Optimization of molecular tools for high-throughput genetic screening

Nicolas Erard

Clare Hall College

Cancer Research UK Cambridge Institute

University of Cambridge

Submitted for the degree of

Doctor of Philosophy

January 5, 2018

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as specified in the text.

This work is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as specified in the text

The length of this dissertation does not exceed the word limit of 60,000 words as specified by the University of Cambridge Degree Committee.

Abstract

Forward genetic screening allows for the identification of any genes important for a particular biological process or phenotype. While the power of this approach is broadly agreed on, the efficacy of currently available tools limits the strength of conclusions drawn from these experiments.

This thesis describes a method to optimize molecular tools for high-throughput screening, both for shRNA and sgRNA based reagents. Using large shRNA efficacy datasets, we first designed an algorithm predicting the potency of shRNAs based on sequence determinants. Combined with a novel shRNA backbone that further improves the processing of synthetic shRNAs, we built a library of potent shRNAs to reliably and efficiently knock-down any gene in the human and mouse genomes. We then went on to apply a similar approach to identify sgRNAs with increased activity. We complemented this with conservation and repair prediction to increase the likelihood of generating functional knock-outs. With these tools in hand, we constructed, sequence-verified and validated arrayed shRNA and sgRNA libraries targeting any protein coding gene in the human genome. These resources allow large-scale screens to be performed in a multiplexed or arrayed format in a variety of biological contexts.

I have also applied these tools to identify therapeutic targets to circumvent cancer resistance to treatment in two different contexts. To overcome the shortfalls of single target therapy, I have developed multiplexed multidimensional shRNA screening strategy, where two genes are knocked down simultaneously in each cell. This strategy allows the identification of gene pairs that could be targeted in tandem to maximize therapeutic benefits. As a proof of concept, I have used it with a subset of druggable genes in melanoma cell lines. Moreover, we have applied our genome wide shRNA libraries to a different resistance context, stroma-mediated resistance to gemcitabine in PDAC. In this project, we performed screens in a PDAC-CAF coculture setting to try and identify cancer vulnerabilities specifically in the presence of stroma.

Overall, the tools developed in this thesis allow for the efficient knockdown or knockout of any gene, both in an individual or combinatorial setting. Apart from providing a resource that will be useful for many fields, we have performed several proof-of-concept studies where we have applied our tools to identify potential cancer drug targets.

Acknowledgements

I would first like to thank Greg Hannon for inviting me to join his lab and giving me the opportunity to work on such exciting and challenging projects. I would not have learned as much had it not been for the freedom and resources that he provided me with.

I also would like to specially thank Simon Knott, with whom I have worked very closely for the past four years. His mentorship and support throughout my PhD have been invaluable. He has taught me skills ranging from basic molecular biology to experimental design and computational analysis. I would also like to acknowledge Graham Mills. Through our collaboration on the pancreatic cancer project, we have shared the joys of cell-culture based RNAi screening in the CL2 rooms of the CRUK CI. Thanks also to all the members of the Hannon lab for creating a stimulating and friendly environment to learn in.

All of these experiments would not have been possible without an efficient, well stocked laboratory. For that reason, I would also like to thank Sabrina Boettcher and Nikki Coutts for always being there dealing with a large, busy lab.

A special thanks to Ann Kaminski, for taking care of all student paperwork so efficiently, especially when the lab moved from the US to Cambridge.

Finally I would like to thank my family and especially my parents who have supported me in this endeavor.

Contents

List of Abbreviations

bp	Base pairs
CAF	Cancer associated fibroblast
CRISPR	Clustered regularly interspaced short palindromic repeats
CRISPRa	CRISPR activation
CRISPRi	CRISPR interference
crRNA	CRISPR RNA
DMEM	Dulbecco's modified eagle medium
DMSO	Dimethyl sulfoxide
DSB	Double strand break
DTR	Diphteria toxin receptor
FSM	Frameshift mutation
GFP	Green fluorescent protein
LTR	Long terminal repeat
miRNA	MicroRNA
MMEJ	Microhomology mediated end joining
MOI	Multiplicity of infection
mRNA	Messenger RNA
NHEJ	Non-homologous end joining
PAM	Protospacer adjacent motif
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction

PDAC	Pancreatic ductal adenocarcinoma
PGK	Phosphoglycerate kinase promoter
RFP	Red fluorescent protein
RISC	RNA-induced silencing complex
RNAi	RNA interference
RNAseq	RNA sequencing
SFFV	Spleen focus forming virus
sgRNA	Short guide RNA
shRNA	Short hairpin RNA
siRNA	Small interfering RNA
tracRNA	Trans-activating CRISPR RNA
TSS	Transcription start site

Chapter 1

Introduction

Genetic screens are a powerful tool to investigate gene function on a genome-wide scale. Early forward genetic screens relied on DNA mutagens to introduce random mutations in an organism's genome (E.M. Jorgensen & Mango, 2002; St Johnston, 2002; Kile & D.J. Hilton, 2005; Forment et al., 2016). Following selection of individuals with aberrant phenotypes, the causal mutation can be mapped using crosses and linkage analysis. Such screens have been performed in numerous organism and have identified genes involved in for example development, signaling and embryonic patterning (Nüsslein-Volhard & Wieschaus, 1980; Vitaterna et al., 1994). However, the mapping of the mutations is time consuming and labor intensive. The discovery of RNA interference (RNAi) and its harnessing as a tool to knock down genes using sequence-specific RNA triggers has revolutionized genetic screening (Hannon, 2002). More recently, the repurposing of the bacterial CRISPR-Cas adaptive immune system to perform guided gene knockouts, as well as transcriptional repression or activation offers new reagents for functional genetics (Shalem, Sanjana, & Zhang, 2015). In this introduction, I will discuss the mechanisms of RNAi and CRISPR-Cas systems as well as the design of potent effector molecules and their use in genome-wide genetic screens.

1.1 RNA interference

1.1.1 Mechanism

RNA interference is an evolutionary conserved pathway that regulates gene expression and mediates resistance to foreign genetic material. This phenomenon was first observed in petunia plants by Napoli and Jorgensen when the introduction of a transgene to deepen flower color led to the unexpected generation of flowers with variegated or no pigmentation (Napoli, Lemieux, & R. Jorgensen, 1990). The triggers of RNAi were then discovered by Fire and Mello when they showed that long double-stranded RNAs (dsRNA) could elicit sequence-specific gene silencing in *Caenorhabditis elegans* (Fire et al., 1998). These long dsRNA triggers can be delivered by direct injection or by feeding *C. elegans* bacteria expressing dsRNA (Timmons & Fire, 1998). The potency and specificity of RNAi coupled with complete sequencing of the *C. elegans* genome rapidly allowed genome-scale loss-of-function screening to be performed (E.M. Jorgensen & Mango, 2002). Although dsRNAs were subsequently used as RNAi triggers in *Drosophila melanogaster*, this approach was initially hindered in mammalian cells as long dsRNAS trigger the interferon pathway. Further studies in plants and Drosophila cells identified small interfering RNAs (siRNA) as RNAi mediators (Hamilton & Baulcombe, 1999; P. Zamore et al., 2000; Bernstein et al., 2001). These 22 nucleotide RNAs are loaded into the RNA Interference Silencing Complex (RISC) and guide it to degrade messenger RNA (mRNA). Synthetically synthesized siRNAs were shown to trigger RNAi in mammalian cells without interferon pathway activation (Elbashir, Lendeckel, & Tuschl, 2001) and can be transfected into cells to knock-down any gene of interest (Caplen et al., 2001).

The understanding of the endogenous RNAi pathway has led to the development of other synthetic RNAi triggers mimicking microRNAs (miRNA), an endogenous class of dsRNAs (Reinhart et al., 2000; Zeng, Wagner, & Cullen, 2002). Canonical miRNAs are expressed as long poly-adenylated dsRNA transcripts (primary miRNA) from Pol-II promoters (Lagos-Quintana et al., 2001; Bartel, 2004). Stem-loop, hairpin-like structures are embedded in pri-miRNA transcripts and are flanked by single-stranded RNA segments (Han et al., 2006). pri-miRNAs go through a two-step maturation process (Denli et al., 2004; Han et al., 2006). First, primary miRNAs are processed in the nucleus by the Microprocessor complex formed by Drosha, an

RNAse III, and DGCR8/Pasha, a dsRNA binding protein (Y. Lee et al., 2003; Gregory et al., 2004). This complex recognizes the stem-loop structures present in the pri-miRNA and cleaves 11bp away from the stem-ssRNA junction. The resulting product, termed precursor miRNA (pre-miRNA) is about 70nt long, and has a 5′-phosphate group and a 3′ OH. Correctly processed pre-miRNAs are exported to the cytoplasm by exportin 5 (Yi et al., 2003; Lund et al., 2004). Pre-miRNA are then processed by Dicer, which recognizes the 5′ and 3′ ends of the pre-miRNA and cleaves 22 nucleotides from the 5′ end, removing the loop and producing an imperfect dsRNA duplex of the miRNA guide strand and its complement, the passenger strand (Bernstein et al., 2001). The guide strand is then preferentially loaded into Ago2, in the context of RISC which is guided to its target mRNA, resulting in target cleavage or translational repression (figure 1.1) (Hammond et al., 2001; Hutvagner & P.D. Zamore, 2002; Martinez et al., 2002; Khvorova, Reynolds, & Jayasena, 2003; Schwarz et al., 2003).

Artificial short-hairpin RNAs (shRNA) similar to endogenous miRNAs have successfully been used to trigger RNAi (Paddison, Caudy, et al., 2002; Brummelkamp, Bernards, & Agami, 2002; Paddison, Silva, et al., 2004; Berns et al., 2004; Paddison, M. Cleary, et al., 2004). The first generation of shRNAs mimicked pre-miRNAs, with stems and loops of varying lengths (Brummelkamp, Bernards, & Agami, 2002; Berns et al., 2004; Paddison, Silva, et al., 2004; Root et al., 2006). These shRNAs were expressed from Pol III promoters, as RNA polymerase III can be used for precise initiation and termination of RNA transcripts. shRNAs expressed in this way are directly processed by Dicer to produce siRNA triggers. The second generation of shRNAs mimic more closely the endogenous miRNA genes (Zeng, Wagner, & Cullen, 2002; Silva, M.Z. Li, et al., 2005). They rely on embedding the sequence of the target of interest in backbones of primary miRNA such as miR-30. These artificial miRNAs can be expressed from Pol II promoters and go through the same two-step maturation process as endogenous miRNAs. Second-generation shRNAs were shown to be processed more efficiently and generated a larger amount of mature small RNAs, leading to increased suppression of target mRNAs (S.R. Knott et al., 2014). Although long dsRNAs, shRNAs, and miRNAs follow different maturation steps, it is now clear that they elicit RNAi through a common biochemical pathway, with an siRNA as the key intermediary.

1.1.2 siRNA and shRNA design

In mammalian cells, siRNA and shRNA-based RNAi quickly emerged as an essential tool. Yet implementing RNAi silencing on a genome-wide scale reliably proved challenging. Whereas long dsRNA can be used for gene-silencing in *C. elegans* and *D. Melanogaster*, discrete, short sequences are used as silencing triggers in mammalian cells. In many cases, siRNAs or shRNAs designed against genes have not been effective (Khvorova, Reynolds, & Jayasena, 2003). At first, design rules were drawn from the understanding of the mechanisms of the RNAi pathway (Schwarz et al., 2003; Khvorova, Reynolds, & Jayasena, 2003; Zeng & Cullen, 2003; Ui-Tei et al., 2004). Although both strands of siRNA duplexes can be loaded into RISC, examination of endogenous miRNA showed that thermodynamical asymmetry between the ends of each strand could be used as a predictor of which strand would be loaded preferentially: the strand with the less stable 5′ end (Schwarz et al., 2003; Khvorova, Reynolds, & Jayasena, 2003). This was confirmed in *in vitro* RNAi experiments in which siRNAs were mutated and cleavage of both sense and antisense target mRNA was measured (Schwarz et al., 2003). A systematic analysis of 180 siRNAs targeting regions of the firefly luciferase and the human cyclophilin B mRNA further enriched the list of features predicting siRNA efficacy (Reynolds et al., 2004). Most of these features aimed at increasing the RISC loading bias towards the antisense strand: high stability of the 5′ end of the sense strand and low stability of the 5′ antisense strand both promote incorporation of the antisense strand. Low overall G/C content was also shown to be a predictor of siRNA efficacy, as well as some base preferences at nucleotides 1, 10, 13 and 19 (Reynolds et al., 2004).

Further studies attempted to predict effective siRNAs based on sequence features but used a relatively low number of siRNA efficacy data points, and the training sets of siRNAs were sometimes designed following the rules mentioned above which biased predictions. Subsequent approaches, however, used larger and unbiased datasets: BIOPREDsi, an artificial neural network algorithm, was trained on a dataset of 2,182 randomly selected siRNAs (Huesken et al., 2005). To assay the potency of these shRNAs, a fluorescent reporter system was used. By flanking a reporter gene with the siRNA target sequence, fluorescence levels upon siRNA transfection can be used as a proxy for siRNA efficacy. Each of these siRNAs were assayed using a high-throughput fluorescent reported gene system. Although BIOPREDsi successfully predicted activity of independent siRNAs sets, neural networks are "black box" algorithms which make interpretation of the algorithm's parameters difficult. A reanalysis of the same dataset using the LASSO method for constructing a linear model, with additional input variables such as frequency of k-mers within the guides further improved prediction accuracy and provided insight into specific sequence determinants of effective siRNAs (Vert et al., 2006). The development of reverse-transfection reagents further increased the toolkit available to rapidly identify effective siRNAs (Ziauddin & D. Sabatini, 2001; R.Z. Wu, Bailey, & D.M. Sabatini, 2002; R. Kumar, Conklin, & Mittal, 2003). Using a microarrayer, an arrayed mix of RNAi probe, siRNA target-GFP fusion and an internal RFP control can be spotted on a glass slide. Upon transfection of cells on top of the slide, each spotted siRNA can be scored by comparing the intensity of the GFP fluorescence to the RFP. Using this technique, a large number of siRNA or shRNAs can be tested in vivo in a semi-automatic fashion (R. Kumar, Conklin, & Mittal, 2003).

siRNA design algorithms can be applied to shRNAs and were first used to design shRNA genome-wide libraries (Berns et al., 2004; Paddison, Silva, et al., 2004). However, shRNAs go through multiple additional maturation steps and specific algorithms taking these criteria into account needed to be developed. The siRNA design algorithms mentioned above relied on machine-learning and large training datasets, which were not initially available for shRNAs. Novel strategies to monitor shRNA potency took advantage of high-throughput sequencing to generate thousands of data points in a single experiment (Fellmann, Zuber, et al., 2011). These sensor assays rely on cloning in the same vector an shRNA under control of an inducible vector as well as the shRNA target site in the 3′ UTR of a fluorescent reporter. Such constructs can be generated in a pooled fashion, cloned in viral vectors and transduced in cells. Upon shRNA expression induction, fluorescence is monitored by flow cytometry. The extent of the fluorescent reporter knockdown is correlated to the shRNA potency. Cells with varying degrees of fluorescence can be sorted by flow-cytometry and the shRNAs from each population can be sequenced by high-throughput sequencing. Such datasets have been used to train a highly accurate random-forest based classification algorithm (shERWOOD) (S.R. Knott et al., 2014) and support-vector machine classifier (SplashRNA) (Pelossof et al., 2017). Combining these computational shRNA selection strategies with optimized miRNA backbones has allowed for the creation of state-of-the-art genome-wide shRNA libraries (Fellmann, Hoffmann, et al., 2013; S.R. Knott et al., 2014).

While RNAi provides a convenient method to knock-down any gene of interest, off-target effects can make genome-scale RNAi experiments challenging to analyze. Although this issue was initially overlooked as early studies showed that a single base pair mismatch altered significantly siRNA potency and expression profiling using DNA microarrays showed little off-target effects. The first siRNA rules thus used BLAST to select unique genomic sequences to avoid off-target effect (Berns et al., 2004). Subsequent studies outlined that a perfect match between the siRNA guide's core seed region (bases 2-8) was necessary for efficient gene-silencing but that there was high tolerance for mismatches outside of this region (A.L. Jackson, Bartz, et al., 2003; A.L. Jackson, Burchard, et al., 2006; Birmingham et al., 2006). Complementarity between the seed region of siRNA and 3′ UTRs of multiple genes was then identified as one of the sequence-dependent source of off-target effects. Second-generation shRNA design algorithm take this into account and remove shRNAs mapping multiple times to the genome, allowing up to three mismatches outside of the seed region.

Another source of off-target effects lies in the saturation of the endogenous miRNA pathway by synthetic small RNAs. Introducing high levels of RNAi triggers has been shown to perturb gene regulation by miRNA in cells and to induce toxicity in animals (Grimm et al., 2006; Khan et al., 2009). These effects can be reduced by using miRNA based shRNAs and could be potentially eliminated by using recent shRNA backbones and more effective RNAi trigger that repress gene efficiently at lower concentrations. Off-target effects are nevertheless unlikely to be completely eliminated and false positives in large scale RNAi experiments can be avoided by increasing the number of shRNA targeting each gene of interest (Echeverri et al., 2006).

By combining rules to maximize siRNA/shRNA activity and minimize off-target effects, large collections of sequence-verified shRNA vectors covering all genes in the human and mouse genome have been generated (Silva, M.Z. Li, et al., 2005; Root et al., 2006; S.R. Knott et al., 2014). These collections allow for flexible targeting of any gene in a one-by-one setting as well as multiplexed genome-wide RNAi screens.

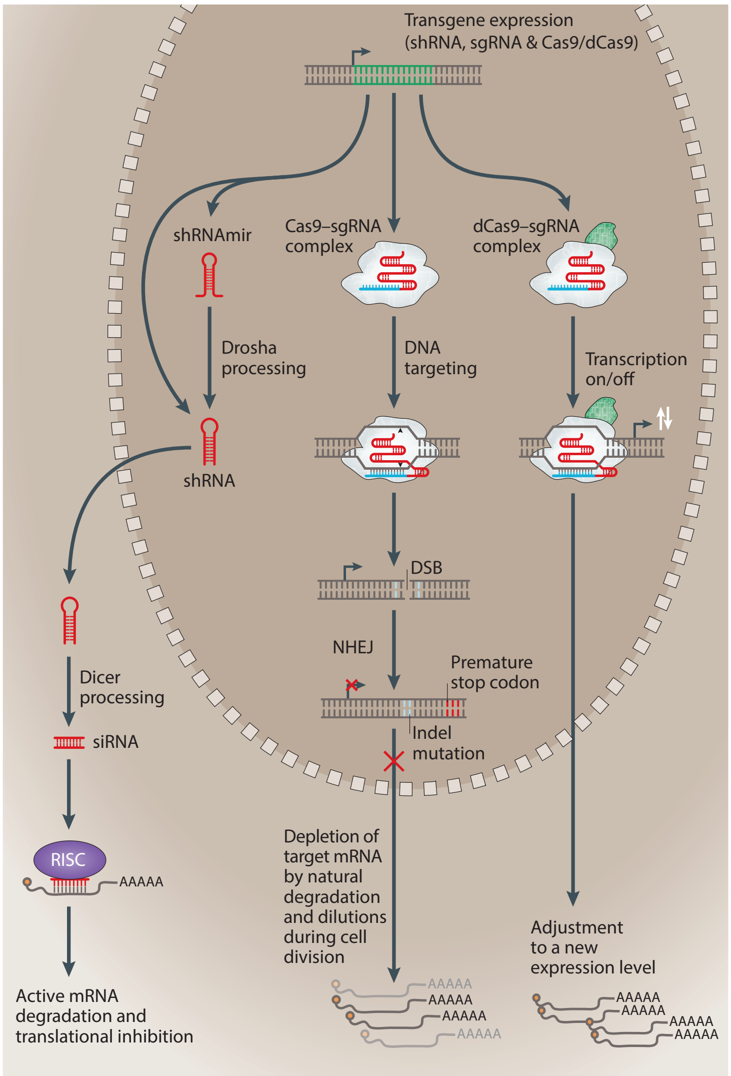


FIGURE 1.1: **Gene perturbation using RNAi or CRISPR/Cas systems** For RNAi, synthetic shRNA can be expressed from PolII promoters. Primary transcripts follow a multi-step maturation process resulting in the production of an siRNA complementary to target mRNA sequence. Target silencing is achieved by RISC loaded with the siRNA. For CRISPR/Cas systems, an sgRNA with a 20bp sequence complementary to DNA target site can be expressed from PolIII promoters, and form a complex with Cas9. The sgRNA-Cas9 complex will then generate a DSB at target site. This DSB is repaired by the error-prone NHEJ pathway which can introduce indels and frame-shift mutations. CRISPR/Cas systems can also be used to modulate transcription levels using catalytically inactive Cas9 (dCas9) fused with transcription activators or repressors. Adapted from Shalem, Sanjana, & Zhang, 2015.

1.2 Gene perturbation using the CRISPR/Cas systems

In recent years, sequence-specific RNA-guided endonucleases have emerged as a complementary tool to RNAi to perform gene perturbation at the DNA level. In particular, the repurposing of the microbial adaptive immune system CRISPR (clustered regularly interspaced short palindromic repeat) has allowed for both targeted gene mutagenesis as well as transcriptional activation or repression (figure 1.1) (Shalem, Sanjana, & Zhang, 2015).

1.2.1 Mechanism

CRISPR/Cas systems are an adaptive immune system present in most archaea and many bacterial species. A CRISPR locus is comprised of CRISPR arrays, short variable sequences called spacers, separated by short direct repeats, and is flanked by diverse cas genes (Koonin & Makarova, 2013; Barrangou & Marraffini, 2014). CRISPR/Cas immunity involves three distinct phases: adaptation, expression and immunity. Each of these phases requires distinct cas genes. In the adaptation phase, short sequences, called protospacers, from invading viruses and plasmids are incorporated in the CRISPR array (F.J. Mojica et al., 2005; Barrangou, Fremaux, et al., 2007; Bolotin et al., 2005). This step is followed by CRISPR RNA (crRNA) biogenesis (Brouns et al., 2008; Deltcheva et al., 2011), which involves the transcription of CRISPR array to yield long precursor RNAs that are further processed by Cas ribonucleases to generate small crRNAs. These crRNAs comprise sequences from invading plasmid or phage and guide Cas nucleases for specific cleavage of complementary sequences in the immunity phase (Brouns et al., 2008; Garneau et al., 2010; Jinek, Chylinski, et al., 2012; Samai et al., 2015).

The adaptation phase is thought to be mediated by a Cas1 and Cas2 protein complex and is shared by most CRISPR/Cas systems (Yosef, Goren, & Qimron, 2012; Nunez, Kranzusch, et al., 2014; Nunez, A.S. Lee, et al., 2015). Although the precise of mechanism of how spacers are acquired and integrated into the CRISPR loci remains to be determined, the selection of new protospacers relies on the presence of highly conserved motifs termed protospacer adjacent motifs (PAMs) (F. Mojica et al., 2009; Bolotin et al., 2005; Garneau et al., 2010; Shah et al., 2013). These motifs are generally conserved 2-5nt long sequences immediately flanking protospacer sequences that are to be integrated in CRISPR arrays. PAMs have also been implicated in the immunity phase as mutation of PAMs in target plasmids prevents cleavage by the Cas9 nucleases (Semenova et al., 2011). As PAM sequences are absent from CRISPR arrays, this mechanism allows the CRISPR/Cas system to distinguish self from non-self and prevents cleavage of spacer sequences in CRISPR arrays.

Great diversity between CRISPR/Cas systems has been observed at the level of crRNA biogenesis and targeting, and these systems have been divided in two classes, according to the number of proteins involved in targeting: class I systems rely on multiple proteins and the crRNA to form an effector complex (Zhao et al., 2014; R.N. Jackson et al., 2014) whereas class II systems utilize a large Cas protein in conjunction with the guide crRNA (Jinek, Chylinski, et al., 2012). Multiple and diverse class I systems have been characterized outlining the diverse architecture of CRISPR/Cas (Makarova et al., 2011; Wiedenheft, Sternberg, & Doudna, 2012).

The most well studied Class I types are type I and type III. In type I systems, the pre-crRNA is cleaved by the Cas6 endoribonuclease, one of the proteins of the CRISPR-associated complex for antiviral defense (Cascade) (R. Wang et al., 2010). The complex then recognizes and binds the target DNA homologous to the crRNA and recruits Cas3 which catalyzes target cleavage (Beloglazova et al., 2011; Huo et al., 2014). In type III systems, cleaved pre-crRNA are incorporated in an effector complex containing Cas10 that mediates target binding and cleavage. Type III Cas-systems have been shown to target both RNA and DNA (Hale et al., 2012; Staals et al., 2014; Samai et al., 2015; T.Y. Liu, Iavarone, & Doudna, 2017).

Class II CRISPR system rely on a single subunit crRNA-effector module, and are divided in types II and V. Most type II loci also encode a trans-activation CRISPR RNA (tracrRNA). The tracrRNA is partially complementary to the repeats in CRISPR arrays. Processing of the pre-crRNA requires annealing of the tracrRNA to the repeats, formation of a complex with Cas9 and cleavage by a host RNAse III protein (Deltcheva et al., 2011). Cas9, in complex with the crRNA and tracrRNA, then identifies the DNA target and mediates its cleavage (Jinek, Chylinski, et al., 2012). Type V systems, more recently identified in *Francisella novicida*, rely on CpfI for both processing and interference (Zetsche, Gootenberg, et al., 2015; Fonfara et al., 2016). CpfI cleaves pre-crRNA, to generate intermediate crRNA that are further matured and subsequently guide the single RNA-CpfI complex to its target.

While both class I and class II CRISPR/Cas systems rely on RNA-guided endonucleases to provide adaptive immunity to bacteria or archaea, the effectors of class II systems are single proteins rather than complexes making them more convenient to use as molecular tools in different organisms. Using type II CRISPR systems as-is would still require 4 components: two RNAs, the tracrRNA and crRNA guide as well as two proteins, RNAse III and Cas9. The development of chimeric guide RNAs has greatly simplified this system. *In vitro* experiments showed that the mature tracrRNA:crRNA duplex could be replaced by a single chimeric RNA, also known as single-guide RNA (sgRNA), that maintains the secondary structures needed for Cas9 cleavage as well as the 20nt needed to guide it specifically to its target (Jinek, Chylinski, et al., 2012; S. Cho et al., 2013). This technical advance has greatly contributed to the widespread use of *S. pyogenes* Cas9 coupled with sgRNAs as a genetic engineering tool.

1.2.2 Genome engineering of mammalian cells

Gene editing

In bacteria and archaea, type II CRISPR systems neutralize invading genetic elements by inducing DNA double strand breaks at the target site. Two domains within Cas9 proteins, the HNH and RuvC domain, each mediate the cleavage of one of the DNA strand, leading to a double strand break (DSB) 3 base pairs upstream of the PAM motif (Garneau et al., 2010; Jinek, Chylinski, et al., 2012). This interference provides bacteria resistance to bacteriophages or invading plasmids. In mammalian cells, DSBs can be repaired mainly by non-homologous end-joining (NHEJ) (Lieber et al., 2003) or by homology-directed repair (HDR) (Jasin & Rothstein, 2013). The NHEJ pathway is error prone and an insertion-deletion (indel) mutation can be introduced upon repair of the DSB. Targeting Cas9 to coding exons of genes of interest can lead to the introduction of frame-shift mutations and premature stop codons resulting in production of non-functional proteins and nonsense mediated-decay of the mRNA. Although on average only two-thirds of repairs would produce frame-shift mutations, large indels might produce non-functional proteins. This approach was first used in mammalian cells to generate knock-outs by co-expressing Cas9 and sgRNAs to the CLTA gene and the AAVS1 locus, with a knock-out efficiency close to 10% (Jinek, East, et al., 2013; Mali, L. Yang, et al., 2013). Other studies used the complete *S. pyogenes* CRISPR system, Cas9, tracrRNA, CRISPR array and RNAse III, to perform multiplexed genome engineering of EMX1 and PVALB (Cong et al., 2013). Although both of these strategies can be used to efficiently knock-out a gene, the chimeric gRNA/Cas9 system is more convenient as it requires only two components. This strategy was rapidly shown to work with high efficiency in mouse embryonic stem cells and embryos and could be used to knock-out multiple genes simultaneously, in a one-step injection of Cas9 mRNA and multiple sgRNAs (B. Shen et al., 2013; H. Wang et al., 2013). Although all the examples above rely on repair by the NHEJ pathway, precise genome-editing can be achieved by co-delivering a repair donor. The frequency of the integration of the donor is increased by the DSB generated by Cas9 and this allows faster generation of engineered cell lines or mice than by conventional homologous recombination. Furthermore, single stranded DNA oligos can be used as template donors to introduce point-mutations, protein tags, or LoxP sites efficiently and conveniently (H. Wang et al., 2013; H. Yang et al., 2013).

Gene activation and repression using CRISPR/Cas9

As an RNA-guided endonuclease, the CRISPR/Cas9 system has further evolved to become a powerful platform for gene regulation. Cas9 can be inactivated by introducing two mutations in its catalytic domains (D10A and H841A) (Jinek, Chylinski, et al., 2012). The inactive Cas9 (dCas9) can then be fused to any transcription factor to target it to a locus of interest and mediate gene activation or repression. The repressive effect of dCas9 was first shown in Escherichia Coli: targeting dCas9 to the promoter region or non-template strand of a transcript leads to the down-regulation of the mRNA. This is thought to be due to steric hindrance of RNA polymerase by Cas9 during initiation and elongation (Qi et al., 2013). dCas9 alone also hinders mRNA transcription in mammalian cells but to a lesser extent (2-fold reduction in human cells, 1000-fold reduction in bacteria) (Qi et al., 2013). To increase this effect in mammalian cells, dCas9 was fused to transcriptional repressors such as KRAB. The dCas9-KRAB fusion was shown to repress a control GFP construct robustly when targeted to the locus (Gilbert, Larson, et al., 2013). Although early studies required several sgRNAs targeting the same promoter to recruit sufficient dCas9-KRAB to silence the gene, novel sgRNA selection rules have reduced this requirement and robust knock-down can be now be achieved with one sgRNA

(Gilbert, Horlbeck, et al., 2014; Horlbeck et al., 2016). Other dCas9 fusions have been used for transcriptional activation and gain-of-function screens (Perez-Pinera et al., 2013; A.W. Cheng et al., 2013; Mali, Aach, et al., 2013; Konermann et al., 2015). These large-scale experiments were previously limited by the use of cDNA overexpression libraries, which do not always recapitulate isoform complexity. To activate gene expression, dCas9 can be fused to a single, or combinations of various activation domains such as the one from VP16 (Maeder et al., 2013), p65 and HSP1 (Konermann et al., 2015). Similarly to CRISPRi, early experiments in mammalian cells required at least 3 to 4 sgRNAs targeting the dCas9-activator fusion to the same promoter to achieve gene activation. Two strategies were explored to increase the number of synthetic activator recruited to the locus. The first, named SunTag, takes advantage of single-chain variable fragment antibodies (scvF) in which the variable regions of the light and heavy chain of the antibody have been fused. This antibody can be expressed from cells and fused to proteins of interest (Tanenbaum et al., 2014). The SunTag is comprised of an array of peptides recognized by such antibodies. A dCas9-SunTag fusion can thus recruit multiple copies of an scvF-activator which leads to gene activation using a single sgRNA. A second strategy is to modify the sgRNA scaffold. The addition of MS2-binding RNA hairpins to the parts of the sgRNA protruding from the Cas9 protein can direct the recruitment of MS2-activator fusions to dCas9. Using two MS2 loops to direct MS2-p65-HSF1 to the sgRNA-dCas9-VP64 duplex was shown to greatly improve gene activation of targets (Konermann et al., 2015). Both strategies where used in genome-scale studies where additional rules for sgRNA selection where identified to robustly activate any genes (Gilbert, Horlbeck, et al., 2014; Konermann et al., 2015). Customizing the sgRNA scaffold provides great flexibility in directing transcription factors to dCas9. The use of wide range of RNA recruitment structures, such as MS2, PP7 and com, can be used to encode both target site and effector in the sgRNA. By combining both CRISPRi and CRISPRa, this allows for the investigation of transcriptional networks by simultaneously activating and repressing different genes (Zalatan et al., 2015).

Epigenome editing

Recent studies have also shown that CRISPR/Cas9 can be used to modify epigenetic marks. dCas9 fused with the histone demethylase LSD1, was shown to remove activating H3K4 activating marks and repress gene levels when targeted specifically to enhancers (Kearns et al., 2015). Conversely, genes can be activated by using the catalytic core of human acetyltransferase p300 to acetylate H3K27 at promoters and enhancers (I.B. Hilton et al., 2015). Similar strategies have been devised to edit methylation marks by using dCas9-Tet1 or dCas9-DNMT3 fusions (J.I. McDonald et al., 2016; Vojta et al., 2016; X. Xu et al., 2016; X. Liu et al., 2016; Stepper et al., 2017). Although these experiments expand the range of tools used to modify gene expression using CRISPR/Cas9 systems, they have mostly been used on a rather small number of genomic loci and it is unclear whether they are efficient enough to perform genome-wide scale experiments.

1.2.3 sgRNA design

The CRISPR/Cas system allows gene perturbation in multiple ways, however they all require precise focusing of Cas9 to the intended target. Achieving maximal on-target efficiency and minimizing off-target effects requires careful selection of the effector sequence, especially for genome-wide experiments. The search space for potent guides is significantly smaller for CRISPR/Cas systems as target sites need to be flanked by specific PAM motifs. The PAM of *S. pyogenes* Cas9 was first both computationally and experimentally identified: it recognizes spacers flanked by NGG or, to a lesser extent NAG (Garneau et al., 2010; Jinek, Chylinski, et al., 2012). The PAMs from other Cas orthologs have been characterized in high-throughput assays to extend the repertoire of targetable sequences and provide orthogonal Cas systems. Cpf1's PAM, TTTN allows for example to edit AT rich regions where SpCas9 cannot be used (Esvelt et al., 2013). Initial rules for sgRNAs were primarily based on GC content and position of the target site in the exon structure, as introducing frameshift mutations in the first exon increase the likelihood of generating indels that lead to loss of protein function. Comparing the efficacy of sgRNAs targeting essential genes such as ribosomal genes refined some of these parameters and allowed for the design of support-vector-machine classifiers to predict sgRNA potency (T.

Wang et al., 2014). The next generation of sgRNA design algorithms relied on specific sensor assays similar to the ones used in RNAi to acquire large data sets of sgRNA efficacy data (Doench, Hartenian, et al., 2014; Chari et al., 2015). Two approaches were explored, one study tiled cell surface markers with libraries of sgRNA that were delivered using lentiviruses. Using flow cytometry, cells that had lost the marker could be sorted and the sgRNAs sequenced. Knockout of the surface marker was used as a measue of the sgRNA's potency (Doench, Hartenian, et al., 2014). A second study generated libraries of sgRNAs and corresponding target sites, and delivered both libraries to HEK293 cells. Direct high-throughput sequencing of the target sites and analysis of the mutations rates was used to rank sgRNAs (Chari et al., 2015). Both of these studies trained machine-learning algorithms and identified sequence determinants of sgRNA efficacy. In addition to sequence based paramaters, additional variables such as chromatin accessibility at the target site can also be considered when selecting target sites (X. Wu et al., 2014; Moreno-Mateos et al., 2015; H. Xu et al., 2015).

Similarly, CRISPRi and CRISPRa reagents have been optimized to allow the use of single rather than multiple sgRNAs to activate or repress genes. As the mechanism of action is different from the active Cas9, additional parameters were shown to be critical. They include distance from the target site to the transcription start site (TSS) and nucleosome positioning (Maeder et al., 2013; Gilbert, Larson, et al., 2013). Although in intial CRISPRi/a experiments, sgRNAs needed to be targeted to the non-template DNA strand to exhibit activity, Cas9-transcription factor fusions do not have this requirement and large-scale experiments showed that the DNA strand that was targeted did not correlate with sgRNA activity (Qi et al., 2013; Gilbert, Larson, et al., 2013). The impact of the distance to TSS and nucleosome positioning were validated by analyzing CRISPRi/a straight lethal and ricin susceptibility screens and these criteria, as well as sequence features were used to design genome-wide libraries for gene activation and repression (Gilbert, Horlbeck, et al., 2014; H. Xu et al., 2015; Horlbeck et al., 2016).

Early CRISPR/Cas9 mediated cleavage experiments both in vitro and in vivo showed that Cas9 target recognition tolerated mismatches (Jinek, Chylinski, et al., 2012; Mali, L. Yang, et al., 2013). Perfect target-sgRNA complementarity was initially thought to be required for cleavage in a "seed" region of 13bp proximal to the PAM. These observations were refined by using a large number of mutated guides, with up to 5 mismatches with target sequence (Hsu et al., 2013; Fu et al., 2013). Although mismatches in the seed sequence greatly reduce Cas9 efficiency, they are tolerated at specific positions and cleavage was observed with as many as three non-contiguous mismatches. Furthermore, small insertions or deletions in the sgRNA do not hamper target cleavage (Lin et al., 2014). To limit off-target cleavage, initial sgRNA design algorithm removed sgRNAs mapping to multiple genomic loci with less than three mismatches, especially at the 5′ end of the guide RNA (Hsu et al., 2013). Other approaches to characterize CRISPR/Cas9 off-target effects used dCas9 fused to HA tags to identify Cas9 binding sites by ChIP-Seq (Kuscu et al., 2014; X. Wu et al., 2014). These studies uncovered large numbers of off-target binding sites, recognized by dCas9 using the 5 bp proximal to the PAM. Additionally, chromatin accessibility was found to be a strong indicator of binding as potential off-target sites outside of DNAse I hypersensitive regions were significantly less bound. Estimating indels generated by active Cas9 at these off-target sites however showed that they are very rarely mutated above background level, and although Cas9 binds region with limited sequence-complementarity, perhaps more extensive pairing is required for target cleavage. Whole genome profiling of DSB repairs using high-throughput sequencing further identified off-target sites undetected *in silico* and highlighted the variability of off-target effects depending on sgRNA sequence (Pattanayak et al., 2013; X. Wang et al., 2015; S.W. Cho et al., 2014). The specificity of other Cas9 orthologues such as *Staphylococcus aureus* Cas9 or from other CRISPR-Cas types such as CpfI have also been characterized and shown to levels of specificity similar to SpCas9 (Kleinstiver et al., 2016; D. Kim et al., 2016; Ran, Cong, et al., 2015).

In addition to filtering multi-mapper sgRNAs, chemical modifications in the sgRNA and modification of the Cas9 protein have been used to decrease off-target effects. sgRNAs truncated at their 5′ end are slightly less active than their full length counterparts but reduce substantially cleavage at off target sites. The extension of loop sequences and mutations in the sgRNA scaffold were used in imaging experiments to direct a dCas9-eGFP fusion to target genomic sequences more selectively. To perform genome-editing with high specificity, a Cas9 nickase, in which only one of the Cas9 catalytic sites is active can be used (Cong et al., 2013; Jinek, Chylinski, et al., 2012). To generate a DSB, a pair of target sites in close proximity needs to be cleaved simultaneously, thus increasing specificity (Ran, Hsu, et al., 2013; Mali, Aach, et al., 2013).

1.3 High-throughput screening format and strategies

With optimized RNAi or CRISPR libraries in hand, large scale screens can be performed efficiently. For mammalian cells based experiments, two different formats are mainly used: arrayed or pooled screens. Depending on the number of target genes, arrayed screens are performed in multi-well plate formats, in which each well receives a different effector. Depending on the phenotype considered, transient transfections of RNA effectors or plasmid-based shRNAs or sgRNAs can be used, or viral vectors can be deployed when stable expression is needed. Using this format, the effect of a knock-down/knockout on a phenotype can be assessed by using plate-readers or automated microscopy, which allows screening of a large variety of different phenotypes. This strategy was for example used to screen 8000 genes for regulators of NF-kB transcription activity using as a luciferase reporter assay (Brummelkamp, Nijman, et al., 2003; Zheng et al., 2004). The use of microscopy as a read-out allows scoring of phenotypes using a greater number of parameters, temporal or spatial. This has been applied to profile genes involved in complex phenotypes such as endocytosis by transfecting cells with genome-wide arrayed siRNA libraries and tracking intake of fluorescent transferrin and epidermal growth-factor by microscopy (Collinet et al., 2010). In addition to using high-content imaging to score phenotypes, screening in an arrayed format allows greater flexibility to deliver multiple siRNAs, or an siRNA in combination with small molecule inhibitors, to the cells to study synthetic lethal effects or genetic interactions (Laufer et al., 2013). To reduce the amount of screening reagents that are necessary to conduct large scale arrayed screens, another approach is to use solid-phase reverse transfections using siRNA microarrays. This technique relies on spotting effectors in discrete locations on glass slides, and culturing cells on these slides. A different gene is targeted in each cluster of cells depending on its location on the spotted glass slide, which allows to multiplex thousands of knock-downs per slide. One drawback is that only microscopy read-outs can be used to identify both the position of the cells on the glass slide and the phenotype precisely. This strategy has for example been used extensively in the MitoCheck project to perform genome-wide siRNA screens to identify genes involved in cell division (Neumann et al., 2010).

Pooled screens in constrast rely on delivering multiple effector molecules using lentiviral vectors to large number of cells simultaneously rather than in a one-by-one format. This greatly simplifies the screening process but can only be applied to phenotypes that can be scored by enrichment or depletion of effectors in the pool of cells or for which cells of interest can be sorted to identify target genes. Early pooled screens relied on a positive selection to identify cells in which the shRNA targeted a gene relevant to the phenotype (Berns et al., 2004; Kolfschoten et al., 2005). Berns et al. for example identified components of the p53 pathway by infecting pools of human fibroblasts with shRNA libraries and selecting and extracting the shRNA sequence from colonies that bypassed p53-dependent proliferation arrest (Berns et al., 2004). Such experiments where however limited to phenotypes for which the knockdown of specific genes gave cells a competitive advantage. To perform negative-selection screen and identify lethal shRNAs targeting genes important for cell growth or survival, the abundance of each shRNA in the pool of cell needs to be tracked over time. Hits in this type of screens are genes targeted by shRNAs that drop-out over the course of the experiment.

An initial strategy to estimate the abundance of each shRNA in a given sample was to barcode each sequence uniquely. After amplification of these barcodes from the genomic DNA of the pool of cells, drop-outs can be identified by competitive hybridization of pre- and post-treatment/culture barcodes on custom DNA microarrays (Paddison, Silva, et al., 2004; Silva, M.Z. Li, et al., 2005; Ngo et al., 2006). This allows for example the identification of cancer cell and proliferation and survival genes in a systematic manner. Barcoding each shRNA vector uniquely prior to the screen requires large sequencing efforts and the generation of arrayed shRNA libraries. Subsequent approaches relied on the hybridization of the variable region of the shRNA directly, simplifying the generation of large-scale libraries (Schlabach et al., 2008; Luo et al., 2009). Micro-array quantification of pooled screens shRNAs has now been replaced by high-throughput sequencing that allows blind quantification of any number of sequences (S.R. Knott et al., 2014). The flexibility and parallel nature of pooled screens, using shRNAs or sgRNAs as effectors, have been particularly useful in cancer research to identify novel therapeutic targets and cancer vulnerabilities (Schlabach et al., 2008; E. McDonald et al., 2017).

Genome-wide drop-out screens can be relatively easily performed in many fast growing cancer cell lines and can be used to identify both straight-lethal and sensitizer hits when the screen is performed while the cells are treated with a drug to identify both single and combinatorial drug targets in different genetic backgrounds (Luo et al., 2009; Weissmueller et al., 2014; Manchado et al., 2016). The screening approaches described above have mostly been applied to cell culture, however they have also been transposed to in vivo systems that can be more physiologically relevant than 2D cultures. For in vivo experiments, shRNAs or sgRNAs can be delivered directly to adult mice tissue using viruses, or introduced by transplantation of cells pre-infected with the effector molecules. Large-scale in vivo drop out screens have allowed for the identification of new drug targets such as Ptpn2 as a cancer immunotherapy target (Manguso et al., 2017). Gain-of-function genome-wide scale have also been performed *in vivo* to identify genes involved in metastatic processes in lung cancer or liver regeneration (Wuestefeld et al., 2013; Chen et al., 2015).

Pooled screens can be used to simultaneously interrogate large sets of genes in a scalable and efficient manner when compared to arrayed screens. However, they are limited to low-content readouts such as growth rate, or variations in expression of a fluorescent marker, as opposed to the large number of features that can be read out by high-content microscopy in an arrayed setting. Novel screening strategies rely on single-cell sequencing to combine the scale of pooled screens with the complex readouts possible in arrayed screens (Dixit et al., 2016). These experiments rely on the use of expressed barcodes identifying uniquely the molecule used for the perturbation. For example, libraries of barcoded sgRNAs can be transduced into cells. After selection and culture, the transcriptome of the infected cells can be analyzed by single cell sequencing which allows recovery of both the sgRNA delivered to the cell and the transcriptional profile. Large-scale studies using this strategy have so far been performed using CRISPR and CRISPRi (Dixit et al., 2016; Adamson et al., 2016). Although analysis of these experiments are still challenging, they will surely allow faster characterization of hits as well as combinatorial pooled screens in which multiple genes are perturbed simultaneously.

To sum up, the strength of conclusions drawn from the RNAi or CRISPR experiments described above is highly dependent on the potency of the shRNAs or sgRNAs used to trigger knockown or knockout. In this thesis, I will first present work on optimizing both shRNA and sgRNA design (chapter 1 and 2). With these optimized tools in hand, I performed more complex loss-of-function studies, involving the simultaneous suppression of multiple genes in a cell, aiming at identifying combinatorial therapeutics to overcome resistance mechanisms(chapter 1), as well as interrogating a stroma-mediated gemcitabine resistance mechanism in pancreatic cancer (chapter 3).

Chapter 2

Combinatorial drug target discovery by multiplex 2D RNAi screens

This project was started in Cold Spring Harbor and transferred to Cambridge when the lab moved in September 2014. The work described in the first part of this chapter was the result of a collaboration. As part of the shRNA library optimization, Dr. Simon Knott designed shERWOOD, the machine learning algorithm to predict potent shRNA. The validation screens of shERWOOD versus other existing tools, as well as the genome wide validation of ultramir versus the previous backbone were performed by Ashley Maceli (figures 2.6, 2.7A). The rest of the results, including the experimental design of the ultramiR backbone and building of the genome wide libraries were the result of my own work. The work describing improved libraries and their validation was published in Molecular Cell (see Appendix A).

As part of the shERWOOD algorithm and 2D shRNA vector validation, Dr. Elvin Wagenblast performed the qPCRs shown in figure 2.7B and 2.8. The design and construction of the 2D shRNA vector are my own work. The initial vector construction efforts started prior to this PhD thesis and some of the spacer length optimization work (figure 2.4A) was part of my MSc Thesis, although the data has been re-analyzed for the purpose of this thesis. The subsequent large scale 2D screening efforts, including design of the library, sequencing method and performing the screens are all my own work.

2.1 Introduction

The development of targeted therapies aimed at oncogenic drivers has profoundly changed cancer treatment. These new therapies target specific pathways and molecules to which cancer cells have become addicted to sustain their growth, rather than any rapidly dividing cells as in standard chemotherapy. Targeted therapies also allow for rational drug design and are tailored to each tumor's specific genetic mutations and oncogenic drivers, maximizing therapeutic benefit and reducing treatment side effects.

Using targeted therapies has proven successful for several types of cancer where small molecules have been used to target commonly mutated oncogenes such as *Her2*, *BRAF* or *MEK* (Fisher & Larkin, 2012; Gajria & Chandarlapaty, 2011). However, although patients initially respond to treatment, most develop drug resistance followed by tumor relapse. The resistance mechanism can be explained by different factors. Following exposure to targeted therapy, cancer cells can take advantage of pathway plasticity and redundancy to rescue the loss of the targeted pathway. One example of such phenomenon is *Her2* over-expressing breast cancer. In these tumors, growth factor signaling pathways are stimulated and treatments such as Trastuzumab have been used to directly targeted *Her2* (Yu & Hung, 2000). Cancer cells can however acquire deleterious mutations in the tumor supressor *PTEN* leading to the reactivation of the PI3K pathway despite HER2 inhibition (Gajria & Chandarlapaty, 2011). To overcome such resistance mechanism, it is necessary to move towards combinatorial therapies that target multiple pathways simultaneously. This project aims at developing strategies to discover such combinatorial therapeutic target therapies using melanoma as a model.

Melanoma is the rarest but most aggressive of skin cancers (Schadendorf et al., 2015). It is caused by the unimpeded proliferation of melanocytes, the cells responsible for skin pigmentation. Progression of the disease can be classified in five stages, according to the depth at which the tumor has spread (figure 2.1). From stage 0 to stage 2, the tumor is contained locally and can be removed by surgery. At stage 3, cancer cells have spread deeper in the skin and adjacent lymph vessels and glands. Stage 3 tumors can still be treated by surgery but patients relapse in 50% of cases. At stage 4, cancer cells have spread to other parts of the body and the tumors are treated by chemotherapy or targeted therapy. The five year survival is then of only 15% (Holmes, 2014).

High-throughput sequencing studies of numerous tumors have allowed for the discovery of several genes implicated in disease progression (Davies et al., 2002; Hodis et al., 2012). Most of these genes trigger cell growth by up regulating the MAPK/ERK signaling pathway

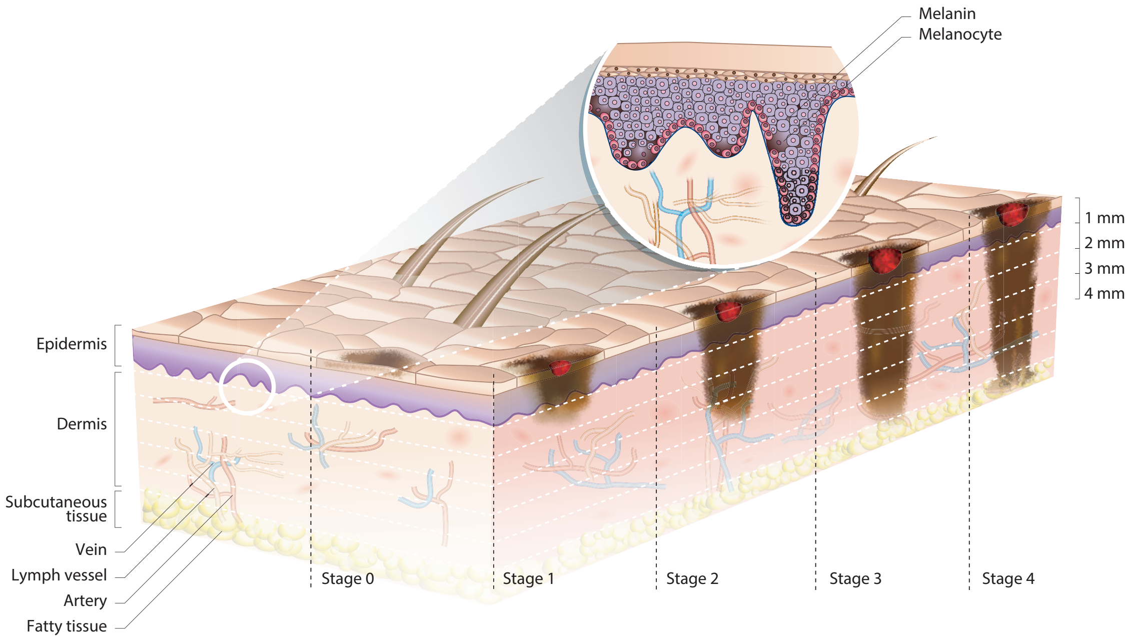


FIGURE 2.1: **Melanoma progression**, adapted from (Holmes, 2014)

(figure 2.2). *BRAF* is the most commonly mutated oncogene in melanoma as 50% of tumors harbor the BRAF^{V600E} activating point mutation (Curtin et al., 2005). Although introduction of mutated *BRAF* in melanocytes is not sufficient to induce invasive melanoma, simultaneous deletion of *p16* or *PTEN* does induce oncogenic transformation (Dankort et al., 2009). Another gene involved in the MAPK/ERK pathway, *NRAS*, is mutated in 20% of melanomas and activates both the MAPK pathway as well as the PI3K pathway ('t Veer et al., 1989) Oncogenes commonly mutated in other cancer types such as *MYC*, *CDKN2A*^{p16} and *PTEN* are also found to be mutated in a lower percentage of melanomas (Curtin et al., 2005; Goel et al., 2006).

Great efforts have been deployed to design small molecules targeting these mutated genes. Vemurafenib, a BRAF inhibitor specifically targeting BRAF^{V600E} is currently used as treatment for melanomas harboring this mutation (Fisher & Larkin, 2012). Despite this drug being effective the initial phase of treatment, most patient relapse after six to eight months. Several resistance mechanism to targeted BRAF^{V600E} inhibition have been identified. The targeted MAPK pathway can be reactivated by up regulation of COT1 (Johannessen et al., 2010), and PDGFR or IGF1R signaling (Villanueva et al., 2010). Mutations of *MEK*, downstream of *BRAF*

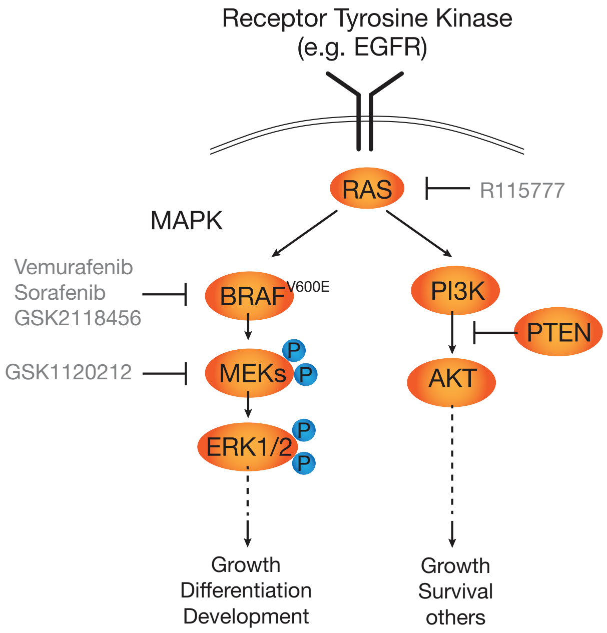


FIGURE 2.2: **The MAPK pathway** Upregulation of the MAPK pathway is common in melanoma and leads to uncontrolled growth of the tumor cells.

in the MAPK pathway can restore the pathway activation and confer resistance to BRAF inhibitors (Emery et al., 2009). To counter such resistance mechanisms, new targets that can be used in multi-therapy treatments need to be discovered.

To identify combinations of drugs deleterious to cancer cells, new assays need to be developed. Such experiments have been performed at large scale in yeast cells for which millions of double mutants have been generated (Bandyopadhyay et al., 2010; Costanzo et al., 2016). In mammalian cells, they have mainly been conducted in an arrayed format, by transfecting pairs of effectors such as siRNAs and measuring phenotype by high-content microscopy (Billmann et al., 2016). Others have relied on delivery of combinations of drug to cells to identify synergistic targets. These strategies are complex to set up and generally require automation platforms, limiting the number of combination of targets that can be tested simultaneously.

Here, we will take advantage of the high-throughput functional genomics tools developed in the lab, particularly RNA interference (RNAi) screening (Paddison, Caudy, et al., 2002; Chang, Elledge, & Hannon, 2006). The first aim of this project aims is to adapt single shRNA expression vector to perform combinatorial shRNA RNAi screens, in which different pairs of genes are targeted in each cell. Several studies report the use of vectors harboring two shRNA, expressed bicistronically and successfully targeting pairs of genes (Chicas et al., 2010). We will further optimize such vectors to ensure that both shRNAs are processed and elicit knockdown at the same level. Finally, we will design a strategy to efficiently clone and sequence highly complex pools of shRNA pairs to allow for simultaneous screening of up to 20 000 combinations of targets in the same experiment.

With these optimized vectors in hand, we will screen libraries of pairs of “druggable genes" in four melanoma cell lines with different genetic backgrounds. Hits in these screens will be further validated *in vitro* using CRISPR/Cas9 system, a genetic tool orthogonal to RNAi. Using this approach, we will focus on interfering with entire molecular networks and will provide insights on critical pathway nodes that could be targeted in combination to impede the proliferation of cancer cells.

2.2 Material and methods

2.2.1 Cloning of shRNAs in the ultramiR backbone

shRNAs sequences were predicted using the shERWOOD algorithm (S.R. Knott et al., 2014) and 97-mers (TGCTGTTGACAGTGAGCGNNNNNNNNNNNNNNNNNNNNNNTAGTGAAGCCACAGATGTANNNNNNNNNNNNNNNNNNNNNNTGCCTACTGCCTCGGA) ordered from Integrated DNA technologies. The shRNAs were amplified individually by PCR using the Haiprin-HpaI-F primer (5′-CTGGGATTACTTCTTCAGGTTAACCCAACAGAAGGCTAAAGAAGGTATATTGCTGTTGACAGTGAGCG) and the Hairpin-HpaI-R primer (5′-AGAGATAGCAAGGTATTCAGTTTTAGTAAACAAGATAATTGCTCCTAAAGTAGCCCCTTGAAGTCCGAGGCAGTAGGC). The 97-mers were amplified by PCR using KOD (Takara) and purified using a QIAquick PCR purification kit. The amplified shRNAs were cloned by Gibson assembly (NEB) in a vector harboring an ultramiR cassette, digested with HpaI (NEB). 1 μL of the Gibson assembly reaction was transformed by electroporation into Endura electrocompetent cells (Lucigen). Cells were plated on LB-Amp agar plates and grown overnight at 30°C. Colonies were picked, grown, mini-prepped and the full ultramiR backbone was Sanger sequenced by GATC-Biotech.

2.2.2 Barcoding shRNAs in the 2D-sequencing vector

The 2D-sequencing vector is based on the MSCV-derived retroviral LMN vector backbone, and harbors an ultramiR cassette flanked by two 25bp barcodes. To build this vector, a gBlock (IDT) comprised of two AT rich regions from the yeast genome separated by an EcoRI and XhoI restriction site were cloned LMN, previously digested using BglII and MluI, by Gibson assembly. Then, an ultramiR cassette was amplified by PCR using two ultramer oligos (IDT) harboring a 25bp random barcode flanked by the Illumina Truseq or SBS3 sequence (forward primer: TATTCTATTCCTTGCCTTACTTTTCTTATGAATTCNNNNNNNNNNNNNNNNNNNNNNNNNAGATCGGAAGAGCACACGTCTGAACTCCAGTCACCATGTCTGTTCTCTACATTGCTTTTTATTGGATCCTATTGGTTTTCCTAACCAACGCGCTGCACAAGATCTTGTTTGAATGAGGCTTCAGTACT, reverse primer: TATATGTACTTATACGGATGTTATTACTCGAGNNN

NNNNNNNNNNNNNNNNNNNNNNAGATCGGAAGAGCGTCGTGTAGGGAAAGAGT GTAGATCTCGGTGGTCGCCGTATCATTTTTGCGGCCGCTAGTCTGTCTAAATGTGCAAT GGGAGCAGAACGCGTAAAGTGATTTAATTTATACCATTTTAATTCAGCT. This barcoded ultramiR cassette was cloned by Gibson assembly in the vector created as described above using the EcoRI and XhoI site. The reaction was transformed in highly competent MegaX DH10B T1R bacterial cells (Life Technologies). More than 5 million independent transformation events were obtained.

The 240 shRNAs targeting druggable genes or olfactory receptors were cloned as a pool in the 2D-seq shRNAs as described in section 2.2.1. For each clone, the full sequence of the shRNA cassette and the flanking barcodes were Sanger sequenced by the Beckman Coulters Genomics facility. Clones for which there was no mutations in the ultramiR cassettes and the GC content of each barcode was lower than 80% were kept to build dual shRNA libraries.

2.2.3 Cloning of dual shRNA libraries

The MSCV-derived retroviral LMN vector backbone was adapted for the expression of pairs of hairpins. In this vector both shRNA cassettes are driven by the 5′ long terminal repeat. GFP and neomycin resistance gene are driven by the PGK promoter. Hairpins in the 2D Sequencing vector were amplified individually from sequence verified glycerol stocks. The 5′ shRNAs were amplified using primers HP1-F (5′-GGATCCTATTGGTTTTCCTAACC) and HP1-R (5′-TCGCGATGAGAAAAAGCCG), and the 3′ shRNAs using primers HP2-F (5′-ATGCATCTTG GCGGCTTT) and HP2-R (5′-GCGGCCGCTAGTCTGTCTAA). A limiting primer concentration of 2μM and 35 cycles of amplification were used to equalize the amount of amplicons across all PCR reactions. PCR products of the first or second shRNA were pooled and loaded on a 1% agarose gel. For each pool, the band ~600bp was excised and the DNA was purified using a QIAGEN Gel extraction kit. The two shRNA pools were cloned in a retroviral expression vector, previously digested with BamHI and NotI (NEB), by a three-way Gibson assembly. The reaction was transformed in highly competent MegaX DH10B T1R bacterial cells from Life Technologies. Enough electroporations were done to have at least 1000 times the number of different shRNA pairs independent transformation events.

2.2.4 Cloning of dual sgRNA libraries

The pCRoatan-dualSgRNA vector (Erard, S.R.V.R. Knott, & Hannon, 2017) was adapted to clone barcoded pairs of sgRNAs from DNA chips. DNA chips (CTGTTGACAGTGAGCGG AAGACgctctaaaacNNNNNNNNNNNNNNNNNNNNCGGTGTGAGACGAGCGTACGCG TNNNNNNNNNNNNNNNNNNNNCTCGAGACAGGGGTCTCAGTCGGNNNNNNNN NNNNNNNNNNNNgttttaGTCTTCTGCCTACTGCCTCGGA, Ns are placeholders for the first sgRNA, the 20nt barcode and the second sgRNA) were ordered from Custom Array, Inc and amplified by PCR. The amplicons were cloned in an intermediary cloning vector (pCR-Blunt II-TOPO, Thermi Fischer) by ligation using the SpeI and ApaI sites. Next, the hU6 promoter and a Zeocin resistance cassette were amplified from pCRoatan-dualPromoter (Erard, S.R.V.R. Knott, & Hannon, 2017) (forward-primer: AGTACCGTCTCTGGTGTTTCGTCCTTTC-CACAAG, reverse-primer: ATGAACGCGTAGTGCGGATCCTGCAGCACGTGTT) and ligated into the intermediary cloning vector harboring the chips with the BsmbI and Mlui restriction sites. The cU6 promoter was similarly amplified from pCRoatan-dualPromoter (forward-primer: ATCGATCTCGAGGCGCCGCCGCTCCTTCAGGCA, reverse-primer: TGATCCTGG TCTCACGACTAAGAGCATCGAGACTGC) and cloned in the plasmid by ligation using the BsaI and XhoI restriction sites. The full sgRNA1-hU6-barcode-cU6-sgRNA2 cassette was excised from the intermediary cloning vector using the BbsI restriction sites and ligated in pCRoatan-dualSgRNA by ligation. For each cloning step, at number of colony greater than 100 times the number of different sequences in the chip were obtained to keep the full complexity of the library.

2.2.5 Cell culture

Melanoma cell lines A-375, SK-Mel-5, SK-Mel-28, and WM-266-4 were purchased from ATCC and grown at 37°C in DMEM supplemented with 10% FBS and 50U/mL of penicillin-streptomycin or MEM supplemented with 10% FBS, sodium pyruvate, and 50U/mL of penicillin-streptomycin, following supplier instructions. Platinum-A retroviral packaging cell lines were purchased from Cell Biolabs and grown at 37°C, in DMEM supplemented with 10% FBS, 50U/mL of penicillin-streptomycin, 1 μg/mL puromycin and 10 μg/mL blasticidin. Human epidermal melanocytes (HEMa-LP) were purchased from Life Technologies and grown at 37°C in Medium 254 supplemented with Human Melanocyte Growth Supplement-2. 4T1 cells were purchased from ATCC and grown at 37°C in RPMI-1640 supplemented with 10% FBS and 50U/mL of penicillin-streptomycin. The 4T1-T cell line is a clonal line derived from the 4T1 line (Wagenblast et al., 2015).

2.2.6 Virus production and cell infection

Platinum-A cells were transfected with plasmids containing the VSVG receptor and the retroviral plasmid of interest, using the calcium-phosphate transfection method (Wigler et al., 1978). The supernatant containing the retrovirus was collected two days after transfection, filtered and stored at 4°C until use. Cells were infected with an average representation of 1000 independent integrations per shRNA, and with a multiplicity of infection of one. Two days after infection, cells were selected with 400 μg/mL neomycin.

2.2.7 Small RNA cloning

Small RNA cloning of mature miRNA expressed from the dual shRNA vector

Small RNAs were cloned using the method described in (Malone et al., 2012). Briefly, 4μg of total RNA was spiked with ^{32}P-labeled 19- and 30-mer, loaded on a 12% polyacrylamide gel and run at 12W for 1.5h. After exposure to a photo-screen, the gel bands corresponding to RNAs ranging from 19 to 30bp were excised and incubated overnight with agitation in 400μL of 0.4M NaCl. The RNAs were then extracted from the supernatant by ethanol precipitation. The 3′ sequencing adapter was ligated to the sample by incubating the size selected RNAs for two hours at room temperature with 1μL of 50μM adapters (/5rApp/TGGAATTCTCGGGTG CCAAGG/3ddC/), 2 μL of T4 RNA ligase Truncated (NEB), 2μL of ATP-free T4 RNA Ligase buffer, and 2μL of DMSO, in a total volume of 20μL. After ligation, the samples were loaded on a polyacrilamide gels and the 41 to 52bp RNAs were extracted as previously. Next, the 5′ adapter was ligated by incubating the samples at 37°C for two hours with 2μL of T4 RNA ligase (NEB), 2μL of T4 RNA ligase buffer, 2μL of DMSO and 1μL 50μM 5′ adapters (GUUCAGAGUUCUACAGUCCGACGAUCU), in a total volume of 20μL. Following ligation, the samples were run on a polyacrilamide gel and the ligated RNAs of size 68 to 79bp were extracted as previously. For reverse-transcription, the RNAs were incubated at 70°C for 30s with 2μL of 10mM dNTPs and 1.5μL of RT primer (GCCTTGGCACCCGAGAATTCCA, complementary to the 3′ adapter). The samples were placed on ice for one minute, and 1μL of SuperScript III reverse-transcriptase (Invitrogen), 4μL of 5X first strand buffer, 1μL of 0.1M DTT and 1μL of RNAse inhibitor was added. The reaction was placed at 50°C for an hour, and the enzyme was heat-inactivated at 70°C for 5min. The cDNA was then amplified by PCR using KOD (Takara) according to the manufacturer's instructions (forwad primer : AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA, reverse primer: CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA). Each sample was barcoded using the 6bp flanking the TruSeq Illumina sequence in the reverse primer. The amplified DNA was loaded on a 1.5% agarose gel and the sequences ranging from 135 to 150bp were excised and purified using a QIAGEN gel extraction kit. The small RNA libraries were quantified by qPCR and sequenced on Illumina GAII or MiSeq platforms.

Small RNA cloning of mature miRNA derived from a miR30 or UltramiR scaffold

Small RNAs ($<$ 200bp) were extracted from 10^7 using the mirVana miRNA isolation kit (ThermoFischer). Libraries were generated using 100ng of small RNAs and the TruSeq small RNA library preparation kit (Illumina) and sequenced on the Illumina MiSeq.

10μg of total RNA extracted using Trizol reagent were used as input. The sequence of the linkers and PCR primers was adapted to allow for multiplexing of small RNA libraries and sequencing using the Illumina technology.

2.2.8 Total RNA sequencing

For each cell line, total RNA was extracted using Trizol reagent. 500ng of RNA was used as input for the Nugen Ovation RNA-Seq System v2. The cDNA was then sheared to an average size of 200bp using a Covaris and Illumina sequencing adaptors were ligated. Libraries were pooled and sequenced on an Illumina Hi-Seq. The sequencing yielded at least 20 million 75bp paired-end reads for each library.

2.2.9 Differential expression analysis

Reads were aligned to the hg19 genome (Lander et al., 2001) using Tophat 2.0.12 (Trapnell, Pachter, & Salzberg, 2009), with default parameters. For all libraries, more than 80% of reads mapped. Aligned reads were assigned to exons using HTSeq 0.6.0 (Anders, Pyl, & Huber, 2015) and differentially expressed genes were called using DESeq 1.14.0 (Anders & Huber, n.d.), with a FDR < 0.05.

2.2.10 Screen sequencing

RNAi screens

The barcodes identifying the shRNAs were amplified from screen genomic DNA using primers P5-SBS3 (5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCT TCCGATCT and P7-TruSeq-BCX (5′-CAAGCAGAAGACGGCATACGAGATNNNNNNTGT GACTGGAGTTCAGACGTGTGCTCTTCCG, where the six Ns are the barcode used for multiplexing). For each time point, at least 200 μg of genomic DNA as input. The amplicon was gel extracted, purified and quantified by qPCR. Libraries were pooled, and sequenced on an Illumina High-Seq. At least 20 million reads per time point mapped to the expected barcodes.

CRISPR/Cas9 screens

The barcodes identifying the sgRNA pairs were amplified from screen genomic DNA using primers P5-EM7 (5′-AATGATACGGCGACCACCGAGATCTACACCGATGATTAATTGTCA ACACGTGCTGCAGACGCGT) and P7-BCX-cU6 (5′-CAAGCAGAAGACGGCATACGAGAT NNNNNNCTGAAGGAGCGGCGGCGCCTCGAG), and processed similarly to RNAi screen samples. Libraries were sequenced on the Illumina MiSeq using a custom read1 primer (EM7-Seq-primer-Read1: CGATGATTAATTGTCAACACGTGCTGCAGACGCGT) and a custom index read primer (cU6-index-primer: CTCGAGGCGCCGCCGCTCCTTCAG).

2.2.11 Screen analysis

For both RNAi and CRISPR screens, reads were mapped to an index of all barcodes expected to be in the pool using bowtie (Langmead et al., 2009) allowing for one mismatch. Libraries were normalized by library size and for each construct depletion rates were calculated by dividing counts at the final timepoint by counts at the initial timepoint. These log-ratios were further normalized so that the control population had a mean of 0 and a variance of one. Constructs were classified as depleted with an FDR cutoff of 0.1 using an empirical Bayes moderated t-test (Smyth, 2004).

2.3 Results

2.3.1 Dual shRNA expressing vector

Vector design

In order to knockdown two independent targets in the same cell, previous studies have relied on co-transfection of multiple targeting molecules such as siRNAs or drugs. This can be done in an arrayed format to achieve large scale combinatorial screening. To perform such experiments in a pooled format, my first aim was to develop a viral vector that I could easily and stably deliver to cells, and from which two shRNAs could be expressed. This vector had to fulfill two requirements. First, the two shRNAs needed to be expressed and processed at similar levels and mediate significant knockdown for both genes simultaneously. Second, since in pooled screens the construct is stably integrated in the genomic DNA of the cells, I needed to be able to estimate the abundance of each shRNA pair in a pool of cells by sequencing. Others have targeted pairs of gene and obtained significant knockdown levels by expressing two shRNAs in a bicistronic fashion (Chicas et al., 2010; Wuestefeld et al., 2013). Following these results, we modified the MSCV-derived vector backbone used for single shRNA screening, named LMN, to express pairs of shRNAs from the LTR promoter. This vector additionally harbors a fluorescent protein and a drug resistance gene expressed from the PGK promoter to allow for tracking and selection of infected cells (figure 2.3).

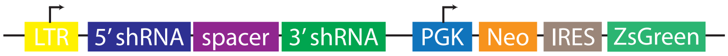


FIGURE 2.3: **A dual shRNA expression vector** Schematic of the vector used to expressed pairs of shRNA. A DNA spacer is introduced between the shRNA cassette to allow efficient processing of both shRNAs. LTR: retroviral long terminal repeat, PGK: mouse phosphoglycerate kinase 1 promoter, Neo: Neomycin resistance gene, IRES: Internal ribosobal entry site, ZsGreen: ZsGreen1 fluorescent protein. Black arrows indicate transcription initiation sites.

Empirical evidence in our lab and others has shown that separating the shRNA from its promoter by a few hundred base pairs can increase processing and overall knockdown efficiency. We hypothesized that to get similar levels of mature guides for both shRNAs, they needed to be separated by a DNA spacer. I thus cloned vectors with two p16 targeting shRNAs (p16 shRNA1 and p16 shRNA2) separated with spacers ranging from 0 to 225bp. To rule out any positional effect, two vectors were built for each spacer length: shRNA1-spacer-shRNA2 as well as shRNA2-spacer-shRNA1.

To compare the efficiency of the shRNAs depending on spacer length, we quantified the abundance of mature guides targeting the p16 genes by small RNA cloning. We transduced all constructs, as well as vectors expressing each single shRNA as a controls, in the A38-5 pancreatic adenocarcinoma cell line, and following selection, prepared small RNA libraries. We then compared the abundance of p16 shRNA1 or p16 shRNA2 guide reads in the samples expressing pairs or single shRNAs (figure 2.4A). To account for library size, the reads from each sample were normalized by the median number of reads of the 20 most expressed endogenous miRNAs. The ratio of normalized counts of each p16 shRNA in dual to single hairpin constructs was calculated to compare processing in both contexts. This ratio represents the level at which the shRNA is processed relative to its maximum potential in single constructs. The results of this first experiment shows that as the spacer length increases, so does the abundance of processed shRNA guides in the dual constructs. For the 225bp construct, both shRNAs are on averaged processed at 20% of the level of the control single shRNA vector.

One caveat of this experiment is that pancreatic adenocarcinoma is frequently associated with p16 mutation and knocking down this gene in the A38-5 cell line can potentially bias the results (Caldas et al., 1994). To further study the expression of dual shRNAs, we tested additional spacers, with a length ranging from 200 to 800bp, with two KRAS targeting hairpins. The spacer sequences used were derived from either the Kanamycin resistance gene or A/T rich regions of the yeast genome. These experiments were performed in two non-KRAS dependent cell lines: human embryonic kidney 293T and ERC. ERCs are derived from the chicken fibroblast DF-1 cell line and has been used extensively in sensor assays to identify potent shRNAs (Fellmann, Zuber, et al., 2011).

The use of these longer spacers further increases processing of the pair of shRNAs (figure 2.4B and 2.4C), although the abundance of mature shRNA reads decreases for spacers longer than 600bp. For all constructs, the spacer does not seem to interfere with the relative processing of the first or second shRNA as both are processed at similar levels regardless of their position in the vector. In paired constructs, the KRAS shRNA-1 is consistently more processed than the KRAS shRNA-2 when normalized to the single shRNA control. To explain this difference, I examined the raw ratio (i.e. not normalized by the single hairpin construct ratio) of read counts for the KRAS shRNAs, averaged for both cell lines (figure 2.4D). The KRAS shRNA-2 was significantly more processed in single constructs than the KRAS shRNA-1, and such strong levels of expression are not reached in the dual shRNA construct.

Overall, when all miRNAs were ranked by number of reads, the KRAS or p16 mature guides expressed from dual shRNAs constructs were in the top 40 most expressed miRNAs (supplementary figure 2.1). In light of these results, we chose a 400bp spacer. This was the longest length that we could use to sequence pairs of shRNAs by paired-end sequencing using Illumina technology at that time. For this spacer length, both shRNAs are processed on average at 25% the level of the single shRNA constructs which we deemed sufficient to induce potent knockdown of both genes.

UltramiR, a variant miRNA scaffold increases shRNA potentcy

Performing combinatorial screens with RNAi requires both expressed shRNAs to elicit efficient knockdown for the genes of interest. Figure 2.4D highlights the great variability in processing of different hairpins, which can be a problem for combinatorial screens in which both expressed

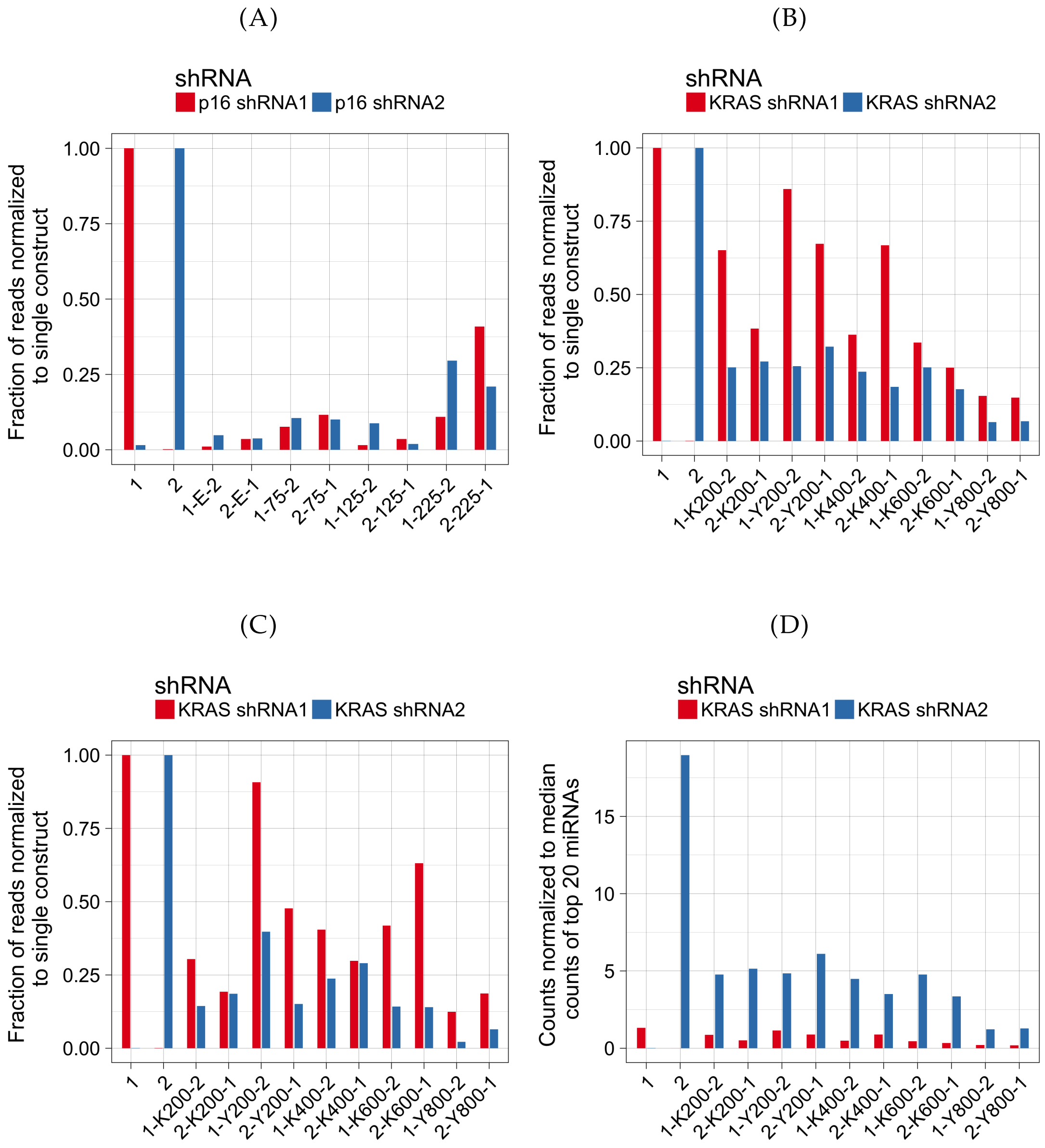


FIGURE 2.4: **Optimization of spacer length to increase processing of shRNA pairs** **(A)** Constructs harboring a pair of p16 targeting shRNAs were transduced in A38-5 cells and small RNA cloning was performed. Shown is the ratio of mature p16 shRNA read counts over the median number of reads of the top 20 microRNAs, normalized by the single shRNA construct ratio. The x-axis indicates which shRNA was placed in the 5′ cassette and how long the spacer was in bp, and which shRNA was placed in the 3′ cassette. **(B), (C)** The same experiment was performed using KRAS targeting hairpins in non KRAS dependent cells either human 293T (B), or chicken ERC (C). The length of the spacer is preceded by a K if the spacer sequence was part of the Kanamycin resistance gene or by a Y if it came from the yeast genome **(D)** Shown are the ratio of mature KRAS shRNA reads over the median number of reads of the top 20 microRNAs, averaged for the both cell lines.

hairpins need to have similar activity. At the same time as I was developing the dual shRNA vectors, I was working with others in the lab on new strategies to select and express potent shRNAs. This involved both designing new backbones to improve shRNA processing and algorithms taking advantage of large shRNA efficacy datasets and machine learning techniques to predict shRNA potency *in silico* based on sequence characteristics.

All of the shRNA experiments described above rely on the miR-30 scaffold that had been modified to allow rapid cloning of any variable region. The endogenous sequence had been modified to include an EcoRI and XhoI restriction site flanking the stem of the shRNA to facilitate cloning. Later studies of Drosha processing sequence determinants by *in vitro* cleavage analysis of highly complex variable synthetic pri-miRNA libraries had identified a conserved CNNC motif important for processing 17bp downstream of the hairpin stem (Auyeung et al., 2013). In the miR30-based shRNA scaffold, the first C of this motif had been replaced by an A to create the EcoRI restriction site, which might have reduce the processing of the pri-miRNA. Others had reported that moving the EcoRI site in the miR30 scaffold to recreate this conserved motif increased small RNA levels and lead to increased knockdown (Fellmann, Hoffmann, et al., 2013). We reasoned that eliminating the restriction sites altogether would be best. Therefore, we created a scaffold called ultramiR, based on the endogenous miR-30 backbone, in which shRNAs can be cloned in the cassette using scarless Gibson assembly.

To test processing of ultramiR shRNAs, we cloned hairpins targeting RPA3 or the Renilla luciferase in the standard miR30 scaffold and in ultramiR. These constructs were transduced at low MOI (<0.3 to guarantee only one infection event per cell), in duplicates, in the human 293T cell line and in the chicken fibroblast derived ERC cell line that was used previously in sensor assays (Fellmann, Zuber, et al., 2011). Following selection for infected cells, we analyzed the levels of mature shRNAs by small RNA sequencing. To normalize across libraries and compare the two backbones, we calculated the log-fold enrichment of Renilla or RPA3 counts relative to the 66^{th} quantile of endogenous miRNA counts (figure 2.5). When shRNAs were placed in the ultramiR backbone, the levels of mature small RNAs were significantly increased compared to the standard miR-30 backbone.

Concurrently, Dr Knott, a member of the Hannon lab, was working on designing algorithms to predict shRNA potency. Using large datasets of shRNA efficacy datapoints generated

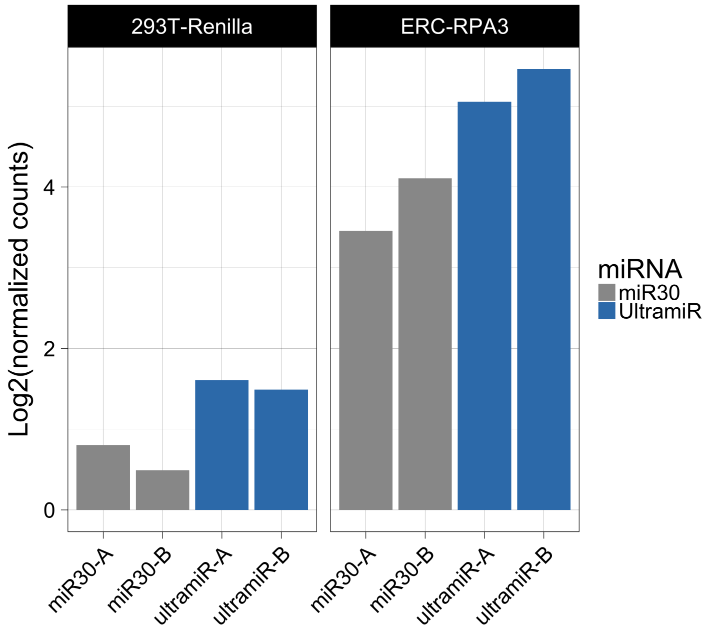


FIGURE 2.5: **The UltramiR scaffold increases shRNA processing** shRNAs targeting RPA3 or the Renilla luciferase were cloned in the standard mir30 scaffold and in ultramiR. After transduction at low MOI in 293T and DF-1 cells, small RNAs were sequenced. shRNA guide counts were normalized by calculating the log-fold enrichment relative to the 66^{th} quantile of endogenous miRNA

by sensor assay, he built a machine-learning based algorithm termed shERWOOD that classifies shRNAs by potency using sequence determinants (S.R. Knott et al., 2014). This selection algorithm was extensively validated by performing test loss-of-function screens with libraries harboring ∼22 000 shRNAs targeting essential genes and 4 000 targeting olfactory receptors. These shRNAs were selected either from the TRC library, using the Designer of Small Interfering RNAs (DSIR) or using shERWOOD. The three libraries were transduced at low MOI in the pancreatic adenocarcinoma A385 cell line. Two days after infection, half of the cells were collected for a reference timepoint, and the rest were grown for ∼12 doublings and collected for the final timepoint. The abundance of each shRNA in both timepoints was estimated by high-throughput sequencing and depletion log ratios were calculated. The shRNAs selected by shERWOOD and targeting consensus-essential genes showed increased log-fold change in the depletion screens when compared to the shRNAs selected by others algorithms (figure 2.6A). Additionally, the percentage of shRNAs depleted for each essential gene was increased (figure

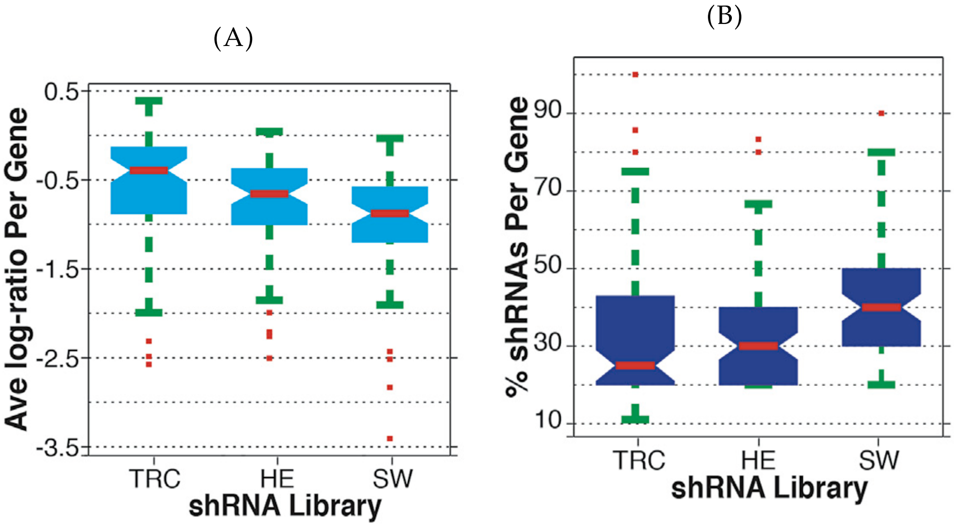


FIGURE 2.6: **Validation of shERWOOD, an shRNA potency prediction algorithm** Libraries from the TRC or Hannon-Elledge (HE) collection, or designed using shERWOOD (SH), targeting essential and olfactory genes were screened in the A385 cell line. Shown in **(A)** is the distribution of log-fold changes of shRNAs targeting consensus-essential genes. Shown in **(B)** is the percentage of shRNAs targeting consensus-essential genes that were depleted.

2.6B). Overall, shERWOOD was shown to be highly efficacious at selecting potent shRNAs for any gene of interest.

Up to this point, all validation experiments had been performed in a mir30 shRNA based backbone. We reasoned that using ultramiR would further increase shRNA potency. To test this, we performed a side-by-side depletion screens with shRNAs targeting essential genes or olfactory receptors as described above. The same set of shRNAs predicted by shERWOOD were cloned in mir30 and ultramiR backbones, and screened in A385 cells as described above. The number of shRNAs per gene depleting as well as the degree at which they depleted was compared for both backbones (figure 2.7A). UltramiR shRNAs were found to be significantly more depleted than miR30 shRNAs (from 0.95 to 1.05, rank-sum test, $p < 0.01$) and for each essential gene, a larger percentage of ultramiR shRNAs was depleted (from 42% to 51%, rank-sum test, $p < 0.01$).

We further assessed the potency of shERWOOD-UltramiR shRNAs by measuring the reduction of mRNA levels they induced. We cloned four shRNAs with the highest shERWOOD

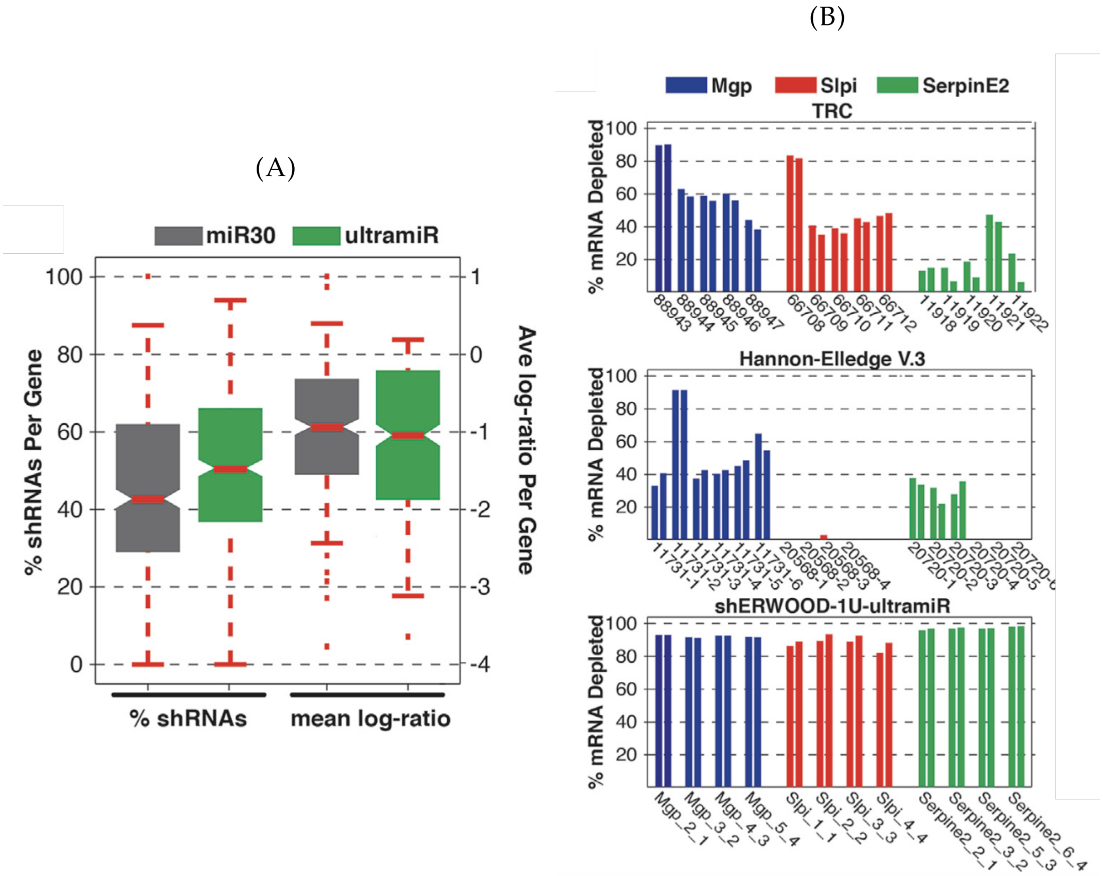


FIGURE 2.7: **Validation of the UltramiR scaffold** **(A)** A library harboring shRNAs targeting essential genes and olfactory receptors was cloned in a mir30 and ultramiR backbone and screened in A385 cells. Shown is, for both libraries, the percentage of shRNA per essential gene that was found to be depleted in the screen, as well as the degree at which these shRNAs were depleted. **(B)** Knockdown effiencies for shRNAs targeting mouse Mgp, Slpi and SerpinE2. shRNAs sequences were obtained from the TRC, the V.3 Hannon-Elledge library or selected using shERWOOD.
]

prediction scores targeting mouse Mgp, SerpinE2 and Slpi into an MSCV-based ultramiR vector. To compare these constructs to existing RNAi libraries, we obtained the current TRC (five shRNA per gene) and V.3 Hannon-Elledge (six shRNA per gene for Mgp and SerpinE2, four for Slpi) library constructs targeting these genes. These constructs, including empty vectors as a negative control, were packaged in viruses and mouse 4T1 cells were infected at single copy. Following selection for infected cells, mRNA levels of each gene was assessed by quantitative RT-PCR (RT-qPCR) and compared to the levels in the corresponding non-targeting control. Both TRC and V.3 Hannon-Elledge shRNAs showed relatively modest knockdown levels with all shRNAs mediating less than 50% target knockdown except for shRNAs 88493 and 66708 (TRC) and 11731-2 (V.3 Hannon-Elledge). In comparison, all shRNAs selected using shERWOOD in the ultramiR scaffold reduced target mRNA levels by over 80% (figure 2.7B). Overall, our data shows that selecting guide sequences using shERWOOD and placing them in the ultramiR scaffold is a robust strategy to consistently generate potent shRNAs with high level of target knockdown.

As a final test of our dual shRNA vector, we combined the shERWOOD-Ultramir shRNA strategy with the dual shRNA expression vector described above. The mir30 cassettes were switched to ultramiR cassettes and they were separated by the spacer of ∼400bp for which mature small RNA levels were optimal. Rather than performing small RNA cloning, we wanted to assess the level of knockdown induced by both shRNAs. We thus cloned two pairs of mouse SerpinE2 and Slpi hairpins in the dual shRNA vector, transduced the 4T1 cell line and a 4T1 derived cell line (4T1-T) and measured mRNA levels by qPCR. An empty vector was used as a control, and log-fold changes of mRNA levels for the dual constructs samples compared to the control were computed (figure 2.8). For both cell lines, a knockdown level of at least 65% was observed for both shRNAs. This data, together with the small RNA cloning experiments (figure 2.4), shows that a vector expressing pairs of shRNAs separated by a spacer can be used to robustly knockdown pairs of genes simultaneously.

2.3.2 Screen sequencing

RNAi loss-of-function screens are generally analyzed by comparing the abundance of shRNAs in a population of cells at a reference start point and after growing the cells. shRNAs that have a

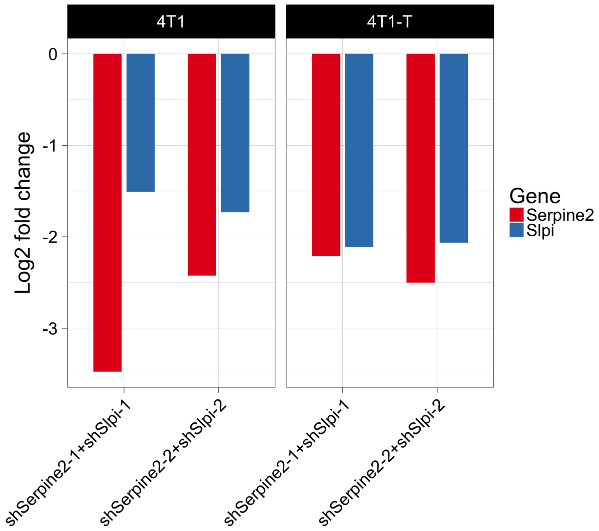


FIGURE 2.8: **Knock-down of pairs of genes using the dual UltramiR shRNA expression vector** Pairs of shRNAs targeting SerpinE2 and Slpi were transduced in 4T1 and 4T1-T cells. Shown is the log2 fold change of mRNA knockdown levels in cells infectected with dual shRNA constructs compared to empty constructs.

negative effect on cell growth or survival are depleted from the pool of cells. High-throughput sequencing provides a convenient way to estimate said abundance. When the shRNAs are delivered by viruses and integrated stably into the genome, the guide sequence can be amplified by PCR, Illumina adapter sequences added, and the shRNA counts can be obtained after sequencing. This experimental setup can in principle be adapted easily to sequence dual shRNA screen.

My initial approach was to take advantage of paired-end sequencing technology, for which both ends of a DNA molecule can be sequenced simultaneously. This requires extracting the variable sequences of both guide RNAs from genomic DNA by PCR. I attempted this on genomic DNA harboring a pair of p16 targeting shRNAs separated by a 225bp (constructs 1-225-2 and 2-225-1, figure 2.4A). Despite trying a large number of conditions and primers, I was never able to PCR a complete shRNA to shRNA fragment. I reasoned that the PCR amplification was perhaps hindered by the complex secondary structure of the two hairpins or of the spacer when denatured. The sequence of the longer spacers (figures 2.4B, 2.4C) was thus amplified from A/T rich regions of the yeast genome or of the Kanamycin resistance gene and showing few secondary structures as predicted by mFold (Zuker, 2003). However, even for these constructs, amplifying both shRNAs simultaneously proved to be unfeasible.

Since each shRNA could be amplified separately, I reasoned that one of the shRNA could be barcoded with a random 25bp sequence that would be placed in between the pair of hairpins. Sequencing each pair would then require two paired-end sequencing runs. The first run would be to sequence the 5′ shRNA and the barcode, and the second to sequence the barcode and the 3′ shRNA. Putting together all the data would allow to count each pair in the pooled screen. To clone such barcoded construct, I first ligated the first pool of shRNAs in the vector, then PCRed the second pool with a primer harboring 25 random nucleotides as a barcode, and added the amplicon to the vector. The barcode thus identifies the second shRNA. The resulting construct harbored all pair-wise combinations of the shRNAs, with a 25bp barcode in between. For this strategy to be viable, each barcode needs to identify uniquely the second shRNA. To test this approach, I cloned a dual shRNA library with 300 shRNAs, for a total of 90 000 different pairs. Despite using a 25bp random barcode, which represents more than a trillion different possibilities, I found upon sequencing of the library that most barcodes were shared across multiple 3′ shRNAs (figure 2.9). The second shRNAs were not uniquely identified by the barcode which made it impossible to extract the pairing information from the two sequencing runs.

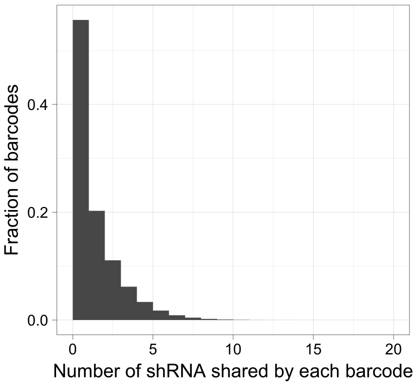


FIGURE 2.9: **Barcoding of shRNAs using random 25bp sequences** A dual shRNA library harboring 90 000 different combinations was built. The 3′ shRNA of dual shRNA constructs was barcoded using a 25bp sequence of random nucleotides. The barcode and 3′ shRNA were sequenced using Illumina paired-end sequencing. Shown is the number of shRNAs that were associated with a given barcode.

As barcoding shRNAs by PCR using primers with stretches of random nucleotides proved unfeasible, I decided to barcode all the shRNAs one by one, before assembling the dual vector.

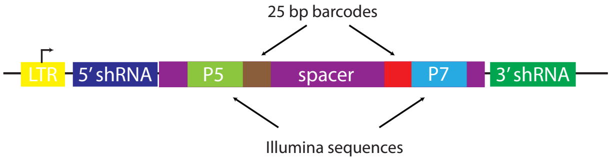


FIGURE 2.10: **Barcoding pairs of shRNAs using two 25bp barcodes** Schematic of the modified dual shRNA vector. ShRNAs are separated by a 500bp spacer. Each shRNA is uniquely barcoded by a 25bp sequence, flanked by Illumina adapter sequences to facilitate generation of high-throughput sequencing libraries.

The spacer was modified to allow for barcoding of both shRNAs (figure 2.10). Pairs of shRNAs in this vector can be identified by sequencing both barcodes simultaneously in a paired-end sequencing run. An additional advantage of barcoding both shRNAs is that it reduced greatly the size of the amplicon that needs to be sequenced as the barcodes are in the middle of the spacer. As the sequences are flanked by Illumina adapter sequences, libraries can be generated in a one-step 30 cycle PCR as opposed to the two 25 cycles PCRs that are generally necessary to sequence miR-30-based shRNA screens, which helps reduce PCR bias. This strategy was tested with a pair of barcoded p16 shRNA transduced in A385 cells. Recovery of the pair of barcodes could be easily achieved by PCR. Barcoding the shRNAs before hand with a unique sequence comes at the cost of more complicated cloning steps for dual shRNA libraries but facilitates the extraction of the information from the cells.

2.3.3 Cloning of complex 2D shRNA plasmid libraries

Most of the designs mentioned above were tested on a single pair of shRNAs. The scale of the screens that I wanted to perform required cloning of complex pools of pairs of shRNAs in the vector to be feasible. Barcoding the two hairpins makes the cloning of such libraries challenging: each of the barcodes needs to identify uniquely an shRNA and be different enough from all the other barcodes to avoid confusion during sequencing. In addition, the cloning needs to be efficient enough that all pairs are present at similar abundance in the final dual shRNA library. To address this issue, I developed a two-step cloning strategy (figure 2.11). In the first step, all of the shRNAs of interest are cloned in a “sequencing vector". This vector holds an ultramiR cassette, flanked by spacer sequences including two 25bp random barcode. The empty

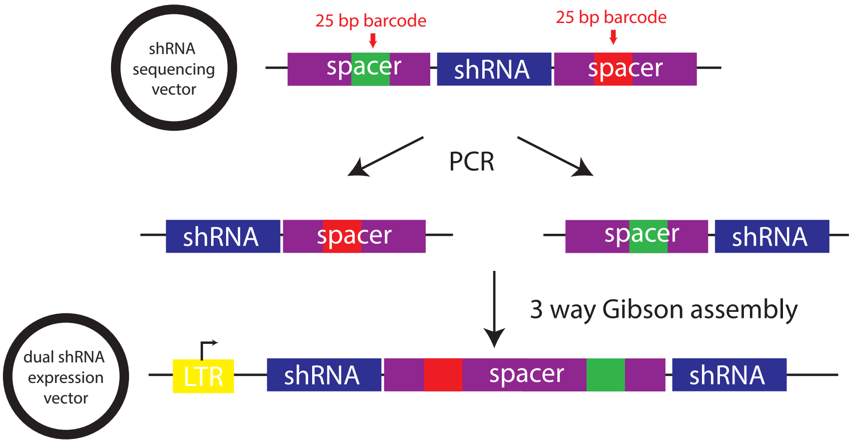


FIGURE 2.11: **Efficient cloning of dual shRNA libraries** Pairs of shRNAs are cloned in the 2D shRNA expression vector in two steps. First, shRNAs of interest are cloned as pools in a sequencing vector with two 25bp barcodes. After transformations, colonies are picked and Sanger sequenced to identify which barcodes are linked to which shRNAs. In the second step, the shRNAs with the first or second barcode are amplified by PCR and all pair-wise combinations are assembled by Gibson assembly in the expression vector.

sequencing vector was generated as a complex pool of vectors with millions of different barcode combinations. The shRNAs of interest can be cloned as a pool in this vector, and colonies can be picked, grown, prepped and sent for Sanger sequencing. The Sanger sequencing covers the entire ultramiR cassette as well as the two regions with barcodes. Each barcodes needs to robustly identify an shRNA, even when sequencing errors are introduced. To filter barcodes with similar sequences, we used the Levenshtein distance, which measures the number of edits (insertions, deletions or substitutions) needed to transform one string of characters to another. Constructs for which the Levenshtein edit distance of any barcode to another in the library is less than three were discarded. Barcodes that had a GC content greater than 80% were also removed. Glycerol stocks of constructs passing these criteria were arrayed in 96-well plates.

To clone all the pair-wise combinations of the barcoded hairpins in the 2D shRNA expression vector, two fragments of the sequencing vector need to be amplified by PCR: the shRNA and the downstream barcode or the shRNA and the upstream barcode. These two PCRs are performed for each arrayed shRNA in the sequencing vector, with a limiting amount of primers to obtain similar amounts of amplicons regardless of the concentration of the template. After

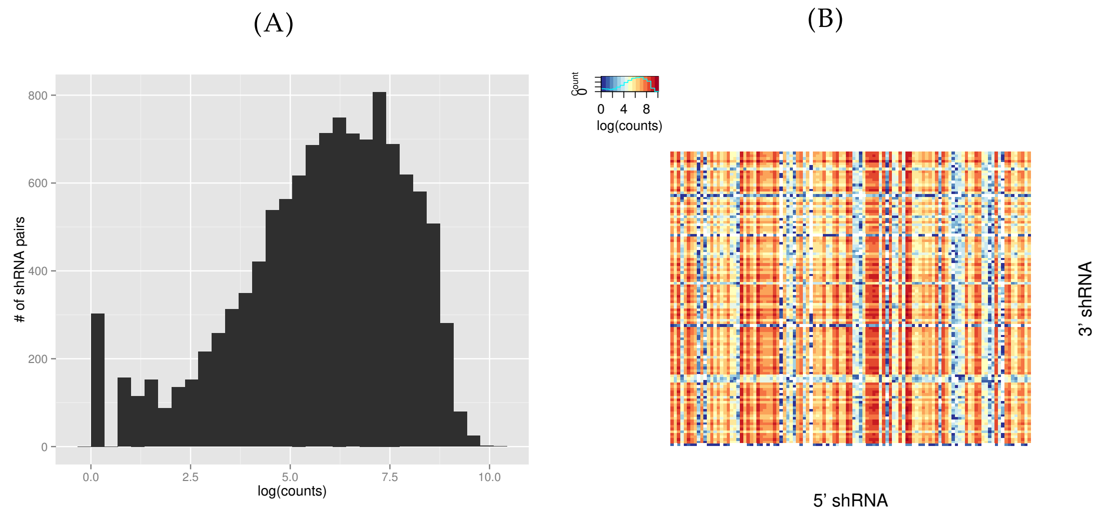


FIGURE 2.12: **Validation of the dual shRNA library cloning strategy** ~110 shRNAs were cloned in the 2D vector for a total number of ~ 12 000 pairs. The abundance of each pair was estimated by high-throughput sequencing. Shown in **(A)** is the distribution of pair counts. **(B)** Heatmap representation of counts for each shRNA pair.

purification, all of the amplicons are pooled and cloned into the expression vector in a three-way Gibson assembly, a highly efficient cloning method allowing for the assembly of multiple overlapping DNA fragments. Using this method, all pair-wise combinations of shRNAs are generated and each shRNA can be identified using its flanking 25bp barcode.

To validate this design, ~110 shRNAs were cloned in the sequencing vector and Sanger-sequenced. A 2D dual shRNA library comprising ~12 000 pairs was assembled. The abundance of each pair in the pool of vectors was estimated by sequencing the barcodes on an Illumina MiSeq.

The distribution of counts per hairpin pairs shows that most pairs are similarly represented in the pool (figure 2.12A). Furthermore, more than 90% of all possible pairs were successfully cloned (figure 2.12B). The missing pairs were mostly due to shRNAs been absent altogether at one or the other position (blue rows in the heatmap) which mostly likely is due to failed amplification during the PCR of the shRNA from the sequencing vector.

Overall, the expression vector and cloning strategy I developed allows the efficient cloning of complex dual shRNA libraries. The optimized spacer length and miRNA scaffold enables robust knockdown of each target gene. Finally, the barcoding strategy I designed enables multiplexed 2D RNAi screens to be carried out and analyzed easily.

2.3.4 Multiplexed 2D shRNA screens of melanoma cell lines

The first gene set I wished to interrogate in melanoma was comprised of all “druggable genes". I thought that focusing on genes for which small molecule inhibitors existed or were being developed would yield greater therapeutic benefits and allow for validation of hits *in vivo* with drug combinations. Four melanoma cell lines were selected for this initial screen: A-375, WM-266-4, SK-Mel-5 and SK-Mel-8. These cell lines were selected from a larger panel of 10 cell lines for their screenability: they can be infected with retroviruses, grow rapidly which reduces the overall length of the screen, and their size is relatively small which allows a large number of cells to be grown on a given area of cell culture plates. Additionnally, these cell lines exhibit a variety of mutations in genes commonly altered in melanoma (table 2.1). Although all of these cell lines have the most common BRAFF V600E mutation, they belong to different melanoma subtypes as classified by large genomic studies and based on alteration in other genes such as *TP53*, *CDKN2A* and *PTEN* (Cancer Genome Atlas Network, 2015; Hayward et al., 2017).

	A-375	SK-Mel-5	SK-Mel-28	WM-266-4
BRAF	V600E/V600E	V600E/WT	V600E/V600E	V600E/V600E
TP53	WT	WT	L145R/L145R	WT
CDKN2A	E61*/E61*	Δ/Δ	WT	WT
PTEN	WT	WT	WT	Δ/WT

TABLE 2.1: **Genotype of the four melanoma cell lines used for the screen**

Gene set and shRNA selection

The number of constructs in a dual shRNA library grows quadratically with the number of single shRNAs considered. Based on the experience of the lab with RNAi screens, pooled screens are feasible if the total number of shRNAs used is less than ∼70 000. Screening in smaller pools reduces the noise inherent to such large-scale experiments. To keep the representation of all shRNAs in the pool of cells, generally 1000 cells per shRNA are kept throughout the screen. Limiting the screen to a targeted set of genes thus also limits the number of cells that need to be grown once and makes the experiment practical. For dual shRNA screens, I wanted to limit the number of target genes to 100 to 200, for a total pairwise complexity of 10 000 to 40 000 constructs.

The first gene set we wished to interrogate in the four selected melanoma cell line comprised all “druggable genes". As a basis for this set, we chose to use the Sophic Druggable Genome database that reports around 4 000 genes as druggable. This list was generated by integrating several databases reporting such genes (Hopkins & Groom, 2002; Russ & Lampel, 2005). In addition, this gene list was enriched by adding genes for which the protein sequence shared similarities with known druggable genes and by computationally text mining the literature (Sophic Alliance Inc, 2010).

To further reduce the size of this gene set, we considered genes that were over-expressed in our melanoma cell lines compared to melanocytes. To assess gene expression levels, I performed total RNA sequencing on the five cell lines, in duplicates. 112 genes were found to be significantly over-expressed in at least three of the four melanoma cell lines when compared to the melanocytes (figure 2.13A). Out of this list 67 genes were over-expressed in all four cell lines.

A pathway analysis of these 112 genes shows that the list includes mostly genes involved in metabolism, immunity and signaling pathways commonly dysregulated in cancer such as the EGFR or ERBB2 pathway (figure 2.13B).

To select shRNAs targeting these 112 genes, we used shRNA sensor data that had been generated to validate the shERWOOD algorithm. In this assay, the top 10 shRNAs as predicted by shERWOOD for all 2 000 druggable genes had been tested for activity *in vivo*. For RNAi screen, multiple targeting molecules per genes are generally used to reduce false-positives. As 10 shRNAs per gene had been validated *in vivo*, we selected the two best scoring shRNAs in the sensor assay to build dual shRNA libraries for the subset of the druggable genome wished to interrogate. The two sets of 112 hairpins were cloned into the sequencing vector and two pools of $\sim$12 000 pair-wise combinations were assembled as described above. I decided to screen the two libraries independently to lower the total number of constructs per experiment.

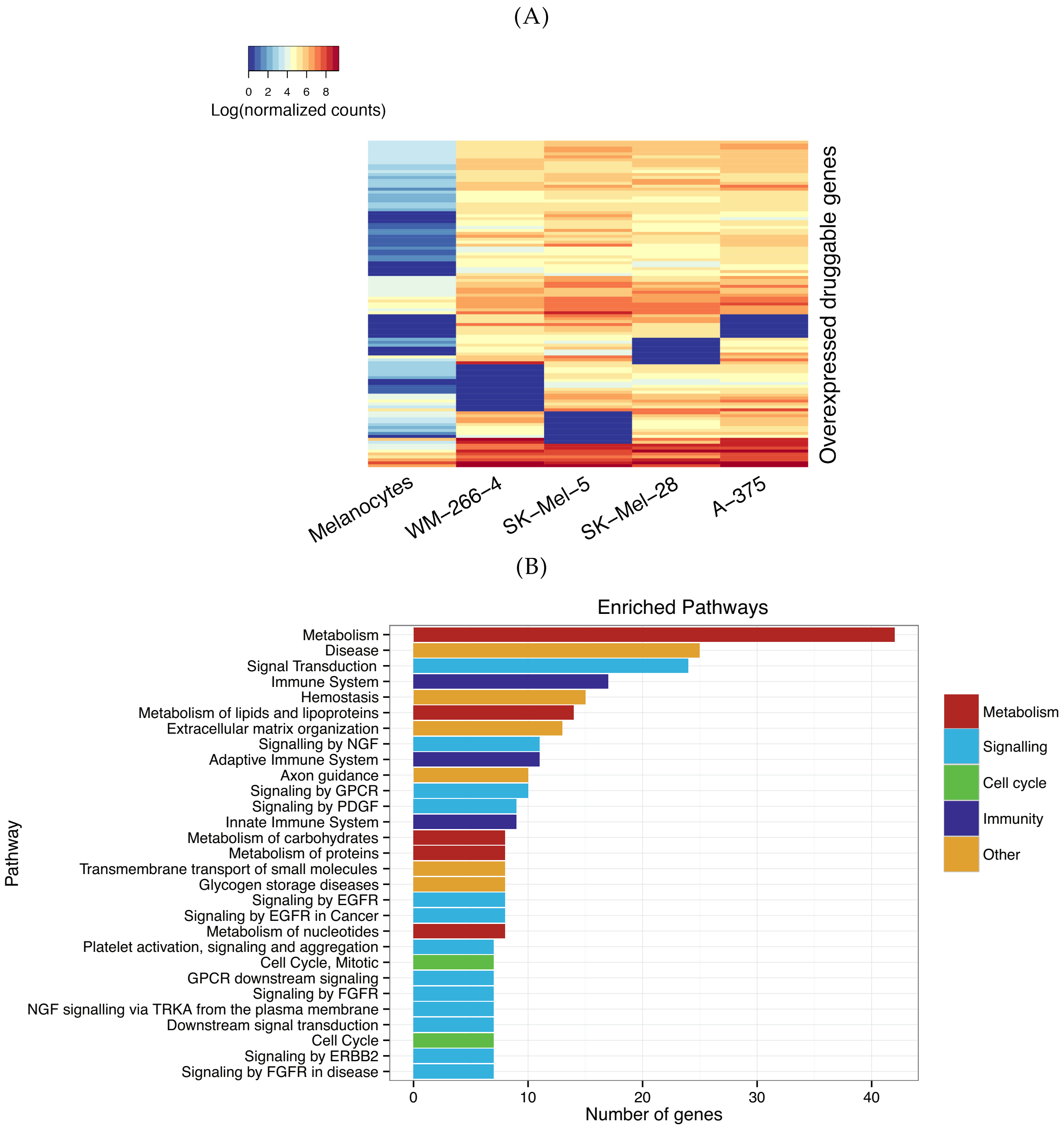


FIGURE 2.13: **Selection of a gene set for combinatorial screening of melanoma cell lines (A)** Total RNA sequencing was performed on A-375, Sk-Mel-5, Sk-Mel-28, WM-266-4 and melanocytes. 112 druggable genes were found to be significantly over-expressed in at least three of these four cell lines compared to the melanocytes. Shown is the log of the normalized counts of these 112 genes for these cell lines. **(B)** Pathway analysis of the 112 over-expressed genes in the four melanoma cell line compared to melanocytes.

Screening of the melanoma cell lines

The two 2D shRNA libraries were packaged into retroviruses used to infect the A-375, SK-Mel-5, SK-Mel-28 and WM-266-4 cell lines. Infections were performed at low MOI to reduce the likelihood of double infection events. At least 30 million infected cells were obtained for each cell line to keep the representation of each shRNA pair in the pool. Two days after infection, half of the cells were collected as a reference timepoint (T0). The rest of the cells were selected with Neomycin and grown for at least 12 doublings and a final time point was collected (T12). This was done in triplicate. For samples infected with the first library, genomic DNA was extracted and the pairs of barcodes identifying the shRNAs were amplified by PCR and sequenced on a Illumina HiSeq. Each sample was sequenced with at least 15 million reads. The abundance of each shRNA pair was measured by each pair's number of reads mapping to the corresponding barcodes. The log-fold change of counts of each construct between T12 and T0 was calculated. An empirical Bayes moderated t-test was applied to the log-fold changes of the three replicates to estimate the significance of the depletion or enrichment (R limma package (Smyth, 2004)). The log-fold change for all pairs in the first library is shown in figure 2.14.

Since the dual shRNA library harbor all pair-wise combinations of hairpins, the lethality of knocking down a single gene can be assessed by examining the log-fold change of pairs harboring two shRNAs targeting the same gene. As a sanity check for this first screen, I looked at the depletion status of three well-studied genes in melanoma or other cancers: CTGF, PLK1 and CDK4. Connective tissue growth factor (CTGF) has recently been identified as a therapeutic target for melanoma and significantly reduces tumor sizes and metastasis upon knockdown. Polo-like Kinase 1 is a kinase involved in the G2/M transition. It has been identified as a therapeutic target in lung, colon and pancreatic cancer. CDK4 is a key player in the cyclin D-CDK4/Rb pathway which is misregulated in 90% of melanoma and promoting cell-cycle progression. The knockdown of this gene was expected to be lethal especially in cells in which CDKN2A which generally inhibits CDK4 was mutated (A-375 and SK-Mel-5). The three shRNA pairs targeting these genes are significantly depleted across all cell lines, with the exception of the SK-Mel-5 cell line, which is resistant to PLK1 knockdown and SK-Mel-28, which is resistant to CDK4 knockdown.

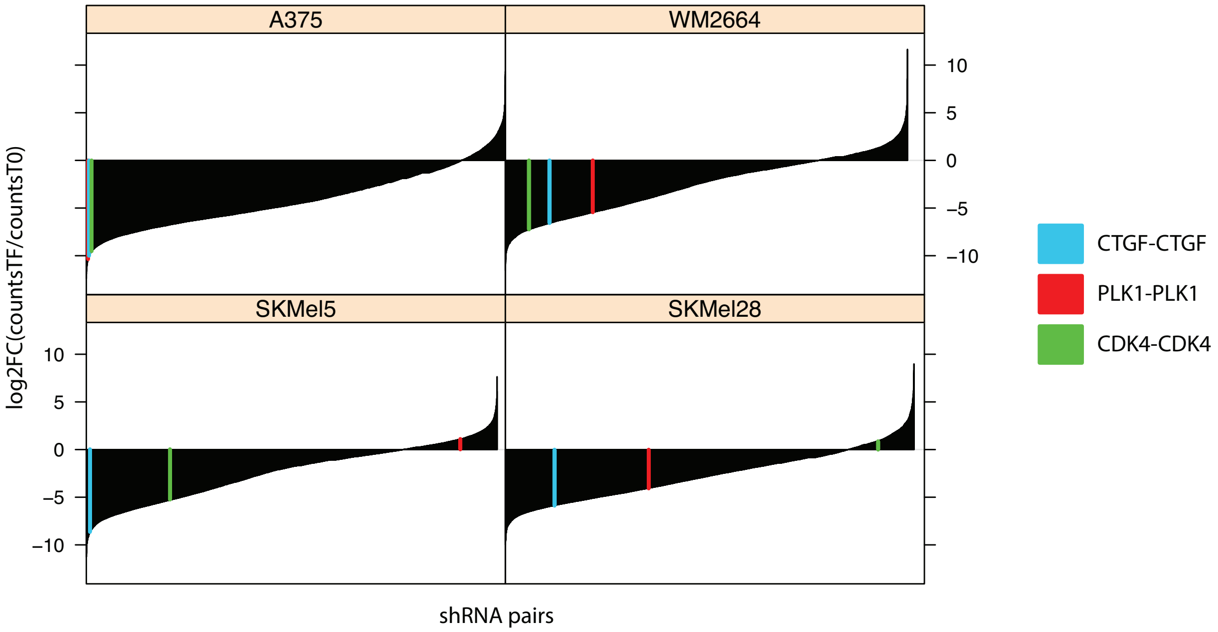


FIGURE 2.14: **Combinatorial screening of four melanoma cell lines** A dual shRNA library harboring all possible pairs of 112 shRNAs targeting a subset of the druggable was screened in A-375, SK-Mel-5, SK-Mel-28 and WM-266-4. Shown is the log2 fold change of the abundance of each pair between the final and initial timepoint. Highlighted are the fold changes of pairs harboring two shRNAs targeting CDK4, CTGF or PLK1.

Although these results seemed promising, we realized after the laboratory move to Cambridge that the Institute's tissue culture incubators had not been calibrated in several years. CO_2 levels were found to vary wildly between 1 and 10%. As a result, the data from this initial screen could not be used and the second screen was not sequenced. One surprising observation that came out of this experiment was that in all four cell lines, nearly 80% of shRNA pairs were depleted. Negative controls were not initially included in the dual shRNA libraries as in large scale RNAi screens a large number of shRNAs are generally not depleted. This may not be the case here as the number of target gene is limited to a small number of druggable genes, selected based on their expression profiles, and if any of these genes is essential for cell proliferation, the 225 pairs harboring an shRNA targeting these genes at the first or second position would be depleted. Since the screens needed to be repeated, I added 20 shRNAs targeting olfactory receptors. These olfactory receptors are not expressed in the melanoma cell lines so targeting them should not impede or favor the proliferation of the cells. This increases slightly the complexity of the 2D libraries, up to ∼15 000 constructs. For the rest of this chapter, I focused on the A-375 cell line as it is the fastest growing of the four chosen lines and the most amenable to high throughput screening.

A-375 cells were transduced with both dual shRNA libraries and the screen was performed and sequenced as described above. Figure 2.15 shows the results of the screen, represented as the depletion ratios of each constructs versus their abundance at the initial timepoint, for both libraries.

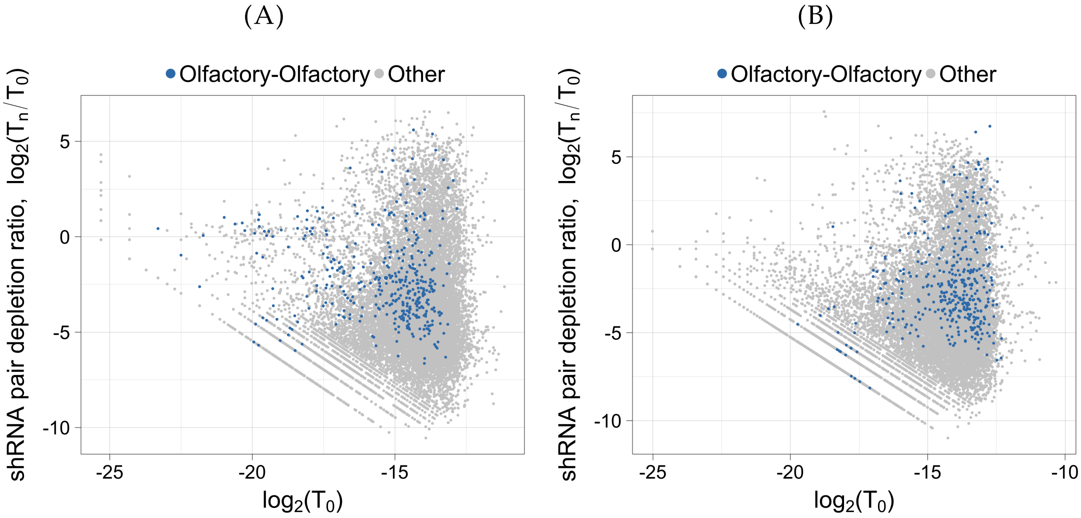


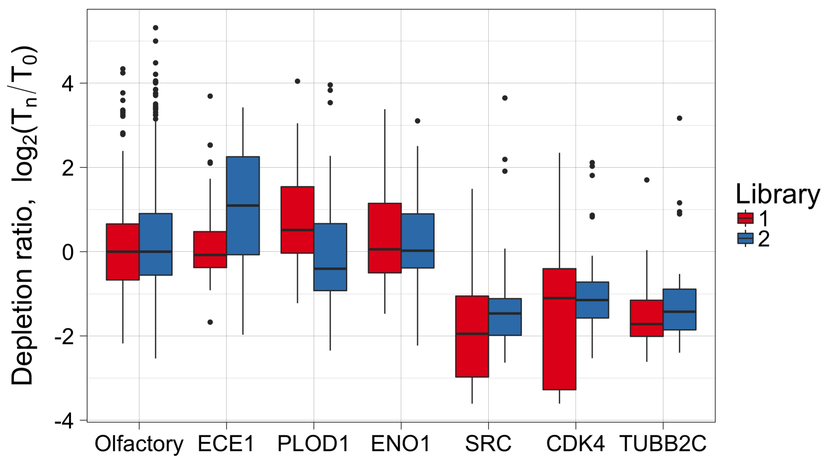
FIGURE 2.15: **Combinatorial screening of the A-375 melanoma cell lines** Two dual shRNA libraries harboring all possible pairs of 112 shRNAs targeting a subset of the druggable genome as well as 20 olfactory receptor shRNAs were screened in A-375. Shown on the y-axis is the log fold change of read counts for each pair at the final timepoint compared to the initial timepoint. The x-axis is the log of counts at the initial points. Pairs composed of two shRNAs targeting olfactory receptors are shown in blue. **(A)** is the first library and **(B)** the second. The log-fold change of the three replicates was averaged.

As observed in the previous screens, more than 80% were depleted in these screens. Surprisingly, the pairs harboring two olfactory receptors were overall depleted as well in both libraries (median log2 fold-change of -2.6 for library 1 and -2.8 for library 2 for the olfactory-olfactory distribution). This can perhaps be explained by the outgrowth of a large number of constructs that grow faster than the olfactory targeting controls. In both libraries, at least 800 constructs are enriched over 4 fold in the final timepoint compared to the T0 and end up taking a large fraction of reads. The enriched pairs were generally not consistent between libraries so I focused on the depleted constructs. To normalize across each replicate, the counts were scaled using the mean and standard deviation of the olfactory-olfactory targeting pairs. Following this normalization, the limma package was used as described above to identify significantly depleted constructs (FDR < 0.05).

A pair of shRNAs can be depleted if one or both of the targeted genes impede cell proliferation. For follow-up experiments, I wished to focus on pairs for which the deleterious effect of knocking down a pair of gene was synergistic and especially on pairs targeting genes that were only depleted when knocked-down in combination. As these libraries include 20 shRNAs targeting olfactory gene, each druggable gene shRNA is paired in 40 pairs with an olfactory gene shRNA. The depletion rates of these constructs can be used to assess the lethality of druggable gene targeting shRNAs. A druggable gene was considered a “hit" if in both libraries the mean of the distribution of depletion rates of all pairs harboring an shRNA targeting this gene and another targeting an olfactory receptor was significantly lower than 0 (t-test, p-val < 0.05). A representative example of the distribution of three genes classified as “hits" or not is shown in figure 2.16A. Overall, 32 druggable genes where found to be necessary for cell proliferation. The knock-down of another 43 genes had no impact on cell growth (figure 2.16B). The remaining 37 genes could not be classified in these two categories as results from both shRNA libraries were not in agreement. This can potentially be a consequence of off-target effects.

To select pairs to investigate further, I chose constructs that were significantly depleted and harboring two shRNAs that were not hits when paired with olfactory receptors. Additionally, at least one pair of shRNA (either shRNA1-shRNA2 or shRNA2-shRNA1) needed to be depleted in both libraries for that pair of genes to be considered in validation experiments. Using such criteria, I wished to select pairs of genes for which knockdown induces synergistic deleterious effect on the proliferation of A-375 cells. An analysis of the data of both libraries showed that around 230 pairs of genes depleted only when both genes were targeted in combination. To further select the most promising candidates for one by one follow up experiments, I decided to perform a second validation screen to eliminate any false positives. 2D shRNA libraries can be built very efficiently when all pair-wise combinations of two sets of genes need to be interrogate. However, generating smaller libraries with specific shRNA pairs requires one

(A)



(B)

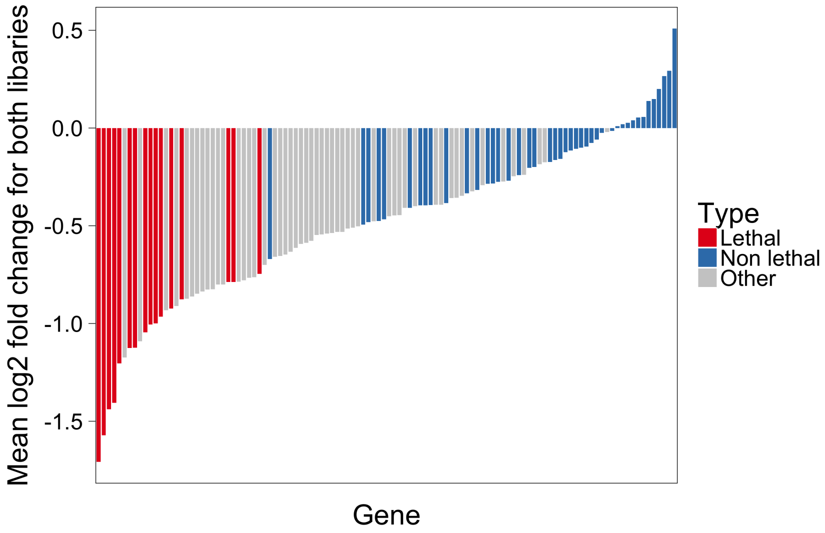


FIGURE 2.16: **Identification of straight lethal hits in the combinatorial screens (A)** A druggable gene was considered a lethal hit if in both libraries the mean of distribution of depletion rates of all pairs harboring an shRNA targeting this gene and another targeting an olfactory receptor was significantly lower than 0 (t-test, p-val $<$ 0.05). Distributions of olfactory receptors and 3 genes classified as non-lethal (first four) and 3 genes considered to be lethal (last 3). **(B)** Average log2 fold change of all constructs harboring an shRNA targeting a druggable gene and another targeting an olfactory receptors. Bars are colored by gene lethality as described in (A).

by one cloning of all pairs before pooling. As I was concurrently designing vectors for harboring pairs of short guide RNAs (sgRNAs) to be used for CRISPR/Cas9 experiments (chapter 3), I decided to validate the dual shRNA screen with a combinatorial CRISPR screen. The variable region of sgRNAs is short compared to a full hairpin (20bp vs 63bp for passenger, guide and loop). The smaller length of the variable region allows for printing complex libraries of pairs of sgRNAs on DNA chips for which the total size of the oligo is limited. (M.A. Cleary et al., 2004). These CRISPR libraries can be cloned in expression vectors harboring two U6 promoters driving the expression of the sgRNAs independently. The vector I used will be described in details in chapter 3, but its schematic map is shown in figure 2.17 for clarity purposes.

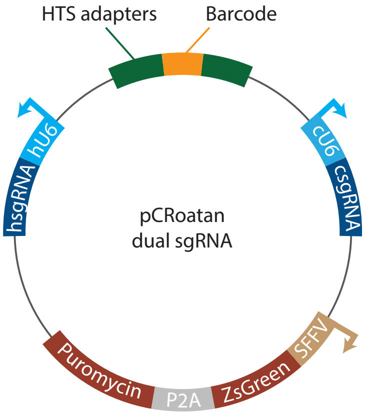


FIGURE 2.17: **A dual sgRNA expressing vector to knock-out pairs of genes** Schematic map of the lentiviral vector used to express pairs of sgRNAs (hU6: human U6 promoter, cU6: chicken U6 promoter, HTS: high-throughput sequencing, SFFV: spleen focus-forming virus promoter).

In this dual sgRNA lentiviral expression vector, the first sgRNA is driven by a human U6 promoter and the second by a chicken U6 promoter. As both sgRNAs are separated by more than 1kb, they cannot be sequenced at the same time using paired-end sequencing for which the amplicon needs to be smaller than 700bp. A 20bp barcode was thus added to the vector to identify each pair of sgRNAs. The two sgRNAs as well as the barcode can be printed on the same oligo so the 20bp identifying each sgRNA pair are known by design. Restriction sites are added between the first sgRNA and the barcode and between the barcode and the second sgRNA to add the promoters. The cloning of these libraries in the lentiviral expression vector requires four steps. First the chips are cloned in an intermediary vector lacking any outside cutter restriction enzyme sites that are necessary to perform this cloning. Next, the human

U6 promoter and sequencing adapters are cloned next to the first sgRNA and the barcode, followed by the remaining adapters and chicken U6 promoters between the barcode and second sgRNA. Finally the complete sgRNAs,barcode and promoters fragment is inserted in the lentiviral vector.

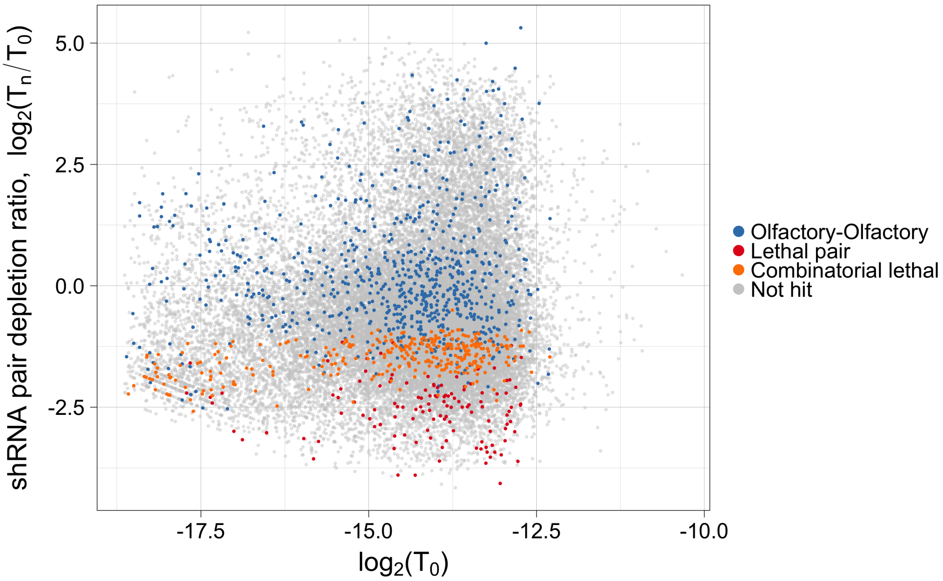


FIGURE 2.18: **Selection of gene pairs to target in validation experiments** shRNA pairs depletion ratio versus log counts at the initial timepoint, for both libraries. The colors corespond to olfactory-olfactory pairs (blue), pairs targeting two lethal genes (red) and combinatorial lethal pairs that will be considered for follow-up experiments (orange).

For this validation screen, the library was composed of the ∼230 pairs identified as synergistically deleterious. As a positive control, the 40 pairs of lethal genes that were most depleted consistently in both dual shRNA libraries were also added to the library. The depletion rates of both the pairs of interest as well as the positive controls is shown in figure 2.18. As a negative control, 264 pairs harboring two sgRNAs targeting olfactory receptor genes were included. For each druggable gene in the library, pairs for which one sgRNA targeted the druggable gene and the other an olfactory receptor were added to estimate the effect of knocking out the gene on its own. To control for false positive and false negative effects, 8 different pairs of sgRNAs for each gene pair were included in the library. In total the dual sgRNA library harbors over 3 800 pairs. Each pair was uniquely barcoded with a 20bp long sequence. These barcodes were selected by randomly generating 10 000 20mer and selecting sequences that had a GC content

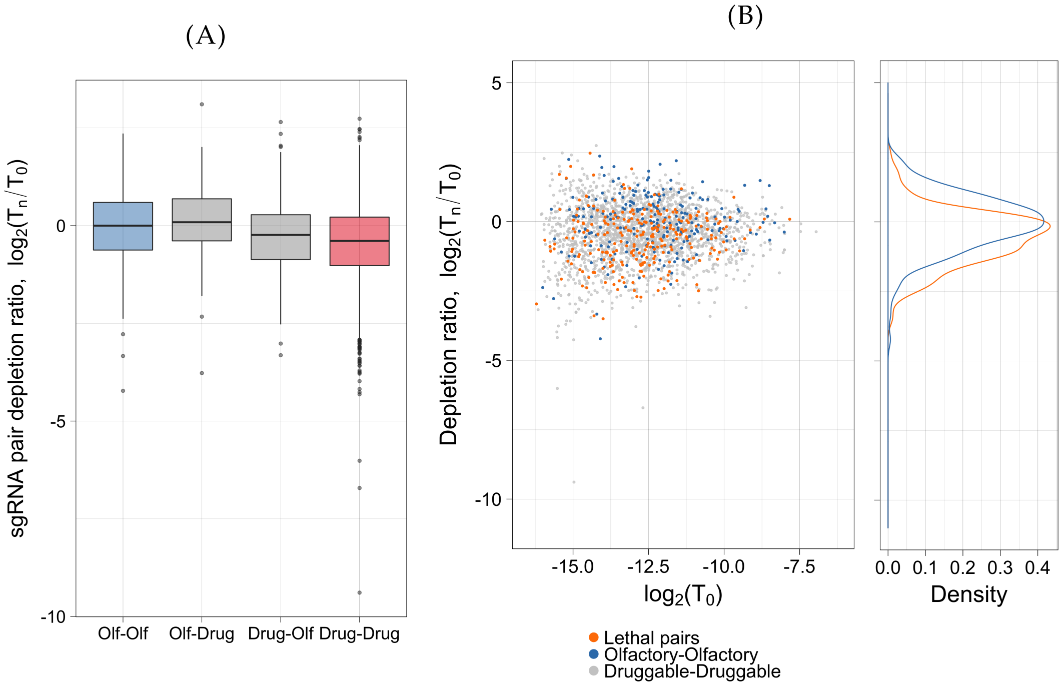


FIGURE 2.19: **Validation of combinatorial hits from the RNAi screen using CRISPR/Cas9 (A)** Depletion ratio of sgRNA constructs targeting pairs of olfactory genes (Olf-Olf), pairs of druggable gene (Drug-Drug), or one of each (Drug-Olf, Olf-Drug). **(B)** MA plot all constructs. Colors show the lethal pairs identified in the RNAi screen (positive controls) or pairs targeting two olfactory receptor genes (negative control).

between 30 and 70%. Any sequence with a Levenshtein edit distance of less than three to any other barcode was also removed.

To perform the CRISPR screen, a cell line stably expressing the Cas9 protein was generated. A-375 cells were first transduced with the lentiCas9-Blast plasmid (Sanjana, Shalem, & Zhang, 2014). Following selection, cells were single cell sorted in a 96-well plate. To select clones for which Cas9 was highly active, clones were transduced with a plasmid expressing the fluorescent protein ZsGreen and an sgRNA targeting ZsGreen expressed from a U6 promoter. For each clone, percentage of fluorescent cells was tracked over time and use to estimate the activity of the Cas9 protein. A clone for which at least 60% of the cells lost ZsGreen signal was selected to perform the screen (data not shown).

The library of dual sgRNA was packaged into lentiviruses, and transduced in A-375-Cas9 cell line at low MOI, in triplicates. As done for the RNAi screens, half of the cells were collected two days after infection for an initial timepoint, and the rest was grown for ∼12 doublings and harvested for a final timepoint. For all timepoints, the barcodes identifying the pairs of sgRNA were amplified by PCR and sequenced on an Illumina MiSeq. For each pair, depletion rates were measures using the limma package as described previously for the 2D RNAi screen.

To validate my approach of using a orthogonal CRISPR screen to validate hits from a dual shRNA screen, I first compared depletion rates of pairs harboring two sgRNAs targeting druggable genes to the ones comprised of one or two olfactory receptor targeting sgRNAs. As all the validation pairs with two druggable gene targeting sgRNAs were chosen because they depleted in the RNAi screen, they should in principle also deplete in the CRISPR screen. As expected, constructs with two sgRNAs directed against druggable gene were overall significantly more depleted than the negative control (figure 2.19A rank-sum test, pval < 0.01). 72 druggable genes were present in all the different pairs of this screen, and amongst those only 13 were found to be lethal in the RNAi screen which explains why the pairs harboring only one druggable gene deplete less than those with two (figure 2.19). Forty positive control pairs (ie. pairs that were consistently and significantly depleted in the RNAi screen) were included in the validation screen for a total of 320 positive control constructs (8 different sgRNA pairs per gene pairs). Surprisingly, most of these positive controls did not hit in the CRISPR screen. Their distribution is slightly more depleted than the olfactory-olfactory population but this is not significant (figure 2.19B). 274 positive-control constructs out of 320 were present above threshold in the T0 and only 64 of these pairs were significantly depleted in the final timepoint. Overall, only 9 of the 40 positive control gene pairs had at least half of the targeting constructs depleted.

To further investigate this issue, I examined the depletion rates of the pairs for which both sgRNAs target one of the 11 lethal gene identified in the RNAi screens (figure 2.20). Although one or two constructs were significantly depleted for all genes, only two genes had more than half of the targeting constructs significantly depleted. This can perhaps be explained by the strong variability of the construct log-fold changes that was observed in the three screen replicates so the depletions are not statistically significant.

Although some of the positive controls that depleted significantly in the RNAi screen did not in the CRISPR screen, this could be explained by the difference in effector selection and mechanism of perturbation. All the shRNAs used in the RNAi screen were validated in a

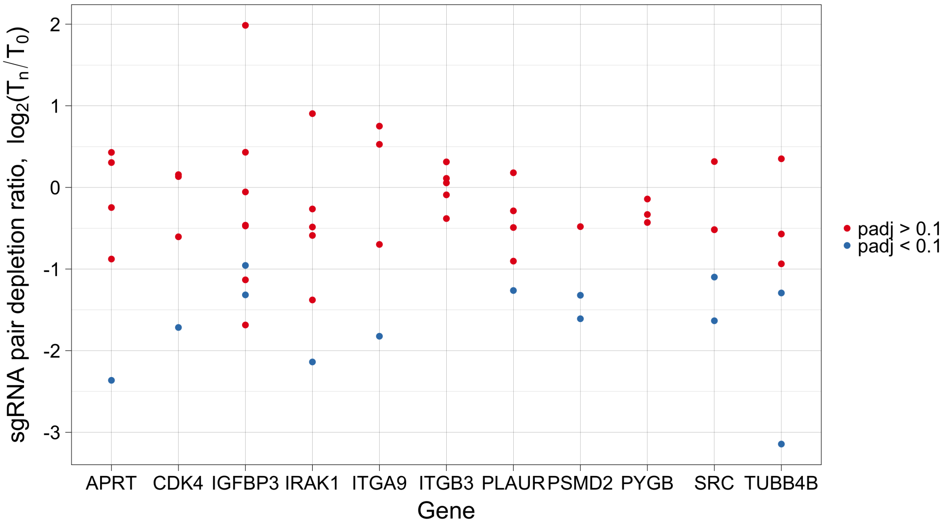


FIGURE 2.20: **Depletion of positive controls in the CRISPR/Cas9 validation screen** Depletion rates of sgRNA constructs targeting lethal genes identified in the RNAi screen.

sensor assay whereas the sgRNAs were predicted *in silico* and not tested *in vitro* before the screen. In addition, targeting a gene with CRISPR/Cas9 does not always lead to a functional knock-out in all the cells as this depends on the scar and mutations left on the loci following the repair of the DSB. Knocking-out two genes simultaneously most likely leads to the generation of mosaics of cells in which both, only one, or none of the genes are mutated, especially if the potency of sgRNAs has not been tested. This would in turn lead to an increased false-negative rate in a pooled CRISPR screen.

Despite these limitations, some true-positives can be identified in the validation CRISPR screen (figure 2.21). These pairs of genes were targeted by at least three independent significantly depleted dual sgRNA constructs. None of the constructs targeting these genes individually depleted significantly in both screens. Although some genes among these hits, such as the Growth factor receptor-bound protein 2 (GRB2) (Lowenstein et al., 1992; A. Cheng et al., 1998; Giubellino et al., 2007) and the receptor tyrosine-protein kinase ERBB3 (Sergina et al., 2007; Engelman et al., 2007; Miller et al., 2009), have been previously shown to be involved in cancer proliferation or resistance mechanisms to treatment in a variety of cancers, most have not been studied in melanoma. As all of these pairs have only been identified as lethal in pooled screen and further experiments need to be performed, in one-by-one assay, using shRNAs, sgRNAs

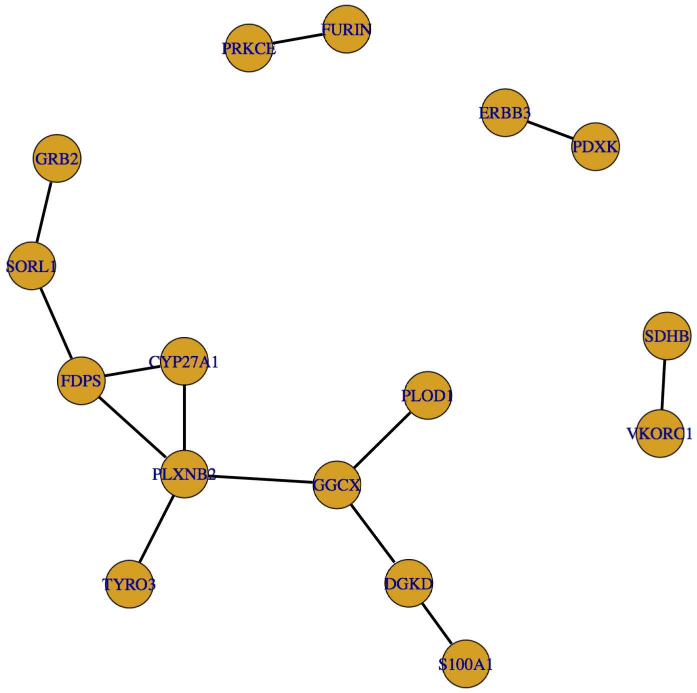


FIGURE 2.21: **Synthetic lethal combinations identified in both RNAi and CRISPR screen** Graph representation of all combinatorial pairs identified as synthetic lethal in both the RNAi and CRISPR screen. Pairs (genes connected by black lines) shown here deplete when in combination, but constructs targeting each gene in individual knock-down/knock-outs do not deplete.

or small molecule inhibitors, to confirm their combinatorial lethality.

2.4 Discussion

Here, I have outlined an approach to perform high-throughput combinatorial RNAi screens using an optimized expression vector harboring pairs of shRNAs. In this vector, two shRNAs, separated by a spacer are expressed from the same promoter. To increase mature small RNA levels of both shRNAs, I tested a range of spacer length ranging from 0 to 800nt. Spacers of 200 to 500nt allowed good processing of both hairpins and I selected the 400nt spacer for to use in the expression vector. In addition, I designed a cloning strategy that allows large libraries comprised of all pair-wise combinations of sets of genes to be easily assembled.

Our initial experiments with miR30 based shRNAs highlight the importance of selecting potent shRNAs and optimizing synthetic miRNA backbones for efficient small RNA production. Building on large shRNA efficacy datasets, shERWOOD, a machine learning algorithm predicting hairpin potency was developed. Coupled with a canonical miR30 backbone, named ultramiR, it permits any gene to be knocked-down with high efficiency. These tools have wide applications beyond dual shRNA screens, both for genome-wide screens and one-by-one experiments. We have thus built a fifth-generation genome-wide library of sanger-sequenced verified shERWOOD shRNAs in the ultramiR scaffold. These libraries harbor on average 5 shRNA for each gene of the human and mouse genome for a total of ∼75 000 human and ∼60 000 mouse shRNAs. The availability of individual sequence-verified shRNAs also allows screens to be performed in an arrayed format, or the generation of custom pools of shRNAs targeting a specific gene-set. This was particularly useful to obtain some of the shRNAs used in this project.

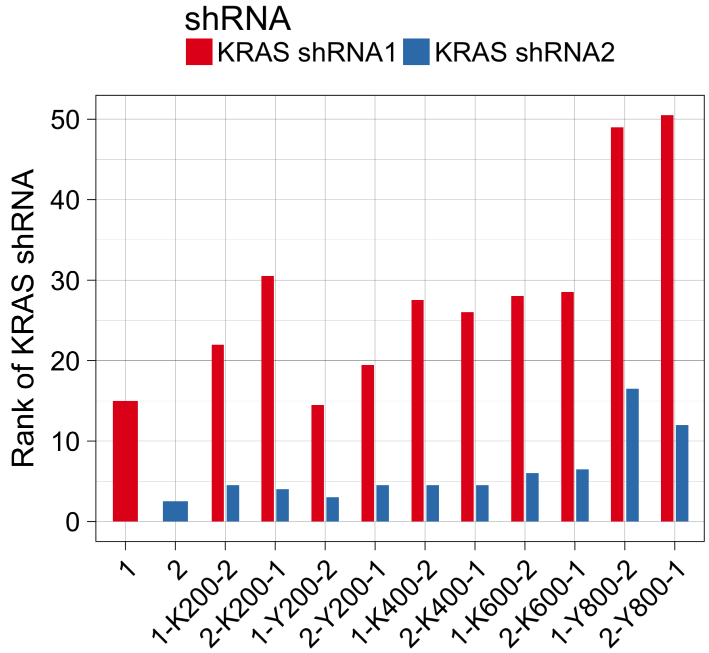
With optimized dual shRNA expression vector in hand, the first gene set we interrogated was comprised of all pair-wise combination of druggable genes over-expressed in four melanoma cell line compared to melanocytes. For each pair of genes, two different pairs of shRNA were included in two libraries to minimize false-positive and false-negative effects. Each library was screened in parallel in the A-375 cell line. From the screen, ∼300 deleterious gene-interactions consistently observed in both libraries were identified. To validate these results, I performed an orthogonal CRISPR/Cas9 screen using pCRoatan, a lentiviral vector expressing pairs of sgRNAs that I and others were developing for another project (chapter 3). Some inconsistencies were observed between the CRISPR screen and the RNAi screens, especially for some constructs that consistently depleted with shRNAs but not with sgRNA. This can perhaps be explained by the different molecular mechanisms used to manipulate gene expression. Knocking down a gene with shRNAs leads to a widespread reduction of the target mRNA to similar levels across all cells. However, the targeted mRNA is generally not removed entirely and some level of expression can remain. In contrast, CRISPR experiments using fluorescent reporters to track knockout shows that knockout effiency varies cell to cell, perhaps depending on Cas9 level of expression within each cell, although this can be somewhat mitigated by generating clonal cell lines expressing Cas9. Double-strand breaks generated by Cas9 are repaired by the error-prone non-homologous end-joining pathway which leads to the introduction of indels at target site. These indels can induce both frame-shift as well as in-frame repairs with varying efficiency depending on the sgRNA used as a guide. A pool of cells with a Cas9 induced knock-out is thus a mosaic of cells each harboring a different genomic scar at the targeted locus, which can lead to a variety of phenotypic outcome. This is particularly true when Cas9 is directed to two genes simultaneously, which can lead to four different repair events in each cell. Improvements in sgRNA design is likely to further improve such combinatorial CRISPR screen by increasing editing efficacy. Additionally, the potency of shRNAs used in the RNAi screen had been experimentally tested using a sensor assay. Perhaps using similar high-throughput assays to select sgRNAs will improve the consistency of CRISPR screens.

One of the advantages of using CRISPR technology rather than RNAi in combinatorial experiments is that guides have a single variable region of 20bp which allows multiple sgRNA to be printed on the same DNA chip. This allows great flexibility in library design and constructs can be barcoded with known sequences which eliminates the intermediary cloning step required to barcode pairs of shRNA. Although this is technically feasible for two hairpins, sequencing of many shRNA constructs shows that when shRNAs are cloned from chips, a large number of constructs have errors when Sanger-sequenced. This can be due to the complex structure hairpins which makes DNA synthesis less reliable or the Gibson assembly cloning. The intermediary cloning step is thus necessary to generated dual shRNA libraries as it selects for error-free construct thus limiting potential false-negative effects of having mutations in the hairpin sequences.

Combining the results of both RNAi and CRISPR screens, we focused of pairs of genes that showed synergistic deleterious effect on cell proliferation. Although all selected pairs pass stringent criteria in terms of number of significantly depleted construct per pair, depletion rates, and increased lethality of combination when compared to individual targeting of both genes, further experiments need to be performed to validate and characterize these hits. As a first step, I am performing one-by-one dual knockdown experiment to eliminate hits that were an artifact of the large scale screen that can be noisy given the total number of constructs considered. In addition, some pairs can already be targeted by small molecule inhibitors, which provides a third orthogonal validation strategy. Although four cell lines had initially been selected, the screens have only been performed on A-375, and lethal pairs should be tested on a wider range of cell lines.

The tools developed in the project will allow for the high-throughput discovery of targets to be used in combinatorial drug therapies. This approach focuses on interfering with entire molecular networks and will provide insights on critical pathway nodes that could be targeted in combination to overcome resistance mechanisms that are observed in single target treatments.

2.5 Supplementary data



SUPPLEMENTARY FIGURE 2.1: **Abundant mature miRNA guides are produced from pairs of shRNAs expressed from the same promoter** Constructs harboring a pair of KRAS targeting hairpins separated by a spacer of various length were transduced in A-375 and ERC cells. Following selection, small RNA cloning was performed. miRNA guides were ranked by read counts. Shown is this rank for mature KRAS shRNA guides, averaged for both cell lines.

Chapter 3

A CRISPR resource for individual, combinatorial, or multiplexed gene knockout

Since this project was performed in collaboration with Dr. Simon Knott, I will describe our individual contributions. Dr. Knott designed the machine learning algorithm and integrated the different parameters that make up the final CRoatan libraries (figures 3.1, 3.2, 3.3, 3.4, 3.7). I designed and validated the dual expression vector, built the genome wide libraries and performed all experiments. The results of this work were published in Erard et al (2017) Mol Cell (See Appendix B). For the purpose of this thesis, I analyzed the data to generate all the figures in this chapter except the ones cited above.

3.1 Introduction

Genetic screens have proven to be invaluable tools to investigate gene function on a genome-wide scale. However, the strength of the conclusions drawn from these experiments greatly depends on the efficacy of the effector used for targeting. To address this issue, multiple algorithms to select potent effectors were designed and experimentally validated for shRNA based gene silencing (Fellmann, Zuber, et al., 2011; Fellmann, Hoffmann, et al., 2013; S.R. Knott et al., 2014). Concurrently, novel shRNA backbones and expression strategies were devised to increase on-target effect and specificity.

As the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) bacterial system was repurposed to perform targeted gene knock-outs in many organism, similar approaches were taken to identify potent short guide RNAs (sgRNAs). For both several Type II and Type V CRISPR systems, large sgRNA potency datasets were generated and used to train algorithms to predict the efficacy of an sgRNA based on its sequence characteristics (Doench, Hartenian, et al., 2014; Chari et al., 2015). Unlike RNAi that mediates mRNA decay, Cas9 induces double-strand breaks at the target site. The phenotypic outcome of targeting a locus with CRISPR thus depends not only on the efficacy of the gRNA at guiding the Cas9 to the locus, but also on the characteristics of the genomic scar left by the repair of the DSB. To maximize the likelihood of generating functional knockouts, initial studies focused on selecting target sites in the first genic exons to increase the likelihood of an indel producing premature stop codons leading to the elimination of transcripts by the nonsense-mediated decay pathway (Doench, Hartenian, et al., 2014). Others focused on targeting protein functional domains in which the loss of a few amino-acids would be deleterious even if the DSB was repaired in frame (Shi et al., 2015).

CRISPR induced DSB were thought to be repaired predominantly by the non-homologous end-joining (NHEJ) pathway, making any prediction on the outcome of the repair impossible. However, sequencing of large numbers of these genomic scars has shown that the repair can be guided by micro-homologies between DNA end when the DSB site is flanked by short homologous stretches (Bae et al., 2014). In this case, the likelihood and characteristics of each repair resolution can be predicted using the length and GC content of the homologous sequences, as well as their distance from the DSB break. Using these predictions, target sites for which the repair will lead to a frame-shift mutation can be selected.

To increase the efficacy of CRISPR reagents, different sgRNA expression strategies have been explored. Multiple independent sgRNAs have been targeted to genes by using crRNA arrays mimicking closely the prokaryotic gRNA maturation pathway, both using Cas9 with an RNAse III and with CpfI (Cong et al., 2013; Zetsche, Heidenreich, et al., 2016). Alternatively, vectors harboring sgRNAs driven from multiple U6 promoters can be designed (Vidigal & Ventura, 2015). Although both of these strategies have mostly been used to either perform combinatorial screens or to ablate long non-coding sequences, targeting multiple sgRNA to the same gene should in principle lead to a greater likelihood of functional knockout.

Here, we combined the strategies identifying highly active sgRNA sequences and predicting the impact of a DSB repair on target protein to select putative target sites based on sequence, amino-acid conservation at target site, and frame-shit mutation likelihood. Additionally, we designed a novel vector expressing two sgRNAs, and developed an algorithm to select pairs of sgRNAs that would have synergistic deleterious effects when simultaneously targeting the same gene. To validate our approach experimentally and compare it to existing sgRNA selection algorithms, we performed five multiplexed loss-of-function screens in two cell lines.

3.2 Materials and methods

3.2.1 Random Forest Training and Scoring

10 random forests were constructed for each of the Doench et al and Chari et al datasets. For each data type, sgRNAs in the top 75th and bottom 50th percentile for each gene were classified as potent and weak, respectively. The 10 forests were trained using the Matlab treeBagger package (1000 trees per forest). Forests were trained using incrementally increasing penalties for false-positive classifications (1.2, 1.4, 1.6, 1.8, 2, 2.2, 2.4, 2.6, 2.8, 3). During training forests are constructed using the 28 overlapping 3mers of each target as features, and the class of the target (potent or weak) as the output.

When a new target is being scored, it is decomposed into 28 3mers, and these are given to each of the 20 forests (10 corresponding to the Doench et al data and 10 to the Chari et al. data) as input. The target is then assigned a value between 0 and 10 corresponding to the highest stringency forest it was assigned as potent by. For example, if a target was called potent by a Doench forest that was trained with a penalty of 2.2 (6th lowest) and a Chari forest trained with a penalty of 2 (5th lowest), the target would receive a score of 5.

3.2.2 sgRNA-pair Scoring

For each gene, all pairwise scores were calculated for the top 10 CRoatan scoring sgRNAs. All sgRNA pairs begin with a score of 0. Overlapping pairs are assigned a final score of 0. Pairs that are less than 10kb apart with DSB-DSB distances that are not divisible by 3 are assigned a score of 2.5 if they target the same transcripts. Scores are incremented by 1 if pairs have imbalanced CRoatan scores (one less than 7 and one greater than 7). This scoring matrix is then given as input to the maximum weighted matching algorithm (matlab maxWeightMatching).

3.2.3 Cell lines

CRISPR/Cas9 screens were performed in melanoma A-375 (ATCC CRL-1619, female) and chronic myelogenous leukemia K-562 (ATCC CCL-243, female) cell lines. A-375 were grown at 37°C in Dulbecco's Modified Eagle Medium (DMEM), supplemented with 10% Fetal Bovine Serum and 50U/mL penicillin/streptomycin. K-562 were grown at 37°C in RPMI1640 supplemented with 10% FBS and 50U/mL penicillin/streptomycin. The 293FT cell line (Thermo-Fischer) was grown at 37°C in DMEM supplemented with 10% FBS and penicillin/streptomycin.

A-375 cells were infected at low multiplicity of infection (MOI) by virus produced using lentiCas9-Blast (Addgene # 52962) (Sanjana et al., 2014) and selected using blasticidin (10μg/mL). Following 10 days of selection, single cells were sorted using the FACSAria IIU cell sorter (BD Biosciences) into 96-well plates. 10 A-375-Cas9 clones were tested for Cas9 functionality by infection with a vector expressing ZsGreen and an sgRNA targeting ZsGreen. Knockout efficiency was estimated by flow cytometry after 14 days. One of the A-375-Cas9 clonal lines exhibiting more than 50% knockout of ZsGreen in this assay was selected for further experiments. The K-562 clonal cell line expressing Cas9 was kindly gifted by Dr. Vakoc (Cold Spring Harbor Laboratory).

3.2.4 sgRNA Library Construction

For single sgRNA libraries, sgRNA sequences were predicted using existing algorithms (RNA-Configurator, sgRNAScorer, GPP web portal and CRoatan) and oligonucleotides containing these sequences were ordered from Integrated DNA Technologies (IDT, Table S1). These molecules were amplified by PCR (forward primer (FP): TTACCGTAACTTGAAAGTATTTCGATTTCTTG-GCTTTATATATCTTGTGGAAAGGACGAAACACCG, reverse primer (RP): GGACTAGCCT-TATTTTAACTTGCTATTTCTAGCTCTAAAAC) and cloned by Gibson assembly into a 3rd generation lentiviral vector harboring a U6 promoter, an sgRNA backbone, and a ZsGreen-P2A-PuromycrinR transcript driven by a spleen focus-forming virus promoter (pCRoatan-singleSgRNA).

For dual sgRNA libraries, sgRNA sequences were predicted using CRoatan. Primers containing these sequences were ordered from IDT (Table S1) and used to amplify a hU6-EM7-ZeocinR-cU6 cassette (pCRoatan-dualPromoter). The amplicon was digested with BbsI (NEB) and ligated into a 3rd generation lentiviral vector (pCRoatan-dualSgRNA) previously digested with BsmBI (ThermoFischer).

Combinatorial sgRNA libraries were built using DNA chips (CustomArray, Inc.) containing 10K molecules harboring a barcode and two flanking sgRNA sequences (Table S2). Chips were amplified by 5 separate 18-cycle PCRs to ensure high-complexity end product. The amplicons were first cloned by ligation into an intermediate cloning vector (pCR-BluntII TOPO based) using SpeI (NEB) and ApaI (NEB). Subsequently, the hU6 and cU6 promoters driving the sgRNAs were added to the vector. The hU6 promoter was amplified from lentiCrisprv2 (Addgene # 52961) by PCR (FP: AGTACCGTCTCTGGTGTTTCGTCCTTTCCA-CAAG, RP: GTACCTACGCGTGAGGGCCTATTTCCCATGATTC), and cloned by ligation using the BsmBI (ThermoFischer) and MluI (NEB) restriction sites. The cU6 promoter (cU6-3, Kudo et al., 2005) was amplified from a gBlock (IDT) by PCR (FP: ATCGATCTCGAGGCGC-CGCCGCTCCTTCAGGCA, RP: TGATCCTGGTCTCACGACTAAGAGCATCGAGACTGC), and cloned by ligation using the BsaI (NEB) and XhoI (NEB) restriction sites. Following these three steps, the full sgRNA1-hU6-EM7-ZeocinR-Barcode-cU6-sgRNA2 cassette was digested from the intermediate cloning vector using BbsI and ligated in the lentiviral expression vector (pCRoatan-dualSgRNA) as described previously. All transformations were performed with Invitrogen's MegaX DH10B T1 electro-competent cells using a Bio-Rad Gene Pulser Xcell and Bio-Rad Gene Pulser 1 mm cuvettes for electroporation. For each library, a minimum of 10 million successfully transformed cells were obtained.

3.2.5 sgRNA Library Screening

sgRNA libraries were packaged using the 293FT cell line (Thermo Fischer). Cells were co-transfected with library vector (60μg), pMDL (12.5μg), CMV-Rev (6.5μg) and VSV-G (9μg) by calcium phosphate transfection. The media was replaced at 14h and virus was collected at 36h and filtered using a 0.45μM syringe filter (Millex-HV, EMD Millipore). Viral infections were performed at an MOI of 0.3 to ensure a maximum of one sgRNA integration per cell. sgRNA representation in the infected population was maintained at a minimum of 1000 infected cells per sgRNA at each passage. All screens were performed in triplicates. Two days after infection, cells were collected for a reference time point. After 12 doublings, cells were harvested for a final time point. Infected cells were selected using Puromycin (1μg/mL) after the initial time point and throughout the screen.

3.2.6 CRISPR/Cas9 Library processing and analysis

Following cell harvests, DNA was extracted using the QIAGEN QIAamp DNA Blood Midi kit. For each sample, sgRNA molecules or barcodes identifying sgRNA pairs were extracted from the genomic DNA in 24 separate 30-cycle PCR reactions in which 2μg of DNA input was included. Illumina adapters were included in the PCR primers (Table S3). Libraries were sequenced using custom read one primers on the Illumina MiSeq or HiSeq platforms. Following sequencing, reads were trimmed to a length of 20bp and sgRNA counts were extracted using the bowtie algorithm (Langmead et al., 2009). For each sgRNA or sgRNA pairs, log fold change values were calculated by dividing the abundance after twelve doublings by the abundance at the reference timepoint, two days after infection (Knott et al., 2014).

3.2.7 Dual-sgRNA genomic scar analysis

200,000 A-375-Cas9 and K-562-Cas9 cells were transduced with CRoatan constructs targeting 3 different olfactory receptor genes. Following selection with Puromycin cells were grown for 12 doublings and then harvested for analysis. DNA was extracted using the QIAGEN QIAamp DNA Blood Midi kit. The target region, including 50bp upstream and downstream of both sgRNA target sites was amplified by PCR, in 16 25-cycle PCR reactions in which 500ng of

DNA input was included (Table S3). Following purification using the QIAquick PCR Purification Kit, Illumina adapters were added via PCR and samples were processed on the Illumina MiSeq platform using paired-end reads of 200bp to cover both sgRNA target sites. Reads were mapped to the relevant genomic region using the bwa mem algorithm and cut types were analyzed and counted using the CIGAR string of the alignment (Li et al., 2009).

3.3 Results

3.3.1 sgRNA selection strategy

A random-forest algorithm to predict sgRNA efficacy

Predicting sgRNA efficacy from sequence determinants *in silico* requires large training datasets of paired sequence-efficacy datapoints. Such datasets did not exists when this project started and our first approach was to adapt the shRNA sensor assay that had been developed by the Hannon lab and others to CRISPR/Cas9 (Fellmann, Zuber, et al., 2011). This assay relies on expressing in the same cell an shRNA (under control of an inducible promoter) and a fluorescent protein flanked by the target site of the shRNA. The efficacy of the shRNA can be assessed by measuring the drop in fluorescence when the induced. Our strategy was thus to pair both sgRNA and target site on the same lentiviral vector, transduce Cas9 expressing cells and estimate sgRNA efficacy by deep sequencing mutated target sites. However, two studies using sgRNA efficacy datasets to build algorithms to predict sgRNA strength were published shortly after and we decided to use the datasets they had generated. Doench et al. used a library of 3000 sgRNAs tiling nine mouse and human cell surface proteins (Doench, Hartenian, et al., 2014). Following delivery of the sgRNAs, target-negative cells were FACS-sorted. The relative abundance of each sgRNA in the target-negative population compared to the unsorted population can be used to measure effector potency. Chari et al. used a library-on-library approach by first stably transducing cells with a synthesized library of 1400 target sites (Chari et al., 2015). They then transfected a library of the corresponding sgRNAs and measured for each sgRNA CRISPR-induced indels by high-throughput sequencing of target sites. One caveat of this experimental design is that each cell receives a large number of sgRNA-expressing plasmid, leading to high sgRNA concentration in each cell as opposed to when a single U6-sgRNA cassette is integrated in the genome. Additionally, large number of these sgRNAs are likely to be non-targeting but will compete for binding to Cas9 proteins reducing the number of targeting sgRNA-Cas9 complexes in the cell.

Using these two datasets, we developed a random forest-based sgRNA prediction algorithms. All 3-mers of the sgRNA binding site, with an additional 4 base pairs upstream and 6 base pairs downstream (this includes the PAM) were used as features. For each dataset, 10 different random-forests were train to discriminate weak from potent sgRNAs using increasing penalties for false positive predictions. The random-forests are thus increasingly stringent at identifying potent sgRNAs. When predicting the potency of new sgRNA sequences, the highest stringency level they pass on both sets of random-forests is used as a score. To validate this scoring system, we used a subset of sgRNAs from the Doench et al. dataset that were withheld during training. The percent-ranks of sgRNAs in the *in vivo* assay were compared to the scores assigned by the prediction algorithm (figure 3.1).

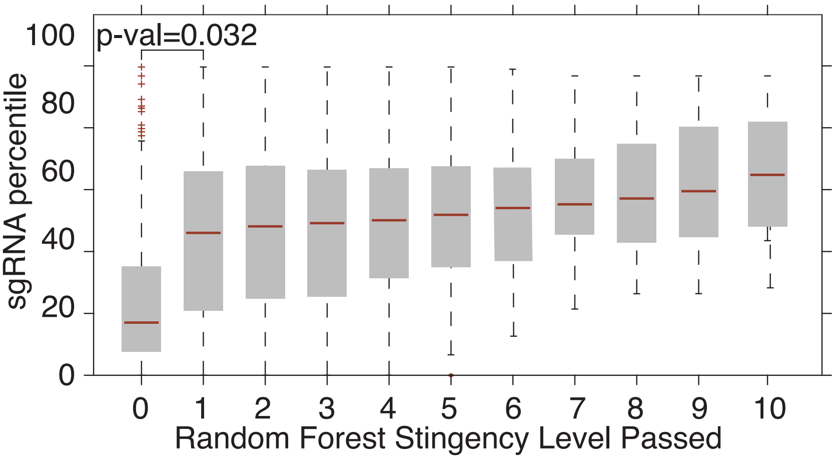


FIGURE 3.1: **A random forest-based algorithm to predict sgRNA potency** Activity of the sgRNAs from the Doench et al. dataset, stratified by the random forest stringency level passed. For all the boxplots in this chapter the the edges of the box are the 25th and 75th percentiles. The error bars extend to the values q3 + w(q3 - q1) and q1 - w(q3 - q1), where w is 1.5 and q1 and q3 are the 25th and 75th percentiles. sgRNAs passing the first stringency level are significantly more active (rank-sum test, p-val=0.032). The sgRNA percent-rank was calculated by ranking sgRNAs targeting the same gene.

A significant gain in activity was observed for sgRNAs passing the minimum stringency threshold (rank sum test, p-val = 0.032). Higher scores correlate with an increase in sgRNA potency, although this increase is not statistically significant.

Amino-acid conservation based target site selection

Unlike with shRNAs, where targeting leads to mRNA cleavage, the phenotypic consequence of targeting a locus with an sgRNA depends on the genomic scar left by the DSB repair. CRISPR/Cas9 induced mutations can be analyzed by high-throughput sequencing and have been shown to be mostly deletions of varying length, which can cause frame-shifts if the length is not a multiple of 3. If not, only a few amino-acids are deleted from the original protein. To maximize the impact of such a deletion, previous studies have targeted Cas9 to known protein functional domain (Shi et al., 2015). As most proteins lack verified domain annotations, this strategy cannot be easily implemented to build genome-scale sgRNA libraries. An alternative is to predict the effect of an indel or substitution of any amino-acid in silico using SIFT, PolyPhen-2 or PROVEAN (P. Kumar, Henikoff, & Ng, 2009; Adzhubei et al., 2010; Choi et al., 2012). Here, we used scores generated using PROVEAN (Protein Variation Effect Analyzer) as it performs similarly to other software in specificity and sensitivity, and scores for any amino-acid single substitution or deletion are available for all human proteins from Ensembl 66. For any given protein sequence, PROVEAN uses BLASTP to create a set of other know proteins of any species with a sequence identity over 80%. The predicted effect of an amino-acid variation is then predicted based on the change of similarity between the protein and the selected set before and after the mutation. This score is *in fine* based on the conservation of amino-acid residues and we used it as a metric to rank sgRNA target sites according to the predicted deleterious effect of an amino-acid deletion. If the predicted DSB break was within a codon, we used the PROVEAN score for a deletion of that amino-acid, if not, we averaged the scores for the deletion of the flanking amino-acids (figure 3.2A).

An analysis of the Doench et al. datasets shows that these scores are correlated with the measured functional knock-out rate (figure 3.2B). We thus considered the PROVEAN score of neighbouring amino-acids when selecting target sites to maximize the deleterious effect of an indel on protein function.

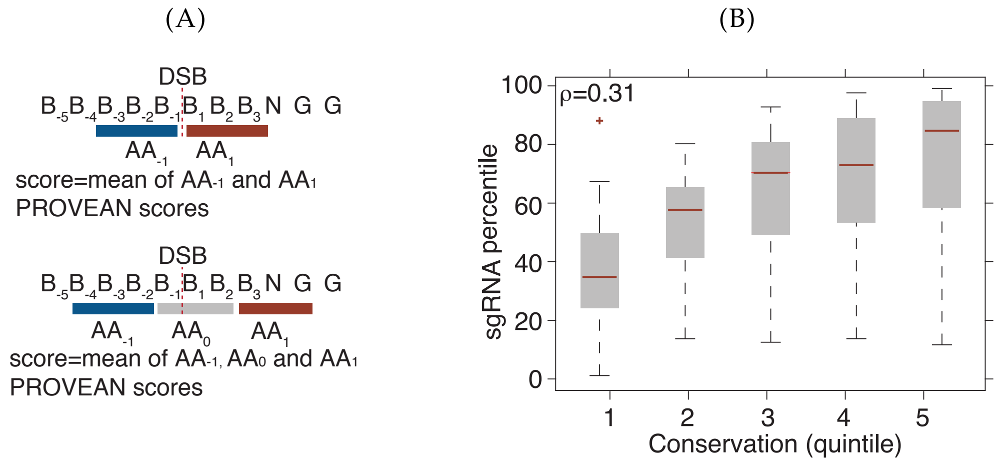


FIGURE 3.2: **Selection of target sites using amino-acid conservation (A)** Strategy used to assign conservation scores target sites using PROVEAN scores from neighboring amino-acids. **(B)** Efficacy of sgRNAs from the Doench et al. dataset stratified by conservation score (ρ=0.317).

Predicting frame-shift mutations inducing repairs

CRISPR/Cas9 DSB were thought to be predominantly repair by the NHEJ pathway, which makes predicting the resulting genomic scar impossible. However, recent high-throughput characterization of large numbers scars left by CRISPR/Cas9 show that some repairs are mediated by microhomology-mediated end joining (MMEJ) (Bae et al., 2014). When the DSB target site is flanked by small homologous regions, the repair can be predicted (figure 3.3A). We reasoned that targeting sites with a high likelihood of an MMEJ mediated repair leading to a frame-shift mutation would increase the rate of functional knockout. Similarly to Bae el al., we developed a linear regression model to predict the likelihood of a micro-homology guided repair resolution based on the length, GC content and distance to the DSB of the corresponding homologous region. For target sites with a MMEJ likelihood above the median of all target sites, the likelihood of a FSM was measured as the fraction of predicted resolutions corresponding to FSMs. An analysis of the Doench datasets shows that efficacy is increased when target sites with an FSM score greater than 66% are chosen (figure 3.3B). In addition to conservation scores, FSM-likelihood of targeted locus can also be used as a criteria for sgRNA selection.

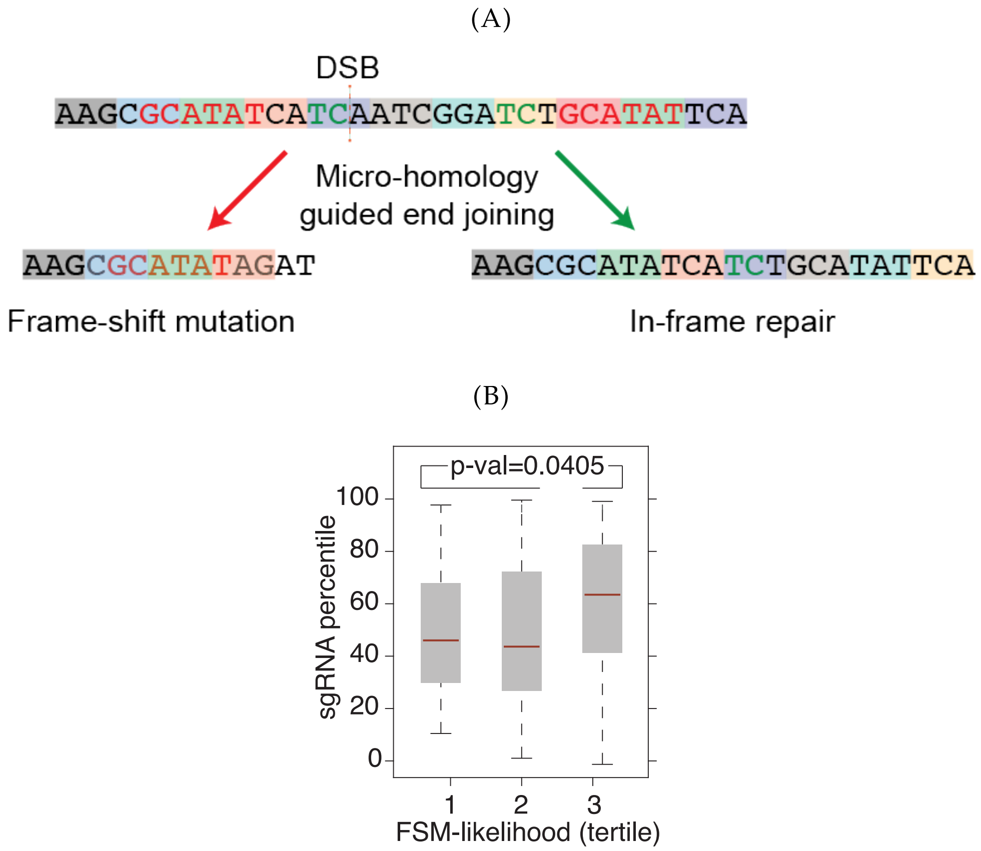


FIGURE 3.3: **Selction of target sites based on frame-shift mutation likelihood upon DSB repair** **(A)** Cas9-induced double-strand breaks can be repaired by micro-homology mediated end joining, the proportion frame-shift mutations or in-frame repairs can be predicted based on the length, GC content and distance to the double-strand break of the corresponding homologous region **(B)** Potency of sgRNAs from the Doench et al. dataset stratified by frame-shit mutation likelihood. (rank-sum test p-val=0.0405 for tertile 3 sgRNAs compared to tertile 1 and 2 sgRNAs).

CRoatan: an algorithm to select potent sgRNAs

sgRNA potency, conservation of amino-acids at target site and predicted FSM scores can all be used to select sgRNAs generating functional knock-outs efficiently. To consolidate these three components in a single predictive algorithm, we first separated the sgRNAs into three groups based on their random-forest scores: group A, with a score of 0, group B with a score of 1 to 5 and group C with a score of 6 to 10. These groups received a baseline score of 0, 3 and 6 respectively. Within each group, sgRNAs were ranked according to their conservation and FSM rate scores. For conservation, the target site was deemed conserved if its score was higher than the median score for all Cas9 sites of human CDS. For FSM rate, the threshold was set to a FSM score greater than 66%. For each sgRNA, an additional score of 1, 2 and 3 was added to the baseline score if neither of the FSM or conservation threshold were passed, 1 if one of these threshold was passed and 2 if both were passed. sgRNAs in groups A, B or C were thus respectively assigned scores of 1 to 3, 4 to 6 or 7 to 9 (figure 3.4A). Using these algorithm, which we called CRoatan, we ranked the sgRNAs in the Doench et al datasets. CRoatan scores correlate well with the observed experimental efficacy (figure 3.4B). We then used CRoatan to predict the most ten most potent sgRNA for all protein-coding genes in human refseq annotation and identified potent sgRNAs for each target.

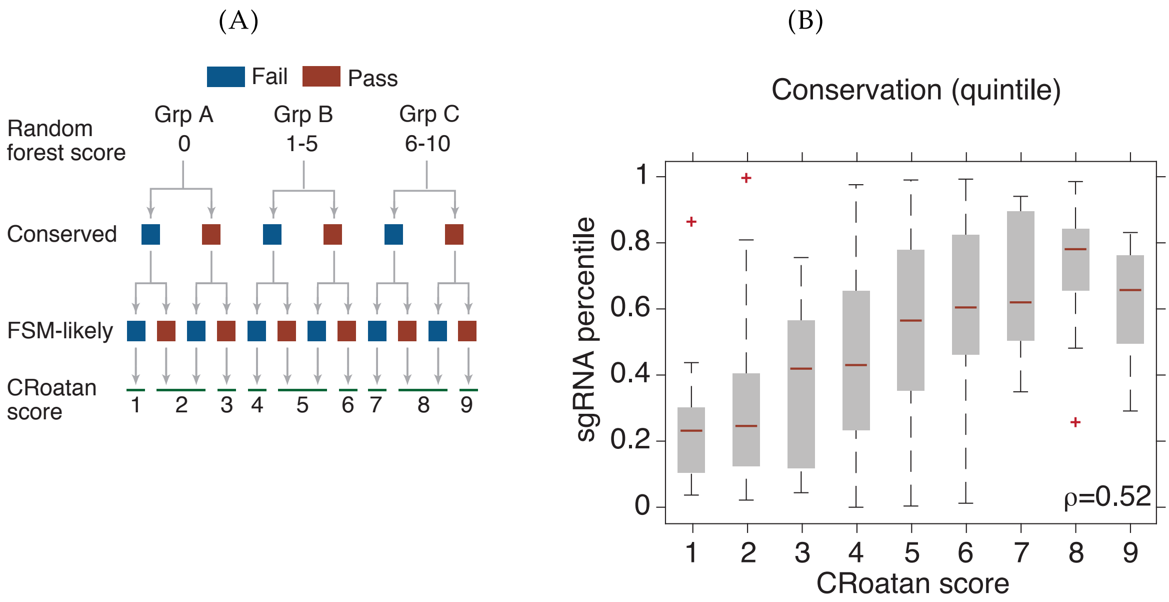


FIGURE 3.4: **CRoatan: an algorithm identifying potent sgRNAs** **(A)** Diagram of the strategy used to combine random forest,conservation and frame-shift mutation likelihood scores in a single CRoatan score. **(B)** Re-analysis of the depletion scores of the sgRNAs in the Doench et al datasets, ordered by their CRoatan score (ρ=0.562).

Experimental validation

To compare CRoatan to other existing sgRNA selection algorithms, we built 4 different CRISPR libraries and performed loss-of-function screens. Each library was composed of 20 essential genes and 20 non-essential genes, with 5 sgRNA per gene. Essential genes were chosen based on previous RNAi screens in cancer cell lines and olfactory receptor genes were chose as non-essential controls. A different prediction software was used to select sgRNAs for each library: sgRNA Scorer, Gene Perturbation Platform Web Portal (GPP WP), RNA-Configurator (Dharmacon) and CRoatan. Each library was independently cloned in a lentiviral vector harboring a human U6 promoter to express the sgRNA and a spleen-focused forming virus (SFFV) promoter driving the expression of ZsGreen, a fluorescent marker, and a puromycin resistance marker in a bi-cistronic fashion. The libraries were packaged in lentiviruses and transduced into the melanoma A375 and the chronic myelogenous leukemia K562 cell line. Both cell lines expressed Cas9 constitutively. Following transduction, cells were selected and cultured for around 12 doublings. The depletion of each construct was measured by calculating log-ratios of abundance after 12 doublings to abundance 2 days after infection. To compare each library, these log-ratios were normalized using the log-ratio of the sgRNAs targeting the non-essential genes. Analysis of the depletion of constructs targeting essential genes in the CRoatan library shows that the observed depletion correlates with CRoatan scores (figure 3.5). However, depletion rates did not correlate with conservation score and FSM mutation alone.

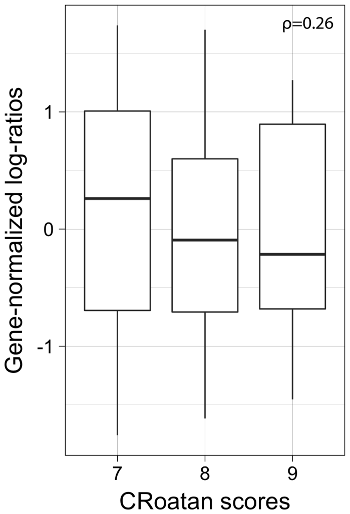


FIGURE 3.5: **Validation of CRoatan using a CRISPR/Cas9 depletion screen** Gene by gene Z-score normalized depletion rates of sgRNAs targeting essential genes, stratified by CRoatan score (ρ=0.26). Depletion rates were calculated as the average of the log-ratios in the screens conducted in the A375 and K562 cell lines.

We then compared the depletion of sgRNAs targeting essential genes in the four libraries. CRoatan constructs were found to be significantly more depleted than RNA-Configurator and sgRNAScorer constructs (rank-sum pval of 0.042 and 0.026 respectively, figure 3.6). Although CRoatan constructs were on average more depleted than GPP constructs, this difference was not statistically significant (rank-sum pval=0.19). Overall, this demonstrates the effectiveness of CRoatan as an sgRNA selection algorithm (figure 3.6).

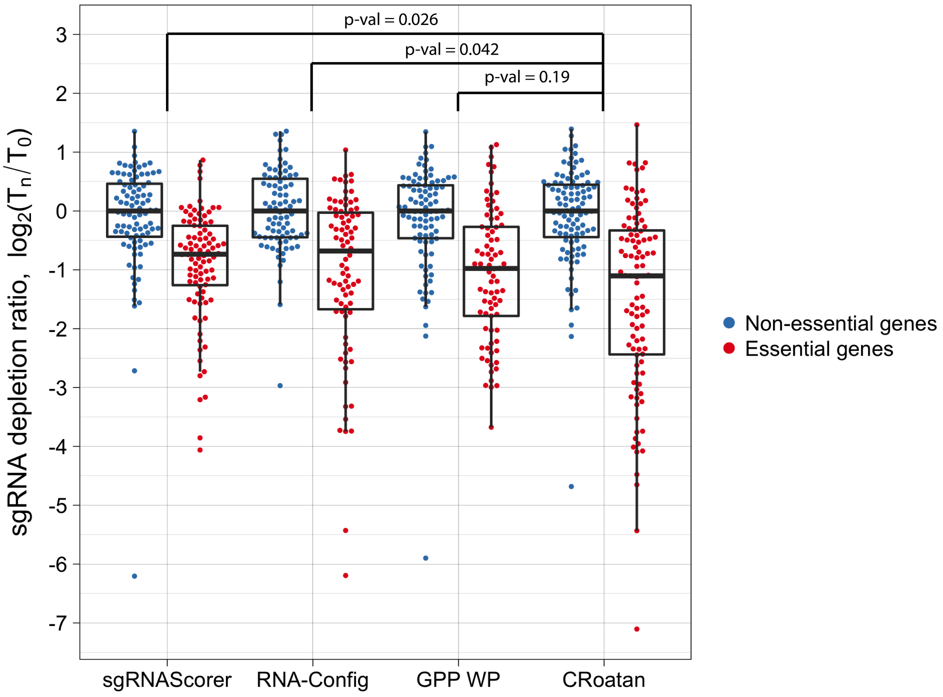


FIGURE 3.6: **Comparaison of CRoatan to other sgRNA selection tools** Libraries of sgRNAs targeting essential or non-essential genes were screened in A375 and K562. Shown are the depletion log-ratios of sgRNAs targeting essential and non-essential genes, averaged for both screens. Libraries were designed using sgRNAScorer, RNA-Configurator, GPP WP or CRoatan. sgRNAs targeting essential genes are more active in the CRoatan library

3.3.2 Dual-gRNA expression strategy

Vector design

To further increase the efficacy of CRoatan constructs, we sought to target each gene with multiple sgRNAs simultaneously, and designed a vector expressing pairs of sgRNA. This vector is based on the one used in the screens above (figure 3.7). Each sgRNA is expressed from an independent U6 promoter. To lower the risk of recombination, the first sgRNA is driven by a human U6 promoter and the second from a chicken U6 promoter that had previously been used in RNAi experiments to express first-generation shRNAs. We placed these promoters in opposite orientation to prevent transcriptional interference. We also added a barcode identifying each pair uniquely, flanked by Illumina sequencing adapter sequences to facilitate sequencing of high-throughput screening experiments. The small size of sgRNA variable sequence (20bp) allows for printing both sgRNAs and the barcode on DNA chips on the same oligo. Complex libraries targeting one or two genes can be easily assembled by first cloning the chip in a lentiviral vector harboring two sgRNA barcodes and sequentially adding the hU6 and cU6 promoters. To create combinatorial libraries from existing construct, the second sgRNA cassette can be randomly shuffled. The two sgRNA variable region are too far apart to sequence them simultaneously by deep sequencing, but the barcode and second sgRNA sequence can be used to uniquely identify each pair.

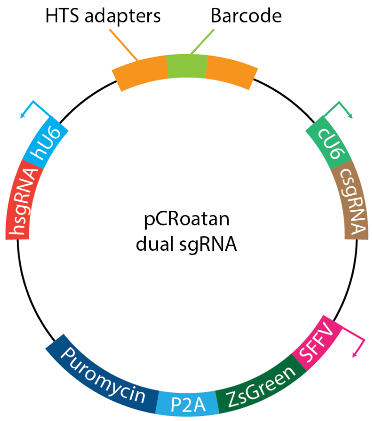


FIGURE 3.7: **A dual-sgRNA expression vector** Schematic map of the lentiviral vector used to express pairs of sgRNAs (hU6: human U6 promoter, cU6: chicken U6 promoter, HTS: high-throughput sequencing, SFFV: spleen focus-forming virus promoter).

Dual-gRNA induced genomic scars

Targeting CRISPR to two adjacent sites can theoretically allow for the deletion of large fragments of genes which would be effective to disrupt gene function. To quantify the likelihood of such events as well as the mutation rates generated by this dual-sgRNA vector at both target sites, we generated 3 dual constructs against olfactory receptors. Olfactory receptors were chosen for this test as they generally have only one exon which makes sequencing of genomic scars at two target sites possible using paired-end Illumina sequencing. Following transduction in A375-Cas9 and K562-Cas9 cells, target regions were amplified by PCR and sequenced. We then classified the observed genomic scars in five types: no mutation, hU6-sgRNA or cU6-sgRNA indel if only one of the target site was mutated, hU6-sgRNA and cU6-sgRNA indel if both sites were mutated simultaneously and fragment deletion when the sequence between the two sgRNA target site was entirely removed. For all three sgRNA pairs, we observed at least 50% mutation rate in both cell lines. Surprisingly, the predominant genomic scar is fragment deletion resulting in the removal of up to 200 base pairs in the loci (figure 3.8). Although this rate of total fragment deletion is likely to decrease when the DSB sites are further apart, targeting pairs of sgRNA to the same exon can result in large deletions that would impact greatly protein functionality.

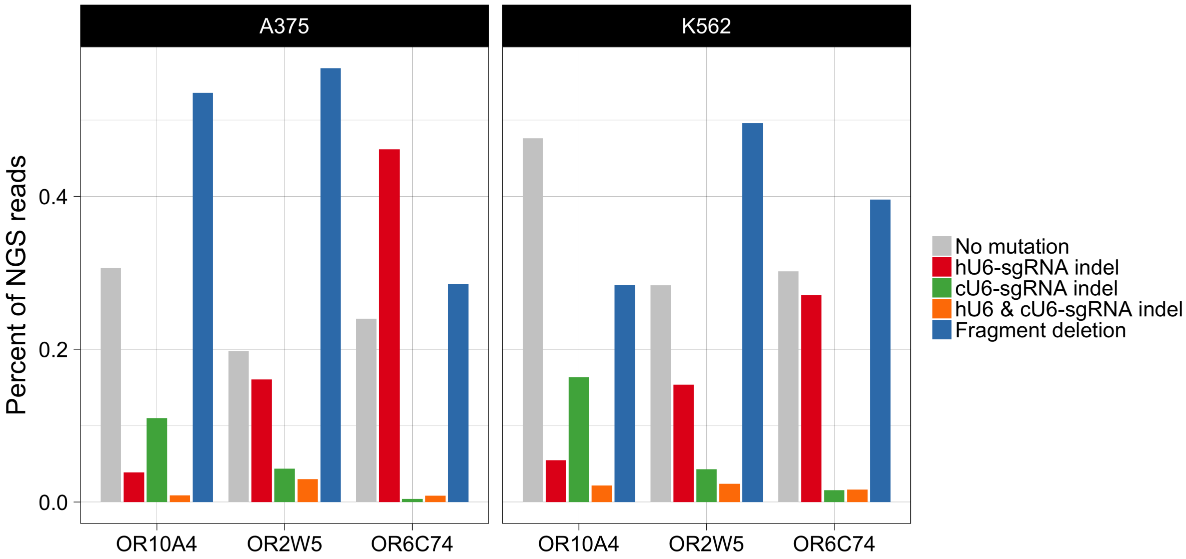


FIGURE 3.8: **Simultaneous targeting of olfactory receptors with multiple sgRNAs** Deep sequencing analysis of deletion patterns induced by targeting pairs of sgRNAs to olfactory receptor genes. hU6-sgRNA or cU6-sgRNA indels are indels where only the site targeted by the hU6 or cU6 driven sgRNA are mutated. hU6 & cU6-sgRNA indels are indels where both sites are mutated, and frament deletions repairs for which the full sequence between the target sites is removed.

A careful analysis of repairs for which only a single site was mutated confirms that a significant fraction of repair-resolutions are MMEJ mediated and can be predicted, validating our strategy to score target site based on the likelihood of MMEJ mediated frame shift repair-resolutions. However, when large fragment deletions are observed, the most common genomic scar is blunt-end joining of the predicted DSB (figure 3.9). Following this observation, we reasoned that increased efficiency could be obtained by selecting pairs for which the number of nucleotides between the predicted DSBs is not divisible by 3. In this case, a large fragment can be deleted and the resulting repair is likely to induce a frame-shift mutation.

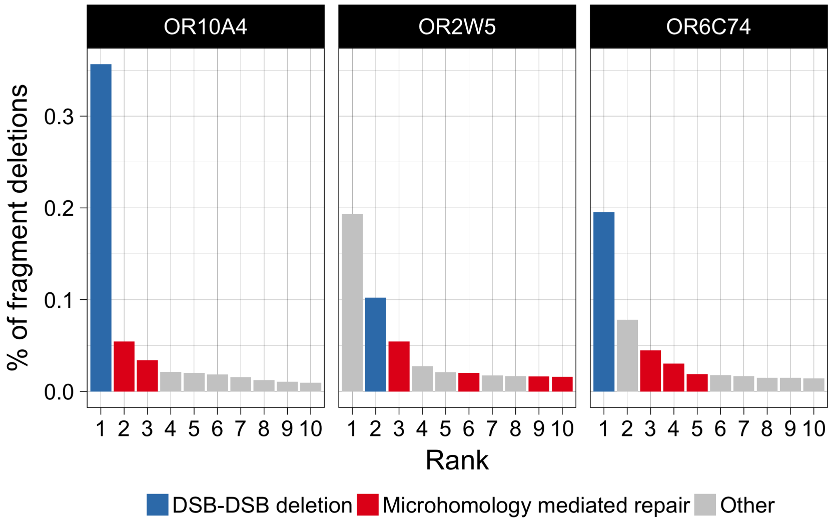


FIGURE 3.9: **Predictable genomic scars upon targeting of sites with pairs of sgRNAs** Analysis of the fragment deletions described in figure 3.8. The rate of occurrence of the top 10 unique deletion pattern is shown, as measured by their average frequency in infected A375 and K562 cells. Deletions corresponding to the precise excision of the fragment ranging from the DSB sites of the two sgRNAs are labeled DSB-DSB deletion. Other deletions for which additional bases are removed are Others. Deletions that could be predicted by examination of small homologous regions upstream of the first target site and downstream of the second target site are also annotated.

sgRNA pairing

To pair sgRNAs for each target gene, we first selected the top 20 sgRNAs as predicted by CRoatan. Sequences mapping to multiple regions of the genomes with mismatches were removed to produce a short-list of 10 potent and specific sgRNAs. A 10x10 pairwise score matrix is then calculated with all possible pairs. To maximize the likelihood of having synergistic deleterious effect within each pair, additional selection criteria were implemented. Pairs in which the target sites are overlapping were removed as this would cause similar effects as using a single sgRNA. For most genes, we were not able to find 10 sgRNAs with the strongest score. In this case, the pairing could create pairs with strong sgRNAs and pairs with weak sgRNAs that would not lead to efficient knockouts. To ensure that each pair harbors a potent sgRNA, the score of unbalanced pairs was incremented. The score of sgRNAs pairs targeting the same exon or two different exons shared by a set of isoforms was also increased to promote pairs where both sgRNAs target each isoform simultaneously. Additionally, target sites for which the predict DSB to DSB distance corresponds to a frame shift mutation as mentioned above were preferred. After scoring of each pair, a weighed graph was built and a maximum matching algorithm was applied to identify the coupling with the highest sum of pair scores (figure 3.10).

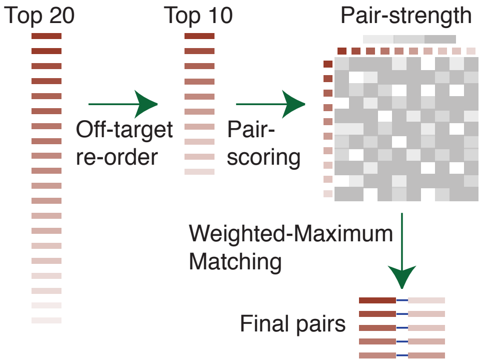


FIGURE 3.10: **sgRNA pairing algorithm used to design five constructs per gene** For each gene, the 20 most potent sgRNAs as predicted by CRoatan are considered. After filtering sgRNAs to reduce off-target effects, all pairs are scored in a 10x10 matrix. Scores are adjusted to increase the likelihood of each pair receiving at least one potent sgRNA. The score of pairs targeting the same exon and for which the DSB to DSB distance is not divisible by three is also incremented. A weighted-maximum matching algorithm is then used to select five pairs per gene.

Combinatorial CRISPR/Cas9 screen

To assess the effectiveness of our dual-sgRNA strategy, we designed a combinatorial CRISPR screen. 100 sgRNAs targeting 10 essential genes and 10 non-essential genes were paired, for a total of 10 000 combinations. This library was screened in A375-Cas9 and the abundance log-ratios of each sgRNA pair were calculated as described above (reference to paragraph). The large number of constructs in this library allowed us to evaluate effectiveness of all components of CRoatan, except the random-forest selection as all selected sgRNAs were in the highest scoring group. Since each sgRNA targeting an essential gene is paired with 10 sgRNAs targeting non-essential gene sgRNAs twice (10 with the essential-gene targeting sgRNA driven by the hU6 promoter, 10 by the cU6 promoter), we measured the efficacy of the sgRNA by averaging the log-ratios of these 10 pairs. This provides a more robust measurement of the strength of each sgRNA than the unique depletion rate we obtained in the CRoatan single-sgRNA screen.

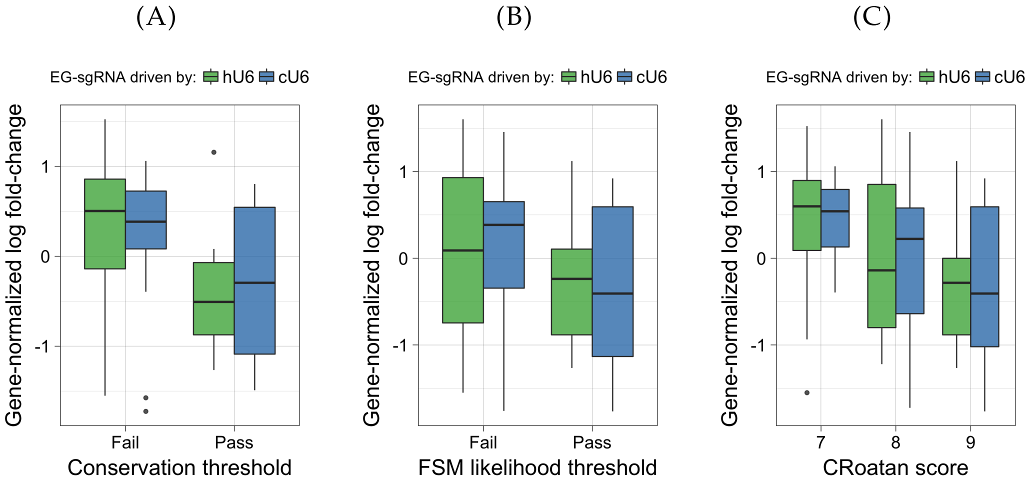


FIGURE 3.11: **Assessment of the strategies used to select potent sgRNAs** Gene-normalized depletion rates of sgRNAs stratified by fail/pass of the conservation threshold **(A)**, of the frame-shift likelihood score threshold **(B)**, or by CRoatan scores **(C)**. The conservation threshold is the median conservation score of all Cas9 target sites in human CDS. For frame-shift mutation likelihood, a score greater than 66% was required to pass.

We calculated the average depletion rate of each essential gene targeting constructs and used z-scores to compare sgRNAs targeting the same gene. When an essential gene sgRNA is paired with a non essential gene sgRNA, it can be expressed from the human U6 promoter or the chicken U6 promoter. To avoid any bias in promoter strength, we separated these two cases. We then separated sgRNA constructs in two groups: pass or fail conservation threshold or frame-shift mutation likelihood threshold. sgRNAs passing the amino-acid conservation or the frame-shift mutation threshold deplete significantly more, regardless of whether the sgRNA is expressed from the hU6 or cU6 promoters (figure 3.11A and 3.11B). This indicates that both of these strategies are useful in predicting potent sgRNAs. Similarly to figure 3.5, we observed a correlation between the predicted CRoatan scores and the *in vivo* activity of sgRNAs (figure 3.11C).

Next, to confirm that using pairs of sgRNAs on the same gene increases the likelihood of functional knockout, we compared the depletion rates of constructs in which an essential-gene sgRNA is paired with a non-essential gene sgRNA to constructs where it is paired with one of the 4 other sgRNAs targeting the same gene. For each of the 10 essential gene targeted in this screen, we compared depletion rates when the gene was targeted by pairs of sgRNAs or with a single sgRNA. For most genes, a stronger phenotype was observed when the gene was targeted by pairs of sgRNAs (figure 3.12A, rank-sum p-val < 0.001).

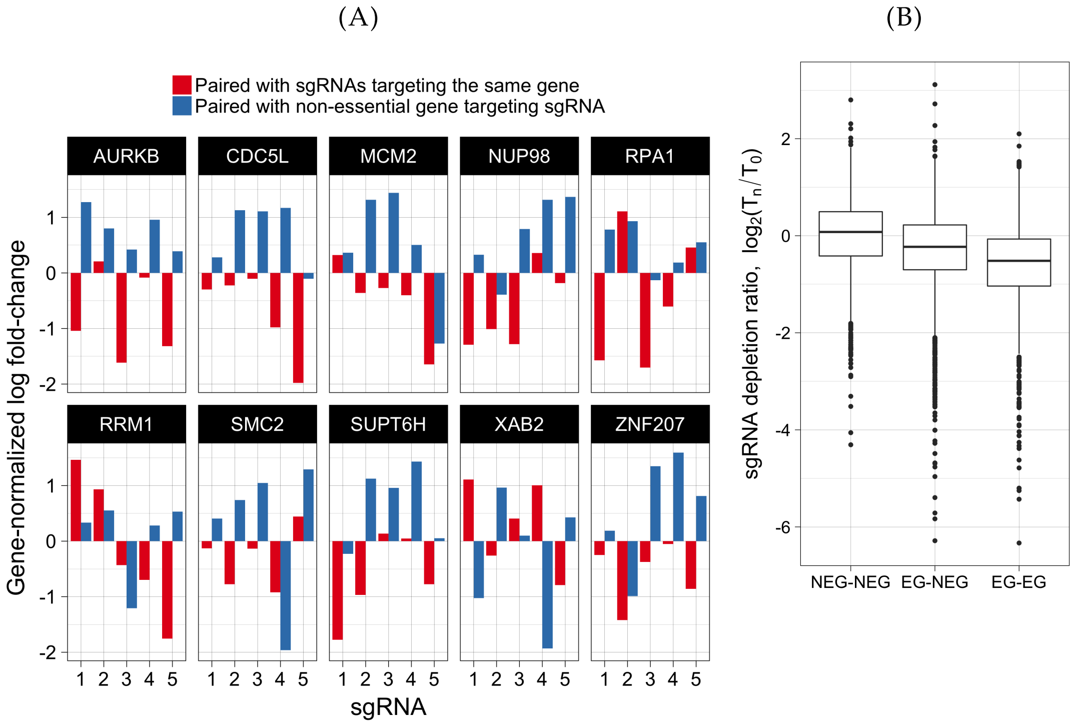


FIGURE 3.12: **Experimental validation of the dual-sgRNA expression constructs** **(A)** Average depletion rates essential gene sgRNAs when they are paird with a non-essential gene sgRNA (blue) or with another sgRNA targeting the same essential gene (red). **(B)** Depletion log-rations of sgRNAs, EG-EG: constructs with pairs of sgRNAs targeting essential gene, NEG-EG: constructs with one sgRNA targeting an essential gene and one a non-essential gene, NEG-NEG: constructs with two sgRNAs targeting non-essential genes.

We extended this comparison to pairs of sgRNAs targeting different essential gene. We found that depletion levels are increased significantly with the number of essential gene sgRNAs in the construct. This combinatorial screen allowed us to estimate the relevance of each component of our computational selection process as well as our expression vector. Both the conservation and the frame-shift mutation likelihood scores were found to positively contribute in the selection of potent sgRNAs. Using pairs of sgRNAs yields stronger phenotypic outcome, here measured as construct depletion rate. Furthermore, this expression vector can be used to knock-out multiple gene simultaneously.

3.3.3 Library validation

As a final validation of our dual-sgRNA library, we designed a dual-sgRNA CRISPR library harboring 200 constructs targeting the 20 essential and 20 non-essential genes that were screened initially (figure 3.6). Both CRoatan and the pairing strategy described previously (figure 3.10) were used to identify the most potent 5 pairs per gene. This library was screened in A375-Cas9 and K562-Cas9 cells and depletion log-ratios were measured for each construct as described above (figure 3.6). Dual CRoatan constructs were significantly more depleted than single sgRNAs. This was true when all the other prediction algorithms we tested were considered together (rank-sum test, p-val<0.0001) or separately (rank-sum test, sgRNAScorer: p-val=$1.5e^{-5}$, RNA-Config: p-val=0.0014, GPP WP: p-val=0.015) (figure 3.13A). Although CRoatan dual sgRNAs had higher depletion rates than CRoatan single sgRNAs, this difference was not significant (rank-sum test, p-val=0.4). Loss-of-function genetic screens are generally performed to identify genes implicated in a phenotype of interest without any prior knowledge. In these screens, “hits" are selected for follow-up experiments using criteria such as the number of construct per gene significantly depleted, or with a depletion rate above a predefined threshold. We performed such an analysis on the 5 depletion screens. We considered an increasingly large percentage of the most depleted sgRNA constructs (top 10% to top 50%), and called a gene a hit if at least two constructs targeting that gene were in the considered fraction. Using this analysis, we can estimate the true-positive and false-positive rate as the fraction of essential or non-essential genes identified as hits. In both sensitivity and specificity, the dual CRoatan library performed better than any other library (figure 3.13B).

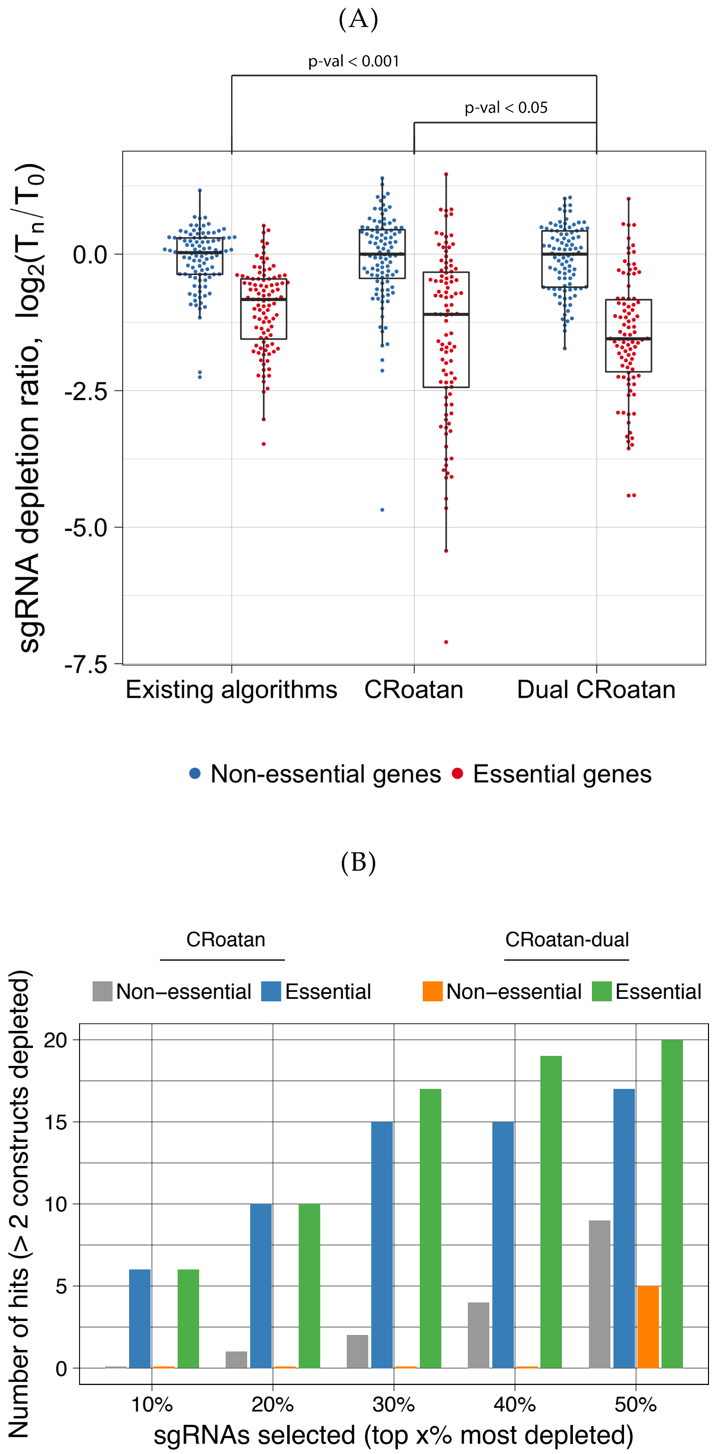


FIGURE 3.13: **Dual-CRoatan Constructs Provide Superior CRISPR-Based Gene Targeting (A)** Depletion rates of CRISPR constructs targeting essential and non-essential genes in loss-of-function screens. sgRNA sequences were either selected using existing algorithms (RNA-Config, sgRNAScorer, GPP WP) or using CRoatan. Vector harboring one (existing algorithms and CRoatan) or two (Dual CRoatan) sgRNAs were used. **(B)** Number of genes identified as hits when an increasing fraction of depleted sgRNA constructs is considered. To be classified as a hit, two constructs targeting that gene need to be depleted.

3.4 Discussion

CRISPR systems have been used extensively to manipulate the genome of multiple organisms, both in one-by-one experiments and large-scale genetic screens. Here, we leveraged available sgRNA efficacy datasets to build a random forest based machine learning algorithm to predict sgRNA potency. We combined this approach with stringent selection of putative target site by conservation and frame-shift mutation repair likelihood to maximize the phenotypic consequence of a DSB on protein function. Furthermore, we have demonstrated that significant gains in efficacy can be gained when multiple sgRNAs simultaneously target Cas9 to the target gene. We have thus designed a vector expressing pair of sgRNAs from two independent U6 promoters and implement an algorithm identifying pairs of guides with synergistic deleterious effect. This vector can also be used to study to study complex regulatory networks and genetic interactions in combinatorial screens.

To compare our library to preexisting sgRNA prediction tools, we performed six side-by-side loss of function genetic screens. CRoatan reagents showed significant greater efficacy, both in terms of depletion rates, sensitivity and specificity. Nevertheless, further improvements could surely be implemented. In our combinatorial CRISPR screen (figure 3.12B), we noticed that for pairs with only one essential gene targeting sgRNA, depletion rates were slightly reduced when that sgRNA was expressed from the chicken U6 promoter as opposed to the human U6 promoter. Thus, sequencing of the expressed guide RNAs could be performed to compare these two promoters as well as others that have been used such as the mouse U6 promoter and select the one leading to optimal levels of sgRNA expression.

Based on these results, we have used CRoatan to predict the most potent sgRNAs pairs for all human genes and we are building a sequence-verified genome-wide arrayed library. The availability of individual arrayed construct allows for large scale screening of complex phenotypes for which pooled screens are not amenable. Smaller pools of constructs can also be easily assembled for use in targeted screening approaches or *in vivo* screens for which the number of sgRNAs that can be used in one experiment is limited.

Overall, we hope that this toolkit will allow for individual and combinatorial gene knockouts to be carried out on a large-scale, both in multiplexed and arrayed formats. The current

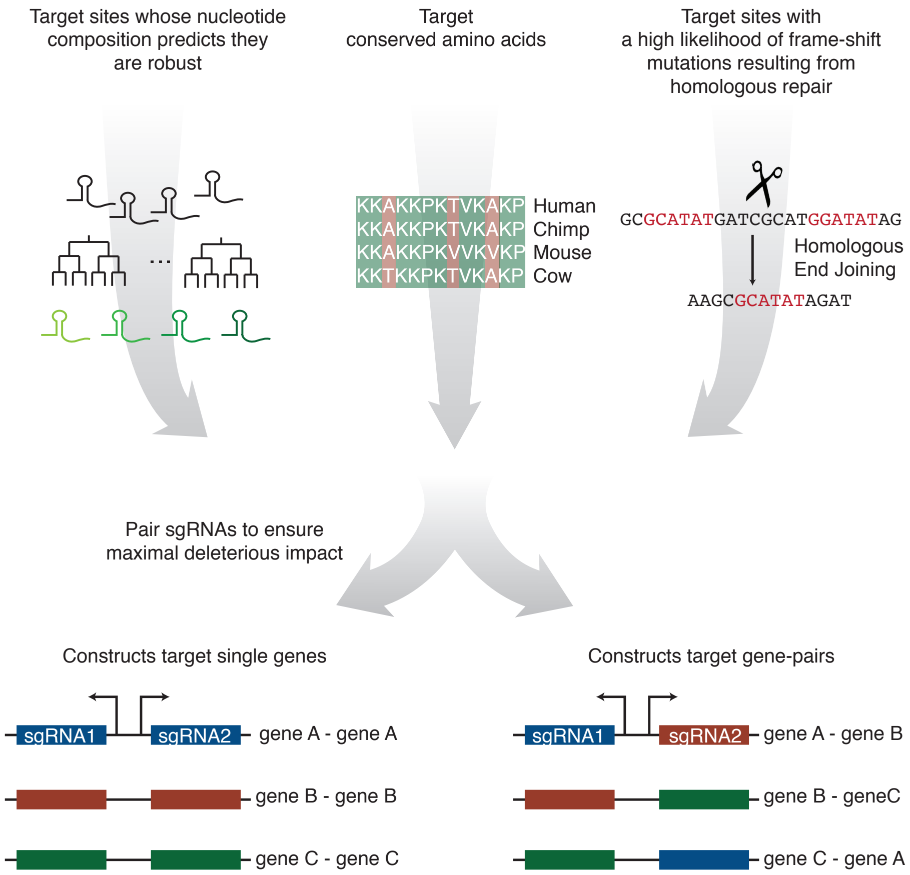

Target sites whose nucleotide composition predicts they are robust
Target conserved amino acids
Target sites with a high likelihood of frame-shift mutations resulting from homologous repair
KKAKKPKTVKAKP Human
KKAKKPKTVKAKP Chimp
KKAKKPKVVKVKP Mouse
KKTKKPKTVKAKP Cow
GCGCATATGATCGCATGGATATAG
Homologous End Joining
AAGCGCATATAGAT
Pair sgRNAs to ensure maximal deleterious impact
Constructs target single genes
Constructs target gene-pairs
sgRNA1
sgRNA2
gene A - gene A
gene B - gene B
gene C - gene C
sgRNA1
sgRNA2
gene A - gene B
gene B - geneC
gene C - gene A

FIGURE 3.14: **Graphical summary of the CRoatan sgRNA selection algorithm**

library design includes five dual sgRNA CRoatan constructs targeting all refseq annotated human coding sequences. Around 50 000 constructs have been sequence-verified at present, and we aim to complete this set to reach 100 000 sgRNA pairs targeting ∼20 000 genes.

Chapter 4

Investigating stroma-mediated gemcitabine resistance in pancreatic ductal adenocarcinoma

This project was carried out in collaboration with Graham Mills, PhD student in Prof. Duncan Jodrell's laboratory. For the purpose of this thesis, I will outline my specific contributions to this project. The isolation of the cancer-associated fibroblasts and cancer cell lines, as well as the characterization of the gemcitabine resistance phenotype was performed by the Jordell lab (data used for figure 4.2). To perform the genome-wide RNAi screen, I cloned the shRNA libraries, produced retrovirus, and infected the cancer cells. The subsequent passaging of cells in mono and coculture was a collaborative effort between Graham and myself, as well as the gDNA extraction and PCR of shRNAs. Esin Orhan, a master student I supervised, also screened one of the 8 shRNA pools. The screen analysis and figures presented here are the result of my own work.

4.1 Introduction

Pancreatic ductal adenocarcinoma (PDAC) is one of the most lethal human cancers, in part because it is resistant to many chemotherapeutic treatments (Kamisawa et al., 2016). Over the past decades, the main treatment has remained gemcitabine, a cytidine analogue, as well as more recently, combinatorial therapies such as FOLFIRINOX and gemcitabine with nab-paclitaxel (Burris et al., 1997; Vaccaro, Sperduti, & Milella, 2011; Von Hoff et al., 2013). Although these new treatments have increased patient outcome, their high toxicity is often not well tolerated and new therapies are needed.

One of the characteristics of pancreatic cancer is a strong desmoplastic reaction in which a dense and complex stroma surrounds the tumors. This fibrous tissue is composed of a variety of cell types such as pancreatic stellate cells, fibroblasts, immune cells and extracellular matrix, and has been shown to promote cancer progression and resistance to therapy (Feig et al., 2012). Activated pancreatic stellate cells (PSC, or cancer-associated fibroblasts) have been identified as one of the main drivers of the stromal reaction in PDAC (M. Apte et al., 2004; Bachem et al., 2005). PSCs are present in the pancreas in a quiescent state and are activated upon injury as well as in response to cytokines or to growth factors produced by cancer cells such, as PDGF, VEGF or TGFB-1 (Bachem et al., 2005; Vonlaufen et al., 2008). Once activated, PSCs assume a myofibroblast-like shape and secrete large amounts of extracellular matrix proteins, growth-factors and cytokines. PSCs have been shown to promote cancer progression both *in vitro* and *in vivo*. *In vitro* cultures of PDAC cell lines with activated PSC conditionned media lead to an increase in the cancer cells proliferation, motility and resistance to gemcitabine (Hwang et al., 2008; Z. Xu et al., 2010; Vonlaufen et al., 2008). Furthermore, subcutaneous or orthotopic injection of cancer cells with activated PSCs increases tumor size and metastasic burden *in vivo* (Hwang et al., 2008; Z. Xu et al., 2010; Bachem et al., 2005; Vonlaufen et al., 2008).

The dense stroma associated with PDAC also obstructs tumor vasculature which impedes drug delivery. This was first shown in genetically engineered mouse models recapitulating the progression of PDAC by inhibiting the pro-stromal Hedgehog signalling pathway using a Smoothened inhibitor (Olive et al., 2009). This led to stromal depletion and increased tumor vasculature and gemcitabine delivery. Further studies aimed at removing the stroma by targeting hyaluronan, an abundant extracellular matrix component whose accumulation leads to elevated interstitial fluid pressure, compressing blood vessels (Provenzano et al., 2012; Jacobetz et al., 2013). Its enzymatic removal using hyaluronidase greatly improves response to chemotherapies in mice. Although stroma ablation leads to sensitivity to treatment on the short-term, genetic deletion of Hedgehog in PDAC mouse models led to more aggressive tumors with increased metastatic burden, indicating that the stroma could also play a role in controlling tumor growth (Rhim et al., 2014; Ozdemir et al., 2014). Clinical trials targeting the

Hedgehog pathway in combination with gemcitabine have also failed to increase patient survival, which suggests that stromal depletion to increase drug delivery might not be a viable therapeutic strategy (E.J. Kim et al., 2014).

Regardless of its role as a physical barrier to drug delivery, the stroma has also been shown to confer tumor cells of multiple cancers innate resistance to therapy in *in vitro* coculture experiments (Straussman et al., 2012; Wilson et al., 2012). In PDAC, an *in vivo* study has shown that targeting the vitamin-D receptor expressed by PSC reprograms PSC towards their quiescent state and increases tumor sensitivity to gemcitabine (Sherman et al., 2014). Targeting specific pathways in activated PSC rather than removing the stroma can thus be leveraged to alleviate PDAC resistance to treatment. In this project, we took advantage of a CAF-PDAC coculture system that recapitulates stroma-mediated gemcitabine resistance to perform a high-throughput RNAi screen. With this strategy, we hope to identify genes and pathways involved in this resistance mechanism that could be used in targeted therapies in combination with gemcitabine to treat PDAC.

4.2 Material and methods

4.2.1 Genome-wide shRNA libraries

shRNA libraries in the MSCV-based LMN vector were ordered from Transomics Technologies. These libraries harbor shRNAs predicted by the shERWOOD algorithm (S.R. Knott et al., 2014). 8 pools of ∼10 000 shRNAs covering 19 000 mouse genes were used. Each pool was subcloned into an MSCV-based LMH vector. In these vectors, shRNAs are expressed from the long terminal repeat (LTR) retroviral promoter, and mCherry and a Hygromycin resistance gene are expressed from the mouse phosphoglycerate kinase 1 (PGK) promoter in a bicistronic fashion. To preserve complexity and avoid the introduction of mutations, the shRNAs were excised from the LMN vectors by digestion using BglII and MluI (NEB). After gel extraction, the miR cassettes were ligated into LMH previously cut with the same enzymes. Ligation reaction were purified on MinElute columns (QIAGEN) and transformed into MegaX DH10B T1R electrocompetent cells (Life Technologies). For each pool, at least 5 million independent transformants were obtained.

4.2.2 Cell lines

Both cancer and cancer associated fibroblasts cell lines were provided by the Jodrell Laboratory, CRUK Cambridge Institute. The K8484 cell line was established from a pancreatic ductal adenocarcinoma of a KRAS^{G12D}; p53^{R172H};PDX-Cre (KPC) genetically engineered mouse (Hingorani et al., 2005). Primary cancer associated fibroblasts were isolated using a differential adhesion protocol from KPC tumor tissue. Both K8484 and mCAFs cell lines were grown in grown at 37°C, with 5% CO_2 in DMEM supplemented with 10% FBS and 50U/mL of penicillin-streptomycin. mCAFs were engineered to be sensitive to diphteria toxin by transduction with a retroviral vector in which the diphteria toxin receptor is driven by the LTR. To select transduced cell, the vector also expresses the ZsGreen fluorescent protein and a Neomycin resistance gene driven by the PGK promoter. To make the mCAFs-DTR cell line, mCAFs were transduced at low multiplicity of infection (MOI) and selected for 10 days with 300 μg/mL of Geneticin (Life technologies). The Phoenix-Eco (ATCC CRL-3214) packaging cells were grown at 37°C with 5% CO_2 in DMEM supplemented with 10% FBS and 50U/mL of penicillin-streptomycin.

4.2.3 Coculture assay

20 000 mCAFs-DTR (counted on the Beckman Vi-Cell) were plated in 96-well plates. The following day, 2 000 K8484-LMH-mCherry cells were seeded on top of the mCAFs-DTR-ZsGreen (cocultures) or in empty wells (monocultures). On the third day, cells were dosed with gemcitabine concentrations ranging from 0.001 to 30 μM. 72h later, the cells were harvested and the cancer cells expressing mCherry were counted by flow cytometry on the Miltenyi Biotech MacsQuant analyzer. For these drug curves, the number of cancer cells for each gemcitabine concentration was normalized to the number of cells in the DMSO control.

4.2.4 Virus production

Phoenix-Eco cells were plated at 50% confluency in 15cm tissue culture dishes. The following day, cells were transfected using the calcium-phosphate transfection method (). A transfection mix was prepared with viral vector (60μg), VSV-G (7.5μg), 20nM Pasha siRNA (200 μL), 2M calcium chloride (187.5μL). Water was added to reach a total volume of 1.5mL. This mixture was added to the same volume of 2xHBS while it was being bubbled. 2XHBS buffer (50mM HEPES, 280mM NaCl, 1.5mM Na_2P0_4, 12mM Dextrose, 10mM KCl) was prepared before hand. Several aliquots were prepared and adjusted to pH ranging from 6.9 to 7.1 using NaOH. Each aliquot was tested by test transfection and only the best one was kept. After 60s bubbling, the mixture was incubated for 15min at room temperature and added to the Phoenix-Eco packaging cells, with 17mL of fresh media and 7.5μL of 100mM chloroquin. Media was replaced 14 hours later, and Sodium Butyrate was added (1000X stock at 1M). Thirty hours later, virus was collected, filtered using a 0.45μm filter (EMD-Millipore) and stored at 4°C.

4.2.5 Screen infection and selection

For each shRNA library, virus titers were first estimated by test infecting K8484 cells with different volumes of virus, and the volume needed to infect 30% cells was calculated. Large-scale infections were then performed in 15cm plates. In each plate, 10M K8484 cells were incubated with the appropriate volume of virus, 8μg/mL polybrene in 16mL of media. Infection percentage was measure two days later by flow cytometry on the Miltenyi Biotech MacsQuant analyzer. At least 1000 infected cells per shRNA were obtained. All the infections were done in triplicates. K8484 infected with LMH based constructs were then selected for a week with 500μg/mL Hygromycin B (ThermoFischer scientific).

4.2.6 Monoculture and coculture screens

Once selected, infected K8484 cells of each replicate were split in 5 samples: two monoculture screens (plus and minus gemcitabine), two for the coculture screens (plus and minus gemcitabine) and one as an initial timepoint. For monoculture screen, 5M infected K8484 were plated on 5 15cm dishes. The following day, 40μL of 5μM gemcitabine (final concentration of

10nM) or 40μL of DMSO was added. Cells were dosed for 72h, harvested, counted (Vi-cell) and and 5M cells were plated and dosed again. This was repeated 6 times, after which cells were harvested for a final timepoint. For coculture screens, 5M infected K8484 and 50M mCAFs-DTR were plated on 5 15cm dishes. 40μL of 50μM gemcitabine (final concentration of 100nM) or 40μL of DMSO was added the following day and cells were dosed for 72h. Cells were then harvested, counted and relative numbers of cancer and CAFs was assessed using the Miltenyi Biotech MacsQuant analyzer. 5M cancer cells were plated and dosed again. When necessary, fresh mCAFs were added to bring up the number of CAFs to 50M. This was repeated six times. At the end of the screen, 100ng/mL of dipheteria toxin (Sigma Aldrich) was added to selectively kill mCAFs-DTR. After 4 days, the remaining cells were harvested for a final timepoint.

4.2.7 Screen sequencing

gDNA was extracted from samples using the QIAmp DNA Blood Maxi kit (Qiagen). shRNAs were amplified by PCR (forward primer: CAGAATCGTTGCCTGCACATCTTGGAAAC, reverse primer: CTGCTAAAGCGCATGCTCCAGACTGC). For each sample, 32 50μL PCRs (KOD, Takara) with 2μg of gDNA as input were performed. 1mL of this first PCR was purified on 4 Qiagen PCR purification columns. Illumina adpaters were added in a secondary PCR (forward primer P7-BCX-TS-Mir-Loop: CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTG GAGTTCAGACGTGTGCTCTTCCGATCTTAGTGAAGCCACAGATGTA where Ns are the 6bp sequencing index, reverse primer P5-mir3: AATGATACGGCGACCACCGAGATCTACA CCAGCAGTATGTTGAAGTCCGAGGCAGTAGGCA). To maintain complexity, 2 PCRs with 500ng of amplicon from the first PCR were done. The amplicon from the secondary PCR was run on a 1.5% agarose gel and the band ∼150nt was excised, and the DNA was purified using a Qiagen MinElute purification column. Libraries were then quantified using a Qubit fluorometer (ThermoScientific) and pooled. The pool was then quantified precisely by qPCR (KAPA Illumina library quantification kit) and sequenced on an Illumina HiSeq with a custom read 1 primer (CAGCAGTATGTTGAAGTCCGAGGCAGTAGGCA). Each sample was sequenced at a depth of at least 10 million reads.

4.2.8 Screen analysis

Reads were mapped to a custom index made of the mouse shRNA sequences present in all pools using bowtie (Langmead et al., 2009). Differentially expressed shRNAs were called using DESeq2 (Love, Huber, & Anders, 2014) or by selecting shRNA for which the ZScore of the log2 fold change ratio of normalized reads before and after treatment was lower than -1.96. For both Z-Score and DESeq2 analysis a gene was considered a hit if at least two shRNAs targeting it were hits. Genes for which only one shRNA were included in the library were called hits if that shRNA was depleted. To perform gene-set enrichment analysis, shRNAs were ranked using the log2 fold change ratio of read counts before and after treatment. Then, each gene was assigned a rank equal to the average rank of the two most depleted shRNA targeting that gene. Gene set enrichment analysis (GSEA) was then performed using GSEAPreRanked with the ranked list as input, and the hallmark and KEGG pathway gene sets as databases (Kanehisa & Goto, 2000; Liberzon et al., 2015; Subramanian et al., 2005).

4.3 Results

4.3.1 A coculture assay recapitulates stroma-mediated resistance *in vitro*

To study stroma-mediated gemcitabine resistance of pancreatic cancer cells, the Jordell lab has developed a 2D coculture assay. In this assay, tumor cells express a fluorescent protein and are plated on top of a confluent monolayer of cancer-associated fibroblasts. The growth of the cancer cells and their response to drug treatment can be tracked using the intensity of the fluorescent marker (figure 4.1). One of the cell lines that has been extensively tested with this assay is the pancreatic ductal adenocarcinoma (PDAC) derived K8484 cell line. It was established from the PDAC of the genetically engineered KRAS^{G12D}; p53^{R172H};Pdx-1-Cre (KPC) mouse model. In this model, the Cre recombinase is placed under control of the Pdx-1 promoter, specific to pancreatic progenitor cells. Upon Cre expresssion, pancreatic cells express the mutant KRAS^{G12D} and p53^{R172H} alleles, which leads to the development of metastatic PDACs that recapitulate the evolution of the human cancer.

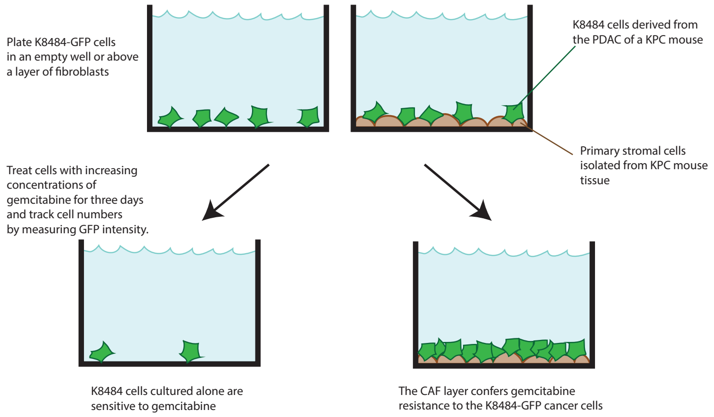


Plate K8484-GFP cells in an empty well or above a layer of fibroblasts
K8484 cells derived from the PDAC of a KPC mouse
Primary stromal cells isolated from KPC mouse tissue
Treat cells with increasing concentrations of gemcitabine for three days and track cell numbers by measuring GFP intensity.
K8484 cells cultured alone are sensitive to gemcitabine
The CAF layer confers gemcitabine resistance to the K8484-GFP cancer cells

FIGURE 4.1: **Schematic of the cancer cell and cancer associated fibroblast coculture assay** This assay developped by the Jordell lab to study stroma-mediated resistance to gemcitabine treatment

The coculture of K8484 cells with αSMA^{+} fibroblasts derived from KPC shows that the cancer become resistance to gemcitabine treatment when compared to cells grown in monocultures (figure 4.2). The Jordell lab has done complementary experiments to characterize this phenotype and shown that it is transient, relies on cell-to-cell contact and is specific to cancer associated fibroblast as it was not observed when coculture assays were performed with stromal cells isolated from normal fibroblasts (data not shown).

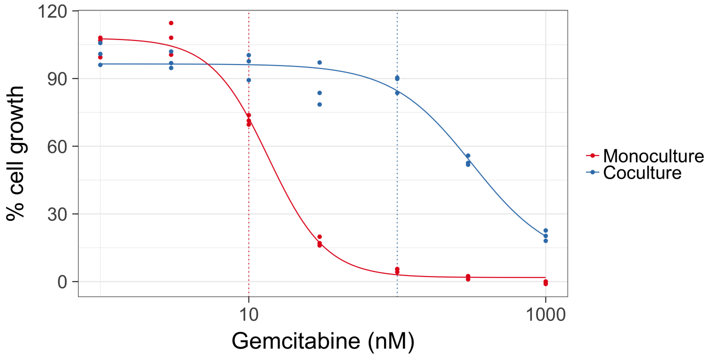


FIGURE 4.2: **K8484 PDAC cells cocultured with cancer associated fibroblasts are resistant to gemcitabine** Dose response curves of K8484-GFP cells in monoculture or grown on a monolayer of CAFs. Cells were treated with increasing concentration of gemcitabine and the number of K8484-GFP cells was assessed after three days of culture using the Pherastar plate reader. The dotted lines indicate the concentration of gemcitabine used in the monoculture screen (blue) or in the coculture screen (red).

To investigate the mechanism and genes involved in this stroma-mediated gemcitabine resistance phenotype, we performed a genome-wide RNAi depletion screen. Several approaches could be taken to perform this screen: targeting the fibroblasts, or the cancer cells, in pools or in an arrayed format. Targeting the fibroblasts requires this screen be done in 96- or 384-well plates, and the number of surviving K8484 cells following gemcitabine treatment could have been used to identify hits. However, testing $\sim$80 000 shRNAs using this assay would have been infeasible so we preferred to perform a pooled RNAi screen in the K8484 cell line. The coculture assay described above was routinely done in 96-well plate, a format not suitable for the large amount of cells that need to be cultured during large scale screens. We implemented several modifications in the protocol to adapt it to large tissue culture dishes and to allow extraction and sequencing of shRNAs from large pools of infected cancer cells in coculture with CAFs.

4.3.2 Genome-wide RNAi screens of K8484 PDAC cells in mono- and coculture

RNAi screens are generally analyzed by estimating the abundance of each shRNA in the initial and final timepoint using high-throughput sequencing. To this end, the shRNAs integrated in the genome of infected cells are amplified from genomic DNA by PCR. This proved difficult for the timepoint of coculture screens as the pools of cells comprised 10% infected cancer cells and 90% cancer associated fibroblasts, diluting the target loci. As this PCR needs to be as efficient as possible to avoid skewing the distribution of shRNA counts by PCR bias, the cancer associated fibroblasts needed to be eliminated from the pool of cells before PCR. To this end, we transduced the CAF cell line with a plasmid harboring the human diphteria toxin receptor (DTR) gene, a ZsGreen fluorescent protein, and a Neomycin resistance gene. (Saito et al., 2001). As naive murine cells are resistant to diphtheria toxin because they lack the receptor, DTR expressing cells can be killed selectively and rapidly. When treated with diphteria toxin, 80 to 95% of expressing DTR cells were found to be killed within 4 days while survival rates of naive cells were unaffected. We thus used this strategy at the end of coculture screens to enrich for K8484 infected with shRNAs before harvesting and PCR amplification.

With these cell lines in hands, we performed a coculture assay in 15cm dishes and observed the same stroma-induced gemcitabine resistance. This confirmed that this experiment could be scaled up and suitable for high-throughput pooled screens. To specifically identify genes

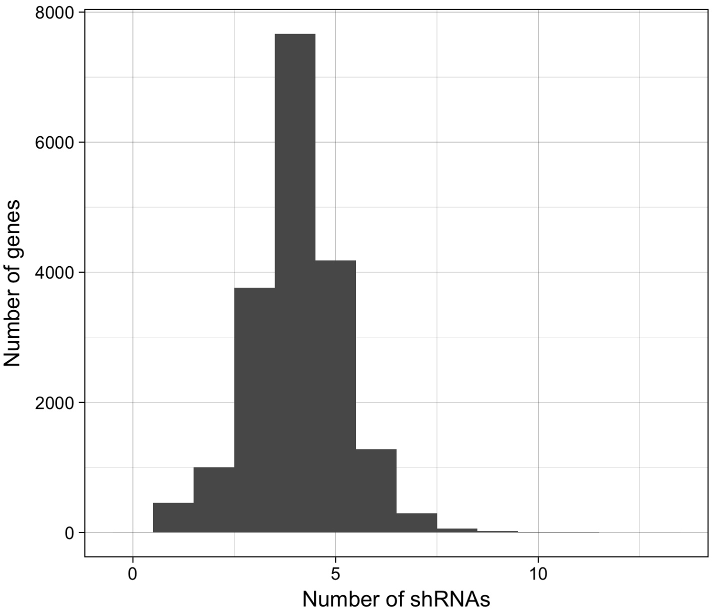


FIGURE 4.3: **Multiple shRNA per gene were used to target each gene in the RNAi screen** Shown is the distribution of number of shRNAs per gene in the RNAi library used for the screens

involved in the cross-talk between CAFs and cancer cells, we decided to perform 4 different screens: monoculture and coculture, both on and off gemcitabine. Hits in the monoculture and coculture off-drug screens would be genes required for proliferation of the cancer cells in mono- or coculture independently of gemcitabine treatment. Hits in the monoculture screen would be genes that sensitize the cancer cells to gemcitabine. The results from these three screens can thus be used to filter the hit list of the coculture on-drug screen and identify genes that only hit when the cells are grown in coculture and with gemcitabine.

Two different gemcitabine concentrations were chosen for mono- and coculture experiments, based on the drug curves (figure 4.2). For monoculture screens, we chose a drug concentration of 10nM which corresponds to the IC80. At this concentration, only cells harboring shRNAs targeting gemcitabine sensitizer genes would be significantly depleted from the pool. K8484 in coculture are completely resistant to 10nM gemcitabine so the concentration was increased to 100nM for cocultures screens. This concentration is sufficient to kill 90% of cancer cells in monoculture and will cause cell death in coculture if the shRNA interferes with the resistance phenotype.

The mouse genome-wide shRNA library we used for this screen was provided by Transomic technologies, which distributes the tools I described in Chapter 1. To reduce the number of cells that needed to be cultured, the screen was split in 8 pools of 10 000 shRNAs. All the hairpins in library were chosen using the shERWOOD algorithm and are cloned in the Ultramir backbone (S.R. Knott et al., 2014). It covers over 18 500 mouse genes with an average of 4 shRNAs per gene to control for off-target effects (figure 4.3). This library was originally cloned into LMN, a MSCV-based retroviral vector in which the shRNAs are expressed from the LTR viral promoter, and the ZsGreen fluorescent protein as well as the Neomycin resistance gene are driven bicistronically by the PGK promoter. As the mCAFs-DTR cell line expresses ZsGreen as well, we subcloned the shRNA library into LMH, a vector similar to LMN where ZsGreen was replaced with mCherry and the Neomycin resistance gene by a Hygromycin resistance gene. This facilitates relative counting of CAFs and infected cancer cells by flow-cytometry, the cancer cells expressing mCherry and the CAFs-DTR ZsGreen.

Each library pool was packaged in retrovirus and transduced in K8484 cells in triplicates, at low multiplicity of infection to avoid double infection events. The number of infected cells was superior to 500 times the number of shRNAs in each pool to keep the full complexity of the libraries. After selection with hygromycin, each replicate was split into 5 samples: monoculture on and off drug, coculture on and off drug, and the rest of the cells were collected as an initial timepoint. For cocultures, mCAFs-DTR were added to reach a ratio of 1/10 cancer cells to fibroblasts. The next day, gemcitabine or DMSO was added to all screens, and cells were cultured for 72hours. Cells were then split and replated, this was repeated 6 times to allow for dilution of cells for which growth was hampered by the knockown. For coculture screens, the relative percentage of cancer cells to fibroblasts was estimated by flow cytometry and fresh mCAFs-DTR were added when necessary to keep the 1/10 ratio (figure 4.4).

After 6 gemcitabine dosings, monocultures were harvested for a final timepoint. Cocultures were treated with diphteria toxin for 4 days and harvested. For all samples and timepoints, genomic DNA was extracted and abundance of shRNAs estimated by high-throughput sequencing. Depletion ratios were then calculated for each shRNA by dividing counts at the final timepoint by counts at the initial timepoint.

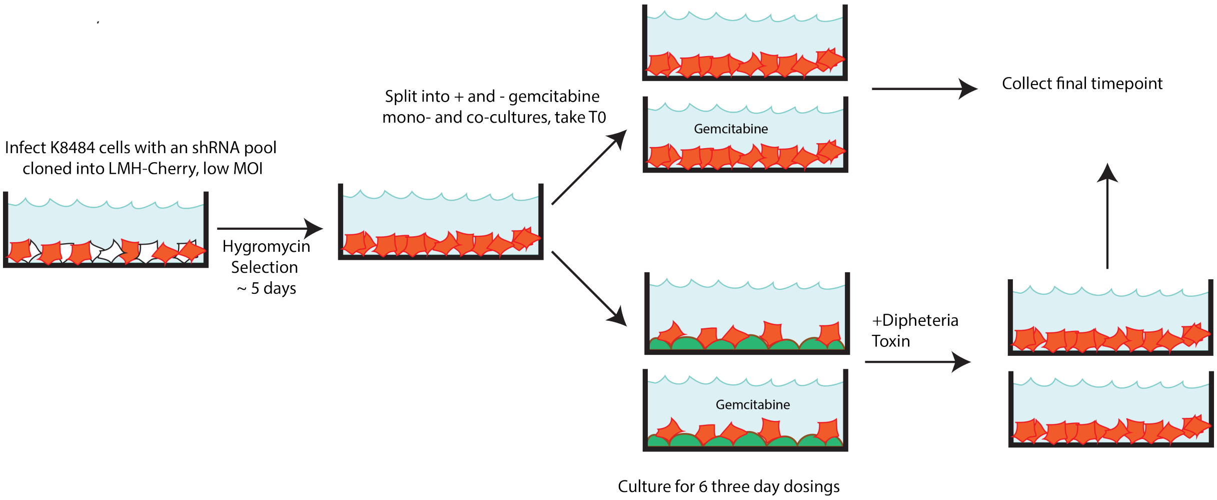

Infect K8484 cells with an shRNA pool cloned into LMH-Cherry, low MOI
Hygromycin Selection ~ 5 days
Split into + and - gemcitabine mono- and co-cultures, take T0
Gemcitabine
Collect final timepoint
+Dipheteria Toxin
Gemcitabine
Culture for 6 three day dosings

FIGURE 4.4: **Coculture and monoculture genome-wide RNAi screens** Large scale RNAi screens can be performed can be performed on PDAC-CAFs coculture by introducing the diphteria toxin receptor gene in CAFs to negatively select them at the end of the experiment.

Screen analysis

To identify shRNAs consistently depleted in the three replicates of each samples, we used two different statistical tools to call hits: DESeq2 and a Z-Score based ranking. DESeq2 performs differential expression analysis by fitting negative binomial generalized linear models to each shRNAs and uses a Wald test for significance testing. Because the genes targeted in each pool are different, normalization across pools was not possible so we applied DESeq2 to each pool of shRNAs independently. As an output, this analysis provides normalized log2-fold change for each shRNA as well as whether they are statistically significantly enriched or depleted (figure 4.5A).

To validate the hits identified by DESeq2, we performed a second analysis based on Z-Scores (figure 4.5B). For each condition, the median of the log-ratios of the three replicates was used as a measure of shRNA depletion/enrichments. Z-scores were then calculated for all shRNAs of the pool using this value, and shRNAs with a z-score below -1.96 or above 1.96 were deemed significantly enriched or depleted (this corresponds to a p-value < 0.05 in a normal distribution setting).

Although little is known about the genes implicated in the stroma-mediated gemcitabine

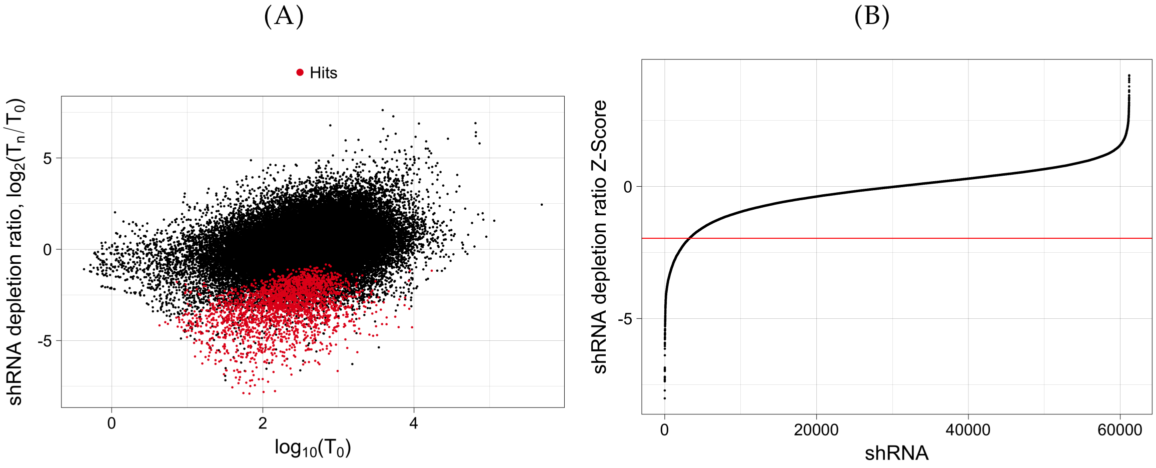


FIGURE 4.5: **Screen hit calling using DESeq or a Z-Score analysis** K8484 cells were screened with a genome-wide RNAi libraries, in monoculture and coculture, with and without gemcitabine. Screens were analysed using DEseq (A), or using Z-Scores (B). Shown in (A) are the result of the DESeq analysis for the coculture screen with gemcitabine. Shown on the y-axis is the log fold change of read counts for each pair at the final timepoint compared to the initial timepoint. The x-axis is the log of counts at the initial timepoint. Colored are significantly depleted shRNA (padj <0.05). Shown in (B) is the result of the Z-Score analysis of the same screen. The y-axis shows each shRNA depletion Z-Score. A cutoff of -1.96 was set to call an shRNA a hit.

resistance phenotype, genes such as Checkpoint Kinase 1 (CHK1) and the Serine/threonine-protein kinase ATR have been identified as sensitizers to gemcitabine have been identified (Prevo et al., 2012; Koh et al., 2015). For both the DESeq2 and Z-Score analysis, most of the shRNAs targeting these genes were strongly depleted which gave us confidence with respect to the relevance of our assay.

When we compared the shRNAs identified as hits in both of our analysis, we found that only ∼15% were common to both (figure 4.6). One characteristic of DESeq2 is that it performs automatic count outlier detection and removal. shRNAs with highly variable log2-fold changes, for example that are depleted in 2 of the replicates but not in the third one would be removed by DESeq2 while they could be called hits in the Z-Score analysis as it considers the median of the three replicates. On the other hand, our Z-Score threshold selects for strongly depleted shRNAs compared to the rest of the population while DESeq2 also can also identify constructs that are depleted less strongly if that depletion is consistent across replicates.

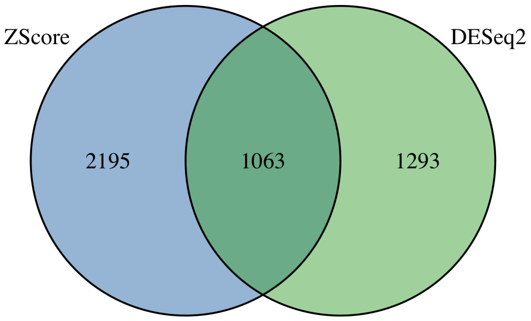


FIGURE 4.6: **Different shRNAs are identified as hits using DESeq or a Z-Score analysis** Venn diagram of shRNAs identified as hits using both analysis methods described in figure 4.5.

The two analysis described above focus on scoring depleted shRNAs. However we are mostly interested in finding which genes are responsible for our phenotype. As the shRNA library harbors multiple shRNAs targeting each genes (figure 4.3), we used a cut-off of at least 2 significantly depleted shRNAs per gene to call gene-level hits. Genes targeted by only one or two shRNA were also considered hit if one of the shRNA was significantly depleted. Although these cutoffs are rather loose, we wished to reduce our false-negative rates at this stage and not remove shRNAs targeted with few shRNAs. This certainly increases the number of false-positives but these can be removed in validation experiments.

Gene hits were called based on the shRNA-level results from the Z-Score and DESeq2 analysis, for each condition (figure 4.7). Overall, the Z-Score based analysis identified a larger number of gene hits than the DESeq one which is coherent based on the number of shRNAs classified as significantly depleted by both methods. For both analysis, ∼300 genes were identified as specifically depleted in the coculture plus gemcitabine sample. Surprisingly, over 150 genes were hits only in the coculture minus screen. These hits are in theory not lethal genes as these would be shared with the monoculture minus gemcitabine sample. As the cancer cells are competing with the fibroblasts in the cocultures, these are perhaps weaker hits that only

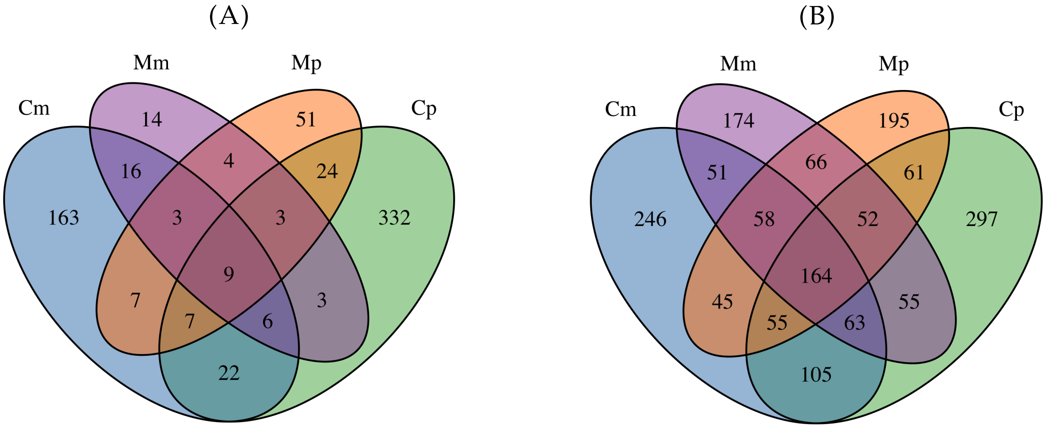


FIGURE 4.7: **Identification of gene hits in the coculture and monoculture screens** Venn diagram of genes identified as hits using the DESeq (A) or ZScore analysis (B). A gene was considered a hit if at least 2 shRNAs were significantly depleted, or one shRNA if the gene was targeted by a unique shRNA. C and M stand for coculture and monoculture. m and p stand for minus gemcitabine or plus gemcitabine.

appear when additional selection pressure is applied. However, it remains unclear why these hits are not also shared with the coculture plus gemcitabine sample.

To identify pathways implicated in the resistance phenotype, we performed gene-set enrichment analysis in the hits of the four screens, using the hallmark and KEGG pathway gene sets (Kanehisa & Goto, 2000; Liberzon et al., 2015; Subramanian et al., 2005) (figure 4.8). In both cases, pathways related to cell proliferation and maintenance such as DNA replication, cell cycle, or proteasome and ribosomal activity were significantly enriched in hits of all screens. However, no pathway was identified as significantly enriched only in hits of samples treated with gemcitabine which could have provided insights into genes to target in follow-up experiments.

Before performing one-by-one validation experiments, we wished to repeat the screen on a smaller scale to confirm that the hits identified in the large-scale experiments were true-positives. For this secondary screen, we selected all genes identified as lethal in the monoculture and coculture plus gemcitabine screen. To control for false-positives, we ordered 5 shRNA per gene and added 200 shRNAs targeting human olfactory receptors, for a total of ∼6000 shRNAs. These experiments are currently ongoing.

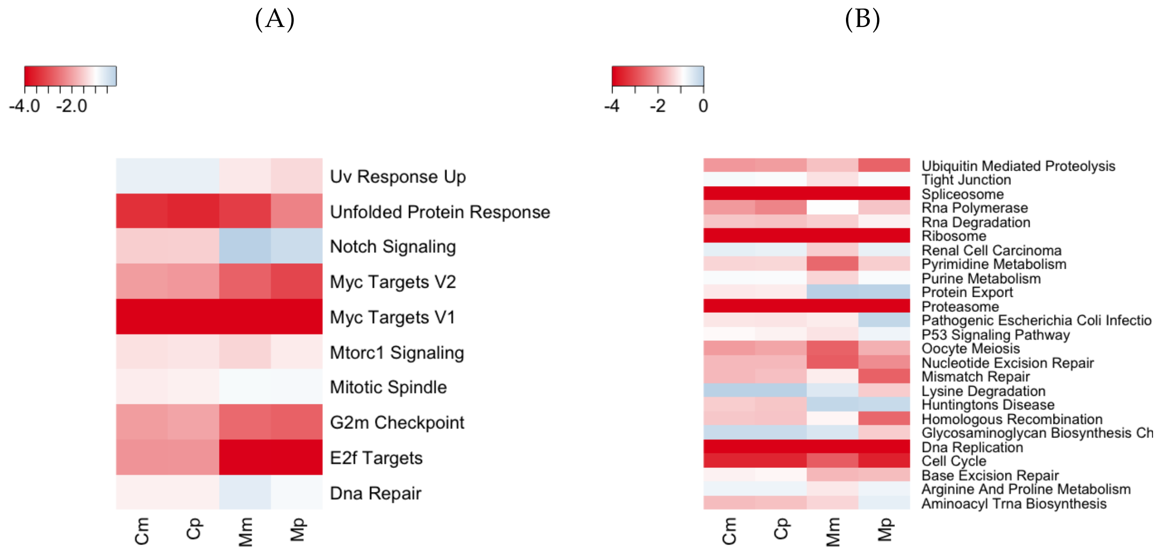


FIGURE 4.8: **Gene-set enrichment analysis identifies significantly depleted pathways** Genes were pre-ranked using the average rank of the the two most depleted shRNA targeting each gene. The hallmark gene set (A), or the KEGG gene set (B) were used. Shown are pathways significantly enriched in depleted shRNAs (FDR $<$ 0.1). Color indicates log10 of the p-value.

4.4 Discussion

One of the challenges of investigating stroma-mediated gemcitabine resistance mechanism is the lack of assays that can be leveraged to perform high-throughput experiments. Here, we adapted a CAF-PDAC *in vitro* coculture assay to perform a genome-wide RNAi screen to identify sensitizers to gemcitabine. This required scaling up the format of the assay as well engineering the CAFs to add the diphteria toxin receptor to specifically extract cancer cells from the coculture. The scale and length of the RNAi screen certainly introduces additional noise to the experiment, as only one in ten cells in culture harbors an shRNA and some may be depleted during passaging of the cocultures. However, essential pathways as well as some genes involved in gemcitabine resistance in monocultures were found to be depleted which gives us some confidence in the hits that will be identified in this assay. It is still too early to pinpoint genes involved in the resistance mechanism, but the secondary RNAi screen will hopefully be helpful in identifying sensitizers to gemcitabine in the monoculture screens as well as players involved in the CAF-cancer cell crosstalk in the coculture screens.

Chapter 5

Conclusions

The aim of this thesis was to optimize molecular biology tools to perform large-scale genetic screening, both in a combinatorial or single target setting. Genetic screens have been widely used to chart genome-phenotype interactions. However, the strength and reproducibility of the results drawn from these experiments greatly depends on the potency of the reagents used to elicit knock-down or knock-out. I first worked on developing genome-wide libraries of highly potent shRNAs. Active sequences can be selected efficiently by acquiring large empirical shRNA efficacy datasets to train machine-learning algorithms (Fellmann, Zuber, et al., 2011; S.R. Knott et al., 2014). These sensor assays, combined with optimized expression vectors and synthetic miRNA backbones, allowed us to generate compact (ie with a small number of highly efficient shRNAs targeting each gene) shRNA libraries that improved the accuracy of RNAi dropout screens (S.R. Knott et al., 2014). Building on this experience, we applied the same reasoning to build CRISPR libraries by using published sgRNA activity datasets (Doench, Hartenian, et al., 2014; Chari et al., 2015) and designing vectors expressing pairs of sgRNAs to maximize knock-out likelihood.

One of the differences in predicting potent shRNA vs sgRNA stems from the differences in targeting mRNA or DNA. For shRNAs, targeting any sequence to the mRNA will lead to its complete degradation. The effect of an sgRNA, however, is not only dependent on the efficacy of guiding Cas9 to the loci and on its cleavage but also on the repair of the double-strand break and its effect on the coding sequence. Although most previously published sgRNA design algorithm relied mostly on predicting sgRNA efficacy (Doench, Hartenian, et al., 2014; Chari et al., 2015), we have shown that targeting conserved amino-acid increases protein functional knock-out rate by increasing the likelihood that even the removal of a few base-pairs during DSB repair will be deleterious to protein function. Moreover, although DSBs where thought to be mainly repaired by NHEJ, leading to repair pattern difficult to predict, recent studies have revealed that the DSB is flanked by small stretches of homologous sequences, it can be repaired by HEJ in a predictable manner (Bae et al., 2014). We have taken advantage of this knowledge and we showed that selecting target sites that would induce predictable frame-shift mutations also results in increased knock-out rates.

The development of new expression strategies for shRNAs and sgRNAs has also driven the increase in their specificity and potency. For shRNAs, one improvement was the use of mir30-based shRNAs expressed from Pol II promoters mimicking endogenous miRNA (Silva, M.Z. Li, et al., 2005). These constructs were shown to be more efficient than simple hairpin designs similar to pre-miRNAs. We have also shown here that the potency of miR-30 based constructs can be further increased by using scar-free cloning which allows removal of any restriction sites in the scaffold of the mir30 shRNA. Others have achieved similar results by moving the restriction sites used for cloning away for the hairpin structure to facilitate processing (Fellmann, Hoffmann, et al., 2013). Similarly, several chimeric sgRNA backbones replacing the crRNA/tracRNA hybrid have been tested to maximize knock-out rates using Cas9 (Jinek, Chylinski, et al., 2012; S. Cho et al., 2013). In additions to these backbone optimizations, vectors expressing multiple sgRNAs from Pol III promoters or as crRNA arrays have been used to direct Cas9 to multiple genes (Cong et al., 2013; Zetsche, Heidenreich, et al., 2016; Vidigal & Ventura, 2015). We used a similar strategy here to focus Cas9 to multiple sites within the same gene. This significantly increases functional consequences when compared to single sgRNA targeting.

Here, we have combined both machine-learning and expression strategies to create efficient shRNA and sgRNA constructs to maximize gene perturbation at the mRNA or gDNA level. Using these tools, we have generated libraries of constructs targeting all protein coding genes in the human or mouse genome. Each shRNA or sgRNA library is comprised of over $\sim$100 000 sequence-verified constructs targeting $\sim$ 20 000 genes. These libraries are available in an arrayed format and can be used to perform large-scale pooled or arrayed screens.

Moving forward, several improvements to both selection algorithms as well as expression strategies could be implemented to further increase the efficacy of these tools. For both shRNAs and sgRNAs, design algorithms rely on large efficacy datasets and their accuracy depends on the quality and size of that data. Subsequent shRNA predictors to the one presented here have used larger integrated datasets and two-tiered SVMs to increase accuracy. Similarly, increasing datapoints could yield better predictions of potent sgRNAs, as current predictors have relied on datasets of less than 2000 sequences. In addition, the development of sensor type assays for sgRNAs similar to the ones used shRNAs could possibly lead to better predictions. In this study, we used datasets from Doench et al and Chari et al. Doench et al targeted nine surface proteins with tiled libraries of sgRNAs and assessed knockout rates using flow-cytometry. The assay was performed over two weeks which possibly may limit separation of potent *vs* less potent sgRNA and, as endogenous loci are targeted, other factors such as locus accessibility could be at play. Chari et al first introduced synthetic target-site libraries in cells before transfecting sgRNA libraries, and used high-throughput sequencing to assess knock-out. Each cell thus received a large number of different sgRNAs that compete for Cas9 binding which could also introduce a bias in the selection. In my opinion, an ideal sensor experiment would be to deliver both the sgRNA and the target site in the same vector and use direct sequence-based readout. This would mimic closely the way sgRNAs are used in genetic screens (one sgRNA expressed in each cell) and sequencing the target sites allows to score both the frequency of indel generation as well as the type of genomic scars that are introduced by the repair. This could also be used to study micro-homology directed end joining repair with larger datasets.

In addition to larger and more accurate training datasets, improving sgRNA expression could lead to better knock-out rates. In our current library, pairs of sgRNA, we used a human and a chicken U6 promoter to avoid any recombination. However, combinatorial experiments with lethal and non-lethal sgRNA expressed from either the cU6 or hU6 promoter show that the activity of sgRNAs expressed from the cU6 promoters is lower. We have not yet determined whether this is due to transcription levels or precision of the transcription from the cU6 promoter. Identifying or testing new U6 promoters could perhaps also contribute to higher efficacy in future sgRNA libraries.

Genetic screens have been extremely powerful tools to identify cancer dependencies and therapeutic targets that can be leveraged in the clinic (Schlabach et al., 2008). However, one of the limitations of single-targeted therapies is the appearance of resistance cancer cells following treatment. This can be due to the acquisition of new mutations as well as rewiring of cellular pathways to compensate for the inhibition of the targeted one. To reduce the development of these resistance mechanisms, combinatorial targeted therapies can be used. Synthetic lethal screens, in which the screen is performed on drug or on a modified cell line, have been widely used to identify lethal combinations in a systematic manner (Luo et al., 2009; Manchado et al., 2016). However, they require one of the targets to be known. To be able to test large number of target combinations simultaneously, the strategy outlined in this thesis relies on expressing pairs of shRNA from the same promoter. By separating both shRNAs by ∼500bp, similar levels of miRNA are generated, and both genes can be knocked down to similar levels. To use these vectors in high-throughput pooled screens, the sequence of both shRNAs needs to be amplified from genomic DNA of pools of cells which was surprisingly challenging. After exploring several alternatives, I had to rely on uniquely barcoding each shRNA and sequencing pairs of barcode to estimate the abundance of the pairs in the pool of cells. This complicates the generation of dual shRNA libraries but greatly simplifies screen sequencing when compared to classic single shRNA screen. As a proof of principle, a screen of ∼10 000 pairs of shRNAs targeting druggable gene was performed in a melanoma cell line. This initial screen yielded over 300 synthetic lethal pairs. To validate the hits from this initial RNAi screen and remove false positives, I performed a secondary CRISPR screen using the vectors described above. Overall, 13 pairs were identified as strong hits in both shRNA and CRISPR screens. Although these hits remain to be validated in one-by-one assays, these results indicate the viability of our approach to identify synthetic lethal pairs of genes on a large-scale.

Although targeting multiple pathways in cancer cells can overcome resistance mechanisms, growing evidence highlights the role of the tumor micro-environment in protecting the tumor from therapy (M.V. Apte et al., 2013). Thus, addressing the specific advantages that stromal cells confer to cancer cells can be another powerful approach. As part of this thesis, we have leveraged RNAi tools we have developed to perform a genome-wide screen in a PDAC-CAFs coculture model. This model recapitulates the stroma-mediated gemcitabine resistance of PDAC that was observed *in vivo*. The hits from this screen are still being validated, but this systematic approach will hopefully provide insights in the pathways involved in the crosstalk between cancer and stromal cells.

Overall, we have combined biological insights, novel expression strategies and empirical efficacy measurements together with machine-learning algorithms to develop potent molecular tools to perform gene perturbation. The tools developed in this thesis allow for efficient knockdown or knockout of genes both in an individual or combinatorial setting. This toolkit provides a resource of arrayed, sequence-verified libraries of constructs targeting any gene in the human genome. It could be used in many areas of biology to uncover genotype to phenotype relationships. In this thesis, I have used these tools to identify combinatorial drug targets in melanoma and investigate stroma-mediated gemcitabine resistance in PDAC. Although the results of these studies need to be further validated, they provide a proof of concept of how these resources can be applied.

Appendix A

A Computational Algorithm to Predict shRNA Potency

Molecular Cell
Resource

A Computational Algorithm to Predict shRNA Potency

Simon R.V. Knott,[1,3] Ashley R. Maceli,[1,3] Nicolas Erard,[1,3] Kenneth Chang,[1,3] Krista Marran,[1] Xin Zhou,[1] Assaf Gordon,[1] Osama El Demerdash,[1] Elvin Wagenblast,[1] Sun Kim,[1] Christof Fellmann,[1,4] and Gregory J. Hannon[1,2,*]

[1]Watson School of Biological Sciences, Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA

[2]Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK

[3]Co-first author

[4]Present address: Mirimus, Inc., 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA

*Correspondence: hannon@cshl.edu

http://dx.doi.org/10.1016/j.molcel.2014.10.025

SUMMARY

The strength of conclusions drawn from RNAi-based studies is heavily influenced by the quality of tools used to elicit knockdown. Prior studies have developed algorithms to design siRNAs. However, to date, no established method has emerged to identify effective shRNAs, which have lower intracellular abundance than transfected siRNAs and undergo additional processing steps. We recently developed a multiplexed assay for identifying potent shRNAs and used this method to generate ~250,000 shRNA efficacy data points. Using these data, we developed shERWOOD, an algorithm capable of predicting, for any shRNA, the likelihood that it will elicit potent target knockdown. Combined with additional shRNA design strategies, shERWOOD allows the ab initio identification of potent shRNAs that specifically target the majority of each gene's multiple transcripts. We validated the performance of our shRNA designs using several orthogonal strategies and constructed genome-wide collections of shRNAs for humans and mice based on our approach.

INTRODUCTION

The discovery of RNAi promised a new era in which the power of genetics could be applied to model organisms for which large-scale studies of gene function were previously inconvenient or impossible (Berns et al., 2004; Brummelkamp et al., 2002; Chuang and Meyerowitz, 2000; Fire et al., 1998; Gupta et al., 2004; Hannon, 2002; Kamath et al., 2003; Kambris et al., 2006; Paddison et al., 2004; Sánchez Alvarado and Newmark, 1999; Svoboda et al., 2000; Timmons and Fire, 1998; Tuschl et al., 1999; Zender et al., 2008). It quickly became clear that implementing RNAi, especially on a genome-wide scale, could be challenging. This was particularly true for applications in mammalian cells in which discrete sequences, in the form of small interfering RNAs (siRNAs) or short hairpin RNAs (shRNAs), were used as silencing triggers (Brummelkamp et al., 2002; Elbashir et al., 2001; Paddison et al., 2002). The overall degree of knockdown achieved was found to vary tremendously depending on the precise sequence of the small RNA that is loaded into the RNAi effector complex (RISC) (Chiu and Rana, 2002; Khvorova et al., 2003; Schwarz et al., 2003). However, the nature of sequence and structural motifs that favor RISC loading and high turnover target cleavage has yet to be fully revealed (Ameres and Zamore, 2013).

Early studies aimed at optimizing RNAi in mammals used endogenous microRNAs as a guide for the design of effective artificial RNAi triggers (Khvorova et al., 2003; Reynolds et al., 2004; Schwarz et al., 2003; Ui-Tei et al., 2004; Zeng and Cullen, 2003). Canonical microRNAs are processed by a two-step nucleolytic mechanism (Seitz and Zamore, 2006). The initial cleavage of the primary microRNA (miRNA) transcript in the nucleus by the microprocessor yields a short, often imperfect hairpin loop, the pre-miRNA (Denli et al., 2004; Lee et al., 2003). This is exported to the cytoplasm, where a second cleavage by Dicer and its associated cofactors yields a short duplex of ~19-20 nucleotides with two nucleotide 3′ overhangs (Bernstein et al., 2001; Grishok et al., 2001; Hutvágner et al., 2001; Ketting et al., 2001; Lund et al., 2004; Yi et al., 2003). This duplex serves as a substrate for preferential loading of one strand into Argonaute proteins in the context of RISC (Hammond et al., 2001; Hutvágner and Zamore, 2002; Khvorova et al., 2003; Martinez et al., 2002; Schwarz et al., 2003).

An examination of the sequences of endogenous miRNAs indicated that thermodynamic asymmetry between the two ends of the short duplex was a strong predictor of which strand would be accepted by Argonaute as the "guide" (Khvorova et al., 2003; Schwarz et al., 2003). Applying this insight to artificial triggers, initially in the form of siRNAs, validated the generality of this observation, and thermodynamic asymmetry became a key guiding principle of both siRNA and shRNA design (Reynolds et al., 2004; Silva et al., 2005). Subsequent studies of the structure of the Ago-small RNA complex have also indicated a sequence preference for a 5′ terminal U that fits into a binding pocket in the mid-domain of the Argonaute protein (Seitz et al., 2008; Wang et al., 2008).

In many ways, siRNAs gain entry into RISC in mammals by simulating the end product of the two-step miRNA processing pathway. shRNAs, which mimic either the primary miRNA or pre-miRNA, must be processed nucleolytically prior to RISC loading (Brummelkamp et al., 2002; Cullen, 2006; Paddison et al., 2002). Therefore, shRNAs are likely subject to additional constraints that lead to efficient recognition by Drosha and Dicer. We do not yet understand the selection rules for effective flux through the miRNA biogenesis pathway and, therefore, cannot predict ab initio which transcripts will produce small RNAs. However, studies of Drosha in particular have implicated patterns of conservation and base pairing in the basal stem, regions adjacent to the Drosha cleavage site, as determinants of efficient pre-miRNA cleavage (Auyeung et al., 2013; Chen et al., 2004; Han et al., 2006; Seitz and Zamore, 2006). Elements within the hairpin loop have also been shown to have an impact on both Drosha efficiency and its site preference (Han et al., 2006; Zhang and Zeng, 2010).

Several attempts have been made to extract predictive rules for the design of effective small RNAs from endpoint silencing data. The first serious attempt applied artificial neural networks to a set of ~2,000 paired data points, associating the sequence of siRNA guides with a corresponding knockdown measurement (established using fluorescent reporters) (Huesken et al., 2005). Experience in the field supported the effectiveness of BIOPREDSi. However, access to the algorithm eventually became impossible. The same data set was subsequently used to produce a second algorithm, Designer of Small Interfering RNA (DSIR), which included additional input variables (the frequency of each nucleotide, each 2-mer, and each 3-mer within the guide) (Vert et al., 2006). To accommodate this large number of parameters, linear modeling was performed using Lasso regression (a form of linear regression that iteratively decreases the use of nonpredictive variables in the linear model) (Tibshirani, 1996).

siRNA design algorithms could be applied for the design of shRNAs, and these did inform the design of genome-wide shRNA collections (Berns et al., 2004; Paddison et al., 2004). However, the prognostic power of siRNA design algorithms is compromised for shRNA design. shRNAs, expressed from RNA polII or polIII promoters, reach lower intracellular concentrations than transfected, synthetic siRNAs (Berns et al., 2004; Paddison et al., 2004). Moreover, shRNAs have additional constraints for effective processing. Therefore, it was imperative that shRNA-specific algorithms be developed.

The generation of accurate siRNA design algorithms was only made possible with the creation of large training data sets. So far, a corresponding shRNA data set has been lacking. Recently, we developed a “sensor” method that allows for the parallel assessment of shRNA potencies on a massive scale (Fellmann et al., 2011). Using the sensor approach, we interrogated ~250,000 shRNAs for their effectiveness in the reporter setting. We used this data set to train a machine learning algorithm for potent shRNA prediction. We tested this algorithm, which we termed, shERWOOD, both at the level of individual shRNAs and at the level of optimized shRNA mini libraries. We demonstrated that, by applying computational shRNA selection in combination with a set of target selection heuristics and an optimized microRNA scaffold, we are able to create highly potent shRNAs. We built upon this result to design and construct next-generation shRNA libraries targeting the constitutive exomes of mice and humans. Predictions for other organisms and custom shRNA designs are also made available via a web-based version of shERWOOD.

RESULTS

Neighboring Positions of the Target Sequence Are Predictive of shRNA Strength

As a prelude to creating an shRNA design algorithm, we first developed a large-scale sensor data set in which shRNA potency was measured and associated with sequence information. To perform the assay, we synthesized 12 sets of ~25,000 constructs that included a doxycycline-inducible shRNA and a GFP-tagged shRNA target sequence located downstream of a constitutive promoter (Fellmann et al., 2011). Libraries were packaged and infected (at single copy) into a reporter cell line. In the absence of doxycycline, GFP was detectable in each cell. However, in the presence of doxycycline, the shRNAs became expressed, and the resultant GFP signal was reduced in a manner proportional to shRNA potency. Using fluorescence-activated cell sorting, cells with low GFP levels, in the presence of drug, were gathered and analyzed via next generation sequencing (NGS) to determine which shRNAs became enriched (i.e., which shRNAs have high potency). Operating iterative cycles of this assay has been shown to identify extremely potent constructs (Fellmann et al., 2011).

We next wished to extract which sequence characteristics were most predictive of shRNA efficacy. This subset of characteristics could then be employed as inputs during machine learning. We first developed a method to consolidate the different sensor data points into a single value for each shRNA (see Supplemental Information available online). These accurately capture the enrichment pattern of individual iterations of the sensor in one single value, therefore allowing downstream machine learning to proceed more easily (Figure 1A). Analysis of the coefficients used to consolidate the sensor data shows that information from the final sensor iteration contributes the most to the final potency value. However, information from the second iteration is also included (Figure S1A).

To distinguish discretely between strong and weak shRNAs, we applied an empirical Bayes moderated t test to the shRNA potency measurements extracted from two biological replicates (Smyth, 2004). Strong and weak shRNAs were those that were enriched or depleted, respectively, with a false discovery rate (FDR) < 0.05.

To test individual nucleotide positions for their predictive capacity, we compared, at each position in the target sequence, each nucleotide’s enrichment and or depletion levels in the potent compared with the weak shRNAs (Figure 1B; Figure S1B; binomial test, FDR < 0.05; Vacic et al., 2006). In general, low GC content is predictive of high efficacy, with the exception of the third nucleotide inside the guide target, which shows a strong selection for cytosine. Also of note is a lack of enrichment for thymidine at the 22nd position of the guide target (corresponding to the first position of the guide). This arose because our input data sets were derived from shRNAs preselected by DSIR.

We next tested whether any pairs of positions had predictive capacity for shRNA strength beyond what was expected based

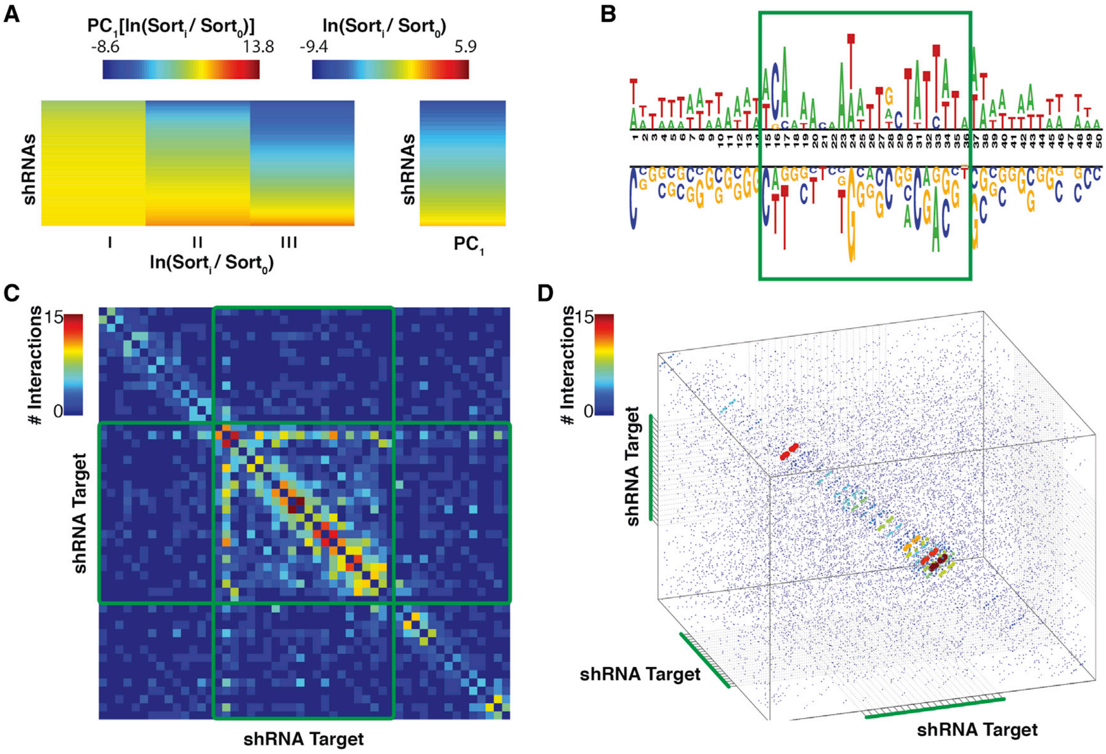


Figure 1. Identification of Sequence Characteristics Predictive of shRNA Efficacy

(A) shRNA score determination via sensor NGS data. On the left is a heatmap representation of normalized shRNA read counts for each on-dox sensor sort. The right panel represents shRNA potencies, calculated by extracting the first principal component of the left panel matrix.

(B) A nucleotide logo representing enriched (top) and depleted (bottom) nucleotides ($p < 0.05$) in potent shRNAs.

(C) A heatmap demonstrating the predictive capacity (with respect to shRNA potency) of each pair of positions within the target region. Heatmap cells are colored to represent the number of nucleotide combinations that were significantly predictive ($p < 0.05$) at each position-pair.

(D) The predictive capacity of each triplet of positions within the target region. Data point colors and sizes represent the number of nucleotide triplets that were significantly predictive ($p < 0.05$) at each position triplet.

on their individual predictive power. To calculate a measurement for each position pair, we applied linear regression to identify synergistic predictive capacity (p value < 0.05; Supplemental Experimental Procedures). Following this, each position pair was assigned a value equal to the sum of nucleotide combinations that were predictive of shRNA potency when assessed at the two positions (Figure 1C; Figure S1C). For a given position within the target, the most predictive partner is the neighboring nucleotide. An exception to this trend is observed in the positions corresponding to the shRNA guide seed, where predictive position pairs are also observed in nucleotides separated by up to four bases.

Finally, we wished to determine whether triplets of positions showed a similar trend to that observed in the pairwise analysis. For this, we performed a modified version of the linear regression tests described above, where triplets instead of pairs of nucleotides were assessed for synergistic predictive capacity. As with the pairwise analysis, neighboring triplets of positions within the target show strong predictive power compared with triplets of nonneighboring positions (Figure 1D). Furthermore, the distance between predictive triplets is also extended slightly in the guide seed region of the shRNA.

A Sensor-Based Computational Algorithm to Predict shRNA Efficacy

Because sequence-based characteristics correlated with shRNA efficiency, we sought to apply machine learning to the sensor-derived efficacy measurements. The goal was to develop a computational algorithm that would predict, for any target sequence, the potency of a corresponding shRNA. We reasoned that the best machine learning tool to apply to this task was random forest regression analysis (Breiman, 2001). The reasons for this decision were two-fold. First, there is no decrease in the accuracy of random forests when the number of input variables is large. Second, the architecture of the algorithm takes into account increases in accuracy that can be achieved by analyzing combinations of input variables.

Our training data set was of two distinct types. One comprised an unbiased set of shRNAs that tiled every nucleotide of nine genes (Fellmann et al., 2011). A second comprised a larger set of shRNAs preselected by the DSIR algorithm (described above). We therefore chose to separate data corresponding to each input class and to train separate forests. We also chose to separate data based on the 5′ nucleotide of the guide. This was done for two reasons. First, previous studies, supported by structural insights, had suggested that the 5′ nucleotide of the guide was a prominent determinant of small RNA potency (Fellmann et al., 2011; Frank et al., 2010; Khvorova et al., 2003; Reynolds et al., 2004). Therefore, training forests individually for shRNAs initiating with each base focused the prediction process on additional determinants. Moreover, the DSIR-based predictions were already heavily biased toward U and A at the 5′ position. In fact, the bias was so strong that we did not have sufficient data to train 5′C and 5′G forests for these data sets. This meant that, in the first pass, we trained six independent modules.

In each module, input data were composed of individual base information as well as all neighboring pairs of bases throughout the guide sequence. In addition, the set of triplet position/nucleotide combinations found to be predictive, as assessed by linear regression, were also included (Figure 1D). After training each of the modules, we sought to determine which input variables were relied most heavily upon. For each module, each variable was permuted across observations, and the resultant reduction in predictive capacity was recorded at each regression tree. The resultant changes were then averaged across trees, and that mean was normalized by their standard deviation. The triplet variables were heavily relied upon (Figure S2A), particularly the triplet corresponding to shRNA guide positions 2–4.

To consolidate these modules, a second-tier random forest was trained using the first-tier outputs, the corresponding shRNA guide base information, and a set of thermodynamic properties extracted from each shRNA (e.g., enthalpy, entropy). We named the compiled algorithm shERWOOD.

To test the prognostic power of shERWOOD, we took advantage of the unbiased nature of the tiled shRNA sensor data. For each of the nine genes represented, we independently trained a shERWOOD algorithm without the data corresponding to that gene. We could then test shERWOOD performance against experimental data in a manner that was not skewed by the use of those data for training. We saw an overall Pearson correlation of 0.72 between experimentally derived potency measurements and computational predictions (Figure 2A). For comparison, DSIR achieves a correlation of 0.4, and a prior shRNA prediction algorithm trained on a subset of the sensor data used in this study achieves 0.56 (Matveeva et al., 2012; Vert et al., 2006). This indicates that shERWOOD achieves a roughly 180% increase in performance over currently existing siRNA prediction algorithms and a 126% increase in efficacy over existing shRNA-specific prediction algorithms.

We supplemented shERWOOD with additional heuristics to maximize the probability of successfully reducing protein levels in most cell and tissue types. The complex nature of alternative splicing patterns provided a strong motivation for directing shRNAs against constitutive exons. We therefore developed a strategy that iteratively searches for regions within a gene that are shared by at least 80% of transcripts (Supplemental Experimental Procedures). This algorithm also tests whether high-potency shRNAs have the potential to cosuppress paralogous genes. Considered together, these strategies have the potential to maximize the probability of biologically meaningful results from studies using shRNAs.

Benchmarking shERWOOD

To assess the performance of the shERWOOD algorithm, we felt that it was necessary to test a large number of shRNAs for their biological effects because one can find anecdotal evidence for excellent performance for nearly any algorithm or strategy. We therefore chose ~2,200 genes based on their enrichment in gene ontology (GO) categories likely to impact the growth and survival of cells in culture (Figure 2B). As controls, particularly for the likelihood of off-target effects, we included 400 olfactory receptor genes. Olfactory receptors are expressed only in olfactory neurons, and even then, they display allelic choice so that only one paralog is expressed per cell. Therefore, shRNAs targeting olfactory receptors are highly unlikely to have relevant, on-target biological effects in any cell line screened in vitro. To benchmark the performance of shERWOOD, we compared a focused mini library predicted with this algorithm to two widely used genome-wide collections, namely The RNAi Consortium (TRC) collection distributed by Sigma-Genosys and the so-called Hannon-Elledge V3 library distributed presently by GE Dharmacon (K.C., unpublished data). To produce the shERWOOD-based library and a deeper simulation of the V3 library, we used either shERWOOD or DSIR to predict their top 10 scoring shRNAs for our test genes. The sequences of TRC shRNAs are listed on a public web portal, and we selected all listed shRNAs for each gene. In the case of TRC shRNAs, it was necessary to adapt them to a 22-base pair stem for placement into the miR-30 context.

For each test library, we synthesized 27,000 oligonucleotides in solid phase on microarrays (Cleary et al., 2004). These were cleaved, amplified, and cloned directly into a miR-30 scaffold within a murine stem cell virus (MSCV)-based retroviral vector without sequence validation. In this arrangement, the primary shRNA was transcribed from the long terminal repeat (LTR) promoter, whereas GFP and Neomycin resistance were expressed separately as a bicistronic transcription unit from the phosphoglycerate kinase promoter (PGK) (Figure S2D). Pilot sequencing showed that each library was of similar quality and representation.

Each library was infected separately into the pancreatic ductal adenocarcinoma cell line A385. Two days after infection, cells were collected for a reference time point, and, after ~12 doublings, cells were again harvested for a final time point (Supplemental Experimental Procedures). shRNA representation was determined following amplification of hairpin inserts from genomic DNA (Sims et al., 2011), and, after processing, shRNA read counts were compared between the initial and final time points (Supplemental Experimental Procedures; Figures S2E–S2G).

To enable direct comparisons between libraries, we censored the shERWOOD- and DSIR-based libraries on a per gene basis to contain the same number of hairpins as were available in the TRC library, keeping those with the best algorithmic scores. We then selected the consensus set of “essential” genes, accepting only those where at least two hairpins in each library passed the statistical threshold (FDR < 0.1). As expected, the resulting set of genes that were important for the growth and survival of A385 was depleted of olfactory receptor shRNAs

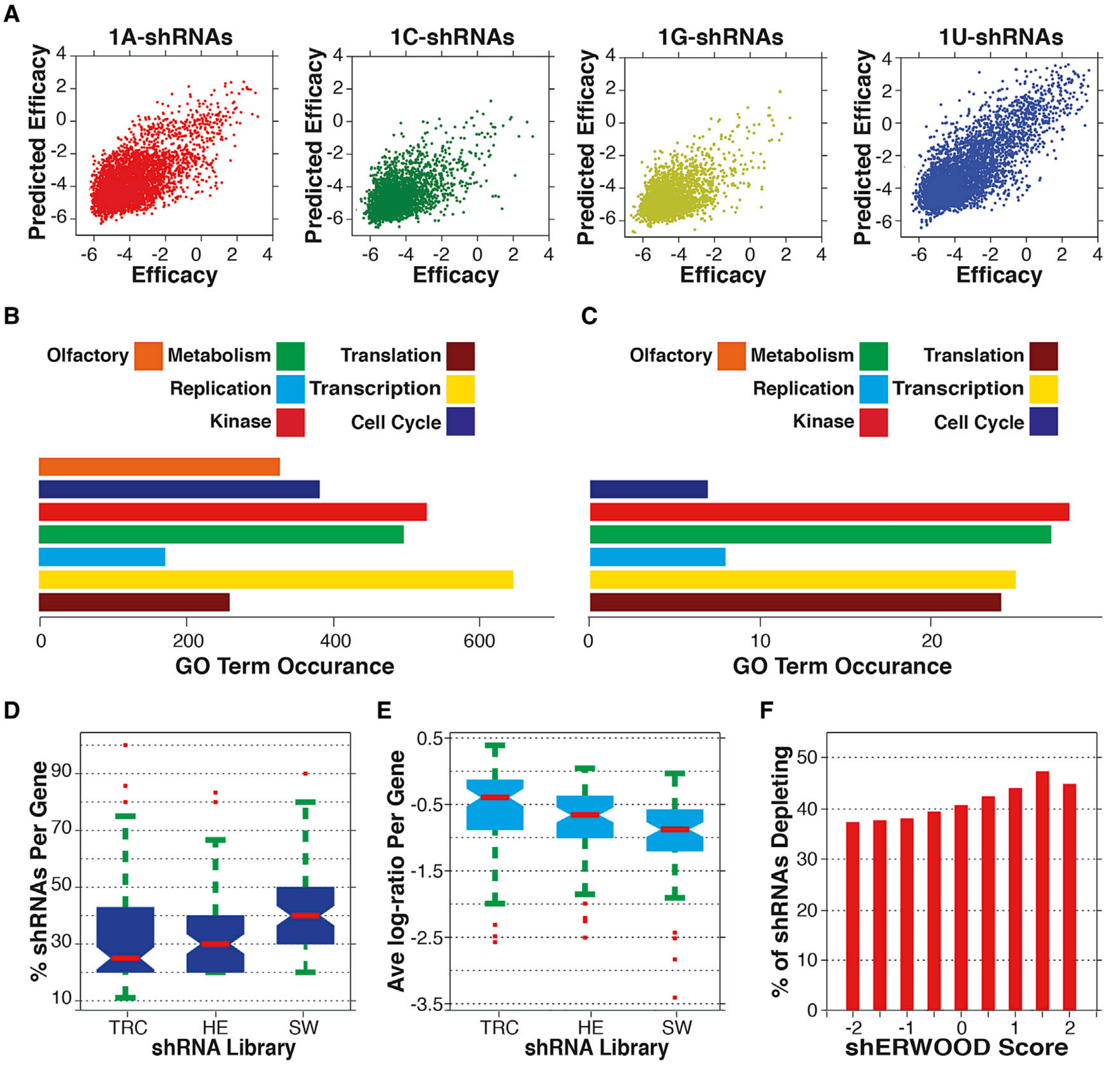


Figure 2. Construction and Validation of an shRNA-Specific Predictive Algorithm

(A) Consolidated cross-validation of predictions versus sensor scores for all shRNAs in the Fellmann et al. (2011) data set (shRNAs are separated by the guide 5′ nucleotide).

(B) GO term instances associated with the targeted gene set selected for shRNA validation screens.

(C) GO term instances associated with genes for which at least two hairpins were significantly depleted in each of the TRC, Hannon-Elledge (HE), and shERWOOD (SW) validation screens.

(D) The percentage of shRNAs targeting consensus-essential genes that were depleted in each of the TRC, HE, and shERWOOD shRNA screens. The plot was made with the Matlab Boxplot function using default parameters. The edges of the box are the 25th and 75th percentiles. The error bars extend to the values q3 + w(q3 − q1) and q1 − w(q3 − q1), where w is 1.5 and q1 and q3 are the 25th and 75th percentiles.

(E) Average log-fold change for shRNAs targeting consensus-essential genes (per gene) for each of the TRC, EH, and shERWOOD validation screens. The plot was made with the Matlab Boxplot function using default parameters. The edges of the box are the 25th and 75th percentiles. The error bars extend to the values q3 + w(q3 − q1) and q1 − w(q3 − q1), where w is 1.5 and q1 and q3 are the 25th and 75th percentiles.

(F) The percentage of shRNAs corresponding to consensus-essential genes that, for any given shERWOOD score, were depleted in the shERWOOD validation screen.

(Figure 2C). In contrast, the set of consensus-essential genes was enriched for GO terms associated with translation.

To benchmark shRNA selection strategies against each other, we determined the percentage of shRNAs in each mini library that scored for each consensus essential gene. For the TRC library, 24% of shRNAs achieved significant depletion, whereas 31% of DSIR-predicted sequences and 40% of shERWOOD-based hairpins scored (Figure 2D). We also considered performance from the perspective of median log-fold depletion. For the TRC collection, the average log-fold change was −0.4. For DSIR, this rose to −0.62, and it increased further to −0.78 for shERWOOD shRNAs (Figure 2E). We note that this type of analysis slightly favors the library with the weakest overall shRNAs because it will be this collection that sets entry criteria for the consensus-essential gene set.

To assess whether shERWOOD scores were a proxy for shRNA potency, we examined the relationship between the shERWOOD score and the probability of being significantly depleted for each consensus-essential gene. For this, we analyzed all ten shERWOOD predictions using a sliding scale of shERWOOD score cutoffs (Figure 2F). As an example, considering shRNAs with a score greater than 0.5, the likelihood that an shRNA will be depleted if it targets one of our consensus essential genes is 42%. Again, this underestimates the information content of shERWOOD scores because, in the cumulative plot shown, the minimum number of scoring hairpins for a given gene, irrespective of scores, is two (i.e., 20%).

Structure-Guided Insights Expand the shRNA Prediction Space

Regardless of the accuracy of predictive models, we sometimes found it difficult to identify potent shRNAs because of search space restrictions imposed by sequence constraints (e.g., GC content), gene length, or the complexity of alternative splicing patterns. We therefore sought ways to expand the sequence space to which we could apply the shERWOOD approach. Analysis of miRNA seed sequences as well as other data have suggested that the first base of the small RNA guide does not pair with its target (Lai, 2002; Lewis et al., 2005; Yuan et al., 2006). Structural studies have supported this hypothesis by showing that the first base of the guide is tightly bound within a pocket in the mid-domain of Ago proteins (Figure S3A; Elkayam et al., 2012; Frank et al., 2010; Nakanishi et al., 2012; Wang et al., 2008). Because the first base of the guide is a strong contributor to shRNA efficacy, we reasoned that we could expand the range of possible effective shRNAs by simply changing the first base of all potential guides to a U, promoting their binding to RISC, and, theoretically, not altering target site choice. We will henceforth refer to this as the 1U strategy. A simulated construction of a human genome-wide shRNA library demonstrates that, when this strategy is implemented, predicted shRNA potencies increase dramatically, particularly for short GC-rich genes (Figure S3B).

To test the 1U strategy in a high-throughput manner, we constructed a sensor library where the top 15 shRNAs targeting a set of ~2,000 “druggable” genes were predicted using the 1U strategy. The constructs were designed so that the shRNAs contained the 1U conversion and the target sites contained the endogenous base. shRNA potencies were extracted as described in Figures 1 and 3A. The distribution indicates that ~50% of the shRNAs were strong or very strong (knockdown efficiency > 75%) based on the scores of control shRNAs that were assayed in parallel. When shRNAs were separated into native and artificial 1U sets and the score distributions were plotted, we were surprised to see a significant reduction in the efficacy of the nonnative 1U shRNAs (Figure 3B; Wilcoxon rank-sum test, p value < 0.01). This was strongly suggesting that RISC interacts not only with the 1U of the guide but also with the first base of the target site.

We therefore stratified 1U shRNAs into four sets based on their endogenous 5′ nucleotide (Figure 3C). This analysis indicated that only a subset of shRNAs performs well when a 1U switch is made (based on the bimodal distributions for endogenous 1A, 1C, and 1G shRNAs) but that the subset that does perform well is predicted to be quite efficacious by the sensor assay. This bimodal distribution is not observed for shERWOOD-selected endogenous 1U shRNAs, and we see that the majority of this shRNA class are efficient.

Given these results, we sought to determine whether we could predict sequences for which a 1U conversion would result in a highly effective shRNA. We fit a Gaussian mixture model to the sensor scores (Figure S3C) and applied this model to assign shRNAs into one of the two resultant populations (Figure S3D). Following clustering, we applied a binomial test separately for shRNAs where the endogenous base was 1A, 1C, 1G, and 1U to determine whether any nucleotides were enriched/depleted in the strong shRNAs with respect to weak shRNAs. All sets show a strong enrichment for U in the target region corresponding to the shRNA guide positions 3, 7, and 8 (Figure 3D). There is also a strong selection for Cs in the target region-corresponding position 19 of the endogenous 1A, 1C, and 1G shRNA guides.

These results prompted us to develop a computational algorithm that could both select the strongest endogenous 1U shRNAs and identify which endogenous 1C, 1G, and 1A shRNAs were likely to yield potent 1U-converted molecules. Data points for which the mixed Gaussian clustering resulted in less than a 70% confidence group assignment were censored (Figure S3E). We trained a random forest using the 22 nucleotides of the endogenous base as well as all neighboring pairs of nucleotides as input and the corresponding 1U conversion sensor scores as output. The algorithm was able to achieve 80% specificity while maintaining 50% sensitivity. Notably, we were able to increase the specificity to 85% through the supplemental application of previously reported rules for shRNA selection (Figure 3E; Fellmann et al., 2011; Matveeva et al., 2012).

To validate this addition to the shERWOOD algorithm, we performed an shRNA screen as described above, in which shRNAs were selected with the 1U strategy with or without applying the additional filter. We also applied this variant of the algorithm to the shRNA screen described in Figure 2. We found that, when additional filters were applied to the 1U strategy, shRNAs targeting our set of consensus-essential genes showed a significantly higher percentage of depleted shRNAs per gene (Wilcoxon rank-sum test, p < 0.01) and a stronger mean depletion, as measured by log ratio (Wilcoxon rank-sum test, p < 0.01; Figure 3F).

A Variant miRNA Scaffold Increases shRNA Potency

Recently completed studies of evolutionarily conserved determinants of Drosha processing raised the possibility that the placement of the EcoRI site in the standard miR-30 scaffold might have reduced the efficiency of pre-miRNA cleavage (Auyeung et al., 2013). Others have reported that alternatively positioning the EcoRI site within the scaffold increases small RNA levels, presumably by improving biogenesis. This led to overall more potent knockdown (Fellmann et al., 2013). We therefore chose to create shRNAs by Gibson assembly, removing restriction sites altogether from the shRNA scaffold (Figure S4). We felt that this was the surest way to avoid any unanticipated effects of altering processing signals. We termed this scaffold ultramiR.

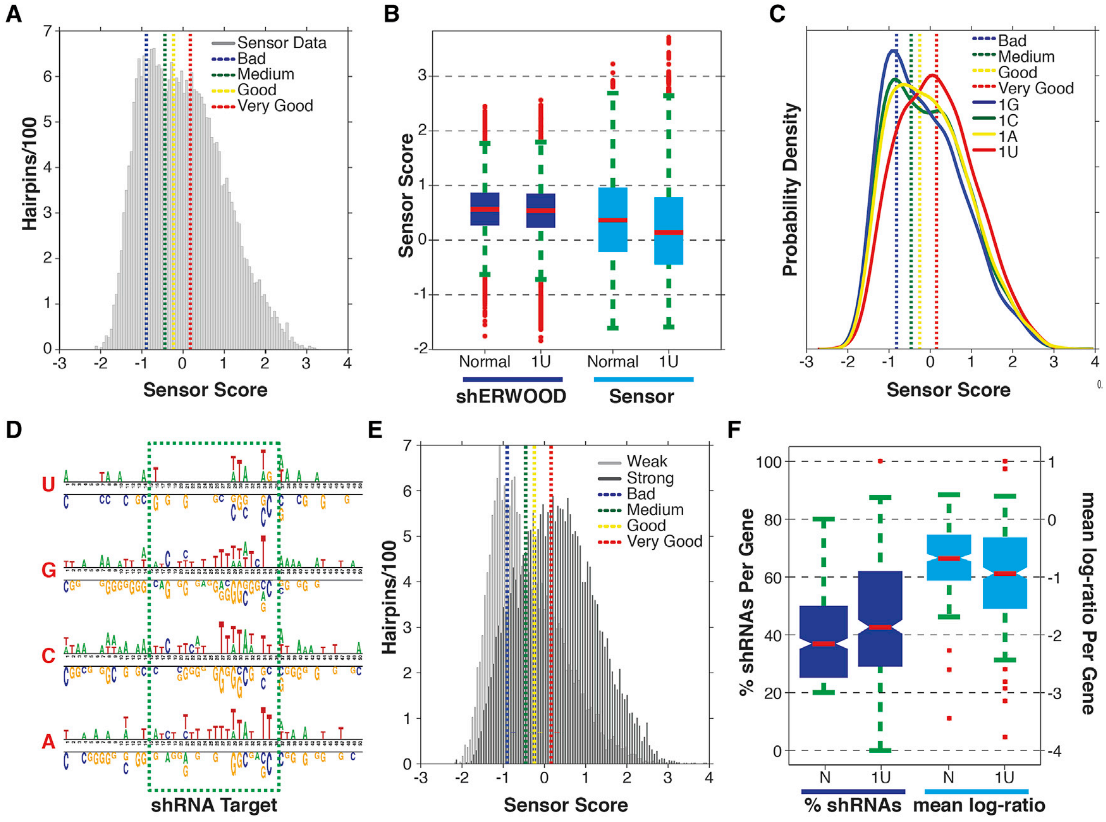


Figure 3. Structure-Guided Maximization of shRNA Prediction Space

(A) Histogram of sensor scores for the top 15 shRNAs as identified by the shERWOOD-1U strategy, targeting ~2000 druggable genes. Overlaid are the mean sensor scores for control shRNAs representing poor, medium, potent, and very potent shRNAs (with mean knockdown efficiencies of 25%, 50%, 75%, and >90%, respectively).

(B) The distribution of shERWOOD-1U prediction scores for shRNAs where endogenous 1U shRNAs are separated from endogenous non-1U shRNAs. Sensor scores for endogenous 1U and non-1U shRNAs are displayed on the left. The plot was made with the Matlab Boxplot function using default parameters. The edges of the box are the 25th and 75th percentiles. The error bars extend to the values q3 + w(q3 − q1) and q1 − w(q3 − q1), where w is 1.5 and q1 and q3 are the 25th and 75th percentiles.

(C) Distribution of sensor scores for shERWOOD-1U-selected shRNAs, separated by endogenous guide 5′ nucleotides.

(D) A nucleotide logo representing enriched (top) and depleted (bottom) nucleotides ($p < 0.05$) in potent shERWOOD-1U-selected shRNAs (separated by endogenous guide 5′ nucleotides).

(E) The distribution of sensor scores for shRNAs classified as weak and potent by a random forest classifier trained on the shERWOO-1U sensor data.

(F) The distributions of the percentage of shERWOOD- and shERWOOD-1U-selected shRNAs targeting consensus-essential genes that were depleted in validation screens (left). In addition, normalized log-fold changes of shRNAs, identified under each selection scheme, are displayed (right). The plot was made with the Matlab Boxplot function using default parameters. The edges of the box are the 25th and 75th percentiles. The error bars extend to the values q3 + w(q3 − q1) and q1 − w(q3 − q1), where w is 1.5 and q1 and q3 are the 25th and 75th percentiles.

To test ultramiR performance, we inserted two shRNAs, targeting luciferase or mouse RPA3, into the standard scaffold and into ultramiR. These constructs were packaged and infected in duplicate (multiplicity of infection [MOI] < 0.3) into human embryonic kidney 293T (HEK293T) cells and the modified DF1 reporter line used for the sensor screen, respectively (Fellmann et al., 2011). Following selection for singly infected cells, we analyzed the levels of mature shRNAs by small RNA sequencing (Malone et al., 2012). shRNA guide counts were normalized across libraries by determining their log-fold enrichment relative to the 66th quantile of endogenous microRNA levels. A comparison of the normalized shRNA values indicated that, when shRNAs were placed into the ultramiR scaffold, mature small RNA levels were increased significantly relative to levels observed using the standard miR-30 scaffold (Figure 4A). Notably, the performance of ultramiR and the previously described alternate scaffold, miR-E, were indistinguishable (data not shown).

To provide a more rigorous test of ultramiR performance, we created a variant of the shERWOOD-selected 1U strategy shRNA library and compared its performance to that of the same library in the standard scaffold. Considering the consensus-essential gene set, over half of all shRNAs in the library were depleted significantly (Figure 4B). This substantial improvement (from 42% to 51%, Wilcoxon rank-sum test, $p < 0.01$) was accompanied by a greater degree of mean log-fold

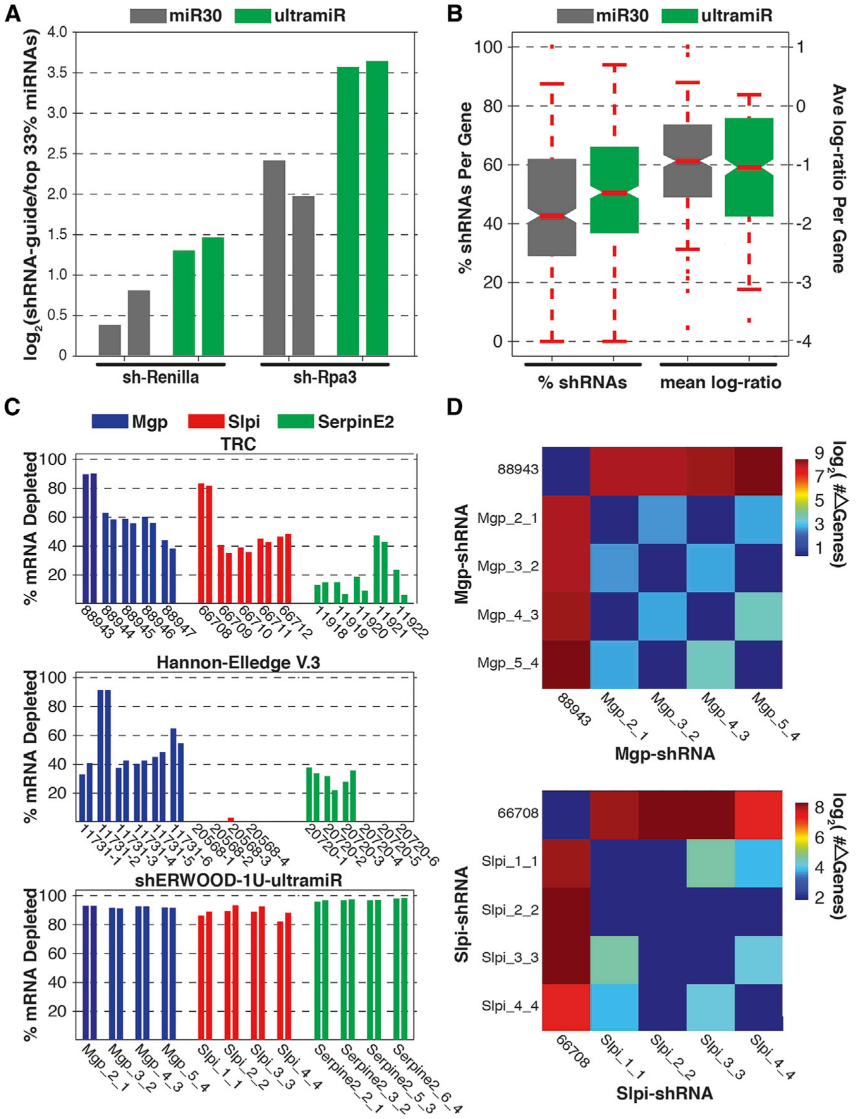


Figure 4. Validation of an Alternative Mir Scaffold

(A) Relative abundances of processed guide sequences for two shRNAs (as determined via small RNA cloning and NGS analysis) when cloned into traditional miR30 and ultramiR scaffolds. Values represent the log-fold enrichment of shRNA guides with respect to sequences corresponding to the ten most abundant microRNAs.

(B) Distributions of the percentage of shHERWOOD-1U-selected shRNAs targeting consensus-essential genes that were depleted in validation screens when shRNAs were placed into miR30 and ultramiR scaffolds. Log-fold changes for the same constructs are displayed on the left. The plot was made with the Matlab Boxplot function using default parameters. The edges of the box are the 25th and 75th percentiles. The error bars extend to the values q3 + w(q3 − q1) and q1 − w(q3 − q1), where w is 1.5 and q1 and q3 are the 25th and 75th percentiles.

(C) Knockdown efficiencies for shRNAs targeting the mouse genes Mgp, Slpi, and Mgp. shRNAs assessed were those contained within the TRC collection, those initially designed for the Hannon-Elledge V.3 library, and those designed using the current strategies. The TRC and Hannon-Elledge V.3 shRNAs are housed within each library's lentiviral vectors, whereas the shERWOOD-1U-selected shRNAs are housed within an ultramiR scaffold in a retroviral vector. Ultramir is constitutively expressed from the LTR.

(D) The number of differentially expressed genes (>2-fold change and FDR < 0.05) identified through pairwise comparisons of the cell lines corresponding to Mgp and Slpi knockdown by the shERWOOD-1U-selected shRNAs and the TRC shRNAs 88943 and 66708.

depletion for each construct (from −0.95 to −1.05, Wilcoxon rank-sum test, $p < 0.01$).

We also tested a limited number of individual shRNAs for their potency by measuring reductions in target mRNA levels. We selected the four shRNAs with the highest shERWOOD scores for mouse Mgp, Serpine2, and Slpi. These were cloned into an MSCV-based ultramiR vector in which hygromycin resistance and mCherry were also expressed as a bicistronic transcript from the PGK promoter. We also chose to compare these shRNAs to those developed using previous library construction strategies. For this, we obtained the current TRC (five shRNAs per gene) and V.3 Hannon-Elledge (six shRNAs per gene) library constructs targeting these genes. For the Hannon-Elledge library, because there were not four precloned shRNAs for each gene, we assembled the remaining shRNAs that were designed as part of that library but never constructed. We failed to clone two constructs (both targeting Slpi) after multiple attempts, meaning that only four V3 constructs were tested for that gene. Mouse 4T1 cells were infected at single copy, and knockdown was tested following selection of infected cells. The TRC library is carried within a vector lacking a fluorescent marker. We therefore calibrated infection levels to achieve single copy by comparison with parallel infections and selections with V3 constructs. The knockdown efficiency of each shRNA was assessed by comparing transcript levels (via quantitative PCR) to those in cells infected with corresponding empty vectors. The TRC shRNAs showed modest knockdown in most cases, with only two shRNAs showing more than 80% of transcript reduction (88943 and 66708, Figure 4C). The Hannon-Elledge V.3 shRNAs produced relatively modest levels of knockdown. In comparison the majority of shRNAs designed using the strategies outlined here reduced target mRNA levels by over 80%, with most reducing target mRNA levels by more than 90% (Figure 4D). Considered together, our data indicate that the combined use of shERWOOD and the ultramiR scaffold consistently produces highly potent shRNAs.

To assess the specificity of shRNA knockdown, we performed RNA sequencing (RNA-seq) on all cell lines expressing shERWOOD-ultramiR shRNAs targeting Slpi and Mgp and the two cell lines harboring TRC constructs 88943 and 66708, which target Mgp and Slpi, respectively. Even in the absence of off-target effects, the silencing of a gene through RNAi will likely elicit biological effects that result in changes in the abundance of other mRNAs. Unlike so-called "off-target" effects, phenotypic effects that emanate from on-target silencing should be consistent for all efficacious shRNAs. Therefore, by comparing the expression profiles of cells harboring different shRNAs corresponding to a single gene, one should be able to infer the scope of off-target effects for each construct. The shRNAs that show the greatest propensity to off-targets will be those that create expression profiles most dissimilar to the mean profile.

When either Mgp or Slpi were silenced using the strategies outline here, the expression profiles in the resultant lines were found to be highly similar. Less than 25 genes were altered in their expression (DESeq, fold change > 2 and FDR < 0.05) between any pair of corresponding lines. However, when these were compared with lines that had Mgp or Slpi silenced using potent TRC constructs, a significant difference in expression profiles was observed. Over 500 genes are altered in the line where Mgp has been silenced using the TRC constructs, and approximately 250 are altered in the line expressing the TRC Slpi-shRNA (Figure 4D).

These results could reflect our current strategies for reducing off-targeting or our use of a microRNA-based scaffold. Recently, others have observed strong phenotypic changes, related to microRNA dysregulation, when U6 driven stem-loop shRNAs were expressed in cells where the target gene had been deleted (Baek et al., 2014). In contrast, when these same shRNAs were expressed from a microRNA scaffold, the phenotype was not observed. Overall, the aforementioned analysis indicates that shRNAs produced using the strategies outlined in this report, when expressed in an ultramiR scaffold, show strong knockdown capacity and limited off-target effects.

DISCUSSION

The application of RNAi in mammalian cells promised a revolution in understanding gene function and in the discovery and validation of therapeutic targets. Although the impact of RNAi has been enormous, there has also been substantial frustration in attempts to fully realize the potential of this technology. Many different sequences often need to be tested to obtain one that potently suppresses expression, a problem that is particularly acute with shRNAs expressed from single-copy transgenes. This, and the resulting variability in the quality of publicly available genome-wide shRNA collections, has caused consternation, particularly when very similar shRNA screens carried out by different investigators yield largely nonoverlapping results (Babij et al., 2011; Luo et al., 2009; Scholl et al., 2009). We tried to address problems with current shRNA technologies by optimizing target sequence choice and small RNA production.

We leveraged our prior development of a high-throughput assay for testing shRNA potency to develop a computational algorithm capable of accurately predicting the outcome of the sensor screen and, in turn, predicting potentially potent shRNAs. Through iterative cycles of training and refinement we produced a tool that permits highly efficacious shRNAs to be generated for nearly any gene.

We validated the performance of our approach and benchmarked it against current tools using nonsequence verified, focused shRNA libraries. Based on our analyses, we can now generate shRNA libraries where nearly 60% of all hairpins targeting essential genes are strongly depleted in multiplexed screens. This means that, for any library containing, on average, four hairpins per gene, most bona fide hits will be identified by multiple hairpins, greatly reducing the probability of false-positive calls. Because our libraries were used in their raw forms, we feel that this is a lower boundary of performance because sequence-validated and arrayed collections will not contain a mixture of shRNA variants generated by synthesis and PCR errors.

Given the promise of our approach, we have undertaken the construction of fourth- and fifth-generation sequence-verified shRNA libraries targeting the mouse and human genomes. The fourth generation toolkit takes advantage of shERWOOD in a canonical miR-30 scaffold and currently comprises over 75,000 shRNAs targeting human genes and 40,000 shRNAs targeting mouse genes. The fifth-generation toolkit places shERWOOD shRNAs in the ultramiR scaffold and is presently ~50% complete.

We have predicted shERWOOD shRNAs targeting constitutive exons of annotated human, mouse, and rat protein coding genes, and these are available via a web portal (http://sherwood.cshl.edu:8080/sherwood/). We have additionally made shERWOOD available as a web-based tool for custom shRNA prediction, for example for the design of shRNAs for other model organisms or for specific mRNA isoforms or noncoding RNAs.

Overall, we feel that the combination of improvements to shRNA technologies described herein creates a next-generation RNAi toolkit that will produce more reliable outcomes for investigators, whether applied on a gene-by-gene basis or in the context of unbiased, genome-wide screens.

EXPERIMENTAL PROCEDURES

Cell Lines

The sensor algorithm was performed using Eco-rtTA-chicken (ERC) cells (derived from DF-1 chicken embryonic fibroblasts) (Fellmann et al., 2011). All shRNA screens were performed in the pancreatic adenocarcinoma cell line A385 (Cui et al., 2012). Small RNA analysis for RPA2 shRNAs was performed in the ERC cell line (Fellmann et al., 2011) and in HEK293T cells for the *Renilla* shRNAs. Individual shRNA knockdown experiments were performed in the 4T1 murine mammary cancer cell line (Dexter et al., 1978).

Vectors

All RNAi screens and small RNA cloning experiments were performed with an MSCV-based retroviral vector harboring a bicistronic transcript (eGFP-IRES-Neomycin) downstream of the PGK promoter (Figure S2D). Single-target knockdown experiments for shERWOOD-ultramiR shRNAs were performed with a similar vector, where Neomycin is replaced with Hygromycin, and enhanced GFP is replaced with mCHERRY. Single-target knockdown experiments for the Hannon-Elledge V3 and TRC shRNAs were performed with the GIPZ and pLKO.1 vectors, respectively (GE Dharmacon).

shRNA Library Construction

To ensure high-complexity end products, all shRNA libraries were amplified from raw chip material using 16 separate reactions with 22 PCR cycles. For each reaction, 1 μl of 100 μM chip material was used. All transformations were performed with Invitrogen's MegaX DH10B T1 electrocompetent cells using a Bio-Rad Gene Pulser Xcell and Bio-Rad Gene Pulser 1 mm cuvettes for electroporation. For each library, a minimum of 25 M successfully transformed cells were obtained.

shRNA Library Screening

shRNA libraries were packaged using the Platinum-A retrovirus packaging cell (Cell Biolabs). Cells were cotransfected with glycoprotein G of the vesicular stomatitis virus and siRNAs targeting the shRNA processing protein Pasha (QIAGEN). Viral infections were performed at an MOI of 0.3 to ensure a maximum of one shRNA infection per cell. shRNA representation in the infected cell population was maintained at a minimum of 1,000 infected cells per shRNA on each passage. All screens were performed in triplicate. Two days after infection, cells were collected for a reference time point, and, after ~12 doublings, cells were again harvested for a final time point. Neomycin selection began after the initial time point and continued throughout the screens.

shRNA Library Processing and Analysis

Following cell harvests, DNA was extracted with the QIAGEN QIAamp DNA Blood Maxi kit. For each sample, shRNA molecules were extracted from genomic DNA in 96 separate 25-cycle PCR reactions where 2 μg of input DNA was included in each reaction. Following this initial PCR, Illumina adapters were added via PCR, and samples were processed on the Illumina Hi-Seq-2.0 platform (read depth was maintained at ~1,000 short reads per shRNA). Following sequencing, shRNA counts were extracted with the bowtie algorithm (allowing zero mismatches) and normalized by their total counts. Log-fold changes demonstrated a GC bias in the control shRNA population (Figure S2E). To remove this bias, a 1° polynomial was fit to each screen replicate's log-fold change versus GC content data, and this curve was then subtracted from each data point (Figure S2F). Following this, values were further normalized so that the control population had a population variance of one. shRNAs were classified as depleted with an FDR cutoff of 0.1 using an empirical Bayes moderated test (Figure S2G; Smyth, 2004).

For further details, see the Supplemental Experimental Procedures.

ACCESSION NUMBERS

All raw and processed data are available through the National Center for Biotechnology Information under the accession number GSE62189.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures and four figures and can be found with this article online at http://dx.doi.org/10.1016/j.molcel.2014.10.025.

AUTHOR CONTRIBUTIONS

S.R.V.K. and G.J.H. designed the experiments and wrote the manuscript. S.R.V.K. designed the algorithm. A.M. performed all shRNA screens. A.M. and X.Z. performed the 1U sensor experiments. N.E. performed all small RNA cloning. S.R.V.K., A.M., and N.E. constructed the sequence-verified libraries. K.C. and K.M. performed the DSIR –sensor experiments. S.R.V.K. and A.G. developed and implemented the exon inclusion and off-target minimization strategies. A.G. and O.E.D. designed the shERWOOD website. E.W. and S.K. performed the individual knockdown experiments. S.R.V.K. and S.K. performed the RNA-seq experiments.

ACKNOWLEDGMENTS

This work was supported by the Howard Hughes Medical Institute, by grants from the NIH (to G.J.H.), and by a gift from Kathryn W. Davis. S.R.V.K. is supported by a fellowship from The Hope Funds for Cancer Research. E.W. is supported by a fellowship from the Boehringer Ingelheim Foundation. We thank all members of the Hannon laboratory for help with initial library construction. We would also like to thank Blake Simmons (Transomics Technologies) for help with constructing sequence-verified versions of the libraries. Finally, we thank Vasily Vagin for helpful comments on this manuscript. S.R.V.K. and G.H. are associated with Transomic Technologies, who have commercialized libraries constructed using the shERWOOD and ultramiR design strategies.

Received: April 3, 2014
Revised: September 11, 2014
Accepted: October 23, 2014
Published: November 26, 2014

REFERENCES

Ameres, S.L., and Zamore, P.D. (2013). Diversifying microRNA sequence and function. Nat. Rev. Mol. Cell Biol. *14*, 475–488.

Auyeung, V.C., Ulitsky, I., McGeary, S.E., and Bartel, D.P. (2013). Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing. Cell *152*, 844–858.

Babij, C., Zhang, Y., Kurzeja, R.J., Munzli, A., Shehabeldin, A., Fernando, M., Quon, K., Kassner, P.D., Ruefli-Brasse, A.A., Watson, V.J., et al. (2011). STK33 kinase activity is nonessential in KRAS-dependent cancer cells. Cancer Res. *71*, 5818–5826.

Baek, S.T., Kerjan, G., Bielas, S.L., Lee, J.E., Fenstermaker, A.G., Novarino, G., and Gleeson, J.G. (2014). Off-target effect of doublecortin family shRNA on neuronal migration associated with endogenous microRNA dysregulation. Neuron *82*, 1255–1262.

Berns, K., Hijmans, E.M., Mullenders, J., Brummelkamp, T.R., Velds, A., Heimerikx, M., Kerkhoven, R.M., Madiredjo, M., Nijkamp, W., Weigelt, B., et al. (2004). A large-scale RNAi screen in human cells identifies new components of the p53 pathway. Nature *428*, 431–437.

Bernstein, E., Caudy, A.A., Hammond, S.M., and Hannon, G.J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature *409*, 363–366.

Breiman, L. (2001). Random forests. Machine Learning *45*, 5–32.

Brummelkamp, T.R., Bernards, R., and Agami, R. (2002). A system for stable expression of short interfering RNAs in mammalian cells. Science *296*, 550–553.

Chen, C.Z., Li, L., Lodish, H.F., and Bartel, D.P. (2004). MicroRNAs modulate hematopoietic lineage differentiation. Science *303*, 83–86.

Chiu, Y.L., and Rana, T.M. (2002). RNAi in human cells: basic structural and functional features of small interfering RNA. Mol. Cell *10*, 549–561.

Chuang, C.F., and Meyerowitz, E.M. (2000). Specific and heritable genetic interference by double-stranded RNA in Arabidopsis thaliana. Proc. Natl. Acad. Sci. USA *97*, 4985–4990.

Cleary, M.A., Kilian, K., Wang, Y., Bradshaw, J., Cavet, G., Ge, W., Kulkarni, A., Paddison, P.J., Chang, K., Sheth, N., et al. (2004). Production of complex nucleic acid libraries using highly parallel in situ oligonucleotide synthesis. Nat. Methods *1*, 241–248.

Cui, Y., Brosnan, J.A., Blackford, A.L., Sur, S., Hruban, R.H., Kinzler, K.W., Vogelstein, B., Maitra, A., Diaz, L.A., Jr., Iacobuzio-Donahue, C.A., et al. (2012). Genetically defined subsets of human pancreatic cancer show unique in vitro chemosensitivity. Clinical cancer research *18*, 6519–6530.

Cullen, B.R. (2006). Induction of stable RNA interference in mammalian cells. Gene Ther. *13*, 503–508.

Denli, A.M., Tops, B.B., Plasterk, R.H., Ketting, R.F., and Hannon, G.J. (2004). Processing of primary microRNAs by the Microprocessor complex. Nature *432*, 231–235.

Dexter, D.L., Kowalski, H.M., Blazar, B.A., Fligiel, Z., Vogel, R., and Heppner, G.H. (1978). Heterogeneity of tumor cells from a single mouse mammary tumor. Cancer Res. *38*, 3174–3181.

Elbashir, S.M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., and Tuschl, T. (2001). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. Nature *411*, 494–498.

Elkayam, E., Kuhn, C.D., Tocilj, A., Haase, A.D., Greene, E.M., Hannon, G.J., and Joshua-Tor, L. (2012). The structure of human argonaute-2 in complex with miR-20a. Cell *150*, 100–110.

Fellmann, C., Zuber, J., McJunkin, K., Chang, K., Malone, C.D., Dickins, R.A., Xu, Q., Hengartner, M.O., Elledge, S.J., Hannon, G.J., and Lowe, S.W. (2011). Functional identification of optimized RNAi triggers using a massively parallel sensor assay. Mol. Cell *41*, 733–746.

Fellmann, C., Hoffmann, T., Sridhar, V., Hopfgartner, B., Muhar, M., Roth, M., Lai, D.Y., Barbosa, I.A., Kwon, J.S., Guan, Y., et al. (2013). An optimized microRNA backbone for effective single-copy RNAi. Cell Reports *5*, 1704–1713.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature *391*, 806–811.

Frank, F., Sonenberg, N., and Nagar, B. (2010). Structural basis for 5′-nucleotide base-specific recognition of guide RNA by human AGO2. Nature *465*, 818–822.

Grishok, A., Pasquinelli, A.E., Conte, D., Li, N., Parrish, S., Ha, I., Baillie, D.L., Fire, A., Ruvkun, G., and Mello, C.C. (2001). Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. Cell *106*, 23–34.

Gupta, S., Schoer, R.A., Egan, J.E., Hannon, G.J., and Mittal, V. (2004). Inducible, reversible, and stable RNA interference in mammalian cells. Proc. Natl. Acad. Sci. USA *101*, 1927–1932.

Hammond, S.M., Boettcher, S., Caudy, A.A., Kobayashi, R., and Hannon, G.J. (2001). Argonaute2, a link between genetic and biochemical analyses of RNAi. Science *293*, 1146–1150.

Han, J., Lee, Y., Yeom, K.H., Nam, J.W., Heo, I., Rhee, J.K., Sohn, S.Y., Cho, Y., Zhang, B.T., and Kim, V.N. (2006). Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. Cell *125*, 887–901.

Hannon, G.J. (2002). RNA interference. Nature *418*, 244–251.

Huesken, D., Lange, J., Mickanin, C., Weiler, J., Asselbergs, F., Warner, J., Meloon, B., Engel, S., Rosenberg, A., Cohen, D., et al. (2005). Design of a genome-wide siRNA library using an artificial neural network. Nat. Biotechnol. *23*, 995–1001.

Hutvágner, G., and Zamore, P.D. (2002). A microRNA in a multiple-turnover RNAi enzyme complex. Science *297*, 2056–2060.

Hutvágner, G., McLachlan, J., Pasquinelli, A.E., Bálint, E., Tuschl, T., and Zamore, P.D. (2001). A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. Science *293*, 834–838.

Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., et al. (2003). Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature *421*, 231–237.

Kambris, Z., Brun, S., Jang, I.H., Nam, H.J., Romeo, Y., Takahashi, K., Lee, W.J., Ueda, R., and Lemaitre, B. (2006). Drosophila immunity: a large-scale in vivo RNAi screen identifies five serine proteases required for Toll activation. Current biology: CB *16*, 808–813.

Ketting, R.F., Fischer, S.E., Bernstein, E., Sijen, T., Hannon, G.J., and Plasterk, R.H. (2001). Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. Genes Dev. *15*, 2654–2659.

Khvorova, A., Reynolds, A., and Jayasena, S.D. (2003). Functional siRNAs and miRNAs exhibit strand bias. Cell *115*, 209–216.

Lai, E.C. (2002). Micro RNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation. Nat. Genet. *30*, 363–364.

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Rådmark, O., Kim, S., and Kim, V.N. (2003). The nuclear RNase III Drosha initiates microRNA processing. Nature *425*, 415–419.

Lewis, B.P., Burge, C.B., and Bartel, D.P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell *120*, 15–20.

Lund, E., Güttinger, S., Calado, A., Dahlberg, J.E., and Kutay, U. (2004). Nuclear export of microRNA precursors. Science *303*, 95–98.

Luo, J., Emanuele, M.J., Li, D., Creighton, C.J., Schlabach, M.R., Westbrook, T.F., Wong, K.K., and Elledge, S.J. (2009). A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene. Cell *137*, 835–848.

Malone, C., Brennecke, J., Czech, B., Aravin, A., and Hannon, G.J. (2012). Preparation of small RNA libraries for high-throughput sequencing. Cold Spring Harbor protocols *2012*, 1067–1077.

Martinez, J., Patkaniowska, A., Urlaub, H., Lührmann, R., and Tuschl, T. (2002). Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. Cell *110*, 563–574.

Matveeva, O.V., Nazipova, N.N., Ogurtsov, A.Y., and Shabalina, S.A. (2012). Optimized models for design of efficient miR30-based shRNAs. Front. Genet. *3*, 163.

Nakanishi, K., Weinberg, D.E., Bartel, D.P., and Patel, D.J. (2012). Structure of yeast Argonaute with guide RNA. Nature *486*, 368–374.

Paddison, P.J., Caudy, A.A., Bernstein, E., Hannon, G.J., and Conklin, D.S. (2002). Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells. Genes Dev. *16*, 948–958.

Paddison, P.J., Silva, J.M., Conklin, D.S., Schlabach, M., Li, M., Aruleba, S., Balija, V., O'Shaughnessy, A., Gnoj, L., Scobie, K., et al. (2004). A resource for large-scale RNA-interference-based screens in mammals. Nature *428*, 427–431.

Reynolds, A., Leake, D., Boese, Q., Scaringe, S., Marshall, W.S., and Khvorova, A. (2004). Rational siRNA design for RNA interference. Nat. Biotechnol. *22*, 326–330.

Sánchez Alvarado, A., and Newmark, P.A. (1999). Double-stranded RNA specifically disrupts gene expression during planarian regeneration. Proc. Natl. Acad. Sci. USA *96*, 5049–5054.

Scholl, C., Fröhling, S., Dunn, I.F., Schinzel, A.C., Barbie, D.A., Kim, S.Y., Silver, S.J., Tamayo, P., Wadlow, R.C., Ramaswamy, S., et al. (2009). Synthetic lethal interaction between oncogenic KRAS dependency and STK33 suppression in human cancer cells. Cell *137*, 821–834.

Schwarz, D.S., Hutvágner, G., Du, T., Xu, Z., Aronin, N., and Zamore, P.D. (2003). Asymmetry in the assembly of the RNAi enzyme complex. Cell *115*, 199–208.

Seitz, H., and Zamore, P.D. (2006). Rethinking the microprocessor. Cell *125*, 827–829.

Seitz, H., Ghildiyal, M., and Zamore, P.D. (2008). Argonaute loading improves the 5′ precision of both MicroRNAs and their miRNA* strands in flies. Current biology: CB *18*, 147–151.

Silva, J.M., Li, M.Z., Chang, K., Ge, W., Golding, M.C., Rickles, R.J., Siolas, D., Hu, G., Paddison, P.J., Schlabach, M.R., et al. (2005). Second-generation shRNA libraries covering the mouse and human genomes. Nat. Genet. *37*, 1281–1288.

Sims, D., Mendes-Pereira, A.M., Frankum, J., Burgess, D., Cerone, M.A., Lombardelli, C., Mitsopoulos, C., Hakas, J., Murugaesu, N., Isacke, C.M., et al. (2011). High-throughput RNA interference screening using pooled shRNA libraries and next generation sequencing. Genome Biol. *12*, R104.

Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology *3*, Article3.

Svoboda, P., Stein, P., Hayashi, H., and Schultz, R.M. (2000). Selective reduction of dormant maternal mRNAs in mouse oocytes by RNA interference. Development *127*, 4147–4156.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. J. R. Stat. Soc. Ser. A Stat. Soc. *58*, 267–288.

Timmons, L., and Fire, A. (1998). Specific interference by ingested dsRNA. Nature *395*, 854.

Tuschl, T., Zamore, P.D., Lehmann, R., Bartel, D.P., and Sharp, P.A. (1999). Targeted mRNA degradation by double-stranded RNA in vitro. Genes Dev. *13*, 3191–3197.

Ui-Tei, K., Naito, Y., Takahashi, F., Haraguchi, T., Ohki-Hamazaki, H., Juni, A., Ueda, R., and Saigo, K. (2004). Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference. Nucleic Acids Res. *32*, 936–948.

Vacic, V., Iakoucheva, L.M., and Radivojac, P. (2006). Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. Bioinformatics *22*, 1536–1537.

Vert, J.P., Foveau, N., Lajaunie, C., and Vandenbrouck, Y. (2006). An accurate and interpretable model for siRNA efficacy prediction. BMC Bioinformatics *7*, 520.

Wang, Y., Sheng, G., Juranek, S., Tuschl, T., and Patel, D.J. (2008). Structure of the guide-strand-containing argonaute silencing complex. Nature *456*, 209–213.

Yi, R., Qin, Y., Macara, I.G., and Cullen, B.R. (2003). Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes Dev. *17*, 3011–3016.

Yuan, Y.R., Pei, Y., Chen, H.Y., Tuschl, T., and Patel, D.J. (2006). A potential protein-RNA recognition event along the RISC-loading pathway from the structure of A. aeolicus Argonaute with externally bound siRNA. Structure (London, England: 1993) *14*, 1557–1565.

Zender, L., Xue, W., Zuber, J., Semighini, C.P., Krasnitz, A., Ma, B., Zender, P., Kubicka, S., Luk, J.M., Schirmacher, P., et al. (2008). An oncogenomics-based in vivo RNAi screen identifies tumor suppressors in liver cancer. Cell *135*, 852–864.

Zeng, Y., and Cullen, B.R. (2003). Sequence requirements for micro RNA processing and function in human cells. RNA *9*, 112–123.

Zhang, X., and Zeng, Y. (2010). The terminal loop region controls microRNA processing by Drosha and Dicer. Nucleic Acids Res. *38*, 7689–7697.

Appendix B

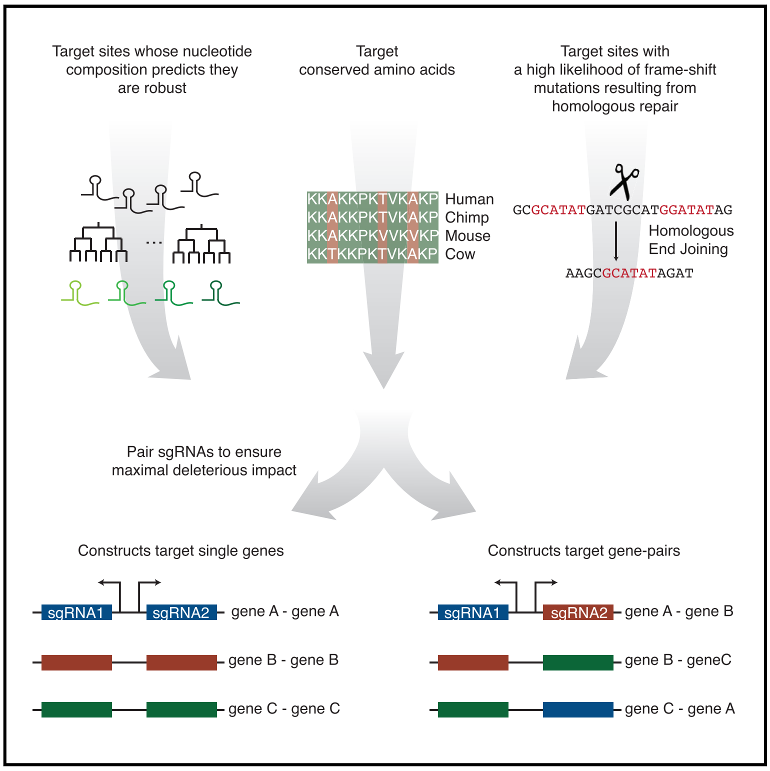
A CRISPR Resource for Individual, Combinatorial, or Multiplexed Gene Knockout

Technology

Molecular Cell

A CRISPR Resource for Individual, Combinatorial, or Multiplexed Gene Knockout

Graphical Abstract


Target sites whose nucleotide composition predicts they are robust
Target conserved amino acids
Target sites with a high likelihood of frame-shift mutations resulting from homologous repair
KKAKKPKTVKAKP Human
KKAKKPKTVKAKP Chimp
KKAKKPKVVKVKP Mouse
KKTKKPKTVKAKP Cow
GCGCATATGATCGCATGGATATAG
Homologous End Joining
AAGCGCATATAGAT
Pair sgRNAs to ensure maximal deleterious impact
Constructs target single genes
Constructs target gene-pairs
sgRNA1
sgRNA2
gene A - gene A
gene B - gene B
gene C - gene C
gene A - gene B
gene B - geneC
gene C - gene A

Highlights

- Target conservation and target-flanking homologous sequences impact sgRNA potency
- Functional impact is increased when Cas9 is focused to multiple sites in the target
- A resource for individual and combinatorial CRISPR screens is presented

Authors

Nicolas Erard, Simon R.V. Knott, Gregory J. Hannon

Correspondence

simon.knott@cshs.org (S.R.V.K.), greg.hannon@cruk.cam.ac.uk (G.J.H.)

In Brief

Erard et al. present an algorithm for predicting sgRNA potency that they combine with expression strategies to generate a CRISPR resource for performing individual, combinatorial, or multiplexed gene knockout in human cells. The resource is compared to other published tools through comparative multiplexed screens. **Please note that a Correction has been appended to this article**.

Erard et al., 2017, Molecular Cell *67*, 348–354
July 20, 2017 © 2017 The Authors. Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.molcel.2017.06.030

CellPress

Molecular Cell
Technology

A CRISPR Resource for Individual, Combinatorial, or Multiplexed Gene Knockout

Nicolas Erard,[1,5] Simon R.V. Knott,[1,2,3,5,*] and Gregory J. Hannon[1,2,4,6,*]
[1]Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK
[2]Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA
[3]Cedars-Sinai Medical Institute, 8700 Beverly Boulevard, Los Angeles, CA 90048, USA
[4]New York Genome Center, 101 6th Avenue, New York, NY 10013, USA
[5]These authors contributed equally to this work
[6]Lead Contact
*Correspondence: simon.knott@cshs.org (S.R.V.K.), greg.hannon@cruk.cam.ac.uk (G.J.H.)
http://dx.doi.org/10.1016/j.molcel.2017.06.030

SUMMARY

We have combined a machine-learning approach with other strategies to optimize knockout efficiency with the CRISPR/Cas9 system. In addition, we have developed a multiplexed sgRNA expression strategy that promotes the functional ablation of single genes and allows for combinatorial targeting. These strategies have been combined to design and construct a genome-wide, sequence-verified, arrayed CRISPR library. This resource allows single-target or combinatorial genetic screens to be carried out at scale in a multiplexed or arrayed format. By conducting parallel loss-of-function screens, we compare our approach to existing sgRNA design and expression strategies.

INTRODUCTION

Genetic screens have played a fundamental role in charting genotype-phenotype interaction maps for a variety of organisms (Carpenter and Sabatini, 2004). However, confounding factors, such as non-uniformity in the efficacies of targeting molecules, have limited the depth to which data from such studies can be interpreted. These problems have been somewhat mitigated for short hairpin RNA (shRNA)-based gene silencing because, after several rounds of optimization, experimentally validated algorithms for selecting potent guide sequences have been developed (Fellmann et al., 2011; Knott et al., 2014; Pelossof et al., 2017). Similar approaches have been applied for selecting Cas9 guide RNAs (sgRNAs) for use with the type II clustered regularly interspaced short palindromic repeats (CRISPR) system, where large sgRNA potency datasets were used to train prediction algorithms (Chari et al., 2015; Doench et al., 2014, 2016). However, unlike with mRNA cleavage, Cas9-induced double-strand breaks (DSBs) leave a genomic scar whose characteristics determine the phenotypic consequences of targeting a locus.

The distance of the target from the translation start site is anticorrelated with sgRNA efficacy, probably because N terminus proximal frameshift mutations (FSMs) are more likely to induce nonsense-mediated mRNA decay or the production of truncated nonfunctional proteins (Doench et al., 2014). Non-homologous end joining (NHEJ) was thought to act as the predominant repair mechanism at Cas9-induced DSBs; this made predicting the likelihood of an FSM, for a given target, impossible. However, deep sequencing of these genomic scars has revealed that some homologous end joining (HEJ) contributes to repair of Cas9 cleavage events (Bae et al., 2014). Here the frequencies of specific repair resolutions are dependent on the length, guanine-cytosine (GC) content, and distance from the cut site of the two DSB-flanking homologous loci, suggesting that these likelihoods can be estimated. Finally, sgRNAs that focus Cas9 to functional domains provide a greater probability of phenotypic impact, likely because in-frame mutations in these regions have a greater potential to disrupt protein function (Shi et al., 2015).

The implementation of optimized effector expression strategies should also drive the efficacy of CRISPR knockout assays. Systems have been developed in which multiple RNA polymerase III promoters drive independent sgRNAs (Vidigal and Ventura, 2015). Alternatively, others have shown that Cpf1 can be focused to multiple targets in cells that express crRNA arrays harboring independent sgRNAs (Zetsche et al., 2017). These tools have primarily been applied in order to characterize combinatorial gene interactions and to delete non-coding sequences. However, these strategies may also aid in studies where single gene knockouts are desired in each cell, as the simultaneous focusing of Cas9 to multiple sites within the target should elicit greater functional consequences.

DESIGN

Not all of the strategies outlined above have been experimentally validated, nor have they been integrated into a consolidated framework for constructing sgRNA expression vectors. We reasoned that a gain in sgRNA efficacy could be achieved by combining current selection methods with strategies to maximize the likelihood of functionally deleterious genomic scars. We developed an sgRNA selection algorithm that identifies putative targets based on predictive nucleotide combinations, the likelihood of an FSM, and whether the target lies in a functional domain. For effector delivery, we have developed a system that allows for the simultaneous expression of two independent sgRNAs from each construct. With the goal of expressing two

© 2017 The Authors. Published by Elsevier Inc.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

guides for a single target in each construct, we have developed a computational algorithm that optimizes the likelihood of synergistic deleterious effects. These methods have been validated through a reanalysis of pre-existing data and by carrying out comparative multiplexed CRISPR screens. We have predicted construct designs for all protein-coding human genes and made these available via a web portal (http://croatan.hannonlab.org/).

RESULTS

gRNA Selection Strategy

Two datasets of sgRNA efficacy have been used to develop existing selection algorithms. Doench et al. (2014) assessed the potency of sgRNAs in libraries that tiled cell surface proteins. There the abundance of integrated sgRNAs in FACS-isolated, target-negative cells was used as a measure of effector strength. Chari et al. (2015) infected cells with scrambled Cas9 targets and then transfected the same cells with corresponding sgRNAs. Here target mutation rates were the readout for efficiency. We developed a random-forest-based sgRNA prediction tool using these two datasets for training. For each dataset, ten random forests were trained to separate potent and weak guides, which were preclassified based upon a top- and bottom-40% efficacy cutoff, respectively. All 3mers in the region spanning four nucleotides upstream and six nucleotides downstream of the sgRNA binding site were used as input. The ten random forests were trained using incrementally increasing penalties for false-positive predictions. Thus, those trained with higher values were more stringent in assigning potency to a target. When analyzing new sgRNAs, sequences receive scores equal to the highest stringency level they pass in both random forest sets. This scoring system was applied to sgRNAs in the Doench tiling set that were withheld during training, and a significant difference in efficacy is observed when comparing sgRNAs that pass versus those that fail the minimum-stringency threshold (Figure S1A, rank-sum p value < 0.01). Beyond this, increments in prediction values are not matched with significant efficiency gains, although scores do correlate with potency globally.

The advantage of focusing Cas9 to known functional protein domains has been previously recognized (Shi et al., 2015). However, as many genes lack well-defined domain information, this strategy is not easily applied to the construction of genome-scale sgRNA collections. As a surrogate, we used amino-acid conservation at the Cas9 cut-site to guide sgRNA selection. We assigned scores to targets based on the predicted deleterious effects of DSB-proximal amino acid substitutions, which were calculated using the protein variation effect analyzer (PROVEAN) algorithm (Figure S1B; Choi et al., 2012). A reanalysis of the Doench tiling set shows that, for sgRNAs that pass the minimum random forest stringency threshold, these scores are correlated with the probability of inducing a measurable phenotype (Figure 1A, Spearman correlation [ρ] = 0.32).

Others have demonstrated that repair at Cas9-induced DSBs is partially driven by HEJ (Figure S1C; Bae et al., 2014). Using deep-sequencing data of Cas9 targets, we developed a linear regression model to predict the likelihood of homology-guided repair resolutions based on the length, GC content, and distance to the DSB of the corresponding homologous loci. The overall likelihood of an FSM at a target is measured as the fraction of predicted resolution scores that correspond to FSMs (Figure S1D, ρ = 0.74). This is only relevant for targets where homologous repair is likely. Thus, a lower-limit cutoff equal to the median of likelihood sums for HEJ-guided resolutions at human CDS Cas9 targets is applied as well. A reanalysis of the Doench dataset demonstrates that, for sgRNAs that pass the minimum random forest stringency threshold, a gain in efficacy can be attained by selecting targets where there is >66% chance that an FSM will occur (Figure 1B, rank-sum p value < 0.05).

To consolidate these predictive component algorithms, we first group sgRNAs based on the stringency level they passed during random-forest analysis (groups A, B, and C, Figure S1E). Within each group, sgRNAs are ranked based on their passing conservation and FSM-likelihood threshold tests. The median score of all human CDS Cas9 sites is the lower-limit threshold for conservation. We set a threshold of 66% to qualify sgRNAs as being likely to induce an FSM. sgRNAs in group A are given a score between one and three based on their passing zero, one, or two of the conservation and FSM-likelihood tests. With these same tests, sgRNAs in groups B and C are assigned scores between four and six and between seven and nine, respectively. A reanalysis of the Doench tiling set with this algorithm, which we call CRoatan, demonstrates that scores correlate strongly with potency (Figure 1C, ρ = 0.52). When CRoatan was applied to identify ten sgRNAs for each protein-coding gene in the refseq annotation, the algorithm could identify high-scoring sgRNAs for each target (Figure S1F).

To evaluate CRoatan empirically, we constructed four CRISPR libraries whose output would inform on the quality of the tool. Each library was composed of 200 sgRNAs targeting 20 essential and 20 nonessential genes (EG and NEG, respectively; five sgRNAs per gene). EGs were identified in a summary analysis of independent shRNA screens, and olfactory-receptor genes served as NEGs (Marcotte et al., 2012). For each library, a different sgRNA selection tool was used to define inclusion: gene perturbation platform (GPP; Doench et al., 2014), sgRNAScorer (Chari et al., 2015), Edit-R (Dharmacon), and CRoatan. Libraries were cloned into a lentiviral backbone where human U6 drives sgRNA expression and where a zsGreen-P2A-Puromycin bicistronic transcript is expressed from the spleen focus-forming virus promoter (SFFV). Libraries were packaged and infected into A-375 melanoma and K-562 leukemia cells, and following selection with puromycin, the cells were passaged for ~12 doublings. Normalized log ratios were then calculated based on construct abundances in the infected and final cell populations (Knott et al., 2014).

To assess the effectiveness of our effector selection strategies, we calculated gene-normalized depletion scores for all EG-sgRNAs in the CRoatan library. We could not test the initial grouping strategy, as all sgRNAs were group C members (Figure S1E). The depletion rates of EG-sgRNAs were correlated with CRoatan score (Figure 1D, ρ = 0.52). Depletion rates were not found to correlate with conservation or FSM-likelihood scores alone. When EG-sgRNA depletion rates were compared among the four libraries, CRoatan sgRNAs were found to be significantly more reduced in representation than those identified with the sgRNAScorer and Edit-R tools (Figure 1E, rank-sum p value < 0.05). CRoatan EG-sgRNAs were

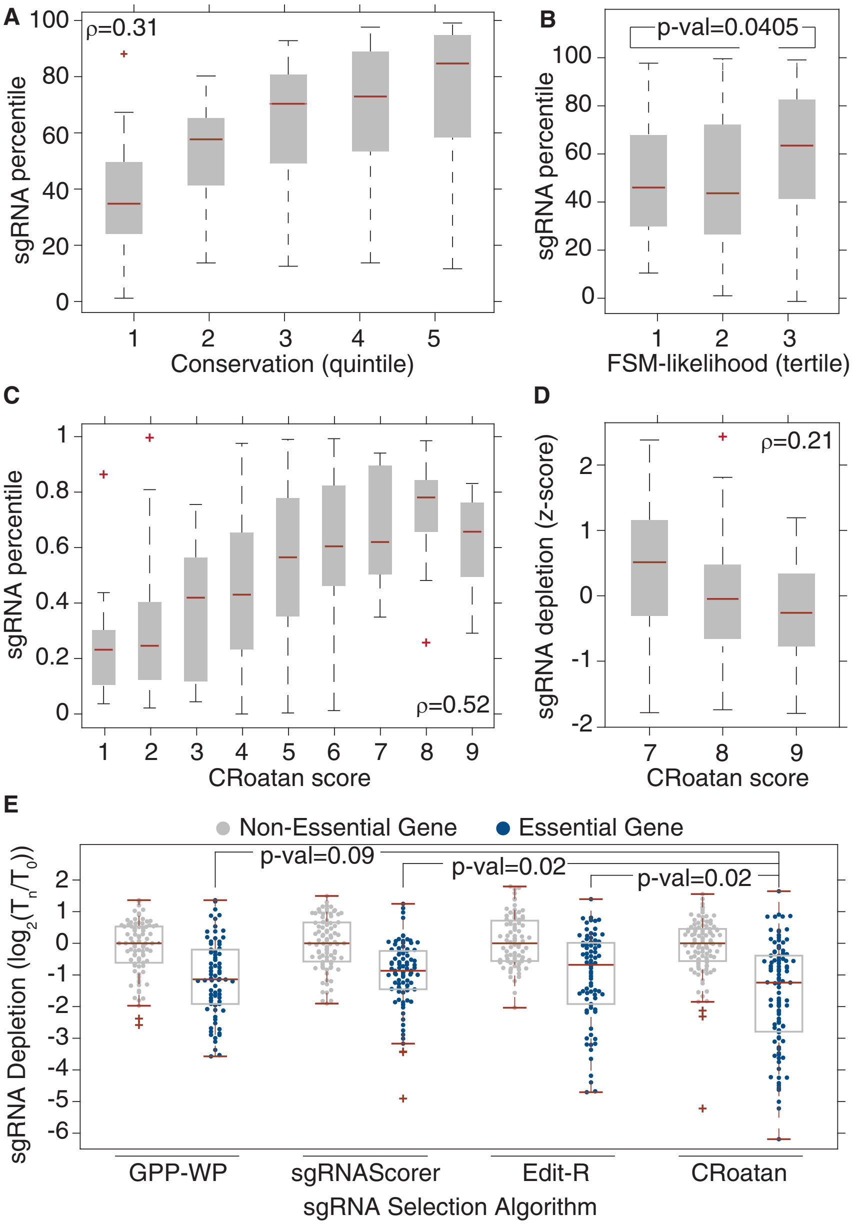


Figure 1. CRoatan, an Algorithm for Identifying Potent sgRNAs

(A) The potency of sgRNAs analyzed in Doench et al. stratified by conservation score (calculated as described in Figure S1B, ρ = 0.32). An sgRNA percentile is the percentile rank of an sgRNA relative to all other effectors targeting the same gene. This plot, as all others in the figure, was generated with the MATLAB boxplot function using default parameters. The edges of the box are the 25th and 75th percentiles. The error bars extend to the values q3 + w(q3 − q1) and q1 − w(q3 − q1), where w is 1.5 and q1 and q3 are the 25th and 75th percentiles.

(B) Efficacy percentiles of the sgRNAs analyzed in Doench et al. when stratified by the likelihood of frameshift mutations (FSM likelihood) at the corresponding target site (rank-sum p value = 0.0405 for tertile 3 versus tertile 1 and 2 sgRNAs).

(C) Efficacy percentiles of the sgRNAs analyzed in Doench et al. when stratified by the consolidated CRoatan algorithm (ρ = 0.52).

(D) *Z*-score-normalized depletion rates of EG-sgRNAs when stratified by CRoatan score (ρ = 0.21). Depletion rates were calculated as the average log ratio in screens carried out in A-375 and K-562 cells.

(E) Depletion rates of NEG- and EG-targeting sgRNAs in screens corresponding to those described in (D). sgRNA libraries were designed using the GPP-WP, sgRNAScorer, and Edit-R algorithms (rank-sum p value = 0.0942 for GPP-WP, 0.0209 for sgRNAScorer, and 0.0233 for Edit-R).

more depleted than those identified with the GPP algorithm; however, this difference was not statistically significant (rank-sum p value > 0.05).

Dual-sgRNA Expression Constructs

We reasoned that a higher frequency of deleterious mutations could be inflicted by simultaneously focusing multiple independent sgRNAs to each gene target. Toward this end, we constructed a lentiviral vector harboring two divergent U6 promoters, where the 5′ promoter was human and the 3′ promoter was chicken (Figure 2A, hU6 and cU6, respectively). These were chosen to reduce the probability that recombination would eliminate critical elements of the cassette. Between the promoters is an identification barcode, which is bordered by Illumina adapters, for sequencing-based quantification of construct abundances. The vector also harbors a bicistronic zsGreen-P2A-Puromycin transcript that is expressed from SFFV.

We designed an algorithm to pair sgRNAs for a target within the dual-U6 vector to maximize the probability of synergistic deleterious effects. The algorithm receives as input ten sgRNAs, which have been extracted from the top-20 CRoatan-scoring effectors, after they have been reranked to reflect off-target likelihoods (Figure 2B; Knott et al., 2014). A 10 × 10 pairwise score matrix is then calculated using a heuristic scoring algorithm (Figure S2A). sgRNAs with overlapping targets are not considered for pairing. To ensure that each construct harbors at least one potent effector, the algorithm increments the score of sgRNA pairs with unbalanced CRoatan scores. sgRNA pairs are also increased in their scores if they target the same exon or two exons that contribute to a common set of isoforms. Finally, we predicted that DSB-DSB blunt-end joining would be the predominant repair resolution in cases where simultaneous cleavage events caused the target-flanked region to be deleted. Thus, the score is also increased for each pair whose deletion fragment length corresponds to an FSM. After the sgRNA pairs have been scored, a weighted maximum matching algorithm is applied to identify the coupling with the highest sum of pair scores.

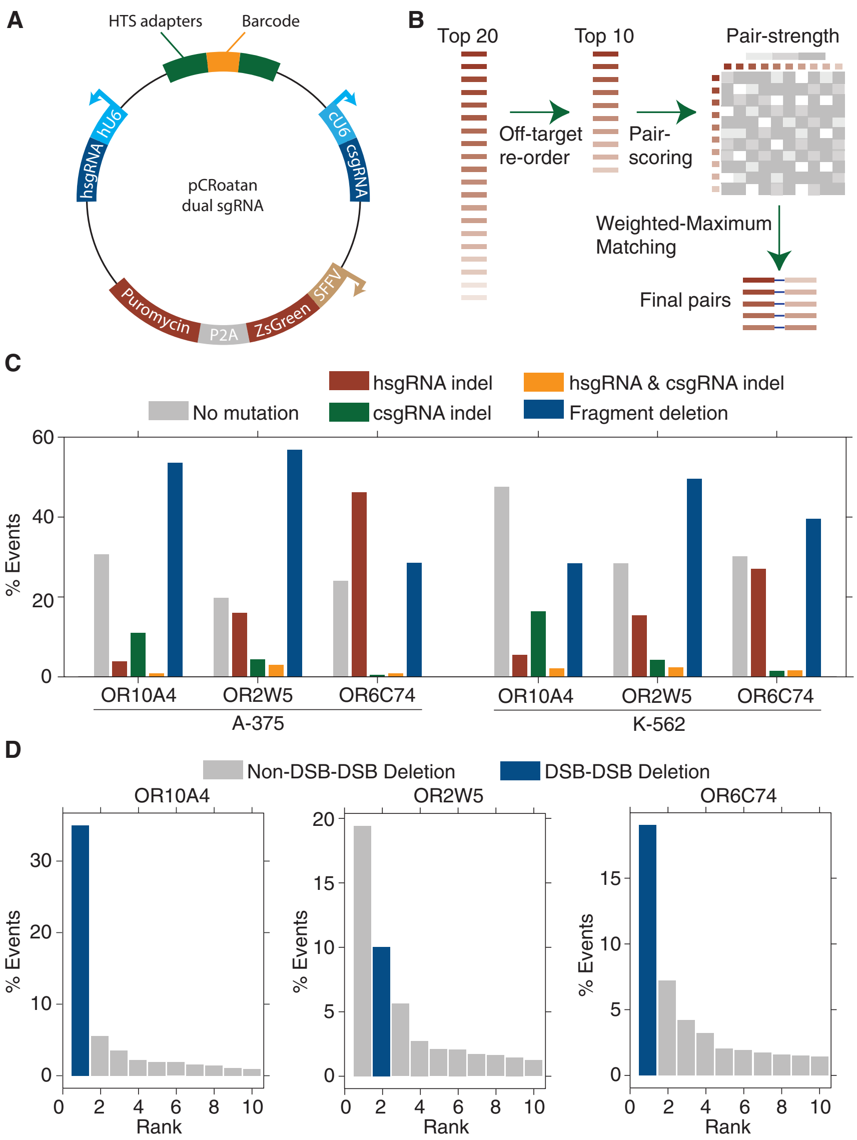


Figure 2. Simultaneous Targeting with Multiple sgRNAs Results in Predictable Genomic Scars

(A) Schematic map of the lentiviral, dual-sgRNA expression vector with relevant features highlighted. hU6, human U6 promoter; cU6, chicken U6 promoter; hsgRNA, human U6-promoter-driven sgRNA; csgRNA, human U6-promoter-driven sgRNA; HTS, high-throughput sequencing adapters; SFFV, spleen focus-forming virus promoter.

(B) sgRNA pairing algorithm used to design five targeting constructs for a gene. The top 20 sgRNAs for each gene are filtered to a set of 10 to reduce the probability of off-targeting effects. Pairs within these 10 are then scored using the set of heuristics defined in Figure S2. The resultant pairing matrix is then used as input for a maximum-weighted matching algorithm to define a final set of 5 sgRNA pairs.

(C) Paired-end sequencing analysis of genomic scars left after dual-CRoatan NEG-targeting constructs have been infected into A-375 and K-562 cells. hsgRNA and csgRNA indels are where only one of the two targeted regions shows mutational burden in an HTS fragment. hsgRNA and csgRNA indel counts represent cases where both targets have indels, and fragment deletions are where the region between the two targets is deleted.

(D) Analysis of the genomic scars described in (C) that correspond to fragment deletions between two sgRNA target sites. The top ten most frequent deletions are shown with their corresponding rate of occurrence, as measured by their average frequency in infected A-375 and K-562 cells. Scars that result from exact deletion of the double-strand-break-flanked fragment are annotated as DSB-DSB deletions. Scars where, in addition to the fragment deletion, other bases are inserted or deleted are annotated as non-DSB-DSB deletions.

To test our library assembly strategy, we cloned dual-CRoatan constructs for three olfactory receptor genes (OR10A4, OR2W5, and OR6C74) and infected A-375 and K-562 cells with these. These sgRNA pairs were chosen for the short distance between their corresponding targets, which allows simultaneous analysis of both sites with Illumina paired-end sequencing. We profiled the genomic scars that had been left after infection and found that high rates of mutation existed for all sgRNA pairs (Figure 2C). Fragment deletion between the two targets was the predominant scar. A deeper analysis revealed that, in these cases, the most commonly observed resolution was the predicted blunt-end joining of the two DSBs (Figures 2D and S2B).

High-Throughput Analysis of Library Efficacy

To evaluate our strategy more broadly, we designed a combinatorial CRISPR screening library whose output would inform on the contributions that the CRoatan algorithm, as well as the dual-sgRNA expression system, made to reagent efficacy. The library was composed of 100 sgRNAs targeting 20 EGs and 20 NEGs. sgRNAs were cloned into both the hU6 and cU6 positions, which resulted in a final library harboring 10,000 sgRNA pairs. The constructs were screened in A-375 cells, and these experiments were processed in the same manner as those experiments described in Figures 1D and 1E.

To assess the impact of the CRoatan algorithm constituents, we calculated gene-normalized depletion scores for constructs harboring one EG-sgRNA, as depletion rates could be attributed directly to the efficacy of this effector for these constructs. In contrast to the single-sgRNA CRoatan screen described in Figures 1D and 1E, here depletion rates were significantly greater for EG-sgRNAs that passed the conservation and FSM-likelihood thresholds, indicating that these two strategies contributed positively (Figures S3A and S3B, Friedman p value < 0.05). We reason that this correlation was observable here, and not in the initial CRoatan screen, because for each EG-sgRNA, the score was calculated as the average depletion rate of the 100 constructs in which it was paired with an NEG-sgRNA. Also, as was the case

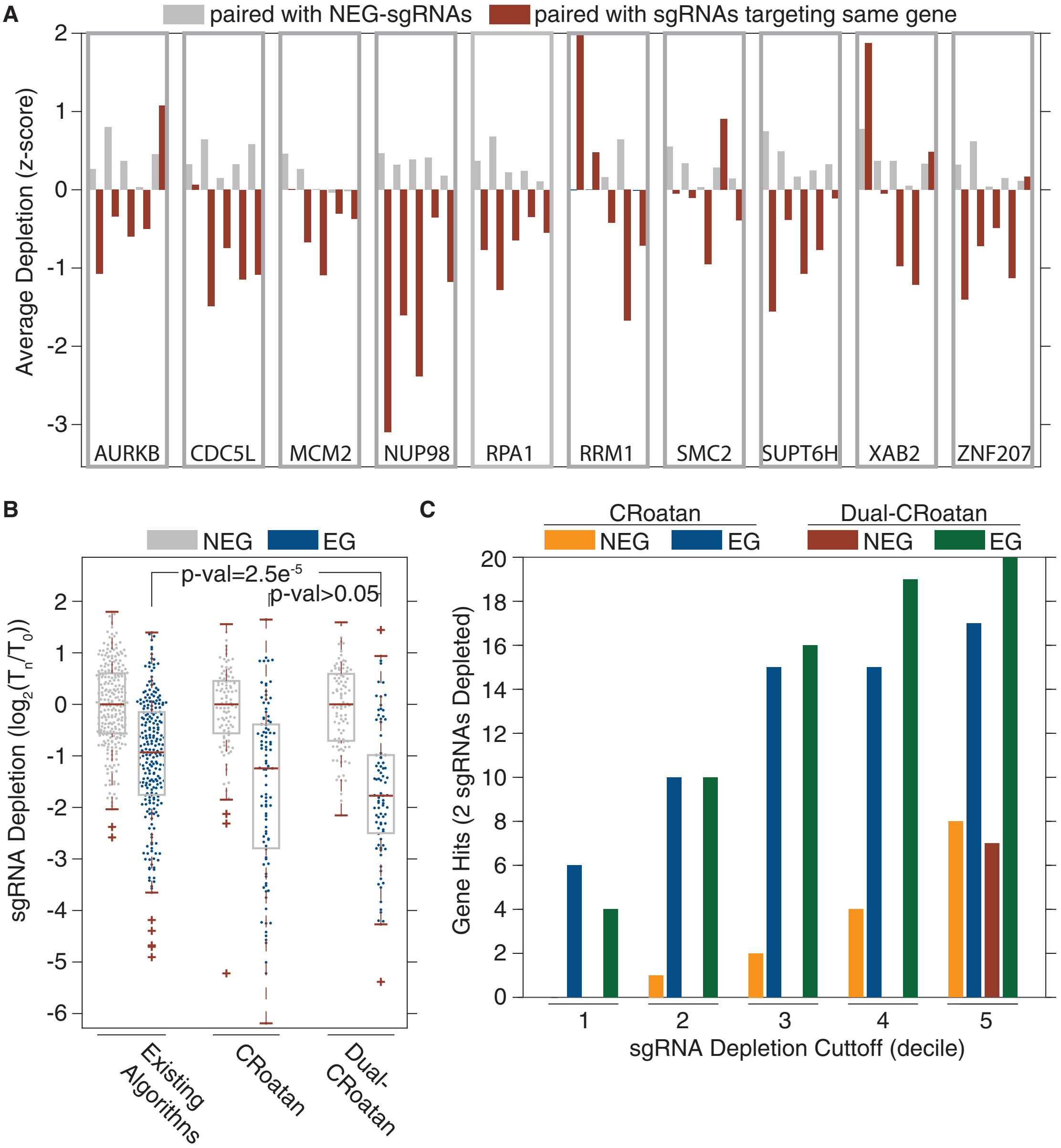


Figure 3. Dual-CRoatan Constructs Provide Superior CRISPR-Based Gene Targeting

(A) Average depletion rates for each EG-sgRNA when it is paired with NEG-sgRNAs (gray) and when it is paired with sgRNAs targeting the same EG (brown). sgRNAs are grouped based on the gene target; rank-sum p value = 0.0006.

(B) Depletion rates of NEG- and EG-targeting CRISPR constructs in negative-selection screens. Shown are the consolidated depletion rates for single-sgRNA constructs selected using pre-existing tools (GPP, sgRNAScorer, or Edit-R algorithms) as well as the rates for CRoatan single-sgRNA constructs and CRoatan dual-sgRNA constructs (dual-CRoatan, rank-sum p value = $2.5e^{-5}$ for existing algorithms and p value > 0.05 for single-CRoatan constructs). Depletion rates were calculated as the average log ratio in screens carried out in A-375 and K-562 cells. This plot was generated with the MATLAB boxplot function using default parameters. The edges of the box are the 25th and 75th percentiles. The error bars extend to the values q3 + w(q3 − q1) and q1 − w(q3 − q1), where w is 1.5 and q1 and q3 are the 25th and 75th percentiles.

(C) Gene-level analysis of CRoatan and CRoatan dual-sgRNA construct depletion rates. Using the average depletion rate for each construct in A-375 and K-562 cells, gene "hits" were calculated using a series of stringencies (top 10%, 20%, 30%, 40%, and 50% most-depleted sgRNAs). For a gene to be called a hit at a given stringency, a minimum of two constructs need to be depleted beyond the stringency level.

for the initial CRoatan screen, depletion rates correlated positively with CRoatan score (Figure S3C, Friedman p value < 0.01).

A significant increase in depletion levels was also observed when constructs harboring two EG-sgRNAs were compared to those harboring one or zero EG-RNAs (Figure S3D, rank-sum p value < 0.01). This was also evident at the individual sgRNA level. For each EG-sgRNA, we calculated the mean depletion rate of constructs where it was paired with an NEG-sgRNA and also where it was paired with one of the other four sgRNAs that target the same gene. Nearly all of the EG-sgRNAs elicited a more robust phenotype when they were paired with other sgRNAs targeting the same gene (Figure 3A, rank-sum p value < 0.001).

As a final test of our consolidated strategy, we constructed a CRISPR library using the CRoatan algorithm and the pairing principles outlined in Figures 2A, 2B, and S2A (dual-CRoatan). Each construct in the library harbors two sgRNAs that together target one of the 10 EGs or 10 NEGs described in Figures 1D and 1E for multiplexed mutagenesis. The library was screened and analyzed as was described for these earlier experiments. The EG-targeting dual-CRoatan constructs had significantly higher depletion rates than the single-sgRNA constructs. This was true when all sgRNAs identified with existing algorithms were considered together and also when sgRNAs identified with the GPP, sgRNAScorer, and Edit-R algorithms were considered separately (rank-sum p values = $2.4e^{-5}$, 0.005, $3.9e^{-5}$, and 0.002, respectively). EG-targeting constructs in the dual-CRoatan library were more depleted than their counterparts in the CRoatan library; however, this difference was deemed statistically insignificant (rank-sum p value > 0.05). Finally, we analyzed the CRoatan and dual-CRoatan screens to identify gene-level "hits." Using a two-construct minimum threshold to identify a gene as depleted, we calculated false-positive and true-positive rates at a series of construct depletion cutoffs. This analysis demonstrated the superiority of the dual-CRoatan library in terms of both sensitivity and specificity (Figure 3C).

DISCUSSION

The CRISPR-Cas9 system has been applied to a variety of molecular manipulations, with the most common being perturbation of gene function in mammalian cells. This can be achieved by inducing mutations in target gene coding sequences or by focusing transcriptional regulators to gene promoters. Others have demonstrated, through a set of parallel loss-of-function screens, that mutagenesis is more effective at ablating gene function. Here we have combined machine-learning and sgRNA-expression strategies to create CRISPR constructs that maximize the likelihood of mutation-based functional silencing. Through a set of parallel genetic screens, we have demonstrated that these reagents are significantly more efficacious than other available tools. Based upon these results, we have assembled a sequence-verified collection of CRISPR constructs using these design principles.

We have demonstrated that a significant gain in efficacy is attained when two independent sgRNAs simultaneously focus Cas9 to the target gene. Thus, we have designed the library such that two sgRNAs with high prediction scores are expressed from each construct (Figure S3E). An added benefit of this strategy is that constructs can be easily manipulated to target gene pairs to interrogate synthetic interactions. This feature will be particularly useful for identifying parallel or related molecular pathways with combinatorial screens. Another feature of the toolkit is the availability of individual sequence-verified constructs, which allows large-scale screens to be carried out in an arrayed format.

Overall, we hope that this toolkit will be of benefit to the scientific community, as it will allow individual and combinatorial gene knockouts to be carried out on a large scale in both multiplexed and arrayed formats. The library design includes five constructs for each protein coding human Refseq gene. At present, the library is comprised of ∼50,000 sequence-verified constructs; the goal is to complete the collection at five constructs per ∼20,000 predicted genes.

LIMITATIONS

At the date of publication, half of the ∼100,000 construct designs in the human library had been sequence verified and included in the physical resource. Thus, there is poor coverage, in terms of targeting molecules, for a subset of genes. Current coverage statistics are reported on the following web portal: http://croatan.hannonlab.org.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Cell Lines
- METHOD DETAILS
  - Random Forest Training and Scoring
  - sgRNA-Pair Scoring
  - sgRNA Library Construction
  - sgRNA Library Screening
  - CRISPR/Cas9 Library Processing and Analysis
  - Dual-sgRNA Genomic Scar Analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND SOFTWARE AVAILABILITY
- ADDITIONAL RESOURCES
  - Detailed Protocol
  - Online Resource

SUPPLEMENTAL INFORMATION

Supplemental Information includes three figures, three tables, and supplemental text and can be found with this article online at http://dx.doi.org/10.1016/j.molcel.2017.06.030.

AUTHOR CONTRIBUTIONS

N.E., S.R.V.K., and G.J.H. designed the experiments and wrote the manuscript. S.R.V.K. designed the prediction algorithm and analyzed the CRISPR screens. N.E. performed all experiments. S.R.V.K. and G.J.H. supervised all experiments.

ACKNOWLEDGMENTS

This work was supported in part by funding from CRUK. We thank the CRUK Genomics facility for providing sequencing services and members of the Hannon and Knott labs for helpful discussions. S.R.V.K. and G.H. are associated with Transomic Technologies, which has commercialized libraries constructed using the CRoatan design strategy.

Received: February 21, 2017
Revised: April 21, 2017
Accepted: June 23, 2017
Published: July 20, 2017; corrected online: September 21, 2017

REFERENCES

Bae, S., Kweon, J., Kim, H.S., and Kim, J.S. (2014). Microhomology-based choice of Cas9 nuclease target sites. Nat. Methods *11*, 705–706.

Carpenter, A.E., and Sabatini, D.M. (2004). Systematic genome-wide screens of gene function. Nat. Rev. Genet. *5*, 11–22.

Chari, R., Mali, P., Moosburner, M., and Church, G.M. (2015). Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. Nat. Methods *12*, 823–826.

Choi, Y., Sims, G.E., Murphy, S., Miller, J.R., and Chan, A.P. (2012). Predicting the functional effect of amino acid substitutions and indels. PLoS ONE *7*, e46688.

Doench, J.G., Hartenian, E., Graham, D.B., Tothova, Z., Hegde, M., Smith, I., Sullender, M., Ebert, B.L., Xavier, R.J., and Root, D.E. (2014). Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nat. Biotechnol. *32*, 1262–1267.

Doench, J.G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E.W., Donovan, K.F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., et al. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat. Biotechnol. *34*, 184–191.

Fellmann, C., Zuber, J., McJunkin, K., Chang, K., Malone, C.D., Dickins, R.A., Xu, Q., Hengartner, M.O., Elledge, S.J., Hannon, G.J., and Lowe, S.W. (2011). Functional identification of optimized RNAi triggers using a massively parallel sensor assay. Mol. Cell *41*, 733–746.

Knott, S.R., Maceli, A.R., Erard, N., Chang, K., Marran, K., Zhou, X., Gordon, A., El Demerdash, O., Wagenblast, E., Kim, S., et al. (2014). A computational algorithm to predict shRNA potency. Mol. Cell *56*, 796–807.

Kudo, T., and Sutou, S. (2005). Usage of putative chicken U6 promoters for vector-based RNA interference. J. Reprod. Dev. *51*, 411–417.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Marcotte, R., Brown, K.R., Suarez, F., Sayad, A., Karamboulas, K., Krzyzanowski, P.M., Sircoulomb, F., Medrano, M., Fedyshyn, Y., Koh, J.L., et al. (2012). Essential gene profiles in breast, pancreatic, and ovarian cancer cells. Cancer Discov. *2*, 172–189.

Pelossof, R., Fairchild, L., Huang, C.H., Widmer, C., Sreedharan, V.T., Sinha, N., Lai, D.Y., Guan, Y., Premsrirut, P.K., Tschaharganeh, D.F., et al. (2017). Prediction of potent shRNAs with a sequential classification algorithm. Nat. Biotechnol. *35*, 350–353.

Sanjana, N.E., Shalem, O., and Zhang, F. (2014). Improved vectors and genome-wide libraries for CRISPR screening. Nat. Methods *11*, 783–784.

Shi, J., Wang, E., Milazzo, J.P., Wang, Z., Kinney, J.B., and Vakoc, C.R. (2015). Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. Nat. Biotechnol. *33*, 661–667.

Vidigal, J.A., and Ventura, A. (2015). Rapid and efficient one-step generation of paired gRNA CRISPR-Cas9 libraries. Nat. Commun. *6*, 8083.

Zetsche, B., Heidenreich, M., Mohanraju, P., Fedorova, I., Kneppers, J., DeGennaro, E.M., Winblad, N., Choudhury, S.R., Abudayyeh, O.O., Gootenberg, J.S., et al. (2017). Multiplex gene editing by CRISPR-Cpf1 using a single crRNA array. Nat. Biotechnol. *35*, 31–34.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Raw and analyzed data	This paper	GEO: GSE97434
Human reference genome NCBI build 37, GRCh37	Genome Reference Consortium	http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/
Experimental Models: Cell Lines		
A-375	ATCC	CRL-1619
A-375-Cas9	This paper	
K562-Cas9	Gift by the Vakoc Laboratory (CSHL)	
Oligonucleotides		
Primers used to generate sgRNAs libraries, see Table S1	This paper	N/A
Sequenced included in the DNA chip used to clone combinatorial sgRNA libraries, see Table S2	This paper	N/A
Primers used to amplify sgRNAs from gDNA and sequence targeted loci, see Table S3	This paper	N/A
Recombinant DNA		
pCRoatan-dualSgRNA	This paper	N/A
pCRoatan-singleSgRNA	This paper	N/A
pCRoatan-dualPromoter	This paper	N/A
Software and Algorithms		
Bowtie	Langmead et al., 2009	RRID: SCR_005476
Bwa	Li and Durbin, 2009	RRID:SCR_010910
Other		
Resource website for the paper	This paper	http://croatan.hannonlab.org/
CRoatan dual-sgRNA cloning protocol	This paper	Methods S1

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Gregory Hannon (greg.hannon@cruk.cam.ac.uk).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell Lines

CRISPR/Cas9 screens were performed in melanoma A-375 (ATCC CRL-1619, female) and chronic myelogenous leukemia K-562 (ATCC CCL-243, female) cell lines. A-375 were grown at 37C in DMEM, supplemented with 10% FBS and penicillin/streptomycin. K-562 were grown at 37C in RPMI1640 supplemented with 10% FBS and penicillin/streptomycin. The 293FT cell line (Thermo-Fischer) was grown at 37C in DMEM supplemented with 10% FBS and penicillin/streptomycin.

A-375 cells were infected at low MOI by virus produced using lentiCas9-Blast (Addgene #52962) (Sanjana et al., 2014) and selected using blasticidin (10 μg/mL). Following 10 days of selection, single cells were sorted using the FACSAria IIU cell sorter (BD Biosciences) into 96-well plates. 10 A-375-Cas9 clones were tested for Cas9 functionality by infection with a vector expressing ZsGreen and an sgRNA targeting ZsGreen. Knockout efficiency was estimated by flow cytometry after 14 days. One of the A-375-Cas9 clonal lines exhibiting more than 50% knockout of ZsGreen in this assay was selected for further experiments. The K-562 clonal cell line expressing Cas9 was kindly gifted by Dr. Vakoc (Cold Spring Harbor Laboratory).

METHOD DETAILS

Random Forest Training and Scoring

Ten random forests were constructed for each of the Doench et al. and Chari et al. datasets. For each data type, sgRNAs in the top- and bottom-40th percentile for each gene were classified as potent and weak, respectively. The 10 forests were trained using the MATLAB treeBagger package (1000 trees per forest). Forests were trained using incrementally increasing penalties for false-positive classifications (0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2). During training forests are constructed using the 28 overlapping 3mers of each target as features, and the class of the target (potent or weak) as the output.

When a new target is being scored, it is decomposed into 28 3mers, and these are given to each of the 20 forests (10 corresponding to the Doench et al. data and 10 to the Chari et al. data) as input. The target is then assigned a value between 0 and 10 corresponding to the highest stringency forest it was assigned as potent by. For example, if a target was called potent by a Doench forest that was trained with a penalty of 1.2 (6th lowest) and a Chari forest trained with a penalty of 1 (5th lowest), the target would receive a score of 5. The data presented in Figures S1A and 1C were calculated using out-of-bag random forest predictions with default MATLAB parameters.

sgRNA-Pair Scoring

For each gene, all pairwise scores were calculated for the top 10 CRoatan scoring sgRNAs. All sgRNA pairs begin with a score of 0. Overlapping pairs are assigned a final score of 0. Pairs that are less than 10kb apart with DSB-DSB distances that are not divisible by 3 are assigned a score of 2.5 if they target the same transcripts. Scores are incremented by 1 if pairs have imbalanced CRoatan scores (one less than 7 and one greater than 7). This scoring matrix is then given as input to the maximum weighted matching algorithm (MATLAB maxWeightMatching).

sgRNA Library Construction

For single sgRNA libraries, sgRNA sequences were predicted using existing algorithms (Edit-R, sgRNAScorer, GPP web portal and CRoatan) and oligonucleotides containing these sequences were ordered from Integrated DNA Technologies (IDT, Table S1). These molecules were amplified by PCR (forward primer (FP): TTACCGTAACTTGAAAGTATTTCGATTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCG, reverse primer (RP): GGACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAAC) and cloned by Gibson assembly into a 3rd generation lentiviral vector harboring a U6 promoter, an sgRNA backbone, and a ZsGreen-P2A-Puromycrin^{R} transcript driven by a spleen focus-forming virus promoter (pCRoatan-singleSgRNA).

For dual sgRNA libraries, sgRNA sequences were predicted using CRoatan. Primers containing these sequences were ordered from IDT (Table S1) and used to amplify a hU6-EM7-Zeocin^{R}-cU6 cassette (pCRoatan-dualPromoter). The amplicon was digested with BbsI (NEB) and ligated into a 3rd generation lentiviral vector (pCRoatan-dualSgRNA) previously digested with BsmBI (ThermoFischer).

Combinatorial sgRNA libraries were built using DNA chips (CustomArray, Inc.) containing 10K molecules harboring a barcode and two flanking sgRNA sequences (Table S2). Chips were amplified by 5 separate 18-cycle PCRs to ensure high-complexity end product. The amplicons were first cloned by ligation into an intermediate cloning vector (pCR-BluntII TOPO based) using SpeI (NEB) and ApaI (NEB). Subsequently, the hU6 and cU6 promoters driving the sgRNAs were added to the vector. The hU6 promoter was amplified from lentiCrisprv2 (Addgene #52961) by PCR (FP: AGTACCGTCTCTGGTGTTTCGTCCTTTCCACAAG, RP: GTACCTACGCGTGAGGGCCTATTTCCCATGATTC), and cloned by ligation using the BsmBI (ThermoFischer) and MluI (NEB) restriction sites. The cU6 promoter (cU6-3, Kudo and Sutou, 2005) was amplified from a gBlock (IDT) by PCR (FP: ATCGATCTCGAGGCGCCGCCGCTCCTTCAGGCA, RP: TGATCCTGGTCTCACGACTAAGAGCATCGAGACTGC), and cloned by ligation using the BsaI (NEB) and XhoI (NEB) restriction sites. Following these three steps, the full sgRNA1-hU6-EM7-Zeocin^{R}-Barcode-cU6-sgRNA2 cassette was digested from the intermediate cloning vector using BbsI and ligated in the lentiviral expression vector (pCRoatan-dualSgRNA) as described previously. All transformations were performed with Invitrogen's MegaX DH10B T1 electro-competent cells using a Bio-Rad Gene Pulser Xcell and Bio-Rad Gene Pulser 1 mm cuvettes for electroporation. For each library, a minimum of 10 million successfully transformed cells were obtained.

sgRNA Library Screening

sgRNA libraries were packaged using the 293FT cell line (Thermo Fischer). Cells were co-transfected with library vector (60 μg), pMDL (12.5 μg), CMV-Rev (6.5 μg) and VSV-G (9 μg) by calcium phosphate transfection. The media was replaced at 14h and virus was collected at 36h and filtered using a 0.45 μM syringe filter (Millex®-HV, EMD Millipore). Viral infections were performed at an MOI of 0.3 to ensure a maximum of one sgRNA integration per cell. sgRNA representation in the infected population was maintained at a minimum of 1000 infected cells per sgRNA at each passage. All screens were performed in triplicates. Two days after infection, cells were collected for a reference time point. After ~12 doublings, cells were harvested for a final time point. Infected cells were selected using Puromycin (1 μg/mL) after the initial time point and throughout the screen.

CRISPR/Cas9 Library Processing and Analysis

Following cell harvests, DNA was extracted using the QIAGEN QIAamp DNA Blood Midi kit. For each sample, sgRNA molecules or barcodes identifying sgRNA pairs were extracted from the genomic DNA in 24 separate 30-cycle PCR reactions in which 2 μg of DNA input was included. Illumina adapters were included in the PCR primers (Table S3). Libraries were sequenced using custom read one primers on the Illumina MiSeq or HiSeq platforms. Following sequencing, reads were trimmed to a length of 20bp and construct counts were extracted using the bowtie algorithm (Langmead et al., 2009). Constructs were then filtered based on a minimum read-count threshold of 50 in the reference sample. Corresponding log-fold change values were then calculated by dividing the abundance after twelve doublings by the abundance at the reference time point, two days after infection (Knott et al., 2014).

Dual-sgRNA Genomic Scar Analysis

200,000 A-375-Cas9 and K-562-Cas9 cells were transduced with CRoatan constructs targeting 3 different olfactory receptor genes. Following selection with Puromycin cells were grown for ~12 doublings and then harvested for analysis. DNA was extracted using the QIAGEN QIAamp DNA Blood Midi kit. The target region, including 50bp upstream and downstream of both sgRNA target sites was amplified by PCR, in 16 25-cycle PCR reactions in which 500ng of DNA input was included (Table S3). Following purification using the QIAquick PCR Purification Kit, Illumina adapters were added via PCR and samples were processed on the Illumina MiSeq platform using paired-end reads of 200bp to cover both sgRNA target sites. Reads were mapped to the relevant genomic region using the bwa mem algorithm and cut types were analyzed and counted using the CIGAR string of the alignment (Li and Durbin, 2009).

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical parameters such as definition of center, error bars and significance are reported in the main text, figures and figure legends. Data are judged to be significant when $p < 0.05$ by the rank-sum test or the Friedman test. Statistical significance analysis was performed in MATLAB using the freidman and ranksum functions.

DATA AND SOFTWARE AVAILABILITY

All raw and processed data have been deposited in the National Center for Biotechnology Information Gene Expression Omnibus under accession number GSE97434. All code will be made available for non-commercial use upon request.

ADDITIONAL RESOURCES

Detailed Protocol

A detailed protocol describing the cloning of pairs of sgRNAs in the pCRoatan-dualSgRNA expression vector is provided in the Methods S1.

Online Resource

Detailed cloning protocols, plasmid maps and construct designs for all protein coding human genes are available via a web portal: http://croatan.hannonlab.org.

Molecular Cell
Correction

A CRISPR Resource for Individual, Combinatorial, or Multiplexed Gene Knockout

Nicolas Erard, Simon R.V. Knott,* and Gregory J. Hannon*
*Correspondence: simon.knott@cshs.org (S.R.V.K.), greg.hannon@cruk.cam.ac.uk (G.J.H.)

http://dx.doi.org/10.1016/j.molcel.2017.08.027

(Molecular Cell *67*, 348–354; July 20, 2017)

In our manuscript, we erroneously labeled sequences that were downloaded using the Dharmacon CRISPR RNA Configurator website as being selected with the Edit-R algorithm. These sequences were not selected by the Edit-R algorithm, nor were any Dharmacon products used in this manuscript. As a result, we are unable to draw conclusions regarding the efficacy of the Edit-R algorithm relative to the CRoatan algorithm. We apologize for any confusion this may have caused.

© 2017 The Author(s). Published by Elsevier Inc.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Bibliography

Adamson, B., Norman, T.M., Jost, M., Cho, M.Y., Nuñez, J.K., Chen, Y., Villalta, J.E., Gilbert, L.A., Horlbeck, M.A., Hein, M.Y., et al. (2016). A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* **167**: 1867–1882.e21.

Adzhubei, I., Schmidt, S., Peshkin, L., Ramensky, V., Gerasimova, A., Bork, P., Kondrashov, A., & Sunyaev, S. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* **7**: 248–249.

Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**: R106.

Anders, S., Pyl, P., & Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**: 166–169.

Apte, M.V., Wilson, J.S., Lugea, A., & Pandol, S.J. (2013). A starring role for stellate cells in the pancreatic cancer microenvironment. *Gastroenterology* **144**: 1210–9.

Apte, M., Park, S., Phillips, P., Santucci, N., Goldstein, D., Kumar, R., Ramm, G., Buchler, M., Friess, H., McCarroll, J., et al. (2004). Desmoplastic reaction in pancreatic cancer: role of pancreatic stellate cells. *Pancreas* **29**: 179–87.

Auyeung, V.C., Ulitsky, I., McGeary, S.E., & Bartel, D.P. (2013). Beyond Secondary Structure: Primary-Sequence Determinants License Pri-miRNA Hairpins for Processing. *Cell* **152**: 844–858.

Bachem, M.G., Schünemann, M., Ramadani, M., Siech, M., Beger, H., Buck, A., Zhou, S., Schmid-Kotsas, A., & Adler, G. (2005). Pancreatic carcinoma cells induce fibrosis by stimulating proliferation and matrix synthesis of stellate cells. *Gastroenterology* **128**: 907–21.

Bae, S., Kweon, J., Kim, H., & Kim, J. (2014). Microhomology-based choice of Cas9 nuclease target sites. *Nat. Methods* **11**: 705–706.

Bandyopadhyay, S., Mehta, M., Kuo, D., Sung, M.K., Chuang, R., Jaehnig, E.J., Bodenmiller, B., Licon, K., Copeland, W., Shales, M., et al. (2010). Rewiring of genetic networks in response to DNA damage. *Science* **330**: 1385–9.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., & Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**: 1709–12.

Barrangou, R. & Marraffini, L.A. (2014). CRISPR-Cas Systems: Prokaryotes Upgrade to Adaptive Immunity. *Mol. Cell* **54**: 234–244.

Bartel, D.P. (2004). MicroRNAs Genomics, Biogenesis, Mechanism, and Function. *Cell* **116**: 281–297.

Beloglazova, N., Petit, P., Flick, R., Brown, G., Savchenko, A., & Yakunin, A.F. (2011). Structure and activity of the Cas3 HD nuclease MJ0384, an effector enzyme of the CRISPR interference. *EMBO J.* **30**: 4616–27.

Berns, K., Hijmans, E., Mullenders, J., Brummelkamp, T.R., Velds, A., Heimerikx, M., Kerkhoven, R.M., Madiredjo, M., Nijkamp, W., Weigelt, B., et al. (2004). A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* **428**: 431–7.

Bernstein, E., Caudy, A., Hammond, S., & Hannon, G. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409**: 363–6.

Billmann, M., Horn, T., Fischer, B., Sandmann, T., Huber, W., & Boutros, M. (2016). A genetic interaction map of cell cycle regulators. *Mol. Biol. Cell* **27**: 1397–407.

Birmingham, A., Anderson, E.M., Reynolds, A., Ilsley-Tyree, D., Leake, D., Fedorov, Y., Baskerville, S., Maksimova, E., Robinson, K., Karpilow, J., et al. (2006). 3 UTR seed matches, but not overall identity, are associated with RNAi off-targets. *Nat. Methods* **3**: 199–204.

Bolotin, A., Quinquis, B., Sorokin, A., & Ehrlich, S. (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**: 2551–61.

Brouns, S.J.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J., Snijders, A.P., Dickman, M.J., Makarova, K.S., Koonin, E.V., & Oost, J. van der (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**: 960–4.

Brummelkamp, T.R., Bernards, R., & Agami, R. (2002). A System for Stable Expression of Short Interfering RNAs in Mammalian Cells. *Science* **296**: 550–553.

Brummelkamp, T.R., Nijman, S.M., Dirac, A.M., & Bernards, R. (2003). Loss of the cylindromatosis tumour suppressor inhibits apoptosis by activating NF-kappaB. *Nature* **424**: 797–801.

Burris, H., Moore, M., Andersen, J., Green, M., Rothenberg, M., Modiano, M., Cripps, M., Portenoy, R., Storniolo, A., Tarassoff, P., et al. (1997). Improvements in survival and clinical benefit with gemcitabine as first-line therapy for patients with advanced pancreas cancer: a randomized trial. *J. Clin. Oncol.* **15**: 2403–13.

Caldas, C., Hahn, S.A., Costa, L.T. da, Redston, M.S., Schutte, M., Seymour, A.B., Weinstein, C.L., Hruban, R.H., Yeo, C.J., & Kern, S.E. (1994). Frequent somatic mutations and homozygous deletions of the p16 (MTS1) gene in pancreatic adenocarcinoma. *Nat. Genet.* **8**: 27–32.

Cancer Genome Atlas Network (2015). Genomic Classification of Cutaneous Melanoma. *Cell* **161**: 1681–96.

Caplen, N., Parrish, S., Imani, F., Fire, A., & Morgan, R. (2001). Specific inhibition of gene expression by small double-stranded RNAs in invertebrate and vertebrate systems. *Proc. Natl. Acad. Sci. U.S.A.* **98**: 9742–7.

Carpenter, A.E. & Sabatini, D.M. (2004). Systematic genome-wide screens of gene function. *Nat. Rev. Genet.* **5**: 11–22.

Chang, K., Elledge, S.J., & Hannon, G.J. (2006). Lessons from Nature: microRNA-based shRNA libraries. *Nat. Methods* **3**: 707–14.

Chari, R., Mali, P., Moosburner, M., & Church, G.M. (2015). Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nat. Methods* **12**: 823–826.

Chavez, A., Tuttle, M., Pruitt, B.W., Ewen-Campen, B., Chari, R., Ter-Ovanesyan, D., Haque, S.J., Cecchi, R.J., Kowal, E.J.K., Buchthal, J., et al. (2016). Comparison of Cas9 activators in multiple species. *Nat. Methods* **13**: 563–567.

Chen, S., Sanjana, N.E., Zheng, K., Shalem, O., Lee, K., Shi, X., Scott, D.A., Song, J., Pan, J.Q., Weissleder, R., et al. (2015). Genome-wide CRISPR Screen in a Mouse Model of Tumor Growth and Metastasis. *Cell* **160**: 1246–1260.

Cheng, A.W., Wang, H., Yang, H., Shi, L., Katz, Y., Theunissen, T.W., Rangarajan, S., Shivalila, C.S., Dadon, D.B., & Jaenisch, R. (2013). Multiplexed activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. *Cell Res.* **23**: 1163–71.

Cheng, A., Saxton, T., Sakai, R., Kulkarni, S., Mbamalu, G., Vogel, W., Tortorice, C., Cardiff, R., Cross, J., Muller, W., & Pawson, T. (1998). Mammalian Grb2 regulates multiple steps in embryonic development and malignant transformation. *Cell* **95**: 793–803.

Chicas, A., Wang, X., Zhang, C., McCurrach, M., Zhao, Z., Mert, O., Dickins, R.A., Narita, M., Zhang, M., & Lowe, S.W. (2010). Dissecting the Unique Role of the Retinoblastoma Tumor Suppressor during Cellular Senescence. *Cancer Cell* **17**: 376–387.

Cho, S.W., Kim, S., Kim, Y., Kweon, J., Kim, H.S., Bae, S., & Kim, J.S. (2014). Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res.* **24**: 132–41.

Cho, S., Kim, S., Kim, J., & Kim, J. (2013). Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nat. Biotechnol.* **31**: 230–232.

Choi, Y., Sims, G.E., Murphy, S., Miller, J.R., & Chan, A.P. (2012). Predicting the Functional Effect of Amino Acid Substitutions and Indels. *Plos One* **7**: e46688.

Cleary, M.A., Kilian, K., Wang, Y., Bradshaw, J., Cavet, G., Ge, W., Kulkarni, A., Paddison, P.J., Chang, K., Sheth, N., et al. (2004). Production of complex nucleic acid libraries using highly parallel in situ oligonucleotide synthesis. *Nat. Methods* **1**: 241–8.

Collinet, C., Stöter, M., Bradshaw, C.R., Samusik, N., Rink, J.C., Kenski, D., Habermann, B., Buchholz, F., Henschel, R., Mueller, M.S., et al. (2010). Systems survey of endocytosis by multiparametric image analysis. *Nature* **464**: 243–9.

Cong, L., Ran, F., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., & Zhang, F. (2013). Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**: 819–23.

Costanzo, M., VanderSluis, B., Koch, E.N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S.D., et al. (2016). A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**: aaf1420.

Curtin, J.A., Fridlyand, J., Kageshita, T., Patel, H.N., Busam, K.J., Kutzner, H., Cho, K.H., Aiba, S., Bröcker, E.B., LeBoit, P.E., et al. (2005). Distinct sets of genetic alterations in melanoma. *N. Engl. J. Med.* **353**: 2135–47.

Dankort, D., Curley, D.P., Cartlidge, R.A., Nelson, B., Karnezis, A.N., Damsky, W.E., You, M.J., DePinho, R.A., McMahon, M., & Bosenberg, M. (2009). Braf(V600E) cooperates with Pten loss to induce metastatic melanoma. *Nat. Genet.* **41**: 544–52.

Davies, H., Bignell, G.R., Cox, C., Stephens, P., Edkins, S., Clegg, S., Teague, J., Woffendin, H., Garnett, M.J., Bottomley, W., et al. (2002). Mutations of the BRAF gene in human cancer. *Nature* **417**: 949–54.

Delás, M.J., Sabin, L.R., Dolzhenko, E., Knott, S.R., Munera Maravilla, E., Jackson, B.T., Wild, S.A., Kovacevic, T., Stork, E.M., Zhou, M., et al. (2017). lncRNA requirements for mouse acute myeloid leukemia and normal differentiation. *Elife* **6**: e25607.

Deltcheva, E., Chylinski, K., Sharma, C.M., Gonzales, K., Chao, Y., Pirzada, Z.A., Eckert, M.R., Vogel, J., & Charpentier, E. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**: 602–7.

Denli, A.M., Tops, B.B., Plasterk, R.H., Ketting, R.F., & Hannon, G.J. (2004). Processing of primary microRNAs by the Microprocessor complex. *Nature* **432**: 231–235.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**: 1853–1866.e17.

Doench, J.G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E.W., Donovan, K.F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., et al. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**: 184–191.

Doench, J.G., Hartenian, E., Graham, D.B., Tothova, Z., Hegde, M., Smith, I., Sullender, M., Ebert, B.L., Xavier, R.J., & Root, D.E. (2014). Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.* **32**: 1262–7.

Du, D., Roguev, A., Gordon, D.E., Chen, M., Chen, S.-H., Shales, M., Shen, J., Ideker, T., Mali, P., Qi, L.S., & Krogan, N.J. (2017). Genetic interaction mapping in mammalian cells using CRISPR interference. *Nat. Methods* **14**: 577–580.

Echeverri, C.J., Beachy, P.A., Baum, B., Boutros, M., Buchholz, F., Chanda, S.K., Downward, J., Ellenberg, J., Fraser, A.G., Hacohen, N., et al. (2006). Minimizing the risk of reporting false positives in large-scale RNAi screens. *Nat. Methods* **3**: 777–9.

Elbashir, S., Lendeckel, W., & Tuschl, T. (2001). RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev.* **15**: 188–200.

Emery, C.M., Vijayendran, K.G., Zipser, M.C., Sawyer, A.M., Niu, L., Kim, J.J., Hatton, C., Chopra, R., Oberholzer, P.A., Karpova, M.B., et al. (2009). MEK1 mutations confer resistance to MEK and B-RAF inhibition. *Proc. Natl. Acad. Sci. U.S.A.* **106**: 20411–6.

Engelman, J.A., Zejnullahu, K., Mitsudomi, T., Song, Y., Hyland, C., Park, J.O., Lindeman, N., Gale, C.M., Zhao, X., Christensen, J., et al. (2007). MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling. *Science* **316**: 1039–43.

Erard, N., Knott, S.R.V.R., & Hannon, G.J. (2017). A CRISPR Resource for Individual, Combinatorial, or Multiplexed Gene Knockout. *Mol. Cell* **67**: 348–354.e4.

Esvelt, K.M., Mali, P., Braff, J.L., Moosburner, M., Yaung, S.J., & Church, G.M. (2013). Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nat. Methods* **10**: 1116–21.

Feig, C., Gopinathan, A., Neesse, A., Chan, D.S., Cook, N., & Tuveson, D.A. (2012). The Pancreas Cancer Microenvironment. *Clin. Cancer Res.* **18**: 4266–4276.

Fellmann, C., Hoffmann, T., Sridhar, V., Hopfgartner, B., Muhar, M., Roth, M., Lai, D.Y., Barbosa, I.A.A., Kwon, J.S., Guan, Y., et al. (2013). An optimized microRNA backbone for effective single-copy RNAi. *Cell Rep.* **5**: 1704–13.

Fellmann, C., Zuber, J., McJunkin, K., Chang, K., Malone, C.D., Dickins, R.A., Xu, Q., Hengartner, M.O., Elledge, S.J., Hannon, G.J., & Lowe, S.W. (2011). Functional identification of optimized RNAi triggers using a massively parallel sensor assay. *Mol. Cell* **41**: 733–46.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., & Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391**: 806–811.

Fisher, R. & Larkin, J. (2012). Vemurafenib: a new treatment for BRAF-V600 mutated advanced melanoma. *Cancer Manag. Res.* **4**: 243–52.

Fonfara, I., Richter, H., Bratovič, M., Le Rhun, A., & Charpentier, E. (2016). The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. *Nature* **532**: 517–21.

Forment, J.V., Herzog, M., Coates, J., Konopka, T., Gapp, B.V., Nijman, S.M., Adams, D.J., Keane, T.M., & Jackson, S.P. (2016). Genome-wide genetic screening with chemically mutagenized haploid embryonic stem cells. *Nat Chem Biol* **13**: 12–14.

Fu, Y., Foden, J.A., Khayter, C., Maeder, M.L., Reyon, D., Joung, J., & Sander, J.D. (2013). High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.* **31**: 822–6.

Gajria, D. & Chandarlapaty, S. (2011). HER2-amplified breast cancer: mechanisms of trastuzumab resistance and novel targeted therapies. *Expert Rev. Anticancer Ther.* **11**: 263–75.

Garneau, J.E., Dupuis, M.-È.È., Villion, M., Romero, D.A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadán, A.H., & Moineau, S. (2010). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**: 67–71.

Gilbert, L.A., Horlbeck, M.A., Adamson, B., Villalta, J.E., Chen, Y., Whitehead, E.H., Guimaraes, C., Panning, B., Ploegh, H.L., Bassik, M.C., et al. (2014). Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159**: 647–61.

Gilbert, L.A., Larson, M.H., Morsut, L., Liu, Z., Brar, G.A., Torres, S.E., Stern-Ginossar, N., Brandman, O., Whitehead, E.H., Doudna, J.A., et al. (2013). CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**: 442–51.

Giubellino, A., Gao, Y., Lee, S., Lee, M.J., Vasselli, J.R., Medepalli, S., Trepel, J.B., Burke, T.R., & Bottaro, D.P. (2007). Inhibition of tumor metastasis by a growth factor receptor bound protein 2 Src homology 2 domain-binding antagonist. *Cancer Res.* **67**: 6012–6.

Goel, V.K., Lazar, A.J., Warneke, C.L., Redston, M.S., & Haluska, F.G. (2006). Examination of mutations in BRAF, NRAS, and PTEN in primary cutaneous melanoma. *J. Invest. Dermatol.* **126**: 154–60.

Gregory, R.I., Yan, K.-p., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N., & Shiekhattar, R. (2004). The Microprocessor complex mediates the genesis of microRNAs. *Nature* **432**: 235–240.

Grimm, D., Streetz, K.L., Jopling, C.L., Storm, T.A., Pandey, K., Davis, C.R., Marion, P., Salazar, F., & Kay, M.A. (2006). Fatality in mice due to oversaturation of cellular microRNA/short hairpin RNA pathways. *Nature* **441**: 537–41.

Hale, C.R., Majumdar, S., Elmore, J., Pfister, N., Compton, M., Olson, S., Resch, A.M., Glover, C.V., Graveley, B.R., Terns, R.M., & Terns, M.P. (2012). Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. *Mol. Cell* **45**: 292–302.

Hamilton, A.J. & Baulcombe, D.C. (1999). A Species of Small Antisense RNA in Posttranscriptional Gene Silencing in Plants. *Science* **286**: 950–952.

Hammond, S., Boettcher, S., Caudy, A., Kobayashi, R., & Hannon, G. (2001). Argonaute2, a link between genetic and biochemical analyses of RNAi. *Science* **293**: 1146–50.

Han, J., Lee, Y., Yeom, K.H., Nam, J.W., Heo, I., Rhee, J.K., Sohn, S.Y., Cho, Y., Zhang, B.T., & Kim, V. (2006). Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell* **125**: 887–901.

Hannon, G.J. (2002). RNA interference. *Nature* **418**: 244–251.

Hayward, N.K., Wilmott, J.S., Waddell, N., Johansson, P.A., Field, M.A., Nones, K., Patch, A.M., Kakavand, H., Alexandrov, L.B., Burke, H., et al. (2017). Whole-genome landscapes of major melanoma subtypes. *Nature* **545**: 175–180.

Hilton, I.B., D'Ippolito, A.M., Vockley, C.M., Thakore, P.I., Crawford, G.E., Reddy, T.E., & Gersbach, C.A. (2015). Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat. Biotechnol.* **33**: 510–7.

Hingorani, S.R., Wang, L., Multani, A.S., Combs, C., Deramaudt, T.B., Hruban, R.H., Rustgi, A.K., Chang, S., & Tuveson, D.A. (2005). Trp53R172H and KrasG12D cooperate to promote chromosomal instability and widely metastatic pancreatic ductal adenocarcinoma in mice. *Cancer Cell* **7**: 469–83.

Hodis, E., Watson, I.R., Kryukov, G.V., Arold, S.T., Imielinski, M., Theurillat, J.P., Nickerson, E., Auclair, D., Li, L., Place, C., et al. (2012). A landscape of driver mutations in melanoma. *Cell* **150**: 251–63.

Holmes, D. (2014). The cancer that rises with the sun. *Nature* **515**: S110–1.

Hopkins, A.L. & Groom, C.R. (2002). The druggable genome. *Nat. Rev. Drug. Discov.* **1**: 727–730.

Horlbeck, M.A., Gilbert, L.A., Villalta, J.E., Adamson, B., Pak, R.A., Chen, Y., Fields, A.P., Park, C.Y., Corn, J.E., Kampmann, M., & Weissman, J.S. (2016). Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *Elife* **5**: e19760.

Hsu, P.D., Scott, D.A., Weinstein, J.A., Ran, F., Konermann, S., Agarwala, V., Li, Y., Fine, E.J., Wu, X., Shalem, O., et al. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**: 827–32.

Huesken, D., Lange, J., Mickanin, C., Weiler, J., Asselbergs, F., Warner, J., Meloon, B., Engel, S., Rosenberg, A., Cohen, D., et al. (2005). Design of a genome-wide siRNA library using an artificial neural network. *Nat. Biotechnol.* **23**: 995–1001.

Huo, Y., Nam, K.H., Ding, F., Lee, H., Wu, L., Xiao, Y., Farchione, M., Zhou, S., Rajashankar, K., Kurinov, I., et al. (2014). Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. *Nat. Struct. Mol. Biol.* **21**: 771–7.

Hutvagner, G. & Zamore, P.D. (2002). A microRNA in a Multiple-Turnover RNAi Enzyme Complex. *Science* **297**: 2056–2060.

Hwang, R.F., Moore, T., Arumugam, T., Ramachandran, V., Amos, K.D., Rivera, A., Ji, B., Evans, D.B., & Logsdon, C.D. (2008). Cancer-associated stromal fibroblasts promote pancreatic tumor progression. *Cancer Res.* **68**: 918–26.

Jackson, A.L., Bartz, S.R., Schelter, J., Kobayashi, S.V., Burchard, J., Mao, M., Li, B., Cavet, G., & Linsley, P.S. (2003). Expression profiling reveals off-target gene regulation by RNAi. *Nat. Biotechnol.* **21**: 635–7.

Jackson, A.L., Burchard, J., Schelter, J., Chau, B., Cleary, M., Lim, L., & Linsley, P.S. (2006). Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity. *RNA* **12**: 1179–87.

Jackson, R.N., Golden, S.M., Erp, P.B. van, Carter, J., Westra, E.R., Brouns, S.J., Oost, J. van der, Terwilliger, T.C., Read, R.J., & Wiedenheft, B. (2014). Crystal structure of the CRISPR RNA–guided surveillance complex from {<}i{>}Escherichia coli{<}/i{>}. *Science* **345**: 1473–1479.

Jacobetz, M.A., Chan, D.S., Neesse, A., Bapiro, T.E., Cook, N., Frese, K.K., Feig, C., Nakagawa, T., Caldwell, M.E., Zecchini, H.I., et al. (2013). Hyaluronan impairs vascular function and drug delivery in a mouse model of pancreatic cancer. *Gut* **62**: 112–20.

Jasin, M. & Rothstein, R. (2013). Repair of Strand Breaks by Homologous Recombination. *Csh Perspect Biol* **5**: a012740.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., & Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**: 816–21.

Jinek, M., East, A., Cheng, A., Lin, S., Ma, E., & Doudna, J. (2013). RNA-programmed genome editing in human cells. *Elife* **2**: e00471.

Johannessen, C.M., Boehm, J.S., Kim, S.Y., Thomas, S.R., Wardwell, L., Johnson, L.A., Emery, C.M., Stransky, N., Cogdill, A.P., Barretina, J., et al. (2010). COT drives resistance to RAF inhibition through MAP kinase pathway reactivation. *Nature* **468**: 968–72.

Jorgensen, E.M. & Mango, S.E. (2002). The art and design of genetic screens: caenorhabditis elegans. *Nat. Rev. Genet.* **3**: 356–69.

Kabadi, A.M., Ousterout, D.G., Hilton, I.B., & Gersbach, C.A. (2014). Multiplex CRISPR/Cas9-based genome engineering from a single lentiviral vector. *Nucleic Acids Res.* **42**: e147.

Kamisawa, T., Wood, L.D., Itoi, T., & Takaori, K. (2016). Pancreatic cancer. *Lancet* **388**: 73–85.

Kanehisa, M. & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**: 27–30.

Kearns, N.A., Pham, H., Tabak, B., Genga, R.M., Silverstein, N.J., Garber, M., & Maehr, R. (2015). Functional annotation of native enhancers with a Cas9-histone demethylase fusion. *Nat. Methods* **12**: 401–403.

Khan, A.A., Betel, D., Miller, M.L., Sander, C., Leslie, C.S., & Marks, D.S. (2009). Transfection of small RNAs globally perturbs gene regulation by endogenous microRNAs. *Nat. Biotechnol.* **27**: 549–55.

Khvorova, A., Reynolds, A., & Jayasena, S.D. (2003). Functional siRNAs and miRNAs exhibit strand bias. *Cell* **115**: 209–16.

Kile, B.T. & Hilton, D.J. (2005). The art and design of genetic screens: mouse. *Nat. Rev. Genet.* **6**: 557–67.

Kim, D., Kim, J., Hur, J.K., Been, K.W., Yoon, S.H., & Kim, J.S. (2016). Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells. *Nat. Biotechnol.* **34**: 863–8.

Kim, E.J., Sahai, V., Abel, E.V., Griffith, K.A., Greenson, J.K., Takebe, N., Khan, G.N., Blau, J.L., Craig, R., Balis, U.G., et al. (2014). Pilot clinical trial of hedgehog pathway inhibitor GDC-0449 (vismodegib) in combination with gemcitabine in patients with metastatic pancreatic adenocarcinoma. *Clin. Cancer Res.* **20**: 5937–5945.

Kim, H.K., Song, M., Lee, J., Menon, V.A., Jung, S., Kang, Y.-M., Choi, J.W., Woo, E., Koh, H.C., Nam, J.-W., & Kim, H. (2016). In vivo high-throughput profiling of CRISPR-Cpf1 activity. *Nat. Methods* **14**: 153–159.

Kleinstiver, B.P., Tsai, S.Q., Prew, M.S., Nguyen, N.T., Welch, M.M., Lopez, J.M., McCaw, Z.R., Aryee, M.J., & Joung, J.K. (2016). Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells. *Nat. Biotechnol.* **34**: 869–74.

Knott, S.R., Maceli, A.R., Erard, N., Chang, K., Marran, K., Zhou, X., Gordon, A., El Demerdash, O., Wagenblast, E., Kim, S., et al. (2014). A computational algorithm to predict shRNA potency. *Mol. Cell* **56**: 796–807.

Koh, S.B., Courtin, A., Boyce, R.J., Boyle, R.G., Richards, F.M., & Jodrell, D.I. (2015). CHK1 Inhibition Synergizes with Gemcitabine Initially by Destabilizing the DNA Replication Apparatus. *Cancer Res.* **75**: 3583–95.

Kolfschoten, I.G., Leeuwen, B. van, Berns, K., Mullenders, J., Beijersbergen, R.L., Bernards, R., Voorhoeve, P., & Agami, R. (2005). A genetic screen identifies PITX1 as a suppressor of RAS activity and tumorigenicity. *Cell* **121**: 849–58.

Konermann, S., Brigham, M.D., Trevino, A.E., Joung, J., Abudayyeh, O.O., Barcena, C., Hsu, P.D., Habib, N., Gootenberg, J.S., Nishimasu, H., et al. (2015). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* **517**: 583–8.

Koonin, E.V. & Makarova, K.S. (2013). CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. *RNA Biol* **10**: 679–86.

Kumar, P., Henikoff, S., & Ng, P. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**: 1073–1081.

Kumar, R., Conklin, D.S., & Mittal, V. (2003). High-throughput selection of effective RNAi probes for gene silencing. *Genome Res.* **13**: 2333–40.

Kuscu, C., Arslan, S., Singh, R., Thorpe, J., & Adli, M. (2014). Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nat. Biotechnol.* **32**: 677–83.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., & Tuschl, T. (2001). Identification of Novel Genes Coding for Small Expressed RNAs. *Science* **294**: 853–858.

Lander, E., Linton, L., Birren, B., Nusbaum, C., Zody, M., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**: 1–10.

Laufer, C., Fischer, B., Billmann, M., Huber, W., & Boutros, M. (2013). Mapping genetic interactions in human cancer cells with RNAi and multiparametric phenotyping. *Nat. Methods* **10**: 427–31.

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Rådmark, O., Kim, S., & Kim, N.V. (2003). The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425**: 415–419.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–9.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., & Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**: 417–425.

Lieber, M.R., Ma, Y., Pannicke, U., & Schwarz, K. (2003). Mechanism and regulation of human non-homologous DNA end-joining. *Nat Rev Mol Cell Bio* **4**: nrm1202.

Lin, Y., Cradick, T.J., Brown, M.T., Deshmukh, H., Ranjan, P., Sarode, N., Wile, B.M., Vertino, P.M., Stewart, F.J., & Bao, G. (2014). CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Res.* **42**: 7473–85.

Liu, T.Y., Iavarone, A.T., & Doudna, J.A. (2017). RNA and DNA Targeting by a Reconstituted Thermus thermophilus Type III-A CRISPR-Cas System. *PLoS ONE* **12**: e0170552.

Liu, X., Wu, H., Ji, X., Stelzer, Y., Wu, X., Czauderna, S., Shu, J., Dadon, D., Young, R.A., & Jaenisch, R. (2016). Editing DNA Methylation in the Mammalian Genome. *Cell* **167**: 233–247.e17.

Love, M.I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**: 550.

Lowenstein, E., Daly, R., Batzer, A., Li, W., Margolis, B., Lammers, R., Ullrich, A., Skolnik, E., Bar-Sagi, D., & Schlessinger, J. (1992). The SH2 and SH3 domain-containing protein GRB2 links receptor tyrosine kinases to ras signaling. *Cell* **70**: 431–42.

Lund, E., Güttinger, S., Calado, A., Dahlberg, J.E., & Kutay, U. (2004). Nuclear Export of MicroRNA Precursors. *Science* **303**: 95–98.

Luo, J., Emanuele, M.J., Li, D., Creighton, C.J., Schlabach, M.R., Westbrook, T.F., Wong, K.K., & Elledge, S.J. (2009). A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene. *Cell* **137**: 835–48.

Maeder, M.L., Linder, S.J., Cascio, V.M., Fu, Y., Ho, Q.H., & Joung, J. (2013). CRISPR RNA-guided activation of endogenous human genes. *Nat. Methods* **10**: 977–9.

Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J., Wolf, Y.I., Yakunin, A.F., et al. (2011). Evolution and classification of the CRISPR–Cas systems. *Nat. Rev. Microbiol.* **9**: 467–477.

Mali, P., Aach, J., Stranges, P., Esvelt, K.M., Moosburner, M., Kosuri, S., Yang, L., & Church, G.M. (2013). CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.* **31**: 833–8.

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., & Church, G.M. (2013). RNA-guided human genome engineering via Cas9. *Science* **339**: 823–6.

Malone, C., Brennecke, J., Czech, B., Aravin, A., & Hannon, G.J. (2012). Preparation of Small RNA Libraries for High-Throughput Sequencing. *Cold Spring Harb. Protoc.* **2012**: pdb.prot071431.

Manchado, E., Weissmueller, S., Morris, J.P., Chen, C.C., Wullenkord, R., Lujambio, A., Stanchina, E. de, Poirier, J.T., Gainor, J.F., Corcoran, R.B., et al. (2016). A combinatorial strategy for treating KRAS-mutant lung cancer. *Nature* **534**: 647–51.

Manguso, R.T., Pope, H.W., Zimmer, M.D., Brown, F.D., Yates, K.B., Miller, B.C., Collins, N.B., Bi, K., LaFleur, M.W., Juneja, V.R., et al. (2017). In vivo CRISPR screening identifies Ptpn2 as a cancer immunotherapy target. *Nature* **547**: 413–418.

Martinez, J., Patkaniowska, A., Urlaub, H., Lührmann, R., & Tuschl, T. (2002). Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. *Cell* **110**: 563–74.

McDonald, E., Weck, A. de, Schlabach, M.R., Billy, E., Mavrakis, K.J., Hoffman, G.R., Belur, D., Castelletti, D., Frias, E., Gampa, K., et al. (2017). Project DRIVE: A Compendium of Cancer Dependencies and Synthetic Lethal Relationships Uncovered by Large-Scale, Deep RNAi Screening. *Cell* **170**: 577–592.e10.

McDonald, J.I., Celik, H., Rois, L.E., Fishberger, G., Fowler, T., Rees, R., Kramer, A., Martens, A., Edwards, J.R., & Challen, G.A. (2016). Reprogrammable CRISPR/Cas9-based system for inducing site-specific DNA methylation. *Biol. Open* **5**: 866–74.

Miller, T.W., Pérez-Torres, M., Narasanna, A., Guix, M., Stål, O., Pérez-Tenorio, G., Gonzalez-Angulo, A.M., Hennessy, B.T., Mills, G.B., Kennedy, J., et al. (2009). Loss of Phosphatase and Tensin homologue deleted on chromosome 10 engages ErbB3 and insulin-like growth factor-I receptor signaling to promote antiestrogen resistance in breast cancer. *Cancer Res.* **69**: 4192–201.

Mojica, F., Díez-Villaseñor, C., García-Martínez, J., & Almendros, C. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**: 733–40.

Mojica, F.J., Díez-Villaseñor, C., García-Martínez, J., & Soria, E. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol.* **60**: 174–82.

Moreno-Mateos, M.A., Vejnar, C.E., Beaudoin, J.D., Fernandez, J.P., Mis, E.K., Khokha, M.K., & Giraldez, A.J. (2015). CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat. Methods* **12**: 982–8.

Napoli, C., Lemieux, C., & Jorgensen, R. (1990). Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in trans. *Plant Cell Online* **2**: 279–289.

Neumann, B., Walter, T., Hériché, J.K., Bulkescher, J., Erfle, H., Conrad, C., Rogers, P., Poser, I., Held, M., Liebel, U., et al. (2010). Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature* **464**: 721–7.

Ngo, V.N., Davis, R., Lamy, L., Yu, X., Zhao, H., Lenz, G., Lam, L.T., Dave, S., Yang, L., Powell, J., & Staudt, L.M. (2006). A loss-of-function RNA interference screen for molecular targets in cancer. *Nature* **441**: 106–10.

Nunez, J.K., Kranzusch, P.J., Noeske, J., Wright, A.V., Davies, C.W., & Doudna, J.A. (2014). Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nat. Struct. Mol. Biol.* **21**: 528–34.

Nunez, J.K., Lee, A.S., Engelman, A., & Doudna, J.A. (2015). Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature* **519**: 193–8.

Nüsslein-Volhard, C. & Wieschaus, E. (1980). Mutations affecting segment number and polarity in Drosophila. *Nature* **287**: 795–801.

Olive, K.P., Jacobetz, M.A., Davidson, C.J., Gopinathan, A., McIntyre, D., Honess, D., Madhu, B., Goldgraben, M.A., Caldwell, M.E., Allard, D., et al. (2009). Inhibition of Hedgehog signaling enhances delivery of chemotherapy in a mouse model of pancreatic cancer. *Science* **324**: 1457–61.

Ozdemir, B.C., Pentcheva-Hoang, T., Carstens, J.L., Zheng, X., Wu, C.C., Simpson, T.R., Laklai, H., Sugimoto, H., Kahlert, C., Novitskiy, S.V., et al. (2014). Depletion of carcinoma-associated fibroblasts and fibrosis induces immunosuppression and accelerates pancreas cancer with reduced survival. *Cancer Cell* **25**: 719–34.

Paddison, P.J., Caudy, A.A., Bernstein, E., Hannon, G.J., & Conklin, D.S. (2002). Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells. *Genes Dev.* **16**: 948–58.

Paddison, P.J., Cleary, M., Silva, J.M., Chang, K., Sheth, N., Sachidanandam, R., & Hannon, G.J. (2004). Cloning of short hairpin RNAs for gene knockdown in mammalian cells. *Nat. Methods* **1**: 163–7.

Paddison, P.J., Silva, J.M., Conklin, D.S., Schlabach, M., Li, M., Aruleba, S., Balija, V., O'Shaughnessy, A., Gnoj, L., Scobie, K., et al. (2004). A resource for large-scale RNA-interference-based screens in mammals. *Nature* **428**: 427–31.

Pattanayak, V., Lin, S., Guilinger, J.P., Ma, E., Doudna, J.A., & Liu, D.R. (2013). High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat. Biotechnol.* **31**: 839–43.

Pelossof, R., Fairchild, L., Huang, C.H., Widmer, C., Sreedharan, V.T., Sinha, N., Lai, D.Y., Guan, Y., Premsrirut, P.K., Tschaharganeh, D.F., et al. (2017). Prediction of potent shRNAs with a sequential classification algorithm. *Nat. Biotechnol.* **35**: 350–353.

Perez-Pinera, P., Kocak, D.D., Vockley, C.M., Adler, A.F., Kabadi, A.M., Polstein, L.R., Thakore, P.I., Glass, K.A., Ousterout, D.G., Leong, K.W., et al. (2013). RNA-guided gene activation by CRISPR-Cas9-based transcription factors. *Nat. Methods* **10**: 973–6.

Prevo, R., Fokas, E., Reaper, P.M., Charlton, P.A., Pollard, J.R., McKenna, W., Muschel, R.J., & Brunner, T.B. (2012). The novel ATR inhibitor VE-821 increases sensitivity of pancreatic cancer cells to radiation and chemotherapy. *Cancer Biol. Ther.* **13**: 1072–81.

Provenzano, P.P., Cuevas, C., Chang, A.E., Goel, V.K., Von Hoff, D.D., & Hingorani, S.R. (2012). Enzymatic targeting of the stroma ablates physical barriers to treatment of pancreatic ductal adenocarcinoma. *Cancer Cell* **21**: 418–29.

Qi, L.S., Larson, M.H., Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P., & Lim, W.A. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**: 1173–83.

Ran, F., Cong, L., Yan, W.X., Scott, D.A., Gootenberg, J.S., Kriz, A.J., Zetsche, B., Shalem, O., Wu, X., Makarova, K.S., et al. (2015). In vivo genome editing using Staphylococcus aureus Cas9. *Nature* **520**: 186–91.

Ran, F., Hsu, P.D., Lin, C.Y., Gootenberg, J.S., Konermann, S., Trevino, A.E., Scott, D.A., Inoue, A., Matoba, S., Zhang, Y., & Zhang, F. (2013). Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154**: 1380–9.

Reaper, P.M., Griffiths, M.R., Long, J.M., Charrier, J.D., Maccormick, S., Charlton, P.A., Golec, J.M., & Pollard, J.R. (2011). Selective killing of ATM- or p53-deficient cancer cells through inhibition of ATR. *Nat. Chem. Biol.* **7**: 428–30.

Reinhart, B., Slack, F., Basson, M., Pasquinelli, A., Bettinger, J., Rougvie, A., Horvitz, H., & Ruvkun, G. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* **403**: 901–6.

Reynolds, A., Leake, D., Boese, Q., Scaringe, S., Marshall, W.S., & Khvorova, A. (2004). Rational siRNA design for RNA interference. *Nat. Biotechnol.* **22**: 326–330.

Rhim, A.D., Oberstein, P.E., Thomas, D.H., Mirek, E.T., Palermo, C.F., Sastra, S.A., Dekleva, E.N., Saunders, T., Becerra, C.P., Tattersall, I.W., et al. (2014). Stromal elements act to restrain, rather than support, pancreatic ductal adenocarcinoma. *Cancer Cell* **25**: 735–47.

Root, D.E., Hacohen, N., Hahn, W.C., Lander, E.S., & Sabatini, D.M. (2006). Genome-scale loss-of-function screening with a lentiviral RNAi library. *Nat. Methods* **3**: 715–9.

Rudalska, R., Dauch, D., Longerich, T., McJunkin, K., Wuestefeld, T., Kang, T.W., Hohmeyer, A., Pesic, M., Leibold, J., Thun, A. von, et al. (2014). In vivo RNAi screening identifies a mechanism of sorafenib resistance in liver cancer. *Nat. Med.* **20**: 1138–46.

Russ, A.P. & Lampel, S. (2005). The druggable genome: an update. *Drug Discov. Today* **10**: 1607–1610.

Saito, M., Iwawaki, T., Taya, C., Yonekawa, H., Noda, M., Inui, Y., Mekada, E., Kimata, Y., Tsuru, A., & Kohno, K. (2001). Diphtheria toxin receptor-mediated conditional and targeted cell ablation in transgenic mice. *Nat. Biotechnol.* **19**: 746–50.

Samai, P., Pyenson, N., Jiang, W., Goldberg, G.W., Hatoum-Aslan, A., & Marraffini, L.A. (2015). Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity. *Cell* **161**: 1164–1174.

Sanjana, N.E., Shalem, O., & Zhang, F. (2014). Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods* **11**: 783–784.

Schadendorf, D., Fisher, D.E., Garbe, C., Gershenwald, J.E., Grob, J.J., Halpern, A., Herlyn, M., Marchetti, M.A., McArthur, G., Ribas, A., et al. (2015). Melanoma. *Nat. Rev. Dis. Primers* **1**: 15003.

Schlabach, M.R., Luo, J., Solimini, N.L., Hu, G., Xu, Q., Li, M.Z., Zhao, Z., Smogorzewska, A., Sowa, M.E., Ang, X.L., et al. (2008). Cancer proliferation gene discovery through functional genomics. *Science* **319**: 620–4.

Schwarz, D.S., Hutvágner, G., Du, T., Xu, Z., Aronin, N., & Zamore, P.D. (2003). Asymmetry in the assembly of the RNAi enzyme complex. *Cell* **115**: 199–208.

Semenova, E., Jore, M.M., Datsenko, K.A., Semenova, A., Westra, E.R., Wanner, B., Oost, J. van der, Brouns, S.J.J., & Severinov, K. (2011). Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc. Natl. Acad. Sci. U.S.A.* **108**: 10098–103.

Sergina, N.V., Rausch, M., Wang, D., Blair, J., Hann, B., Shokat, K.M., & Moasser, M.M. (2007). Escape from HER-family tyrosine kinase inhibitor therapy by the kinase-inactive HER3. *Nature* **445**: 437–41.

Shah, S.A., Erdmann, S., Mojica, F.J., & Garrett, R.A. (2013). Protospacer recognition motifs: mixed identities and functional diversity. *RNA Biol* **10**: 891–9.

Shalem, O., Sanjana, N.E., Hartenian, E., Shi, X., Scott, D.A., Mikkelsen, T.S., Heckl, D., Ebert, B.L., Root, D.E., Doench, J.G., & Zhang, F. (2014). Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. *Science* **343**: 84–87.

Shalem, O., Sanjana, N.E., & Zhang, F. (2015). High-throughput functional genomics using CRISPR-Cas9. *Nat. Rev. Genet.* **16**: 299–311.

Shen, B., Zhang, J., Wu, H., Wang, J., Ma, K., Li, Z., Zhang, X., Zhang, P., & Huang, X. (2013). Generation of gene-modified mice via Cas9/RNA-mediated gene targeting. *Cell Res.* **23**: 720–3.

Shen, J., Zhao, D., Sasik, R., Luebeck, J., Birmingham, A., Bojorquez-Gomez, A., Licon, K., Klepper, K., Pekin, D., Beckett, A.N., et al. (2017). Combinatorial CRISPR-Cas9 screens for de novo mapping of genetic interactions. *Nat. Methods* **14**: 573–576.

Sherman, M.H., Yu, R.T., Engle, D.D., Ding, N., Atkins, A.R., Tiriac, H., Collisson, E.A., Connor, F., Van Dyke, T., Kozlov, S., et al. (2014). Vitamin D receptor-mediated stromal reprogramming suppresses pancreatitis and enhances pancreatic cancer therapy. *Cell* **159**: 80–93.

Shi, J., Wang, E., Milazzo, J., Wang, Z., Kinney, J., & Vakoc, C. (2015). Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. *Nat. Biotechnol.* **33**: 661–667.

Silva, J.M., Li, M.Z., Chang, K., Ge, W., Golding, M.C., Rickles, R.J., Siolas, D., Hu, G., Paddison, P.J., Schlabach, M.R., et al. (2005). Second-generation shRNA libraries covering the mouse and human genomes. *Nat. Genet.* **37**: 1281–8.

Silva, J.M., Marran, K., Parker, J.S., Silva, J., Golding, M., Schlabach, M.R., Elledge, S.J., Hannon, G.J., & Chang, K. (2008). Profiling essential genes in human mammary cells by multiplex RNAi screening. *Science* **319**: 617–20.

Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**: Article3.

Sophic Alliance Inc (2010). The Integrated Druggabe Genome.

St Johnston, D. (2002). The art and design of genetic screens: Drosophila melanogaster. *Nat. Rev. Genet.* **3**: 176–88.

Staals, R.H., Zhu, Y., Taylor, D.W., Kornfeld, J.E., Sharma, K., Barendregt, A., Koehorst, J.J., Vlot, M., Neupane, N., Varossieau, K., et al. (2014). RNA targeting by the type III-A CRISPR-Cas Csm complex of Thermus thermophilus. *Mol. Cell* **56**: 518–30.

Stepper, P., Kungulovski, G., Jurkowska, R.Z., Chandra, T., Krueger, F., Reinhardt, R., Reik, W., Jeltsch, A., & Jurkowski, T.P. (2017). Efficient targeted DNA methylation with chimeric dCas9-Dnmt3a-Dnmt3L methyltransferase. *Nucleic Acids Res.* **45**: 1703–1713.

Straussman, R., Morikawa, T., Shee, K., Barzily-Rokni, M., Qian, Z.R., Du, J., Davis, A., Mongare, M.M., Gould, J., Frederick, D.T., et al. (2012). Tumour micro-environment elicits innate resistance to RAF inhibitors through HGF secretion. *Nature* **487**: 500–4.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., & Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102**: 15545–50.

't Veer, L. van, Burgering, B., Versteeg, R., Boot, A., Ruiter, D., Osanto, S., Schrier, P., & Bos, J. (1989). N-ras mutations in human cutaneous melanoma from sun-exposed body sites. *Mol. Cell. Biol.* **9**: 3114–6.

Tanenbaum, M.E., Gilbert, L.A., Qi, L.S., Weissman, J.S., & Vale, R.D. (2014). A protein-tagging system for signal amplification in gene expression and fluorescence imaging. *Cell* **159**: 635–46.

Ui-Tei, K., Naito, Y., Takahashi, F., Haraguchi, T., Ohki-Hamazaki, H., Juni, A., Ueda, R., & Saigo, K. (2004). Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference. *Nucleic Acids Res.* **32**: 936–48.

Timmons, L. & Fire, A. (1998). Specific interference by ingested dsRNA. *Nature* **395**: 854–854.

Trapnell, C., Pachter, L., & Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: 1105–1111.

Vaccaro, V., Sperduti, I., & Milella, M. (2011). FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer. *N. Engl. J. Med.* **365**: 768–9, author reply 769.

Vert, J.P., Foveau, N., Lajaunie, C., & Vandenbrouck, Y. (2006). An accurate and interpretable model for siRNA efficacy prediction. *BMC Bioinformatics* **7**: 520.

Vidigal, J.A. & Ventura, A. (2015). Rapid and efficient one-step generation of paired gRNA CRISPR-Cas9 libraries. *Nat. Commun.* **6**: 8083.

Villanueva, J., Vultur, A., Lee, J.T., Somasundaram, R., Fukunaga-Kalabis, M., Cipolla, A.K., Wubbenhorst, B., Xu, X., Gimotty, P.A., Kee, D., et al. (2010). Acquired resistance to BRAF inhibitors mediated by a RAF kinase switch in melanoma can be overcome by cotargeting MEK and IGF-1R/PI3K. *Cancer Cell* **18**: 683–95.

Vitaterna, M., King, D., Chang, A., Kornhauser, J., Lowrey, P., McDonald, J., Dove, W., Pinto, L., Turek, F., & Takahashi, J. (1994). Mutagenesis and mapping of a mouse gene, Clock, essential for circadian behavior. *Science* **264**: 719–25.

Vojta, A., Dobrinić, P., Tadić, V., Bočkor, L., Korać, P., Julg, B., Klasić, M., & Zoldos, V. (2016). Repurposing the CRISPR-Cas9 system for targeted DNA methylation. *Nucleic Acids Res.* **44**: 5615–28.

Von Hoff, D.D., Ervin, T., Arena, F.P., Chiorean, E., Infante, J., Moore, M., Seay, T., Tjulandin, S.A., Ma, W.W., Saleh, M.N., et al. (2013). Increased survival in pancreatic cancer with nab-paclitaxel plus gemcitabine. *N. Engl. J. Med.* **369**: 1691–703.

Vonlaufen, A., Joshi, S., Qu, C., Phillips, P.A., Xu, Z., Parker, N.R., Toi, C.S., Pirola, R.C., Wilson, J.S., Goldstein, D., & Apte, M.V. (2008). Pancreatic stellate cells: partners in crime with pancreatic cancer cells. *Cancer Res.* **68**: 2085–93.

Wagenblast, E., Soto, M., Gutiérrez-Ángel, S., Hartl, C.A., Gable, A.L., Maceli, A.R., Erard, N., Williams, A.M., Kim, S.Y., Dickopf, S., et al. (2015). A model of breast cancer heterogeneity reveals vascular mimicry as a driver of metastasis. *Nature* **520**: 358–62.

Wang, H., Yang, H., Shivalila, C.S., Dawlaty, M.M., Cheng, A.W., Zhang, F., & Jaenisch, R. (2013). One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153**: 910–8.

Wang, R., Preamplume, G., Terns, M.P., Terns, R.M., & Li, H. (2010). Interaction of the Cas6 Riboendonuclease with CRISPR RNAs: Recognition and Cleavage. *Structure* **19**: 257–264.

Wang, T., Wei, J.J., Sabatini, D.M., & Lander, E.S. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**: 80–4.

Wang, X., Wang, Y., Wu, X., Wang, J., Wang, Y., Qiu, Z., Chang, T., Huang, H., Lin, R.J., & Yee, J.K. (2015). Unbiased detection of off-target cleavage by CRISPR-Cas9 and TALENs using integrase-defective lentiviral vectors. *Nat. Biotechnol.* **33**: 175–8.

Weissmueller, S., Manchado, E., Saborowski, M., Morris, J.P., Wagenblast, E., Davis, C.A., Moon, S.H., Pfister, N.T., Tschaharganeh, D.F., Kitzing, T., et al. (2014). Mutant p53 drives pancreatic cancer metastasis through cell-autonomous PDGF receptor beta signaling. *Cell* **157**: 382–394.

Wiedenheft, B., Sternberg, S.H., & Doudna, J.A. (2012). RNA-guided genetic silencing systems in bacteria and archaea. *Nature* **482**: 331–8.

Wigler, M., Pellicer, A., Silverstein, S., & Axel, R. (1978). Biochemical transfer of single-copy eucaryotic genes using total cellular DNA as donor. *Cell* **14**: 725–31.

Wilson, T.R., Fridlyand, J., Yan, Y., Penuel, E., Burton, L., Chan, E., Peng, J., Lin, E., Wang, Y., Sosman, J., et al. (2012). Widespread potential for growth-factor-driven resistance to anticancer kinase inhibitors. *Nature* **487**: 505–9.

Wu, R.Z., Bailey, S.N., & Sabatini, D.M. (2002). Cell-biological applications of transfected-cell microarrays. *Trends Cell Biol.* **12**: 485–8.

Wu, X., Scott, D.A., Kriz, A.J., Chiu, A.C., Hsu, P.D., Dadon, D.B., Cheng, A.W., Trevino, A.E., Konermann, S., Chen, S., et al. (2014). Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat. Biotechnol.* **32**: 670–6.

Wuestefeld, T., Pesic, M., Rudalska, R., Dauch, D., Longerich, T., Kang, T.-W.W., Yevsa, T., Heinzmann, F., Hoenicke, L., Hohmeyer, A., et al. (2013). A Direct in vivo RNAi screen identifies MKK4 as a key regulator of liver regeneration. *Cell* **153**: 389–401.

Xu, H., Xiao, T., Chen, C.H., Li, W., Meyer, C.A., Wu, Q., Wu, D., Cong, L., Zhang, F., Liu, J.S., et al. (2015). Sequence determinants of improved CRISPR sgRNA design. *Genome Res.* **25**: 1147–57.

Xu, X., Tao, Y., Gao, X., Zhang, L., Li, X., Zou, W., Ruan, K., Wang, F., Xu, G.L., & Hu, R. (2016). A CRISPR-based approach for targeted DNA demethylation. *Cell Discov.* **2**: 16009.

Xu, Z., Vonlaufen, A., Phillips, P.A., Fiala-Beer, E., Zhang, X., Yang, L., Biankin, A.V., Goldstein, D., Pirola, R.C., Wilson, J.S., & Apte, M.V. (2010). Role of Pancreatic Stellate Cells in Pancreatic Cancer Metastasis. *Am. J. Pathology* **177**: 2585–2596.

Yang, H., Wang, H., Shivalila, C.S., Cheng, A.W., Shi, L., & Jaenisch, R. (2013). One-step generation of mice carrying reporter and conditional alleles by CRISPR/Cas-mediated genome engineering. *Cell* **154**: 1370–9.

Yi, R., Qin, Y., Macara, I.G., & Cullen, B.R. (2003). Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev.* **17**: 3011–6.

Yosef, I., Goren, M.G., & Qimron, U. (2012). Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic Acids Res.* **40**: 5569–76.

Yu, D. & Hung, M. (2000). Overexpression of ErbB2 in cancer and ErbB2-targeting strategies. *Oncogene* **19**: 6115–21.

Zalatan, J.G., Lee, M.E., Almeida, R., Gilbert, L.A., Whitehead, E.H., La Russa, M., Tsai, J.C., Weissman, J.S., Dueber, J.E., Qi, L.S., & Lim, W.A. (2015). Engineering complex synthetic transcriptional programs with CRISPR RNA scaffolds. *Cell* **160**: 339–50.

Zamore, P., Tuschl, T., Sharp, P., & Bartel, D. (2000). RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101**: 25–33.

Zender, L., Xue, W., Zuber, J., Semighini, C.P., Krasnitz, A., Ma, B., Zender, P., Kubicka, S., Luk, J.M., Schirmacher, P., et al. (2008). An oncogenomics-based in vivo RNAi screen identifies tumor suppressors in liver cancer. *Cell* **135**: 852–64.

Zeng, Y. & Cullen, B.R. (2003). Sequence requirements for micro RNA processing and function in human cells. *RNA* **9**: 112–23.

Zeng, Y., Wagner, E.J., & Cullen, B.R. (2002). Both natural and designed micro RNAs can inhibit the expression of cognate mRNAs when expressed in human cells. *Mol. Cell* **9**: 1327–33.

Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., Oost, J. van der, Regev, A., et al. (2015). Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* **163**: 759–71.

Zetsche, B., Heidenreich, M., Mohanraju, P., Fedorova, I., Kneppers, J., DeGennaro, E., Winblad, N., Choudhury, S., Abudayyeh, O., Gootenberg, J., et al. (2016). Multiplex gene editing by CRISPR-Cpf1 using a single crRNA array. *Nat. Biotechnol.* **35**: 31–34.

Zhao, H., Sheng, G., Wang, J., Wang, M., Bunkoczi, G., Gong, W., Wei, Z., & Wang, Y. (2014). Crystal structure of the RNA-guided immune surveillance Cascade complex in Escherichia coli. *Nature* **515**: 147–50.

Zheng, L., Liu, J., Batalov, S., Zhou, D., Orth, A., Ding, S., & Schultz, P.G. (2004). An approach to genomewide screens of expressed small interfering RNAs in mammalian cells. *Proc. Natl. Acad. Sci. U.S.A.* **101**: 135–40.

Ziauddin, J. & Sabatini, D. (2001). Microarrays of cells expressing defined cDNAs. *Nature* **411**: 107–10.

Zuber, J., Shi, J., Wang, E., Rappaport, A.R., Herrmann, H., Sison, E.A., Magoon, D., Qi, J., Blatt, K., Wunderlich, M., et al. (2011). RNAi screen identifies Brd4 as a therapeutic target in acute myeloid leukaemia. *Nature* **478**: 524–8.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**: 3406–3415.