates with poor survival in MDS. Conditional deletion of Msi2 in a mouse model of MDS results in a rapid loss of MDS haematopoietic stem and progenitor cells (HSPCs) and reverses the clinical features of MDS. Inversely, inducible overexpression of MSI2 drives myeloid disease progression. The MDS HSPCs remain dependent on MSI2 expression after disease initiation. Furthermore, MSI2 expression expands and maintains a more activated (G1) MDS HSPC. Gene expression profiling of HSPCs from the MSI2 MDS mice identifies a signature that correlates with poor survival in MDS patients. Overall, we identify a role for MSI2 in MDS representing a therapeutic target in this disease.

[1] Molecular Pharmacology and Center for Cell Engineering, Center for Stem Cell Biology, Center for Experimental Therapeutics, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA. [2] Memorial Sloan Kettering Cancer Center, Cancer Biology Program, New York, New York 10065, USA. [3] Computational Biology Program Memorial Sloan Kettering Cancer Center, Memorial Sloan Kettering Institute, New York, New York 10065, USA. [4] Memorial Sloan Kettering Cancer Center, Human Oncology and Pathogenesis Program, New York, New York 10065, USA. [5] Memorial Sloan Kettering Cancer Center, Department of Medicine, Leukemia Service, New York, New York 10065, USA. [6] Department of Animal Biology, Department of Cell and Developmental Biology and Institute for Regenerative Medicine, Schools of Veterinary Medicine and Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [7] Montefiore Medical Center, Albert Einstein College of Medicine, Bronx, New York 10461, USA. [8] Joan and Sanford I. Weill Department of Medicine, Weill Cornell Medical College, New York, New York 10065, USA. [9] Division of Hematology and Medical Oncology, Department of Medicine and Pharmacology, Weill Cornell Medical College, Cornell University, New York, New York 10065, USA. [10] Sylvester Comprehensive Cancer Center, Department of Medicine, Miller School of Medicine, University of Miami, Miami, Florida 33136, USA. * These authors contributed equally to this work. Correspondence and requests for materials should be addressed to M.G.K. (email: Kharasm@mskcc.org).

The majority of haematological disorders involving the myeloid lineage are thought to be of stem cell origin, including myeloproliferative diseases, myelodysplastic syndromes, acute myeloid leukaemia and acquired or heritable bone marrow failure syndromes[1–3]. In each instance, dysregulation of normal stem cell function is thought to contribute to the disease phenotype. Moreover, stem cell characteristics are modulated by a variety of developmental pathways and regulators. Recent studies of MSI2 in normal and malignant hematopoietic stem cell (HSC) biology suggested that MSI2 might play a role in myelodysplastic syndromes (MDS)[4–11]. It was previously reported that *MSI2* expression in MDS was reduced in patients with low-risk and high-risk MDS compared with normal CD34 cells[7]. However, in this study there was a subset of MDS patients with excess blasts with increased *MSI2* (ref. 7). The functional importance of MSI2 in MDS therefore remains unclear. We examine previously published expression data sets and patient samples to find that MSI2 is increased in high-risk MDS patients. Additionally, we utilize MSI2 loss and gain of function approaches in the context of a mouse model of MDS and find that MSI2 is required for MDS.

Results

Elevated MSI2 expression predicts poor survival in MDS. In our examination of a previously published expression data set, we found that *MSI2* expression was increased in $CD34^+$ population in high-risk MDS patients (refractroy anemia with excess blasts; RAEB) compared with healthy individuals that were not age matched or Low-Risk MDS (Refractory Anemia; RA or refractory anemia with ringed sideroblasts; RARS), Fig. 1a)[12]. Elevated MSI2 levels correlated with a poor clinical survival (Fig. 1b and Supplementary Fig. 1a). In line with the microarray data, high-risk MDS patients had increased intracellular MSI2 in their $CD34^+CD38^-$ cells compared with low-risk MDS patients and healthy individuals (Fig. 1c,d). Altogether, the MDS patient data suggests that the level of *MSI2* expression correlates with disease subtype and clinical outcome. In contrast to the acute myelogenous leukemia (AML) patient data, where elevated *MSI2* expression correlates with FLT3-ITD/NPM1 mutations[5,8,9,11], MDS patients do not typically harbour these mutations. Due to the low number of patients with recurrent mutations in this study, we are unable to correlate MSI2 levels with individual mutations (Supplementary Table 1).

Msi2 is required for MDS. To test if Msi2 could be functionally important in MDS, we utilized a murine model of MDS. The *NUP98-HOXD13* transgenic model (*NHD13*) recapitulates many of the salient features of MDS, including neutropenia, lymphopenia and hypercellular or normocellular bone marrow at 4–7 months[13–16]. Also, 12–17% of the marrow contains dysplastic erythroid, myeloid and rare megakaryocytic cell types[13]. Similar to patients with MDS, a significant cohort of the primary mice can progress and develop an aggressive AML. However, if the bone marrow of *NHD13* mice is transplanted, the recipient animals succumb to a fully penetrant but non-lethal form of MDS that rarely progresses to AML (ref. 15). Although the *NHD13*

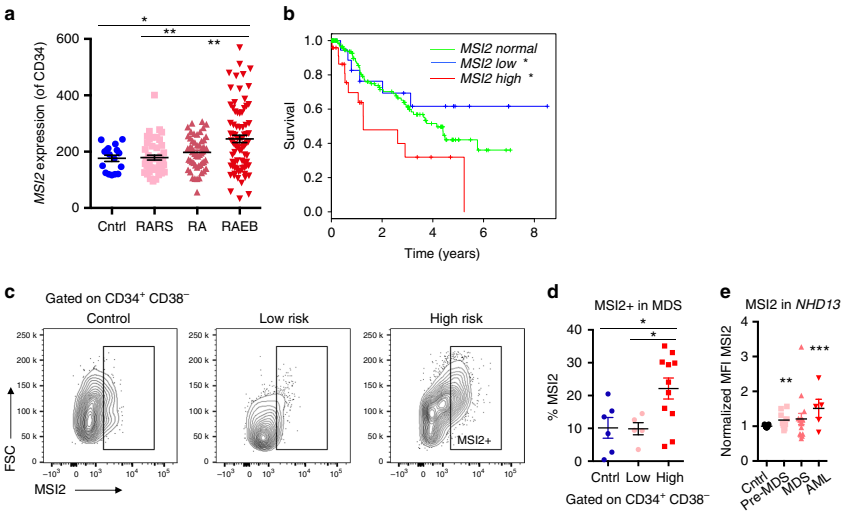


Figure 1 | Elevated MSI2 expression predicts poor survival in MDS. (**a**) Microarray expression data ($CD34^+$ population) from normal elderly individuals ($CD34+$; $n=17$) and MDS ($n=183$), RARS, RA, RAEB, $^*P<0.05$, $^{**}P<0.01$, $^{***}P<0.001$ Student's *t*-test mean ± s.e.m[12]. (**b**) Overall survival in MDS patients stratified by *MSI2* expression (as high (*Z*-score > 1), low (*Z*-score < −1), or normal (−1 ≤ *Z*-score ≤ 1) log-rank test. (**c**). Representative flow cytometric analysis of independent patient cohort of primary patients samples gated on $CD34^+CD38^-$ and stained for intracellular MSI2, age-matched elderly individuals ($n=6$), low risk (RA or RARS; $n=5$) and high risk (RAEB-1, RAEB-2 $n=10$) (**d**). MSI2 positivity summarized from gating of patients in **c**. (**e**) MSI2 intracellular levels in the *NHD13* MDS/AML animal model. Cells are initially gated on MSI2-positive cells (Supplementary Fig. 1 for gating) and median fluorescence intensity (MFI) is normalized to the control (C57BL6 mice), $n=19$. Pre-MDS represents the analysis preformed in *NHD13* primary or transplanted mice within 1–2 months of birth and before MDS onset ($n=9$), the MDS mice that are older than 2 months or primary transplanted and have low WBC, ($n=16$) and AML samples are from *NHD13* mice that have transformed to AML ($n=5$), **d** and **e** $^*P<0.05$, $^{**}P<0.01$, $^{***}P<0.001$ Student's *t*-test horizontal line is the mean ± s.e.m.

transplanted bone marrow cells engraft poorly, they still retain the clinical features of MDS (~10–20% peripheral blood chimerism)[15]. Utilizing intracellular staining for MSI2, we found a significant albeit modest increase in MSI2 levels in the bone marrow of 44% of NHD13 pre-MDS, 50% of MDS, and 80% of AML animals (Fig. 1e and Supplementary Fig. 1b). The significant increase in MSI2 was also observed within the sorted progenitors from pre-MDS animals (Supplementary Fig. 1c,d).

In agreement with MDS patient data, we observed an increase in the expression of MSI2 in the *NHD13* mice during disease progression. These data suggested that altering MSI2 levels in the *NHD13* model could alter the disease fate. To test this hypothesis, *Msi2* conditional knockout were crossed with the *NHD13* mice and then transplanted into congenic recipients (Fig. 2a,b). The chimerism in the peripheral blood and at the level of the haematopoietic stem and progenitor cell (HSPC) was significantly reduced one month after pIpC-mediated deletion (Fig. 2c,d). *Msi2* deletion resulted in the loss of the *NHD13*-expressing cells and a reversal of MDS-like disease that included an increase in white blood cell (WBC) counts, red blood cells and platelets (Fig. 2e–g). When the mice were analyzed 14 months after transplantation there was a trend towards reduced spleen weight, normalized WBC counts and significantly reduced chimerism in the HSPCs and in progenitors (Supplementary Fig. 2a–c). Despite the fact that some of the mice had detectable donor chimerism, the donor myeloid cells were mainly absent and the few remaining donor cells retained MSI2 expression indicating that the deleted cells were selected against (Supplementary Fig. 2d,e). Nevertheless, these mice did not have detectable dysplastic cells or leukaemia in their bone marrow (Supplementary Fig. 2f).

MSI2 overexpression in MDS drives transformation. We next assessed if forced MSI2 expression could alter the disease course using the same model of MDS. Of note, MSI2 overexpression by itself does not result in leukaemic transformation[5]. We utilized our previously described inducible MSI2 overexpressing mouse model that can be controlled with doxycycline (*KH2-Col1A1-tet-on-MSI2/ROSA26*-rTTA) and crossed them with the *NHD13* mice[5]. Control (C57BL/6), *NHD13* or *NHD13/MSI2* bone marrow was transplanted into congenic mice and then allowed to engraft before MSI2 was induced (Fig. 3a). After 5 months the *NHD13/MSI2* overexpression mice started to succumb to lethal myeloid diseases while the *NHD13* mice had symptoms of a mild MDS. The *NHD13/MSI2* mice had reduced WBC (5/25) or elevated WBC counts, reduced red blood cell counts, increased mean corpuscular volume and increased chimerism in the blood at 5 months post-transplantation compared with the control and the *NHD13* mice (Fig. 3b–e). We observed increased immature myeloid cells in the peripheral blood, and all of the MSI2 overexpressing *NHD13* mice eventually succumbed to various lethal myeloid diseases including MPN/MDS or an AML/MDS with a median latency of 228 days (Fig. 3f–h and Supplementary Fig. 3a–g). At end point, we found that the *NHD13/MSI2* mice had a more severe disease burden based on increased spleen and liver weights compared with the control and *NHD13* mice (Fig. 3i,j). The *NHD13* mice

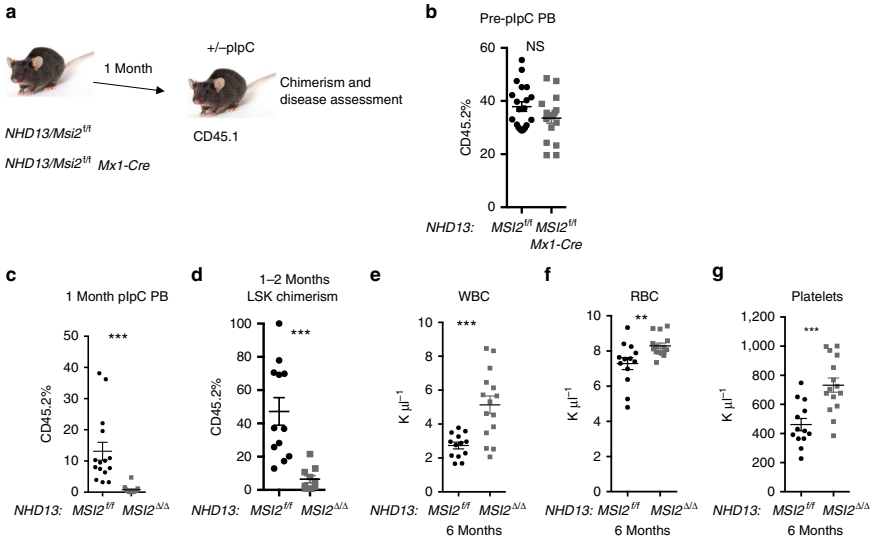


Figure 2 | *Msi2* is required to maintain MDS. (**a**) Transplant scheme for Msi2 conditional knockout in *NHD13* MDS model. (**b**) Peripheral blood chimerism 4 weeks posttransplantation as described in **a**, (*NHD13/Msi2*$^{f/f}$ $n=20$ and *NHD13/MSI2*$^{f/f}$ *Mx1-Cre;* $n=18$). (**c**) Peripheral blood chimerism 8-weeks posttransplantation (4-weeks post-pIpC treatment; *NHD13/Msi*$^{f/f}$ $n=15$ and *NHD13/Msi2*$^{\Delta/\Delta}$*Mx1-Cre;* $n=14$). (**d**) Chimerism within the haematopoietic stem and progenitor cell compartment from bone marrow aspirates performed 8-weeks posttransplantation, gated on Lin-Sca1$^+$Kit$^+$ cells, (*NHD13/Msi2*$^{f/f}$ $n=12$ and *NHD13/Msi2*$^{\Delta/\Delta}$*Mx1-Cre;* $n=10$). (**e**) WBC. (**f**) Red blood cells (RBC). (**g**) Platelets at 6 months posttransplant, **f**–**h**, (analysed at 6 months post-pIpC, *NHD13/Msi2*$^{f/f}$ $n=13$ and *NHD13/Msi2*$^{\Delta/\Delta}$*Mx1-Cre;* $n=15$). Data in **b**–**g** are represented two independent transplants $^*P<0.05$, $^{**}P<0.01$, $^{***}P<0.001$, Student's *t*-test horizontal line is the mean ± s.e.m.

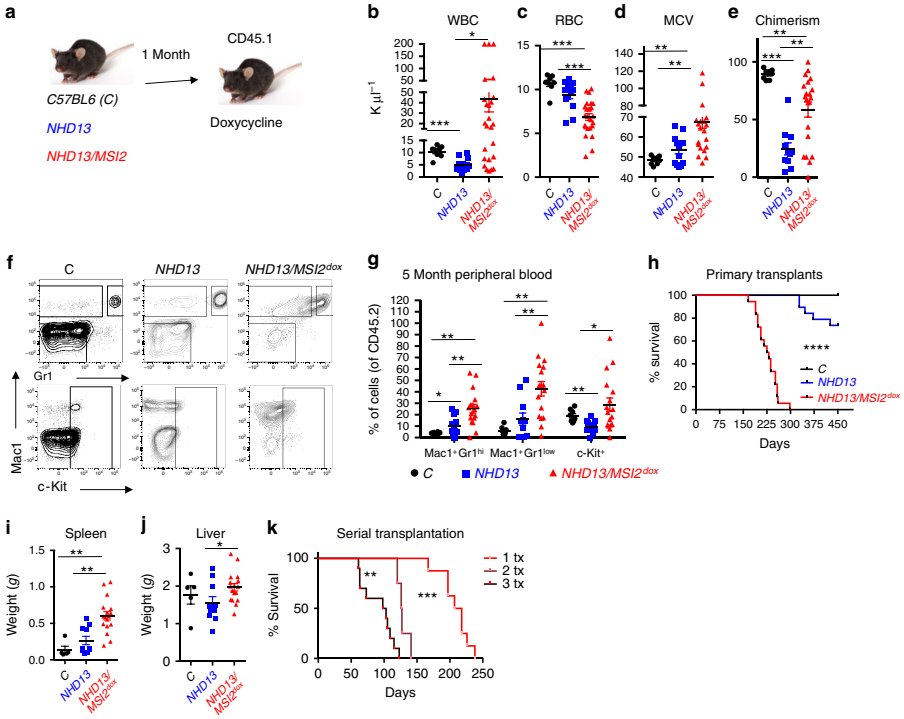


Figure 3 | Sustained MSI2 overexpression transforms MDS to a lethal AML. (**a**) Experimental scheme for MSI2 overexpression in the *NHD13* transplant model. (**b**) Peripheral blood analysis 5 months posttransplantation, including WBC. (**c**) Red blood cell count (RBC). (**d**) Mean corpuscular volume (MCV). Data in **b**–**d**, (C; $n=8$, *NHD13*; $n=13$, *NHD13/MSI2*; $n=25$). (**e**) Chimerism (CD45.2) in the peripheral blood 5 months posttransplantation, (C; $n=8$, *NHD13*; $n=11$, *NHD13/MSI2*; $n=21$). (**f**) Representative Mac1, Gr1, and c-Kit staining of peripheral blood from the bone marrow. Mice were analysed 5 months posttransplant and gated on CD45.2$^+$ cells. (**g**) Immunophenotyping of peripheral blood 5 months posttransplant. Samples gated as in **e**, data in **f** and **g**, (C; $n=8$, *NHD13*; $n=11$, *NHD13/MSI2*; $n=17$). (**h**) Survival curves of *NHD13* with MSI2 overexpression combined from two independent transplants mice (C; $n=8$, *NHD13*; $n=15$, *NHD13/MSI2*^{Dox}; $n=18$). (**i**) Spleen and (**j**) liver weights from healthy mice analysed at the end point or moribund mice, (C; $n=5$, *NHD13*; $n=10$, *NHD13/MSI2*; $n=17$). (**k**) Survival analysis of serially transplanted myeloid disease in the *NHD13/MSI2*^{Dox} mice (primary donor; $n=8$, secondary; $n=4$ and tertiary transplants; $n=10$; transplanted from two independent donors). (**l**) Experimental scheme for testing MSI2 dependence of *NHD13/MSI2* AML. The bone marrow from moribund primary *NHD13/MSI2* transplanted animals was secondarily transplanted into congenic mice and treated accordingly. (**m**) Survival analysis for doxycycline on/off secondary transplants, (Dox^{ON}; $n=8$ and Dox^{OFF}; $n=8$ from two independent donors and transplants), Arrows 1, 2 and 3 indicate three representative mice that are described in **n** and **o**. (**n**) Intracellular MSI2 staining by flow cytometry in fixed bone marrows of representative moribund mice described in **m**. (**o**) Representative CD45.1/2 staining of mice at different points in the survival curve from **m**. (**p**) Survival analysis of doxycycline treated *NHD13/MSI2* secondary transplants relative to transplants whose recipients had a 2-week delay in doxycycline treatment posttransplant, Dox^{ON}; $n=9$ and Dox^{OFF}; $n=10$ from two independent transplants). Data in **b**–**k**, are represented two independent transplants $^*P<0.05$, $^{**}P<0.01$, $^{***}P<0.001$, **b**–**e**, **g**,**i** and **j** are Student's *t*-test and horizontal line is the mean ± s.e.m and **h**,**k**,**m** and **p**, *P* values assessed by log-rank.

showed signs of a MPN/MDS or MDS disease, but only 3 out 15 mice died of a characterized myeloid disease (AML/MDS $n=2$, MDS $n=1$, and one mouse was found dead and another died of a non-myeloid disease; Fig. 3i,j). Serial transplantation of the *NHD13/MSI2* demonstrated reduced latency further supporting the idea that MSI2 overexpression resulted in a clonal myeloid disease (Fig. 3k and Supplementary Fig. 3h). We secondarily transplanted the *NHD13* AMLs and then compared the disease burden to the secondary transplants from the *NHD13/MSI2* mice. Despite the fact that both groups had myeloid disease, the *NHD13/MSI2* group retained their more aggressive phenotype compared with the *NHD13* AMLs indicated by the increased spleen and liver weights (Supplementary Fig. 3i,j).

MSI2 maintains activated MDS stem and progenitor cells. We then determined if MSI2 overexpression was required to maintain the disease. Thus, we transplanted *NHD13/MSI2* overexpressing mice into secondary recipients with or without doxycycline feed. Mice that were maintained on doxycycline and expressed MSI2 rapidly formed a lethal myeloid disease, while the majority of mice that were no longer being induced survived significantly

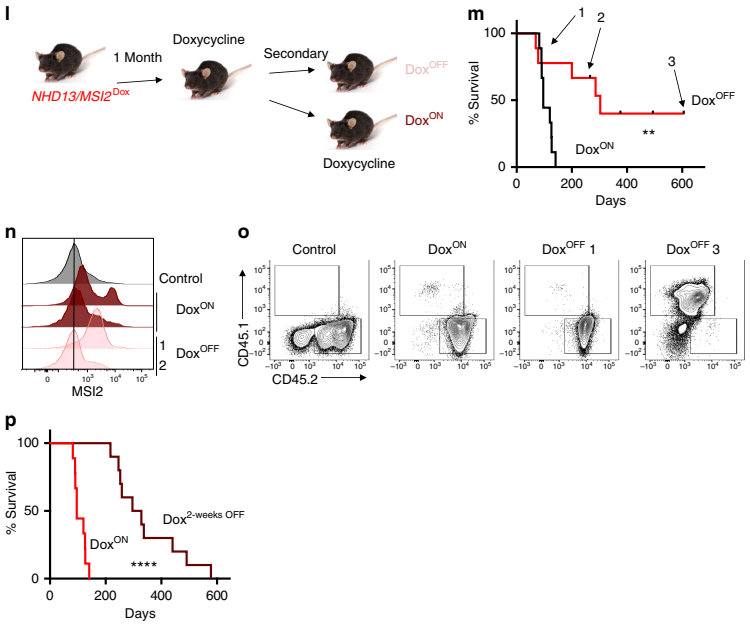


Figure 3 | continued.

longer (median, 96 versus 303 days; Fig. 3l,m and Supplementary Fig. 3k–t). Moreover, in mice that were no longer induced and died at the same time as the mice in the induced group, we were still able to detect high levels of intracellular MSI2 suggesting selection for constitutive activation *in vivo* (Fig. 3n). Similarly, a mouse that died later also demonstrated leaky MSI2 expression albeit at lower levels compared with the mouse that died earlier. Overall, in all the mice that died of leukaemia, MSI2-positive cells were detectable. However, mice killed at the experimental end point that remained disease free, we found that the chimerism was either low or undetectable (Fig. 3o). To further examine if transient withdrawal of MSI2 expression could also delay the leukaemia, we transplanted *NHD13/MSI2* cells and waited 2 weeks to induce MSI2 expression (Fig. 3p). We observed a delay in the myeloid leukaemia (312 days compared with 96 days) in the control, and these leukaemias relapsed with MSI2 positivity, providing evidence that MSI2 overexpression must be sustained to maintain disease.

NHD13 haematopoietic cells have a block in their differentiation at the HSC to multipotent progenitors (MPP) stage and have dramatically reduced numbers of HSPCs (ref. 16; Fig. 4a,b). MSI2 induction for 5 days resulted in an increase in the percentage of phenotypic LSKs (Lin-Sca1$^+$Kit$^+$ cells) and a decrease in the phenotypic HSCs, but no difference in the frequency of the myeloid or erythroid progenitors (Fig. 4a–c and Supplementary Fig. 4a). Interestingly, if the 5-day-induced *NHD13/MSI2* cells were then transplanted in the absence of doxycycline to turn off MSI2, we observed reduced chimerism at 1 month (Fig. 4d). Alternatively, when non-induced bone marrow was engrafted and then activated for MSI2, the LSK compartment was expanded at 1 month (Supplementary Fig. 4b), and in the diseased *NHD13/MSI2* mice LSKs and myeloid progenitors (granulocyte-monocyte progenitor (GMP) and common myeloid progenitor (CMP)) were increased compared with the *NHD13* mice (Fig. 4e and Supplementary Fig. 4c). Similarly to the 5-day induction the chimerism of the phenotypic HSCs (LSK$^+$CD150$^+$;CD48) was reduced in the diseased *NHD13/MSI2* compared with the *NHD13* and the control animals. We then profiled the cell cycle status of the HSPCs using Brdu incorporation and Hoechst staining and found reduced cell death (sub-G1) and increased percentage of cells in G1, which suggests the accumulation of more activated HSPCs (Fig. 4f,g). Taken together with the previous data, MSI2 expression maintains a more aggressive myeloid disease and a more activated HSPC.

To further characterize how MSI2 alters the *NHD13* MDS programme in the dysregulated stem cell compartment, we performed transcriptome profiling in the HSPCs (LSK) from transplanted mice after 3 months of doxycycline administration and before the mice demonstrate any disease phenotype. To elucidate the *NHD13/MSI2* expression programme, we utilized a generalized linear model that identified 891 significant genes (q-value <0.01, generalized linear model), of which 137 genes were upregulated (log2 fold change >0) and 754 genes were downregulated (log2 fold change ≤ 0). We then matched the gene signature to human homologues (690 genes; Supplementary Data 1) and created a heatmap after unsupervised hierarchical clustering, which separated the samples into their respective groups (Fig. 4h).

To functionally annotate our RNA-sequencing, we performed gene set enrichment analysis[17] on all curated gene sets in the molecular signatures database (http://www.broadinstitute.org/msigdb; 3,256 gene sets) combined with an additional set of relevant gene sets (92 gene sets from our experimentally derived or published haematopoietic self-renewal and differentiation signatures[4,17]; rank list; Supplementary Data 2). We found 14 gene sets that were enriched for genes that were upregulated and

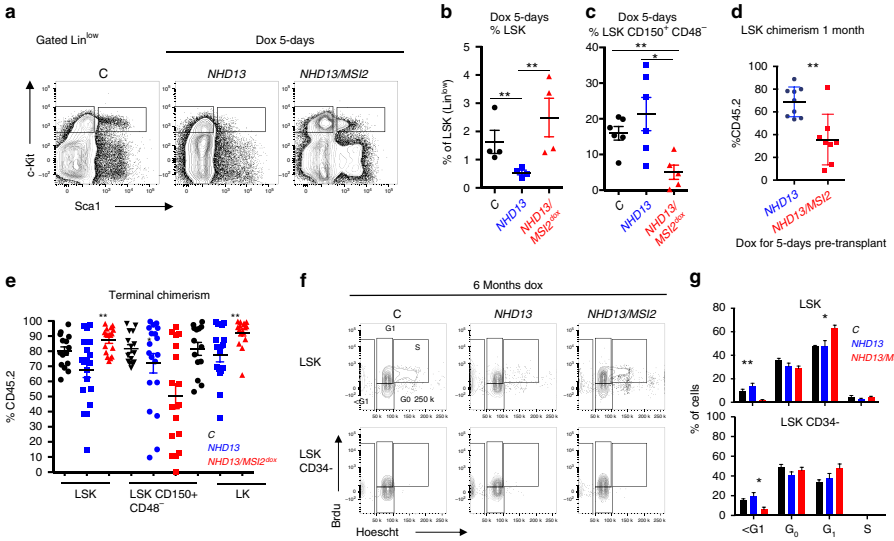


Figure 4 | MSI2 overexpression leads to haematopoietic progenitor stem cell expansion. (**a**) Representative flow cytometry of primary animals treated with doxycycline (Dox) for 5 days and live and Lin$^-$ gated. (**b**) Percentage of LSK population among the Lin^{low} compartment from primary mice (C; *NHD13* and *NHD13/MSI2*^{Dox}; $n=6$, $n=6$ and $n=5$ from five independent experiments). (**c**) Percentage of the HSC (LSK$^+$CD150$^+$CD48$^-$) compartment gated from the LSK$^+$ at 1 month transplantation after 5 day dox administration in the primary animals, data representative (C; *NHD13* and *NHD13/MSI2*^{Dox} same as **b**). (**d**) Chimerism (CD45.2) in the LSK compartment at 1 month transplantation after 5 day dox administration in the primary animals, data representative of two independent transplants (*NHD13*; $n=9$, *NHD13/MSI2*^{Dox}; $n=8$), (**e**) Terminal chimerism from transplants in Fig. 3 in the gated populations, (C, *NHD13*, *NHD13/MSI2*^{Dox}; $n=11$-16 combined from two independent transplants). (**f**) Representative flow cytometric plots from transplanted mice analysed (combined experiments from 3 and 7 months posttransplant) that were injected with Brdu 24 h and then sorted for LSK cells and gated accordingly. (**g**) Data represented in **f**, ($n=3$ for each group combined from two independent experiments). (**h**) RNA-sequencing of LSK sorted cells from primary transplanted animals (4 months posttransplantation and 3 months post MSI2 induction) before disease initiation underwent unsupervised clustering of the differentially expressed genes with human homologues, (C; $n=3$, *NHD13*; $n=2$, *NHD13/MSI2*^{Dox}; $n=4$). (**i**–**k**) Gene set enrichment analysis (GSEA) from ranked list of *NHD13/MSI2*^{Dox}*/NHD13* versus Control. (**l**) Unsupervised clustering of the mouse *NHD13* signature overlapped with MDS patients. (**m**) *MSI2* expression (*Z*-score) separated based on clustering in **l**. (**n**) Survival of MDS patients based on clusters from the *NHD13/MSI2* signature (**l**,**m**). Data in **b**–**e**,**g** and **m** $^*P<0.05$, $^{**}P<0.01$, $^{***}P<0.001$, are Student's *t*-test and the horizontal line is the mean ± s.e.m and **n**, *P* values are displayed and calculated with log-rank.

29 gene sets enriched for downregulated genes (Supplementary Tables 2 and 3). The top ranked gene sets included enrichment in an *NRAS* activated signature[18], a reduced quiescent phenotype[19] and a more progenitor-like cell (Fig. 4i–k). Taken together with our phenotypic analysis of the HSPC compartment, MSI2 induction increases the cells that are in G1, switching them to a less quiescent and more progenitor-like gene expression signature.

To determine if the MSI2 signature from the murine model of MDS corresponds to patients with MDS, we overlapped the *NHD13/MSI2* RNA-seq (690 genes) with microarray data from control ($n=17$) and MDS patients ($n=183$) (refs 12,20). After unsupervised clustering of the human microarray data, we obtained four distinct clusters (Fig. 4l). Patients with elevated MSI2 expression were mainly found in Cluster-2, which predicted a poor survival compared with the other clusters (Fig. 4m,n). Our study demonstrates an important functional role of MSI2 in MDS (Supplementary Fig. 4d).

Discussion

In summary, we found that elevated MSI2 expression predicts poor prognosis in MDS and is required for maintaining the diseased MDS stem cell. Cooperativity with *NHD13* has been associated with various factors including FLT3, MEIS1, P16 and TP53 (refs 16,21–23). MSI2 overexpression can act as a cooperating oncogene and drive transformation, accelerate leukaemia and increase disease burden in the context of a MDS mouse model. Additionally, we found reduced apoptosis and a more activated stem cell, suggesting that the altered HSPC may contribute to disease progression. Gene expression profiling of HSPCs from the *NHD13/MSI2* mice generated a signature that overlapped with human MDS and could predict patient outcome. Our lab previously found that MSI2 directly binds to the mRNA of mixed-lineage leukaemia target genes including *Hoxa9*, *Myc* and *Ikzf2*, and regulates the translation of these targets in a mixed-lineage leukaemia-AF9 leukaemia model[9]. Additionally, a recent report showed that Msi2 may regulate the development and propagation of AML through Tetraspanin 3 (refs 24). Future studies will determine how MSI2 alter stem cells in MDS or whether it uses similar mechanisms as in AML.

Several studies have demonstrated that *Msi2* is required for HSPC engraftment[4,10]. It is unclear if in the context of suppressed hematopoiesis where few normal HSCs remain, targeting could result in additional toxicity. However, we found that the HSPC

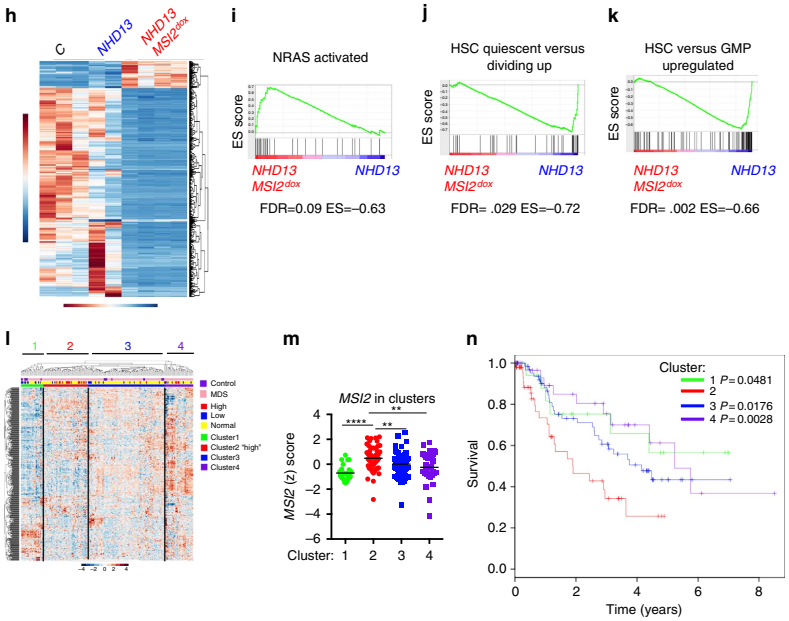


Figure 4 | continued.

population demonstrated a selective advantage and remained addicted to the forced MSI2 expression, as removal of MSI2 overexpression greatly reduced chimerism and reversed the myeloid disease. We propose that the increased expression of MSI2 in the HSPCs in high-risk MDS patients might allow for a therapeutic index in these patients. Our mouse model might provide a context to test how targeting MSI2 might alter the disease. Overall, our study suggests that targeting MSI2 could provide a therapeutic benefit in MDS.

Methods

Transgenic mice. *KH2-Col1A1-tet-on-MSI2/ROSA26*-rTTA transgenic mice[5], backcrossed 10 times to C57BL/6 strain or Msi2 conditional knockout[4] were crossed with Vav-Tg-*NUP98-HOXD13* mice. The primary donors that were used for transplants were either male or female of 3–4-month-old animals[14]. All animal procedures were approved by the Institutional Animal Care and Use Committee at Memorial Sloan Kettering Cancer Center.

Non-competitive transplants. Transplants were performed with 2–3 $\times$ 10^6 bone marrow cells from 12–16-week-old C57BL/6 donor mice mixed with 0.2 $\times$ 10^6 CD45.1$^+$ helper cells injected into the retro-orbital of lethally irradiated B6.SJL-Ptprc^a Pepc^b/BoyJ recipient mice. Secondary transplants were performed by injecting 1 $\times$ 10^6 bone marrow or spleen cells into sublethally irradiated B6.SJL-Ptprc^a Pepc^b/BoyJ mice.

Peripheral blood analysis. Peripheral blood was collected from the facial vein using a lancet or retro-orbital cavity using a heparinized glass capillary tube. A compete peripheral blood count was collected using a Hemavet 950 (Drew Scientific).

Flow cytometry. Flow cytometry experiments were carried out using BD Fortessa, LSRII, or LSRFortessa instruments. Bone marrow and spleen cells collected from mice were subjected to red blood cell lysis before staining. Peripheral blood and leukaemic bone marrow and spleen were immunophenotyped with the following antibodies: CD45.2, CD45.1, Mac1, Gr1, c-Kit, CD71, Ter119 and B220. For stem cell analyses, bone marrow cells were stained with the following antibodies: lineage (Gr1, B220, CD3a, CD4, CD8 and Ter119), Sca1, c-Kit, CD150, CD48, CD16/32 and CD34. MSI2 staining was performed using a rabbit anti-mouse/human MSI2 antibody (Abcam) with a goat anti-rabbit Alexa647 conjugated secondary (Life Technologies). Anti-mouse antibodies were used at 1:200 and secondary antibodies were used at 1:400. Data analyses were performed using the FlowJo software.

Statistical analyses. To compute *P* values for bar graphs, an unpaired 2-tailed Student's *t*-test was used except where stated otherwise. Error bars reflect the s.e.m., except where stated otherwise. In survival curves, significance was calculated using log-rank analysis. Graph Pad Prism 4.0 and the R statistical environment were used to carry out all statistical analyses.

***NHD13* RNA-seq analysis.** *NHD13* mouse RNA-seq raw data were deposited to Gene Expression Omnibus (GEO) GSE76840. Differential analysis of RNA-seq samples utilized the DESeq package for gene expression analysis[25]. False discovery rate correction of *P* values used for all bioinformatics analyses of this study utilized the Benjamini–Hochberg procedure. Mapping between entrez IDs between mouse and human genes was done using the biomaRt R packages[26]. Heatmap clustering and production was done using the heatmap function found within the NMF package in R (ref. 27).

Human data analysis and RNA-seq analysis. The clinical microarray samples consist of 183 and 17 healthy controls from anonymized donors that were not age matched, but included elderly patients who underwent hip replacement. MDS patients and 17 controls samples were publically available on GEO with the reference series tag: GSE19429 (ref. 12). The MDS patient samples were collected from several centres: Oxford and Bournemouth (UK), Duisburg (Germany), Stockholm (Sweden) and Pavia (Italy). This study was approved by the ethics committees (Oxford C00.196, Bournemouth 9991/03/E, Duisburg 2283/03, Stockholm 410/03, Pavia 26264/2002) and informed consent was obtained[12].

The microarray data were downloaded and had gene identifiers in the form of AffyID probes. For *MSI2* we mapped the probes and found that only 4 out of the 9 probes were correctly matched to *MSI2*. We utilized the Spearman correlation coefficient to assess the correlation with the remaining probes and the three probes, which showed a good correlation (1552364_s_at, 243010_at and 243579_at), were then averaged together. These AffyID probes were converted to human and mouse

Entrez IDs, which was done using the biomaRt tool in R. The AffyIDs that had human and mouse Entrez IDs were kept as it indicated that the AffyID corresponded to a human gene that had an orthologue in mouse. In the case that several AffyIDs mapped to a single human Entrez ID the corresponding AffyID rows were combined and a mean of their values were taken. This gave one row of mean expression of all Affy probes that corresponded to a human Entrez gene.

Kaplan–Meier curves were generated to gauge survival probability between the samples that had high, low and normal *MSI2* expression. A heatmap was generated using the statistically significant genes derived from the general linearized model. The rows in this expression matrix were log10 transformed and then a *Z*-score was computed for every element in the row. A heatmap was generated using this matrix of *Z*-scores. This heatmap allowed for row and column clustering. The 200 samples which composed this matrix were labelled according to whether the sample was derived from an MDS patient or a control. Additionally, the samples were labelled according to their French–American–British MDS clinical classification. The samples could take 1 of 4 possible classifications: healthy, RA, RA with excess blasts (RAEB) and RARS. Lastly, the samples were also classified according to their *MSI2* expression levels, which could be classed as high *MSI2* expression, low *MSI2* expression or normal *MSI2* expression. *MSI2* expression was classified as high if the *Z*-score for the *MSI2* in a sample was >1. *MSI2* expression was classified as low if the *Z*-score for the musashi2 gene in a sample was < -1. *MSI2* expression was classified as normal if the *Z*-score for the musashi2 gene in a sample was $-1 \leq \times \leq 1$.

Age-matched normal individuals and primary MDS patient samples. Normal bone marrows from elderly individuals were obtained from hip replacements and MDS patient samples (PBMCs, low risk; RA $n=3$, RARS $n=2$ and high risk; RAEB-1 $n=4$, RAEB-2 $n=6$) were obtained from the Memorial Hospital Tumor Bank under the protocol IRB Waiver Number: WA0260-12 and HBUC: HBS2012060.

Cell Cycle analysis. Before 24 h analysis, mice received an intraperitoneal injection of 1 mg kg^{-1} of BRDU. Mice were killed and Lin-Sca1 + c-Kit + cells were sorted, fixed with 1.6% paraformaldehyde for 15 min, and permeablized with ice-cold methanol. To prevent cell loss, LSKs were mixed with B220 + splenocytes and subsequently stained with CD34 and Hoechst for cell cycle, and then analysed by flow cytometry.

References

1. Shih, A. H. & Levine, R. L. Molecular biology of myelodysplastic syndromes. *Semin. Oncol.* **38,** 613–620 (2011).
2. Will, B. *et al.* Stem and progenitor cells in myelodysplastic syndromes show aberrant stage-specific expansion and harbor genetic and epigenetic alterations. *Blood* **120,** 2076–2086 (2012).
3. Nimer, S. D. Myelodysplastic syndromes. *Blood* **111,** 4841–4851 (2008).
4. Park, S.-M. *et al.* Musashi-2 controls cell fate, lineage bias, and TGF-β signaling in HSCs. *J. Exp. Med.* **211,** 71–87 (2014).
5. Kharas, M. G. *et al.* Musashi-2 regulates normal hematopoiesis and promotes aggressive myeloid leukemia. *Nat. Med.* **16,** 903–908 (2010).
6. Ito, T. *et al.* Regulation of myeloid leukaemia by the cell-fate determinant Musashi. *Nature* **466,** 765–768 (2010).
7. Pereira, J. K. *et al.* Distinct expression profiles of MSI2 and NUMB genes in myelodysplastic syndromes and acute myeloid leukemia patients. *Leuk. Res.* **36,** 1300–1303 (2012).
8. Byers, R. J., Currie, T., Tholouli, E., Rodig, S. J. & Kutok, J. L. MSI2 protein expression predicts unfavorable outcome in acute myeloid leukemia. *Blood* **118,** 2857–2867 (2011).
9. Park, S. M. *et al.* Musashi2 sustains the mixed-lineage leukemia-driven stem cell regulatory program. *J. Clin. Invest.* **125,** 1286–1298 (2015).
10. de Andrés-Aguayo, L. *et al.* Musashi 2 is a regulator of the HSC compartment identified by a retroviral insertion screen and knockout mice. *Blood* **118,** 554–564 (2011).
11. Thol, F. *et al.* Prognostic significance of expression levels of stem cell regulators MSI2 and NUMB in acute myeloid leukemia. *Ann. Hematol.* **92,** 315–323 (2013).
12. Pellagatti, A. *et al.* Deregulated gene expression pathways in myelodysplastic syndrome hematopoietic stem cells. *Leukemia* **24,** 756–764 (2010).
13. Raza-Egilmez, S. Z. *et al.* NUP98-HOXD13 gene fusion in therapy-related acute myelogenous leukemia. *Cancer Res.* **58,** 4269–4273 (1998).
14. Lin, Y.-W., Slape, C., Zhang, Z. & Aplan, P. D. NUP98-HOXD13 transgenic mice develop a highly penetrant, severe myelodysplastic syndrome that progresses to acute leukemia. *Blood* **106,** 287–295 (2005).
15. Chung, Y. J., Choi, C. W., Slape, C., Fry, T. & Aplan, P. D. Transplantation of a myelodysplastic syndrome by a long-term repopulating hematopoietic cell. *Proc. Natl Acad. Sci. USA* **105,** 14088–14093 (2008).
16. Xu, H. *et al.* Loss of p53 accelerates the complications of myelodysplastic syndrome in a NUP98-HOXD13-driven mouse model. *Blood* **120,** 3089–3097 (2012).
17. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102,** 15545–15550 (2005).
18. Croonquist, P. A., Linden, M. A., Zhao, F. & Van Ness, B. G. Gene profiling of a myeloma cell line reveals similarities and unique signatures among IL-6 response, N-ras-activating mutations, and coculture with bone marrow stromal cells. *Blood* **102,** 2581–2592 (2003).
19. Graham, S. M., Vass, J. K., Holyoake, T. L. & Graham, G. J. Transcriptional analysis of quiescent and proliferating CD34 + human hemopoietic cells from normal and chronic myeloid leukemia sources. *Stem Cells* **25,** 3111–3120 (2007).
20. Gerstung, M. *et al.* Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. *Nat. Commun.* **6,** 5901 (2015).
21. Humeniuk, R., Koller, R., Bies, J., Aplan, P. & Wolff, L. Brief report: loss of p15Ink4b accelerates development of myeloid neoplasms in Nup98-HoxD13 transgenic mice. *Stem Cells* **32,** 1361–1366 (2014).
22. Greenblatt, S. *et al.* Knock-in of a FLT3/ITD mutation cooperates with a NUP98-HOXD13 fusion to generate acute myeloid leukemia in a mouse model. *Blood* **119,** 2883–2894 (2012).
23. Slape, C., Liu, L. Y., Beachy, S. & Aplan, P. D. Leukemic transformation in mice expressing a NUP98-HOXD13 transgene is accompanied by spontaneous mutations in Nras, Kras, and Cbl. *Blood* **112,** 2017–2019 (2008).
24. Kwon, H. Y. *et al.* Tetraspanin 3 is required for the development and propagation of acute myelogenous leukemia. *Cell Stem Cell* **17,** 152–164 (2015).
25. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
26. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4,** 1184–1191 (2009).
27. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11,** 367 (2010).

Acknowledgements

We would like to thank Ross Levine for helpful discussions. We would like to thank Aaron Chang for experimental support. We would also like to thank Aly Azeem Khan, Agnes Viale and the MSKCC sequencing core for all their support. M.G.K. is supported by the US National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases; R01DK101989-01A1, National Cancer Institute R01, 1R01CA193842-01, MSK Cancer Center Support Grant/Core Grant (P30 CA008748), Louis V Gerstner Young Investigator Award, Kimmel Scholar Award and V-Scholar Award, Leukemia Research Foundation, C.J.L. was supported by an R01 from the National Cancer Institute (NIH), and a fellowship from the W.W. Smith Charitable Trust.

Author contributions

J.T., T.C.H., E.A., H.X., T.S.B., P.T., R.O., A.C., L.V., S.M.P. and C.P. performed experiments. A.R.P. and C.L. performed computational work and biostatistics. B.H.D. performed histopathology. C.F., M.P., G.R., M.G. and V.M.K. provided clinical MDS samples. C.L. provided critical reagents and discussion. A.V. provided critical project advice. S.D.N. provided support and critical advice; M.G.K. wrote the manuscript, managed the project and supported the work.

Additional information

Accession codes: *NHD13* mouse RNA-seq raw data was deposited to GEO under the accession code GSE76840.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

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

Chapter 5: Ibrutinib

Chronic Lymphocytic Leukemia:

Chronic lymphocytic leukemia (CLL) is a typically indolent liquid malignancy characterized by the uncontrolled expansion and accumulation of B-cell lymphocytes in the bone marrow, blood, and lymph nodes[262]. The aggressive proliferation of B-cells overwhelms the normal constituents cells of the hematopoietic system and thus gives the pathology[263]. Despite the uncontrolled expansion of B-cell lymphocytes, CLL is typically diagnosed through a routine blood test, prior to a patient experiencing any of the symptoms typical of leukemic malignancies such as fatigue, dyspnea, or pancytopenia[264]. Patients with CLL tend to be older and their prognosis varies according to their clinical subtype[265]. However, CLL median survival has a broad range with some patients dying within a couple years to others surviving for decades with the disease[266,267].

While CLL is generally seen as a smoldering leukemia, its prevalence is profound. CLL accounts for twenty-five to thirty percent of all leukemias in the Western world[268]. The median age of CLL diagnosis is seventy, but the medical literature has reports of teenage CLL patients, indicating the disease can affect all age ranges[265,269]. Disease genesis is believed to result from genetic abnormalities in B-cell lymphocytes, but the complete underlying disease mechanism remains poorly understood. Treatments for CLL include cytotoxic chemotherapies along with targeted approaches[270–272]. These targeted drugs primarily go after B-cell surface markers, but a few drugs target molecular pathways critical to cancer cell survival such as the PI3 kinase or Bcl-2 pathways[273–275]. One particular targeted therapy used in the treatment of relapsed and refractory CLL is the drug ibrutinib[276].

Ibrutinib and its Resistance Mechanism:

Ibrutinib is a small molecule inhibitor that covalently binds to Bruton's tyrosine kinase (BTK)[277]. BTK activity is essential for B-cell lymphocyte development and disrupting its activity has clinical benefit for patients

suffering from CLL[278,279]. Specifically, ibrutinib covalently binds the C481 sulfhydryl group of BTK's active site[279]. This irreversible inhibition of the enzyme's active site nullifies BTK's activity and disrupts the production of B-cell lymphocytes. Ibrutinib can force CLL into remission, but 5.3% of all patients on ibrutinib will have disease progression[279]. A mechanism behind this progression was unknown until the present study[279].

In the study a forty-nine year old woman with refractory CLL was placed on ibrutinib and observed to have a positive response to treatment. However, her disease relapsed after twenty-one months despite ibrutinib dose escalation. Two patient blood samples before relapse and after relapse were sent for RNA-sequencing and mutation calling was performed. This analysis demonstrated that a thymidine to adenine mutation in C481 amino acid changed the active site of BTK such that ibrutinib no longer permanently inhibited the binding pocket. With ibrutinib no longer able to inhibit BTK the CLL progressed. This was the first report of a genetic mechanism yielding resistance to ibrutinib[279].

Contribution:

It was my privilege to contribute to the discovery of a genetic resistance mechanism to ibrutinib in the following ways. Initially, when the patient's data was delivered to the laboratory of Dr. Christina Leslie, a targeted search of a pre-selected set of genes was investigated to see if any mutation was present. This initial survey of genes included BTK and the C481S mutation was described. However, to ensure that no other mutations could possibly account for the resistance phenotype, I performed a genome-wide mutation analysis. In this analysis I compared the patient's pre-relapsed samples against her post-relapsed samples and called mutations between the two conditions. The analysis was done on a genome-wide scale with the final result being that no notable genetic changes were observed between the pre and post relapse conditions aside from the C481S mutation. Consequently, the patient's resistance to ibrutinib was determined to be caused by the identified C481S mutation.

portant questions both about the nature of the evidence and about its sufficiency, a topic that has been the subject of inquiry and discussion in the patient-safety community for well over a decade.[1] As noted in the editorial, replications have confirmed substantial effects regarding the use of a surgical checklist, but rigorous randomized trials have not been carried out and are unlikely to be. In contrast to the relatively simple act of providing a new drug or procedure, implementing the surgical checklist calls for performance of a diverse array of 20 or more actions, which can, and should, vary from one institution to another.

An even more important barrier to performing a randomized trial is that implementation of the checklist almost always requires major culture change. Although culture can (and should) be measured, because of its unique nature in a given operating suite (even among individual rooms), the more relevant comparison after implementation of a checklist is with the prior condition, a before-versus-after study, not with other organizations with very different cultures. The key culture change facilitated by the surgical checklist is the development of highly functioning teams, the value of which is well supported by evidence from many venues in and out of health care.

Weiser and Krummel reemphasize the key learnings from all checklist replication studies: success requires great effort directed toward the implementation process and strong leadership, a point also made by Haynes et al., who in addition note that the 9% reduction in mortality that Urbach et al. report could be a significant trend if mortality was followed for a longer time period.

Robblee notes a potential benefit from implementing the surgical checklist that has been underappreciated in the literature: the identification of near misses, which are defined as errors or malfunctions that might well have caused harm if they had not been intercepted. If these events are analyzed, the underlying process failures (so-called latent failures) can often be identified and the process redesigned to prevent the errors from recurring. Indeed, if and when performance of the surgical checklist is fully institutionalized as an integral part of a culture of everyday teamwork in the operating room, it may turn out that one of its major benefits will be identifying opportunities for process improvement.

Lucian L. Leape, M.D.
Harvard School of Public Health
Boston, MA

Since publication of his article, the author reports no further potential conflict of interest.

1. Leape LL, Berwick DM, Bates DW. What practices will most improve safety? Evidence-based medicine meets patient safety. JAMA 2002;288:501-7.

DOI: 10.1056/NEJMc1404583

Ibrutinib Resistance in Chronic Lymphocytic Leukemia

TO THE EDITOR: Ibrutinib, an inhibitor that binds covalently to C481 of Bruton's tyrosine kinase (BTK), has produced remarkable responses in patients with relapsed and refractory chronic lymphocytic leukemia (CLL).[1-4] However, 5.3% of patients have disease progression, and the mechanism of resistance is largely unknown. Herein we describe the mechanism of resistance in such a case.

A 49-year-old woman had a diagnosis of CLL established in 2000. After the failure of multiple treatments, she began receiving ibrutinib at a dose of 560 mg daily in 2010 as part of a phase 1, dose-escalation study of ibrutinib in B-cell cancers.[1] By month 11, a partial response was achieved with an absolute lymphocyte count of 4530 cells per cubic millimeter. Computed tomography at month 18 showed a marked but incomplete reduction of lymphadenopathy. At month 21, a rapidly rising lymphocyte count and progressive lymphadenopathy were noted. Despite a dose escalation to 840 mg daily, CLL progressed during the next 4 weeks (for details, see the Supplementary Appendix, available with the full text of this letter at NEJM.org). Peripheral-blood samples were collected before ibrutinib administration (day –52), while the patient was having a response to the drug (day 472), when progressive disease was first noted (day 589), and before dose escalation (day 616). Figure S1 in the Supplementary Appendix shows the dates of sample collection in relation to the patient's absolute lymphocyte count over the treatment course.

RNA sequencing revealed a thymidine-to-adenine mutation at nucleotide 1634 of the BTK complementary DNA (cDNA) (GenBank accession number, NM_000061.2), leading to a substitution of serine for cysteine at residue 481 (C481S). The mutation was detected in the samples collected

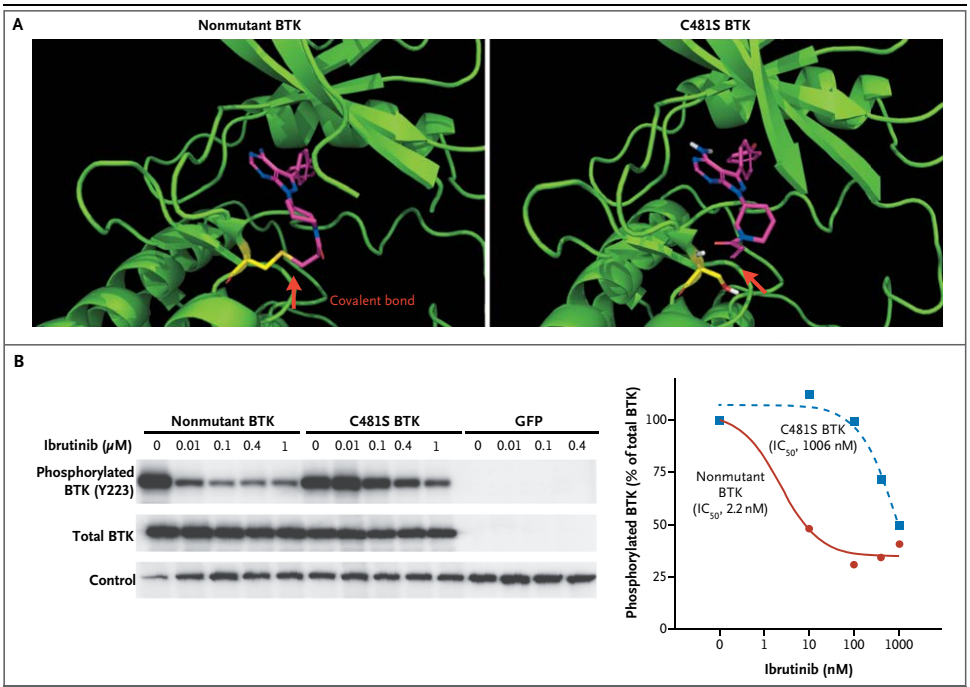


Figure 1. Effect of C481S Mutation of Bruton's Tyrosine Kinase (BTK) on Ibrutinib Binding and the Ability of Ibrutinib to Inhibit BTK Phosphorylation.

Panel A shows structural modeling of nonmutant and mutant BTK with ibrutinib. The red arrows indicate the covalent bond between ibrutinib (purple and blue) and BTK (green and yellow) before and after the mutation. Panel B shows the inhibition of nonmutant BTK or C481S BTK phosphorylation by ibrutinib in HEK 293 cells. The half-maximal inhibitory concentration (IC_{50}) of ibrutinib for inhibition of BTK phosphorylation was analyzed and plotted with GraphPad Prism. GFP denotes green fluorescent protein.

when progressive disease was first noted (88% of reads) and before dose escalation (92% of reads) but not in those collected before ibrutinib administration or while the patient was having a response (Fig. S2A in the Supplementary Appendix). No other genetic changes were identified that correlated with the patient's clinical course in the same manner as the BTK mutation. Sanger sequencing of cDNA verified that the mutation was detected only in the samples collected during relapse (Fig. S2B in the Supplementary Appendix). A more sensitive, allele-specific polymerase-chain-reaction assay (1% analytic sensitivity) detected the mutation in the genomic DNA of samples collected during relapse but not in those collected before ibrutinib administration or while the patient was having a response (Fig. S3 in the Supplementary Appendix).

Ibrutinib binds covalently to the sulfhydryl group of C481 of BTK in the active site, resulting in irreversible inhibition of its kinase activity.[5] Structural modeling suggested that the C481S mutation would disrupt this covalent binding, changing irreversible binding to reversible binding (Fig. 1A). Fluorescently tagged ibrutinib labeled the nonmutant BTK, and the covalent binding that was formed withstood electrophoresis, whereas reversible binding to the C481S or C481A mutant of BTK did not. This showed biochemically the critical role of cysteine in covalent-bond formation (Fig. S4 in the Supplementary Appendix).

Phosphorylation of BTK (pY223) reflects BTK

kinase activity. Introduction of the recombinant nonmutant and C481S BTK constructs into HEK 293 cells showed that phosphorylation of C481S BTK at Y223 became significantly less sensitive to ibrutinib inhibition than the nonmutant BTK did (half-maximal inhibitory concentration, 1006 nM vs. 2.2 nM) (Fig. 1B).

Taken together, our data indicate that the C481S mutation disrupts the covalent binding between BTK and ibrutinib. The impaired binding leads to a loss of inhibition of BTK enzymatic activity that ultimately results in ibrutinib resistance in the patient. Consistent with the findings reported in the *Journal* by Woyach et al.,[6] our studies confirm that BTK is a relevant pharmacologic target of ibrutinib from a genetic perspective.

Richard R. Furman, M.D.
Shuhua Cheng, Ph.D.
Weill Cornell Medical College
New York, NY

Pin Lu, M.D., Ph.D.
University of Chicago
Chicago, IL

Menu Setty, M.S.
Alexendar R. Perez, B.A.
Memorial Sloan-Kettering Cancer Center
New York, NY

Ailin Guo, M.D., Ph.D.
University of Chicago
Chicago, IL

Joelle Racchumi, B.S.
Weill Cornell Medical College
New York, NY

Guozhou Xu, Ph.D.
Hao Wu, Ph.D.
Boston Children's Hospital
Boston, MA

Jiao Ma, Ph.D.
Weill Cornell Medical College
New York, NY

Susanne M. Steggerda, Ph.D.
Pharmacyclics
Sunnyvale, CA

Morton Coleman, M.D.
Weill Cornell Medical College
New York, NY

Christina Leslie, Ph.D.
Memorial Sloan-Kettering Cancer Center
New York, NY

Y. Lynn Wang, M.D., Ph.D.
University of Chicago
Chicago, IL
ylwang@bsd.uchicago.edu

Drs. Furman, Cheng, and Lu contributed equally to this letter.

Supported by grants from the Leukemia and Lymphoma Society and the Prince Family Foundation (both to Dr. Wang).

Disclosure forms provided by the authors are available with the full text of this letter at NEJM.org.

This letter was published on May 28, 2014, and updated on June 6, 2014, at NEJM.org.

1. Advani RH, Buggy JJ, Sharman JP, et al. Bruton tyrosine kinase inhibitor ibrutinib (PCI-32765) has significant activity in patients with relapsed/refractory B-cell malignancies. J Clin Oncol 2013;31:88-94.
2. Byrd JC, Furman RR, Coutre SE, et al. Targeting BTK with ibrutinib in relapsed chronic lymphocytic leukemia. N Engl J Med 2013;369:32-42. [Erratum, N Engl J Med 2014;370:786.]
3. Cheng S, Ma J, Guo A, et al. BTK inhibition targets in vivo CLL proliferation through its effects on B-cell receptor signaling activity. Leukemia 2014;28:649-57.
4. Herman SE, Gordon AL, Hertlein E, et al. Bruton tyrosine kinase represents a promising therapeutic target for treatment of chronic lymphocytic leukemia and is effectively targeted by PCI-32765. Blood 2011;117:6287-96.
5. Honigberg LA, Smith AM, Sirisawad M, et al. The Bruton tyrosine kinase inhibitor PCI-32765 blocks B-cell activation and is efficacious in models of autoimmune disease and B-cell malignancy. Proc Natl Acad Sci U S A 2010;107:13075-80.
6. Woyach JA, Furman RR, Liu T-M, et al. Resistance mechanisms for the Bruton's tyrosine kinase inhibitor ibrutinib. N Engl J Med 2014;370:2286-94.

DOI: 10.1056/NEJMc1402716

CORRECTIONS

Case 12-2014: A 59-Year-Old Man with Fatigue, Abdominal Pain, Anemia, and Abnormal Liver Function (April 17, 2014; 370:1542-50). In Table 2 (page 1547), the mode of inheritance for erythropoietic protoporphyria should have been "Autosomal recessive or X-linked," rather than "Autosomal dominant." The article is correct at NEJM.org.

Case 11-2014: A Man with Traumatic Injuries after a Bomb Explosion at the Boston Marathon (April 10, 2014;370:1441-51). In the legend for Figure 1 (page 1443), the phrase "taken on admission," should be added after "Plain radiographs of the chest (Panel A) and pelvis (Panel B) . . . ," and the phrase "taken after the surgical amputation" should be added after ". . . contrast-enhanced multidetector CT (MDCT) with volume rendering . . . ." The article is correct at NEJM.org.

The Renormalization of Smoking? E-Cigarettes and the Tobacco "Endgame" (January 23, 2014;370:293-5). In Figure 2 (page 294), the bars for use of cigarettes and use of electronic cigarettes should have been shown side by side, rather than stacked, since some students may have been included in both categories. The article is correct at NEJM.org.

Chapter 6: Discussion:

GuideScan

The field of genome engineering has revolutionized the way researchers conduct science. A critical actor in this revolution is CRISPR technology. CRISPR systems allow investigators to efficiently change virtually any genetic locus quickly and easily[90]. Since CRISPR's advent the technology has been applied to numerous fields of biology from neuroscience to cancer science[100,280,281]. It easily allows for the generation of genetic knockouts with sgRNAs targeted against the coding genome. A single advantage of these rapid knockouts is that the augmented pace of genetic screens focused on revealing the genetic actors involved in cellular pathways[97,282,283]. Such screens can now be done through knockout, as opposed to knockdown, studies of target genes. Furthermore, CRISPR gives researchers the ability to dissect the non-coding genome, allowing them to excise or modify long non-coding RNA, enhancer sites, microRNA clusters or single nucleotide polymorphisms with ease[123]. The ability to efficiently and rapidly modify the estimated ninety-eight percent of the genome that does not code for protein is profound[284]. While the overwhelming majority of the human genome is non-coding in nature, a large segment of it undergoes transcription, with some estimates saying as high as over seventy percent[285]. The nature of this transcription and the role it plays in regulating global genetic expression is poorly understood. The ability to understand this element of the genome is greatly advanced by the ability to directly modify it with CRISPR technology.

However CRISPR's potential is not only limited to the research setting. A myriad of diseases from inborn errors of metabolism to cancer have their basis in human genetics. Medicine's ability to combat these aliments is modest, absent a way to treat the underlying genetic condition. CRISPR technologies potentially offer a solution with its ability to target a genetic locus and alter its base sequence. Indeed this potential has already been realized as clinical trials involving CRISPR are already underway in the People's Republic of China and the United States is just about to start their

own[115,286,287]. The ability to manipulate a genome has never been easier. Yet, the power of CRISPR technologies rests on a fundamental assumption: the sgRNA precisely cuts its target.

While previous literature noted several popular sgRNA selection tools failed to return certain worrisome sgRNA off-target sites (such as sites with only one mismatch), the amount of the false negative cut sites returned by these tools was underappreciated[116,124]. The GuideScan software package and web interface illustrated the magnitude of the specificity problem that accompanied these method's returned sgRNAs. The magnitude of underreporting a sgRNA's target space is potentially of particular importance for CRISPR genome-wide screens. Underreporting a target space of a sgRNA can increase the amounts of false positives and false negatives in a screen. For example a sgRNA, thought to be unique, but with multiple perfect target sites could potentially represent a false positive in a negative selection screen. The multiple target sites mean multiple cut sites that could compromise the viability of the cell. The resulting genomic instability would disadvantage the cell and it would appear as a positive result in a negative selection screen. Conversely, this same sgRNA could be a false negative in a positive selection screen for the exact same reason. Consequently, understanding the target space of a given sgRNA is essential for determining sgRNA specificity and therefore the fidelity of the sgRNA to an experimental target. GuideScan provides a solution to this specificity problem in addition to the added flexibility intrinsic to creating custom sgRNA databases.

By construction GuideScan allows for the creation of custom sgRNA database, unique up to a user-defined Hamming distance h, for any arbitrary genome. GuideScan was developed as a general CRISPR database construction tool to facilitate the greater use of CRISPR technologies in the life sciences. While CRISPR has been applied to many model organisms it has not been extended to all. Furthermore, individuals within species are diverse and this diversity exists in everything from cell lines to patients. For CRISPR systems to be applied effectively, the sgRNAs that are used should

be unique to a target genome that varies across individuals. GuideScan provides a solution for this problem in allowing for the generation of custom sgRNA databases for any organism.

GuideScan is a general CRISPR sgRNA database construction software. While it has default parameters that will create a database for the Cas9 enzyme, GuideScan can generate sgRNA databases for any CRISPR system. GuideScan allows a user to specify the location and identity of both canonical and alternative PAM sequences. These PAM sequences can be of any length and there is no limit to the amount of PAM sequences that can be specified. An example of this utility was demonstrated with GuideScan's creation of Cpf1 databases for six model organisms that are hosted on the GuideScan web interface. Regardless of the advances that may come about in the CRISPR field, GuideScan is a generalized software method that is capable of producing sgRNAs to match the advance.

Additionally, the specification of alternative PAM sequences allows investigators to continually develop their CRISPR databases based on new knowledge of existing CRISPR systems. It is known that CRISPR endonucleases tend to recognize an array of PAM sequences with differing efficiencies of binding[117]. For the Cas9 endonuclease the canonical PAM sequence takes the form NGG, but it is also capable of recognizing other PAM sequences, albeit with lower efficiency. In contrast, while the Cpf1 endonuclease has a well-described canonical PAM sequence, its potential alternative PAM sequences are not well described [130]. Additionally, new PAM sequences offering greater target specificity are continually being reported for many CRISPR systems[117]. The ability to regenerate sgRNA databases using new knowledge about existing CRISPR systems is a key advantage that a flexible software package such as GuideScan allows.

Critically, GuideScan exhaustively enumerates the mismatch neighborhoods of all database sgRNAs up to a user defined Hamming distance. The results of this are two fold. First GuideScan produces databases where constituent sgRNAs are guaranteed to be unique up to a

user defined Hamming distance h. Second GuideScan determines the complete set of off-targets for a sgRNA in the database by enumerating off-targets out to a second Hamming distance q (where q > h). This allows a researcher to know the complete target space of a given sgRNA and allows the investigator to design experiments with the confidence that their sgRNAs are uniquely targeting their target site. This complete off-target determination was virtually absent in all other methods and the ability to design sgRNAs databases unique up to a user defined Hamming distance is unique to GuideScan.

The complete determination of off-targets, and the filtering of sgRNAs meeting a user-defined uniqueness standard, makes GuideScan sgRNAs more specific than sgRNAs returned by competing methods. For Cas9 sgRNAs, GuideScan is able to compute target site specificity scores for each sgRNA. This score essentially represents the probability that a given sgRNA will cleave only its target site given information about potential off-target cut sites. This score takes into account the full target space, up to a Hamming distance q, and it is the only score that represents specificity with a full accounting of a sgRNAs' potential target space. GuideScan Cas9 sgRNAs are more specific and more numerous that other competing methods. The fact that GuideScan sgRNAs are more specific than competing tools may also suggest that they are more efficient in cutting their target sites. This added efficiency has nothing to do with the sequence characteristics of the cutting site. Simply, the efficiency may be enhanced because the target space is smaller than for unfiltered sgRNAs. Consequently, the probability of the CRISPR endonuclease binding to a specific target site increases. If cutting efficiency is seen as the probability that a particular target site is cut, then more specific sgRNAs will increase cutting efficiency.

Overall, GuideScan provides a solution to the sgRNA specificity problem in addition to allowing the generation of customizable unique sgRNA databases for any CRISPR system in any organism. It is a generalized sgRNA database construction tool equipped to handle changes in the every

advancing field of CRISPR. Already multiple labs have utilized the software and the web interface receives queries from across the globe. The fact that GuideScan allows for rapid batch queries of unique sgRNA databases represents a substantial advance over other methods, and positions GuideScan as a useful tool to the broader genome engineering community.

RBMX

RBMX is an hnRNP implicated in playing a role regulating the splicing of proteins involved in genome maintenance[184,185,187,189]. Several studies have investigated RBMX's association with various pathologies and it appears to be tied to diseases whose mechanism relies on genomic damage or instability[195–197]. However, the manner through which RBMX influences these diseases remains broadly unknown. The work reported in this thesis sheds some light on the possible means through which RBMX could work in AML.

The knockdown of RBMX globally affects alternative splicing in MOLM13 AML cells. When a gene ontology study was done on all significant differentially spliced genes, the pathways that were highlighted fell into the DNA damage and apoptosis response categories. These pathways were in line with earlier reports that RBMX has a role in maintaining genomic integrity[195–197]. Looking specifically at the alternative splicing events affected by RBMX knockdown it was clear that skipped exon events formed the majority of all significant alternative splicing events. Investigating the genes with significant skipped exons suggested that the pathways affected were again the DNA damage response and apoptosis regulation pathways. Upon deeper investigation it was shown that only those genes that have skipped exon events where the exon was retained in response to RBMX knockdown were the genes capable of producing this pathway result. No other form of alternative splicing event was able to reproduce these pathways, which indicated that the global gene ontology terms were deriving from the significant skipped exon events where exons were being retained in response to RBMX knockdown.

The identities of these genes were interesting as some of them had multiple significant skipped exon events. The gene with the greatest number of significant skipped exon events was CD44, which interestingly has been shown to be crucial for AML pathology[222]. Other interesting genes included POLL and DAXX which both have critical function in DNA repair and apoptosis respectively[226–228]. When a domain analysis was conducted on the significant skipped exon events several of the significantly associated domains illuminated the potential activity the gene was involved with. Specifically, in response to RBMX knockdown the POLL gene undergoes differential splicing at its BRCT domain; a domain implicated in breast cancer pathology[229,230].

RBMX knockdown also affects differential splicing of the MBD1 gene always at its last exon. This gene interacts with H3K9 methyltransferase to lay down H3K9me3 marks in the genome[237]. These marks have a repressive effect on expression[232]. Interestingly, when H3K9me3 marks in k562 cells were overlaid on the significant differential splicing events only three splicing events intersected these marks. All three events were skipped exons where there was increased exon retention in response to RBMX knockdown, and all three marks overlapped distinct genes. One of these genes was MBD1. This suggests that MBD1 may be involved in some form of self-regulation using H3K9me3 marks to repress its own activity as a repressive actor in the genome.

As noted earlier, RBMX levels are elevated in human AML samples. While the exact mechanism by which RBMX is elevated in this malignancy remains to be determined, it is known that RBMX knockdown sensitizes cells to DNA damage[189]. It is possible that RBMX contributes to some form of DNA repair in AML, but its elevation suggests an interesting possibility. Given that RBMX depletion increases a cell's susceptibility to DNA damage, it is reasonable to think that targeting RBMX for knockdown or knockout, either with shRNAs or sgRNAs respectively, in AML cells may sensitize them to chemotherapeutics. This would make the potency of chemotherapy much

greater and could either increase tumor lysis in a patient or reduce the amount of cytotoxic therapy needed to achieve therapeutic effect. AML is an aggressive malignancy, but it may have an Achilles heel in RBMX. A therapeutic pathway may exist with the sequential depletion of RBMX and administration of DNA-damaging chemotherapy.

Musashi2

MUSASHI2 elevation has been noted in various forms of leukemia, but its role in Myelodysplastic Syndrome (MDS) remained largely undefined[183,244–246]. MDS is a disease characterized by bone marrow failure and can result in both chronic and acute forms of leukemia[247,248]. Understanding MDS pathology is critical for prolonging patient survival and safeguarding patients from having their MDS undergo transformation to leukemia. Showing that MUSASHI2 levels correlated with disease survival and that a musashi2 signature from mouse accurately clustered human patient samples into disease cohorts that were significantly different in their MUSASHI2 expression, suggested a role for MUSASHI2 as a potential therapeutic target.

Elevated MUSASHI2 levels correlated with worst survival in patients with MDS. As with RBMX, it may be that the targeted knockdown or knockout, with shRNAs or sgRNAs respectively, of MUSASHI2 has a clinical benefit for MDS patients. It is clear that musashi2 affects MDS pathology and its elevation may therefore be a mechanism through which the disease can be treated. Alternatively, because MUSASHI2 levels correlated with survival, MUSASHI2 can potentially serve as a biomarker for the disease that informs clinical treatment. The idea here would be that those patients with higher MUSASHI2 levels would warrant more aggressive treatment, while those with lower levels may warrant more conservative treatment. Overall the role of MUSASHI2 appears critical not only for the pathology of leukemia, but for MDS as well.

Ibrutinib

Ibrutinib is a targeted therapy used in refractory and aggressive cases of CLL[276]. It has a high patient response rate, but resistance to the drug does occur at a minute, albeit notable frequency[279]. Understanding one such resistance mechanism, through mutating the active site of BTK to which ibrutinib binds, informs both CLL and ibrutinib biology. Furthermore, by understanding the exact BTK resistance mutation, structural modeling can be used to inform the design of better small molecule inhibitors. Ultimately, being one of the first groups to describe the C481S resistance mutation, paved the way for clinicians, researchers, and drug developers to better understand ibrutinib treatment and its potential resistance.

Appendix:

Burrows Wheeler Transformation and FM-Index:

The Burrows Wheeler transformation (BWT) is a method to convert a string into a permutation of itself in an efficient and reversible manner[288]. The main purpose of the BWT is to allow strings to be compressed and amenable to indexing. These attributes are critical for the use of strings in an FM-index. To best understand BWT it is useful to walk through an example of the process.

The set of possible characters shall be restricted to A,C,G,T, which are the symbols for the DNA nucleotides adenine, cytosine, guanine, and thymine respectively. Furthermore, within this set a lexicographical ordering will be imposed such that:

$$A < C < G < T$$

Additionally, one other character, $, will precede all the other characters. This character shall be called the terminator character and will be found naturally at the end of every string.

$$\$ < A < C < G < T$$

Now imagine we have a string, S, such that S = AGTC$. Taking the final character of the string and pre-appending it to the front of the string will give a single rotation of the string S. Computing all rotations of S will give a matrix of rotations.

$AGTC
C$AGT
TC$AG
GTC$A

AGTC$

Sorting this rotation matrix lexicographically yields a shorted rotation matrix.

$AGTC
AGTC$
C$AGT
GTC$A
TC$AG

Extracting the final column of this rotation matrix yields the BWT of string S

S = AGTC$
BWT(S) = C$TAG

When the BWT is applied to a set of strings, those strings will all be sorted lexicographically and these strings can be compressed using run-length encoding procedure. Compression of data is only useful if it can be reversibly decompressed without loss of information. The BWT is reversible through the construction of the rotation matrix. To best understand this reversibility it is important to return to an example string. Imagine there is a string Q such that Q = AATGGC$. Construct a rotation matrix of Q, but before doing this add subscripts to each character enumerating its occurrence in the original string. Consequently, Q = $A_0A_1T_0G_0G_1C_0$. The subscripts do not affect ranking: lexicographically $A_0 = A_1$ in terms of ranking order. The terminator character will not have a subscript since it can only occur once per string.

$$\$A_0A_1T_0G_0G_1C_0$$
$$C_0\$A_0A_1T_0G_0G_1$$

$G_1C_0\$A_0A_1T_0G_0$

$G_0G_1C_0\$A_0A_1T_0$

$T_0G_0G_1C_0\$A_0A_1$

$A_1T_0G_0G_1C_0\$A_0$

$A_0A_1T_0G_0G_1C_0\$$

After sorting the resulting rotation matrix

$\$A_0A_1T_0G_0G_1C_0$

$A_0A_1T_0G_0G_1C_0\$$

$A_1T_0G_0G_1C_0\$A_0$

$C_0\$A_0A_1T_0G_0G_1$

$G_1C_0\$A_0A_1T_0G_0$

$G_0G_1C_0\$A_0A_1T_0$

$T_0G_0G_1C_0\$A_0A_1$

Inspecting the last column of the sorted rotation matrix and breaking up the elements by characters the following order is observed:

A: A_0, A_1 | **C:** C_0 | **G:** G_1, G_0 | **T:** T_0

Doing the same for the first column of the sorted rotation matrix results in:

A: A_0, A_1 | **C:** C_0 | **G:** G_1, G_0 | **T:** T_0

The orderings are identical. This property is characteristic of BWT matrices and is known as LF-Mapping. This ultimately allows the BWT to be reversible. Extract the last column of the BWT(Q) rotation matrix to get:

$C_0\$A_0G_1G_0T_0A_1$

Now re-rank the string according to character occurrence in BWT(Q) [B-rank]

$$C_0\$A_0G_0G_1T_0A_1$$

Look at the sorted BWT rotation matrix once more and, for each row, identify the first and last character. Additionally, add the B-rank that corresponds to each character in the last column of the sorted BWT rotation matrix (Table 11).

Sorted BWT Rotation Matrix

$$\$A_0A_1T_0G_0G_1C_0$$
$$A_0A_1T_0G_0G_1C_0\$$$
$$A_1T_0G_0G_1C_0\$A_0$$
$$C_0\$A_0A_1T_0G_0G_1$$
$$G_1C_0\$A_0A_1T_0G_0$$
$$G_0G_1C_0\$A_0A_1T_0$$
$$T_0G_0G_1C_0\$A_0A_1$$

Table 11: Burrows Wheeler Transform Reversal

First Character in BWT Rotation Matrix Row	Last Character in BWT Rotation Matrix Row	B-Rank
$	C_0	0
A_0	$	0
A_1	A_0	0
C_0	G_1	0
G_1	G_0	1
G_0	T_0	0
T_0	A_1	1

From this matrix one can reconstruct the original string by looking at the first row. Recall rotations position show those characters occurring to the left (last character column) of a queried character (first character in column). Looking at the first column it is noted that C occurs to the left of $. We see that the B-rank for this C is 0 indicating that it is the first C seen in the string. To determine which character is leftward of C we go to the row in the above matrix that begins with C and find that G occurs to the left. The B-rank for this G is 0 so we once again go to the first row in the matrix that starts with G and repeat the process. In this manner the string is reconstructed as follows:

$$C_0\$$$
$$G_1C_0\$$$
$$G_0G_1C_0\$$$
$$T_0G_0G_1C_0\$$$
$$A_1T_0G_0G_1C_0\$$$
$$A_0A_1T_0G_0G_1C_0\$$$

Recall Q = $A_0A_1T_0G_0G_1C_0\$$ and it is evident that the BWT is reversible. Ultimately, the BWT can be utilized to create an efficient data structure that allows for the compression of strings, the indexing of these strings, and the compression of the index itself, which allows for a space efficient and search efficient string data. This efficient data structure is termed the FM-index and can be seen as an extension of the BWT (after the creators: Paolo **F**erragina and Giovanni **M**anzini).

Tree Traversals:

In-Order[289]

In-order traversal begins by starting at the leaf node of the leftmost branch of a tree (Figure 55). Once it visits the leftmost leaf the traversal advances upward one node. If the resulting node has no unvisited children it

is processed. However, if the resulting node has unvisited children the traversal will go to the leaf node of the leftmost unvisited branch and traverse upward. The traversal visits branches from left to right and advances from the leaf nodes to the root node. The traversal continues until all nodes in the tree are visited.

Figure 55: In-order traversal procedure

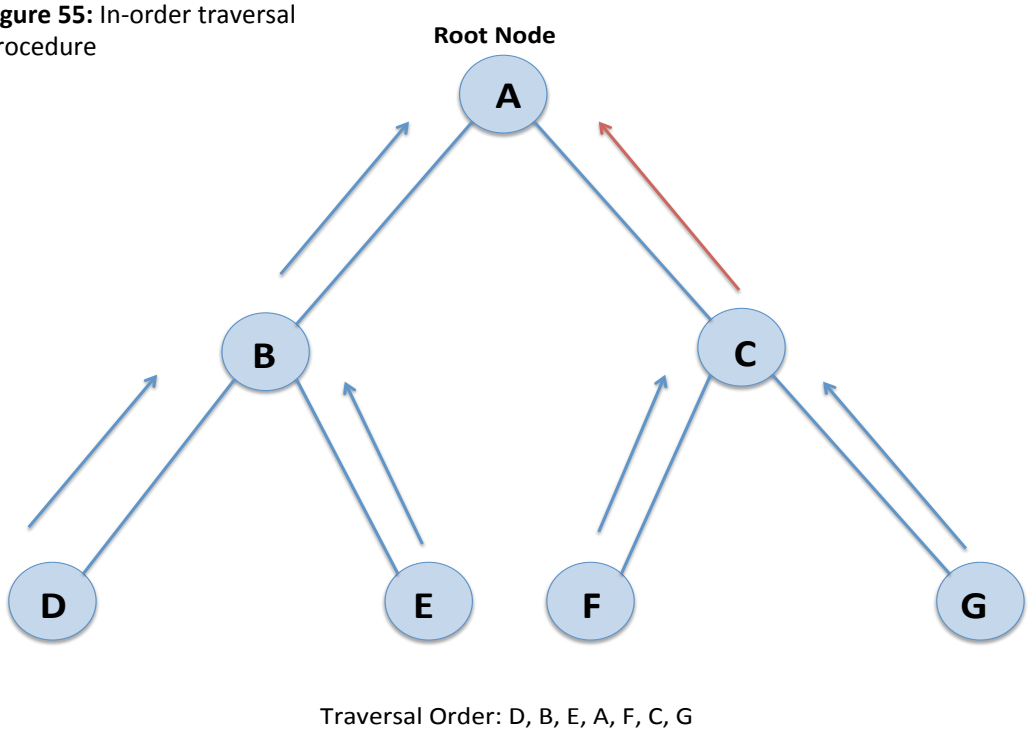
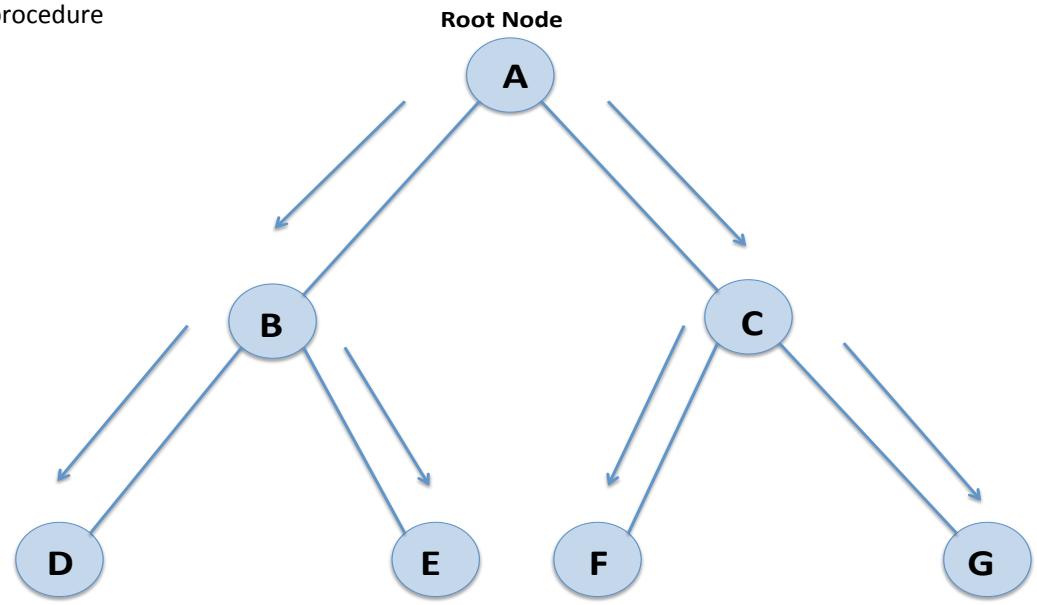


Figure 55: In-order traversal procedure

Pre-Order[289]

Pre-order traversal begins by starting at the root node and transiting down the leftmost branch until the traversal encounters a leaf node (Figure 56). When a leaf node is encountered the traversal backtracks to the nearest parent node with more than one child node and transits down the leftmost branch once more. This procedure is done recursively until all nodes in the tree are visited.

Figure 56: Pre-order traversal procedure



Traversal Order: A, B, D, E, C, F, G

Figure 56: Pre-order traversal procedure

Post-Order[289]

Post-order traversal begins by starting at the leftmost leaf node and transiting upward until a parent node with multiple children nodes are encountered (Figure 57). When such a parent node is encountered the traversal repositions itself at the leftmost unvisited leaf node of the parent node and repeats the process. The procedure will repeat itself recursively until all nodes in the tree are visited.

Figure 57: Post-order traversal procedure

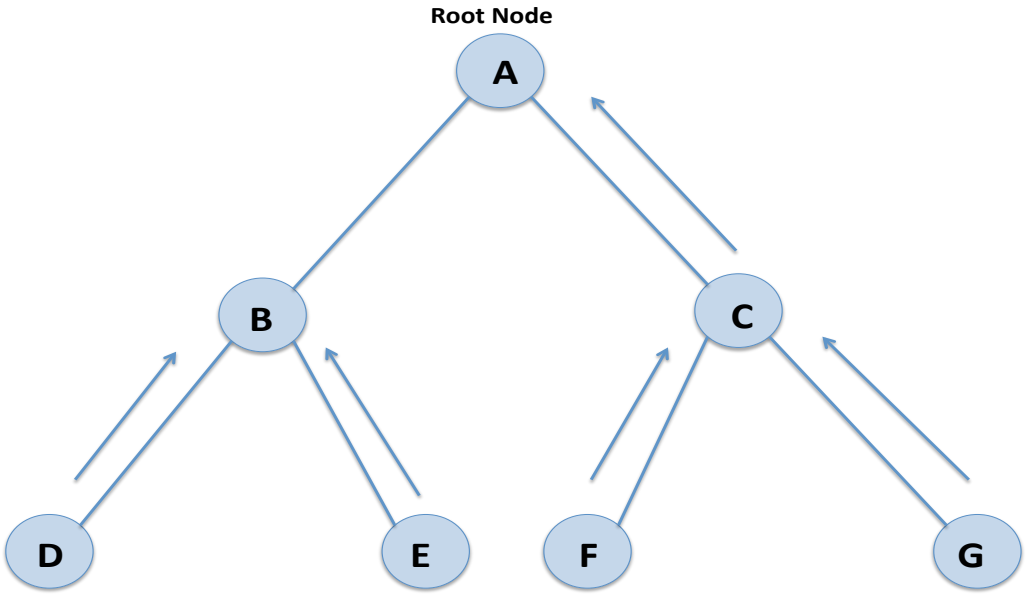


Figure 57: Post-order traversal procedure

Interval Tree:

An interval search tree takes a set of intervals and transforms them into nodes in a search tree (Figure 58)[290]. If an interval has the coordinates [a,b] where a < b then the keys for each node take the value a, even though each node has the information [a,b] stored in it. Additionally, each node stores the maximum value in its subtree and can store additional metadata as well. By construction the left children nodes of a parent have values of a that are less than the parent's value of a. Conversely, the right children nodes of a parent have values of a that are greater than the parent's value of a.

When queries are processed through an interval tree the root node determines if an interval overlaps the root node. If no intersection exists then it takes the query value of a, which shall be called a_q, and determines if a_q is less than the maximum value present in the first node of the left subtree. If a_q is less than the maximum value in the left subtree then the interval moves to the first node in the left subtree and determines if overlap exists. It progresses

down the tree diverting down left and right subtree according the a_q relationship to the subtree's maximum value. This search is done until all overlapping intervals are identified.

Figure 58: Interval tree

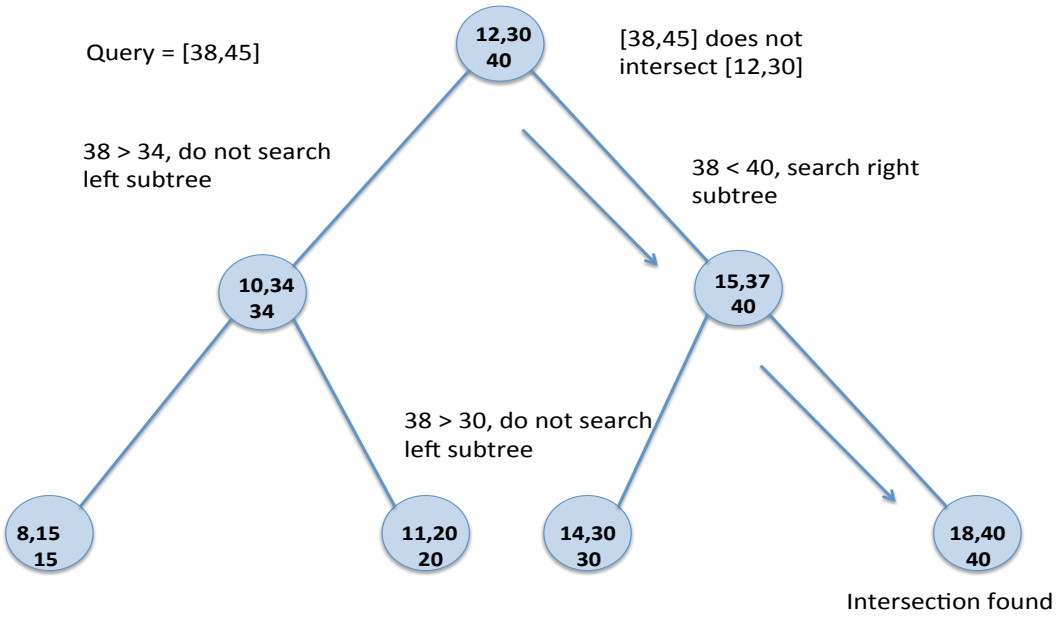


Figure 58: Interval tree diagram

Reverse Phase Protein Array:

Briefly, reverse phase protein array (RPPA) is performed by isolating cell lysate or bodily fluid and placing it on individual segments of a microarray[291]. The set of proteins that are to be quantified is determined a priori and antibodies against these proteins are secured. Across all samples a single antibody is deployed on the microarray[291]. Antibody labeling and quantification can be accomplished through multiple avenues, but all involve the emission of light as an output. This light emission is quantified and is how RPPA's quantitative nature is derived[291]. This procedure is repeated for every protein of interest.

Bibliography

1. Mendel, G. Versuche über Pflanzen-Hybriden. *Verhandlungen des Naturforschenden Vereines, Abhandlungern, Brünn* **4,** 3–47 (1866).
2. Stamhuis, I. H., Meijer, O. G. & Zevenhuizen, E. J. Hugo de Vries on heredity, 1889-1903. Statistics, Mendelian laws, pangenes, mutations. *Isis.* **90,** 238–67 (1999).
3. Simunek, M., Hossfeld, U. & Wissemann, V. 'Rediscovery' revised - the cooperation of Erich and Armin von Tschermak-Seysenegg in the context of the 'rediscovery' of Mendel's laws in 1899-1901. *Plant Biol. (Stuttg).* **13,** 835–41 (2011).
4. Correns, C. E. G. Mendel's Regel über das Verhalten der Nachkommenschaft der Rassenbastarde. in *Gesammelte Abhandlungen zur Vererbungswissenschaft aus Periodischen Schriften 1899–1924* 9–18 (Springer Berlin Heidelberg, 1924). doi:10.1007/978-3-642-52587-2_2
5. Gross, C. Gordon M. Shepherd, Creating Modern Neuroscience: The Revolutionary 1950s. *Soc. Hist. Med.* **24,** 190–191 (2011).
6. Avery, O. T., Macleod, C. M. & McCarty, M. *STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. The Journal of experimental medicine* **79,** (1944).
7. Crick, F. Central dogma of molecular biology. *Nature* **227,** 561–3 (1970).
8. de Wit, E. & de Laat, W. A decade of 3C technologies: insights into nuclear organization. *Genes Dev.* **26,** 11–24 (2012).
9. Li, G.-W. & Xie, X. S. Central dogma at the single-molecule level in living cells. *Nature* **475,** 308–15 (2011).
10. Robinson, V. L. Rethinking the central dogma: noncoding RNAs are

biologically relevant. *Urol. Oncol.* **27,** 304–6

11. Wahlestedt, C. Targeting long non-coding RNA to therapeutically upregulate gene expression. *Nat. Rev. Drug Discov.* **12,** 433–46 (2013).
12. Fu, Y., Dominissini, D., Rechavi, G. & He, C. Gene expression regulation mediated through reversible m^6A RNA methylation. *Nat. Rev. Genet.* **15,** 293–306 (2014).
13. Griffiths AJF, Gelbart WM, Miller JH, et al. Genotypes and Phenotypic Distribution. in *Modern Genetic Analysis* (W. H. Freeman, 1999).
14. Tawata, M., Aida, K. & Onaya, T. Screening for genetic mutations. A review. *Comb. Chem. High Throughput Screen.* **3,** 1–9 (2000).
15. Veltman, J. A. & Brunner, H. G. De novo mutations in human genetic disease. *Nat. Rev. Genet.* **13,** 565–75 (2012).
16. Faruqi, A. F., Seidman, M. M., Segal, D. J., Carroll, D. & Glazer, P. M. Recombination induced by triple-helix-targeted DNA damage in mammalian cells. *Mol. Cell. Biol.* **16,** 6820–8 (1996).
17. Wang, G., Levy, D. D., Seidman, M. M. & Glazer, P. M. Targeted mutagenesis in mammalian cells mediated by intracellular triple helix formation. *Mol. Cell. Biol.* **15,** 1759–68 (1995).
18. Sandor, Z. & Bredberg, A. Deficient DNA repair of triple helix-directed double psoralen damage in human cells. *FEBS Lett.* **374,** 287–91 (1995).
19. Strobel, S. A. & Dervan, P. B. Site-specific cleavage of a yeast chromosome by oligonucleotide-directed triple-helix formation. *Science* **249,** 73–5 (1990).
20. Strobel, S. A., Doucette-Stamm, L. A., Riba, L., Housman, D. E. & Dervan, P. B. Site-specific cleavage of human chromosome 4 mediated by triple-helix formation. *Science* **254,** 1639–42 (1991).
21. Strobel, S. A. & Dervan, P. B. Single-site enzymatic cleavage of yeast genomic DNA mediated by triple helix formation. *Nature* **350,** 172–4 (1991).
22. Wigler, M. *et al.* DNA-mediated transfer of the adenine

phosphoribosyltransferase locus into mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* **76,** 1373–6 (1979).

23. Wigler, M. *et al.* Transformation of mammalian cells with an amplifiable dominant-acting gene. *Proc. Natl. Acad. Sci. U. S. A.* **77,** 3567–70 (1980).
24. Capecchi, M. R. Gene targeting in mice: functional analysis of the mammalian genome for the twenty-first century. *Nat. Rev. Genet.* **6,** 507–12 (2005).
25. Capecchi, M. R. High efficiency transformation by direct microinjection of DNA into cultured mammalian cells. *Cell* **22,** 479–88 (1980).
26. Chang, H. H. Y., Pannunzio, N. R., Adachi, N. & Lieber, M. R. Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nat. Rev. Mol. Cell Biol.* (2017). doi:10.1038/nrm.2017.48
27. Orr-Weaver, T. L. & Szostak, J. W. Yeast recombination: the association between double-strand gap repair and crossing-over. *Proc. Natl. Acad. Sci. U. S. A.* **80,** 4417–21 (1983).
28. Moore, J. K. & Haber, J. E. Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in Saccharomyces cerevisiae. *Mol. Cell. Biol.* **16,** 2164–73 (1996).
29. Boulton, S. J. & Jackson, S. P. Saccharomyces cerevisiae Ku70 potentiates illegitimate DNA double-strand break repair and serves as a barrier to error-prone DNA repair pathways. *EMBO J.* **15,** 5093–103 (1996).
30. Wilson, T. E. & Lieber, M. R. Efficient processing of DNA ends during yeast nonhomologous end joining. Evidence for a DNA polymerase beta (Pol4)-dependent pathway. *J. Biol. Chem.* **274,** 23599–609 (1999).
31. Rodgers, K. & McVey, M. Error-Prone Repair of DNA Double-Strand Breaks. *J. Cell. Physiol.* **231,** 15–24 (2016).
32. Nelson, M. M. C. D. L. *Principles of Biochemisrty*. (2004).
33. Wong, E. A. & Capecchi, M. R. Analysis of homologous recombination in cultured mammalian cells in transient expression and stable

transformation assays. *Somat. Cell Mol. Genet.* **12,** 63–72 (1986).

34. Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, P. W. *Molecular Biology of the Cell*. (Garland Science, 2007).
35. Shrivastav, M., De Haro, L. P. & Nickoloff, J. A. Regulation of DNA double-strand break repair pathway choice. *Cell Res.* **18,** 134–47 (2008).
36. Capecchi, M. R. Altering the genome by homologous recombination. *Science* **244,** 1288–92 (1989).
37. Folger, K. R., Wong, E. A., Wahl, G. & Capecchi, M. R. Patterns of integration of DNA microinjected into cultured mammalian cells: evidence for homologous recombination between injected plasmid DNA molecules. *Mol. Cell. Biol.* **2,** 1372–87 (1982).
38. Hinnen, A., Hicks, J. B. & Fink, G. R. Transformation of yeast. *Proc. Natl. Acad. Sci. U. S. A.* **75,** 1929–33 (1978).
39. Harvey Lodish, Arnold Berk, S Lawrence Zipursky, Paul Matsudaira, David Baltimore, and J. D. *Molecular Cell Biology*. (W. H. Freeman, 2000).
40. Joyner, A. L., Skarnes, W. C. & Rossant, J. Production of a mutation in mouse En-2 gene by homologous recombination in embryonic stem cells. *Nature* **338,** 153–6 (1989).
41. Schwartzberg, P. L., Goff, S. P. & Robertson, E. J. Germ-line transmission of a c-abl mutation produced by targeted gene disruption in ES cells. *Science* **246,** 799–803 (1989).
42. Stoddard, B. L. Homing endonuclease structure and function. *Q. Rev. Biophys.* **38,** 49–95 (2005).
43. Voytas, D. F. Plant genome engineering with sequence-specific nucleases. *Annu. Rev. Plant Biol.* **64,** 327–50 (2013).
44. Fernández-Martínez, L. T. & Bibb, M. J. Use of the meganuclease I-SceI of Saccharomyces cerevisiae to select for gene deletions in actinomycetes. *Sci. Rep.* **4,** 7100 (2014).
45. Epinat, J.-C. *et al.* A novel engineered meganuclease induces

homologous recombination in yeast and mammalian cells. *Nucleic Acids Res.* **31,** 2952–62 (2003).

46. Marcaida, M. J. *et al.* Crystal structure of I-DmoI in complex with its target DNA provides new insights into meganuclease engineering. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 16888–93 (2008).
47. Arnould, S. *et al.* Engineered I-CreI derivatives cleaving sequences from the human XPC gene can induce highly efficient gene correction in mammalian cells. *J. Mol. Biol.* **371,** 49–65 (2007).
48. Chapdelaine, P., Pichavant, C., Rousseau, J., Pâques, F. & Tremblay, J. P. Meganucleases can restore the reading frame of a mutated dystrophin. *Gene Ther.* **17,** 846–58 (2010).
49. Galetto, R., Duchateau, P. & Pâques, F. Targeted approaches for gene therapy and the emergence of engineered meganucleases. *Expert Opin. Biol. Ther.* **9,** 1289–303 (2009).
50. Silva, G. *et al.* Meganucleases and other tools for targeted genome engineering: perspectives and challenges for gene therapy. *Curr. Gene Ther.* **11,** 11–27 (2011).
51. Arnould, S. *et al.* The I-CreI meganuclease and its engineered derivatives: applications from cell modification to gene therapy. *Protein Eng. Des. Sel.* **24,** 27–31 (2011).
52. Fajardo-Sanchez, E., Stricher, F., Pâques, F., Isalan, M. & Serrano, L. Computer design of obligate heterodimer meganucleases allows efficient cutting of custom DNA sequences. *Nucleic Acids Res.* **36,** 2163–73 (2008).
53. Cathomen, T. & Joung, J. K. Zinc-finger nucleases: the next generation emerges. *Mol. Ther.* **16,** 1200–7 (2008).
54. Kim, Y. G., Cha, J. & Chandrasegaran, S. Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 1156–60 (1996).
55. Bitinaite, J., Wah, D. A., Aggarwal, A. K. & Schildkraut, I. FokI dimerization is required for DNA cleavage. *Proc. Natl. Acad. Sci. U. S.*

A. **95,** 10570–5 (1998).

56. Urnov, F. D. *et al.* Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* **435,** 646–51 (2005).
57. Urnov, F. D., Rebar, E. J., Holmes, M. C., Zhang, H. S. & Gregory, P. D. Genome editing with engineered zinc finger nucleases. *Nat. Rev. Genet.* **11,** 636–46 (2010).
58. Sander, J. D. & Joung, J. K. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat. Biotechnol.* **32,** 347–55 (2014).
59. Boch, J. & Bonas, U. Xanthomonas AvrBs3 family-type III effectors: discovery and function. *Annu. Rev. Phytopathol.* **48,** 419–36 (2010).
60. Gaj, T., Gersbach, C. A. & Barbas, C. F. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol.* **31,** 397–405 (2013).
61. Moscou, M. J. & Bogdanove, A. J. A simple cipher governs DNA recognition by TAL effectors. *Science* **326,** 1501 (2009).
62. Boch, J. *et al.* Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **326,** 1509–12 (2009).
63. Christian, M. *et al.* Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics* **186,** 757–61 (2010).
64. Li, T. *et al.* TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. *Nucleic Acids Res.* **39,** 359–72 (2011).
65. Joung, J. K. & Sander, J. D. TALENs: a widely applicable technology for targeted genome editing. *Nat. Rev. Mol. Cell Biol.* **14,** 49–55 (2013).
66. Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. & Nakata, A. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *J. Bacteriol.* **169,** 5429–33 (1987).
67. Mojica, F. J., Díez-Villaseñor, C., Soria, E. & Juez, G. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol. Microbiol.* **36,** 244–6 (2000).

68. Grissa, I., Vergnaud, G. & Pourcel, C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8,** 172 (2007).
69. Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151,** 2551–61 (2005).
70. Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Soria, E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol.* **60,** 174–82 (2005).
71. Pourcel, C., Salvignol, G. & Vergnaud, G. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151,** 653–63 (2005).
72. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315,** 1709–12 (2007).
73. Brouns, S. J. J. *et al.* Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321,** 960–4 (2008).
74. Makarova, K. S., Aravind, L., Wolf, Y. I. & Koonin, E. V. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biol. Direct* **6,** 38 (2011).
75. Makarova, K. S. *et al.* Evolution and classification of the CRISPR-Cas systems. *Nat. Rev. Microbiol.* **9,** 467–77 (2011).
76. Wright, A. V, Nuñez, J. K. & Doudna, J. A. Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. *Cell* **164,** 29–44 (2016).
77. Barrangou, R. & Marraffini, L. A. CRISPR-Cas systems: Prokaryotes upgrade to adaptive immunity. *Mol. Cell* **54,** 234–44 (2014).
78. van der Oost, J., Westra, E. R., Jackson, R. N. & Wiedenheft, B. Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nat. Rev. Microbiol.* **12,** 479–92 (2014).
79. Marraffini, L. A. & Sontheimer, E. J. Self versus non-self discrimination

during CRISPR RNA-directed immunity. *Nature* **463,** 568–71 (2010).

80. Shah, S. A., Erdmann, S., Mojica, F. J. M. & Garrett, R. A. Protospacer recognition motifs: mixed identities and functional diversity. *RNA Biol.* **10,** 891–9 (2013).
81. Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155,** 733–40 (2009).
82. Karvelis, T., Gasiunas, G. & Siksnys, V. Harnessing the natural diversity and in vitro evolution of Cas9 to expand the genome editing toolbox. *Curr. Opin. Microbiol.* **37,** 88–94 (2017).
83. Zhang, Y. *et al.* Comparison of non-canonical PAMs for CRISPR/Cas9-mediated DNA cleavage in human cells. *Sci. Rep.* **4,** 5405 (2014).
84. Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.* **13,** 722–36 (2015).
85. Westra, E. R., Dowling, A. J., Broniewski, J. M. & van Houte, S. Evolution and Ecology of CRISPR. *Annu. Rev. Ecol. Evol. Syst.* **47,** 307–331 (2016).
86. Chylinski, K., Le Rhun, A. & Charpentier, E. The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems. *RNA Biol.* **10,** 726–37 (2013).
87. Deltcheva, E. *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471,** 602–7 (2011).
88. Karvelis, T. *et al.* crRNA and tracrRNA guide Cas9-mediated DNA interference in Streptococcus thermophilus. *RNA Biol.* **10,** 841–51 (2013).
89. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337,** 816–21 (2012).
90. Doudna, J. A. & Charpentier, E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* **346,** 1258096 (2014).
91. Hou, Z. *et al.* Efficient genome engineering in human pluripotent stem

cells using Cas9 from Neisseria meningitidis. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 15644–9 (2013).

92. Esvelt, K. M. *et al.* Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nat. Methods* **10,** 1116–21 (2013).
93. Shi, J. *et al.* Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. *Nat. Biotechnol.* **33,** 661–7 (2015).
94. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339,** 819–23 (2013).
95. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339,** 823–6 (2013).
96. Wang, H. *et al.* One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153,** 910–8 (2013).
97. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343,** 80–4 (2014).
98. Wu, Y. *et al.* Correction of a genetic disease in mouse via use of CRISPR-Cas9. *Cell Stem Cell* **13,** 659–62 (2013).
99. Hsu, P. D., Lander, E. S. & Zhang, F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157,** 1262–78 (2014).
100. Maddalo, D. *et al.* In vivo engineering of oncogenic chromosomal rearrangements with the CRISPR/Cas9 system. *Nature* **516,** 423–7 (2014).
101. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152,** 1173–83 (2013).
102. Larson, M. H. *et al.* CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat. Protoc.* **8,** 2180–96 (2013).
103. Gilbert, L. A. *et al.* CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154,** 442–51 (2013).
104. Guggino, W. B. & Cebotaru, L. Adeno-Associated Virus (AAV) gene

therapy for cystic fibrosis: current barriers and recent developments. *Expert Opin. Biol. Ther.* (2017). doi:10.1080/14712598.2017.1347630

105. Yang, Y. *et al.* Inactivation of E2a in recombinant adenoviruses improves the prospect for gene therapy in cystic fibrosis. *Nat. Genet.* **7,** 362–9 (1994).
106. Holt, N. *et al.* Human hematopoietic stem/progenitor cells modified by zinc-finger nucleases targeted to CCR5 control HIV-1 in vivo. *Nat. Biotechnol.* **28,** 839–47 (2010).
107. Ando, Y. *et al.* Guideline of transthyretin-related hereditary amyloidosis for clinicians. *Orphanet J. Rare Dis.* **8,** 31 (2013).
108. Dryja, T. P. *et al.* A point mutation of the rhodopsin gene in one form of retinitis pigmentosa. *Nature* **343,** 364–6 (1990).
109. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. *Cell* **72,** 971–83 (1993).
110. Mahadevan, M. *et al.* Myotonic dystrophy mutation: an unstable CTG repeat in the 3' untranslated region of the gene. *Science* **255,** 1253–5 (1992).
111. Campuzano, V. *et al.* Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science* **271,** 1423–7 (1996).
112. Riordan, J. R. *et al.* Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245,** 1066–73 (1989).
113. Saiki, R. K. *et al.* Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230,** 1350–4 (1985).
114. Sicinski, P. *et al.* The molecular basis of muscular dystrophy in the mdx mouse: a point mutation. *Science* **244,** 1578–80 (1989).
115. Reardon, S. First CRISPR clinical trial gets green light from US panel. *Nature* (2016). doi:10.1038/nature.2016.20137
116. Tsai, S. Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target

cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.* **33,** 187–97 (2015).

117. Kleinstiver, B. P. *et al.* Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **523,** 481–5 (2015).
118. Cho, S. W. *et al.* Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res.* **24,** 132–41 (2014).
119. Collins, F. S. *et al.* New goals for the U.S. Human Genome Project: 1998-2003. *Science* **282,** 682–9 (1998).
120. Collins, F. S., Morgan, M. & Patrinos, A. The Human Genome Project: lessons from large-scale biology. *Science* **300,** 286–90 (2003).
121. Speir, M. L. *et al.* The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res.* **44,** D717-25 (2016).
122. Tyner, C. *et al.* The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res.* **45,** D626–D634 (2017).
123. Perez, A. R. *et al.* GuideScan software for improved single and paired CRISPR guide RNA design. *Nat. Biotechnol.* **35,** 347–349 (2017).
124. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34,** 184–91 (2016).
125. Doench, J. G. *et al.* Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.* **32,** 1262–7 (2014).
126. Moreno-Mateos, M. A. *et al.* CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat. Methods* **12,** 982–8 (2015).
127. Heigwer, F., Kerr, G. & Boutros, M. E-CRISP: fast CRISPR target site identification. *Nat. Methods* **11,** 122–3 (2014).
128. Bae, S., Park, J. & Kim, J.-S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* **30,** 1473–5 (2014).

129. Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31,** 827–32 (2013).
130. Zetsche, B. *et al.* Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* **163,** 759–71 (2015).
131. Montague, T. G., Cruz, J. M., Gagnon, J. A., Church, G. M. & Valen, E. CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. *Nucleic Acids Res.* **42,** W401-7 (2014).
132. Kleinstiver, B. P. *et al.* Broadening the targeting range of Staphylococcus aureus CRISPR-Cas9 by modifying PAM recognition. *Nat. Biotechnol.* **33,** 1293–1298 (2015).
133. Zhurkin, V. B., Tolstorukov, M. Y., Xu, F., Colasanti, A. V. & Olson, W. K. Sequence-Dependent Variability of B-DNA. in *DNA Conformation and Transcription* 18–34 (Springer US). doi:10.1007/0-387-29148-2_2
134. Cronk, J. Nucleotides and nucleic acids. *Biochemistry* (2016). Available at: http://guweb2.gonzaga.edu/faculty/cronk/CHEM440pub/L04.html. (Accessed: 12th July 2017)
135. Wu, X. *et al.* Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat. Biotechnol.* **32,** 670–676 (2014).
136. Yates, A. *et al.* Ensembl 2016. *Nucleic Acids Res.* **44,** D710-6 (2016).
137. Smit, A. F. The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.* **6,** 743–748 (1996).
138. de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A. & Pollock, D. D. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLoS Genet.* **7,** e1002384 (2011).
139. Kamal, M., Xie, X. & Lander, E. S. A large family of ancient repeat elements in the human genome is under strong selection. *Proc. Natl. Acad. Sci.* **103,** 2740–2745 (2006).
140. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
141. Patro, R. Aligning and Mapping RNA-Seq reads. 63 (2016). Available at: http://robpatro.com/redesign/AlignmentAndMapping.pdf. (Accessed:

5th July 2017)

142. De La Briandais, R. File searching using variable length keys. in *Papers presented at the the March 3-5, 1959, western joint computer conference on XX - IRE-AIEE-ACM '59 (Western)* 295–298 (ACM Press, 1959). doi:10.1145/1457838.1457895
143. Jankowski, R. Advanced data structures by Peter Brass Cambridge University Press 2008. *ACM SIGACT News* **41,** 19 (2010).
144. Hamming, R. W. Error Detecting and Error Correcting Codes. *Bell Syst. Tech. J.* **29,** 147–160 (1950).
145. Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.* **10,** (1966).
146. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
147. Kingsford, C. Interval Trees. *Bioinformatics Lectures* Available at: https://www.cs.cmu.edu/~ckingsf/bioinfo-lectures/intervaltrees.pdf. (Accessed: 10th July 2017)
148. Vidigal, J. A. & Ventura, A. Rapid and efficient one-step generation of paired gRNA CRISPR-Cas9 libraries. *Nat. Commun.* **6,** 8083 (2015).
149. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12,** 656–64 (2002).
150. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
151. Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153,** 307–19 (2013).
152. Kozomara, A. & Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* **39,** D152-7 (2011).
153. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–74 (2012).
154. Martelli, M. P. *et al.* EML4-ALK Rearrangement in Non-Small Cell Lung Cancer and Non-Tumor Lung Tissues. *Am. J. Pathol.* **174,** 661–670

(2009).

155. Lugo, T. G., Pendergast, A. M., Muller, A. J. & Witte, O. N. Tyrosine kinase activity and transformation potency of bcr-abl oncogene products. *Science* **247,** 1079–82 (1990).
156. Johnson, N. A. *et al.* Lymphomas with concurrent BCL2 and MYC translocations: the critical factors associated with survival. *Blood* **114,** 2273–2279 (2009).
157. Argaman, L. *et al.* Novel small RNA-encoding genes in the intergenic regions of Escherichia coli. *Curr. Biol.* **11,** 941–950 (2001).
158. van Bakel, H., Nislow, C., Blencowe, B. J. & Hughes, T. R. Most 'Dark Matter' Transcripts Are Associated With Known Genes. *PLoS Biol.* **8,** e1000371 (2010).
159. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43,** D204-12 (2015).
160. Tilgner, H. *et al.* Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.* **22,** 1616–25 (2012).
161. Lee, Y. & Rio, D. C. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annu. Rev. Biochem.* **84,** 291–323 (2015).
162. Watanabe, Y., Yokobori, S. & Kawarabayasi, Y. [Intron and pre-mRNA splicing of bacteria and Archaea]. *Tanpakushitsu Kakusan Koso.* **47,** 833–6 (2002).
163. Belfort, M. Self-splicing introns in prokaryotes: migrant fossils? *Cell* **64,** 9–11 (1991).
164. Belfort, M. & Weiner, A. Another Bridge between Kingdoms: tRNA Splicing in Archaea and Eukaryotes. *Cell* **89,** 1003–1006 (1997).
165. Tocchini-Valentini, G. D., Fruscoloni, P. & Tocchini-Valentini, G. P. Evolution of introns in the archaeal world. *Proc. Natl. Acad. Sci.* **108,** 4782–4787 (2011).
166. Roy, S. W. & Gilbert, W. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat. Rev. Genet.* **7,** 211–21 (2006).

167. Paushkin, S. V, Patel, M., Furia, B. S., Peltz, S. W. & Trotta, C. R. Identification of a human endonuclease complex reveals a link between tRNA splicing and pre-mRNA 3' end formation. *Cell* **117,** 311–21 (2004).
168. Zillmann, M., Gorovsky, M. A. & Phizicky, E. M. Conserved mechanism of tRNA splicing in eukaryotes. *Mol. Cell. Biol.* **11,** 5410–6 (1991).
169. Belfort, M. Mobile self-splicing introns and inteins as environmental sensors. *Curr. Opin. Microbiol.* **38,** 51–58 (2017).
170. Abelson, J., Trotta, C. R. & Li, H. tRNA Splicing. *J. Biol. Chem.* **273,** 12685–12688 (1998).
171. Pyle, A. M. Group II Intron Self-Splicing. *Annu. Rev. Biophys.* **45,** 183–205 (2016).
172. Lehmann, K. & Schmidt, U. Group II introns: structure and catalytic versatility of large natural ribozymes. *Crit. Rev. Biochem. Mol. Biol.* **38,** 249–303 (2003).
173. Penny, D., Collins, L. J., Daly, T. K. & Cox, S. J. The relative ages of eukaryotes and akaryotes. *J. Mol. Evol.* **79,** 228–39 (2014).
174. Patel, A. A. & Steitz, J. A. Splicing double: insights from the second spliceosome. *Nat. Rev. Mol. Cell Biol.* **4,** 960–70 (2003).
175. Wahl, M. C., Will, C. L. & Lührmann, R. The Spliceosome: Design Principles of a Dynamic RNP Machine. *Cell* **136,** 701–718 (2009).
176. Will, C. L. & Luhrmann, R. Spliceosome Structure and Function. *Cold Spring Harb. Perspect. Biol.* **3,** a003707–a003707 (2011).
177. Fica, S. M. *et al.* RNA catalyses nuclear pre-mRNA splicing. *Nature* **503,** 229–34 (2013).
178. Sammeth, M., Foissac, S. & Guigó, R. A general definition and nomenclature for alternative splicing events. *PLoS Comput. Biol.* **4,** e1000147 (2008).
179. Matlin, A. J., Clark, F. & Smith, C. W. J. Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.* **6,** 386–98 (2005).

180. Black, D. L. Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* **72,** 291–336 (2003).
181. Sakharkar, M. K., Chow, V. T. K. & Kangueane, P. Distributions of exons and introns in the human genome. *In Silico Biol.* **4,** 387–93 (2004).
182. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40,** 1413–5 (2008).
183. Kharas, M. G. *et al.* Musashi-2 regulates normal hematopoiesis and promotes aggressive myeloid leukemia. *Nat. Med.* **16,** 903–908 (2010).
184. Soulard, M. *et al.* hnRNP G: sequence and characterization of a glycosylated RNA-binding protein. *Nucleic Acids Res.* **21,** 4210–7 (1993).
185. Nasim, M. T., Chernova, T. K., Chowdhury, H. M., Yue, B.-G. & Eperon, I. C. HnRNP G and Tra2beta: opposite effects on splicing matched by antagonism in RNA binding. *Hum. Mol. Genet.* **12,** 1337–48 (2003).
186. Heinrich, B. *et al.* Heterogeneous nuclear ribonucleoprotein G regulates splice site selection by binding to CC(A/C)-rich regions in pre-mRNA. *J. Biol. Chem.* **284,** 14303–15 (2009).
187. Matsunaga, S. *et al.* RBMX: a regulator for maintenance and centromeric protection of sister chromatid cohesion. *Cell Rep.* **1,** 299–308 (2012).
188. Lingenfelter, P. A. *et al.* Expression and conservation of processed copies of the RBMX gene. *Mamm. Genome* **12,** 538–45 (2001).
189. Adamson, B., Smogorzewska, A., Sigoillot, F. D., King, R. W. & Elledge, S. J. A genome-wide homologous recombination screen identifies the RNA-binding protein RBMX as a component of the DNA-damage response. *Nat. Cell Biol.* **14,** 318–28 (2012).
190. May, K. M. & Hardwick, K. G. The spindle checkpoint. *J. Cell Sci.* **119,** 4139–42 (2006).
191. Disease, N. I. of A. and M. and S. Handout on Health: Systemic Lupus

Erythematosus. (2016). Available at: https://www.niams.nih.gov/health_info/Lupus/default.asp. (Accessed: 16th July 2017)

192. Mortensen, E. S., Fenton, K. A. & Rekvig, O. P. Lupus nephritis: the central role of nucleosomes revealed. *Am. J. Pathol.* **172,** 275–83 (2008).
193. Kavanaugh, A., Tomar, R., Reveille, J., Solomon, D. H. & Homburger, H. A. Guidelines for clinical use of the antinuclear antibody test and tests for specific autoantibodies to nuclear antigens. American College of Pathologists. *Arch. Pathol. Lab. Med.* **124,** 71–81 (2000).
194. Su, K.-Y. & Pisetsky, D. S. The role of extracellular DNA in autoimmunity in SLE. *Scand. J. Immunol.* **70,** 175–83 (2009).
195. Hirschfeld, M. *et al.* HNRNP G and HTRA2-BETA1 regulate estrogen receptor alpha expression with potential impact on endometrial cancer. *BMC Cancer* **15,** 86 (2015).
196. Shin, K.-H., Kang, M. K., Kim, R. H., Christensen, R. & Park, N.-H. Heterogeneous nuclear ribonucleoprotein G shows tumor suppressive effect against oral squamous cell carcinoma cells. *Clin. Cancer Res.* **12,** 3222–8 (2006).
197. Shin, K.-H. *et al.* p53 promotes the fidelity of DNA end-joining activity by, in part, enhancing the expression of heterogeneous nuclear ribonucleoprotein G. *DNA Repair (Amst).* **6,** 830–40 (2007).
198. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *N. Engl. J. Med.* **368,** 2059–2074 (2013).
199. Prieto, C. *Hemaexplorer Data*. (2017).
200. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
201. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLoS Comput. Biol.* **9,** e1003118 (2013).
202. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).

203. Benjamini Yoav, H. Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc.* **57,** 289–300 (1995).
204. Shen, S. *et al.* rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci.* **111,** E5593–E5601 (2014).
205. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7,** 1009–1015 (2010).
206. D'haeseleer, P. How does gene expression clustering work? *Nat. Biotechnol.* **23,** 1499–1501 (2005).
207. Ringnér, M. What is principal component analysis? *Nat. Biotechnol.* **26,** 303–304 (2008).
208. Angelini, Claudia and De Feis, Italia and Nguyen, Viet Anh and van der Wath, Richard and Liò, P. Combining Replicates and Nearby Species Data: A Bayesian Approach. in *Computational Intelligence Methods for Bioinformatics and Biostatistics: 6th International Meeting, CIBB 2009, Genoa, Italy, October 15-17, 2009, Revised Selected Papers* (ed. Masulli, Francesco and Peterson, Leif E. and Tagliaferri, R.) 191–205 (2010). doi:10.1007/978-3-642-14571-1_14
209. Hu, Y. *et al.* DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. *Nucleic Acids Res.* **41,** e39 (2013).
210. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–515 (2010).
211. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22,** 2008–2017 (2012).
212. Shen, S. *et al.* MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic Acids Res.* **40,** e61 (2012).
213. Wang, X. & Cairns, M. J. SeqGSEA: a Bioconductor package for gene

set enrichment analysis of RNA-Seq data integrating differential expression and splicing. *Bioinformatics* **30,** 1777–9 (2014).

214. Wang, W., Qin, Z., Feng, Z., Wang, X. & Zhang, X. Identifying differentially spliced genes from two groups of RNA-seq samples. *Gene* **518,** 164–70 (2013).
215. Aschoff, M. *et al.* SplicingCompass: differential splicing detection using RNA-seq data. *Bioinformatics* **29,** 1141–8 (2013).
216. Liu, R., Loraine, A. E. & Dickerson, J. A. Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems. *BMC Bioinformatics* **15,** 364 (2014).
217. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10,** 48 (2009).
218. Eibl, R. H. *et al.* Expression of variant CD44 epitopes in human astrocytic brain tumors. *J. Neurooncol.* **26,** 165–70 (1995).
219. Humbert, P., Russell, S. & Richardson, H. Dlg, Scribble and Lgl in cell polarity, cell proliferation and cancer. *Bioessays* **25,** 542–53 (2003).
220. Royer, C. & Lu, X. Epithelial cell polarity: a major gatekeeper against cancer? *Cell Death Differ.* **18,** 1470–7 (2011).
221. Ponta, H., Sherman, L. & Herrlich, P. A. CD44: from adhesion molecules to signalling regulators. *Nat. Rev. Mol. Cell Biol.* **4,** 33–45 (2003).
222. Quéré, R. *et al.* High levels of the adhesion molecule CD44 on leukemic cells generate acute myeloid leukemia relapse after withdrawal of the initial transforming event. *Leukemia* **25,** 515–26 (2011).
223. Griffith, J. S. *et al.* Menstrual endometrial cells from women with endometriosis demonstrate increased adherence to peritoneal cells and increased expression of CD44 splice variants. *Fertil. Steril.* **93,** 1745–9 (2010).
224. Wang, S. J., Wong, G., de Heer, A.-M., Xia, W. & Bourguignon, L. Y. W. CD44 variant isoforms in head and neck squamous cell carcinoma

progression. *Laryngoscope* **119,** 1518–30 (2009).

225. Assimakopoulos, D., Kolettas, E., Patrikakos, G. & Evangelou, A. The role of CD44 in the development and prognosis of head and neck squamous cell carcinomas. *Histol. Histopathol.* **17,** 1269–81 (2002).
226. Yang, X., Khosravi-Far, R., Chang, H. Y. & Baltimore, D. Daxx, a novel Fas-binding protein that activates JNK and apoptosis. *Cell* **89,** 1067–76 (1997).
227. Daley, J. M., Laan, R. L. Vander, Suresh, A. & Wilson, T. E. DNA joint dependence of pol X family polymerase action in nonhomologous end joining. *J. Biol. Chem.* **280,** 29030–7 (2005).
228. Lee, J. W. *et al.* Implication of DNA polymerase lambda in alignment-based gap filling for nonhomologous DNA end joining in human nuclear extracts. *J. Biol. Chem.* **279,** 805–11 (2004).
229. Bork, P. *et al.* A superfamily of conserved domains in DNA damage-responsive cell cycle checkpoint proteins. *FASEB J.* **11,** 68–76 (1997).
230. Yu, X., Chini, C. C. S., He, M., Mer, G. & Chen, J. The BRCT domain is a phospho-protein binding domain. *Science* **302,** 639–42 (2003).
231. Watanabe, Y., Hirai, Y., Honda, T., Tomonaga, K. & Makino, A. X-linked RNA-binding motif protein (RBMX) is required for the maintenance of Borna disease virus nuclear viral factories. *J. Gen. Virol.* **96,** 3198–3203 (2015).
232. Hublitz, P., Albert, M. & Peters, A. H. F. M. Mechanisms of transcriptional repression by histone lysine methylation. *Int. J. Dev. Biol.* **53,** 335–54 (2009).
233. Ota, T. *et al.* Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat. Genet.* **36,** 40–5 (2004).
234. Taskén, K. *et al.* The gene encoding the catalytic subunit C alpha of cAMP-dependent protein kinase (locus PRKACA) localizes to human chromosome region 19p13.1. *Genomics* **36,** 535–8 (1996).
235. Maller, J. L. & Krebs, E. G. Progesterone-stimulated meiotic cell division in Xenopus oocytes. Induction by regulatory subunit and

inhibition by catalytic subunit of adenosine 3':5'-monophosphate-dependent protein kinase. *J. Biol. Chem.* **252,** 1712–8 (1977).

236. Cross, S. H., Meehan, R. R., Nan, X. & Bird, A. A component of the transcriptional repressor MeCP1 shares a motif with DNA methyltransferase and HRX proteins. *Nat. Genet.* **16,** 256–9 (1997).
237. Fujita, N. *et al.* Methyl-CpG binding domain 1 (MBD1) interacts with the Suv39h1-HP1 heterochromatic complex for DNA methylation-based transcriptional repression. *J. Biol. Chem.* **278,** 24132–8 (2003).
238. de Andrés-Aguayo, L., Varas, F. & Graf, T. Musashi 2 in hematopoiesis. *Curr. Opin. Hematol.* **19,** 268–72 (2012).
239. de Andrés-Aguayo, L. *et al.* Musashi 2 is a regulator of the HSC compartment identified by a retroviral insertion screen and knockout mice. *Blood* **118,** 554–64 (2011).
240. Sakakibara, S., Nakamura, Y., Satoh, H. & Okano, H. Rna-binding protein Musashi2: developmentally regulated expression in neural precursor cells and subpopulations of neurons in mammalian CNS. *J. Neurosci.* **21,** 8091–107 (2001).
241. Wuebben, E. L., Mallanna, S. K., Cox, J. L. & Rizzino, A. Musashi2 is required for the self-renewal and pluripotency of embryonic stem cells. *PLoS One* **7,** e34827 (2012).
242. Reya, T., Morrison, S. J., Clarke, M. F. & Weissman, I. L. No Title. *Nature* **414,** 105–111 (2001).
243. Morrison, S. J., Uchida, N. & Weissman, I. L. The Biology of Hematopoietic Stem Cells. *Annu. Rev. Cell Dev. Biol.* **11,** 35–71 (1995).
244. Griner, L. N. & Reuther, G. W. Aggressive myeloid leukemia formation is directed by the Musashi 2/Numb pathway. *Cancer Biol. Ther.* **10,** 979–982 (2010).
245. Zhang, H. *et al.* Musashi2 modulates K562 leukemic cell proliferation and apoptosis involving the MAPK pathway. *Exp. Cell Res.* **320,** 119–127 (2014).
246. Park, S.-M. *et al.* Musashi2 sustains the mixed-lineage leukemia–driven

stem cell regulatory program. *J. Clin. Invest.* **125,** 1286–1298 (2015).

247. Tricot, G., Wolf-Peeters, C. De, Hendrickx, B. & Verwilghen, R. L. Bone marrow histology in myelodysplastic syndromes. *Br. J. Haematol.* **57,** 423–430 (1984).
248. Raza, A. *et al.* Apoptosis in bone marrow biopsy samples involving stromal and hematopoietic cells in 50 patients with myelodysplastic syndromes. *Blood* **86,** 268–76 (1995).
249. West, R. R., Stafford, D. A., White, A. D., Bowen, D. T. & Padua, R. A. Cytogenetic abnormalities in the myelodysplastic syndromes and occupational or environmental exposure. *Blood* **95,** 2093–7 (2000).
250. Nisse, C. *et al.* Exposure to occupational and environmental factors in myelodysplastic syndromes. Preliminary results of a case-control study. *Leukemia* **9,** 693–9 (1995).
251. Levine, E. G. & Bloomfield, C. D. Leukemias and myelodysplastic syndromes secondary to drug, radiation, and environmental exposure. *Semin. Oncol.* **19,** 47–84 (1992).
252. Steensma, D. P. *et al.* Common troublesome symptoms and their impact on quality of life in patients with myelodysplastic syndromes (MDS): Results of a large internet-based survey. *Leuk. Res.* **32,** 691–698 (2008).
253. Society, L. and L. Myelodysplatic Syndromes. Available at: https://www.lls.org/disease-information/myelodysplastic-syndromes. (Accessed: 20th July 2017)
254. Institute, N. C. Myelodysplastic Syndromes Treatment (PDQ®)–Patient Version. Available at: https://web.archive.org/web/20161005015558/https://www.cancer.gov/types/myeloproliferative/patient/myelodysplastic-treatment-pdq. (Accessed: 20th July 2017)
255. Bennett, J. M. *et al.* Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group. *Br. J. Haematol.* **33,** 451–8 (1976).

256. Hoyle, C. F. *et al.* AML associated with previous cytotoxic therapy, MDS or myeloproliferative disorders: results from the MRC's 9th AML trial. *Br. J. Haematol.* **72,** 45–53 (1989).
257. List, A. F. *et al.* Expression of the multidrug resistance gene product (P-glycoprotein) in myelodysplasia is associated with a stem cell phenotype. *Br. J. Haematol.* **78,** 28–34 (1991).
258. Kearns, W. G., Sutton, J. F., Maciejewski, J. P., Young, N. S. & Liu, J. M. Genomic instability in bone marrow failure syndromes. *Am. J. Hematol.* **76,** 220–224 (2004).
259. Pereira, J. K. N. *et al.* Distinct expression profiles of MSI2 and NUMB genes in myelodysplastic syndromes and acute myeloid leukemia patients. *Leuk. Res.* **36,** 1300–3 (2012).
260. Taggart, J. *et al.* MSI2 is required for maintaining activated myelodysplastic syndrome stem cells. *Nat. Commun.* **7,** 10739 (2016).
261. Pellagatti, A. *et al.* Deregulated gene expression pathways in myelodysplastic syndrome hematopoietic stem cells. *Leukemia* **24,** 756–64 (2010).
262. Kipps, T. J. *et al.* Chronic lymphocytic leukaemia. *Nat. Rev. Dis. Prim.* **3,** 16096 (2017).
263. Harris, N. L. *et al.* World Health Organization classification of neoplastic diseases of the hematopoietic and lymphoid tissues: report of the Clinical Advisory Committee meeting-Airlie House, Virginia, November 1997. *J. Clin. Oncol.* **17,** 3835–49 (1999).
264. Hallek, M. *et al.* Guidelines for the diagnosis and treatment of chronic lymphocytic leukemia: a report from the International Workshop on Chronic Lymphocytic Leukemia updating the National Cancer Institute-Working Group 1996 guidelines. *Blood* **111,** 5446–56 (2008).
265. Smith, A., Howell, D., Patmore, R., Jack, A. & Roman, E. Incidence of haematological malignancy by sub-type: a report from the Haematological Malignancy Research Network. *Br. J. Cancer* **105,** 1684–92 (2011).

266. Dreger, P. *et al.* Managing high-risk CLL during transition to a new treatment era: stem cell transplantation or novel agents? *Blood* **124,** 3841–3849 (2014).
267. Kanti R Rai, S. S. Staging and prognosis of chronic lymphocytic leukemia. *UpToDate* (2017). Available at: https://www.uptodate.com/contents/staging-and-prognosis-of-chronic-lymphocytic-leukemia?source=search_result&search=cll prognosis&selectedTitle=1~150. (Accessed: 20th July 2017)
268. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer Statistics, 2017. *CA. Cancer J. Clin.* **67,** 7–30 (2017).
269. Luskin, M. *et al.* CLL/SLL diagnosed in an adolescent. *Pediatr. Blood Cancer* **61,** 1107–1110 (2014).
270. Eichhorst, B. F. *et al.* Fludarabine plus cyclophosphamide versus fludarabine alone in first-line therapy of younger patients with chronic lymphocytic leukemia. *Blood* **107,** 885–91 (2006).
271. Byrd, J. C. *et al.* Randomized phase 2 study of fludarabine with concurrent versus sequential treatment with rituximab in symptomatic, untreated patients with B-cell chronic lymphocytic leukemia: results from Cancer and Leukemia Group B 9712 (CALGB 9712). *Blood* **101,** 6–14 (2003).
272. Keating, M. J. *et al.* Early results of a chemoimmunotherapy regimen of fludarabine, cyclophosphamide, and rituximab as initial therapy for chronic lymphocytic leukemia. *J. Clin. Oncol.* **23,** 4079–88 (2005).
273. Brown, J. R. *et al.* Idelalisib, an inhibitor of phosphatidylinositol 3-kinase p110δ, for relapsed/refractory chronic lymphocytic leukemia. *Blood* **123,** 3390–7 (2014).
274. Administration, F. and D. FDA approves new drug for chronic lymphocytic leukemia in patients with a specific chromosomal abnormality. (2016). Available at: https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm495253.htm. (Accessed: 20th July 2017)

275. McLaughlin, P. *et al.* Rituximab chimeric anti-CD20 monoclonal antibody therapy for relapsed indolent lymphoma: half of patients respond to a four-dose treatment program. *J. Clin. Oncol.* **16,** 2825–2833 (1998).
276. O'Brien, S. *et al.* Ibrutinib as initial therapy for elderly patients with chronic lymphocytic leukaemia or small lymphocytic lymphoma: an open-label, multicentre, phase 1b/2 trial. *Lancet Oncol.* **15,** 48–58 (2014).
277. Akinleye, A., Chen, Y., Mukhi, N., Song, Y. & Liu, D. Ibrutinib and novel BTK inhibitors in clinical development. *J. Hematol. Oncol.* **6,** 59 (2013).
278. Khan, W. N. *et al.* Defective B cell development and function in Btk-deficient mice. *Immunity* **3,** 283–299 (1995).
279. Furman, R. R. *et al.* Ibrutinib Resistance in Chronic Lymphocytic Leukemia. *N. Engl. J. Med.* **370,** 2352–2354 (2014).
280. Swiech, L. *et al.* In vivo interrogation of gene function in the mammalian brain using CRISPR-Cas9. *Nat. Biotechnol.* **33,** 102–106 (2014).
281. Heidenreich, M. & Zhang, F. Applications of CRISPR–Cas systems in neuroscience. *Nat. Rev. Neurosci.* **17,** 36–44 (2015).
282. Koike-Yusa, H., Li, Y., Tan, E.-P., Velasco-Herrera, M. D. C. & Yusa, K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat. Biotechnol.* **32,** 267–273 (2013).
283. Shalem, O. *et al.* Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. *Science (80-.).* **343,** 84–87 (2014).
284. Elgar, G. & Vavouri, T. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends Genet.* **24,** 344–52 (2008).
285. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–108 (2012).
286. Cyranoski, D. Chinese scientists to pioneer first human CRISPR trial. *Nature* **535,** 476–7 (2016).
287. Cyranoski, D. CRISPR gene-editing tested in a person for the first time.

Nature **539,** 479 (2016).

288. Langmead, B. *Introduction to the Burrows-Wheeler Transform and FM Index*. (2013).
289. Wittman, T. Tree Traversal. *UCLA Math* (2015). Available at: https://web.archive.org/web/20150213195803/http://www.math.ucla.edu:80/~wittman/10b.1.10w/Lectures/Lec18.pdf. (Accessed: 20th December 2016)
290. Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and C. S. *Introduction to Algorithms*. (MIT Press).
291. Spurrier, B., Ramalingam, S. & Nishizuka, S. Reverse-phase protein lysate microarrays for cell signaling analysis. *Nat. Protoc.* **3,** 1796–808 (2008).