Computational and experimental methods for CRISPR-based saturation mutagenesis screens

by

Jonathan Yee-Ting Hsu

B.S. Bioengineering
University of California San Diego 2016

Submitted to the Department of Biological Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Biological Engineering
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

December 2021

Author .....
Jonathan Y. Hsu
Department of Biological Engineering
December 7, 2021

Certified by .....
J. Keith Joung
Professor of Pathology
Thesis Supervisor

Certified by .....
Luca Pinello
Associate Professor of Pathology
Thesis Supervisor

Accepted by .....
Katharina Ribbeck
Professor of Biological Engineering
Graduate Program Chair

Thesis Committee Members

James J. Collins
Chairman, Thesis Committee
Professor, Biological Engineering, Medical Engineering and Science
Massachusetts Institute of Technology

J. Keith Joung
Thesis Supervisor
Professor, Pathology
Massachusetts General Hospital and Harvard Medical School

Luca Pinello
Thesis Supervisor
Associate Professor, Pathology
Massachusetts General Hospital and Harvard Medical School

Timothy K. Lu
Member, Thesis Committee
Associate Professor, Biological Engineering, Electrical Engineering and Computer Science
Massachusetts Institute of Technology

Computational and experimental methods for CRISPR-based saturation mutagenesis screens

by
Jonathan Yee-Ting Hsu

B.S. Bioengineering
University of California San Diego 2016

Submitted to the Department of Biological Engineering
on December 7th, 2021 in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy in Biological Engineering

Abstract
Genetic variation is a powerful framework for functional characterization of the human genome. The emergence of CRISPR technology has enabled the efficient and diverse installation of genetic variation *in situ*, leading to its widespread use in functional genomics. The application of high-throughput CRISPR saturation mutagenesis screens for the functional interrogation of the coding and non-coding genome holds great promise in accelerating our understanding of how static DNA sequences encode and influence dynamic processes in human development and disease.

In this thesis, we focus on the development of computational and experimental methods for CRISPR-based saturation mutagenesis screens. First, we developed CRISPR screening uncharacterized region function (CRISPR-SURF), a deconvolution framework for the analysis of CRISPR saturation mutagenesis screens. Drawing inspiration from the field of signal processing, we propose the modeling of CRISPR perturbations across an underlying genomic regulatory signal by means of a convolution operation and apply CRISPR-SURF for the discovery of non-coding regulatory elements involved in gene regulation. Second, we developed PrimeDesign to facilitate the rapid design of prime editing (PE) guide RNAs and demonstrate its utility by using recommended designs to install pathogenic variants in human cells. Complementing PrimeDesign, we developed pegPool as a high-throughput pooled screening strategy for prime editing guide RNA (pegRNA) optimization. We demonstrate the generalizability of pegPool by assessing a total of >18,000 pegRNA designs, with up to 210 designs in a single pool, to identify high efficiency pegRNA constructs targeting genomic sites. Finally, we developed multiplexing of site-specific alterations for *in situ* characterization (MOSAIC) as a rapid non-viral method for saturation mutagenesis screens at single-nucleotide and codon resolution. Using MOSAIC, we demonstrate *in situ* saturation mutagenesis of the *BCR-ABL1* oncogene to identify drug resistant variants and *IRF1* untranslated region (UTR) to map non-coding regulatory elements involved in transcriptional initiation.

Thesis Advisors:	J. Keith Joung	Luca Pinello
Titles:	Professor of Pathology	Associate Professor of Pathology

Acknowledgements

First, I want to thank my advisors, Prof. J. Keith Joung and Prof. Luca Pinello, for being amazing mentors throughout my time at MIT and MGH. Your balanced mentorship encouraged me to be creative in my pursuit of science, and at the same time grounded with scientific precision and integrity. Thank you for the scientific and creative autonomy. Thank you for the dynamic and vibrant research environment. Most importantly, thank you for respecting and trusting me. All these things have grown me into the curious scientist I am today.

I want to thank Prof. Jim Collins and Prof. Tim Lu for their support and guidance, and for serving on my thesis committee. Your perspectives and insights have given me scientific direction throughout my years in graduate school.

I want to thank previous and current lab members in the Joung and Pinello labs who have been like family to me. Without you all, I would not have survived graduate school. Jimmy Guo, Daniel Kim, Julian Grünewald, Karl Petri, Esther Tak, Ken Lam, Kendell Clement, Peter Cabeceiras, Ibrahim Kurt, Russell Walton, Nick Perry, Stacy Francis: thank you for all the exciting scientific discussions and inspiring me to be a better person. Justine Shih: thank you for being the greatest technician and helping me cross the finish line.

I want to thank all my scientific collaborators who made my thesis possible. Charlie Fulco, Mitchel Cole, Matt Canver, Andrew Anzalone, Max Shen, Prof. Eric Lander, Prof. Dan Bauer, Prof. David Liu: thank you for the scientific discussions and contributions which made our work possible.

I want to thank all my friends outside of lab for supporting me in diverse ways. Patrick Holec, Tyler Toth, Miguel Reyes, Emi Lutz, Konstantin Krismer, Daniel Anderson, Tatiana Netterfield, Tiffany Song, Shinyoung Lee, Calvin Lee, Erin Kim, Rachel Park, Uyanga Tsedev, Lavi Erisson, Alex Wang, Vivian Chang, James Oh, Brian Yeo, Kevin Hsu, Aneeq Malik, Cass Bonakdar, John Young: thank you for always being there for me.

Lastly, I want to thank my family for always being my foundation. Mom, Dad, Josh, Milo and Nico: I love you more than anything in the world.

Table of Contents

1 Introduction

1.1 Natural genetic variation

The Human Genome Project led to the first draft sequence of the human genome in 2003[1,2]. Fueled by significant technological advancements in next-generation sequencing (NGS), the number of whole genomes sequenced since is rapidly approaching the first million. This exponential growth in genome sequencing data is accompanied by a growing catalog of natural genetic variation[3–7]. Leveraging natural human genetic variation, genome-wide association studies (GWASs) have successfully discovered genes and biological pathways associated with a wide range of monogenic and complex diseases[8–11]. For example, several GWASs led to the identification of *BCL11A* as a major determinant of fetal hemoglobin (HbF) levels[12–14], and this insight currently serves as the basis for therapeutic strategies aimed at treating sickle cell disease (SCD) and β-thalassemia[15].

While GWASs have generated significant insights connecting genetic variation with various phenotypes, its foundation in natural genetic variation can be limited by low variation frequency and biased mutational spectrum[16–18]. Single-nucleotide polymorphisms (SNPs) make up the most prevalent type of genetic variation observed between whole genomes, but are significantly biased towards transition mutations over transversion mutations[19–21]. This simple mutational bias has the potential to limit the realm of genotype and phenotype associations discoverable by GWASs as transition mutations facilitate only a subset of amino acid changes for any given codon sequence and induce smaller disruptive effects on transcription factor binding to cognate sequence motifs compared to transversion mutations[22]. Furthermore, GWASs are better suited for the identification of a small number of genetic variants with relatively large effect sizes and are more limited in the identification of multiple genetic variants underlying highly polygenic traits[23].

Overall, GWASs have served as a powerful framework for the discovery of genes and pathways associated with dichotomous phenotypes, however genetic variants that aren't observed or exhibit small effect sizes only detectable at the molecular level are missed in these studies. Technologies and methods to functionally characterize genetic variation more sensitively would be complimentary to the insights from GWASs.

1.2 CRISPR-Cas genome and epigenome editing technologies

CRISPR-Cas systems have served as powerful scaffolds for the development of genome and epigenome editing technologies due to its programmable nature for genome targeting. Since the initial demonstration of programmable DNA cleavage by CRISPR-Cas9 nuclease in eukaryotic cells[24–28], several new classes of CRISPR-based genome and epigenome editors have been described, significantly expanding the types of targeted perturbations that can be introduced in eukaryotic cells. The various CRISPR technologies enable the introduction of genetic and epigenetic perturbations across a range of different mutation types and resolutions, and their utility ultimately depends on the application of interest.

CRISPR-Cas nucleases introduce targeted double-stranded DNA breaks (DSBs) into the genome[24–30]. Following endogenous DNA repair in the form of classical nonhomologous end-joining (c-NHEJ) and microhomology-mediated end-joining (MMEJ), these DSBs can lead to uncontrolled insertion and deletion (indel) products[31–33]. Due to the variable nature of indels, CRISPR-Cas nucleases are particularly useful in the disruption of coding and non-coding sequence elements. CRISPR-Cas nucleases are commonly used for genome-wide gene knockout screens because their indel products typically form frameshift mutations which abrogate protein function[34,35], and have also been utilized for the disruption of non-coding regulatory elements such as transcription factor binding sites involved in gene regulation[36,37].

CRISPR base editors (BEs) enable the targeted installation of point mutations without requiring DSBs[38–40]. The basic architecture of BEs consist of a Cas9 nickase fused to a ssDNA deaminase enzyme. Following engagement of BEs to its target DNA, the resulting R-loop creates a substrate for the ssDNA deaminase to catalyze a deamination event, ultimately leading the installation of up to several point mutations within the editing window. The Cas9 nickase introduces a nick on the DNA strand opposite of the deamination event, biasing DNA repair towards incorporation of the point mutation(s). Several types of base editors have been described to date, including cytosine base editors (CBE)[38], adenine base editors (ABEs)[39], and C-to-G base editors (CGBEs)[40], which are able to mediate the installation of all possible transition mutations and limited transversion mutations within a precise editing window. BEs have been utilized for gene disruption through the introduction of nonsense mutations or inactivation of splice sites[41,42], and

have also been used to install human genetic variants for functional characterization[43–45]. In contrast to CRISPR-Cas nucleases, BEs introduce very low levels of indels and are more applicable for studies looking to functionally characterize point mutations instead of insertion and deletion mutations[46–67].

CRISPR prime editors (PEs) can install all possible point mutations and short insertion and deletion edits[68,69]. To date, prime editing represents the most versatile and precise genome editing technology given its mechanism of action in directly writing DNA edits into the genome in a programmable manner. The prime editing system primarily consists of two main components in the prime editor protein and prime editing guide RNA (pegRNA). The prime editor protein is a fusion between a Cas9 nickase and an engineered M-MLV reverse transcriptase. The pegRNA resembles the architecture of the single guide RNA (sgRNA) construct used for CRISPR-Cas nucleases and BEs, but additionally has a 3' extension consisting of two elements in the primer binding site (PBS) and reverse transcription template (RTT). Following prime editor protein and pegRNA complexing and engagement with its target DNA, the Cas9 nickase introduces a single-strand break into the non-target strand (NTS) within the R-loop. This single-strand break liberates a 3' DNA end within the R-loop which can interact with the PBS element from the pegRNA 3' extension. This interaction then serves as a substrate for the engineered M-MLV reverse transcriptase to synthesize DNA according to the genetic information on the RTT element, effectively directly writing a desired edit into the genome based on the design of the RTT. Following dissociation of the prime editor complex from its target DNA, the newly-synthesized DNA flap encoding the desired edit competes with the endogenous DNA flap for incorporation. Depending on which flap is incorporated, the target DNA either reverts to its wild-type sequence (which can then be re-targeted by the prime editor) or incorporates the desired edit (which may or may not be re-targeted depending on the installed edit). The versatility and precision of prime editing technology make it well-suited for high-resolution functional characterization of genetic variants.

CRISPR interference (CRISPRi) and activation (CRISPRa) technologies are capable of modifying target gene expression through steric hinderance or inducing epigenetic changes at target loci[70–78]. CRISPRi and CRISPRa typically involve a catalytically-inactivated Cas9 (dCas9)

fused to an effector protein known to modulate transcription (e.g. VP64, KRAB). Following targeting of CRISPRi or CRISPRa to genomic loci, the introduction of epigenetic changes such as chromatin modifications can lead to modulation of gene expression. These types of CRISPR epigenome editing technologies typically exhibit greater perturbation range than CRISPR genome editing technologies, and are therefore more efficient in mapping non-coding regulatory elements at the cost of resolution.

1.3 CRISPR pooled screens for saturation mutagenesis

CRISPR-Cas genome and epigenome editing technologies have been harnessed to introduce genetic and epigenetic perturbations in human cells for characterization of the genome. The dense tiling of these CRISPR perturbations across target genomic sequences, referred to as CRISPR-based saturation mutagenesis, enables the discovery of critical sequence elements underlying protein function or gene expression[36,43,79–81]. In contrast to massively parallel reporter assays (MPRAs)[82–93] and traditional deep mutational scanning assays, CRISPR-based saturation mutagenesis strategies introduce perturbations *in situ*, and therefore allow for functional characterization of the target sequence within its endogenous context.

CRISPR-based saturation mutagenesis experiments can be conducted in a high-throughput manner using pooled screens[79]. Following the *in silico* design of sgRNAs, construction of the sgRNA library, and lentivirus production, human cells can be transduced with the lentivirus library encoding different sgRNA members at a low multiplicity of infection (MOI) to ensure that majority of cells receive only one stably-integrated sgRNA construct. This experimental setup allows for the assessment of many sgRNA perturbations individually with a single pool of cells. Following a phenotypic selection (e.g. cell viability, drug resistance, fluorescence-activated cell sorting (FACS)), targeted amplicon sequencing of the stably-integrated sgRNA constructs in the population of cells pre- and post-selection can reveal changes in the sgRNA distribution. These changes in the sgRNA distribution reflect potential functional effects of the sgRNA perturbations, and downstream statistical analyses can identify functional genomic regions based on where significant sgRNAs are selectively targeting.

2 Multi-scale discovery of non-coding regulatory elements with CRISPR-SURF

2.1 Correspondence

Tiling screens that use CRISPR–Cas technologies provide a powerful approach for the mapping of regulatory elements to phenotypes of interest[36,80,81,94–96]. Here we present CRISPR screening uncharacterized region function (CRISPR-SURF), a deconvolution framework that can be used to identify functional regulatory regions in the genome from data generated by CRISPR–Cas nuclease, CRISPR interference (CRISPRi), or CRISPR activation (CRISPRa) tiling screens. CRISPR-SURF can be run as a stand-alone command line utility (https://github.com/pinellolab/CRISPR-SURF) or as a web application (http://crisprsurf.pinellolab.org/) (**Supplementary Note 2.1**).

The methodology underlying the CRISPR-SURF framework leverages the concept that single guide RNAs (sgRNAs) represent a functional readout for base pairs within the perturbation range. This range depends on the CRISPR screening approach used: CRISPR–Cas nucleases introduce insertion and deletion (indel) mutations of varying lengths (typically <30 bp, although potentially varying with cell type), whereas CRISPRi and CRISPRa strategies may remodel chromatin structure across hundreds of nucleotides. Importantly, each CRISPR technology offers its own advantage: CRISPRi and CRISPRa strategies increase the likelihood of detecting regulatory elements, given their larger perturbation ranges, whereas CRISPR–Cas nucleases provide higher resolution on the boundaries of regulatory elements, given their sharper perturbation windows. Because each sgRNA perturbs variable-size regions around its target site, the sgRNA data from CRISPR tiling screens can be seen as imprecise measurements of an underlying genomic regulatory signal. To address this variable, we model these imprecise measurements by means of a convolution operation that accounts for the perturbation profiles associated with different CRISPR technologies.

CRISPR-SURF deconvolves tiling screen data to find the genomic regulatory signal that best explains the observed sgRNA scores given the perturbation profile and sgRNA spacing (**Fig.**

2.1). The CRISPR-SURF framework accounts for overlapping perturbation profiles between neighboring sgRNAs and leverages shared information to infer the underlying genomic regulatory signal even from noisy measurements. The exact sgRNA targeting coordinates are also taken into account, thus allowing for location-dependent statistical tests with a power that reflects the local density of sgRNAs in a region. This enables CRISPR-SURF to estimate perturbation-specific and position-specific statistical power for CRISPR tiling screens (**Supplementary Note 2.2**).

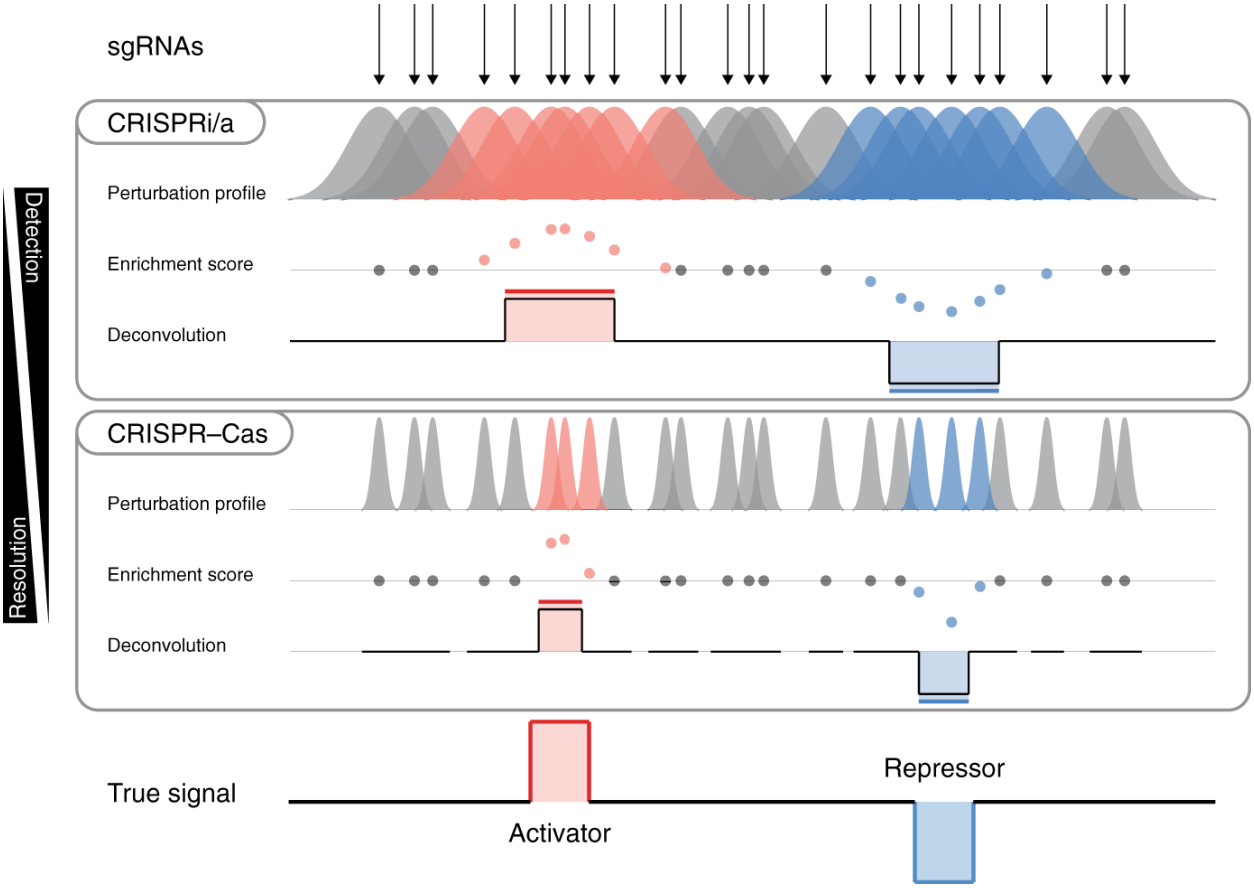


Figure 2.1: CRISPR-SURF deconvolution framework. An illustration of the deconvolution based on sgRNA targeting positions, different perturbation profiles (CRISPRi/a and CRISPR–Cas), and enrichment scores.

We evaluated the performance of CRISPR-SURF by using three published CRISPR tiling screens spanning CRISPR–Cas9[36], CRISPRi[80], and CRISPRa[81] modalities. For all three datasets, CRISPR-SURF reliably identified all of the experimentally validated regulatory elements.

CRISPR-SURF further identified potentially novel regulatory regions supported by both chromatin accessibility and epigenetic marks (**Supplementary Notes 2.3 and 2.4, Supplementary Figs. 2.1–2.3**). We elaborate on key differences between CRISPR-SURF and the analysis methods used in these previous studies in **Supplementary Notes 2.5 and 2.6**.

Furthermore, we carried out two matched CRISPR tiling screens using CRISPR–Cas9 (SpCas9) and CRISPRi (dCas9–KRAB) on the BCL11A locus (**Supplementary Note 2.7**) and found that significant regions identified within previously validated functional enhancers[36,97,98] were narrower in the CRISPR–Cas9 screen than in the CRISPRi screen, consistent with the narrower perturbation profiles of CRISPR–Cas9 indel mutations compared with those of CRISPRi epigenetic modifications (**Supplementary Fig. 2.4**). In summary, CRISPR-SURF leverages the broad CRISPRi and CRISPRa perturbation profile for efficient enhancer discovery and the narrow CRISPR–Cas perturbation profile for high-resolution mapping of critical elements within enhancers.

2.2 Supplementary notes

2.2.1 Supplementary note 2.1: CRISPR-SURF Installation and Usage

All information can be found at our GitHub page: https://github.com/pinellolab/CRISPR-SURF

Installation with Docker

With Docker, no installation is required - the only dependence is Docker itself.
Docker can be downloaded freely here:
https://store.docker.com/search?offering=community&type=edition

To get a local copy of CRISPR-SURF, simply execute the following command:

- `docker pull pinellolab/crisprsurf`

CRISPR-SURF Design

The CRISPR-SURF Design script allows users to design sgRNAs for their CRISPR tiling screens. CRISPR-SURF Design can be run in the terminal with the following command:

```
docker run -v ${PWD}/:/DATA -w /DATA pinellolab/crisprsurf SURF_design [options]
```

Users can specify the following options:

```
-bed, --bed
      Input bed file to design tiling sgRNAs. (Required)
-genome, --genome
      Input genome 2bit file. (Required)
-pams, --pams
      Specification of different CRISPR PAMs where brackets [] allow for multiple nucleotides for a given position (i.e. [ATCG]GG -> NGG, TTT[ACG] -> TTTV, [ATCG]G -> NG). Multiple PAMs separated by spaces can be inputted (i.e. [ATCG]GG TTT[ACG]). (Required)
-orient, --orientations
      Orientation of the spacer sequence relative to the PAM. This must match the length of the -pams option as an orientation must be specified for each PAM. Multiple orientations are separated by spaces (i.e. left right). (Options: left, right | Required)
-guide_l, --guide_length
      Length of the sgRNA to design. (Default: 20)
-g_constraint, --g_constraint
      Constraint forcing the 5' sgRNA bp to be G base. All guides with no 5’ G will be filtered out. (Options: true, false | Default: false)
-out, --out_dir
      Name of output directory. (Default: ./)
```

Running CRISPR-SURF Design Yourself

```
docker run -v ${PWD}/:/DATA -w /DATA pinellolab/crisprsurf SURF_design -bed BED_FILE -genome 2BIT_GENOME_FILE -pams [ATCG]GG TTT[ACG] -orient left right –out example_run
```

IMPORTANT: The BED_FILE and 2BIT_GENOME_FILE must be in the working directory where the command-line code is run.

CRISPR-SURF Count

The CRISPR-SURF Count script generates a required input file, sgRNAs_summary_table.csv, for both the CRISPR-SURF interactive website and command-line deconvolution analysis. CRISPR-SURF Count can be run in the terminal with the following command:

```
docker run -v ${PWD}/:/DATA -w /DATA pinellolab/crisprsurf SURF_count [options]
```

Users can specify the following options:

```
-f, --sgRNA_library
      Input sgRNA library file. Formatting specified below. (Required)
-control_fastqs, --control_fastqs
      List of control FASTQs with sgRNA sequencing prior to selection separated by
spaces (i.e. rep1_control.fastq rep2_control.fastq rep3_control.fastq). (Default:
None)
-sample_fastqs, --sample_fastqs
      List of sample FASTQs with sgRNA sequencing following selection separated by
spaces (i.e. rep1_sample.fastq rep2_sample.fastq rep3_sample.fastq). (Default: None)
-nuclease, --nuclease
      Nuclease used in the CRISPR tiling screen experiment. This information is used
to determine the cleavage index if indels are specified as the perturbation.
(Options: cas9, cpf1 | Default: cas9)
-pert, --perturbation
      Perturbation type used in the CRISPR tiling screen experiment. This information
is used to determine the perturbation index for a given sgRNA. (Options: indel,
crispri, crispra | Default: indel)
-norm, --normalization
      Normalization method between sequencing libraries. (Options: none, median,
total | Default: median)
-count_method, --count_method
      Counting method for sgRNAs from FASTQ. The tracrRNA option aligns a consensus
sequence directly downstream of the sgRNA. The index option uses provided indices to
grab sgRNA sequence from the sequencing reads. (Options: tracrRNA, index | Default:
tracrRNA)
-tracrRNA, --tracrRNA
       If -count_method == tracrRNA. The consensus tracrRNA sequence directly
downstream of the sgRNA for counting from FASTQ. (Default: GTTTTAG)
-sgRNA_index, --sgRNA_index
      If -count_method == index. The sgRNA start and stop indices (0-index) within
the sequencing reads (i.e. 0 20). (Default: 0 20)
-count_min, --count_minimum
      The minimum number of counts for a given sgRNA in each control sample.
(Default: 50)
```

```
-dropout, --dropout_penalty
      The dropout penalty removes sgRNAs that have a 0 count in any of the
control/sample replicates. (Default: True)
-TTTT, --TTTT_penalty
      The TTTT penalty removes sgRNAs that have a homopolymer stretch of Ts >= 4.
(Default: True)
-sgRNA_length, --sgRNA_length
      Length of sgRNAs used in the CRISPR tiling screen experiment. This must match
the sgRNA length provided in the sgRNA library file. (Default: 20)
-reverse, --reverse_score
      Reverse the enrichment score. Generally applied to depletion screens where a
positive score is associated with depletion of a sgRNA. (Default: False)
-out_dir, --out_directory
      The output directory for CRISPR-SURF counts. (Default: ./)
```

To start, you will need one of the following:

- **Option (1)** sgRNA Library File with FASTQs
- **Option (2)** sgRNA Library File with counts

Option (1):

sgRNA Library File Format Example (.CSV):

Chr	Start	Stop	sgRNA_Sequence	Strand	sgRNA_Type
chr2	60717499	60717519	AGCTCTGGAATGATGGCTTA	-	observation
chr2	60717506	60717526	ATTGTGGAGCTCTGGAATGA	+	observation
chr2	60717514	60717534	GGAGTTGGATTGTGGAGCTC	+	observation
chr2	60717522	60717542	AGAAAATTGGAGTTGGATTG	-	negative_control
chr2	60717529	60717549	CTGGAATAGAAAATTGGAGT	+	positive_control

Required Column Names:

- **Chr** - Chromosome

- **Start** - sgRNA Start Genomic Coordinate
- **Stop** - sgRNA Start Genomic Coordinate
- **sgRNA_Sequence** - sgRNA sequence not including PAM sequence
- **Strand** - Targeting strand of the sgRNA
- **sgRNA_Type** - Label for sgRNA type (observation, negative_control, positive_control)

Example CRISPR-SURF Count on Canver et al. 2015[36] for Option (1)

The following command will run CRISPR-SURF Count for Option (1) on provided example data:

```
docker run -v ${PWD}/:/DATA -w /DATA pinellolab/crisprsurf SURF_count -f
/SURF/command_line/exampleDataset/sgRNA_library_file.csv -control_fastqs
/SURF/command_line/exampleDataset/rep1_neg.fastq.gz
/SURF/command_line/exampleDataset/rep2_neg.fastq.gz -sample_fastqs
/SURF/command_line/exampleDataset/rep1_pos.fastq.gz
/SURF/command_line/exampleDataset/rep2_pos.fastq.gz -nuclease cas9 -pert indel
```

Running CRISPR-SURF Count Option (1) Yourself

Place the sgRNA library file and FASTQs in the same directory. The control FASTQs represent the sgRNA distribution prior to selection, while the sample FASTQs represent the sgRNA distribution following selection. Assuming the sgRNA library file is named `sgRNA_library_file.csv`, the FASTQs (2 replicates) are named `rep1_control.fastq`, `rep2_control.fastq`, `rep1_sample.fastq`, `rep2_sample.fastq`, and it's a CRISPR-Cas9 tiling screen, the command-line code would look like:

```
docker run -v ${PWD}/:/DATA -w /DATA pinellolab/crisprsurf SURF_count -f
sgRNA_library_file.csv -control_fastqs rep1_control.fastq rep2_control.fastq -
sample_fastqs rep1_sample.fastq rep2_sample.fastq -nuclease cas9 -pert indel
```

Simply change `-pert indel` to `-pert crispri` or `-pert crispra` for CRISPRi and CRISPRa screens, respectively.

IMPORTANT: The number of control FASTQs must equal the number of sample FASTQs. If a single control FASTQ (i.e. plasmid sequencing) is used for multiple sample FASTQs, just enumerate the `-control_fastqs` option with the same single control FASTQ.

Option (2):

sgRNA Library File Format Example (.CSV):

Chr	Start	Stop	sgRNA_Sequence	Strand	sgRNA_Type	Replicate1_Control_Count	Replicate2_Control_Count	Replicate1_Sample_Count	Replicate2_Sample_Count
chr2	60717499	60717519	AGCTCTGGAATGATGGCTTA	-	observation	322	615	131	403
chr2	60717506	60717526	ATTGTGGAGCTCTGGAATGA	+	observation	365	812	448	227
chr2	60717514	60717534	GGAGTTGGATTGTGGAGCTC	+	observation	86	169	13	129
chr2	60717522	60717542	AGAAAATTGGAGTTGGATTG	-	negative_control	1823	381	1923	321
chr2	60717529	60717549	CTGGAATAGAAAATTGGAGT	+	positive_control	54	124	355	521

Required Column Names:

- **Chr** - Chromosome
- **Start** - sgRNA Start Genomic Coordinate
- **Stop** - sgRNA Start Genomic Coordinate
- **sgRNA_Sequence** - sgRNA sequence not including PAM sequence
- **Strand** - Targeting strand of the sgRNA
- **sgRNA_Type** - Label for sgRNA type (observation, negative_control, positive_control)
- **Replicate1_Control_Count** - sgRNA Count in Replicate 1 Control FASTQ (pre-selection)
- **Replicate2_Control_Count** - sgRNA Count in Replicate 2 Control FASTQ (pre-selection)
- **Replicate1_Sample_Count** - sgRNA Count in Replicate 1 Sample FASTQ (post-selection)
- **Replicate2_Sample_Count** - sgRNA Count in Replicate 2 Sample FASTQ (post-selection)

IMPORTANT: Minimum of two experimental replicates are needed. Additional columns (ReplicateN_Control_Count, ReplicateN_Sample_Count) can be included for more experimental replicates.

Example CRISPR-SURF Count on Canver et al. 2015[36] for Option (2)

The following command will run CRISPR-SURF Count for Option (2) on provided example data:

```
docker run -v ${PWD}/:/DATA -w /DATA pinellolab/crisprsurf SURF_count -f
/SURF/command_line/exampleDataset/sgRNA_library_file_w_counts.csv -nuclease cas9 -
pert indel
```

Running CRISPR-SURF Count Option (2) Yourself

Go into the directory where the sgRNA library file is located. Assuming the sgRNA library file with counts is named `sgRNA_library_file_w_counts.csv` and it's a CRISPR-Cas9 tiling screen, the command-line code would look like:

```
docker run -v ${PWD}/:/DATA -w /DATA pinellolab/crisprsurf SURF_count -f
sgRNA_library_file_w_counts.csv -nuclease cas9 -pert indel
```

Simply change `-pert indel` to `-pert crispri` or `-pert crispra` for CRISPRi and CRISPRa screens, respectively.

IMPORTANT: Additional ReplicateN_Control_Count and ReplicateN_Sample_Count columns can be added depending on the number of replicates used in the experiment. The number of ReplicateN_Control_Count columns must equal ReplicateN_Sample_Count columns. If a single control column (i.e. plasmid count) is used for multiple sample counts, just duplicate the single control column with the appropriate column names.

CRISPR-SURF Deconvolution

The CRISPR-SURF Deconvolution command-line tool takes `sgRNAs_summary_table.csv` (generated from CRISPR-SURF Count) as input. The file requirements are stated below.

Required Column Names:

- **Chr** - Chromosome
- **Start** - sgRNA Start Genomic Coordinate
- **Stop** - sgRNA Start Genomic Coordinate
- **Perturbation_Index** - Genomic coordinate of expected perturbation center (cleavage position for CRISPR-Cas, sgRNA center for CRISPRi/a, editing window for base-editors)
- **sgRNA_Sequence** - sgRNA sequence not including PAM sequence
- **Strand** - Targeting strand of the sgRNA
- **sgRNA_Type** - Label for sgRNA type (observation, negative_control, positive_control)
- **Log2FC_Replicate1** - Replicate 1 Log2FC enrichment score of sgRNA
- **Log2FC_Replicate2** - Replicate 2 Log2FC enrichment score of sgRNA

IMPORTANT: Minimum of two experimental replicates are needed. Additional columns (Log2FC_ReplicateN) can be included for more experimental replicates.

CRISPR-SURF deconvolution can be run in the terminal with the following command:

```
docker run -v ${PWD}/:/DATA -w /DATA pinellolab/crisprsurf SURF_deconvolution
[options]
```

Users can specify the following options:

```
-f, --sgRNAs_summary_table
      Input sgRNAs summary table. Direct output of CRISPR-SURF Count. (Required)
-pert, --perturbation_type
```

The CRISPR perturbation type used in the tiling experiment. (Options: cas9, cpf1, crispri, crispra | Required)

-range, --characteristic_perturbation_range

Characteristic perturbation length. If 0 (default), the -pert argument will be used to set an appropriate perturbation range. (Default: 0)

-scale, --scale

Scaling factor to efficiently perform deconvolution with negligible consequences. If 0 (default), the -range argument will be used to set an appropriate scaling factor. (Default: 0)

-limit, --limit

Maximum distance between two sgRNAs to perform inference on bp in-between. Sets the boundaries of the gaussian profile to perform efficient deconvolution. If 0 (default), the -pert argument will be used to set an appropriate limit. (Default: 0)

-avg, --averaging_method

The averaging method to be performed to combine biological replicates. (Options: mean, median | Default: median)

-null_dist, --null_distribution

The method of building a null distribution for each smoothed beta score. (Options: negative_control, gaussian, laplace | Default: gaussian)

-sim_n, --simulation_n

The number of simulations to perform for construction of the null distribution. (Default: 1000)

-test_type, --test_type

Parametric or non-parametric test for betas. (Options: parametric, nonparametric | Default: parametric)

-lambda_list, --lambda_list

List of lambdas (regularization parameter) separated by spaces to use during the deconvolution step (i.e. 1 2 3 4 5 6 7 8 9 10). If 0 (default), the -pert argument will be used to set a reasonable lambda list. (Default: 0)

-lambda_val, --lambda_val

The lambda value to be used during the deconvolution step. If 0 (default), the -lambda_list argument will be used. (Default: 0)

-corr, --correlation

The Pearson's r correlation coefficient between biological replicates to determine a reasonable lambda for the deconvolution operation. If 0 (default), the -range argument will be used to set an appropriate correlation. (Default: 0)

-genome, --genome

```
      The genome to be used to create the IGV session file. (Options: hg19, hg38,
mm9, mm10, etc. | Default: hg19)
-effect_size, --effect_size
      Effect size to estimate statistical power. (Default: 1)
-padjs, --padj_cutoffs
      List of p-adj. (Benjamini-Hochberg) cut-offs separated by spaces for
determining significance of regulatory regions in the CRISPR tiling screen (i.e. 0.05
0.01 0.001 0.0001). (Default: 0.05 0.01 0.001 0.0001)
-out_dir, --out_directory
      The name of the output directory to place CRISPR-SURF analysis files. (Default:
CRISPR_SURF_Analysis_TIMESTAMP)
```

Example CRISPR-SURF Deconvolution on Canver et al. 2015[36]

The following command will run CRISPR-SURF deconvolution analysis on provided example data:

```
docker run -v ${PWD}/:/DATA -w /DATA pinellolab/crisprsurf SURF_deconvolution -f
/SURF/command_line/exampleDataset/sgRNAs_summary_table.csv -pert cas9
```

Running CRISPR-SURF Deconvolution Yourself

Go into the directory where the sgRNAs summary table is located. Assuming the sgRNAs summary table is named `sgRNAs_summary_table.csv` and it's a CRISPR-Cas9 tiling screen, the command-line call would look like:

```
docker run -v ${PWD}/:/DATA -w /DATA pinellolab/crisprsurf SURF_deconvolution -f
sgRNAs_summary_table.csv -pert cas9
```

Simply change `-pert cas9` to `-pert crispri` or `-pert crispra` for CRISPRi and CRISPRa screens, respectively.

Output Files

1. sgRNAs_summary_table_updated.csv: An updated sgRNAs summary table with deconvolution and p-adj. values.

2. igv_session.xml: An IGV[99] session for the following tracks

- **raw_scores.bedgraph** - sgRNA enrichment scores
- **deconvolved_scores.bedgraph** - deconvolution beta profile
- **positive_significant_regions.bed** - positive significant regions at set FDR
- **negative_significant_regions.bed** - negative significant regions at set FDR
- **neglog10_pvals.bedgraph** - negative log10 p-values for betas
- **statistical_power.bedgraph** - statistical power track at set effect size and FDR

3. significant_regions.csv: List of the significant regions and its associated statistics and supporting sgRNAs.

4. beta_profile.csv: Full deconvolution beta profile with associated statistics.

5. correlation_curve_lambda.csv: The correlation curve generated for determining lambda.

6. crispr-surf_parameters.csv: The CRISPR-SURF analysis parameters used during the analysis session.

7. crispr-surf.log: The log file for CRISPR-SURF analysis.

CRISPR-SURF Interactive Website

In order to make CRISPR-SURF more user-friendly and accessible, we have created an interactive website: http://crisprsurf.pinellolab.org. The website implements all the features of the CRISPR-SURF command-line tool (except CRISPR-SURF Count) and, in addition, provides interactive and exploratory plots to visualize your CRISPR tiling screen data.

The website offers two functions: 1) running CRISPR-SURF on data provided by the user and 2) visualizing CRISPR-SURF analysis on several published data sets, serving as the first database

dedicated to CRISPR tiling screen data. There is a 10,000 sgRNA limitation for analysis with the web application due to server capacity. Analysis of CRISPR tiling screen data with >10,000 sgRNAs requires the use of the command-line tool or provided Docker image.

The web application can also run on a local machine using the provided Docker image we have created. To run the website on a local machine after the Docker installation, execute the following command from the command line:

- `docker run -p 9993:9993 pinellolab/crisprsurf SURF_webapp`

After execution of the command, the user will have a local instance of the website accessible at the URL: http://localhost:9993

2.2.2 Supplementary Note 2.2: CRISPR-SURF Computational Methods

Design sgRNA Tiling Library

CRISPR-SURF provides a tool for the design of a sgRNA libraries for tiling screens. Given a .bed file, a genome, and PAM sequences of interest, the tool simply enumerates all possible targeting sgRNAs where the spacer or PAM sequence overlaps with the target region(s). The tool does not provide a score for the designed sgRNAs. The orientation of the spacer sequence (relative to the PAM), sgRNA length and 5' G filters are other parameters users can use to design their sgRNA library. The design tool supports all PAM sequences (including variants) for all CRISPR-Cas nucleases (Cas9, Cpf1, etc.) and sgRNAs can be designed for multiple PAMs in parallel.

The CRISPR-SURF sgRNA design tool can be used as a command-line tool (**Supplementary Note 2.1**) or on our interactive website at http://crisprsurf.pinellolab.org. On our website, users are provided with intuitive plots to understand the spacing of their tiled sgRNAs. Cumulative distribution functions (CDFs) of the distances between consecutive sgRNAs and a genomic track with sgRNA locations and their expected perturbation profiles are available. The sgRNA library can be downloaded directly from our website.

Data Pre-Processing

Several data pre-processing steps are necessary before performing CRISPR-SURF analysis. Users can either provide FASTQs to perform the data pre-processing steps outlined below with CRISPR-SURF Count, or provide a sgRNA counts file that can be directly analyzed by CRISPR-SURF.

The pre-processing steps can be broken down into (1) sgRNA counting and normalization, (2) sgRNA filtering, and (3) sgRNA enrichment scoring.

(1) sgRNA Counting and Normalization

The sgRNA counting step with FASTQ files is performed with either the tracrRNA sequence (consensus sequence directly downstream of spacer sequence) or sequencing read index. The tracrRNA sequence option allows the user to specify a consensus tracrRNA sequence directly downstream of the sgRNA sequence, allowing for the counting of sgRNAs following the alignment of the tracrRNA sequence to each sequencing read. The sequencing read index option allows the user to specify the sgRNA start and stop indices (0-index) within each sequencing read to count sgRNAs. We discourage the mapping of guide sequences directly to a reference genome since this can lead to ambiguous alignments and incorrect positioning, therefore genomic coordinates are required as input.

(2) sgRNA Filtering

The sgRNA filtering step allows the user to specify penalties associated with sgRNA count minimums, dropouts, and existence of homopolymer T stretches (>3) within the sgRNA sequence. The count minimum penalty filters sgRNAs based on its counts pre-selection to ensure there is sufficient sgRNA representation. The dropout penalty filters sgRNAs with any 0 counts in the post-selection population. The homopolymer T penalty filters sgRNAs with a stretch of >3 Ts as this is a termination signal for RNA pol III.

(3) sgRNA Enrichment Scoring

The sgRNA enrichment scoring step calculates a log_2FC value using the ratio of pre- and post-selection counts for each sgRNA per biological replicate. A pseudo-count of 1 is added to both the pre- and post- selection counts to avoid 0 values.

The CRISPR-SURF Count module can be used to perform all pre-processing steps outlined above. See https://github.com/pinellolab/CRISPR-SURF for more information.

L1-Regularized Deconvolution Framework

The deconvolution framework in CRISPR-SURF leverages L1 regularization and is adapted from the generalized lasso[100]:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{argmin} \left\{ \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|D\beta\|_1 \right\}$$

where $\beta \in \mathbb{R}^p$, $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{nxp}$, $D \in \mathbb{R}^{mxp}$, and $\lambda \geq 0$.

Using the generalized lasso, we encode the deconvolution operation as follows:

$$X = MC$$

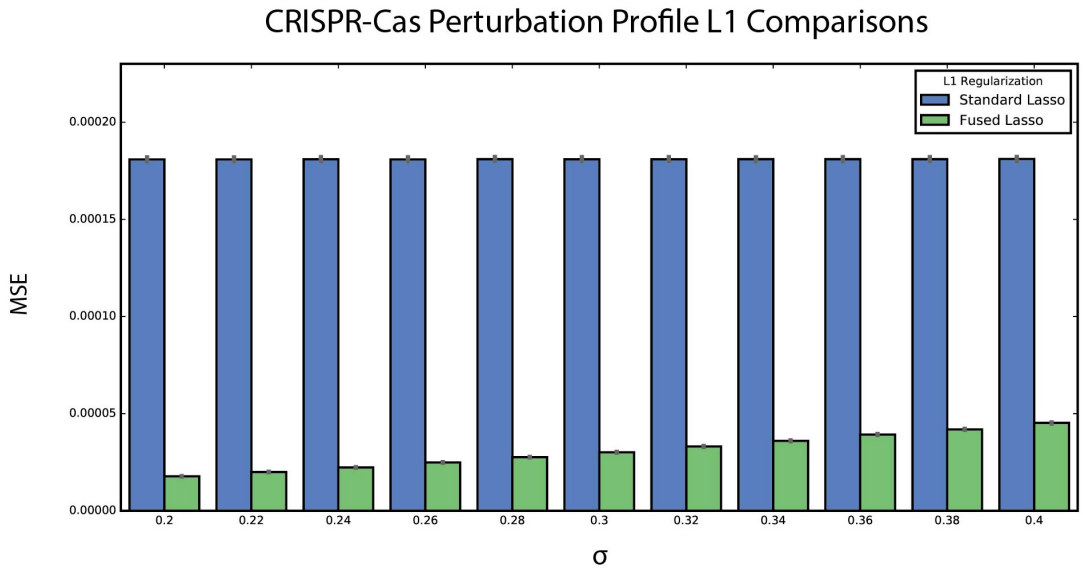
$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{argmin} \left\{ \frac{1}{2} \|y - MC\beta\|_2^2 + \lambda \|D\beta\|_1 \right\}$$

where $M \in \mathbb{R}^{nxp}$ and $C \in \mathbb{R}^{pxp}$.

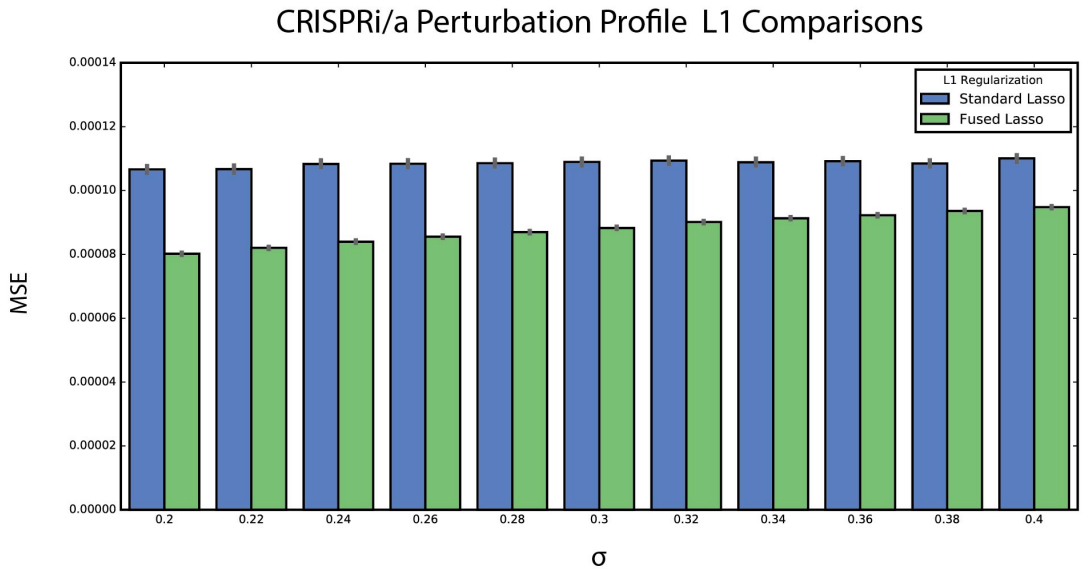
$\hat{\beta}$ is the coefficients vector where $\hat{\beta}_i$ is the inferred functional score for base-pair(s) i, y is a response vector representing the sgRNA enrichment score observations, M is the filtering matrix specifying sgRNA targeting indices, C is the convolution matrix encoding the convolution operation, D is the penalty matrix in the form of a difference matrix, and λ is the regularization parameter tuning the ℓ_1 fusion penalty (fused lasso).

To choose an L1 regularization for the deconvolution framework, we compared deconvolution accuracy by mean-squared error (MSE) between the lasso and fused lasso. Comparisons were performed across varying introduced noise, targeting density (bp per sgRNA), and default CRISPR perturbation profiles; 1000 simulations were performed per comparison (**Supplementary Figures SN2.1 and 2.2**). The fused lasso performed better than the lasso in all direct comparisons and is the L1 regularization choice for the CRISPR-SURF framework. While the lasso is reasonable for feature selection of independent signals, the fused lasso is more suited when there is a natural ordering of the underlying signal (time-series, genomic coordinates, etc.). Due to inherent spatial information in CRISPR tiling screen data, we believe this is the reason the fused lasso outperforms the standard lasso in our application. Additionally, cumulative distribution functions (CDFs) of the MSE highlight CRISPR-SURF's ability to robustly deconvolve a signal (**Supplementary Figures SN2.3 and 2.4**). The CRISPR-Cas nuclease perturbation profile exhibited greatest variance in MSE when targeting density was varied, however, CRISPR-SURF still managed to reconstruct a functional signal in 94.4% of simulations with a targeting density of 50 bp per sgRNA.

a



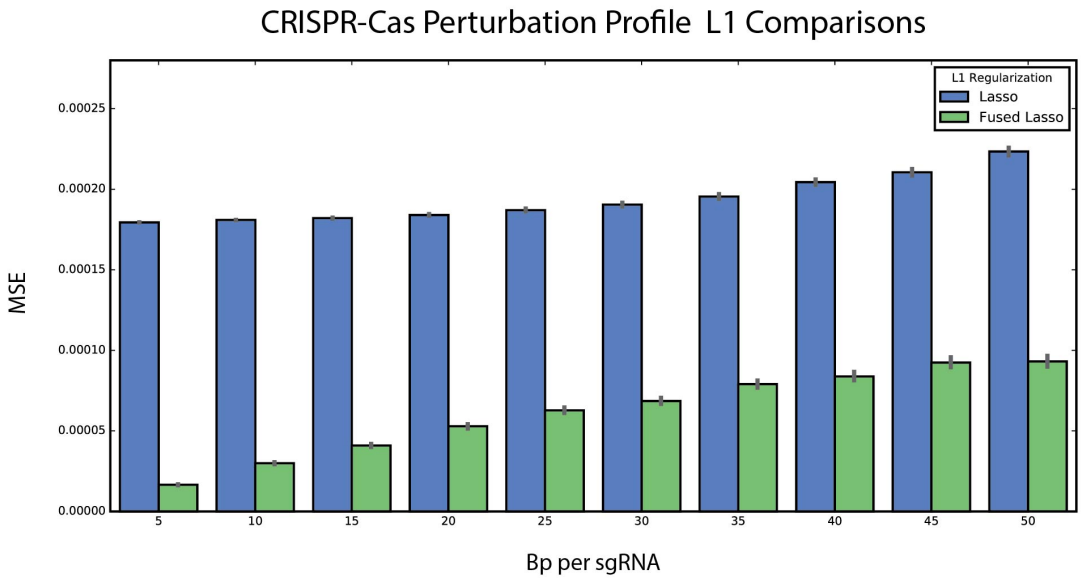
b



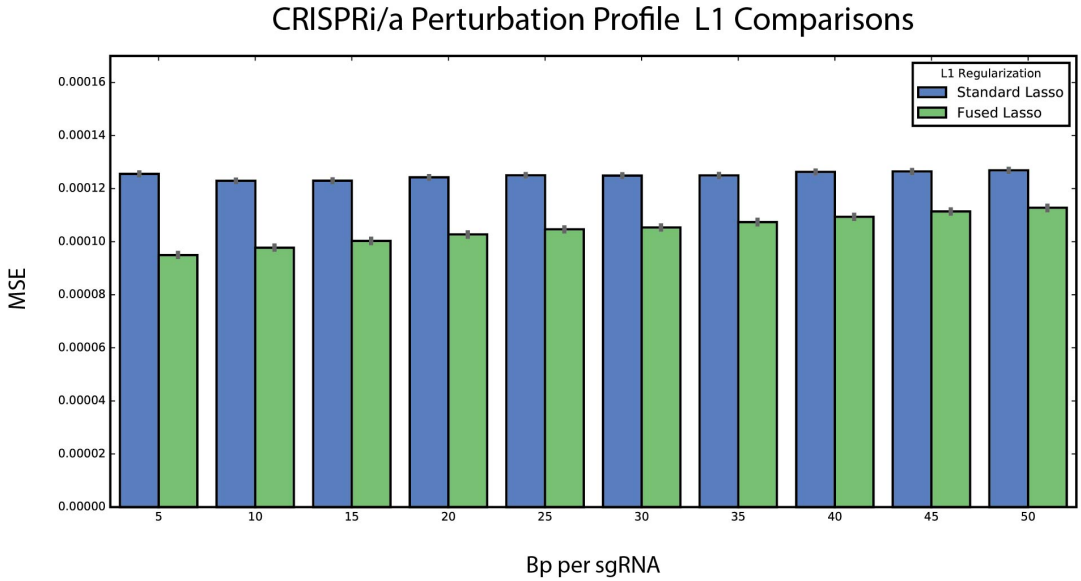
Supplementary Figure SN2.1: Comparison of L1 regularization methods with varying noise. (a) Comparison of deconvolution mean-squared error (MSE) for the lasso and fused lasso L1 regularization methods with varying Gaussian noise for the CRISPR-Cas nuclease perturbation profile. A total of 1000 simulations were performed for each comparison for both the lasso and fused lasso. Grey bars represent 95% confidence intervals. **(b)** Comparison of deconvolution mean-squared error (MSE) for the lasso and fused lasso L1 regularization methods with varying Gaussian noise for the CRISPRi/a perturbation profile. A total of 1000

simulations were performed for each comparison for both the lasso and fused lasso. Grey bars represent 95% confidence intervals.

a

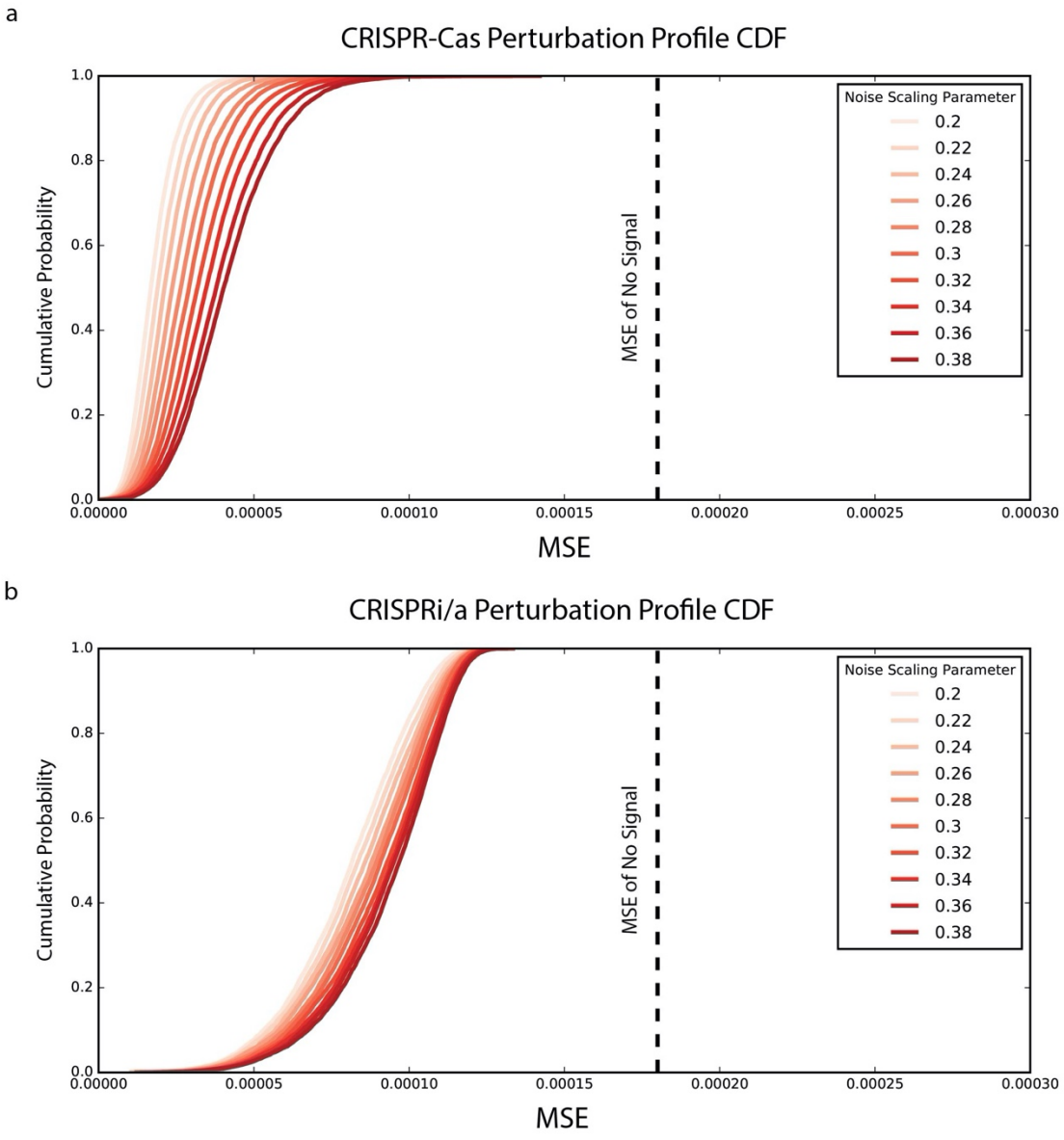


b



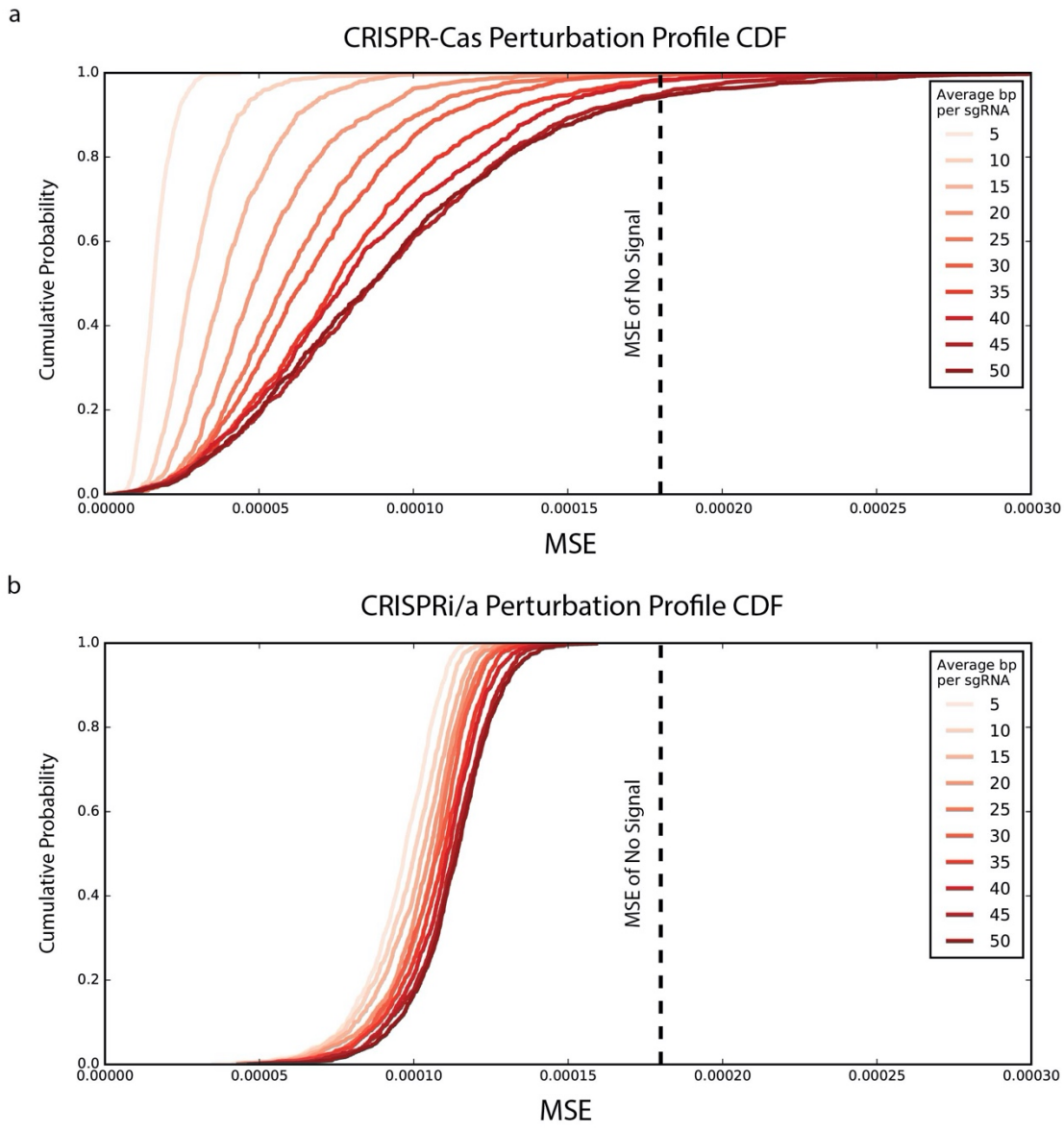
Supplementary Figure SN2.2: Comparison of L1 regularization methods with varying targeting density. (a) Comparison of deconvolution mean-squared error (MSE) for the lasso and fused lasso L1 regularization methods with varying sgRNA targeting density for the CRISPR-Cas nuclease perturbation profile. A total of 1000 simulations were performed for each comparison for both the lasso and fused lasso. Grey bars represent 95% confidence intervals. **(b)**

Comparison of deconvolution mean-squared error (MSE) for the lasso and fused lasso L1 regularization methods with varying sgRNA targeting density for the CRISPRi/a perturbation profile. A total of 1000 simulations were performed for each comparison for both the lasso and fused lasso. Grey bars represent 95% confidence intervals.



Supplementary Figure SN2.3: CDF of deconvolution MSE with varying noise. (a) Cumulative distribution function (CDF) of deconvolution mean-squared error (MSE) with varying Gaussian noise for the CRISPR-Cas nuclease perturbation profile. A total of 1000

simulations were performed for each noise scaling parameter (σ). **(b)** Cumulative distribution function (CDF) of deconvolution mean-squared error (MSE) with varying Gaussian noise for the CRISPRi/a perturbation profile. A total of 1000 simulations were performed for each noise scaling parameter (σ).



Supplementary Figure SN2.4: CDF of deconvolution MSE with varying targeting density. **(a)** Cumulative distribution function (CDF) of deconvolution mean-squared error (MSE) with varying sgRNA targeting density for the CRISPR-Cas nuclease perturbation profile. A total of

1000 simulations were performed for each targeting density (bp per sgRNA). **(b)** Cumulative distribution function (CDF) of deconvolution mean-squared error (MSE) with varying sgRNA targeting density for the CRISPRi/a perturbation profile. A total of 1000 simulations were performed for each targeting density (bp per sgRNA).

Parameterization

The convolution matrix C and regularization parameter λ need to be specified to perform the deconvolution algorithm for CRISPR-SURF analysis. The perturbation profile, encoded within C, represents the perturbation range of the CRISPR screening modality employed. The perturbation profile is represented by a Gaussian window, where a characteristic perturbation length P_L is used to parameterize the Gaussian window G.

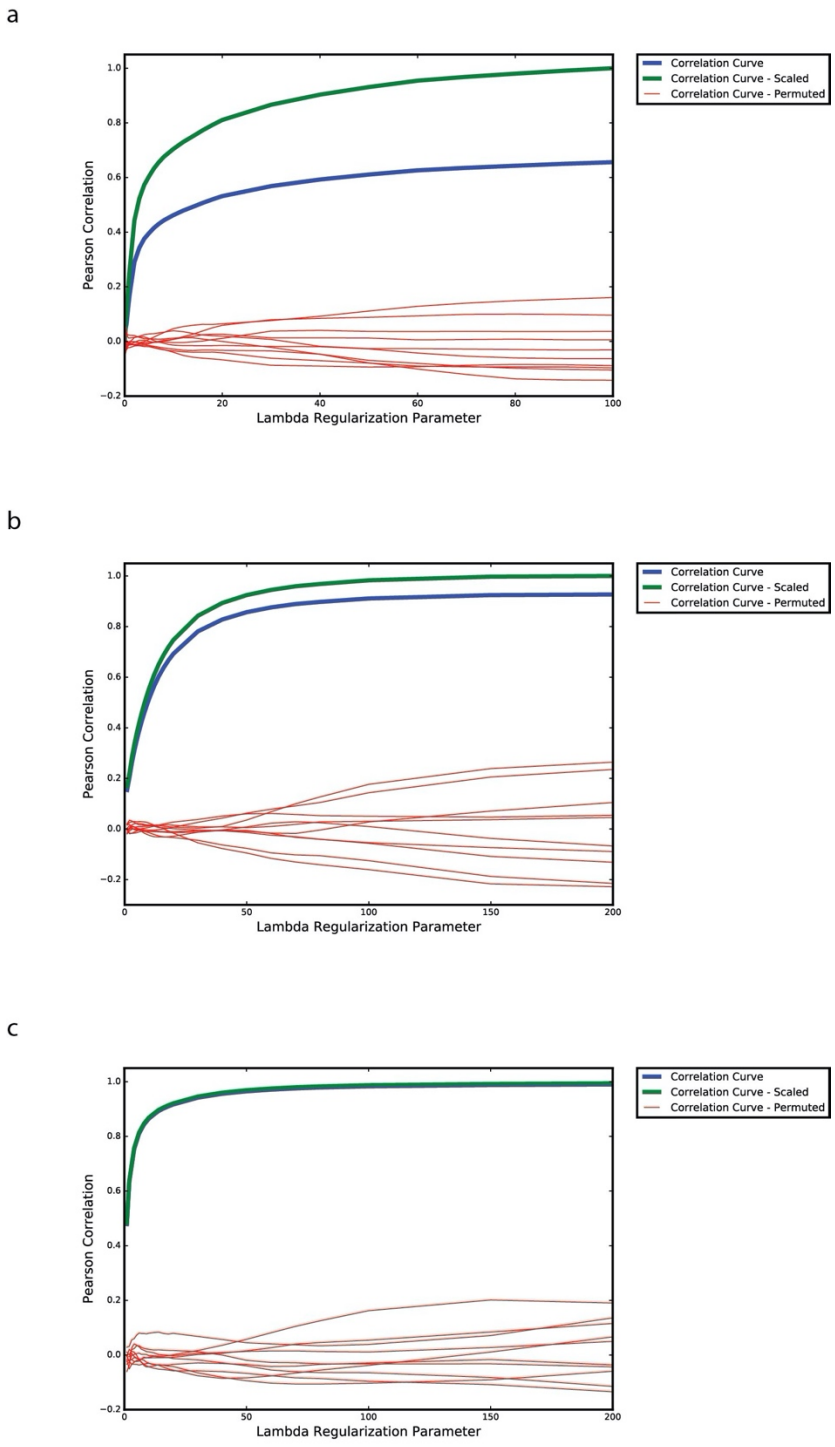
$$G(x, C) = e^{-\frac{x^2}{2C^2}}$$

$$C = \sqrt{-({P_L}^2 / 2 \ln 0.5)}$$

The selection of a characteristic perturbation length P_L is different for varying CRISPR screening modalities. CRISPR-Cas nucleases introduce indel mutations to the DNA sequence and provide a much narrower perturbation profile compared to CRISPRi and CRISPRa strategies which can epigenetically modify the chromatin landscape across hundreds of bp. Through the observation of 96 unique indel distributions for CRISPR-Cas9, the data suggests the average indel length to be around 6 - 12 bp for individual sgRNAs, and a median of 7 bp for the aggregate indel distribution[101]. Based on dCas9, dCas9-KRAB, and dCas9-VP64 characterization for sgRNAs tiled across promoter regions genome-wide, the data suggests dCas9-KRAB (CRISPRi) and dCas9-VP64 (CRISPRa) to exhibit a characteristic perturbation length of at least 200 bp and a total perturbation range of ~1 kb under the assumption that the dCas9 signal gives an estimation of the functional element bounds[102]. Importantly, we acknowledge that our method parameterizes a generalized perturbation profile for CRISPR-Cas, CRISPRi, and CRISPRa strategies, whereas these perturbation profiles may be sgRNA, locus, and cell-type-dependent[103]. We further elaborate on perturbation profiles for different CRISPR technologies in **Supplementary Note**

2.6. In future implementations, we plan on releasing a framework capable of specifying guide-specific perturbation profiles if the parameters underlying the aforementioned dependencies are elucidated.

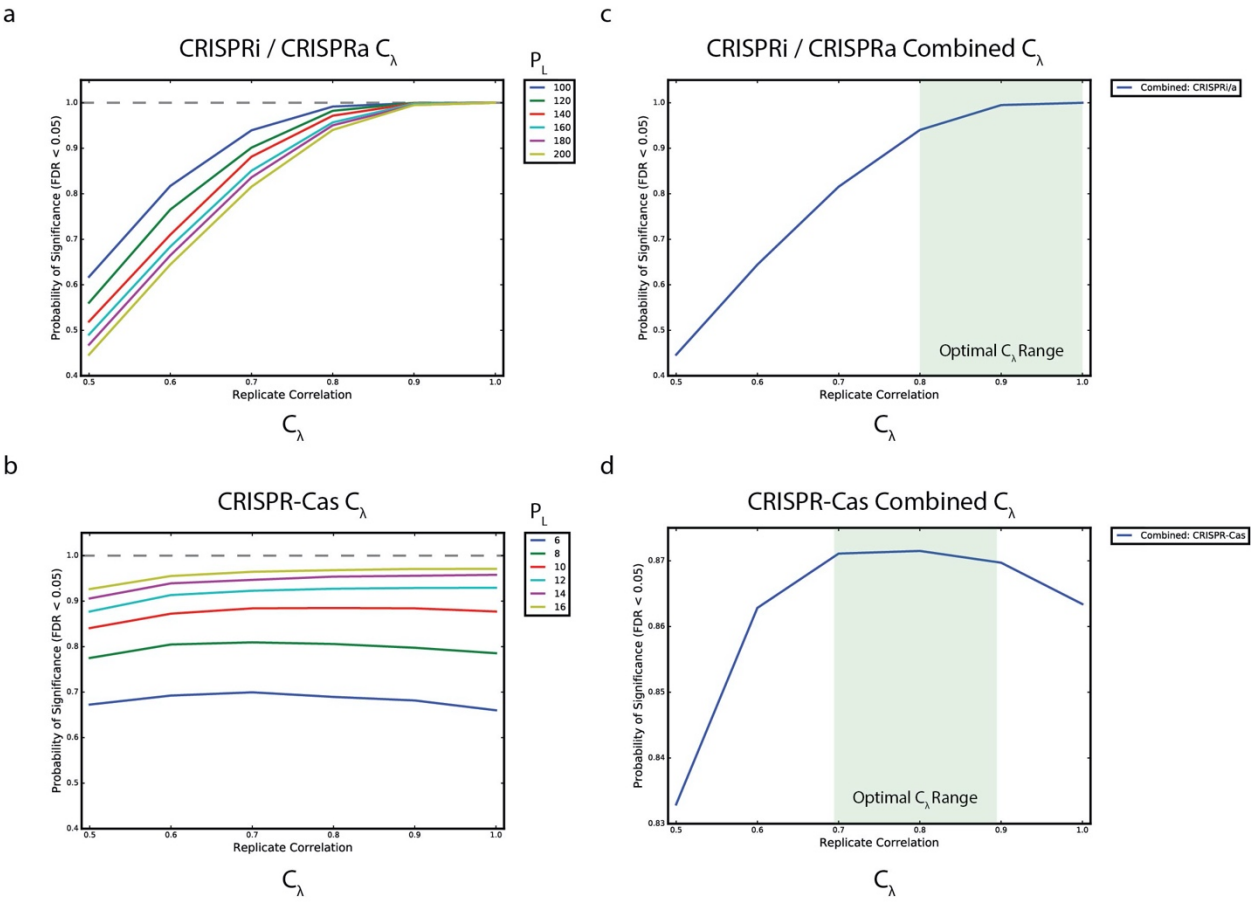
The selection of the regularization parameter λ to tune the ℓ_1 fusion penalty is done heuristically by leveraging information shared between biological replicates. The deconvolution algorithm is performed across an extensive range of λs resulting in a set of corresponding $\hat{\beta}$ coefficient vectors for each biological replicate. The Pearson's r correlation coefficient between replicate $\hat{\beta}$ vectors is then assessed for each λ across all pairwise replicate combinations, and a correlation curve is generated. Under the assumption a signal exists within the deconvolution, the correlation curve rapidly increases and then stabilizes at a near-maximum correlation with increasing λ. Deconvolutions with no signal do not recapitulate the same pattern, and are characterized by low Pearson's r correlation values across the entire λ space, illustrated by random permutations of the sgRNA observations (**Supplementary Figure SN2.5**).



Supplementary Figure SN2.5: Correlation curves of CRISPR-SURF re-analyses. The correlation curves (Pearson's r) of experimental replicates generated in **(a)** Canver et al. 2015[36] (n = 6), **(b)** Fulco et al. 2016[80] (n = 2) and **(c)** Simeonov and Gowen et al. 2017[81] (n = 2) across different λ in CRISPR-SURF analysis. The scaled correlation curve scales the original

correlation curve to have a maximum value of 1. Random permutations of sgRNA enrichment scores were used to view correlation curves with no underlying signal.

The correlation curve is used to identify a reasonable λ under the notion that the initial rapid increase in replicate correlation primarily regularizes noise, and then then stabilizes at a correlation value once λ begins to effectively regularize the true underlying signal. We refer to the correlation value for the identification of λ as C_λ. To generalize a heuristic for identifying λ from correlation curves, we scale the correlation curves and perform simulations to assess the performance of region identification across varying $C_\lambda - \lambda$ relationships. In the simulations, we find that CRISPR-Cas perturbation profiles exhibit an optimum C_λ range of 0.7 to 0.9, while CRISPRi and CRISPRa perturbation profiles exhibit an optimum C_λ range of 0.8 to 1.0 (**Supplementary Figure SN2.6**).



Supplementary Figure SN2.6: Selecting λ across CRISPR screening modalities. Simulations (n = 10000) were performed to assess probability of detecting a signal at varying C_λ values for the identification of λ for both **(a)** CRISPRi/a and **(b)** CRISPR-Cas perturbation profiles. Aggregate curves were generated across reasonable perturbation profiles for each perturbation class and optimum C_λ ranges were established as 0.8 to 1.0 and 0.7 to 0.9 for CRISPRi/a and CRISPR-Cas perturbations, respectively **(c and d)**. C_λ is the scaled correlation curve value used to determine λ, while P_L represents the characteristic perturbation length.

Parameter Robustness

To assess the robustness of parameter selection for both P_L and and C_λ, we vary both parameters in the re-analysis of three published CRISPR tiling screen data sets, and evaluate the results against regulatory regions outlined in the previous studies[36,80,81]. The P_L parameter is varied from 5 to 30 bp for CRISPR-Cas screens, and varied from 100 to 400 bp for CRISPRi and CRISPRa screens. The C_λ parameter is varied across 0.6 to 0.95 for all CRISPR screening modalities. The identification of previously-described functional elements, or reference regions, was perfect within recommended P_L (CRISPR-Cas: 5 - 20 bp, CRISPRi/a: 100 - 300 bp) and C_λ (0.8 - 0.95) values in all three studies (**Supplementary Tables SN3.4 – 3.9**). Significant reference region dropout only occurred at C_λ values of 0.6 and 0.65, which is outside the recommended C_λ values based on simulations described above.

Estimation of FDR

Assessing statistical significance of the resulting $\hat{\beta}$ is done empirically through the generation of β_{null}. The generation of β_{null} is performed by specifying a null distribution for sgRNA enrichment scores S_{null}, and then performing the deconvolution procedure on null observations $y_{null} \in \mathbb{R}^n$ randomly-sampled from S_{null}. The simulations preserve the original sgRNA targeting indices and analysis parameters used in the inference of $\hat{\beta}$.

Following the generation of β_{null}, each $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \dots, \hat{\beta}_p)$ is assessed with its respective $\beta_{null} = (\beta_{null_{1,*}}, \beta_{null_{2,*}}, \beta_{null_{3,*}}, \dots, \beta_{null_{p,*}})$ values to take into account the local spacing of supporting sgRNA observations. P-values are calculated with the following:

$$Pval._i = \frac{2}{N} \min\left\{\text{sum}\left(\beta_{null_{i,*}} \leq \hat{\beta}_i\right), \text{sum}\left(\beta_{null_{i,*}} \geq \hat{\beta}_i\right)\right\}$$

where $Pval._i$ is the p-value for base-pair(s) i, N is the total number of simulations, and $\beta_{null_{i,j}} \in \mathbb{R}^{pxN}$ is the matrix of null βs.

To account for multiple hypothesis testing, the Benjamini-Hochberg (BH) procedure is used to control FDR as it has been shown to work robustly under positive dependency[104].

Estimation of Statistical Power

The statistical power of a CRISPR tiling screen varies across the tiled space due to the non-uniform placement of sgRNAs. With the capability of assessing significance of each $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \ldots, \hat{\beta}_p)$ separately, the deconvolution framework is able to perform density-aware significance tests where a greater number of local sgRNAs increases the local power for detection of functional regulatory regions.

To give an estimation of the power underlying CRISPR tiling screens with our deconvolution framework, we assume homoscedasticity of β. Conceptually, we first replace the sgRNA scores around a position with random samples from the null distribution, then shift these sgRNA scores based on the perturbation profile and position's effect size, and finally assess significance at the position following deconvolution. Power is the fraction of the samples that pass the significance threshold at the position. Formally, we use β_{null} to construct H_0 distributions and estimate H_a as a shift of H_0, with the shift value derived from effect size e. We construct β_{ref} to harbor a functional element with effect size e, and build y_{ref} from the convolution operation between β_{ref} and the perturbation profile G, reflecting the observations of this theoretical functional element. We deconvolve y_{ref}, preserving all parameters in the inference of $\hat{\beta}$, to get $\hat{\beta}_{ref}$ and use $\hat{\beta}_{ref_i}$ as the shift value to estimate statistical power.

For given base-pair(s) i and effect size e, statistical power is estimated with the following steps:

i. Establish $H_{0_i} \in \mathbb{R}^N$ with $\beta_{null_{i,*}}$
ii. Identify critical value α within H_{0_i} that yields significance following BH correction
iii. Construct reference array $\beta_{ref} \in \mathbb{R}^p$ where $\beta_{ref_i} = e$ and $\beta_{ref_{\neq i}} = 0$
iv. Convolve β_{ref} with perturbation profile G used in the inference of $\hat{\beta}$; $\beta_{ref} * G = H$
v. Construct reference response vector y_{ref} from H
vi. Deconvolve y_{ref} with parameters used in the inference of $\hat{\beta}$ to get $\hat{\beta}_{ref}$
vii. Establish $H_{a_i} \in \mathbb{R}^N$ distribution by shifting H_{0_i} distribution by a value of $\hat{\beta}_{ref_i}$
viii. Estimate statistical power with $\frac{1}{N}\text{sum}\left(H_{a_i} \geq \alpha\right)$

2.2.3 Supplementary Note 2.3: Re-Analysis of Published Datasets

Canver et al. 2015[36]: CRISPR-Cas9 Tiling Screen

Enhancer dissection was performed with CRISPR-Cas9 saturating mutagenesis on three previously-described enhancers in DHS +55, +58, and +62 to find critical regions involved in the regulation of *BCL11A*. A total of 5 critical regions were identified in the study: 3 in DHS +55, 1 in DHS +58, and 1 in DHS +62.

All critical regions described were found with CRISPR-SURF analysis, and no additional regions were found (**Supplementary Figure 2.1**).

Fulco et al. 2016[80]: CRISPRi Tiling Screen

Enhancer discovery was performed across the *MYC* locus with CRISPRi in order to find enhancer elements regulating *MYC* expression. In the study, a total of 7 enhancer elements (e1 – e7) and 2 repressive elements (r1 and r2) were identified.

All validated enhancer elements (e1 – e7) and the repressive element (r1) located at the promoter of an isoform of *PVT1* were found with CRISPR-SURF analysis. The second repressive element (r2) described in the study did not reach statistical significance with FDR < 0.05. Two additional elements, one activating (SURF1) and one repressive (SURF2), were found with CRISPR-

SURF. Both the newly-described elements are supported by chromatin accessibility and epigenetic marks in DHS and H3K27ac peaks (**Supplementary Figure 2.2**). The repressive region SURF2 is located at the *CCDC26* promoter. Recent studies have suggested promoter-promoter competition between *PVT1* and *MYC* for an enhancer contact *in cis*, resulting in enhanced cell growth following the introduction of CRISPRi to the *PVT1* promoter[105].

Simeonov and Gowen et al 2017[81]: CRISPRa Tiling Screen

Enhancer discovery was performed across the *IL2RA* locus with CRISPRa in order to find enhancer elements that play a role in regulating *IL2RA* expression. In the study, a total of 6 CRISPRa Responsive Elements (CaREs 1 - 6) and the *IL2RA* TSS were identified to positively regulate *IL2RA* expression.

The *IL2RA* TSS and all CaREs (1 – 6) were identified with CRISPR-SURF analysis. Importantly, CRISPR-SURF uncovered sub-regions, supported by DHS and H3K27ac peaks, within the previously-described CaREs to provide higher-resolution analysis of the CRISPRa tiling screen. Furthermore, CRISPR-SURF identified a pair of repressive elements (SURF3 and SURF4) near the *PFKFB3* promoter (**Supplementary Figure 2.3**).

Supplementary Tables for CRISPR-SURF Re-Analyses

Supplementary Table SN2.1: Replicate correlations in Canver et al. 2015[36] re-analysis

The correlations (Pearson's r) across experimental replicates (n = 6) pre- and post-deconvolution for the Canver et al. 2015 study[36].

Replicate Pair 1	Replicate Pair 2	Pre-Deconvolution Correlation	Post-Deconvolution Correlation
1	2	0.237359026	0.668566608
1	3	0.000266402	0.05723732
1	4	0.286555534	0.780606768
1	5	0.302970456	0.78375776
1	6	0.324100629	0.737622575
2	3	-0.003690412	-0.185419984
2	4	0.311313532	0.629266088

2	5	0.267749374	0.596891714
2	6	0.201920327	0.585492148
3	4	0.065422331	0.122935278
3	5	0.132558463	0.280522327
3	6	0.028803555	0.235850705
4	5	0.616367784	0.898845762
4	6	0.391954497	0.765724068
5	6	0.435460336	0.792012535

Supplementary Table SN2.2: Replicate correlations in Fulco et al. 2016[80] re-analysis

The correlations (Pearson's r) across experimental replicates (n = 2) pre- and post-deconvolution for the Fulco et al. 2016 study[80].

Replicate Pair 1	Replicate Pair 2	Pre-Deconvolution Correlation	Post-Deconvolution Correlation
1	2	0.174314826	0.923716181

Supplementary Table SN2.3: Replicate correlations in Simeonov and Gowen et al. 2017[81] re-analysis

The correlations (Pearson's r) across experimental replicates (n = 2) pre- and post-deconvolution for the Simeonov and Gowen et al. 2017 study[81].

Replicate Pair 1	Replicate Pair 2	Pre-Deconvolution Correlation	Post-Deconvolution Correlation
1	2	0.651495291	0.885857846

Supplementary Table SN2.4: Assessment of characteristic perturbation length for Canver et al. 2015[36]

The characteristic perturbation length (P_L) was varied from 5 to 30 bp for the CRISPR-Cas9 tiling screen to assess impact on the ability to call significant regions (FDR < 0.05). There are a

total of 6 significant reference regions: DHS +55 (3 critical element), DHS +58 (1 critical element), DHS +62 (1 critical element), and *BCL11A* exon 2.

P_L (bp)	Overlap against Reference (max 6)
5	6
7	6
10	6
15	6
20	6
25	6
30	5

Supplementary Table SN3.5: Assessment of characteristic perturbation length for Fulco et al. 2016[80]

The characteristic perturbation length (P_L) was varied from 100 to 400 bp for the CRISPRi tiling screen to assess impact on the ability to call significant regions (FDR < 0.05). There are a total of 9 significant reference regions: e1, e2, e3, e4, e5, e6, e7, *MYC* TSS, *PVT1* TSS.

P_L (bp)	Overlap against Reference (max 9)
100	9
150	9
200	9
250	9
300	9
350	9
400	8

Supplementary Table SN3.6: Assessment of characteristic perturbation length for Simeonov and Gowen et al. 2017[81]

The characteristic perturbation length (P_L) was varied from 100 to 400 bp for the CRISPRa tiling screen to assess impact on the ability to call significant regions (FDR < 0.05). There are a total of

7 significant reference regions: CaRE 1, CaRE 2, CaRE 3, CaRE 4, CaRE 5, CaRE 6, *IL2RA* TSS.

P_L (bp)	Overlap against Reference (max 7)
100	7
150	7
200	7
250	7
300	7
350	7
400	7

Supplementary Table SN3.7: Assessment of $C_\lambda - \lambda$ for Canver et al. 2015[36]

The $C_\lambda - \lambda$ relationship (Pearson's r) was varied from 0.6 to 0.95 for the CRISPR-Cas9 tiling screen to assess impact on the ability to call significant regions (FDR < 0.05). There are a total of 6 significant reference regions: DHS +55 (3 critical element), DHS +58 (1 critical element), DHS +62 (1 critical element), and *BCL11A* exon 2.

C_λ	Overlap against Reference Regions (max 6)
0.6	2
0.65	3
0.7	6
0.75	6
0.8	6
0.85	6
0.9	6
0.95	6

Supplementary Table SN3.8: Assessment of $C_\lambda - \lambda$ for Fulco et al. 2016[80]

The $C_\lambda - \lambda$ relationship (Pearson's r) was varied from 0.6 to 0.95 for the CRISPRi tiling screen to assess impact on the ability to call significant regions (FDR < 0.05). There are a total of 9 significant reference regions: e1, e2, e3, e4, e5, e6, e7, *MYC* TSS, *PVT1* TSS.

C_λ	Overlap against Reference Regions (max 9)
0.6	9
0.65	9
0.7	9
0.75	9
0.8	9
0.85	9
0.9	9
0.95	9

Supplementary Table SN3.9: Assessment of $C_\lambda - \lambda$ for Simeonov and Gowen et al. 2017[81]

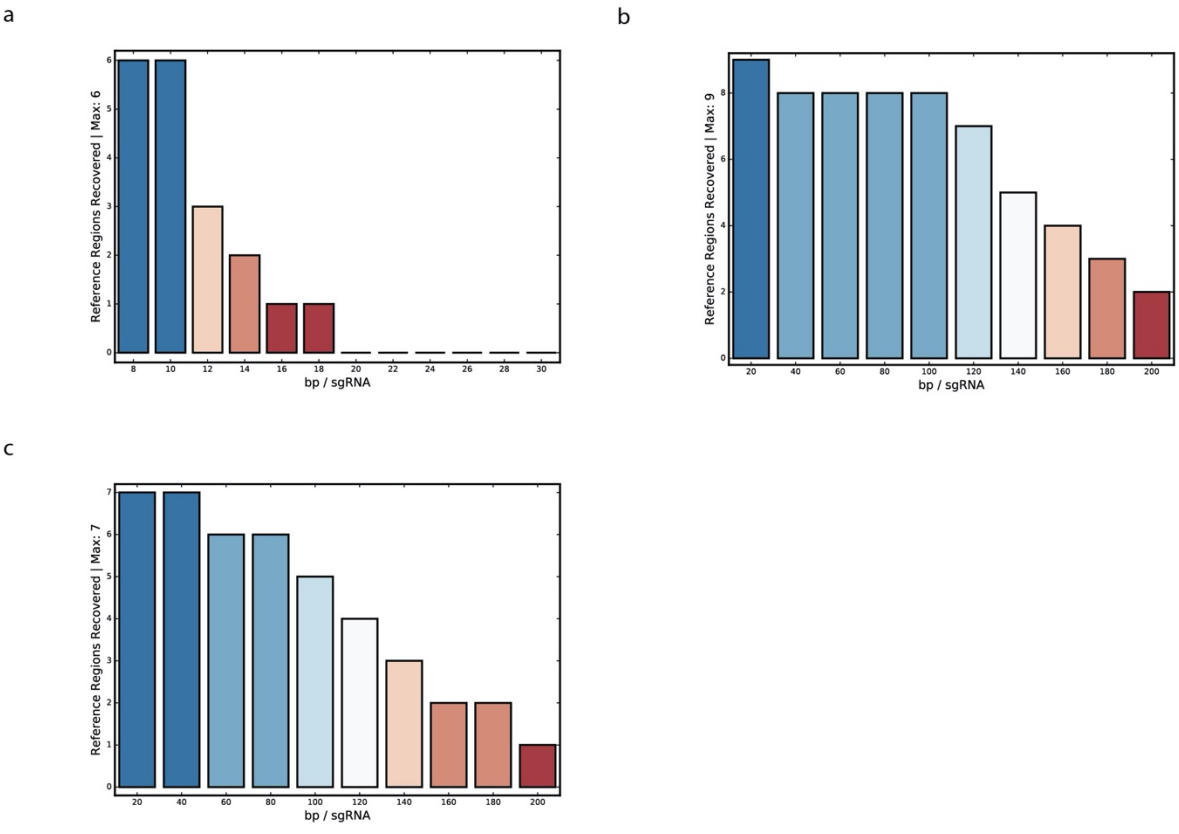
The $C_\lambda - \lambda$ relationship (Pearson's r) was varied from 0.6 to 0.95 for the CRISPRi tiling screen to assess impact on the ability to call significant regions (FDR < 0.05). There are a total of 7 significant reference regions: CaRE 1, CaRE 2, CaRE 3, CaRE 4, CaRE 5, CaRE 6, *IL2RA* TSS.

C_λ	Overlap against Reference Regions (max 7)
0.6	0
0.65	0
0.7	7
0.75	7
0.8	7
0.85	7
0.9	7
0.95	7

2.2.4 Supplementary Note 2.4: Downsampling Simulations

Downsampling sgRNA Library

Simulations were performed to understand the effect of downsampling the sgRNA library, to establish guidelines for more-efficient screening strategies (**Supplementary Figure SN4.1**). The downsampling procedure was performed by aiming to maintain the most-homogenous sgRNA coverage across the total tiling region of interest. In other words, sgRNAs were removed iteratively from the sgRNA library based on the local density of sgRNAs around its target site, determined by the sum of distances of the K nearest sgRNAs to its left and right. The choice of K=5 nearest sgRNAs in both directions allowed a robust and reproducible downsampling procedure. The sgRNA with the smallest distance metric is removed from the sgRNA library, the distance metric is then recalculated for all sgRNAs, and this procedure iterates until a target downsampling value (bp per sgRNA) is achieved.



Supplementary Figure SN4.1: Effects of downsampling sgRNA library. The identification of reference regions (FDR < 0.05) as a function of increasing downsampling of respective sgRNA

libraries in **(a)** Canver et al. 2015[36], **(b)** Fulco et al. 2016[80,81] and **(c)** Simeonov and Gowen et al. 2017. Reference regions are identified if ≥50% of the region is recovered in the CRISPR-SURF analysis. The downsampling metric is defined as the number of bp per sgRNA.

The downsampling simulations were performed on the three studies described in **Supplementary Note 2.3**. The subsequent analyses required ≥50% of the previously-described regions, or reference regions, to overlap the downsampled region calls (FDR < 0.05). The CRISPR-Cas9, CRISPRi, and CRISPRa screens started at a density of 8, 20, and 20 bp per sgRNA.

The CRISPR-Cas9 screen was sensitive to sgRNA downsampling, and was only able to maintain calling all reference regions with 83% (10 bp per sgRNA) of its original sgRNA library. Significant region dropout occurred with 69% (12 bp per sgRNA) of its original sgRNA library, and resulted in the ability to only call 50% of the originally recovered regions. Below 41% (20 bp per sgRNA) of the original sgRNA library, the analysis is not able to recover any reference regions.

For the CRISPRi screen, CRISPR-SURF was able to efficiently call the reference regions, even with aggressive sgRNA downsampling. The dropout of the r1 (PVT1 TSS) element occurred immediately at the first downsampling metric of 40 bp per sgRNA, however, it's important to note this region exhibited the smallest effect size and was not experimentally-validated in the previous study. The other 8 reference regions (*MYC* TSS and e1 – e7) were experimentally-validated elements and were called up to 100 bp per sgRNA. This translates to a downsampled sgRNA library that is only 22% of the original sgRNA library. Below 14% (160 bp per sgRNA) of the original downsampled sgRNA library, <50% of the reference regions were called.

The CRISPRa screen was also fairly efficient in calling reference regions with significant sgRNA downsampling. With a downsampled sgRNA library making up only 48% (40 bp per sgRNA) of the original sgRNA library, all 7 reference regions were called. Below 14% (140 bp per sgRNA) of the original downsampled sgRNA library, <50% of the reference regions were called.

The dropout of reference regions across the CRISPR-Cas9, CRISPRi, and CRISPRa tiling screens is a function of element effect size and width. Elements exhibiting lower regulatory function and supported by fewer sgRNAs were more susceptible to downsampling simulations. The simulations highlight the importance of sgRNA density in CRISPR-Cas tiling screens; a moderate reduction in the original sgRNA library can result in significant reference region dropout. For CRISPRi and CRISPRa tiling screens, there are strong opportunities for the design of more-efficient and cost-effective screens. The downsampling simulations show that reference region identification is nearly perfect (CRISPRi: 8/9 reference regions called, CRISPRa: 7/7 reference regions called) even with less than 50% of the original sgRNA libraries.

Supplementary Tables for sgRNA Downsampling Analyses

Supplementary Table SN4.1: Downsampling sgRNAs assessment for Canver et al. 2015[36]. The sgRNA library was downsampled across 8 to 30 bp per sgRNA for the CRISPR-Cas9 tiling screen to assess impact on the ability to call significant regions (FDR < 0.05). There are a total of 6 significant reference regions: DHS +55 (3 critical element), DHS +58 (1 critical element), DHS +62 (1 critical element), and *BCL11A* exon 2.

Bp per sgRNA	Overlap against CRISPR-SURF Calls (max 6)	Overlap against Reference (max 6)
8	6	6
10	6	6
12	3	3
14	2	2
16	1	1
18	1	1
20	0	0
22	0	0
24	0	0
26	0	0
28	0	0
30	0	0

Supplementary Table SN4.2: Downsampling sgRNAs assessment for Fulco et al. 2016[80].
The sgRNA library was downsampled across 20 to 200 bp per sgRNA for the CRISPRi tiling screen to assess impact on the ability to call significant regions (FDR < 0.05). There are a total of 9 significant reference regions: e1, e2, e3, e4, e5, e6, e7, *MYC* TSS, *PVT1* TSS.

Bp per sgRNA	Overlap against CRISPR-SURF Calls (max 12)	Overlap against Reference (max 9)
20	12	9
40	8	8
60	8	8
80	8	8
100	8	8
120	3	7
140	2	5
160	1	4
180	0	3
200	0	2

Supplementary Table SN4.3: Downsampling sgRNAs assessment for Simeonov and Gowen et al. 2017[81]. The sgRNA library was downsampled across 20 to 200 bp per sgRNA for the CRISPRi tiling screen to assess impact on the ability to call significant regions (FDR < 0.05). There are a total of 7 significant reference regions: CaRE 1, CaRE 2, CaRE 3, CaRE 4, CaRE 5, CaRE 6, *IL2RA* TSS.

Bp per sgRNA	Overlap against CRISPR-SURF Calls (max 22)	Overlap against Reference (max 7)
20	22	7
40	17	7
60	13	6
80	12	6
100	11	5

120	8	4
140	7	3
160	5	2
180	4	2
200	4	1

2.2.5 Supplementary Note 2.5: Limitations of Previous Analysis Methods for CRISPR Tiling Screens

Various analysis methods have been proposed for CRISPR tiling screen data. We focus on describing theoretical concerns of methods from Canver et al. 2015[36] (CRISPR-Cas9), Fulco et al. 2016[80] (CRISPRi), and Simeonov and Gowen et al. 2017[81] (CRISPRa) in order to highlight the motivations for the development of CRISPR-SURF. The analysis methods for CRISPRi/a and CRISPR-Cas9 data are different, therefore we split them up into different sections below.

CRISPRi and CRISPRa Method | Moving Average

To our knowledge, the only method that has been proposed for the analysis of CRISPRi and CRISPRa tiling screens is the moving average. The moving average is a naïve way of smoothing a signal and incorporating spatial information into the analysis by assigning sgRNA scores based on the average of a certain number of surrounding sgRNAs. The number of sgRNAs that go into each "averaging window" is the only parameter for the moving average. Although the moving average is effective for smoothing noisy signals, there are a couple assumptions that are violated when applied to CRISPR tiling screen data.

First, the moving average assumes the sgRNAs are uniformly-spaced across the tiled region by fixing the number of sgRNAs that go into each averaging window. This assumption is violated due to the non-uniform placement of sgRNAs across a tiling region (PAM constraints). This is problematic because regions where fewer sgRNAs can be designed will have much larger averaging window lengths compared to regions where more sgRNAs can be designed. This implies that the perturbation range is connected to sgRNA tiling density. The purpose of the moving average is to combine sgRNA scores with shared information due to their targeting

proximities in order to smooth the signal, however, this quickly becomes problematic because sets of averaged sgRNAs have varying distances between them.

Second, the moving average assumes each sgRNA within an averaging window contributes equally in perturbing a functional element (if one exists) near the center of the averaging window. If a functional element is small in size relative to the averaging window length, this leads to dilution of a signal as sgRNAs near the boundaries of the averaging window will have little effect on the functional element. This is problematic because quantitation for the effect size of a regulatory element is dependent on element length with the use of a moving average.

Lastly, statistical analyses following the moving average are either absent or assume equal statistical power across all genomic regions in previous studies. In Simeonov and Gowen et al. 2017[81], no statistics were provided for the discovery of CRISPRa Responsive Elements (CaREs) 1 – 6 as they were likely called by visual inspection after applying a 5-gRNA moving average to the raw sgRNA enrichment scores. In Fulco et al. 2016[80], a t-test was used to assess significance between scores generated from a 20-gRNA moving average on the tiled genomic region and 20 randomly-selected sgRNAs from the non-targeting sgRNA control population. The use of the t-test in this context assumes equal power across the tiling screen with a set 20 sgRNA observations in both samples. In reality, the power at any given region depends on the number of relevant sgRNA observations. Due to the non-uniform spacing of sgRNAs, the number of sgRNAs that perturb any given genomic region will vary, which is a property that the moving average fails to incorporate into its statistical analysis.

CRISPR-Cas9 Method | Hidden Markov Model

A Hidden Markov Model (HMM) was proposed in Canver et al. 2015[36] to analyze CRISPR-Cas9 tiling screen data. There are many theoretical concerns that arise when using a HMM for the analysis of CRISPR tiling screen data. The proposed HMM architecture requires uniformly spaced observations as input to infer underlying genomic regulatory states, and this is done by pre-processing the sgRNA enrichment scores with LOESS smoothing. The LOESS smoothed signal is then treated as a continuous signal and uniformly sampled as input into the HMM model, completely disregarding the original placement of the sgRNAs. This is problematic

because inference can be performed on genomic regions where sgRNAs aren't actually targeted, and additionally assumes equal statistical power across the tiling screen.

Furthermore, the proposed HMM architecture has very strong limitations in its parameterization that can be broken up into an assumption and initialization problem. The assumption problem with the proposed HMM lies in the fact that a researcher must pre-determine the genomic regulatory states that are possible in the data. Though it is reasonable to assume Neutral, Active, and Repressive states for genomic regulatory regions, these assumptions greatly impact the analysis if the pre-chosen states are not present in the data. For instance, if a Repressive state is specified in the HMM architecture, but a Repressive state is not present in the data (all regulatory regions are Active or Neutral), the HMM will force this state to exist when inferring the genomic regulatory states. The proposed HMM model is also highly-sensitive to parameter initialization, which is required when running the Baum-Welch algorithm to infer the unknown parameters of the HMM. Fine-tuning of the parameter initialization is often required to achieve satisfactory results with the proposed HMM model.

Lastly, it's important to note that the methods described above do not necessarily model CRISPR tiling screen data, but rather focus on data smoothing and subsequent significance testing on the smoothed signal.

2.2.6 Supplementary Note 2.6: Motivation for CRISPR-SURF

The main motivation behind the development of CRISPR-SURF was to address theoretical concerns associated with previously-described methods for the analysis of CRISPR tiling screen data. As mentioned in **Supplementary Note 2.5**, a common limitation in previous methods is the use of arbitrary smoothing approaches as a pre-processing step before statistical analysis. These smoothing operations aggregate information across observations with no understanding of the perturbation range and spacing of sgRNAs, which are key experimental parameters that determine the degree of shared information between sgRNAs and the power underlying a statistical test for a given genomic region. During the development of CRISPR-SURF, we focused on eliminating the need arbitrary smoothing, parameterizing key experimental

parameters into the analysis, and modeling sgRNA enrichment scores as observations stemming from an underlying genomic regulatory signal.

Convolution Operation

In contrast to directly smoothing sgRNA enrichment scores, we focused our modeling approach on reconstructing a genomic regulatory signal (deconvolution) that best explains the observed sgRNA enrichment scores. Conceptually, each sgRNA enrichment score represents a functional read-out for base pairs within its perturbation range. These functional read-outs are a distortion of the underlying genomic regulatory signal because of the variability in editing outcomes for each sgRNA as each sgRNA is represented many times in the experiment. We model the generation of tiled sgRNA enrichment scores by means of a convolution operation because this modeling choice captures the perturbation variability of each sgRNA and preserves spatial information of all the designed sgRNAs. We apply a L1-regularized deconvolution framework to reconstruct the underlying genomic regulatory signal after modeling CRISPR tiling screen data by means of a convolution operation.

Modeling CRISPR tiling screen data by means of a convolution operation allows for several advantages. First, the convolution operation models each sgRNA enrichment score independently. Theoretically, this is important because each cell in the experiment receives a single gRNA, and therefore only experiences the perturbation effects of a single gRNA. Furthermore, modeling the sgRNA enrichment scores as independent allows for the preservation of the exact genomic targets of all the designed sgRNAs as the enrichment scores don't need to be averaged prior to statistical analysis.

Next, the convolution operation readily-adapts to varying sgRNA targeting densities (non-uniform spacing) intrinsic in CRISPR tiling screen data due to sgRNA design limitations (PAM constraints). This is important because the degree of shared information used for the reconstruction of the genomic regulatory signal is finely-tuned based on the local targeting density. For example, genomic region scores with low targeting density will be reconstructed with relatively independent sgRNA observations, while genomic region scores with high targeting density will be reconstructed with a greater degree of shared information between

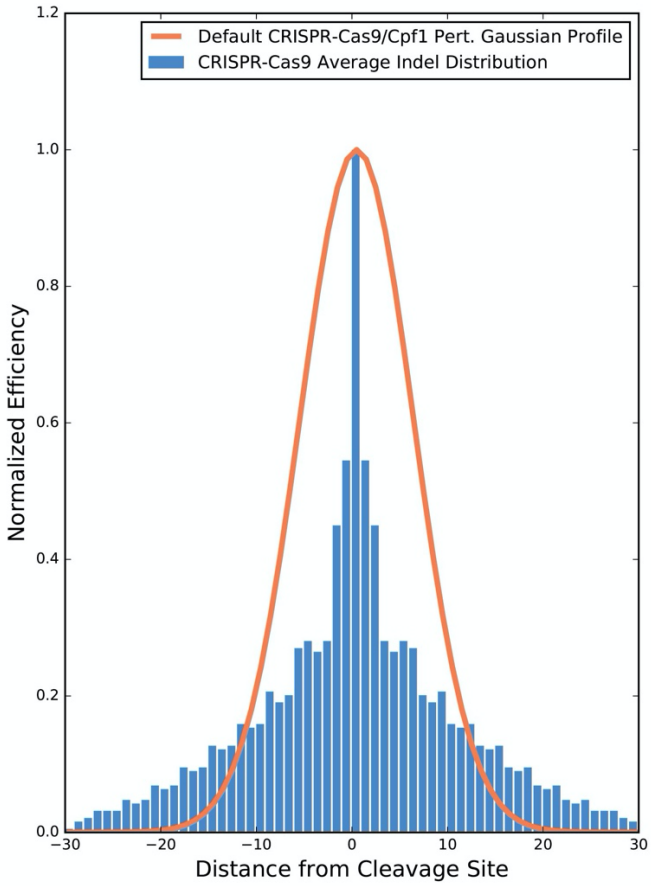
sgRNAs. This is in contrast to the moving average and proposed HMM model which destroys this spatial information (**Supplementary Note 2.5**).

Lastly, the convolution operation allows for adequately-powered statistical tests dependent on the targeting density for a genomic region. The power underlying statistical tests at different regions should vary because of the non-uniform spacing of sgRNAs. For example, a region with high targeting density will have greater power to achieve statistical significance compared to a region with low targeting density because of the increased number of supporting sgRNA observations. The incorporation of statistical power into the analysis provides increased detection sensitivity at regions with high targeting density, and additionally informs on the possibility of false negatives at regions with low targeting density. This is in contrast to the assumptions of equal power for statistical tests with the moving average and proposed HMM model (**Supplementary Note 2.5**).

CRISPR Perturbation Profiles

The usage of the convolution operation to model CRISPR tiling screen data requires knowledge on the different perturbation ranges for different CRISPR technologies; we refer to this as the perturbation profile. Genetic perturbations using CRISPR-Cas nucleases (Cas9, Cas12a, etc.) introduce indel mutations that can be readily observed by targeted amplicon sequencing, while epigenetic perturbations using CRISPRi/CRISPRa remodel chromatin and its effects can be seen in chromatin accessibility assays and ChIP-seq of histone modifications.

CRISPR-Cas genome editing has been well-characterized with targeted amplicon sequencing by next-generation sequencing (NGS) technology. Though indel profiles vary from target to target, the majority of indel mutations are relatively short (<30 bp) and centered around the cleavage site of the CRISPR-Cas nuclease. A recent study characterized the indel profiles of >40,000 sgRNAs and >1,000,000,000 mutational outcomes for CRISPR-Cas9[33]. In **Supplementary Figure SN6.1**, we show this average indel profile overlaid with our default CRISPR-Cas perturbation profile. We provide the average indel profile from this study as a perturbation profile to use in CRISPR-SURF analysis.



Supplementary Figure SN6.1: CRISPR-Cas9 average indel profile and default perturbation profile. An average CRISPR-Cas9 indel profile constructed from >40,000 sgRNAs[33] (blue histogram) overlaid with the default CRISPR-Cas perturbation profile (orange curve) in CRISPR-SURF analysis.

Targeted epigenetic modifications by CRISPRi and CRISPRa have been less-characterized, however, we point to several pieces of experimental evidence that allow us to reasonably infer the perturbation range of these technologies. Chromatin accessibility assays and ChIP-seq of histone modifications have been used to assess the epigenome-modifying capabilities of CRISPRi. In a previous study[103], it's been show that targeting of dCas9-KRAB to enhancer elements results in a decrease in DNase-seq signal (associated with euchromatin) and an increase in H3K9me3 (histone modification associated with heterochromatin). The data presented suggests dCas9-KRAB perturbations spread contiguous H3K9me3 signal spanning ~1.2 kb.

Another previous study examined the effects of both CRISPRi and CRISPRa tiled across the promoter region of genes, and assessed the effects of both epigenetic-editing technologies as a function of distance to the transcription start site (TSS)[102]. Importantly, in this study, dCas9 was used as a control to map functional regions around the TSS. We assume dCas9 does not have the ability to remodel chromatin, and therefore provides relatively fine-mapping of the underlying regulatory region conferring function to the TSS regions. The data suggests that both the CRISPRi and CRISPRa perturbations start affecting functional elements up to ~500 bp away from both directions (left and right of the TSS). Furthermore, we note that there is a monotonic increase in functional signal as the sgRNA target trends closer to the TSS from both directions. The signal peaks when the CRISPRi and CRISPRa sgRNAs target directly over the TSS functional element. This further supports the convolution operation as a reasonable approximation for modeling CRISPR tiling screen data.

In summary, both studies characterizing CRISPRi/a technologies suggest similar perturbation ranges, despite using different cell types and genomic loci. By profiling H3K9me3 marks following introduction of a targeted CRISPRi perturbation, the data suggests contiguous H3K9me3 signal spanning ~1.2 kb stemming from the targeted sgRNA. When assessing CRISPRi and CRISPRa effects as a function of distance from TSS functional elements, the data suggests that both CRISPRi and CRISPRa technologies start perturbing functional elements up to ~500 bp from both directions, leading to a perturbation profile spanning ~1kb.

2.2.7 Supplementary Note 2.7: Experimental Methods

Design and Synthesis of Lentiviral sgRNA Libraries

The sgRNA library for both the CRISPR-Cas9 and CRISPRi screen was constructed analogously to prior screens[36,37]. The summit of every DNase I hypersensitive site (DHS) within the *BCL11A* region (n = 55) was identified from fetal- and adult-derived $CD34^+$ subject to erythroid differentiation. The targeted genomic region included 2 Mb upstream of *BCL11A* encompassing a large deletion proximal to *BCL11A* reported to phenocopy *BCL11A* haploinsufficiency[106]. The regions of DHS summit +/− 200 bp were chosen for saturating mutagenesis based on previous work that suggested functional sequence tended to be located within 200 bp of the peak of DNase I hypersensitivity[36]. Using the *DNA Striker* tool[37], every 20-

mer sequence upstream of an NGG PAM sequence on the sense or anti-sense strand was identified for each *BCL11A* region DHS as well as *BCL11A* exon 2, resulting in the design of 3943 total sgRNAs (including non-targeting negative control guides).

Oligonucleotides were synthesized by microarray. The oligos were batch cloned to lentiGuide-Puro (Addgene plasmid ID 52963) as well as a modified version of lentiGuide-Puro in which the guide RNA scaffold was replaced by a structurally optimized form (A-U flip and stem extension, called combined modification) previously reported to increase the efficiency of Cas9 targeting[107]. Plasmid libraries were sequenced to 1656 and 1392 reads per guide coverage for the original and combined modification libraries, respectively, to demonstrate adequate representation. We generated lentivirus in HEK293T cells and titered on HUDEP-2 cells to identify the amount of virus required to achieve 0.3-0.5 MOI.

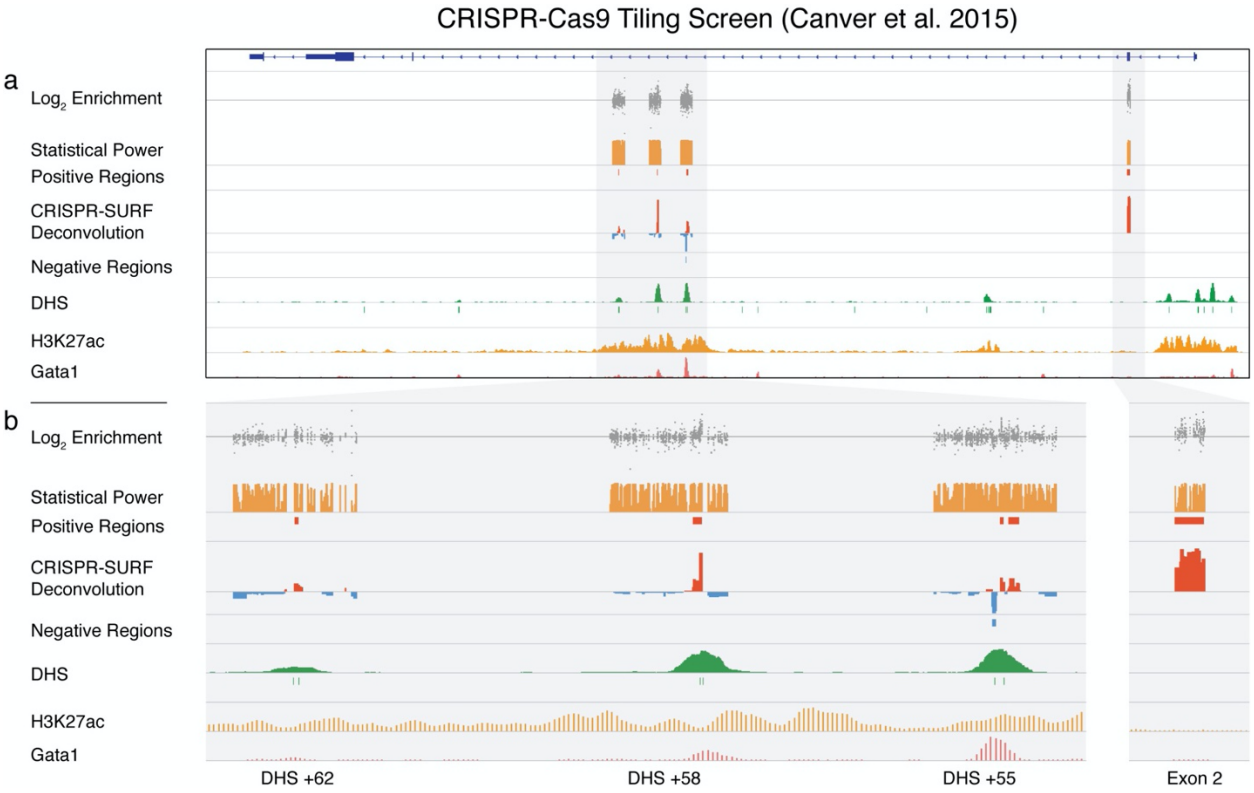
Tiled Pooled CRISPR-Cas9 and CRISPRi screen

HUDEP-2 cells were first transduced with lentiCas9-Blast (Addgene plasmid ID 52962) or pHR-SFFV-dCas9-BFP-KRAB (Addgene plasmid ID 46911) and stably selected with blasticidin 10 mcg/ml or sorted for BFP expression. Subsequently the cells, were transduced with pooled guide RNA lentiviral libraries at MOI<0.5. 24 hours following transduction, the cells were treated with puromycin 1 mcg/ml, and transferred to erythroid differentiation media, with Iscove's Modified Dulbecco's Medium (IMDM) (Life Technologies) supplemented with 330 mg/ml holo-transferrin (Sigma), 10 mg/ml recombinant human insulin (Sigma), 2 IU/ml heparin (Sigma), 5% human solvent detergent pooled plasma AB (Rhode Island Blood Center), 3 IU/ml erythropoietin, 100 ng/ml human SCF, 1 mg/ml doxycycline, 1% L-glutamine, and 2% penicillin/streptomycin.

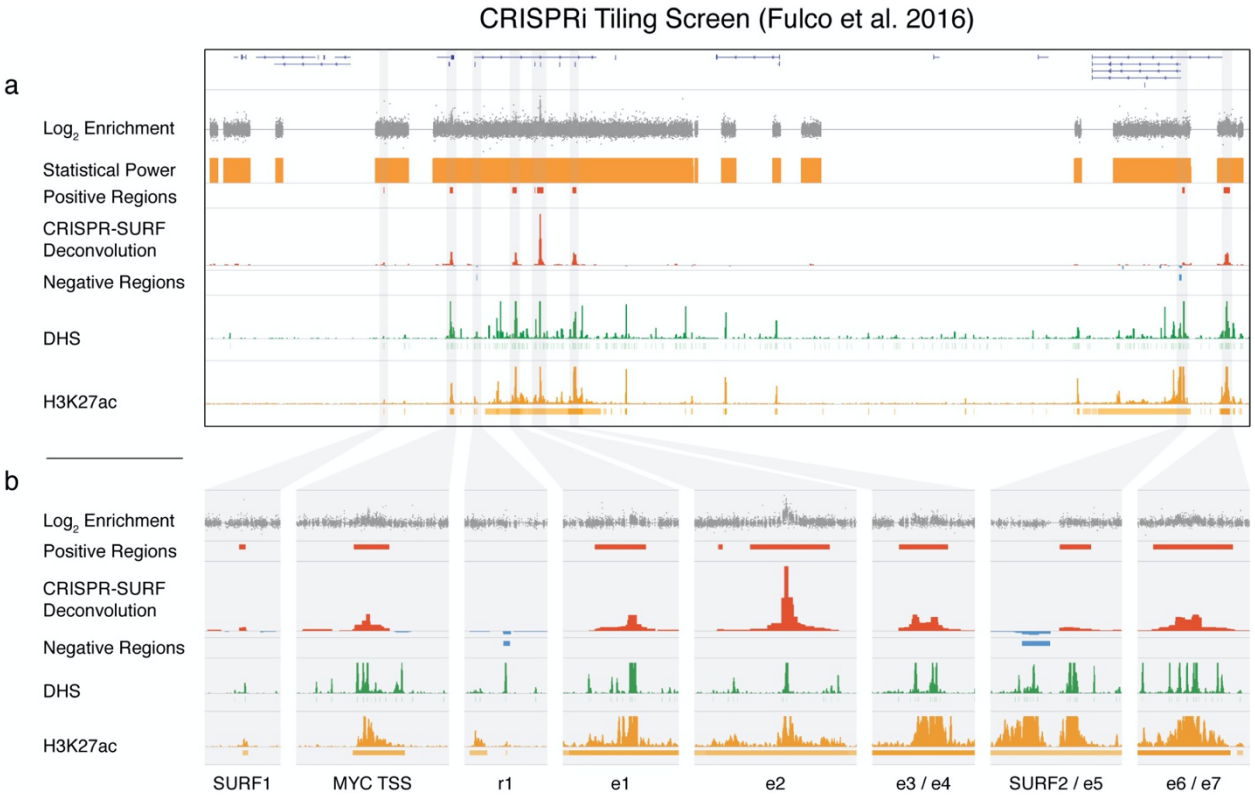
A representation of at least 1000 cells per guide RNA was maintained throughout the experiment. After 12 days, cells were fixed, permeabilized, and stained for intracellular fetal hemoglobin expression. Cells were sorted by flow cytometry to isolate HbF+ cells. In addition, cells prior to sorting (called pre-sort) were collected as a control. Genomic DNA was isolated from the presort and HbF+ populations. PCR amplification of the lentiviral integrants was performed to generate indexed adaptor-flanked amplicons for deep sequencing as previously

described[36]. Since we observed similar performance for the enrichment of positive control and negative control guide RNAs cloned into original lentiGuide-Puro or lentiGuide-Puro with combined modified scaffold, we treated these conditions as technical replicates for further analyses.

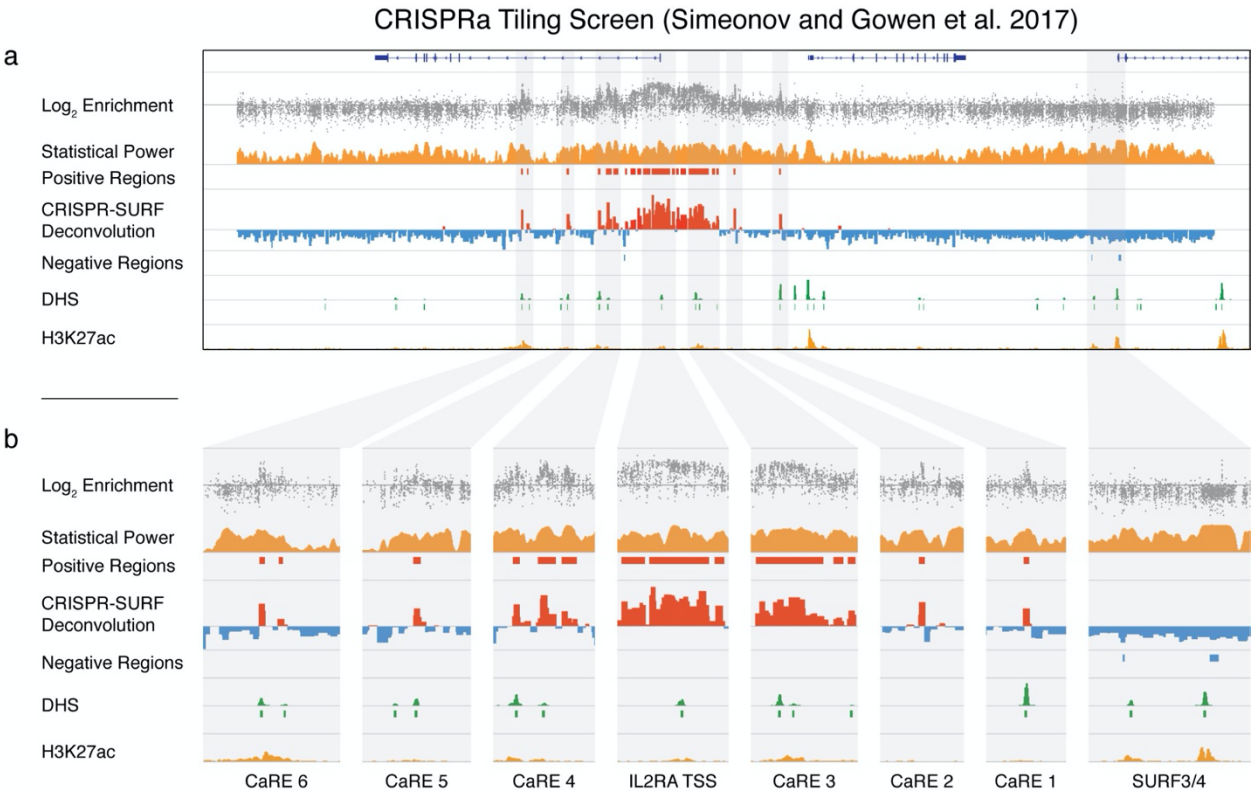
2.3 Supplementary figures



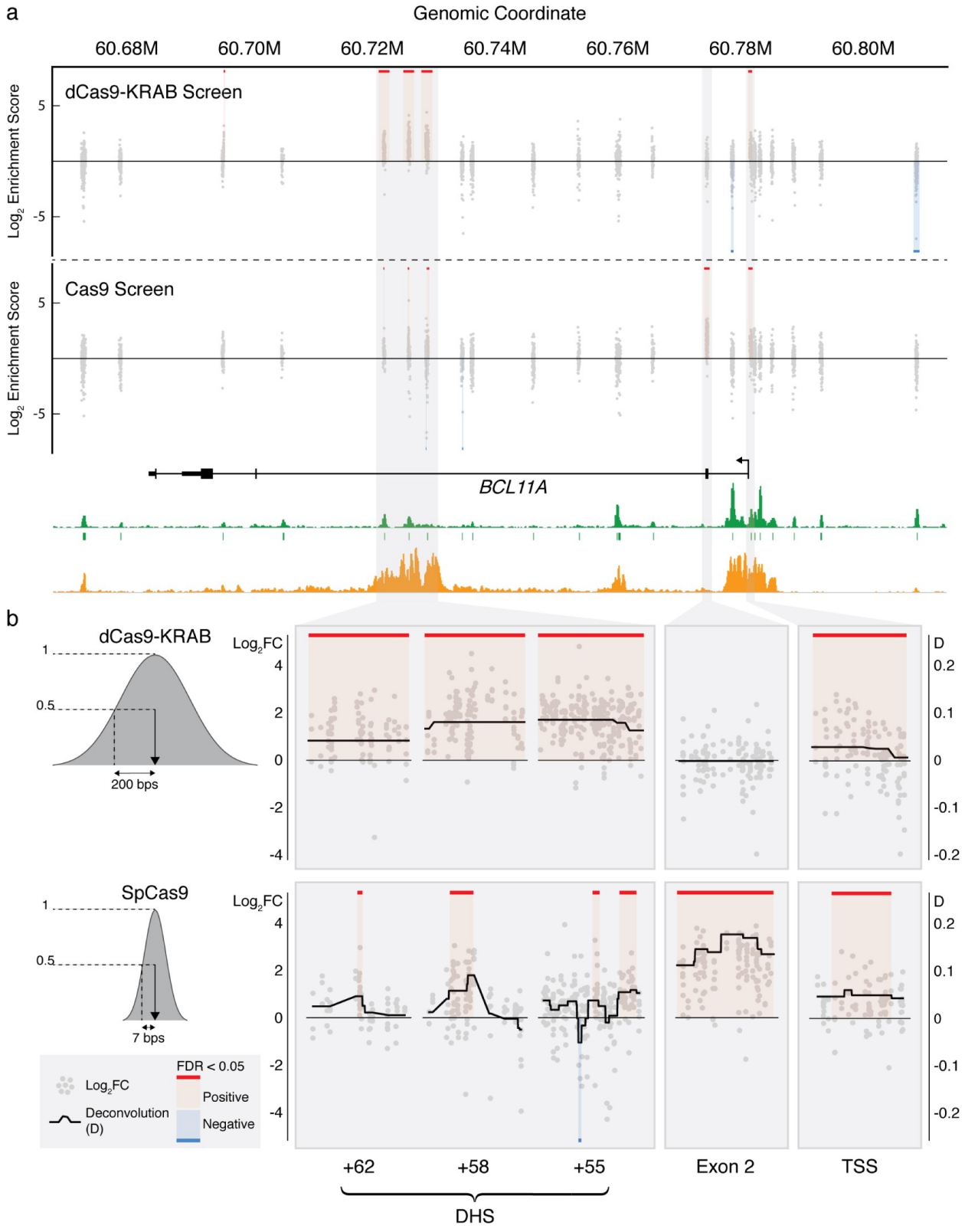
Supplementary Figure 2.1: Reanalysis of a CRISPR–Cas9 tiling screen from Canver et al.[36]. (a) An overview of the BCL11A CRISPR–Cas9 enhancer dissection tiling screen. **(b)** Zoom-in panels of DHS +55, +58, +62, and BCL11A exon 2 to highlight critical regions identified by CRISPR-SURF. All significant regions identified with FDR < 0.05. All panels are shown at same scale.



Supplementary Figure 2.2: Reanalysis of a CRISPRi tiling screen from Fulco et al.[80]. (a) An overview of the MYC CRISPRi enhancer discovery tiling screen. **(b)** Zoom-in panels of MYC TSS, e1–e7, and r1 (regions identified in ref.[80]) along with newly identified regions by CRISPR-SURF (SURF1 and SURF2). All significant regions identified with FDR < 0.05. All panels are shown at same scale.



Supplementary Figure 2.3: Reanalysis of a CRISPRa tiling screen from Simeonov et al.[81]. **(a)** An overview of the *IL2RA* CRISPRa enhancer discovery tiling screen. **(b)** Zoom-in panels of *IL2RA* TSS and CaREs 1–6 (regions identified in ref.[81]) along with regions newly identified by CRISPR-SURF (SURF3 and SURF4). All significant regions identified with FDR < 0.05. All panels are shown at same scale.


a
Genomic Coordinate
60.68M
60.70M
60.72M
60.74M
60.76M
60.78M
60.80M
dCas9-KRAB Screen
Log2 Enrichment Score
5
-5
Cas9 Screen
Log2 Enrichment Score
5
-5
BCL11A
b
dCas9-KRAB
1
0.5
200 bps
Log2FC
4
2
0
-2
-4
D
0.2
0.1
0
-0.1
-0.2
SpCas9
1
0.5
7 bps
Log2FC
4
2
0
-2
-4
D
0.2
0.1
0
-0.1
-0.2
Log2FC
Deconvolution (D)
FDR < 0.05
Positive
Negative
+62
+58
+55
Exon 2
TSS
DHS

Supplementary Figure 2.4: CRISPR-SURF analysis of parallel CRISPRi and CRISPR–Cas9 DHS tiling screens targeted to the *BCL11A* locus. (a) An overview of the *BCL11A* CRISPRi and CRISPR–Cas9 DHS tiling screens. **(b)** Shown are zoom-in panels of *BCL11A* exon 2 and common significant regions (FDR < 0.05) between the CRISPRi and CRISPR–Cas9 tiling screens as determined by CRISPR-SURF.

2.4 Acknowledgements

We thank R. Kurita and Y. Nakamura (RIKEN BioResource Center, Tsukuba, Japan) for sharing HUDEP-2 cells. L.P. is supported by NHGRI Career Development Award R00 HG008399 and CEGS RM1HG009490. D.E.B. is support by NIH R03 DK109232, NIH DP2 OD022716, NIH P01 HL032262, the Burroughs Wellcome Fund, and the Doris Duke Charitable Foundation. J.K.J. is supported by NIH R35 GM118158, NIH RM1 HG009490, and the Desmond and Ann Heathwood MGH Research Scholar Award. M.C.C. is supported by NIDDK Award F30-DK103359. J.M.E. is supported by NIH NHGRI 1K99HG009917-01 and the Harvard Society of Fellows.

2.5 Author contributions

J.Y.H. and L.P. conceived of and developed the CRISPR-SURF framework. M.A.C., M.C.C., and F.S. performed the experiments. C.P.F., D.P., R.F., K.C., J.A.G., L.B., S.H.O., J.M.E., and E.S.L. provided statistical and experimental expertise. J.K.J., L.P., and D.E.B. oversaw the project and offered feedback and guidance. J.Y.H., L.P., D.E.B., and J.K.J. wrote the manuscript with input from all other authors.

3 Rapid and simplified design of prime editing guide RNAs with PrimeDesign

3.1 Abstract

Prime editing (PE) is a versatile genome editing technology, but design of the required guide RNAs is more complex than for standard CRISPR-based nucleases or base editors. Here we describe PrimeDesign, a user-friendly, end-to-end web application and command-line tool for the design of PE experiments. PrimeDesign can be used for single and combination editing applications, as well as genome-wide and saturation mutagenesis screens. Using PrimeDesign, we construct PrimeVar, a comprehensive and searchable database that includes candidate prime editing guide RNA (pegRNA) and nicking sgRNA (ngRNA) combinations for installing or correcting >68,500 pathogenic human genetic variants from the ClinVar database. Finally, we use PrimeDesign to design pegRNAs/ngRNAs to install a variety of human pathogenic variants in human cells.

3.2 Introduction and background

Prime editing is a recently developed class of mammalian cell genome editing technology that enables unprecedented precision in the installation of specific substitutions, insertions, and deletions into the genome[68], offering greater versatility than CRISPR nucleases[24–26] and base editors[38,39]. The most efficient prime editing system described to date (referred to as PE3) consists of three components: a fusion protein of a CRISPR-Cas9 nickase and an engineered reverse transcriptase (RT), a prime editing guide RNA (pegRNA), and a nicking sgRNA (ngRNA) (**Supplementary Fig. 3.1**). The pegRNA targets the Cas9 nickase-RT fusion to a specific genomic locus, but also hybridizes to the nicked single-stranded DNA non-target strand (NTS) within the Cas9-induced R-loop, and serves as a template for reverse transcription to create the "flap" that mediates induction of precise genetic changes (**Supplementary Fig. 3.1a-c**). The ngRNA directs the Cas9 nickase-RT fusion to nick the strand opposite the flap and thereby biases repair towards the desired change encoded in the flap (**Supplementary Fig. 3.1d,e**). The complexity of the PE3 system makes it time-consuming to manually design the required pegRNA and ngRNA components. Beyond the need to design the spacer for both guide

RNAs, there are multiple other parameters that must be accounted for that can impact prime editing efficiencies, including: primer binding site (PBS) length, reverse transcription template (RTT) length, and distance between the pegRNA and ngRNA target sites.

Here we present PrimeDesign, a user-friendly web application (http://primedesign.pinellolab.org/) (**Fig. 3.1**) and command-line tool (https://github.com/pinellolab/PrimeDesign) that automates and thereby simplifies the design of pegRNAs and ngRNAs for single edits, combination edits, and genome-wide and saturation mutagenesis screens. We utilize PrimeDesign to construct PrimeVar, a comprehensive database of candidate prime editing guide RNA (pegRNA) and nicking sgRNA (ngRNA) combinations for installing or correcting >68,500 pathogenic human genetic variants in the ClinVar database. Lastly, we demonstrate the activity of pegRNA and ngRNA designs recommended by PrimeDesign through the installation of human pathogenic variants in human cells.

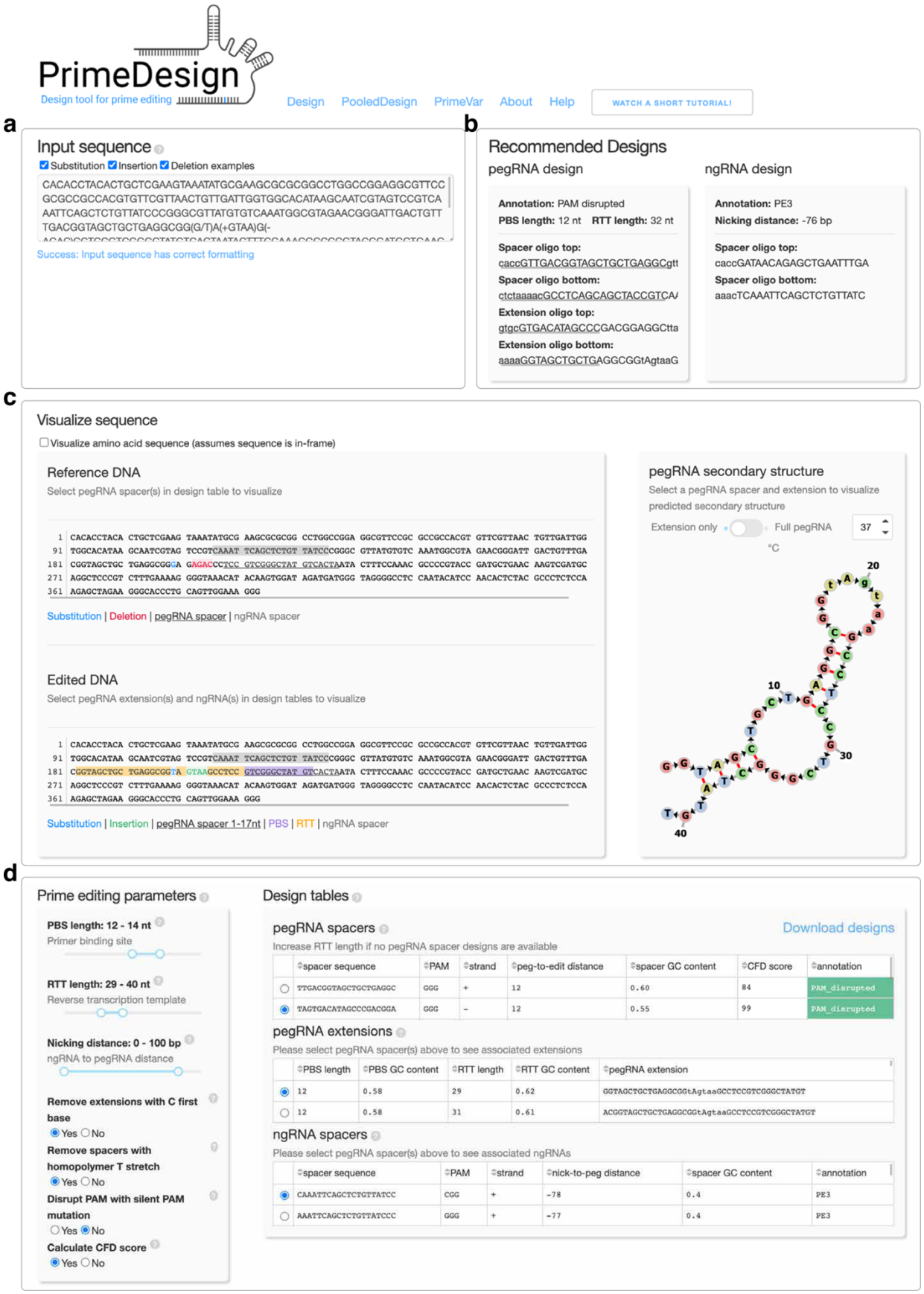

PrimeDesign
Design tool for prime editing
Design
PooledDesign
PrimeVar
About
Help
WATCH A SHORT TUTORIAL!
a
Input sequence
Substitution Insertion Deletion examples
CACACCTACACTGCTCGAAGTAAATATGCGAAGCGCGCGGCCTGGCCGGAGGCGTTCC
GCGCCGCCACGTGTTCGTTAACTGTTGATTGGTGGCACATAAGCAATCGTAGTCCGTCA
AATTCAGCTCTGTTATCCCGGGCGTTATGTGTCAAATGGCGTAGAACGGGATTGACTGTT
TGACGGTAGCTGCTGAGGCGG(G/T)A(+GTAA)G(-
Success: Input sequence has correct formatting
b
Recommended Designs
pegRNA design
Annotation: PAM disrupted
PBS length: 12 nt RTT length: 32 nt
Spacer oligo top:
caccGTTGACGGTAGCTGCTGAGGCgtt
Spacer oligo bottom:
ctctaaaacGCCTCAGCAGCTACCGTCAA
Extension oligo top:
gtgcGTGACATAGCCCGACGGAGGCtta
Extension oligo bottom:
aaaaGGTAGCTGCTGAGGCGGtAgtaaG
ngRNA design
Annotation: PE3
Nicking distance: -76 bp
Spacer oligo top:
caccGATAACAGAGCTGAATTTGA
Spacer oligo bottom:
aaacTCAAATTCAGCTCTGTTATC
c
Visualize sequence
Visualize amino acid sequence (assumes sequence is in-frame)
Reference DNA
Select pegRNA spacer(s) in design table to visualize
1 CACACCTACA CTGCTCGAAG TAAATATGCG AAGCGCGCGG CCTGGCCGGA GGCGTTCCGC GCCGCCACGT GTTCGTTAAC TGTTGATTGG
91 TGGCACATAA GCAATCGTAG TCCGTCAAAT TCAGCTCTGT TATCCCGGGC GTTATGTGTC AAATGGCGTA GAACGGGATT GACTGTTTGA
181 CGGTAGCTGC TGAGGCGGGA GAGACCTCC GTCGGGCTAT GTCACTAATA CTTTCCAAAC GCCCCGTACC GATGCTGAAC AAGTCGATGC
271 AGGCTCCCGT CTTTGAAAAG GGGTAAACAT ACAAGTGGAT AGATGATGGG TAGGGGCCTC CAATACATCC AACACTCTAC GCCCTCTCCA
361 AGAGCTAGAA GGGCACCCTG CAGTTGGAAA GGG
Substitution | Deletion | pegRNA spacer | ngRNA spacer
Edited DNA
Select pegRNA extension(s) and ngRNA(s) in design tables to visualize
1 CACACCTACA CTGCTCGAAG TAAATATGCG AAGCGCGCGG CCTGGCCGGA GGCGTTCCGC GCCGCCACGT GTTCGTTAAC TGTTGATTGG
91 TGGCACATAA GCAATCGTAG TCCGTCAAAT TCAGCTCTGT TATCCCGGGC GTTATGTGTC AAATGGCGTA GAACGGGATT GACTGTTTGA
181 CGGTAGCTGC TGAGGCGGTA GTAAGCCTCC GTCGGGCTAT GTCACTAATA CTTTCCAAAC GCCCCGTACC GATGCTGAAC AAGTCGATGC
271 AGGCTCCCGT CTTTGAAAAG GGGTAAACAT ACAAGTGGAT AGATGATGGG TAGGGGCCTC CAATACATCC AACACTCTAC GCCCTCTCCA
361 AGAGCTAGAA GGGCACCCTG CAGTTGGAAA GGG
Substitution | Insertion | pegRNA spacer 1-17nt | PBS | RTT | ngRNA spacer
pegRNA secondary structure
Select a pegRNA spacer and extension to visualize predicted secondary structure
Extension only
Full pegRNA
37
°C
20
10
30
40
d
Prime editing parameters
PBS length: 12 - 14 nt
Primer binding site
RTT length: 29 - 40 nt
Reverse transcription template
Nicking distance: 0 - 100 bp
ngRNA to pegRNA distance
Remove extensions with C first base
Yes No
Remove spacers with homopolymer T stretch
Yes No
Disrupt PAM with silent PAM mutation
Yes No
Calculate CFD score
Yes No
Design tables
pegRNA spacers
Download designs
Increase RTT length if no pegRNA spacer designs are available
spacer sequence | PAM | strand | peg-to-edit distance | spacer GC content | CFD score | annotation
TTGACGGTAGCTGCTGAGGC | GGG | + | 12 | 0.60 | 84 | PAM_disrupted
TAGTGACATAGCCCGACGGA | GGG | - | 12 | 0.55 | 99 | PAM_disrupted
pegRNA extensions
Please select pegRNA spacer(s) above to see associated extensions
PBS length | PBS GC content | RTT length | RTT GC content | pegRNA extension
12 | 0.58 | 29 | 0.62 | GGTAGCTGCTGAGGCGGtAgtaaGCCTCCGTCGGGCTATGT
12 | 0.58 | 31 | 0.61 | ACGGTAGCTGCTGAGGCGGtAgtaaGCCTCCGTCGGGCTATGT
ngRNA spacers
Please select pegRNA spacer(s) above to see associated ngRNAs
spacer sequence | PAM | strand | nick-to-peg distance | spacer GC content | annotation
CAAATTCAGCTCTGTTATCC | CGG | + | -78 | 0.4 | PE3
AAATTCAGCTCTGTTATCCC | GGG | + | -77 | 0.4 | PE3

Figure 3.1: PrimeDesign web application. (a) PrimeDesign takes a single sequence as input encoding both the original reference and desired edited sequences, **(b)** recommends a candidate pegRNA and ngRNA combination to install the edit of interest, **(c)** provides sequence visualizations of the edit of interest, selected pegRNA and ngRNA designs, and predicted pegRNA secondary structures, and **(d)** enables the interactive design of both pegRNAs and ngRNAs that can be downloaded as a summary table.

3.3 Results

3.3.1 PrimeDesign features

PrimeDesign uses a single input that encodes both the original reference and the desired edited sequences (**Fig. 3.1a and Supplementary Note 3.1**), recommends a candidate pegRNA and ngRNA combination to install the edit of interest (**Fig. 3.1b, Supplementary Fig. 3.2, and Supplementary Note 3.2**), provides sequence visualization of the prime editing event and predicted pegRNA secondary structures (**Fig. 3.1c**), and enumerates all possible pegRNA spacers, pegRNA extensions, and ngRNAs within optimized parameter ranges (previously defined by the Liu group[68]) for installing the desired edit (**Fig. 3.1d**). PrimeDesign enables users to rank pegRNAs based on their predicted specificity (CFD score[108]), provides important annotations for pegRNA (e.g. PAM disruption) and ngRNA (e.g. PE3b) designs, and streamlines the incorporation of PAM-disrupting silent mutations to improve editing efficiency and product purity (**Supplementary Note 3.3**). In addition, PrimeDesign enables the pooled design of pegRNA and ngRNA combinations for genome-wide and saturation mutagenesis screens (http://primedesign.pinellolab.org/pooled), and ranks the designs according to best design practices1. The saturation mutagenesis feature allows for the introduction of mutations at single-base or single-amino acid resolution; PrimeDesign automatically constructs all edits within a user-defined sequence range and generates the designs to install these edits (**Supplementary Note 3.4**).

3.3.2 PrimeVar database

To illustrate the utility of PrimeDesign, we took pathogenic human genetic variants from ClinVar[109] (n = 69,481) and designed candidate pegRNAs and ngRNAs for the correction of these pathogenic alleles. Of these pathogenic variants, we found that 91.7% are targetable by at

least a single pegRNA spacer with a maximum RTT length of 34 nt (**Fig. 3.2a and Supplementary Table 3.1**). An average of 3.7 pegRNA spacers were designed per pathogenic variant, representing multiple options for prime editing to correct each variant. Furthermore, 25.9% of targetable pathogenic variants included at least a single pegRNA that disrupts the PAM sequence, which has been associated with improved editing efficiency and product purity. The PE3b strategy (the design of ngRNAs that preferentially nick the non-edited strand after edited strand flap resolution) is viable for 79.5% of targetable variants (59.7% when only considering mismatches in the seed sequence; **Fig. 3.2b**). Lastly, 11.9% of targetable pathogenic variants are amenable to both the PAM-disrupting and PE3b seed-mismatched strategies.

To make all of these ClinVar prime editing designs more accessible, we constructed PrimeVar (http://primedesign.pinellolab.org/primevar), a comprehensive and searchable database for pegRNA and ngRNA combinations to install or correct >68,500 pathogenic human genetic variants. Using either the dbSNP reference SNP number (rs#) or ClinVar Variation ID, candidate pegRNAs and ngRNAs are readily available across a range of PBS (10–17 nt) and RTT (10–80 nt) lengths.

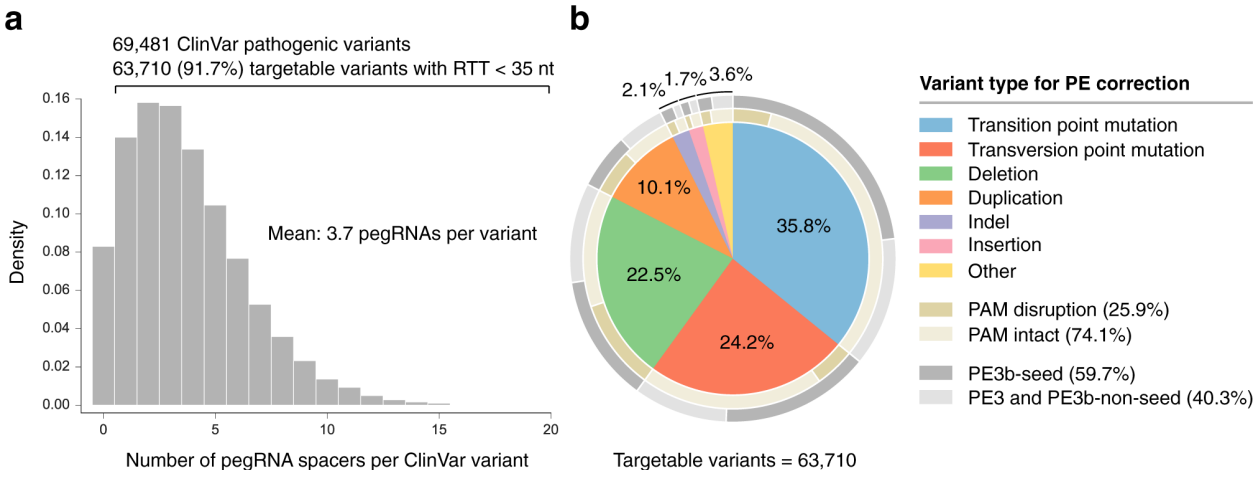


Figure 3.2: PrimeDesign analysis of the ClinVar database. (a) The distribution of the number of designed pegRNA spacers per ClinVar variant. Candidate pegRNAs were determined based on the requirement of RTT length <35 nt and the RT extension to have a minimum homology of 5 nt downstream of the edit. **(b)** The 63,710 (91.7%) targetable ClinVar variants classified by type. The inner ring (gold) represents the proportion of targetable variants by type where at least

one pegRNA could be designed to disrupt the PAM sequence (dark gold). The outer ring (gray) represents the proportion of targetable variants by type where at least one ngRNA could be designed for the PE3b strategy where the mismatch lies in the seed sequence (PAM-proximal nucleotides 1–10) (dark gray). See **Supplementary Table 3.1** for details.

3.3.3 Installation of pathogenic variants in human cells

Lastly, we tested recommended pegRNA and ngRNA combinations from PrimeDesign to install 20 different human pathogenic variants associated with genetic diseases including hemophilia A, Duchenne muscular dystrophy (DMD), MPS I and II, and Fabry disease in HEK293T cells (**Fig. 3.3a, Supplementary Table 3.2, and Supplementary Note 3.2**). We observed installation of the desired edit at mean frequencies of 10% or more for 7 of the 20 (35%) target sites and at mean frequencies of 1–10% for 6 of the 20 (30%) target sites. For a subset of seven of the desired mutations, we designed additional pegRNAs to assess differences between PE3 and PE3b (**Fig. 3.3b**). Generally, we observed mixed trends in the frequencies of the desired edit and a modest reduction in byproducts for PE3b relative to PE3. Lastly, we designed a subset of four additional pegRNAs that introduced PAM-disrupting silent mutations (in addition to the target pathogenic variant) and found that these designs resulted in a mean 1.8-fold increase in the frequency of the desired edit (**Fig. 3.3c**).

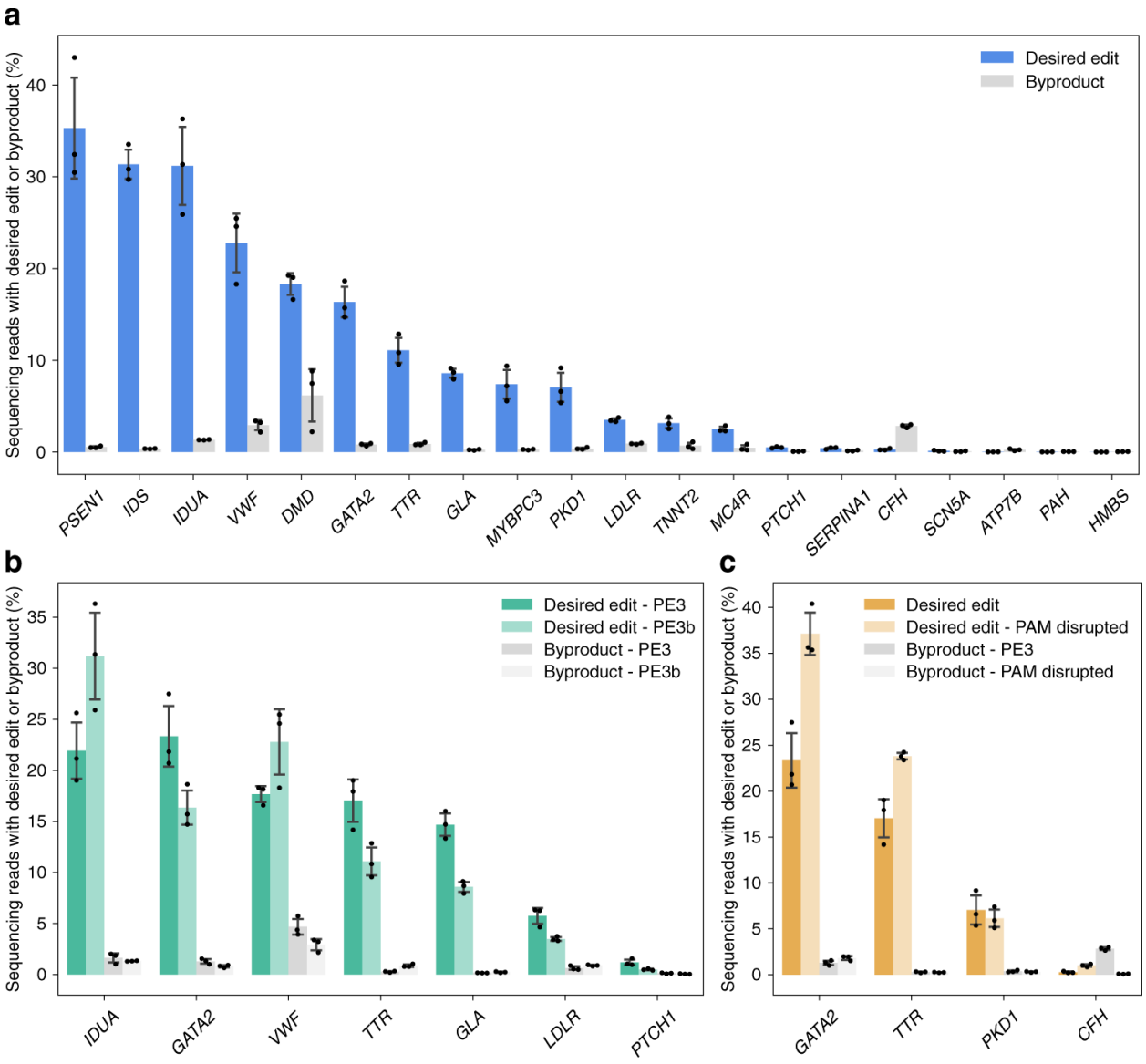


Figure 3.3: Installation of human pathogenic variants in HEK293T cells with PrimeDesign. **(a)** Overview of prime editing efficiencies for the installation of 20 human pathogenic variants in HEK293T cells, using PrimeDesign recommendations. Desired edit refers to sequencing reads containing only the edit of interest, while byproduct refers to sequencing reads containing any mutation(s) outside of only the edit of interest (i.e. indels, desired edit and indels). **(b)** Comparison between PE3 and PE3b editing strategies. **(c)** Assessing the effects of PAM-disrupting silent mutations on prime editing efficiencies. Mean ± s.d. of $n = 3$ independent biological replicates. Some of the data shown in **(a)** are also represented in **(b)** and **(c)**. See **Supplementary Table 3.2** for details.

3.4 Discussion and conclusions

In summary, PrimeDesign is a comprehensive and general method for facile and automated design of pegRNAs and ngRNAs. Our test of pegRNAs and ngRNAs designed by PrimeDesign to create various edits shows that not all designs yield the desired alterations with high frequencies, therefore, users of PrimeDesign may still need to refine pegRNA choices even after testing initial recommendations. Nonetheless, PrimeDesign should greatly simplify the complicated process of designing candidate prime editing components and thereby increase the use of and accessibility to this powerful and important technology[110–112].

3.5 Materials and methods

Molecular cloning

We used a PE2 construct that encodes a P2A-eGFP fusion for cotranslational expression of PE2 and enhanced GFP (eGFP) under control of a CMV promoter (pJUL2440; derived from Addgene no. 132775). For the cloning of pegRNAs (**Supplementary Table 3.2**), double-stranded DNA fragments for the pegRNA scaffold, spacer, and 3′ extension were formed by annealing oligos with compatible overhangs for ligation. The fragments were then ligated using T4 ligase (NEB) and cloned into the BsaI-digested pUC19-based hU6-pegRNA-gg-acceptor entry vector (Addgene no. 132777). For nicking gRNA (ngRNA) cloning, spacer oligos were duplexed and ligated into the BsmbI-digested pUC19-based hU6-SpCas9 gRNA entry vector BPK1520 (Addgene no. 65777). All pegRNA and ngRNA plasmids were transformed into chemically competent E.coli (XL1-Blue, Agilent). Plasmids used for transfection were midi (PE2) or mini prepped (gRNAs) using the Qiagen midi plus or miniprep kits.

Cell culture

STR-authenticated HEK293T cells (CRL-3216) were grown in Dulbecco's modified Eagle medium (DMEM, Gibco) containing 10% fetal bovine serum (FBS, Gibco) and 1% penicillin-streptomycin antibiotic (Gibco). Cells were kept in a 5% CO_2 incubator at 37 °C. Cells were passaged every 2–3 days as cells reached 80% confluency. Cells did not exceed passage 13 for all replicates in this experiment. Mycoplasma testing of the cell culture media took place every 4 weeks with the MycoAlert PLUS mycoplasma detection kit (Lonza) and showed negative results for the duration of this study.

Transfections

HEK293T cells were seeded into 96-well flat-bottom cell culture plates (Corning) for PE treatment at 1.2×10^4 cells/well. Transfections were carried out 18–24 h post-seeding with 30 ng PE2 plasmid, 10 ng pegRNA, and 3.3 ng ngRNA plasmid per transfection (per well, in a 96-well plate). TransIT-X2 (Mirus) was used as the lipofection reagent at 0.3 µL per transfection.

DNA extraction

Post-transfection (72 h), HEK293T cells were washed using 1x PBS (Corning) and lysed with 43.5 µL of gDNA lysis buffer (100 mM Tris, 200 mM NaCl, 5 mM EDTA, 0.05% SDS), 1.25 µL of 1 M DTT (Sigma), and 5.25 µL of Proteinase K per well for 96-well plate experiments. The plates were put into a shaker (500 rpm) at 55 °C overnight, and gDNA was extracted using 1.5x paramagnetic beads. Beads with bound gDNA were washed with 70% ethanol three times using a Biomek FX^P Laboratory Automation Workstation (Beckman Coulter) and then eluted in 35 µL 0.1x EB buffer (Qiagen).

Targeted amplicon sequencing

The gDNA concentrations of several samples from different pegRNAs/replicates were measured using the Qubit dsDNA HS Assay Kit (Thermo Fisher). The first PCR was performed to amplify the genomic regions of interest (200–250 bp) using 10–20 ng of gDNA. Primers for PCR1 included Illumina-compatible adapter sequences (**Supplementary Table 3.2**). A synergy HT microplate reader (BioTek) was then used at 485/528 nm with the Quantifluor dsDNA quantification system (Promega) to measure the concentration of the first PCR products. PCR products from different genomic amplicons were then pooled and cleaned with 0.7x paramagnetic beads. The second PCR was performed to attach unique barcodes to each amplicon using 50–200 ng of the pooled PCR1 products and barcodes that correspond to Illumina TruSeq CD indexes. The PCR2 products were again cleaned with 0.7x paramagnetic beads and measured with the Quantifluor system before final pooling. The final library was sequenced using an Illumina Miseq (Miseq Reagent Kit v.2; 300 cycles, 2 × 150 bp, paired-end). The FASTQ files were downloaded from BaseSpace (Illumina).

Analysis

Amplicon sequencing data were analyzed with CRISPResso version 2.0.42 with HDR mode. Downstream analysis was sourced from 'CRISPResso_quantification_of_editing_frequency.txt.' The frequency of *Desired edit* was determined by taking HDR Unmodified and dividing by Reads_aligned_all_amplicons and the frequency of *Byproduct* was determined by taking the sum of HDR Modified, Reference Modified, Ambiguous and dividing by Reads_aligned_all_amplicons.

PrimeDesign analysis on ClinVar variants

The ClinVar database was accessed April 8th 2020. Variants were filtered with the following conditions: (1) included a valid GRCh38/hg38 coordinate, (2) labeled as Pathogenic for the column "ClinicalSignificance", and (3) contained a unique identifier determined by the concatenation of columns "Name," "RS# (dbSNP)," and "VariationID." All variants with ambiguous IUPAC code were converted into separate entries with non-ambiguous bases for downstream analysis. Following these steps, the total number of ClinVar variants totaled 69,481. Sequence inputs were formatted for all entries for both the installation and correction of these pathogenic variants. After running PrimeDesign on the ClinVar variants, candidate pegRNA designs were filtered with two criteria: (1) maximum RTT length of 34 nt and (2) minimum homology of 5 nt downstream of the edit. The pegRNAs with PAM disrupted annotations have mutations in the dinucleotide GG of the NGG motif, and the ngRNAs with *PE3b*, *PE3b non-seed*, and *PE3b seed* annotations have mismatches anywhere in the protospacer, mismatches outside of PAM-proximal nucleotides 1–10, or mismatches within PAM-proximal nucleotides 1–10, respectively.

Construction of PrimeVar database

The filtered ClinVar variants from the PrimeDesign analysis were used to build a comprehensive database of candidate pegRNA and ngRNA combinations. Prime editing designs are available to install and correct the pathogenic human genetic variants. PrimeDesign was run with a PBS length range of 10–17 nt, RTT length range of 10–80 nt, and ngRNA distance range of 0–100 bp. All of the pegRNA and ngRNA designs for each variant are stored on PrimeVar (http://primedesign.pinellolab.org/primevar).

3.6 Supplementary notes

3.6.1 Supplementary Note 3.1: PrimeDesign input sequence encoding the desired edit

The PrimeDesign input sequence format encodes both the original reference and desired edited sequences. All edits are formatted within a set of parentheses. Substitution edits are encoded by: (ref/edit), where ref is the pre-substitution reference sequence and edit is the postsubstitution edited sequence. Insertion edits are encoded by: (+ins) or (/ins), where ins is the sequence to be inserted into the reference sequence during the editing event. Deletion edits are encoded by: (-del) or (del/), where del is the sequence to be deleted from the reference sequence during the editing event. Substitution, insertion, and deletion edits can be combined for combination edits. All sequences unaffected by editing remain outside of the parentheses. It is recommended to place the intended edit site near the center of the input sequence and have the total input sequence length be >300 bp to ensure thorough design for prime editing. We provide some examples of input sequences below:

Substitution edit:

CACACCTACACTGCTCGAAGTAAATATGCGAAGCGCGCGGCCTGGCCGGAGGCGTT
CCGCGCCGCCAC
GTGTTCGTTAACTGTTGATTGGTGGCACATAAGCAATCGTAGTCCGTCAAATTCAGC
TCTGTTATCCCGG
GCGTTATGTGTCAAATGGCGTAGAACGGGATTGACTGTTTGACGGTAGCTGCTGAGG
CGG(G/T)AGAG
ACCCTCCGTCGGGCTATGTCACTAATACTTTCCAAACGCCCCGTACCGATGCTGAAC
AAGTCGATGCAGG
CTCCCGTCTTTGAAAAGGGGTAAACATACAAGTGGATAGATGATGGGTAGGGGCCT
CCAATACATCCAA
CACTCTACGCCCTCTCCAAGAGCTAGAAGGGCACCCTGCAGTTGGAAAGGG

Insertion edit:

CACACCTACACTGCTCGAAGTAAATATGCGAAGCGCGCGGCCTGGCCGGAGGCGTT
CCGCGCCGCCAC
GTGTTCGTTAACTGTTGATTGGTGGCACATAAGCAATCGTAGTCCGTCAAATTCAGC

TCTGTTATCCCGG
GCGTTATGTGTCAAATGGCGTAGAACGGGATTGACTGTTTGACGGTAGCTGCTGAGG
CGGGA(+GTAA)
GAGACCCTCCGTCGGGCTATGTCACTAATACTTTCCAAACGCCCCGTACCGATGCTG
AACAAGTCGATGC
AGGCTCCCGTCTTTGAAAAGGGGTAAACATACAAGTGGATAGATGATGGGTAGGGG
CCTCCAATACAT
CCAACACTCTACGCCCTCTCCAAGAGCTAGAAGGGCACCCTGCAGTTGGAAAGGG

Deletion edit:
CACACCTACACTGCTCGAAGTAAATATGCGAAGCGCGCGGCCTGGCCGGAGGCGTT
CCGCGCCGCCAC
GTGTTCGTTAACTGTTGATTGGTGGCACATAAGCAATCGTAGTCCGTCAAATTCAGC
TCTGTTATCCCGG
GCGTTATGTGTCAAATGGCGTAGAACGGGATTGACTGTTTGACGGTAGCTGCTGAGG
CGGGAG(-
AGAC)CCTCCGTCGGGCTATGTCACTAATACTTTCCAAACGCCCCGTACCGATGCTGA
ACAAGTCGATGC
AGGCTCCCGTCTTTGAAAAGGGGTAAACATACAAGTGGATAGATGATGGGTAGGGG
CCTCCAATACAT
CCAACACTCTACGCCCTCTCCAAGAGCTAGAAGGGCACCCTGCAGTTGGAAAGGG

Combination edit:
CACACCTACACTGCTCGAAGTAAATATGCGAAGCGCGCGGCCTGGCCGGAGGCGTT
CCGCGCCGCCAC
GTGTTCGTTAACTGTTGATTGGTGGCACATAAGCAATCGTAGTCCGTCAAATTCAGC
TCTGTTATCCCGG
GCGTTATGTGTCAAATGGCGTAGAACGGGATTGACTGTTTGACGGTAGCTGCTGAGG
CGG(G/T)A(+GT AA)G(-
AGAC)CCTCCGTCGGGCTATGTCACTAATACTTTCCAAACGCCCCGTACCGATGCTGA
ACAAGTCGATGC

AGGCTCCCGTCTTTGAAAAGGGGTAAACATACAAGTGGATAGATGATGGGTAGGGG
CCTCCAATACAT
CCAACACTCTACGCCCTCTCCAAGAGCTAGAAGGGCACCCTGCAGTTGGAAAGGG

3.6.2 Supplementary Note 3.2: PrimeDesign recommended pegRNA and ngRNA designs

PrimeDesign provides a pegRNA and ngRNA recommendation to install an edit of interest based on best practices described in Anzalone et al. 2019[68]. The determination of the pegRNA spacer is performed with the following preference: 1) PAM disrupted annotation 2) minimization of distance of nick to edit of interest. The determination of the pegRNA PBS length is performed by calculating the RNA-DNA melting temperatures for all possible PBS lengths, and then choosing the PBS length that is closest to our determined value of 37°C (**Supplementary Figure 3.2**). The determination of the pegRNA RTT length is performed by calculating the edit length and then constructing an RTT with a certain length of homology downstream of the edit: 1) edit length <= 1 is 10 nt homology downstream 2) 1 < edit length <= 5 is 15 nt homology downstream 3) 5 < edit length <= 10 is 20 nt homology downstream 4) 10 < edit length <= 15 is 25 nt homology downstream and 5) edit length > 15 is 35 nt homology downstream. The determination of the ngRNA spacer is performed with the following preference: 1) PE3b seed annotation 2) PE3b non-seed annotation 3) PE3 annotation at a distance as close to 75 bp away from the pegRNA spacer. While the PrimeDesign pegRNA and ngRNA recommendations are informed based on the information to date, we note it should serve as an initial point for design and that further empirical testing of pegRNA and ngRNA designs may be needed to achieve optimal prime editing efficiencies.

3.6.3 Supplementary Note 3.3: Design annotations for pegRNAs and ngRNAs

PrimeDesign provides important annotations during the design of pegRNAs and ngRNAs. For pegRNAs, the different annotations include: PAM intact, PAM disrupted, and PAM disrupted silent mutation. The PAM intact annotation is given to pegRNAs that do not introduce edits into the PAM sequence at positions that have sequence preference, whereas the PAM disrupted annotation is given to pegRNAs that introduce sequence modifications at PAM positions that have sequence preference. For coding sequence edits, PrimeDesign offers a functionality to introduce silent mutations to potentially improve editing efficiency and product purity. When

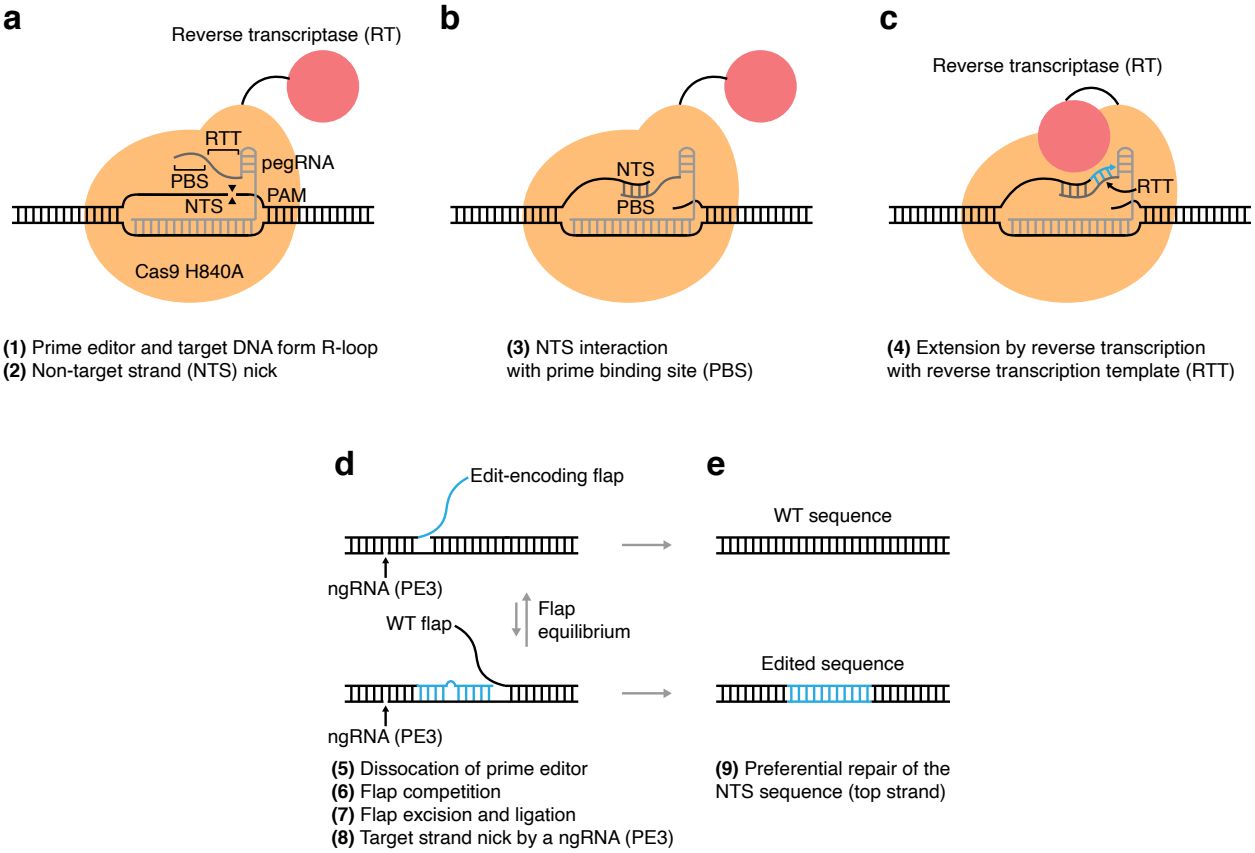
this functionality is turned on and the design is available, the PAM disrupted silent mutation is provided for suitable pegRNA designs. Importantly, the input sequence must be provided inframe in order for this function to work properly. We recommend using the amino acid sequence viewer on our PrimeDesign web application to check whether the input sequence is in-frame, and deleting the left-most bases of the input sequence to achieve the correct frame. PrimeDesign uses the GenScript human codon usage frequency table (https://www.genscript.com/tools/codon-frequency-table) and automatically selects the best codon by frequency to introduce the silent mutation. For ngRNAs, the different annotations include: PE3, PE3b non-seed, and PE3b seed. The PE3 annotation is given to ngRNAs that have a spacer match to both the original reference and desired edited sequences. The PE3b non-seed and PE3b seed annotations are given to ngRNAs that have a spacer that only perfectly matches the desired edited sequence, and therefore preferentially nick the non-edited strand after edited strand flap resolution. The PE3b non-seed ngRNAs contain sequence mismatches to the original reference sequence outside of PAMproximal nucleotides 1-10 (seed region), whereas PE3b seed ngRNAs contain sequence mismatches to the original reference sequence within the seed region. Spacer mismatches in the seed region severely inhibit target DNA binding and cleavage to a larger degree compared to spacer mismatches outside of the seed region. For this reason, PE3b seed ngRNAs may exhibit higher specificity in nicking the non-edited strand after edited strand flap resolution and are therefore more suitable for the PE3b strategy compared to PE3b non-seed ngRNAs.

3.6.4 Supplementary Note 3.4: Genome-wide and saturating mutagenesis designs

PrimeDesign offers the ability to perform pooled designs for genome-wide and saturating mutagenesis screen applications. For each edit of interest, PrimeDesign outputs a user-defined number of pegRNAs (unique spacers) and ngRNAs per pegRNA. These designs are ranked according to general guidelines previously established by the Liu group. Hierarchal ranking of pegRNAs is performed by first using the pegRNA annotations (PAM disrupted -> PAM disrupted silent mutation -> PAM intact), and then using pegRNA-to-edit distances (smallest to largest). Hierarchal ranking of ngRNAs is performed by first using the ngRNA annotations (PE3b seed -> PE3b non-seed -> PE3), and then using deviations from a user-defined ngRNA-to-pegRNA distance parameter (default: 75 bp). PrimeDesign enables streamlined design of

saturation mutagenesis studies with prime editing at single-base and single-amino acid resolution. The PrimeDesign input sequence format for saturation mutagenesis applications is the following: (seq), where seq is the user-defined sequence range of where the saturating mutagenesis will take place. If the "base" option is selected, PrimeDesign will automatically construct all single-base changes (i.e. A -> T,C,G) across the user-defined sequence range. If the "amino acid" option is selected, PrimeDesign will automatically construct all single-amino acid changes (including a stop codon) within the userdefined sequence range. Importantly, the user-defined sequence range must be in-frame in order for this function to work properly. PrimeDesign uses the GenScript human codon usage frequency table (https://www.genscript.com/tools/codon-frequency-table) and automatically selects the best codon by frequency to introduce the amino acid changes.

3.7 Supplementary figures



(1) Prime editor and target DNA form R-loop
(2) Non-target strand (NTS) nick

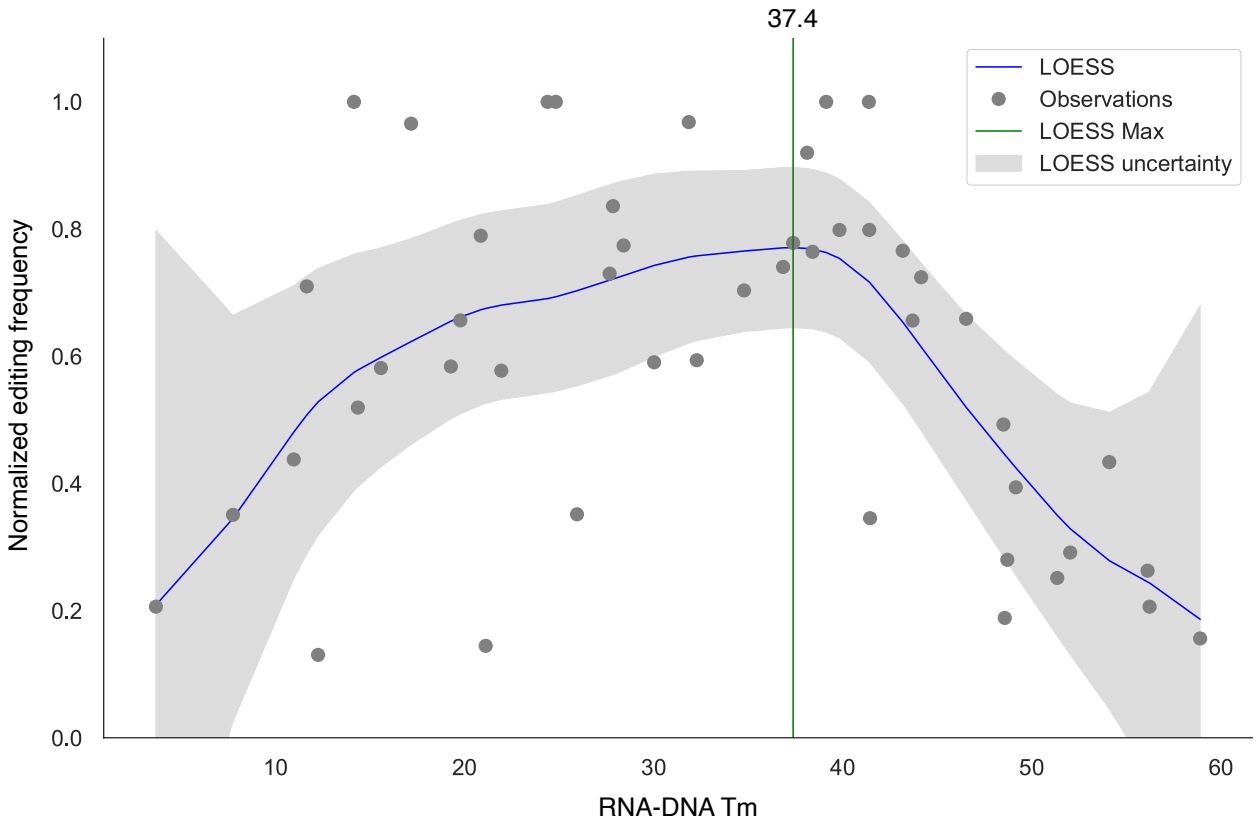
(3) NTS interaction with prime binding site (PBS)

(4) Extension by reverse transcription with reverse transcription template (RTT)

(5) Dissocation of prime editor
(6) Flap competition
(7) Flap excision and ligation
(8) Target strand nick by a ngRNA (PE3)

(9) Preferential repair of the NTS sequence (top strand)

Supplementary Figure 3.1: Overview of prime editing. (a) The prime editor, a fusion protein of a CRISPR-SpCas9 H840A nickase and an engineered Moloney murine leukemia virus reverse transcriptase (MMLV-RT), coupled with a prime editing guide RNA (pegRNA) engages target DNA to form an R-loop and introduces a nick in the nontarget strand (NTS). **(b)** The 3' end of the NTS interacts with the 3' end of the pegRNA extension, called the primer binding site (PBS), and this stabilized interaction allows for **(c)** extension of the NTS by reverse transcription with the reverse transcriptase and reverse transcription template (RTT) of the pegRNA where the edit of interest is encoded in the RTT sequence. **(d)** The prime editor system dissociates from its target DNA and the newly-synthesized edit-encoding strand competes with the wild-type (WT) strand for hybridization; flap excision and ligation occurs, and a target strand (TS) nick is introduced with a nicking sgRNA (ngRNA) for the PE3 strategy. **(e)** The NTS sequence is preferentially repaired with the TS nick, resulting in a bias towards successful editing in scenarios where the edit-encoding flap outcompetes the WT flap for genomic incorporation.



Supplementary Figure 3.2: Relationship between PBS RNA-DNA melting temperature and prime editing efficiency. Reanalysis of previous data[68] was performed to assess the relationship

between the PBS RNA-DNA melting temperature and prime editing efficiency. LOESS was applied and the maximum value of the curve was identified to serve as an initial design recommendation for pegRNA PBS length. The LOESS uncertainty is represented by standard error (SE).

3.8 Supplementary tables

Supplementary Table 3.1.1: PrimeDesign analysis of ClinVar variants (installing variants).

Variant Type	Number of variants	Targetable variants	% Targetable	Mean pegRNAs per variant	Median pegRNAs per variant	Mean pegRNA-to-edit minimum distance
Transition point mutation	24716	23204	93.88%	4.37	4	6.40
Transversion point mutation	16638	15599	93.76%	4.04	4	7.48
Deletion	16402	15303	93.30%	4.19	3	7.95
Duplication	6780	6245	92.11%	4.06	3	9.86
Insertion	1160	1063	91.64%	3.32	3	11.11
Indel	1488	1386	93.15%	3.93	3	9.85
Other**	2297	2105	91.64%	3.38	3	10.80
Total	69481	64905	93.41%	3.95	3	7.65

Variant Type	Median pegRNA-to-edit minimum distance	% PAM disruption	% PE3b	% PE3b-seed	% PAM disruption and PE3b-seed
Transition point mutation	5	41.62%	83.40%	64.99%	21.21%
Transversion point mutation	6	30.73%	80.77%	61.32%	14.69%
Deletion	6	50.19%	77.52%	58.25%	23.72%
Duplication	8	35.26%	74.94%	53.77%	16.33%
Insertion	10	44.50%	75.63%	52.30%	18.06%
Indel	8	56.06%	81.39%	63.71%	28.07%
Other**	10	27.46%	72.26%	45.75%	10.69%
Total	6	40.31%	80.04%	60.58%	19.52%

Supplementary Table 3.1.2: PrimeDesign analysis of ClinVar variants (correcting variants).

Variant Type	Number of variants	Targetable variants	% Targetable	Mean pegRNAs per variant	Median pegRNAs per variant	Mean pegRNA-to-edit minimum distance
Transition point mutation	24723	22832	92.35%	4.10	4	7.64
Transversion point mutation	16760	15396	91.86%	3.95	4	8.06
Deletion	15825	14336	90.59%	3.67	3	10.50
Duplication	7027	6462	91.96%	4.58	4	7.74
Insertion	1187	1099	92.59%	3.57	3	9.31
Indel	1450	1316	90.76%	3.60	3	11.21
Other**	2509	2269	90.43%	3.09	2	12.13
Total	69481	63710	91.69%	3.71	3	8.66

Variant Type	Median pegRNA-to-edit minimum distance	% PAM disruption	% PE3b	% PE3b-seed	% PAM disruption and PE3b-seed
Transition point mutation	5	11.54%	83.02%	64.29%	5.66%
Transversion point mutation	6	18.04%	81.11%	61.39%	8.81%
Deletion	9	43.68%	76.30%	56.22%	19.34%
Duplication	6	47.83%	75.19%	54.27%	21.32%
Insertion	7	36.03%	75.80%	52.87%	15.92%
Indel	10	47.34%	82.37%	64.29%	23.56%
Other**	11	32.26%	65.84%	41.08%	12.43%
Total	7	25.90%	79.50%	59.70%	11.90%

Supplementary Table 3.2: Sequences of pegRNAs, nicking sgRNAs, primers, and amplicons used.

Target gene & edit	pegRNA spacer sequence	PAM	3' extension #1
TNNT2 - delGAG	GAGGGCTCGACG AGAGGAGG	AGG	AGCCTTCCTCCTGTTCTCCTCCTCTCGTCGAG
MYBPC3 - 25bp del	GAAATAAGGTAA AGAGAGGGA	AGG	AACACAGATGTGTCTCCCTGGGTCCCTGCCAGGTC CCCTCTCTTTACCTTAT
MC4R - delCTCT	GCCTGGCCATCA GGAACATG	TGG	GCTCTCATGGCTTCTATGTCCACATGTTCCTGATGG C
SCN5A - delG4CTTCT	GGCCGTGGGATG GGCTTCTG	GGG	GCCATGAAGAAGCTGGGCTCCAAGAAGCCCATCC
TTR - delGTC	GAGTCCCTCATTC CTTGGGAT	TGG	TCCACCACGGCTGTCACCAATCCCAAGGAATGAGG GA
CFH - 24bp del	GAACATGTTGGG ATGGGAAAC	TGG	ACTAAAGTTCTGAATAAAGGTGTGCACTTTATGAT TGATACTCCAGTTTCCCATCCC
GATA2 - delTGT	GTTTCGGCGCCA TAAGGTGG	TGG	GCAAATTGTCAGACGACCACCACCTTATGGCGCC
GLA - del(CAGC)2CC ATG	GATTGGCAAGGA CGCCTACCA	TGG	GGTTGCACATGAAGCGCTCCCAGTGGTAGGCGTCC TTG
PTCH1 - insTCTAC	GAGGTGTCCCGC AAGCCGTTG	AGG	GCCCAGTTCCCTTTCTACGTAGACTCAACGGCTTG CGG
HMBS - dup	GCCCTACCAGGT GCCTCAGGA	AGG	TGCTGCACGATCCCGAGACTTTCGCTGCATCGCTG AAAGGGCCTTCCTGAGGCACCTGG
PKD1 - indel	GCCCAGAGCTGG GGCCCCCA	CGG	GCCCTGCCAGCTCACCTTCCTGCAGCCGTGGGGGC CCCA
SERPINA1 - delAG	GCCTCCTCTGTGA CCCCGGAG	AGG	GCAATGGGGCTGACCTCCGGGGTCACAGA
PAH - delATG	GTATGAATTTTTC ACCCATT	TGG	GCCTCAAGATCTTGATGTTTGTCAGAGCAGGCAGG CTACGTTTATCCAAATGGGTGAAAAATTCATA
VWF - delG	GGTGACTCACCA TGCCCGGG	GGG	GCCTCTGCCCCCGGGCATGGTGAG

DMD - delG	GACTACGAGGCTGGCTCAGGG	GGG	GGACTCCCCCTGAGCCAGCC
PSEN1 - C>T	GAAAGAGCATGATCACATGCT	TGG	AAATATGGCGTCAAGCATGTGATCATGCTCT
IDUA - G>A	GCCGCAGATGAGGAGCAGCTC	TGG	ACACTTCGGCCTAGAGCTGCTCCTCATC
IDS - G>A	GACTGAGGGATGTCTGAAGGC	CGG	GCCAGTATCCCTGGCCTTCAGACATCCCT
ATP7B - delG	GCCAGCAATACCTTTTTCTGC	GGG	GCAGCCTTCCGCAGAAAAAGGTATTGCTG
LDLR - T>A	GCGCGGCGGGGACTGCAGGTA	AGG	AGCAAGCCTTTCCTGCAGTCCC

Target gene & edit	**3' extension #2**	**Primer binding sequence**	**RT template #1**
TNNT2 - delGAG		CCTCTCGTCGAG	AGCCTTCCTCCTGTTCTCCT
MYBPC3 - 25bp del		CTCTCTTTACCTTAT	AACACAGATGTGTCTCCCTGGGTCCCTGCCAGGTCCC
MC4R - delCTCT		GTTCCTGATGGC	GCTCTCATGGCTTCTATGTCCACAT
SCN5A - delG4CTTCT		AAGCCCATCC	GCCATGAAGAAGCTGGGCTCCAAG
TTR - delGTC	TCCACCACGGCTGTCACAAATCCCAAGGAATGAGGGA	CCAAGGAATGAGGGA	TCCACCACGGCTGTCACCAATC
CFH - 24bp del	ACTAAAGTTCTGAATAAAGGTGTGCACTTTATGATTGATACTCGAGTTTCCCATCCC	TCCCATCCC	ACTAAAGTTCTGAATAAAGGTGTGCACTTTATGATTGATACTCCAGTT
GATA2 - delTGT	GCAAATTGTCAGACGACAACCACCTTATGGCGCC	CCTTATGGCGCC	GCAAATTGTCAGACGACCACCA

GLA - del(CAGC)2CCATG		TAGGCGTCCTTG	GGTTGCACATGAAGCGCTCCCAGTGG
PTCH1 - insTCTAC		CGGCTTGCGG	GCCCAGTTCCCTTTCTACGTAGACTCAA
HMBS - dup		TGAGGCACCTGG	TGCTGCACGATCCCGAGACTTTCGCTGCATCGCTGAAAGGGCCTTCC
PKD1 - indel	GCCCTGCCAGCTCACCTTCCCTGCAGTCTTGGGGGCCCCA	GGGCCCCA	GCCCTGCCAGCTCACCTTCCTGCAGCCGTGG
SERPINA1 - delAG		CGGGGTCACAGA	GCAATGGGGCTGACCTC
PAH - delATG		GGGTGAAAAATTCATA	GCCTCAAGATCTTGATGTTTGTCAGAGCAGGCAGGCTACGTTTATCCAAAT
VWF - delG		GGGCATGGTGAG	GCCTCTGCCCCC
DMD - delG		TGAGCCAGCC	GGACTCCCCC
PSEN1 - C>T		ATGTGATCATGCTCT	AAATATGGCGTCAAGC
IDUA - G>A		CTGCTCCTCATC	ACACTTCGGCCTAGAG
IDS - G>A		TTCAGACATCCCT	GCCAGTATCCCTGGCC
ATP7B - delG		GAAAAAGGTATTGCTG	GCAGCCTTCCGCA
LDLR - T>A		CTGCAGTCCC	AGCAAGCCTTTC

Target gene & edit	**RT template #2**	**nicking gRNA #1 spacer**	**nicking gRNA #2 spacer**
TNNT2 - delGAG		GAGGATTGGAAACCCTGATTC	

MYBPC3 - 25bp del		GCATAGTCAGGGACTCTCGT	
MC4R - delCTCT		GCCGCCAAGGTGCCAATATGA	
SCN5A - delG4CTTCT		GCTCCAAGAAGCCCATCCCA	
TTR - delGTC	TCCACCACGGCTGTCACAAATC	GCACGGCTGTCACCAATCCCA	GTATTCACAGCCAACGACTC
CFH - 24bp del	ACTAAAGTTCTGAATAAAGGTGTGCACTTTATGATTGATACTCGAGTT	GAATACTAAAGTTCTGAATAA	
GATA2 - delTGT	GCAAATTGTCAGACGACAACCA	GTCAGACGACCACCACCTTA	GCGCTGCCTTGCCCTCCCAGT
GLA - del(CAGC)2CCATG		GCATGAAGCGCTCCCAGTGGT	GCAATATCTGATACCTGATGC
PTCH1 - insTCTAC		GCCCTTTCTACgtagaCTCAA	GCTGTGATGCTCTTCTACCCT
HMBS - dup		GACTttcgctgcATCGCTGAA	
PKD1 - indel	GCCCTGCCAGCTCACCTTCCTGCAGTCTTGG	GAGCTCACCTTCCTGCAGCCG	
SERPINA1 - delAG		GTCAGCAATGGGGCTGACCTC	
PAH - delATG		GATCTTGATGTTTGTCAGAGC	
VWF - delG		GTCTCTGGCTGCCTCTGCCCC	GCTGACAACCTGCGGGCTGAA
DMD - delG		GAGAATCCTAGCAGATCTTG	
PSEN1 - C>T		GTTATCTAATGGACGACCCCA	

IDUA - G>A		GtAGAGCTGCTCCTCATCTGC	GGCCGGGCCCTGGGGGCGGT
IDS - G>A		GCATTTTCGATTCCGTGACT	
ATP7B - delG		GTCCGCAGAAAAAGGTATTGC	
LDLR - T>A		GCTTTCCTGCAGTCCCCGCCG	GCCATGCTCGCAGCCTCTGCC

Target gene & edit	**FWD primer PCR1**	**REV primer PCR1**
TNNT2 - delGAG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCTGCTGCTCCCTACCTACCTTC	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCTTCAGCCCAGAATCAGGGTTTC
MYBPC3 - 25bp del	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGCTGAACATGCGGAAGCGG	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTTCCAGCCTTGGGCATAGTCAGG
MC4R - delCTCT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAATCAGGATGGTCAAGGTAATCGCTCC	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTTCATTTACTCAGATAGTAGTGCTGTCATCATCTG
SCN5A - delG4CTTCT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGCTGGCCATGTGGCACG	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTTGCCTTCTCTTTGCACTTAGGGGG
TTR - delGTC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTCACAGCCAACGACTCCGGC	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCATATGAGGTGAAAACACTGCTTTAGTAAAAATGG
CFH - 24bp del	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTGTGTAAACGGGGATATCGTCTTTCATCA	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCCACCGGTCTCAGCTTATAATTACATTTTC
GATA2 - delTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGGTCCCCTGGGAGGGGC	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGCCTGCTGACGCTGCCTT
GLA - del(CAGC)2CCATG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAACACATGGAAAAGCAAAGGGAAGGG	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTGCGCTTCGCTTCCTGGC

PTCH1 - insTCTAC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCCCCAGGCTCGTATAGTTGCT	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTCTCTTTCCATTGAAACTGTGATGCTCTTC
HMBS - dup	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGATGTCCTAGGATGTTTTTCCATCAGGG	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTTTCCCAGCTCTCCAAGTCCCC
PKD1 - indel	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCCGCTAAAGGCTGCTCTCTCAAC	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTTGTGACTGATGCTGTGGCAGGT
SERPINA1 - delAG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCCCACACATTCTTCCCTACAGATACC	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGCTTACATTTACCCAAACTGTCCATTACTGG
PAH - delATG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCTAATTCTTACCTGTGTCTTTCTTCTTATCTCGTG	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGGAGAATGATGTAAACCTGACCCACATT
VWF - delG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCCTTGTTTCTTCCTCTCTCTGGCTG	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTTGCGGGCTGAAGGGCTCG
DMD - delG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTTCTTCCTCAAGATCTGCTAGGATTCTCTCT	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAATAAGGGGGGGAAAAAACCAAAACCTTT
PSEN1 - C>T	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCATTATCTAATGGACGACCCCAGGG	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTACGTACAGCTGCCCATCCTTCC
IDUA - G>A	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACTCCTTCACCAAGGGGAGGGG	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGGGCGGTGGGCGCT
IDS - G>A	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGAAGCCAACCCACACAGTATACCTATAGT	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTGAAGCATTTTCGATTCCGTGACTTGGAAG
ATP7B - delG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGAGGCAATCACTGCTGGGCG	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTGGAATTGGGTGCAAAGTCAGCAA
LDLR - T>A	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGGCAGAGGCTGCGAGCA	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTATTACCCCACAAGTCTCCCAGGGA

Target gene & edit	Amplicon sequence	ClinVar ID
TNNT2 - delGAG	GCTTCAGCCCAGAATCAGGGTTTCCAATCCTTTCCCCTAATTTGCT TTCTTCCTCCCTGCTGTAAATCAGGAAGAGAGGGCTCGACGAGAG GAGGAGGAGAACAGGAGGAAGGCTGAGGATGAGGCCCGGAAGA AGAAGGCTTTGTCCAACATGATGCATTTTGGGGGTTACATCCAGA AGGTAGGTAGGGAGCAGCAGG	43648
MYBPC3 - 25bp del	TGCTGAACATGCGGAAGCGGGCGTCTTCTCCCAGGTCCAGGCCAT TCTTGAACCAGGAAATCTTGGGCTATAAATAAGGTAAAGAGAGG GAGGGAAGCCATCCAGGCTGAGAGGGGACCTGGCAGGGACCCAG GGAGACACATCTGTGTTTCTACTCGGGGGGTCCCACGAGAGTCCC TGACTATGCCCAAGGCTGGAAA	177677
MC4R - delCTCT	AATCAGGATGGTCAAGGTAATCGCTCCCTTCATATTGGCACCTTG GCGGATGGCACCAGTGCCGGGGAGGACAGCAATCCTCTTAATGT GAAGCCTGGCCATCAGGAACATGTGGACATAGAGAGAAGCCATG AGAGCCAGCATGGTGAAGAACATGGTGATGAGGCAGATGATGAC AGCACTACTATCTGAGTAAATGA	14316
SCN5A - delG4CTTCT	GGCTGGCCATGTGGCACGAAAGCTTCCCAGGGACCCCAGAAGAT CCTCCCCACTCCCACAAAACCAGGAGCCTGGCTCACCAGGGGCC GTGGGATGGGCTTCTGGGGCTTCTTGGAGCCCAGCTTCTTCATGG CATTGTAGTACTTCTTCTGCTCCTCTGTCATGAAGATGTCCTGGCC CCCTAAGTGCAAAGAGAAGGCA	201571
TTR - delGTC	GCATATGAGGTGAAAACACTGCTTTAGTAAAAATGGAATACTCTT GGTTACATGAAATCCCATCCCTCGTCCTTCAGGTCCACTGGAGGA GAAGTCCCTCATTCCTTGGGATTGGTGACGACAGCCGTGGTGGAA TAGGAGTAGGGGCTCAGCAGGGCGGCAATGGTGTAGCGGCGGGG GCCGGAGTCGTTGGCTGTGAA	13460
CFH - 24bp del	GTGTGTAAACGGGGATATCGTCTTTCATCACGTTCTCACACATTG CGAACAACATGTTGGGATGGGAAACTGGAGTATCCAACTTGTGC AAAAAGATAGAATCAATCATAAAGTGCACACCTTTATTCAGAACT TTAGTATTAAATCAGTTCTCAATTTCATTTTTTATGTATTGTTTTAC TCCTTTTTATTCATACGTAAAATTTTGGATTAATTTGTGAAAATGT AATTATAAGCTGAGACCGGTGGC	16546
GATA2 - delTGT	AGGTCCCCTGGGAGGGGCGGGGTGGCCGGGGCGGGGCGCACTCA CATTGTGCAGCTTGTAGTAGAGGCCACAGGCGTTGCAGACAGGG TCCCCGTTGGCGTTTCGGCGCCATAAGGTGGTGGTTGTCGTCTGA	29722

	CAATTTGCACAACAGGTGCCGGCTCTTCTGGCGGCCGACTGGGAG GGCAAGGCAGCGTCAGCAGGCT	
GLA - del(CAGC)2CC ATG	TTGCGCTTCGCTTCCTGGCCCTCGTTTCCTGGGACATCCCTGGGGC TAGAGCACTGGACAATGGATTGGCAAGGACGCCTACCATGGGCT GGCTGCACTGGGAGCGCTTCATGTGCAACCTTGACTGCCAGGAAG AGCCAGATTCCTGCATCAGGTATCAGATATTGGGTACTCCCTTCC CTTTGCTTTTCCATGTGTTT	10749
PTCH1 - insTCTAC	GCCCCAGGCTCGTATAGTTGCTGCAGATGGTCCTTACTTTTTCAAT TGCCTCCACAAAGTCTGAGGTGTCCCGCAAGCCGTTGAGGTAGAA AGGGAACTGGGCATACTCGATGGGCTCTGCTGCCGGGACTGGAC AGAGAAGGGCACAGGTTAGGAGCAGCCCAGGGTAGAAGAGCAT CACAGTTTCAATGGAAAGAGAG	453836
HMBS - dup	CTTTCCCAGCTCTCCAAGTCCCCAGCCCTCCACAGGTGGAGCACA GGCCCTACCAGGTGCCTCAGGAAGGCCCTTTCAGCGATGCAGCG AAGCAGAGTCTCGGGATCGTGCAGCACACCCACCAGATCCAAGA TGTCCTGGTCCTTGGCTCGCACTTCCACGCCCAAGGCCCCCTGAT GGAAAAACATCCTAGGACATCA	1470
PKD1 - indel	TGTGACTGATGCTGTGGCAGGTCTGAGGAGCTCTGGCCATGGATG GCCCACGTGCTGCTGCCCTACGTCCACGGGAACCAGTCCAGCCCA GAGCTGGGGCCCCCACGGCTGCGGCAGGTGCGGCTGCAGGAAGG TGAGCTGGCAGGGCGTGCCCCAAGACTTAAATCGTTCCTCTTGTT GAGAGAGCAGCCTTTAGCGGA	8197
SERPINA1 - delAG	CCCCACACATTCTTCCCTACAGATACCAGGGTGCAACAAGGTCGT CAGGGTGATCTCACCTTGGAGAGCTTCAGGGGTGCCTCCTCTGTG ACCCCGGAGAGGTCAGCCCCATTGCTGAAGACCTTAGTGATGCCC AGTTGACCCAGGACGCTCTTCAGATCATAGGTTCCAGTAATGGAC AGTTTGGGTAAATGTAAGCT	17980
PAH - delATG	CTAGGAGAATGATGTAAACCTGACCCACATTGAATCTAGACCTTC TCGTTTAAAGAAAGATGAGTATGAATTTTTCACCCATTTGGATAA ACGTAGCCTGCCTGCTCTGACAAACATCATCAAGATCTTGAGGCA TGACATTGGTGCCACTGTCCATGAGCTTTCACGAGATAAGAAGAA AGACACAGGTAAGAATTAGA	604
VWF - delG	CCCTTGTTTCTTCCTCTCTCTGGCTGCACAGCCCCCTCACTCATCC CTGCCTACAAGAAAACTGAAGGGCAGGCACCAGCTCTGTGCCTG GTGACTCACCATGCCCGGGGGGCAGAGGCAGCCAGAGACACAGC CCATGCTCATGCACTCCAGGTCATAGTTCTGGCACGTTTTGGTAC ACTCGAGCCCTTCAGCCCGCA	303

DMD - delG	TTTCTTCCTCAAGATCTGCTAGGATTCTCTCTAGCTCCCCTCTTTC CTCACTCTCTAAGGAAATCAAGATCTGGGCAGGACTACGAGGCT GGCTCAGGGGGGAGTCCTGGTTCAAACTTTGGCAGTAATGCTGGA TTAACAAATGTTCATCATCTCTGGAAAATAAAATCAAAGGTTTTG GTTTTTTCCCCCCCTTATTT	497301
PSEN1 - C>T	ACGTACAGCTGCCCATCCTTCCGGGTATAAAAGCTGACTGACTTA ATGGTAGCCACGACCACCACCATGCAGAGAGTCACAGGGACAAA GAGCATGATCACATGCTTGGCGCCATATTTCAATGTCAGCTCCTC ATCTTCTTCCTCATCTTGCTCCACCACCTGCCGGGAGTTACCCTGG GGTCGTCCATTAGATAATGG	18157
IDUA - G>A	ACTCCTTCACCAAGGGGAGGGGGAGCGAGTGGTGGGAGGCCCGG CCCTGGGTCGGGGGGCGGCTGGGCAACGACCCCACGCGGCGACG GCCCCCCCCCGCCCCGCAGATGAGGAGCAGCTCTGGGCCGAAGT GTCGCAGGCCGGGACCGTCCTGGACAGCAACCACACGGTGGGCG TCCTGGCCAGCGCCCACCGCCCC	11908
IDS - G>A	TGAAGCCAACCCACACAGTATACCTATAGTCTATGGTGCGTATGG AATAGCCCATGATCTTTATATCTTTTAAACTCGGCTTGTCAGAATT CCACTGAGGGATGTCTGAAGGCCGGGGATACTGGCTATAGGCAA TCAGTTCACGGGGATTACCAGGGAGGTACGGATCCTCTTCCAAGT CACGGAATCGAAAATGCTTC	10497
ATP7B - delG	GAGGCAATCACTGCTGGGCGTGGTGCTCTCTGTGGTTTGACCCAC CTCTACTTTTAACCAGCTGCAGAGACAAAAGCCAGCAATACCTTT TTCTGCGGGAAGGCTGCCAGCCTCATTCAGGTGACTGGCCGGTGC ACTCAAAGGGCGCTCACTGTGGGCCAGGATGCCTTCCACGTTGCT GACTTTGCACCCAATTCCAC	88958
LDLR - T>A	TGGCAGAGGCTGCGAGCATGGGGCCCTGGGGCTGGAAATTGCGC TGGACCGTCGCCTTGCTCCTCGCCGCGGCGGGGACTGCAGGTAAG GCTTGCTCCAGGCGCCAGAATAGGTTGAGAGGGAGCCCCCGGGG GGCCCTTGGGAATTTATTTTTTTGGGTACAAATAATCACTCCATCC CTGGGAGACTTGTGGGGTAAT	250987

Target gene & edit	**PrimeDesign Input**
TNNT2 - delGAG	AGCACTGTGCTGGGAGCTACCCTCTCAGAAAGCTCCTTGCTGAGCGGAGAGAAA GCTGAACTCACCCATAAAGACCACAAGCTTCAGCCCAGAATCAGGGTTTCCAAT CCTTTCCCCTAATTTGCTTTCTTCCTCCCTGCTGTAAATCAGGAAGAGAGGGCTC GACGAGAGGAGGAG(-

	GAG)AACAGGAGGAAGGCTGAGGATGAGGCCCGGAAGAAGAAGGCTTTGTCCA ACATGATGCATTTTGGGGGTTACATCCAGAAGGTAGGTAGGGAGCAGCAGGGG TTGCCAGGAGATCCTAGTATAGCCCTGAGGAATGAGGTGTCCACTGCA
MYBPC3 - 25bp del	CATAGATGCCCCCGTCAAAGGGGCAGGGCTTTCTAATCTCCAGAGTCAACACTC CCTGCTTGCTGAACATGCGGAAGCGGGCGTCTTCTCCCAGGTCCAGGCCATTCTT GAACCAGGAAATCTTGGGCTATAAATAAGGTAAAGAGAGGG(- AGGGAAGCCATCCAGGCTGAGAGGG)GACCTGGCAGGGACCCAGGGAGACACA TCTGTGTTTCTACTCGGGGGGTCCCACGAGAGTCCCTGACTATGCCCAAGGCTG GAAACAAACATGGAACCAAGAGTGAGTACCATGGCCCTGCCCAGGGGGAGGAA CCCGGTCCATACACCC
MC4R - delCTCT	TACCATAACATTATGACAGTTAAGCGGGTTGGGATCATCATAAGTTGTATCTGG GCAGCTTGCACGGTTTCAGGCATTTTGTTCATCATTTACTCAGATAGTAGTGCTG TCATCATCTGCCTCATCACCATGTTCTTCACCATGCTGGCTCTCATGGCTTCT(- CTCT)ATGTCCACATGTTCCTGATGGCCAGGCTTCACATTAAGAGGATTGCTGTC CTCCCCGGCACTGGTGCCATCCGCCAAGGTGCCAATATGAAGGGAGCGATTACC TTGACCATCCTGATTGGCGTCTTTGTTGTCTGCTGGGCCCCATTCT
SCN5A - delG4CTTCT	TGCCCTCCATGCTGGGGCCTCTGAGAACCCCAGAATGAGGTTGGTGCCTTCTCTT TGCACTTAGGGGGCCAGGACATCTTCATGACAGAGGAGCAGAAGAAGTACTAC AATGCCATGAAGAAGCTGGGCTCCA(- AGAAGCCCC)AGAAGCCCATCCCACGGCCCCTGGTGAGCCAGGCTCCTGGTTTT GTGGGAGTGGGGAGGATCTTCTGGGGTCCCTGGGAAGCTTTCGTGCCACATGGC CAGCCATCAGAGCCGCTTCACAGTCTTTCAGCCCAGCCTGAGGGGCACTATC
TTR - delGTC	GGAAATGGATCTGTCTGTCTTCTCTCATAGGTGGTATTCACAGCCAACGACTCCG GCCCCCGCCGCTACACCATTGCCGCCCTGCTGAGCCCCTACTCCTATTCCACCAC GGCTGTC(- GTC)ACCAATCCCAAGGAATGAGGGACTTCTCCTCCAGTGGACCTGAAGGACGA GGGATGGGATTTCATGTAACCAAGAGTATTCCATTTTTACTAAAGCAGTGTTTTC ACCTCATATGCTATGTTAGAAGTCCAGGCAGAGACAATAAAACAT
CFH - 24bp del	CCGTGTGTAATATCCCGAGAAATTATGGAAAATTATAACATAGCATTAAGGTGG ACAGCCAAACAGAAGCTTTATTCGAGAACAGGTGAATCAGTTGAATTTGTGTGT AAACGGGGATATCGTCTTTCATCACGTTCTCACACATTGCGAACAACATGTTGG GATGGGAAACTGGAGTATC(- CAACTTGTGCAAAAAGATAGAATC)AATCATAAAGTGCACACCTTTATTCAGAA CTTTAGTATTAAATCAGTTCTCAATTTCATTTTTTATGTATTGTTTTACTCCTTTTT ATTCATACGTAAAATTTTGGATTAATTTGTGAAAATGTAATTATAAGCTGAGAC CGGTGGCTCT

GATA2 - delTGT	GGGAGGGGGGTCGAGGTGGGCGTGGGAGTCCAGCCTGCTGACGCTGCCTTGCC CTCCCAGTCGGCCGCCAGAAGAGCCGGCACCTGTTGTGCAAATTGTCAGACG(-ACA)ACCACCACCTTATGGCGCCGAAACGCCAACGGGGACCCTGTCTGCAACGC CTGTGGCCTCTACTACAAGCTGCACAATGTGAGTGCGCCCCGCCCCGGCCACCC CGCCCCTCCCAGGGGACCTCTGCGCTTTGTGCTGCCAGGCAAGAGG
GLA - del(CAGC)2CCATG	ATGCAGCTGAGGAACCCAGAACTACATCTGGGCTGCGCGCTTGCGCTTCGCTTC CTGGCCCTCGTTTCCTGGGACATCCCTGGGGCTAGAGCACTGGACAATGGATTG GCAAGGACGCCTAC(-CATGGGCTGGCTG)CACTGGGAGCGCTTCATGTGCAACCTTGACTGCCAGGAAG AGCCAGATTCCTGCATCAGGTATCAGATATTGGGTACTCCCTTCCCTTTGCTTTT CCATGTGTTTGGGTGTGTTTGGGGAACTGGAGAGTCTCAACGGGAACAGTTGAG C
PTCH1 - insTCTAC	TTTCATGCAAAGTTCTTCTCTCTTTCCATTGAAACTGTGATGCTCTTCTACCCTGG GCTGCTCCTAACCTGTGCCCTTCTCTGTCCAGTCCCGGCAGCAGAGCCCATCGAG TATGCCCAGTTCCCTTTCTAC(+GTAGA)CTCAACGGCTTGCGGGACACCTCAGAC TTTGTGGAGGCAATTGAAAAAGTAAGGACCATCTGCAGCAACTATACGAGCCTG GGGCTGTCCAGTTACCCCAACGGCTACCCCTTCCTCTTCTGGGAGCAGTACATCG GCCTCCGCCACTGG
HMBS - dup	TGGGGAGCAAGTAGATAGAGGTGGTCCCATGCTTTGCGCCATTGGTTGGGGAAA GATCAGGCCTGATGTCCTAGGATGTTTTTCCATCAGGGGGCCTTGGGCGTGGAA GTGCGAGCCAAGGACCAGGACATCTTGGATCTGGTGGGTGTGCTGCACGATCCC GAGACT(+TTCGCTGC)ATCGCTGAAAGGGCCTTCCTGAGGCACCTGGTAGGGCC TGTGCTCCACCTGTGGAGGGCTGGGGACTTGGAGAGCTGGGAAAGGTGGCAGG GAAGATTTCTTACATGAATGCTCTGTATACAGTGCTAACTCATTCTTGTTGAATG TTGT
PKD1 - indel	CTCACTCGAGGCGGGCATGGGGCAGTAGGGGCTGGAGCGTGTGACTGATGCTGT GGCAGGTCTGAGGAGCTCTGGCCATGGATGGCCCACGTGCTGCTGCCCTACGTC CACGGGAACCAGTCCAGCCCAGAGCTGGGGCCCCCA(-CGGCTGCGGCAGGTG)CGGCTGCAGGAAGGTGAGCTGGCAGGGCGTGCCCCAAG ACTTAAATCGTTCCTCTTGTTGAGAGAGCAGCCTTTAGCGGAGCTCTGGCATCA GCCCTGCTCCCTAGCTGTGTGACCTTTGCCCTCTTAACACCGCCGTTTCCTTCTCT GT
SERPINA1 - delAG	GTCCCAGAAGAACAAGAGGAATGCTGTGCCATGCCTTGAATTTCTTTTCTGCAC GACAGGTCTGCCAGCTTACATTTACCCAAACTGTCCATTACTGGAACCTATGATC TGAAGAGCGTCCTGGGTCAACTGGGCATCACTAAGGTCTTCAGCAATGGGGCTG ACCT(-CT)CCGGGGTCACAGAGGAGGCACCCCTGAAGCTCTCCAAGGTGAGATCACCCT

	GACGACCTTGTTGCACCCTGGTATCTGTAGGGAAGAATGTGTGGGGGCTGCAGC TCTGTCCTGAGGCTGAGGAAGGGGCCGAGGGAAACAAATGAAGAC
PAH - delATG	ATCTTTGGCCTGCGTTAGTTCCAGTGACTGTCTCCTCACCCTCCCCATTCTCTCTT CTAGGAGAATGATGTAAACCTGACCCACATTGAATCTAGACCTTCTCGTTTAAA GAAAGATGAGTATGAATTTTTCACCCATTTGGATAAACGTAGCCTGCCTGCTCT GACAAA(- CAT)CATCAAGATCTTGAGGCATGACATTGGTGCCACTGTCCATGAGCTTTCACG AGATAAGAAGAAAGACACAGGTAAGAATTAGAGGAATTTTGCAACATAAGTAA CTCCACACTGTCTTCATAAAATAACAGCAATATACATATTTAACAG
VWF - delG	AAAAGGAGCCTATCCTGTCGGCCCCCCATGGTCAAGCTGGTGTGTCCCGCTGAC AACCTGCGGGCTGAAGGGCTCGAGTGTACCAAAACGTGCCAGAACTATGACCT GGAGTGCATGAGCATGGGCTGTGTCTCTGGCTGCCTCTGCCCCC(- C)GGGCATGGTGAGTCACCAGGCACAGAGCTGGTGCCTGCCCTTCAGTTTTCTTG TAGGCAGGGATGAGTGAGGGGGCTGTGCAGCCAGAGAGAGGAAGAAACAAGG GCTAAACACGGATAGCAATGGAGGTGGTGCTGGTAATGGGGGCAT
DMD - delG	TTTCTATTTTCAAATACACTCCTGAGTCCCTAACCCCCAAAGCAAAATAAGGGG GGGAAAAAACCAAAACCTTTGATTTTATTTTCCAGAGATGATGAACATTTGTTA ATCCAGCATTACTGCCAAAGTTTGAACCAGGACTCCCCC(- C)TGAGCCAGCCTCGTAGTCCTGCCCAGATCTTGATTTCCTTAGAGAGTGAGGAA AGAGGGGAGCTAGAGAGAATCCTAGCAGATCTTGAGGAAGAAAACAGGTGAGT TTTCTTTCTAGCTTTGTCATTGGTATGCAGAGTGCATACACTTG
PSEN1 - C>T	AATGACAATAGAGAACGGCAGGAGCACAACGACAGACGGAGCCTTGGCCACCC TGAGCCATTATCTAATGGACGACCCCAGGGTAACTCCCGGCAGGTGGTGGAGCA AGATGAGGAAGAAGATGAGGAGCTGACATTGAAATATGGCG(C/T)CAAGCATGT GATCATGCTCTTTGTCCCTGTGACTCTCTGCATGGTGGTGGTCGTGGCTACCATT AAGTCAGTCAGCTTTTATACCCGGAAGGATGGGCAGCTGTACGTATGAGTTTTG TTTTATTATTCTCAAAGCCAGTGTGGCTTTTC
IDUA - G>A	CGCGCTTCCCGGGGTCGGCCTCCGCGTGGCGGGGCCTGGGGACTCCTTCACCAA GGGGAGGGGGAGCGAGTGGTGGGAGGCCCGGCCCTGGGTCGGGGGGCGGCTGG GCAACGACCCCACGCGGCGACGGCCCCCCCCCGCCCCGCAGATGAGGAGCAGC TCT(G/A)GGCCGAAGTGTCGCAGGCCGGGACCGTCCTGGACAGCAACCACACGG TGGGCGTCCTGGCCAGCGCCCACCGCCCCCAGGGCCCGGCCGACGCCTGGCGCG CCGCGGTGCTGATCTACGCGAGCGACGACACCCGCGCCCACCCCAACCG
IDS - G>A	GGACTGCAGGTTCCACCTCGCTGCCCCGTTCCTTCATTTCACGTTGAGCTGTGCA GAGAAGGCAAGAACCTTCTGAAGCATTTTCGATTCCGTGACTTGGAAGAGGATC CGTACCTCCCTGGTAATCCCCGTGAACTGATTGCCTATAGCCAGTATCCC(C/T)G GCCTTCAGACATCCCTCAGTGGAATTCTGACAAGCCGAGTTTAAAAGATATAAA

	GATCATGGGCTATTCCATACGCACCATAGACTATAGGTATACTGTGTGGGTTGG CTTCAATCCTGATGAATTTCTAGCTAACTTTTCTGACATCC
ATP7B - delG	GAACTTGGAACAGAGACCTTGGGATACTGCACGGACTTCCAGGCAGTGCCAGG CTGTGGAATTGGGTGCAAAGTCAGCAACGTGGAAGGCATCCTGGCCCACAGTG AGCGCCCTTTGAGTGCACCGGCCAGTCACCTGAATGAGGCTGGCAGCCTTCC(- C)GCAGAAAAAGGTATTGCTGGCTTTTGTCTCTGCAGCTGGTTAAAAGTAGAGGT GGGTCAAACCACAGAGAGCACCACGCCCAGCAGTGATTGCCTCTGCTGTGCGGC AGACGGTTCATGGCTAAGGCACCCAAGCCTGCCTCCCCACACC
LDLR - T>A	CCGAGTGCAATCGCGGGAAGCCAGGGTTTCCAGCTAGGACACAGCAGGTCGTG ATCCGGGTCGGGACACTGCCTGGCAGAGGCTGCGAGCATGGGGCCCTGGGGCT GGAAATTGCGCTGGACCGTCGCCTTGCTCCTCGCCGCGGCGGGGACTGCAGG(T/ A)AAGGCTTGCTCCAGGCGCCAGAATAGGTTGAGAGGGAGCCCCCGGGGGGCCC TTGGGAATTTATTTTTTTGGGTACAAATAATCACTCCATCCCTGGGAGACTTGTG GGGTAATGGCACGGGGTCCTTCCCAAACGGCTGGAGGGGGCGC

Supplementary Table 3.3: Source data.

Site	Tier	PAM-disrupting	Editing type	Replicate	Edit type	Frequency
TNNT2	1	Y	PE3	1	Desired edit	3.09
TNNT2	1	Y	PE3	2	Desired edit	2.53
TNNT2	1	Y	PE3	3	Desired edit	3.81
MYBPC3	1	Y	PE3	1	Desired edit	9.38
MYBPC3	1	Y	PE3	2	Desired edit	5.57
MYBPC3	1	Y	PE3	3	Desired edit	7.2
MC4R	1	N	PE3	1	Desired edit	2.87
MC4R	1	N	PE3	2	Desired edit	2.29
MC4R	1	N	PE3	3	Desired edit	2.34
SCN5A	1	Y	PE3	1	Desired edit	0.2
SCN5A	1	Y	PE3	2	Desired edit	0.09
SCN5A	1	Y	PE3	3	Desired edit	0.07
TTR	3	Y	PE3b	1	Desired edit	14.04
TTR	3	Y	PE3b	2	Desired edit	9.45
TTR	3	Y	PE3b	3	Desired edit	15.53
TTR	4	Y	PE3	1	Desired edit	23.38
TTR	4	Y	PE3	2	Desired edit	24.24
TTR	4	Y	PE3	3	Desired edit	23.79

TTR	1	N	PE3b	1	Desired edit	10.84
TTR	1	N	PE3b	2	Desired edit	9.56
TTR	1	N	PE3b	3	Desired edit	12.86
TTR	2	N	PE3	1	Desired edit	14.18
TTR	2	N	PE3	2	Desired edit	17.93
TTR	2	N	PE3	3	Desired edit	19.02
CFH	2	Y	PE3	1	Desired edit	0.84
CFH	2	Y	PE3	2	Desired edit	1.02
CFH	2	Y	PE3	3	Desired edit	1.17
CFH	1	N	PE3	1	Desired edit	0.23
CFH	1	N	PE3	2	Desired edit	0.21
CFH	1	N	PE3	3	Desired edit	0.39
GATA2	3	Y	PE3b	1	Desired edit	19.89
GATA2	3	Y	PE3b	2	Desired edit	21.95
GATA2	3	Y	PE3b	3	Desired edit	23.58
GATA2	4	Y	PE3	1	Desired edit	40.4
GATA2	4	Y	PE3	2	Desired edit	35.64
GATA2	4	Y	PE3	3	Desired edit	35.37
GATA2	1	N	PE3b	1	Desired edit	15.71
GATA2	1	N	PE3b	2	Desired edit	14.71
GATA2	1	N	PE3b	3	Desired edit	18.64
GATA2	2	N	PE3	1	Desired edit	27.49
GATA2	2	N	PE3	2	Desired edit	21.83
GATA2	2	N	PE3	3	Desired edit	20.7
GLA	1	Y	PE3b	1	Desired edit	9.13
GLA	1	Y	PE3b	2	Desired edit	8.69
GLA	1	Y	PE3b	3	Desired edit	7.95
GLA	2	Y	PE3	1	Desired edit	14.71
GLA	2	Y	PE3	2	Desired edit	13.33
GLA	2	Y	PE3	3	Desired edit	16.01
PTCH1	1	Y	PE3b	1	Desired edit	0.59
PTCH1	1	Y	PE3b	2	Desired edit	0.48
PTCH1	1	Y	PE3b	3	Desired edit	0.42
PTCH1	2	Y	PE3	1	Desired edit	1.58
PTCH1	2	Y	PE3	2	Desired edit	0.97
PTCH1	2	Y	PE3	3	Desired edit	1.05

HMBS	1	N	PE3b	1	Desired edit	0
HMBS	1	N	PE3b	2	Desired edit	0
HMBS	1	N	PE3b	3	Desired edit	0
PKD1	2	Y	PE3	1	Desired edit	5.92
PKD1	2	Y	PE3	2	Desired edit	5.11
PKD1	2	Y	PE3	3	Desired edit	7.41
PKD1	1	N	PE3	1	Desired edit	6.6
PKD1	1	N	PE3	2	Desired edit	5.37
PKD1	1	N	PE3	3	Desired edit	9.17
SERPINA1	1	Y	PE3b	1	Desired edit	0.47
SERPINA1	1	Y	PE3b	2	Desired edit	0.48
SERPINA1	1	Y	PE3b	3	Desired edit	0.29
PAH	1	N	PE3b	1	Desired edit	0.02
PAH	1	N	PE3b	2	Desired edit	0
PAH	1	N	PE3b	3	Desired edit	0.01
VWF	1	Y	PE3b	1	Desired edit	18.3
VWF	1	Y	PE3b	2	Desired edit	24.59
VWF	1	Y	PE3b	3	Desired edit	25.47
VWF	2	Y	PE3	1	Desired edit	16.58
VWF	2	Y	PE3	2	Desired edit	18.15
VWF	2	Y	PE3	3	Desired edit	18.29
DMD	1	Y	PE3	1	Desired edit	16.63
DMD	1	Y	PE3	2	Desired edit	19.04
DMD	1	Y	PE3	3	Desired edit	19.25
PSEN1	1	Y	PE3	1	Desired edit	43
PSEN1	1	Y	PE3	2	Desired edit	32.44
PSEN1	1	Y	PE3	3	Desired edit	30.47
IDUA	1	Y	PE3b	1	Desired edit	36.31
IDUA	1	Y	PE3b	2	Desired edit	31.35
IDUA	1	Y	PE3b	3	Desired edit	25.9
IDUA	2	Y	PE3	1	Desired edit	21.15
IDUA	2	Y	PE3	2	Desired edit	19.02
IDUA	2	Y	PE3	3	Desired edit	25.62
IDS	1	Y	PE3	1	Desired edit	30.83
IDS	1	Y	PE3	2	Desired edit	29.72
IDS	1	Y	PE3	3	Desired edit	33.53

ATP7B	1	Y	PE3	1	Desired edit	0.01
ATP7B	1	Y	PE3	2	Desired edit	0.02
ATP7B	1	Y	PE3	3	Desired edit	0.01
LDLR	1	N	PE3b	1	Desired edit	3.42
LDLR	1	N	PE3b	2	Desired edit	3.31
LDLR	1	N	PE3b	3	Desired edit	3.75
LDLR	2	N	PE3	1	Desired edit	6.33
LDLR	2	N	PE3	2	Desired edit	4.65
LDLR	2	N	PE3	3	Desired edit	6.27
TNNT2	1	Y	PE3	1	Byproduct	0.35
TNNT2	1	Y	PE3	2	Byproduct	0.59
TNNT2	1	Y	PE3	3	Byproduct	1.1
MYBPC3	1	Y	PE3	1	Byproduct	0.31
MYBPC3	1	Y	PE3	2	Byproduct	0.21
MYBPC3	1	Y	PE3	3	Byproduct	0.28
MC4R	1	N	PE3	1	Byproduct	0.29
MC4R	1	N	PE3	2	Byproduct	0.84
MC4R	1	N	PE3	3	Byproduct	0.22
SCN5A	1	Y	PE3	1	Byproduct	0.12
SCN5A	1	Y	PE3	2	Byproduct	0.04
SCN5A	1	Y	PE3	3	Byproduct	0.03
TTR	3	Y	PE3b	1	Byproduct	0.35
TTR	3	Y	PE3b	2	Byproduct	0.19
TTR	3	Y	PE3b	3	Byproduct	0.34
TTR	4	Y	PE3	1	Byproduct	0.29
TTR	4	Y	PE3	2	Byproduct	0.22
TTR	4	Y	PE3	3	Byproduct	0.26
TTR	1	N	PE3b	1	Byproduct	0.78
TTR	1	N	PE3b	2	Byproduct	0.79
TTR	1	N	PE3b	3	Byproduct	1.05
TTR	2	N	PE3	1	Byproduct	0.34
TTR	2	N	PE3	2	Byproduct	0.31
TTR	2	N	PE3	3	Byproduct	0.21
CFH	2	Y	PE3	1	Byproduct	0.1
CFH	2	Y	PE3	2	Byproduct	0.11
CFH	2	Y	PE3	3	Byproduct	0.07

CFH	1	N	PE3	1	Byproduct	2.9
CFH	1	N	PE3	2	Byproduct	2.99
CFH	1	N	PE3	3	Byproduct	2.64
GATA2	3	Y	PE3b	1	Byproduct	0.53
GATA2	3	Y	PE3b	2	Byproduct	0.83
GATA2	3	Y	PE3b	3	Byproduct	0.86
GATA2	4	Y	PE3	1	Byproduct	1.52
GATA2	4	Y	PE3	2	Byproduct	1.97
GATA2	4	Y	PE3	3	Byproduct	1.98
GATA2	1	N	PE3b	1	Byproduct	0.66
GATA2	1	N	PE3b	2	Byproduct	0.93
GATA2	1	N	PE3b	3	Byproduct	0.85
GATA2	2	N	PE3	1	Byproduct	1.56
GATA2	2	N	PE3	2	Byproduct	1.33
GATA2	2	N	PE3	3	Byproduct	1.01
GLA	1	Y	PE3b	1	Byproduct	0.18
GLA	1	Y	PE3b	2	Byproduct	0.3
GLA	1	Y	PE3b	3	Byproduct	0.24
GLA	2	Y	PE3	1	Byproduct	0.18
GLA	2	Y	PE3	2	Byproduct	0.16
GLA	2	Y	PE3	3	Byproduct	0.16
PTCH1	1	Y	PE3b	1	Byproduct	0.06
PTCH1	1	Y	PE3b	2	Byproduct	0.1
PTCH1	1	Y	PE3b	3	Byproduct	0.05
PTCH1	2	Y	PE3	1	Byproduct	0.18
PTCH1	2	Y	PE3	2	Byproduct	0.08
PTCH1	2	Y	PE3	3	Byproduct	0.14
HMBS	1	N	PE3b	1	Byproduct	0.04
HMBS	1	N	PE3b	2	Byproduct	0.05
HMBS	1	N	PE3b	3	Byproduct	0.03
PKD1	2	Y	PE3	1	Byproduct	0.25
PKD1	2	Y	PE3	2	Byproduct	0.33
PKD1	2	Y	PE3	3	Byproduct	0.36
PKD1	1	N	PE3	1	Byproduct	0.29
PKD1	1	N	PE3	2	Byproduct	0.51
PKD1	1	N	PE3	3	Byproduct	0.34

SERPINA1	1	Y	PE3b	1	Byproduct	0.09
SERPINA1	1	Y	PE3b	2	Byproduct	0.21
SERPINA1	1	Y	PE3b	3	Byproduct	0.12
PAH	1	N	PE3b	1	Byproduct	0.04
PAH	1	N	PE3b	2	Byproduct	0.04
PAH	1	N	PE3b	3	Byproduct	0.05
VWF	1	Y	PE3b	1	Byproduct	2.17
VWF	1	Y	PE3b	2	Byproduct	3.39
VWF	1	Y	PE3b	3	Byproduct	3.22
VWF	2	Y	PE3	1	Byproduct	3.94
VWF	2	Y	PE3	2	Byproduct	4.38
VWF	2	Y	PE3	3	Byproduct	5.74
DMD	1	Y	PE3	1	Byproduct	8.82
DMD	1	Y	PE3	2	Byproduct	2.21
DMD	1	Y	PE3	3	Byproduct	7.48
PSEN1	1	Y	PE3	1	Byproduct	0.46
PSEN1	1	Y	PE3	2	Byproduct	0.66
PSEN1	1	Y	PE3	3	Byproduct	0.48
IDUA	1	Y	PE3b	1	Byproduct	1.36
IDUA	1	Y	PE3b	2	Byproduct	1.3
IDUA	1	Y	PE3b	3	Byproduct	1.32
IDUA	2	Y	PE3	1	Byproduct	2.02
IDUA	2	Y	PE3	2	Byproduct	1.87
IDUA	2	Y	PE3	3	Byproduct	1.02
IDS	1	Y	PE3	1	Byproduct	0.36
IDS	1	Y	PE3	2	Byproduct	0.38
IDS	1	Y	PE3	3	Byproduct	0.33
ATP7B	1	Y	PE3	1	Byproduct	0.24
ATP7B	1	Y	PE3	2	Byproduct	0.35
ATP7B	1	Y	PE3	3	Byproduct	0.13
LDLR	1	N	PE3b	1	Byproduct	0.83
LDLR	1	N	PE3b	2	Byproduct	0.97
LDLR	1	N	PE3b	3	Byproduct	0.9
LDLR	2	N	PE3	1	Byproduct	0.55
LDLR	2	N	PE3	2	Byproduct	0.89
LDLR	2	N	PE3	3	Byproduct	0.5

3.9 Acknowledgements

L.P. is supported by the National Human Genome Research Institute (NHGRI) Career Development Award (R00HG008399), Genomic Innovator Award (R35HG010717) and CEGS RM1HG009490. J.K.J. is supported by NIH R35 GM118158, NIH RM1 HG009490, the Robert B. Colvin, M.D. Endowed Chair in Pathology, and the Desmond and Ann Heathwood MGH Research Scholar Award. D.R.L. is supported by the Merkin Institute of Transformative Technologies in Healthcare, US NIH grants U01AI142756, RM1HG009490, R01EB022376, and R35GM118062, and the HHMI. A.V.A. acknowledges a Jane Coffin Childs postdoctoral fellowship. J.G. was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 416375182.

3.10 Author contributions

J.Y.H. developed PrimeDesign. J.Y.H. and J.G. designed the experiments. R.S. and J.S. performed the experiments and analyzed the data. A.V.A., J.G., K.P., and K.C.L provided feedback during the development of PrimeDesign. M.W.S. contributed to the ClinVar analysis. L.P., J.K.J., and D.R.L. supervised the project and provided feedback and guidance. J.Y.H., L.P., J.K.J., and D.R.L. wrote the manuscript with input from all other authors.

4 *In situ* saturation mutagenesis and optimization of CRISPR prime editing in human cells with MOSAIC

4.1 Abstract

Prime editing offers unprecedented versatility and precision for the installation of DNA edits *in situ*. Here we describe the multiplexing of site-specific alterations for *in situ* characterization (MOSAIC) with prime editing to enable the installation of thousands of defined edits in a pooled fashion. Using MOSAIC, we demonstrate *in situ* saturation mutagenesis of the *BCR-ABL1* oncogene to identify drug resistant variants and *IRF1* untranslated region (UTR) to map non-coding regulatory elements involved in transcriptional initiation. Furthermore, we developed a pooled screening strategy for high-throughput prime editing guide RNA (pegRNA) optimization and demonstrate its utility in assessing >18,000 designs, with up to 210 designs in a single pool, to identify high efficiency pegRNA constructs targeting genomic sites.

4.2 Introduction and background

The ability of prime editing (PE)[68] to install short substitution, insertion, and deletion edits greatly expands the genetic diversity that can be introduced at target genomic loci compared to CRISPR-Cas nucleases[24–26] and base editors (BE)[38–40]. To utilize the versatility and precision of PE for pooled genetic screens, we developed the multiplexing of site-specific alterations for *in situ* characterization (MOSAIC) which enables the installation of thousands of defined edits in a pooled fashion. We showcase the utility of MOSAIC for saturation mutagenesis of both coding and non-coding genomic regions as well as for the high-throughput optimization of pegRNA designs.

4.3 Results

4.3.1 MOSAIC for in situ saturation mutagenesis

For the construction of pooled pegRNA libraries, which we termed pegPools, we developed a polymerase chain reaction (PCR)-based method which does not involve bacterial DNA cloning or virus production. The editing versatility of PE lies in the design of its pegRNA, where the 3’ extension encodes the genetic information to be installed into the genome. Using mixed base

synthesis or oligo pools, we diversified the 3' extension of the pegRNA and assembled the final pegRNA products through two PCR steps to generate the pegPools (**Supplementary Note 4.1**). The purified pegPools were then transfected along with prime editor construct to achieve pooled prime editing in target cells (**Fig. 4.1a**). We demonstrated the utility of pegPools for *in situ* saturation mutagenesis at four genomic sites by installing single- or triple-base substitution edits across sequence windows up to 30 bp (**Fig. 4.1b, Supplementary Table 4.1**). While editing frequencies were relatively uniform across the target editing window for nucleotides outside of the protospacer adjacent motif (PAM), we observed significantly higher editing efficiencies at bases within the PAM compared to bases outside of this region likely due to the ability for re-targeting without PAM disruption (**Fig. 4.1c,d; Supplementary Fig. 4.1a-f**). When we examined the individual alleles, we found the intended DNA diversification within the target editing windows, validating MOSAIC as a strategy for *in situ* saturation mutagenesis (**Fig. 4.1e**). Additionally, we constructed pegPools to introduce randomized insertion edits through mixed base synthesis (oligo orders with IUPAC 'N' bases) of lengths 3, 6, and 9 bp, and were able to install nearly 10,000 unique insertion edits into the genome with a single transfection (**Fig. 4.1f, Supplementary Fig. 4.2a,b, Supplementary Table 4.1**).

To demonstrate the application of MOSAIC for *in situ* saturation mutagenesis of coding sequences, we installed amino acid variants within the tyrosine kinase *ABL1* in chronic myeloid leukemia (CML) cells harboring the *BCR-ABL1* gene fusion. We focused on identifying protein variants resistant to the tyrosine kinase inhibitor imatinib, which is used as a first-line therapy in *BCR-ABL1*-positive CML patients. We screened 84 pegRNA and ngRNA combinations targeting the imatinib binding site in *ABL1* in HEK293T cells and identified several pegRNA-ngRNA combinations that mediated efficient mutagenesis (**Supplementary Fig. 4.3, Supplementary Table 4.2**). Using the most efficient pegRNA and ngRNA combination identified, we constructed a pegPool to introduce codon saturation mutagenesis (NNK mixed base synthesis) spanning amino acids 311-318 corresponding to the imatinib binding site in ABL1. Next, we electroporated this pegPool library, ngRNA, and PE2 plasmid into K562 cells, a CML patient-derived cell line harboring endogenous *BCR-ABL1* (**Fig. 4.1g**). After 3 days, the K562 cells were treated with DMSO (control) or imatinib, and cultured for an additional 7 days. Following targeted amplicon sequencing of *BCR-ABL1* complementary DNA (cDNA) prepared from the

mRNA of edited cells, we observed complete coverage of NNK codons encoding every amino acid variant across all targeted *BCR-ABL1* codons (**Supplementary Fig. 4.4a-g, Supplementary Fig. 4.5a-d**). Overall, genetic variation at residue 315 was significantly enriched in the imatinib treatment group compared to the DMSO control group ($P<0.01$; **Fig. 4.1h**). We identified specific amino acid variants at position 315 that conferred resistance to imatinib treatment, including the T315I "gatekeeper" mutation commonly found in CML patients (**Fig. 4.1i, $p<0.01$; Supplementary Fig. 4.6, Supplementary Table 4.3**). Interestingly, we also identified several other imatinib resistant variants in T315M, T315L, and T315E (significantly enriched in treatment group compared to control, $P<0.01$) that haven't been observed in imatinib-resistant patient samples. We note that these clinically-unobserved amino acid changes require at least 2 base changes from the reference *ABL1* sequence, in contrast to the single base change needed for the clinically-observed T315I mutation. These results demonstrate the capability of MOSAIC to comprehensively screen and identify potential drug resistant variants *in situ* to profile new or existing drugs, and can be broadly applied to characterize resistance susceptibility during drug development.

Additionally, we utilized MOSAIC for *in situ* saturation mutagenesis of non-coding sequences across the 5' untranslated region (UTR) of the transcription factor gene *IRF1* to map regulatory elements involved in its transcriptional initiation (**Supplementary Fig. 4.7a**). We screened 112 pegRNA-ngRNA combinations targeting the *IRF1* 5' UTR in HEK293T cells, chose an efficient pegRNA-ngRNA combination, and constructed a pegPool library installing triple-base randomized substitution edits (NNN mixed base synthesis) across an editing window of 36 bp (**Supplementary Fig. 4.7b, Supplementary Table 4.4**). Following transfection of the pegPool library, ngRNA, and PE2 plasmid into HEK293T cells, we performed targeted amplicon sequencing of both the genomic DNA and cDNA prepared from the edited cells. Based on the proportion of variants on cDNA relative to genomic DNA, we observed statistically significant depletion of two contiguous 3 bp tiles across the entire mutagenesis window (**Supplementary Fig. 4.7c**; $P<0.01$). Notably, these depleted tiles directly overlap the consensus motif of the core downstream promoter element (DPE), supporting its role in the initiation of gene transcription by RNA polymerase II. Thus, we show that MOSAIC can be broadly used to finely-map non-coding regulatory elements involved in gene regulation at base-pair resolution.

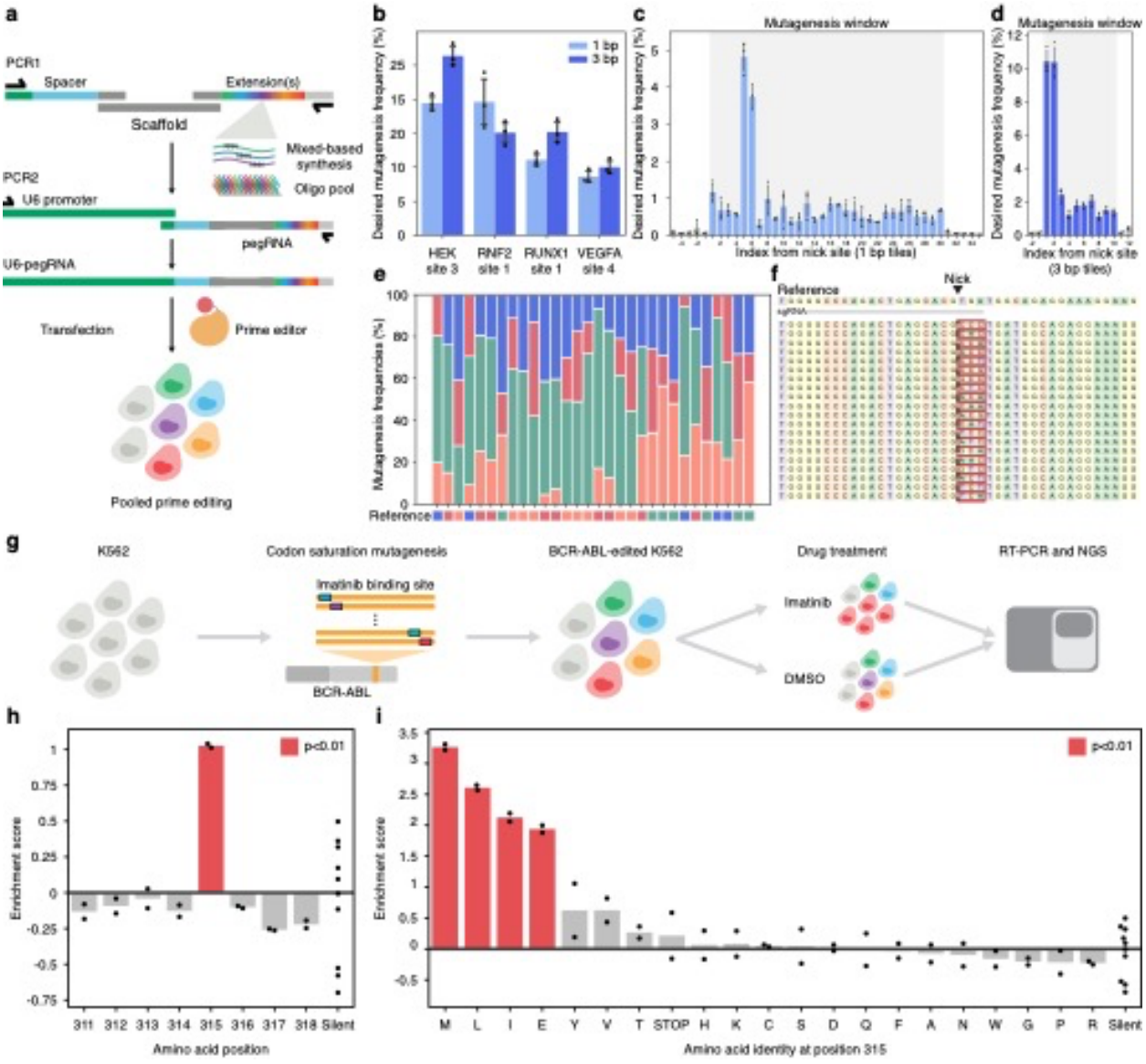


Figure 4.1: Application of MOSAIC for *in situ* saturation mutagenesis. (a) Schematic of the PCR protocol for the rapid generation of pooled pegRNA constructs (pegPools). **(b)** Pooled prime editing efficiencies for the installation of single- or triple-base saturation mutagenesis across genomic targets. **(c,d)** Saturation mutagenesis across HEK site 3 for the installation of single-base and triple-base substitutions. **(e)** Proportion of substitution edits for single-base saturation mutagenesis at HEK site 3. **(f)** Visualization of the top allelic outcomes using a pegPool library encoding randomized insertion edits of length 3. **(g)** Schematic for the *in situ* functional screening of genetic variants associated with imatinib resistance in K562. **(h)** Enrichment scores of amino acid positions within the BCR-ABL1 imatinib binding site

associated with imatinib resistance. **(i)** Enrichment scores of amino acid variants at position 315 associated with imatinib resistance.

4.3.2 MOSAIC for high-throughput pooled pegRNA optimization

Next, we sought to demonstrate the utility of MOSAIC for the high-throughput pooled screening of pegRNA designs. The design of pegRNAs, a major determinant of prime editing efficiencies, are target site-specific and can take on many PBS and RTT length combinations[68,112,113]. We designed specific barcodes for each combination of PBS and RTT length, and reasoned that the installation frequency of these barcodes, encoded by unique editing events, could provide quantitative information on pegRNA design efficiencies in an experiment using pegPools (**Fig. 4.2**). We constructed pegPool libraries consisting of different pegRNA designs using a 4-mer insertion barcoding strategy (**Fig. 4.2a**). To ensure that these barcodes provided quantitative information for the installation of other types of edits of interest, we compared the efficiencies pegRNAs installing barcode insertions to that of various point mutations used in Anzalone *et al*.[68]. We observed a strong correlation between efficiencies of installing barcode insertions and point mutations spanning 90 different pegRNA designs across three different genomic target sites (**Fig. 4.2b, Supplementary Table 4.5**). Subsequently, we sought to establish that the relative editing efficiencies of pegRNAs tested in a pooled format recapitulated their editing efficiencies when tested individually. We constructed three different pegPools with >70 pegRNA designs each and compared the barcode insertion frequencies of the pegRNAs in the pegPool with the editing frequencies of their individually-assessed pegRNA counterparts in HEK293T cells (**Supplementary Table 4.6**). We found strong correlation between the pooled and individual pegRNA efficiencies at all three sites tested (**Fig. 4.2c**). Additionally, we found a high degree of correlation between the editing efficiencies of our pooled pegRNA experiments and pegRNA optimizations performed in Anzalone *et al*.[68] (Spearman R = 0.9; **Supplementary Fig. 4.8**).

To perform pegRNA design characterizations at a larger scale, we designed and constructed pegPool libraries to assess 18,690 unique pegRNA designs across 89 different spacers (210 designs per spacer) (**Supplementary Table 4.7, 4.8**). Following transfection of each pegPool library in HEK293T cells and subsequent targeted amplicon sequencing, we generated individual

pegRNA profiles for each spacer based on the installation frequencies of the barcoded PBS and RTT length combinations (**Supplementary Fig. 4.9**). Using these individual pegRNA profiles, we constructed an averaged pegRNA profile across all spacers with single-nucleotide resolution of the PBS and RTT elements (**Fig. 4.2d**). On average, we observed the highest editing efficiency when the PBS length was 12 nt and RTT was 10 nt (the minimum length tested) (**Supplementary Fig. 4.10**). However, we observed deviations from this averaged profile when we examined the individual pegRNAs, highlighting the target site-specific nature of pegRNA design (**Fig. 4.2e,f**). From the individual pegRNA optimization profiles, we compared the editing frequencies of top-scoring designs versus 25th percentile designs for the installation of all possible single base substitutions and a variety of small insertion and deletion mutations across 20 pegRNA spacers (**Fig. 4.2g; Supplementary Table 4.9, 4.10; Supplementary Note 4.2**). We found that the efficiencies of top-scoring pegRNAs achieved editing efficiencies upwards of 39.8% with PE2 and were on average 2.6-fold more efficient in installing the desired edits compared to the 25th percentile pegRNAs (**Fig. 4.2h**).

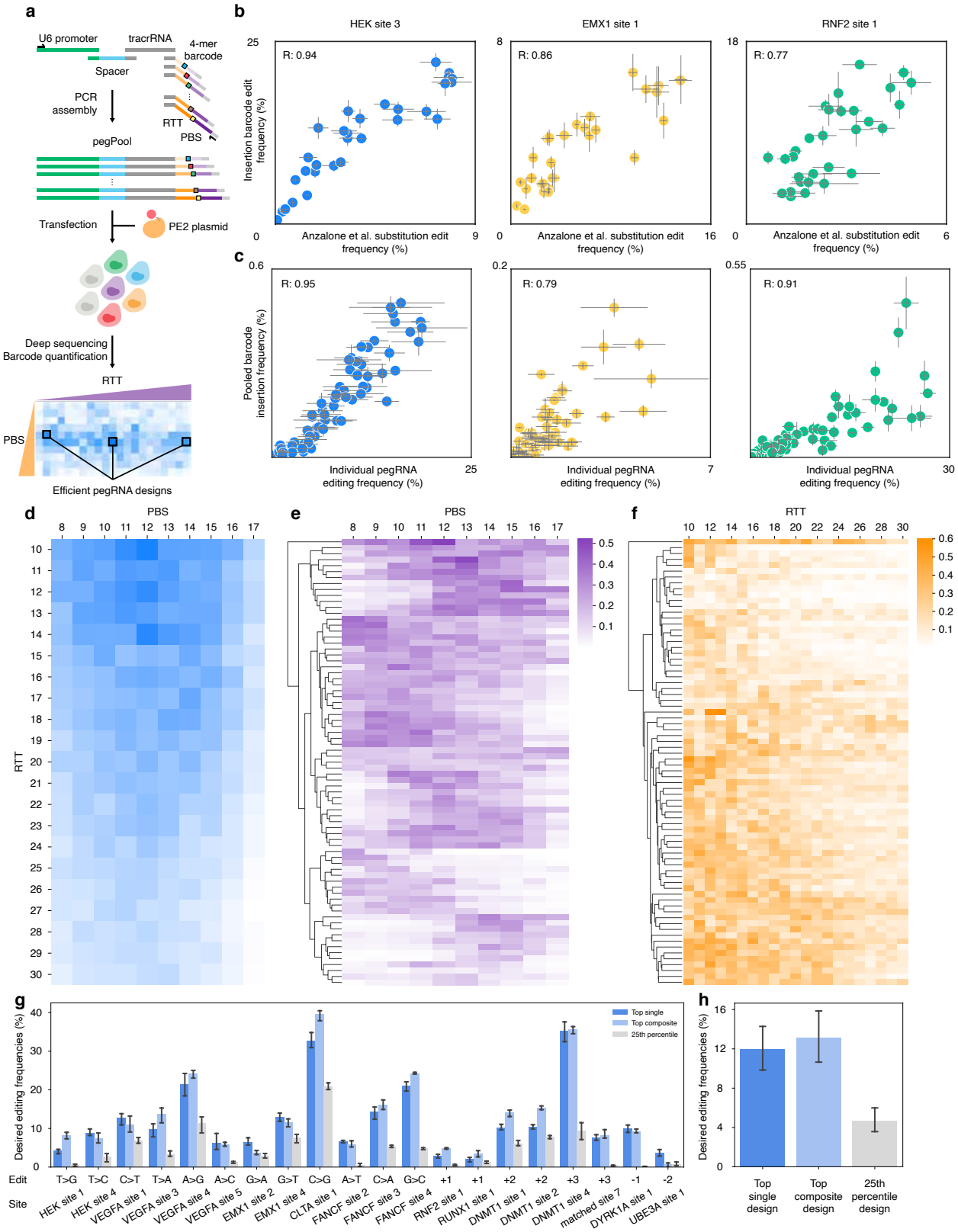


Figure 4.2: High-throughput pooled pegRNA optimization. (a) Overview of MOSAIC for the high-throughput pooled screening of pegRNA designs. **(b)** Scatterplots of prime editing

efficiencies when installing a 4-mer insertion barcode or substitution edit across various pegRNA designs and genomic sites. **(c)** Scatterplots of prime editing efficiencies when installing an insertion barcode with pegRNAs in pooled format or individually. Correlation (R) was determined with Spearman's Rho. **(d)** Averaged pegRNA design profile across varying PBS and RTT lengths. **(e)** Averaged PBS profiles from individual pegRNA design profiles. **(f)** Averaged RTT profiles from individual pegRNA design profiles. **(g)** Comparison of top pegRNA designs to 25th percentile designs from individual pegRNA design profiles. **(h)** Average prime editing efficiencies of top pegRNA designs to 25th percentile designs from individual pegRNA design profiles.

4.4 Discussion and conclusions

In conclusion, MOSAIC is a generalizable method for the installation of thousands of defined edits in a pooled fashion without the need for bacterial DNA cloning or virus production. We envision MOSAIC will greatly enhance the ability for researchers to perform *in situ* saturation mutagenesis screens for the identification of potential drug-resistant variants and critical elements underlying gene regulation, in addition to the general exploration of genetic variation for protein engineering and modification of gene expression. Lastly, our approach also enables high-throughput pooled pegRNA optimization to achieve higher editing efficiencies and vastly expands the range of applications of prime editing.

4.5 Materials and methods

PCR-generated prime editing guide RNAs.

Prime editing guide RNAs (pegRNAs) and nicking sgRNAs (ngRNAs) constructs used in this study were in the form of PCR products. A detailed protocol is available in Supplementary Note 1. Briefly, two sequential PCR assembly steps were performed to construct the pegRNA or ngRNA constructs. The first PCR assembly step assembles a spacer sequence oligo, tracrRNA sequence oligo, and 3' extension sequence oligo(s) (for pegRNAs) into a product consisting of the pegRNA or ngRNA sequence. The second PCR assembly step fuses the U6 promoter sequence with the PCR product from the first assembly step to construct pegRNA or ngRNA constructs capable of expression in human cells. Standard PCR protocol was used for all

reactions and PCR reactions were cleaned up using paramagnetic beads and 75% ethanol washes to isolate desired amplification products.

Human cell culture.

STR-authenticated HEK293T (CRL-3216) and K562 (CCL-243) were used in this study. HEK293T cells were grown in Dulbecco's modified Eagle medium (DMEM) (Gibco) with 10% heat-inactivated fetal bovine serum (FBS) (Gibco) supplemented with 1% penicillin-streptomycin (Gibco) antibiotic mix. K562 cells were grown in Roswell Park Memorial Institute (RPMI) 1640 Medium (Gibco) with 10% FBS supplemented with 1% pen-strep and 1% GlutaMAX (Gibco). Cells were grown at 37 °C in 5% CO_2 incubators and periodically passaged on reaching around 80% confluency. Cell culture media supernatant was tested for mycoplasma contamination every 4 weeks using the MycoAlert PLUS mycoplasma detection kit (Lonza) and all tests were negative throughout the experiments.

Cell transfections.

HEK293T cells were seeded at 1.25×10^4 cells per well into 96-well flat bottom cell culture plates (Corning) or 3×10^5 cells per well into 6-well cell culture plates (Corning). Then, 24 h post-seeding, cells were transfected with 40 ng of prime editor plasmid and 13.3 ng of pegRNA PCR product (and 4.4 ng nicking sgRNA PCR product for PE3) using 0.6 µl of TransIT-X2 (Mirus) lipofection reagent for experiments in 96-well plates, or 1500 ng prime editor plasmid and 500 ng of pegRNA PCR product (and 166.6 ng nicking sgRNA PCR product for PE3) and 15 µl TransIT-X2 for experiments in 6-well plates. K562 cells were electroporated using the SF Cell Line Nucleofector X Kit L (Lonza), according to the manufacturer's protocol with 1×10^6 cells per nucleofection and 6000 ng prime editor plasmid, 1500 ng pegRNA PCR product (and 500 ng nicking sgRNA PCR product for PE3). Then, 72 h post-transfection, cells were lysed for extraction of genomic DNA (gDNA) or RNA. For the drug resistance experiments, K562 cells were treated with 1.25 µM imatinib (Sigma Aldrich) or DMSO (Sigma Aldrich) 72 h post-transfection, and cultured for an additional 7 days before RNA extraction.

DNA extraction.

HEK293T cells were washed with 1× PBS (Corning) and lysed overnight by shaking at 55 °C with 43.5 µl of gDNA lysis buffer (100 mM Tris-HCl at pH 8, 200 mM NaCl, 5 mM EDTA, 0.05% SDS) supplemented with 5.25 µl of 20 mg ml^{-1} Proteinase K (NEB) and 1.25 µL of 1 M DTT (Sigma) per well for experiments in 96-well plates, or with 435 µl DNA lysis buffer, 52.5 µl Proteinase K and 12.5 µl 1 M DTT per well for experiments in 6-well plates. K562 cells were centrifuged for 5 min, media removed and lysed overnight by shaking at 55 °C with 870 µl DNA lysis buffer, 105 µl Proteinase K and 25 µl 1 M DTT per well in 1.5 mL Eppendorf tubes. Subsequently, gDNA was extracted from lysates using 1–2× paramagnetic beads, washed twice with 75% ethanol, and eluted in 50-200 µl of 0.1× EB buffer. DNA extraction was performed using a Biomek FX^P Laboratory Automation Workstation (Beckman Coulter) when using 96-well plates.

RNA extraction and reverse transcription.

At 72 h post-transfection, total RNA was extracted from cells using the NucleoSpin RNA Plus Kit (Clontech, 740984.250). Following RNA extraction, 50–250 ng of purified RNA was used for cDNA synthesis using a High-Capacity RNA-to-cDNA Kit (Thermo Fisher, 4387406). The resulting cDNA was then used downstream as a template for PCR amplification for targeted amplicon sequencing.

Oligonucleotide library design.

An oligonucleotide library containing 21,210 members was synthesized by Agilent (Supplementary Table 8). Each library member contained a spacer sequence, tracrRNA sequence, and pegRNA 3’ extension of varied length. For each spacer sequence, a total of 210 PBS-RTT length combinations were designed for the 3’ extension with PBS lengths spanning 8-17 nt and RTT lengths typically spanning 10-30 nt, where each of these PBS-RTT length combinations were uniquely barcoded with a 4-mer insertion edit placed at the +1 position. Unique flanking sequences were used to amplify sub-libraries from the oligonucleotide pool corresponding to a specific pool of pegRNA designs for a single spacer sequence. Following the amplification of the sub-libraries, subsequent PCR reactions were used to fuse a U6 promoter

element to each of the sub-libraries for expression in human cells. Purified PCR products for these sub-libraries were then used for transfection to perform pooled pegRNA optimizations.

Targeted amplicon sequencing.

The gDNA and cDNA concentrations of samples were measured using the Qubit dsDNA HS Assay Kit (Thermo Fisher). The first PCR was performed to amplify the regions of interest (200–250 bp) using 50–100 ng of gDNA or cDNA. Primers for PCR1 included Illumina-compatible adapter sequences (Supplementary Table 10). A synergy HT microplate reader (BioTek) was then used at 485/528 nm with the Quantifluor dsDNA quantification system (Promega) to measure the concentration of the first PCR products. PCR products from different genomic amplicons were then pooled and cleaned with 0.7x paramagnetic beads. The second PCR was performed to attach unique barcodes to each amplicon using 50–200 ng of the pooled PCR1 products and barcodes that correspond to Illumina TruSeq CD indexes. The PCR2 products were again cleaned with 0.7x paramagnetic beads and measured with the Quantifluor system before final pooling. The final library was sequenced using an Illumina Miseq (Miseq Reagent Kit v.2; 300 cycles, 2 × 150 bp, paired-end). The FASTQ files were downloaded from BaseSpace (Illumina).

Data analysis.

Amplicon sequencing data were analyzed with CRISPResso2 v.2.0.31. Depending on position and length of the intended edit(s), parameters for CRISPResso2 analysis were modified (-w, -wc, -qwc) to center the quantification window around the intended editing window. Downstream analysis was conducted using Python 3.7.6 with data sourced from 'Quantification_window_nucleotide_frequency_table.txt', 'CRISPResso_quantification_of_editing_frequency.txt', and 'Alleles_frequency_table.txt'. Enrichment scores were calculated using the $\log_2$ fold-change of normalized sample counts over control counts.

Statistics and data reporting.

Spearman's rank-order correlation was used for all correlation statistics. A two-tailed Student's *t*-test with *P* values adjusted for multiple testing (Bonferroni) was used to calculate

the *P* values in Fig. 1h,i, Supplementary Fig. 6, Supplementary Fig. 7c, and Supplementary Table 3. The error bars in all dot and bar plots show the standard deviation (s.d.) and were plotted with seaborn (0.10.0). The measure of center for the error bars is the mean. We did not predetermine sample sizes based on statistical methods. Investigators were not blinded to experimental conditions or assessment of experimental outcomes.

4.6 Supplementary notes

4.6.1 Supplementary Note 4.1: PCR-generated prime editing guide RNAs (pegRNAs)

The construction of pegRNAs or ngRNAs by PCR requires two sequential steps. The first PCR steps (labeled PCR1 below) are for the amplification of the "U6 promoter PCR fragment", "pegRNA PCR fragment", and "ngRNA PCR fragment". Following the completion of these first PCR reactions, the desired amplification products are cleaned and isolated with paramagnetic beads and two 75% ethanol washes. The second PCR steps (labeled PCR2 below) fuse the "U6 promoter PCR fragment" to either the "pegRNA PCR fragment" or "ngRNA PCR fragment" to obtain pegRNA and ngRNA constructs, respectively. Following the completion of these second PCR reactions, the desired amplification products are cleaned and isolated with paramagnetic beads and two 75% ethanol washes. These purified PCR products are ready for transfection into mammalian cells. Standard PCR protocol is recommended based on the PCR kit used.

----------PCR1----------

PCR1 reaction to amplify "U6 promoter PCR fragment"

Component	**Volume (µL)**
5X PCR Buffer	10
10 mM dNTPs	1
10 µM *U6 forward primer*	2.5
10 µM *U6 reverse primer*	2.5
U6 promoter-containing plasmid (1-5 ng/uL)	1
Polymerase	0.5
Water	32.5

Total	50

*Clean up with 0.7X paramagnetic beads

PCR1 reaction to assemble “pegRNA PCR fragment”

Component	**Volume (µL)**
5X PCR Buffer	10
10 mM dNTPs	1
10 µM *pegRNA forward primer*	2.5
10 µM *pegRNA reverse primer*	2.5
10 µM *tracrRNA oligo*	1
10 µM *Spacer oligo*	1
10 µM *pegRNA extension oligo 1*	1
10 µM *pegRNA extension oligo 2 (optional)*	0
Polymerase	0.5
Water	30.5
Total	50

*Clean up with 1.2X paramagnetic beads

PCR1 reaction to assemble “ngRNA PCR fragment”

Component	**Volume (µL)**
5X PCR Buffer	10
10 mM dNTPs	1
10 µM *ngRNA forward primer*	2.5
10 µM *ngRNA reverse primer*	2.5
10 µM *tracrRNA oligo*	1
10 µM *Spacer oligo*	1
Polymerase	0.5

Water	31.5
Total	50

*Clean up with 1.2X paramagnetic beads

----------PCR2----------

PCR2 reaction to fuse "U6 promoter PCR fragment" with "pegRNA PCR fragment"

Component	**Volume (µL)**
5X PCR Buffer	10
10 mM dNTPs	1
10 µM *U6 forward primer*	2.5
10 µM *pegRNA reverse primer*	2.5
"U6 promoter PCR fragment" (1-5 ng/uL)	1
"pegRNA PCR fragment" (1-5 ng/uL)	1
Polymerase	0.5
Water	31.5
Total	50

*Clean up with 0.7X paramagnetic beads

PCR2 reaction to fuse "U6 promoter PCR fragment" with "ngRNA PCR fragment"

Component	**Volume (µL)**
5X PCR Buffer	10
10 mM dNTPs	1
10 µM *U6 forward primer*	2.5
10 µM *ngRNA reverse primer*	2.5
"U6 promoter PCR fragment" (1-5 ng/uL)	1
"pegRNA PCR fragment" (1-5 ng/uL)	1
Polymerase	0.5

Water	31.5
Total	50

*Clean up with 0.7X paramagnetic beads

Oligonucleotide and primer sequences

U6 forward primer: CTGTACAAAAAAGCAGGCTTTAAAGGAACCAATTC

U6 reverse primer: GGTGTTTCGTCCTTTCCACAAGATATATAAAGC

pegRNA forward primer: GCTTTATATATCTTGTGGAAAGGACGAAACACC

pegRNA reverse primer: GCAGCACGTGATACACCAAAAAAA

ngRNA forward primer: GCTTTATATATCTTGTGGAAAGGACGAAACACC

ngRNA reverse primer: GCAGCACGTGATACACCAAAAAAAGCACCGACTCGGTGCC

Spacer oligo

ATATCTTGTGGAAAGGACGAAACACCNNNNNNNNNNNNNNNNNNNGTTTTAGAGCTAGA AATAGCAAGTTAAAATAAG

where the Nx20 sequence represents the spacer sequence. Make sure the 5' element is a G base for efficient transcription off the U6 promoter.

tracrRNA oligo

GCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCTATT TCTAGCTCTAAAAC

pegRNA extension oligo 1

AAGTGGCACCGAGTCGGTGC + [insert pegRNA 3' extension sequence 5' to 3'] + TTTTTTTGGTGTATCACGTGCTGC

Depending on the length of the pegRNA extension, additional pegRNA extension oligos may be needed to assemble the full pegRNA 3' extension. For example, if the pegRNA extension is too long to fit on a single oligo and requires two pegRNA extension oligos, it would take on the following form:

pegRNA extension oligo 1

AAGTGGCACCGAGTCGGTGC + [insert pegRNA 3' extension sequence 5' to 3']

pegRNA extension oligo 2

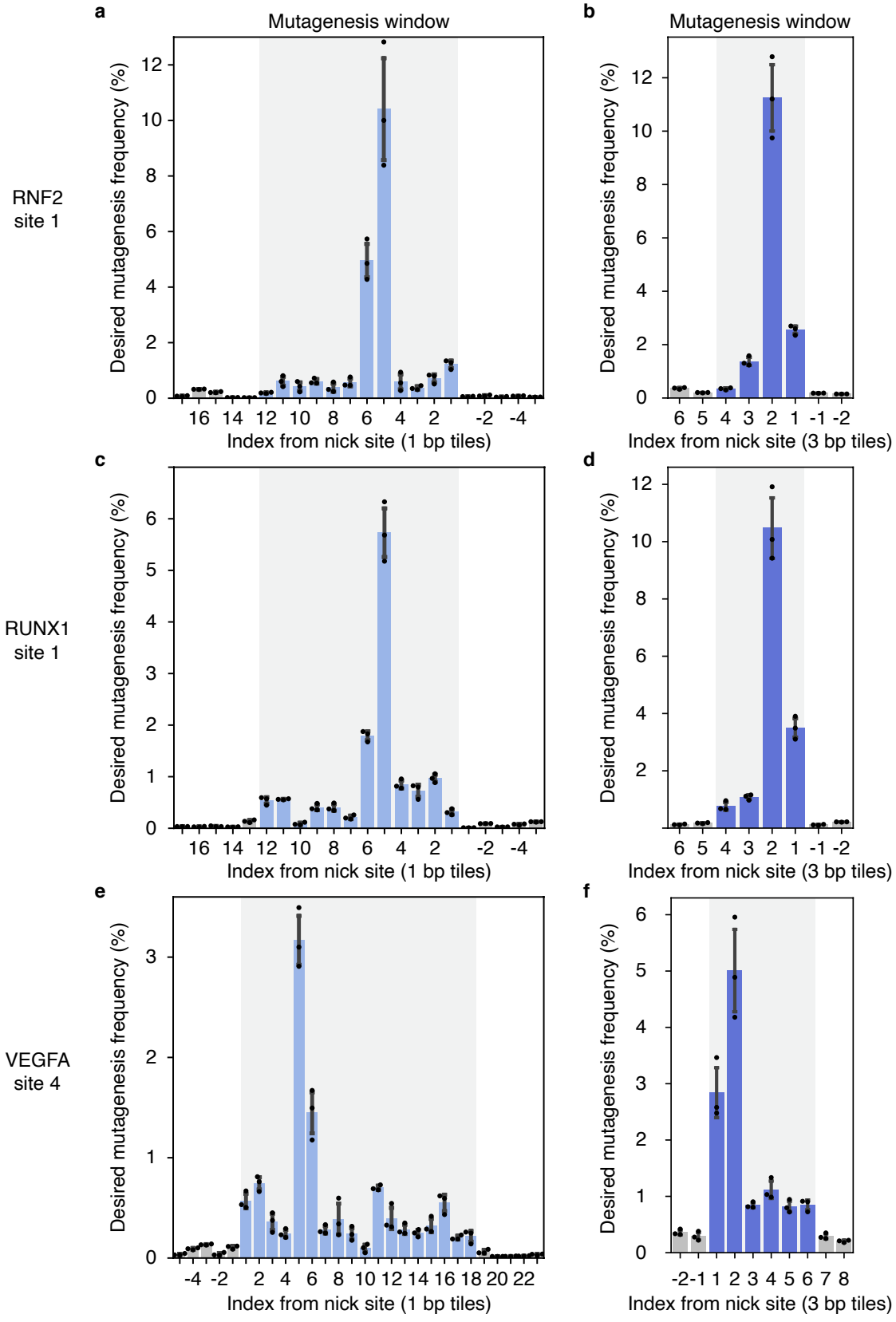
GCAGCACGTGATACACCAAAAAAA + [insert reverse complement pegRNA 3' extension sequence 5' to 3']

Important note: The *pegRNA extension oligos 1 and 2* will require some degree of sequence overlap at the 3' end to mediate successful PCR assembly. We recommend at least 15 bp of overlap (or sufficient melting temperature of ~60°C).

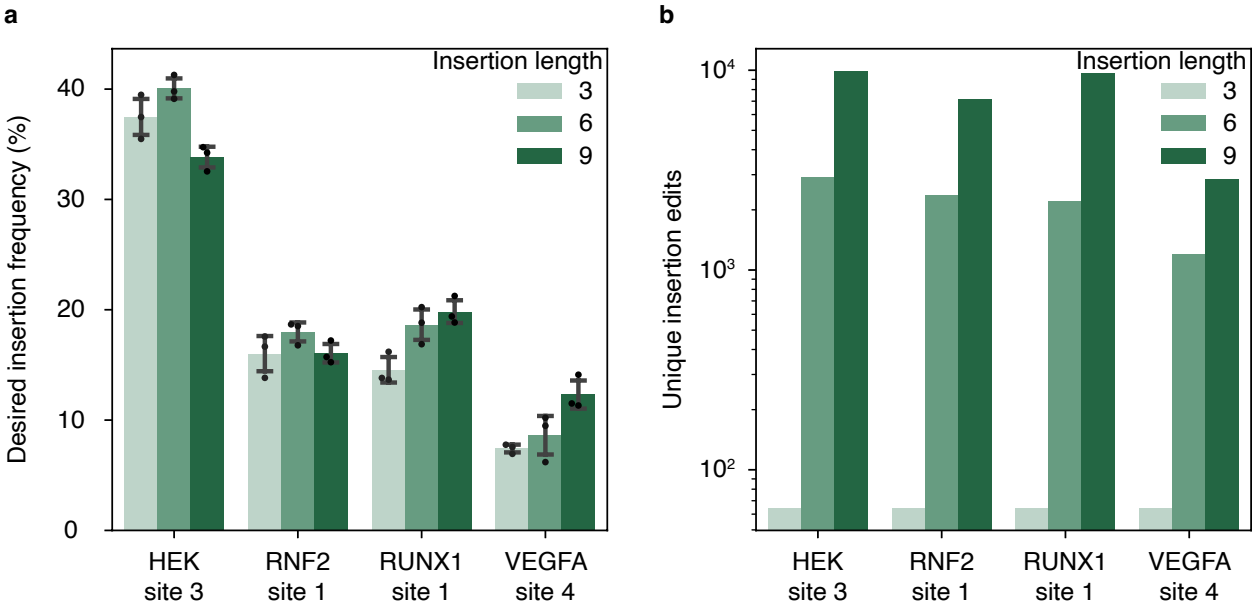
4.6.2 Supplementary Note 4.2: Top single, top composite, and 25th percentile pegRNA designs

The top single pegRNA design refers to the individual PBS-RTT length combination barcode that was observed with highest frequency in the next-generation sequencing data. The top composite pegRNA design utilizes averaged information from an individual pegRNA optimization profile. The composite PBS length is chosen by averaging all values across different RTT lengths for a given PBS length for all the different PBS lengths, and then choosing the PBS length with the highest value. The composite RTT length is chosen by averaging all values across different PBS lengths for a given RTT length for all the different RTT lengths, and then choosing the RTT length with the highest value. The 25th percentile pegRNA designs refer to designs ranked at the 25th percentile within the heatmap.

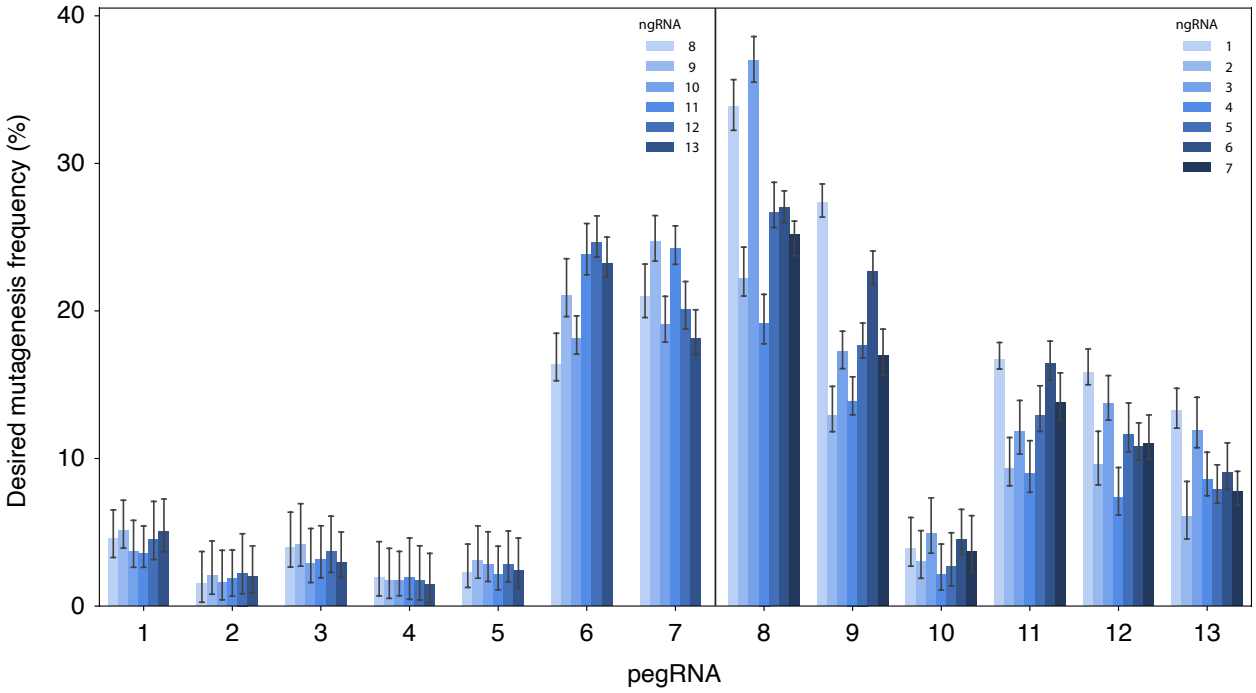
4.7 Supplementary figures



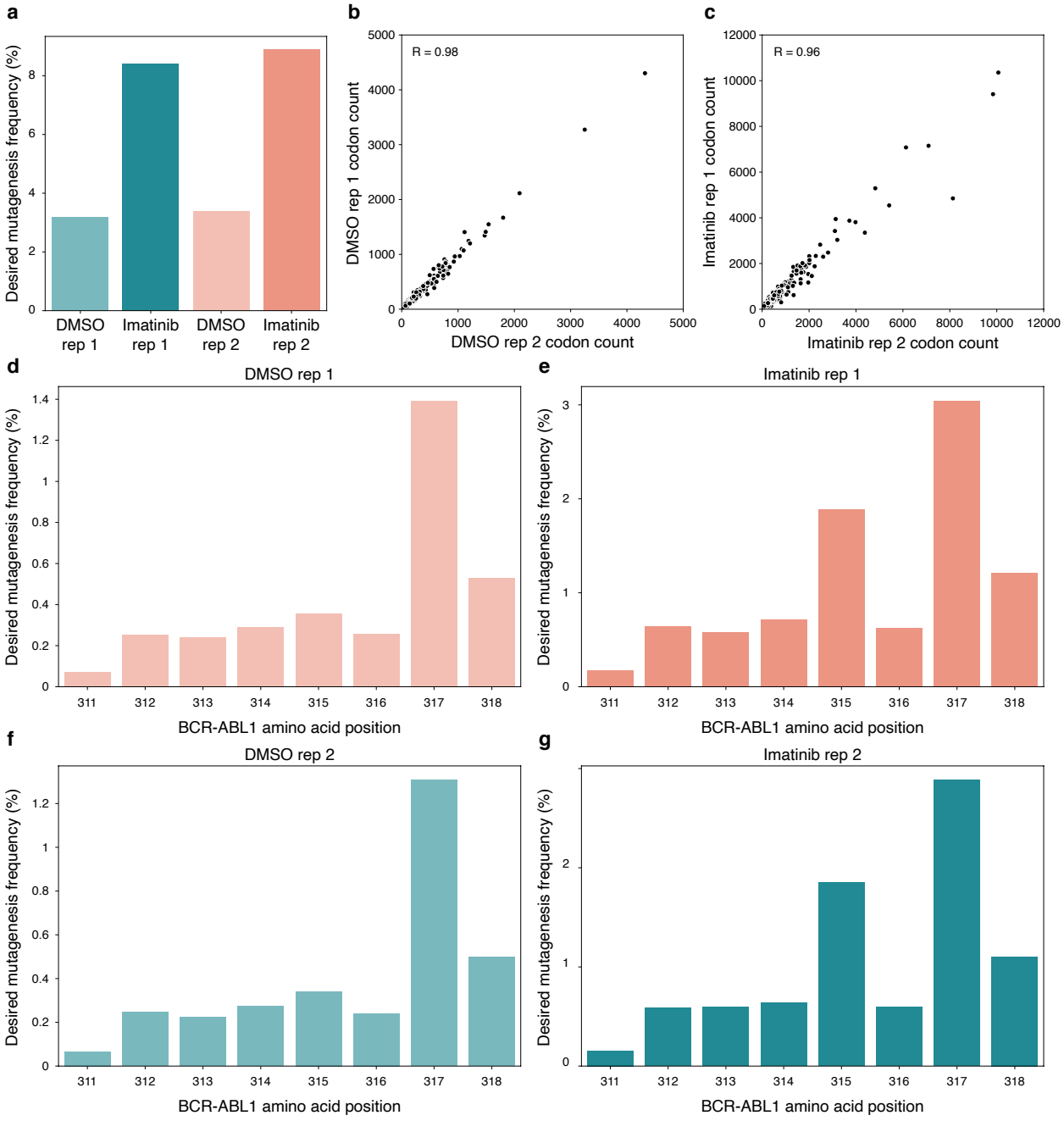
Supplementary Figure 4.1: Saturation mutagenesis efficiencies for single-base and triple-base substitutions. Single-base (**a,c,e**) and triple-base (**b,d,f**) saturation mutagenesis efficiencies across RNF2 site 1, RUNX1 site 1, and VEGFA site 4. Mean ± s.d. of n = 3 independent biological replicates. Design of pegRNAs are available in **Supplementary Table 4.1**.



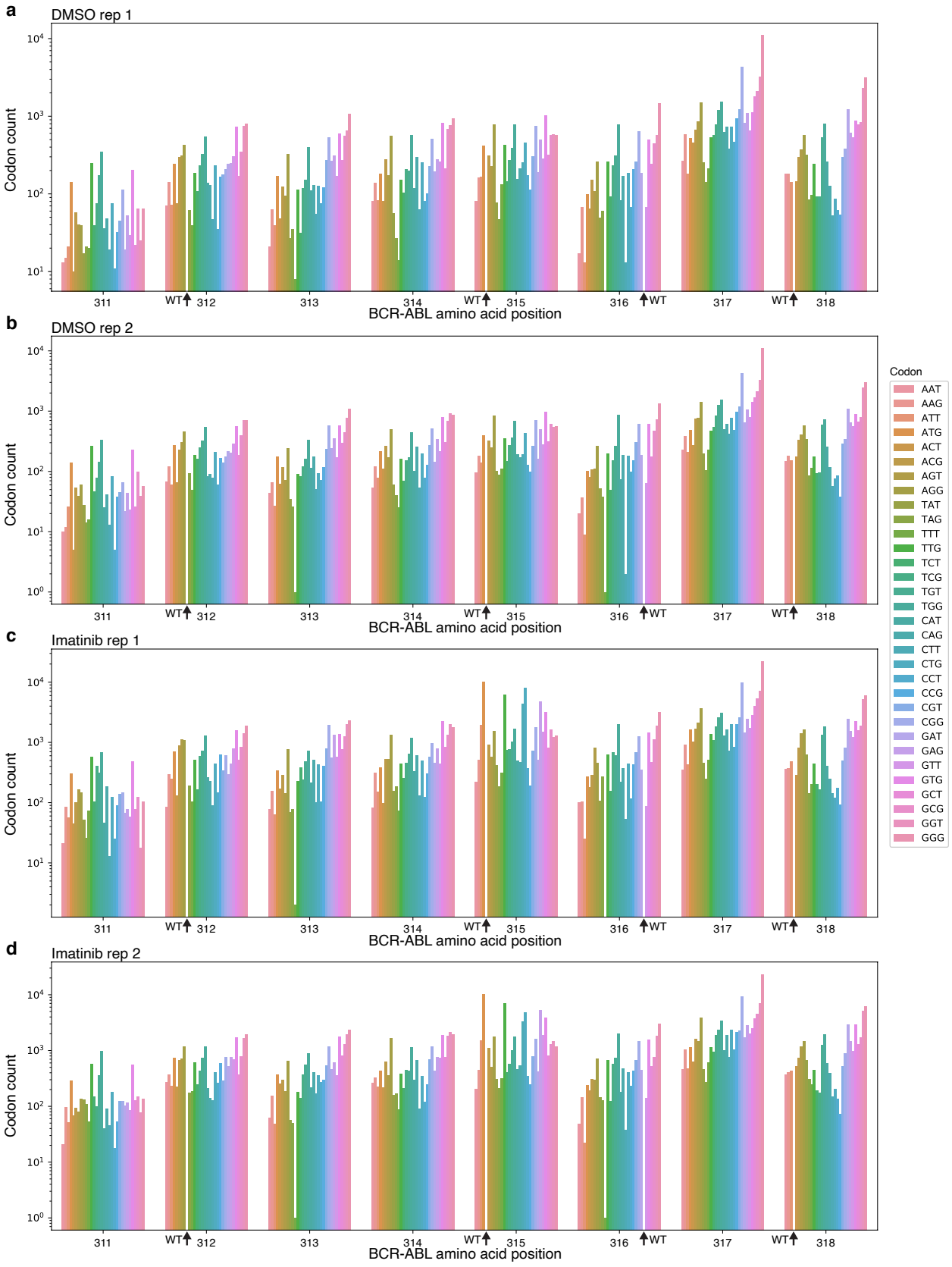
Supplementary Figure 4.2: Installation of randomized insertion edits at genomic targets. Editing efficiencies **(a)** and the count of unique insertion edits **(b)** for randomized insertion edits of various lengths (3, 6, and 9 bp) at HEK site 3, RNF2 site 1, RUNX1 site 1, and VEGFA site 4. Mean ± s.d. of n = 3 independent biological replicates. Design of pegRNAs are available in **Supplementary Table 4.1**.



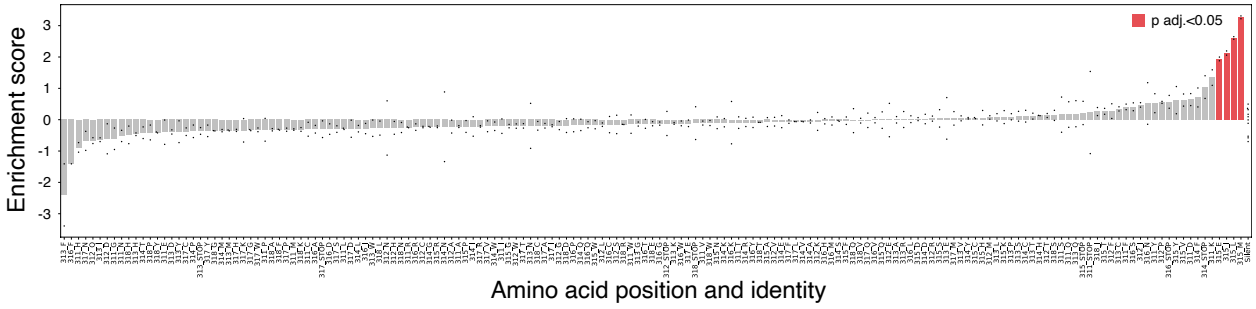
Supplementary Figure 4.3: Screening of pegRNA-ngRNA combinations at the *BCR-ABL1* locus in HEK293T cells. Screening of 84 pegRNA-ngRNA combinations to install a randomized (NNN) substitution edit at amino acid position 315 in *BCR-ABL1*. Experiments were performed in HEK293T cells. Mean ± s.d. of n = 3 independent biological replicates. Design of pegRNAs are available in **Supplementary Table 4.2**.



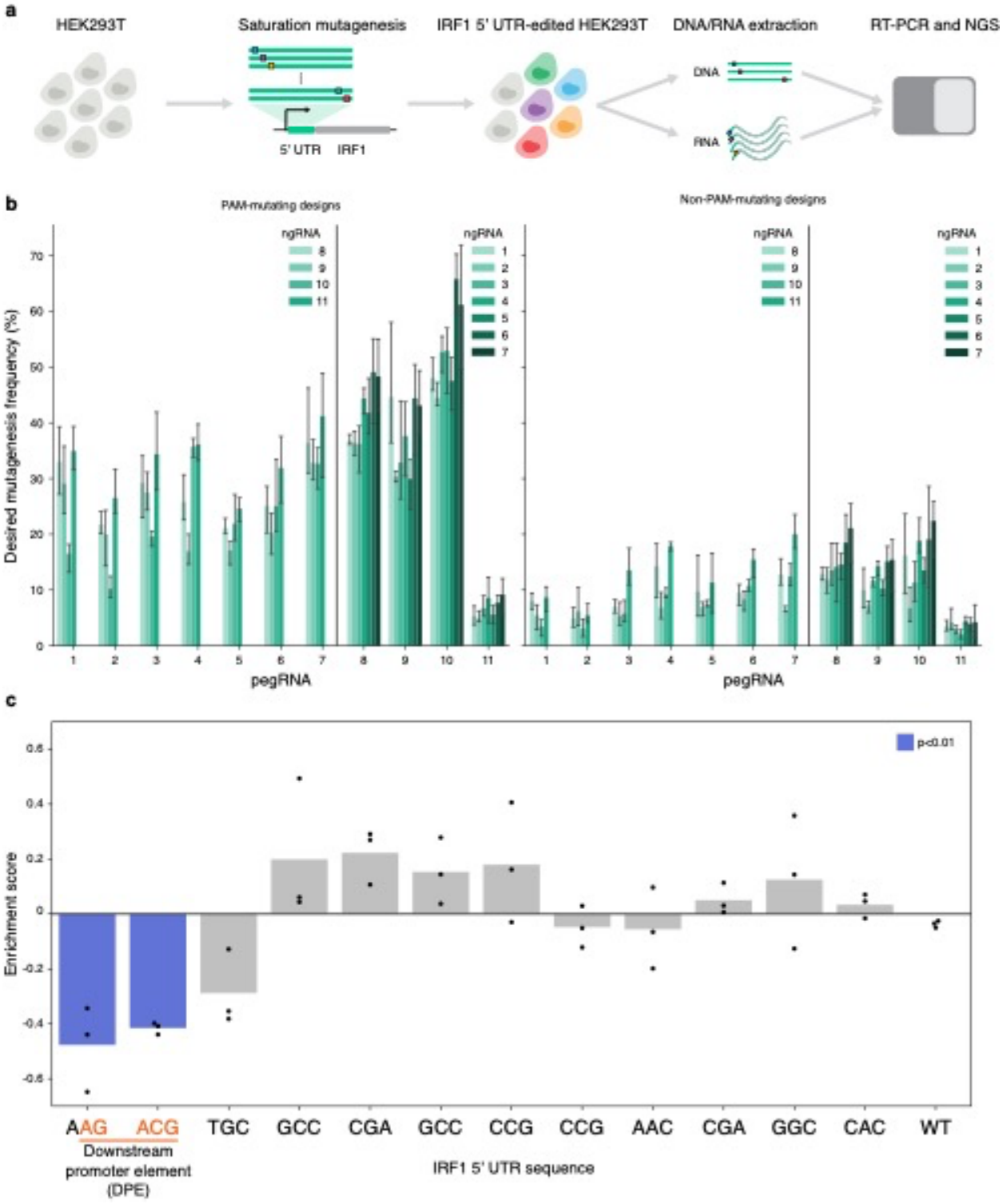
Supplementary Figure 4.4: Overview of saturation mutagenesis efficiencies at BCR-ABL1 with and without imatinib drug treatment in K562 cells. (a) Saturation mutagenesis efficiencies at BCR-ABL1 in K562 cells following 7 days of imatinib or DMSO treatment. Correlation (Spearman's Rho) between **(b)** DMSO and **(c)** imatinib treated samples for the count of variant NNK codons across amino acid positions 311-318 at BCR-ABL1. Saturation mutagenesis efficiencies by amino acid position for **(d,f)** DMSO and **(e,g)** imatinib treated samples.


a
DMSO rep 1
b
DMSO rep 2
c
Imatinib rep 1
d
Imatinib rep 2
Codon count
BCR-ABL amino acid position
311
WT
312
313
314
315
316
317
318
Codon
AAT
AAG
ATT
ATG
ACT
ACG
AGT
AGG
TAT
TAG
TTT
TTG
TCT
TCG
TGT
TGG
CAT
CAG
CTT
CTG
CCT
CCG
CGT
CGG
GAT
GAG
GTT
GTG
GCT
GCG
GGT
GGG

Supplementary Figure 4.5: Count of NNK codons across saturation mutagenesis window at *BCR-ABL1* locus. Count of installed NNK codons across amino acid positions 311-318 within *BCR-ABL1* for **(a,b)** DMSO and **(c,d)** imatinib treated samples. Certain NNK codons at specific amino acid positions are labeled as WT because they are the reference sequence and therefore aren't counted as installed edits.

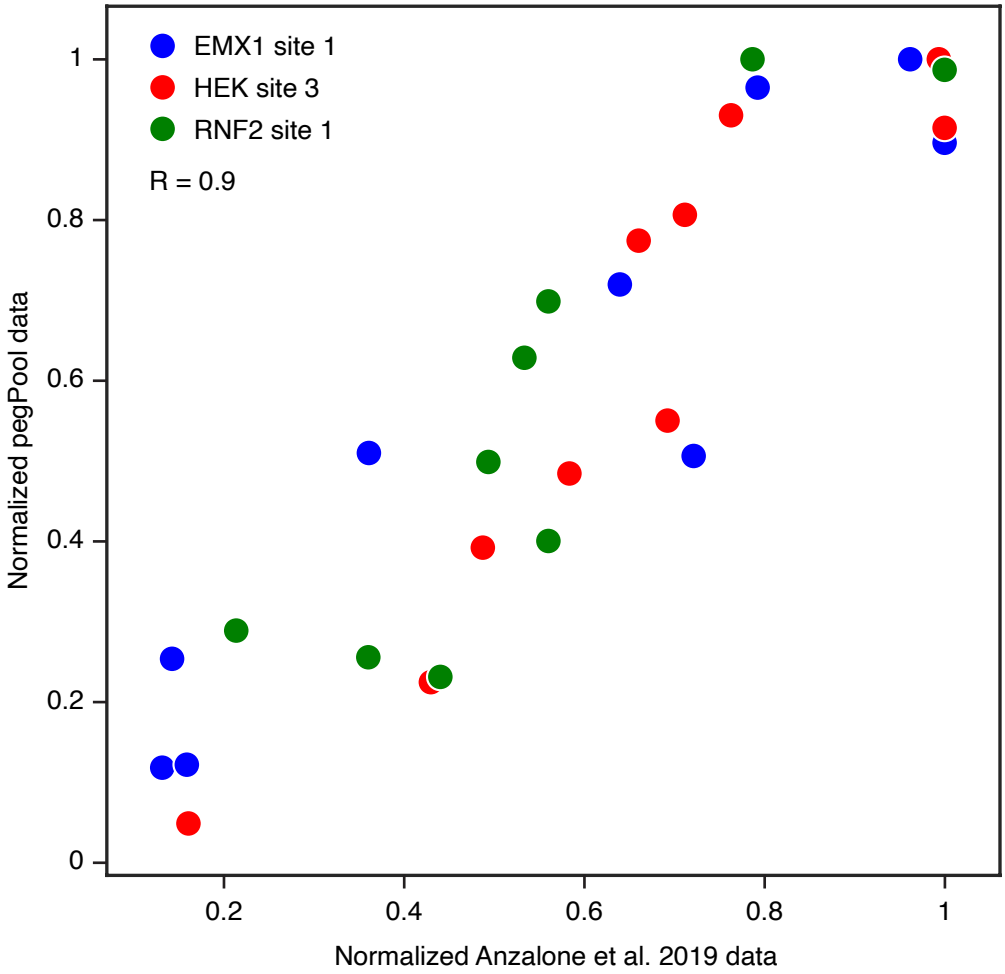


Supplementary Figure 4.6: Enrichment scores for amino acid variants from the *BCR-ABL1* imatinib resistance screen in K562 cells. Enrichment scores for all amino acid variants within residue positions 311-318 in *BCR-ABL1* associated with imatinib resistance in K562 cells. Statistical significance was determined by a t-test comparing any given amino acid variant against a null distribution consisting of NNK codons coding for silent mutations within the BCR-ABL1 protein, followed by multiple testing correction using the Bonferroni method. Statistics are available in **Supplementary Table 4.3**.

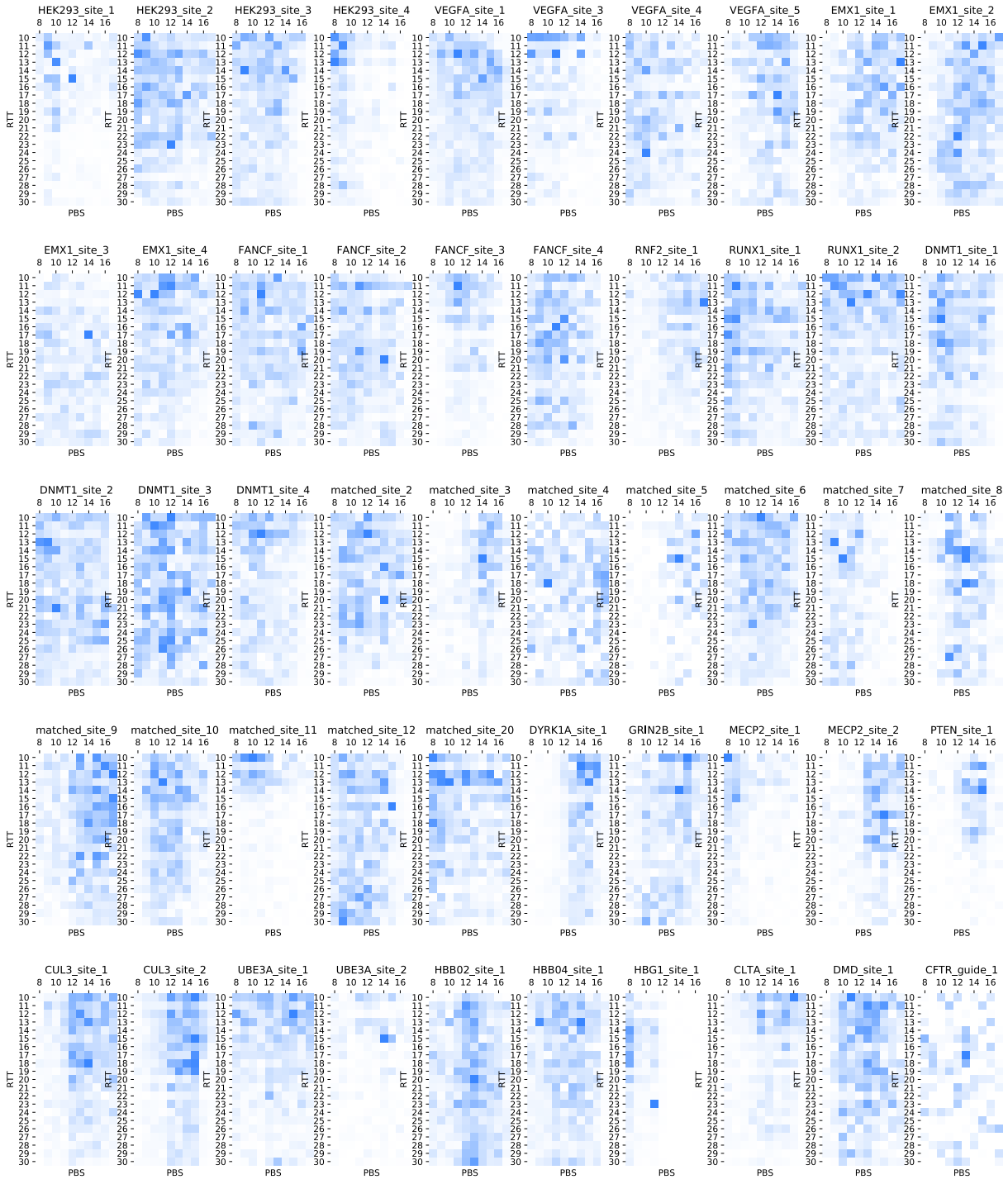


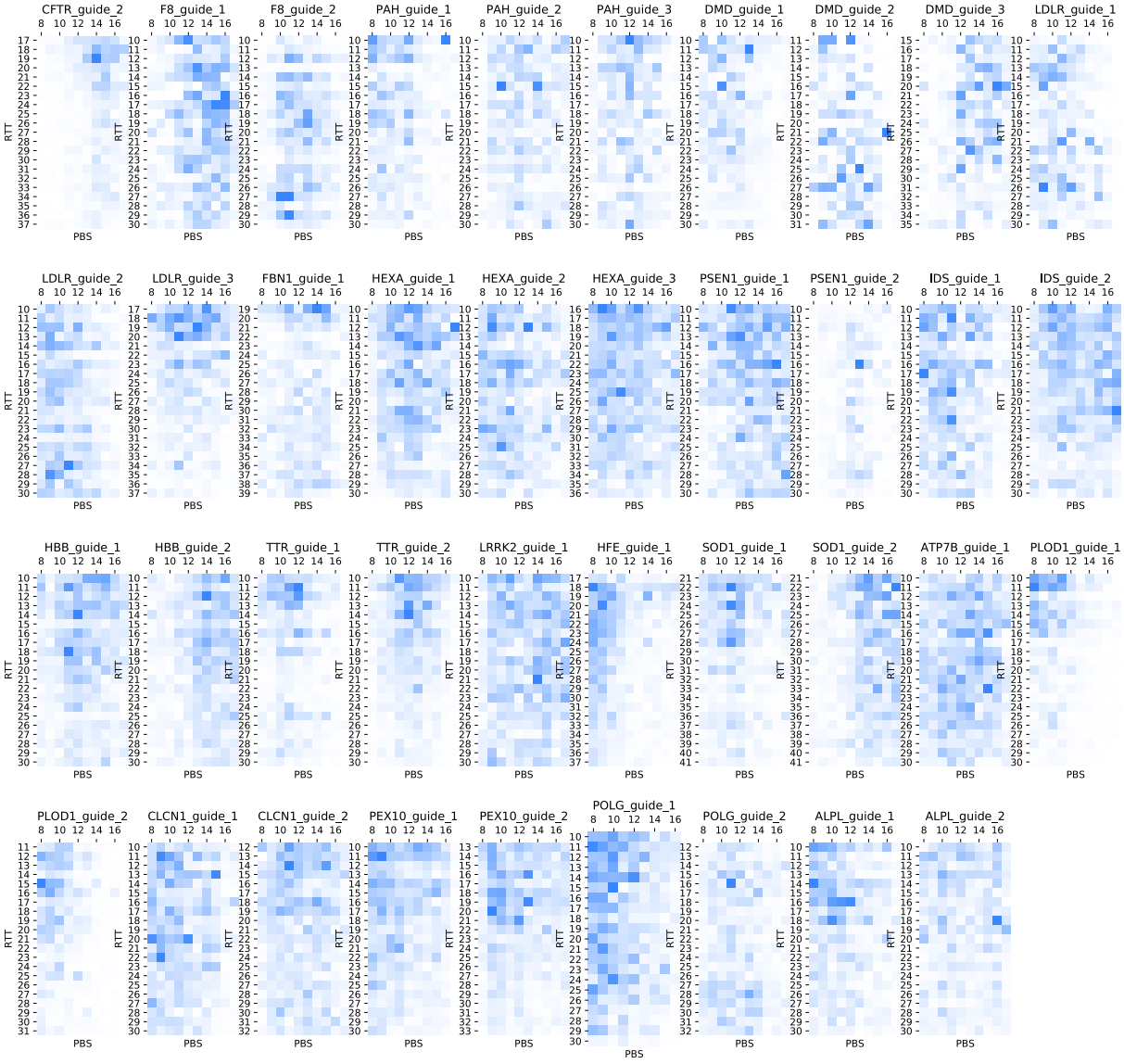
Supplementary Figure 4.7: Application of MOSAIC for the fine-mapping of non-coding regulatory sequences within the *IRF1* 5' UTR. (a) Schematic for *in situ* saturation mutagenesis of the IRF1 5' UTR to discover non-coding regulatory sequence elements involved in its transcription regulation in HEK293T cells. **(b)** Screening of 112 pegRNA-ngRNA

combinations to install randomized substitution edits across the *IRF1* 5' UTR. **(c)** Enrichment scores of 3 bp tiles across the IRF1 5' UTR sequence associated with *IRF1* transcription (mRNA vs. gDNA). Statistical significance was determined by a t-test comparing any given 3-bp mutagenized tile against a null distribution consisting of the reference *IRF1* 5'UTR sequence, followed by multiple testing correction using the Bonferroni method. Mean ± s.d. of n = 3 independent biological replicates. Design of pegRNAs are available in **Supplementary Table 4.4**.

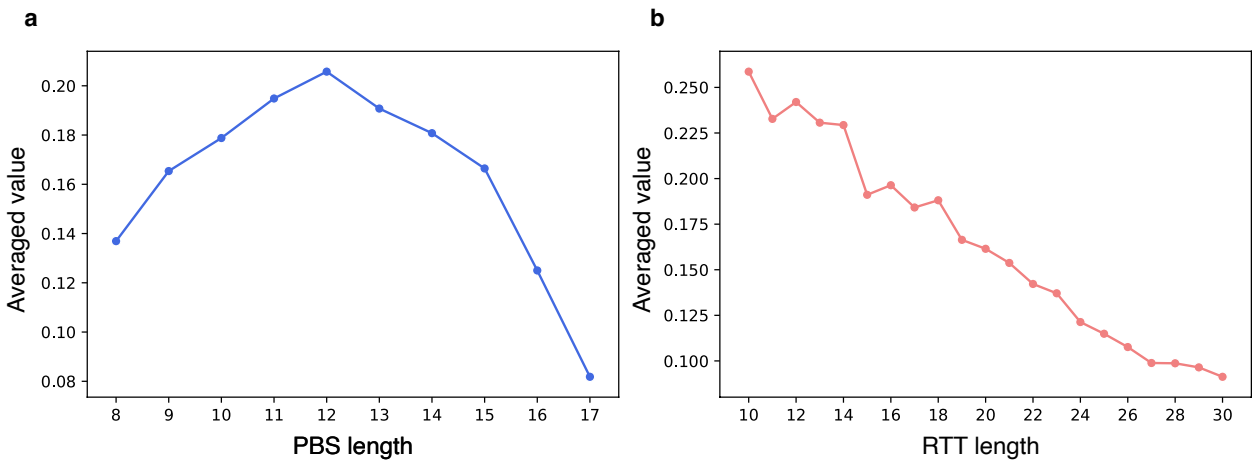


Supplementary Figure 4.8: Correlation between pooled pegRNA optimizations and individual pegRNA optimizations performed in Anzalone et al. 2019. Correlation (Spearman's Rho) between pooled pegRNA optimizations with MOSAIC and individual pegRNA optimizations performed in Anzalone et al. 2019 for EMX1 site 1, HEK site 3, and RNF2 site 1 genomic targets.


HEK293_site_1
HEK293_site_2
HEK293_site_3
HEK293_site_4
VEGFA_site_1
VEGFA_site_3
VEGFA_site_4
VEGFA_site_5
EMX1_site_1
EMX1_site_2
EMX1_site_3
EMX1_site_4
FANCF_site_1
FANCF_site_2
FANCF_site_3
FANCF_site_4
RNF2_site_1
RUNX1_site_1
RUNX1_site_2
DNMT1_site_1
DNMT1_site_2
DNMT1_site_3
DNMT1_site_4
matched_site_2
matched_site_3
matched_site_4
matched_site_5
matched_site_6
matched_site_7
matched_site_8
matched_site_9
matched_site_10
matched_site_11
matched_site_12
matched_site_20
DYRK1A_site_1
GRIN2B_site_1
MECP2_site_1
MECP2_site_2
PTEN_site_1
CUL3_site_1
CUL3_site_2
UBE3A_site_1
UBE3A_site_2
HBB02_site_1
HBB04_site_1
HBG1_site_1
CLTA_site_1
DMD_site_1
CFTR_guide_1
RTT
PBS



Supplementary Figure 4.9: Individual pegRNA optimization profiles across 89 spacer sequences. Individual pegRNA optimization profiles across 89 spacer sequences spanning PBS lengths of 8-17 nt and RTT lengths ranging from 10-41 nt. Each pegRNA optimization profile consists of 210 pegRNA designs.



Supplementary Figure 4.10: Averaged PBS and RTT profiles for pegRNA design. The averaged **(a)** PBS and **(b)** RTT profiles from pegRNA optimization profiles consisting of PBS lengths of 8-17 nt and RTT lengths of 10-30 nt.

4.8 Supplementary tables

Supplementary Table 4.1: Oligonucleotide sequences to construct PCR pegRNAs and ngRNAs for saturation mutagenesis and randomized insertions at HEK site 3, RNF2 site 1, RUNX1 site 1, and VEGFA site 4.

Oligo sequence	Description
ATATCTTGTGGAAAGGACGAAACACCGGCCCAGACTGAGCACGTGA GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	HEK site 3 pegRNA spacer oligo
GGCACCGAGTCGGTGCTGGAHGAAGCAGGGCTTCCTTTCCTCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 1
GGCACCGAGTCGGTGCTGGAGHAAGCAGGGCTTCCTTTCCTCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 2
GGCACCGAGTCGGTGCTGGAGGBAGCAGGGCTTCCTTTCCTCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 3
GGCACCGAGTCGGTGCTGGAGGABGCAGGGCTTCCTTTCCTCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 4
GGCACCGAGTCGGTGCTGGAGGAAHCAGGGCTTCCTTTCCTCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 5
GGCACCGAGTCGGTGCTGGAGGAAGDAGGGCTTCCTTTCCTCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 6
GGCACCGAGTCGGTGCTGGAGGAAGCBGGGCTTCCTTTCCTCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 7

GGCACCGAGTCGGTGCTGGAGGAAGCAHGGCTTCCTTTCCTCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 8
GGCACCGAGTCGGTGCTGGAGGAAGCAGHGCTTCCTTTCCTCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 9
GGCACCGAGTCGGTGCTGGAGGAAGCAGGHCTTCCTTTCCTCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 10
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGDTTCCTTTCCTCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 11
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCVTCCTTTCCTCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 12
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCTVCCTTTCCTCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 13
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCTTDCTTTCCTCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 14
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCTTCDTTTCCTCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 15
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCTTCCVTTCCTCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 16
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCTTCCTVTCCTCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 17
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCTTCCTTVCCTCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 18
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCTTCCTTTDCTCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 19
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCTTCCTTTCDTCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 20
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCTTCCTTTCCVCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 21
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCTTCCTTTCCTDTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 22
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCTTCCTTTCCTCVGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 23
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCTTCCTTTCCTCTHCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 24
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCTTCCTTTCCTCTGDC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 25

GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCTTCCTTTCCTCTGCD ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 26
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCTTCCTTTCCTCTGCC BTCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 27
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCTTCCTTTCCTCTGCC AVCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 28
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCTTCCTTTCCTCTGCC ATDACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 29
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCTTCCTTTCCTCTGCC ATCBCGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 1 bp Tile 30
GGCACCGAGTCGGTGCTGGANNNAGCAGGGCTTCCTTTCCTCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 3 bp Tile 1
GGCACCGAGTCGGTGCTGGAGGANNNAGGGCTTCCTTTCCTCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 3 bp Tile 2
GGCACCGAGTCGGTGCTGGAGGAAGCNNNGCTTCCTTTCCTCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 3 bp Tile 3
GGCACCGAGTCGGTGCTGGAGGAAGCAGGNNNTCCTTTCCTCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 3 bp Tile 4
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCTNNNTTTCCTCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 3 bp Tile 5
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCTTCCNNNCCTCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 3 bp Tile 6
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCTTCCTTTNNNCTGCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 3 bp Tile 7
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCTTCCTTTCCTNNNCC ATCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 3 bp Tile 8
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCTTCCTTTCCTCTGNN NTCACGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 3 bp Tile 9
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCTTCCTTTCCTCTGCC ANNNCGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 3 bp Tile 10
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCTTCCTTTCCTCTGCC ATCANNNCGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 3 bp insertion
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCTTCCTTTCCTCTGCC ATCANNNNNNCGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 6 bp insertion
GGCACCGAGTCGGTGCTGGAGGAAGCAGGGCTTCCTTTCCTCTGCC ATCANNNNNNNNNCGTGCTCAGTCTGTTTTTTTGGTGTATCAC	HEK site 3 extension oligo 9 bp insertion

ATATCTTGTGGAAAGGACGAAACACCGTCAACCAGTATCCCGGTGC GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	HEK site 3 ngRNA spacer oligo
ATATCTTGTGGAAAGGACGAAACACCGTCATCTTAGTCATTACCTG GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	RNF2 site 1 pegRNA spacer oligo
AAGTGGCACCGAGTCGGTGCAADGAACACCTCAGGTAATGACTAAG ATGTTTTTTTGGTGTATCACGTGCTGC	RNF2 site 1 extension oligo 1 bp Tile 1
AAGTGGCACCGAGTCGGTGCAACHAACACCTCAGGTAATGACTAAG ATGTTTTTTTGGTGTATCACGTGCTGC	RNF2 site 1 extension oligo 1 bp Tile 2
AAGTGGCACCGAGTCGGTGCAACGBACACCTCAGGTAATGACTAAG ATGTTTTTTTGGTGTATCACGTGCTGC	RNF2 site 1 extension oligo 1 bp Tile 3
AAGTGGCACCGAGTCGGTGCAACGABCACCTCAGGTAATGACTAAG ATGTTTTTTTGGTGTATCACGTGCTGC	RNF2 site 1 extension oligo 1 bp Tile 4
AAGTGGCACCGAGTCGGTGCAACGAADACCTCAGGTAATGACTAAG ATGTTTTTTTGGTGTATCACGTGCTGC	RNF2 site 1 extension oligo 1 bp Tile 5
AAGTGGCACCGAGTCGGTGCAACGAACBCCTCAGGTAATGACTAAG ATGTTTTTTTGGTGTATCACGTGCTGC	RNF2 site 1 extension oligo 1 bp Tile 6
AAGTGGCACCGAGTCGGTGCAACGAACADCTCAGGTAATGACTAAG ATGTTTTTTTGGTGTATCACGTGCTGC	RNF2 site 1 extension oligo 1 bp Tile 7
AAGTGGCACCGAGTCGGTGCAACGAACACDTCAGGTAATGACTAAG ATGTTTTTTTGGTGTATCACGTGCTGC	RNF2 site 1 extension oligo 1 bp Tile 8
AAGTGGCACCGAGTCGGTGCAACGAACACCVCAGGTAATGACTAA GATGTTTTTTTGGTGTATCACGTGCTGC	RNF2 site 1 extension oligo 1 bp Tile 9
AAGTGGCACCGAGTCGGTGCAACGAACACCTDAGGTAATGACTAAG ATGTTTTTTTGGTGTATCACGTGCTGC	RNF2 site 1 extension oligo 1 bp Tile 10
AAGTGGCACCGAGTCGGTGCAACGAACACCTCBGGTAATGACTAAG ATGTTTTTTTGGTGTATCACGTGCTGC	RNF2 site 1 extension oligo 1 bp Tile 11
AAGTGGCACCGAGTCGGTGCAACGAACACCTCAHGTAATGACTAAG ATGTTTTTTTGGTGTATCACGTGCTGC	RNF2 site 1 extension oligo 1 bp Tile 12
AAGTGGCACCGAGTCGGTGCAANNNACACCTCAGGTAATGACTAAG ATGTTTTTTTGGTGTATCACGTGCTGC	RNF2 site 1 extension oligo 3 bp Tile 1
AAGTGGCACCGAGTCGGTGCAACGANNNCCTCAGGTAATGACTAAG ATGTTTTTTTGGTGTATCACGTGCTGC	RNF2 site 1 extension oligo 3 bp Tile 2
AAGTGGCACCGAGTCGGTGCAACGAACANNNCAGGTAATGACTAA GATGTTTTTTTGGTGTATCACGTGCTGC	RNF2 site 1 extension oligo 3 bp Tile 3
AAGTGGCACCGAGTCGGTGCAACGAACACCTNNNGTAATGACTAAG ATGTTTTTTTGGTGTATCACGTGCTGC	RNF2 site 1 extension oligo 3 bp Tile 4

AAGTGGCACCGAGTCGGTGCAACGAACACCTCAGNNNGTAATGACTAAGATGTTTTTTTGGTGTATCACGTGCTGC	RNF2 site 1 extension oligo 3 bp insertion
AAGTGGCACCGAGTCGGTGCAACGAACACCTCAGNNNNNNGTAATGACTAAGATGTTTTTTTGGTGTATCACGTGCTGC	RNF2 site 1 extension oligo 6 bp insertion
AAGTGGCACCGAGTCGGTGCAACGAACACCTCAGNNNNNNNNNGTAATGACTAAGATGTTTTTTTGGTGTATCACGTGCTGC	RNF2 site 1 extension oligo 9 bp insertion
ATATCTTGTGGAAAGGACGAAACACCGTCAACCATTAAGCAAAACATGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	RNF2 site 1 ngRNA spacer oligo
ATATCTTGTGGAAAGGACGAAACACCGCATTTTCAGGAGGAAGCGAGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	RUNX1 site 1 pegRNA spacer oligo
AAGTGGCACCGAGTCGGTGCTGTDTGAAGCCATCGCTTCCTCCTGAAAATTTTTTTTGGTGTATCACGTGCTGC	RUNX1 site 1 extension oligo 1 bp Tile 1
AAGTGGCACCGAGTCGGTGCTGTCVGAAGCCATCGCTTCCTCCTGAAAATTTTTTTTGGTGTATCACGTGCTGC	RUNX1 site 1 extension oligo 1 bp Tile 2
AAGTGGCACCGAGTCGGTGCTGTCTHAAGCCATCGCTTCCTCCTGAAAATTTTTTTTGGTGTATCACGTGCTGC	RUNX1 site 1 extension oligo 1 bp Tile 3
AAGTGGCACCGAGTCGGTGCTGTCTGBAGCCATCGCTTCCTCCTGAAAATTTTTTTTGGTGTATCACGTGCTGC	RUNX1 site 1 extension oligo 1 bp Tile 4
AAGTGGCACCGAGTCGGTGCTGTCTGABGCCATCGCTTCCTCCTGAAAATTTTTTTTGGTGTATCACGTGCTGC	RUNX1 site 1 extension oligo 1 bp Tile 5
AAGTGGCACCGAGTCGGTGCTGTCTGAAHCCATCGCTTCCTCCTGAAAATTTTTTTTGGTGTATCACGTGCTGC	RUNX1 site 1 extension oligo 1 bp Tile 6
AAGTGGCACCGAGTCGGTGCTGTCTGAAGDCATCGCTTCCTCCTGAAAATTTTTTTTGGTGTATCACGTGCTGC	RUNX1 site 1 extension oligo 1 bp Tile 7
AAGTGGCACCGAGTCGGTGCTGTCTGAAGCDATCGCTTCCTCCTGAAAATTTTTTTTGGTGTATCACGTGCTGC	RUNX1 site 1 extension oligo 1 bp Tile 8
AAGTGGCACCGAGTCGGTGCTGTCTGAAGCCBTCGCTTCCTCCTGAAAATTTTTTTTGGTGTATCACGTGCTGC	RUNX1 site 1 extension oligo 1 bp Tile 9

AAGTGGCACCGAGTCGGTGCTGTCTGAAGCCAVCGCTTCCTCCTGA AAATTTTTTTTGGTGTATCACGTGCTGC	RUNX1 site 1 extension oligo 1 bp Tile 10
AAGTGGCACCGAGTCGGTGCTGTCTGAAGCCATDGCTTCCTCCTGA AAATTTTTTTTGGTGTATCACGTGCTGC	RUNX1 site 1 extension oligo 1 bp Tile 11
AAGTGGCACCGAGTCGGTGCTGTCTGAAGCCATCHCTTCCTCCTGA AAATTTTTTTTGGTGTATCACGTGCTGC	RUNX1 site 1 extension oligo 1 bp Tile 12
AAGTGGCACCGAGTCGGTGCTGTNNNAAGCCATCGCTTCCTCCTGA AAATTTTTTTTGGTGTATCACGTGCTGC	RUNX1 site 1 extension oligo 3 bp Tile 1
AAGTGGCACCGAGTCGGTGCTGTCTGNNNCCATCGCTTCCTCCTGA AAATTTTTTTTGGTGTATCACGTGCTGC	RUNX1 site 1 extension oligo 3 bp Tile 2
AAGTGGCACCGAGTCGGTGCTGTCTGAAGNNNTCGCTTCCTCCTGA AAATTTTTTTTGGTGTATCACGTGCTGC	RUNX1 site 1 extension oligo 3 bp Tile 3
AAGTGGCACCGAGTCGGTGCTGTCTGAAGCCANNNCTTCCTCCTGA AAATTTTTTTTGGTGTATCACGTGCTGC	RUNX1 site 1 extension oligo 3 bp Tile 4
AAGTGGCACCGAGTCGGTGCTGTCTGAAGCCATCGNNNCTTCCTCC TGAAAATTTTTTTTGGTGTATCACGTGCTGC	RUNX1 site 1 extension oligo 3 bp insertion
AAGTGGCACCGAGTCGGTGCTGTCTGAAGCCATCGNNNNNNCTTCC TCCTGAAAATTTTTTTTGGTGTATCACGTGCTGC	RUNX1 site 1 extension oligo 6 bp insertion
AAGTGGCACCGAGTCGGTGCTGTCTGAAGCCATCGNNNNNNNNNCT TCCTCCTGAAAATTTTTTTTGGTGTATCACGTGCTGC	RUNX1 site 1 extension oligo 9 bp insertion
ATATCTTGTGGAAAGGACGAAACACCGATGAAGCACTGTGGGTACG AGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	RUNX1 site 1 ngRNA spacer oligo
ATATCTTGTGGAAAGGACGAAACACCGATGTCTGCAGGCCAGATGA GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	VEGFA site 4 pegRNA spacer oligo
AAGTGGCACCGAGTCGGTGCAATGVGCCATCTGGAGCCCTCATCTG GCCTGCAGATTTTTTTGGTGTATCACGTGCTGC	VEGFA site 4 extension oligo 1 bp Tile 1

AAGTGGCACCGAGTCGGTGCAATGTHCCATCTGGAGCCCTCATCTG GCCTGCAGATTTTTTTGGTGTATCACGTGCTGC	VEGFA site 4 extension oligo 1 bp Tile 2
AAGTGGCACCGAGTCGGTGCAATGTGDCATCTGGAGCCCTCATCTG GCCTGCAGATTTTTTTGGTGTATCACGTGCTGC	VEGFA site 4 extension oligo 1 bp Tile 3
AAGTGGCACCGAGTCGGTGCAATGTGCDATCTGGAGCCCTCATCTG GCCTGCAGATTTTTTTGGTGTATCACGTGCTGC	VEGFA site 4 extension oligo 1 bp Tile 4
AAGTGGCACCGAGTCGGTGCAATGTGCCBTCTGGAGCCCTCATCTG GCCTGCAGATTTTTTTGGTGTATCACGTGCTGC	VEGFA site 4 extension oligo 1 bp Tile 5
AAGTGGCACCGAGTCGGTGCAATGTGCCAVCTGGAGCCCTCATCTG GCCTGCAGATTTTTTTGGTGTATCACGTGCTGC	VEGFA site 4 extension oligo 1 bp Tile 6
AAGTGGCACCGAGTCGGTGCAATGTGCCATDTGGAGCCCTCATCTG GCCTGCAGATTTTTTTGGTGTATCACGTGCTGC	VEGFA site 4 extension oligo 1 bp Tile 7
AAGTGGCACCGAGTCGGTGCAATGTGCCATCVGGAGCCCTCATCTG GCCTGCAGATTTTTTTGGTGTATCACGTGCTGC	VEGFA site 4 extension oligo 1 bp Tile 8
AAGTGGCACCGAGTCGGTGCAATGTGCCATCTHGAGCCCTCATCTG GCCTGCAGATTTTTTTGGTGTATCACGTGCTGC	VEGFA site 4 extension oligo 1 bp Tile 9
AAGTGGCACCGAGTCGGTGCAATGTGCCATCTGHAGCCCTCATCTG GCCTGCAGATTTTTTTGGTGTATCACGTGCTGC	VEGFA site 4 extension oligo 1 bp Tile 10
AAGTGGCACCGAGTCGGTGCAATGTGCCATCTGGBGCCCTCATCTG GCCTGCAGATTTTTTTGGTGTATCACGTGCTGC	VEGFA site 4 extension oligo 1 bp Tile 11
AAGTGGCACCGAGTCGGTGCAATGTGCCATCTGGAHCCCTCATCTG GCCTGCAGATTTTTTTGGTGTATCACGTGCTGC	VEGFA site 4 extension oligo 1 bp Tile 12
AAGTGGCACCGAGTCGGTGCAATGTGCCATCTGGAGDCCTCATCTG GCCTGCAGATTTTTTTGGTGTATCACGTGCTGC	VEGFA site 4 extension oligo 1 bp Tile 13

AAGTGGCACCGAGTCGGTGCAATGTGCCATCTGGAGCDCTCATCTG GCCTGCAGATTTTTTTGGTGTATCACGTGCTGC	VEGFA site 4 extension oligo 1 bp Tile 14
AAGTGGCACCGAGTCGGTGCAATGTGCCATCTGGAGCCDTCATCTG GCCTGCAGATTTTTTTGGTGTATCACGTGCTGC	VEGFA site 4 extension oligo 1 bp Tile 15
AAGTGGCACCGAGTCGGTGCAATGTGCCATCTGGAGCCCVCATCTG GCCTGCAGATTTTTTTGGTGTATCACGTGCTGC	VEGFA site 4 extension oligo 1 bp Tile 16
AAGTGGCACCGAGTCGGTGCAATGTGCCATCTGGAGCCCTDATCTG GCCTGCAGATTTTTTTGGTGTATCACGTGCTGC	VEGFA site 4 extension oligo 1 bp Tile 17
AAGTGGCACCGAGTCGGTGCAATGTGCCATCTGGAGCCCTCBTCTG GCCTGCAGATTTTTTTGGTGTATCACGTGCTGC	VEGFA site 4 extension oligo 1 bp Tile 18
AAGTGGCACCGAGTCGGTGCAATGNNNCATCTGGAGCCCTCATCTG GCCTGCAGATTTTTTTGGTGTATCACGTGCTGC	VEGFA site 4 extension oligo 3 bp Tile 1
AAGTGGCACCGAGTCGGTGCAATGTGCNNNCTGGAGCCCTCATCTG GCCTGCAGATTTTTTTGGTGTATCACGTGCTGC	VEGFA site 4 extension oligo 3 bp Tile 2
AAGTGGCACCGAGTCGGTGCAATGTGCCATNNNGAGCCCTCATCTG GCCTGCAGATTTTTTTGGTGTATCACGTGCTGC	VEGFA site 4 extension oligo 3 bp Tile 3
AAGTGGCACCGAGTCGGTGCAATGTGCCATCTGNNNCCCTCATCTG GCCTGCAGATTTTTTTGGTGTATCACGTGCTGC	VEGFA site 4 extension oligo 3 bp Tile 4
AAGTGGCACCGAGTCGGTGCAATGTGCCATCTGGAGNNNTCATCTG GCCTGCAGATTTTTTTGGTGTATCACGTGCTGC	VEGFA site 4 extension oligo 3 bp Tile 5
AAGTGGCACCGAGTCGGTGCAATGTGCCATCTGGAGCCCNNNTCTG GCCTGCAGATTTTTTTGGTGTATCACGTGCTGC	VEGFA site 4 extension oligo 3 bp Tile 6
AAGTGGCACCGAGTCGGTGCAATGTGCCATCTGGAGCCCTCANNNT CTGGCCTGCAGATTTTTTTGGTGTATCACGTGCTGC	VEGFA site 4 extension oligo 3 bp insertion

AAGTGGCACCGAGTCGGTGCAATGTGCCATCTGGAGCCCTCANNNN NNTCTGGCCTGCAGATTTTTTTGGTGTATCACGTGCTGC	VEGFA site 4 extension oligo 6 bp insertion
AAGTGGCACCGAGTCGGTGCAATGTGCCATCTGGAGCCCTCANNNN NNNNNTCTGGCCTGCAGATTTTTTTGGTGTATCACGTGCTGC	VEGFA site 4 extension oligo 9 bp insertion
ATATCTTGTGGAAAGGACGAAACACCGATGTACAGAGAGCCCAGG GCGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	VEGFA site 4 ngRNA spacer oligo

Supplementary Table 4.2: Oligonucleotide sequences to construct PCR pegRNAs and ngRNAs for saturation mutagenesis at the BCR-ABL1 imatinib binding site.

Oligo sequence	**Description**
TATATCTTGTGGAAAGGACGAAACACCGTTGTTTGTTCAGT TGGGAGGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	BCR-ABL1 peg1/ng1 spacer oligo
ATATCTTGTGGAAAGGACGAAACACCGTGAAGTCCTCGTTG TCTTGTGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	BCR-ABL1 peg2/ng2 spacer oligo
TATATCTTGTGGAAAGGACGAAACACCGTCCTCGTTGTCTT GTTGGCGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	BCR-ABL1 peg3/ng3 spacer oligo
ATATCTTGTGGAAAGGACGAAACACCGTCCTCGTTGTCTTG TTGGCAGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	BCR-ABL1 peg4/ng4 spacer oligo
ATATCTTGTGGAAAGGACGAAACACCGCCTCGTTGTCTTGT TGGCAGGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	BCR-ABL1 peg5/ng5 spacer oligo
ATATCTTGTGGAAAGGACGAAACACCGTGTTGGCAGGGGTC TGCACCGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	BCR-ABL1 peg6/ng6 spacer oligo
TATATCTTGTGGAAAGGACGAAACACCGTTGGCAGGGGTCT GCACCCGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	BCR-ABL1 peg7/ng7 spacer oligo
TATATCTTGTGGAAAGGACGAAACACCGTAGTCCAGGAGGT TCCCGTGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	BCR-ABL1 peg8/ng8 spacer oligo
ATATCTTGTGGAAAGGACGAAACACCGACTCCCTCAGGTAG TCCAGGGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	BCR-ABL1 peg9/ng9 spacer oligo
ATATCTTGTGGAAAGGACGAAACACCGTGCACTCCCTCAGG TAGTCCGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	BCR-ABL1 peg10/ng10 spacer oligo
ATATCTTGTGGAAAGGACGAAACACCGCCTGCCGGTTGCAC TCCCTCGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	BCR-ABL1 peg11/ng11 spacer oligo
ATATCTTGTGGAAAGGACGAAACACCGCCACGGCGTTCACC TCCTGCGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	BCR-ABL1 peg12/ng12 spacer oligo

TATATCTTGTGGAAAGGACGAAACACCGGCCATGTACAGCA GCACCAGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	BCR-ABL1 peg13/ng13 spacer oligo
GGCACCGAGTCGGTGCTCCAGGAGGTTCCCGTAGGTCATGA ACTCNNNGATGATATAGAACGGGGGCTCCCGGGTGCAGA	BCR-ABL1 peg1 extension oligo 1
GGCACCGAGTCGGTGCTCCAGGAGGTTCCCGTAGGTCATGA ACTCNNNGATGATATAGAACGGGGGCTCCCGGGTGCAGA	BCR-ABL1 peg2 extension oligo 1
GGCACCGAGTCGGTGCTCCAGGAGGTTCCCGTAGGTCATGA ACTCNNNGATGATATAGAACGGGGGCTCCCGGGTGCAGA	BCR-ABL1 peg3 extension oligo 1
GGCACCGAGTCGGTGCTCCAGGAGGTTCCCGTAGGTCATGA ACTCNNNGATGATATAGAACGGGGGCTCCCGGGTGCAGA	BCR-ABL1 peg4 extension oligo 1
GGCACCGAGTCGGTGCTCCAGGAGGTTCCCGTAGGTCATGA ACTCNNNGATGATATAGAACGGGGGCTCCCGGGTGCAGA	BCR-ABL1 peg5 extension oligo 1
GGCACCGAGTCGGTGCTCCAGGAGGTTCCCGTAGGTCATGA ACTCNNNGATGATATAGAACGGGGGCTCTCTGGTGCAGA	BCR-ABL1 peg6 extension oligo 1
GGCACCGAGTCGGTGCTCCAGGAGGTTCCCGTAGGTCATGA ACTCNNNGATGATATAGAACGGGGGCTCGCGGGTGCAGA	BCR-ABL1 peg7 extension oligo 1
GGCACCGAGTCGGTGCGGGAGCCCCCGTTCTATATCATCNN NGAGTTCATGACATACGGGAACCTCCTGGA	BCR-ABL1 peg8 extension oligo 1
GGCACCGAGTCGGTGCGGGAGCCCCCGTTCTATATCATCNN NGAGTTCATGACCTACGGGAATCTCCTGGACTACCTGAG	BCR-ABL1 peg9 extension oligo 1
GGCACCGAGTCGGTGCGGGAGCCCCCGTTCTATATCATCNN NGAGTTCATGACCTACGGGAACCTGCTGGACTACCTGAG	BCR-ABL1 peg10 extension oligo 1
GGCACCGAGTCGGTGCGGGAGCCCCCGTTCTATATCATCNN NGAGTTCATGACCTACGGGAACCTCCTGGACTATCTGAG	BCR-ABL1 peg11 extension oligo 1
GGCACCGAGTCGGTGCGGGAGCCCCCGTTCTATATCATCNN NGAGTTCATGACCTACGGGAACCTCCTGGACTACCTGAG	BCR-ABL1 peg12 extension oligo 1
GGCACCGAGTCGGTGCGGGAGCCCCCGTTCTATATCATCNN NGAGTTCATGACCTACGGGAACCTCCTGGACTACCTGAG	BCR-ABL1 peg13 extension oligo 1
CGGAGCCACGTGTTGAAGTCCTCGTTGTCTTGTTGGCAGGG GTCTGCACCCGGGAGCCCC	BCR-ABL1 peg1 extension oligo 2
GCAGCACGTGATACACCAAAAAAAAGTCCTCGTTGTCTCTT TGTTGGCAGGGGTCTGCACCCGGGAGCCCCCGTTCTATA	BCR-ABL1 peg2 extension oligo 2
GCAGCACGTGATACACCAAAAAAACTCGTTGTCTTGTTCTT GGCAGGGGTCTGCACCCGGGAGCCCCCGTTCTATATCAT	BCR-ABL1 peg3 extension oligo 2
GCAGCACGTGATACACCAAAAAAATCGTTGTCTTGTTGCTT GCAGGGGTCTGCACCCGGGAGCCCCCGTTCTATATCATC	BCR-ABL1 peg4 extension oligo 2

GCAGCACGTGATACACCAAAAAAACGTTGTCTTGTTGGCTTCAGGGGTCTGCACCCGGGAGCCCC	BCR-ABL1 peg5 extension oligo 2
GCAGCACGTGATACACCAAAAAAAGCAGGGGTCTGCACCAGAGAGCCCCCGTTC	BCR-ABL1 peg6 extension oligo 2
GCAGCACGTGATACACCAAAAAAACAGGGGTCTGCACCCGCGAGCCCCCG	BCR-ABL1 peg7 extension oligo 2
GCAGCACGTGATACACCAAAAAAATCCAGGAGGTTCCCGTATGTCATGA	BCR-ABL1 peg8 extension oligo 2
GCAGCACGTGATACACCAAAAAAACCTCAGGTAGTCCAGGAGATTCCC	BCR-ABL1 peg9 extension oligo 2
GCAGCACGTGATACACCAAAAAAACTCCCTCAGGTAGTCCAGCAGGTTCCC	BCR-ABL1 peg10 extension oligo 2
GCAGCACGTGATACACCAAAAAAACCGGTTGCACTCCCTCAGATAGTCCAGGAGGTTCCC	BCR-ABL1 peg11 extension oligo 2
GCAGCACGTGATACACCAAAAAAAGGCGTTCACCTCCTGCCGATTGCACTCCCTCAGGTAGTCCAGGAGGTTCCC	BCR-ABL1 peg12 extension oligo 2
CATGTACAGCAGCACCACAGCGTTCACCTCCTGCCGGTTGCACTCCCTCAGGTAGTCCAGGAGGTTCCC	BCR-ABL1 peg13 extension oligo 2
AGGACTTCAACACGTGGCTCCGCTCAAGCCAACTGAACAAACTTTTTTTGGTGTATCACGTGCTGC	BCR-ABL1 peg1 extension oligo 3
AACGCTGTGGTGCTGCTGTACATGTTTTTTTGGTGTATCACGTGCTGC	BCR-ABL1 peg13 extension oligo 3
GGCACCGAGTCGGTGCGGGAGCCCCCGNNKTATATCATCACTGAGTTCATGACATACGGGAACCTCCTGGA	BCR-ABL1 peg8 codon 1 randomization extension oligo 1
GGCACCGAGTCGGTGCGGGAGCCCCCGTTCNNKATCATCACTGAGTTCATGACATACGGGAACCTCCTGGA	BCR-ABL1 peg8 codon 2 randomization extension oligo 1
GGCACCGAGTCGGTGCGGGAGCCCCCGTTCTATNNKATCACTGAGTTCATGACATACGGGAACCTCCTGGA	BCR-ABL1 peg8 codon 3 randomization extension oligo 1
GGCACCGAGTCGGTGCGGGAGCCCCCGTTCTATATCNNKACTGAGTTCATGACATACGGGAACCTCCTGGA	BCR-ABL1 peg8 codon 4 randomization extension oligo 1
GGCACCGAGTCGGTGCGGGAGCCCCCGTTCTATATCATCNNKGAGTTCATGACATACGGGAACCTCCTGGA	BCR-ABL1 peg8 codon 5 randomization extension oligo 1

GGCACCGAGTCGGTGCGGGAGCCCCCGTTCTATATCATCAC TNNKTTCATGACATACGGGAACCTCCTGGA	BCR-ABL1 peg8 codon 6 randomization extension oligo 1
GGCACCGAGTCGGTGCGGGAGCCCCCGTTCTATATCATCAC TGAGNNKATGACATACGGGAACCTCCTGGA	BCR-ABL1 peg8 codon 7 randomization extension oligo 1
GGCACCGAGTCGGTGCGGGAGCCCCCGTTCTATATCATCAC TGAGTTCNNKACATACGGGAACCTCCTGGA	BCR-ABL1 peg8 codon 8 randomization extension oligo 1
GCAGCACGTGATACACCAAAAAAATCCAGGAGGTTCCC	BCR-ABL1 peg8 codon 1-8 randomization extension oligo 2

Supplementary Table 4.3: Statistics for amino acid variation in imatinib resistance *BCR-ABL1* mutagenesis screen.

***BCR-ABL1* amino acid number**	**Amino acid variant**	**Test statistic**	**p. value**	**p. value adjusted**
311	N	-1.484344127	0.16854012	1
311	K	4.316026582	0.00152259	0.251227673
311	I	-0.492408636	0.63306157	1
311	M	-0.855952684	0.41206402	1
311	T	-0.143756474	0.88854846	1
311	S	0.569907586	0.58131699	1
311	R	-0.636488906	0.53875314	1
311	Y	1.749063486	0.11084874	1
311	STOP	0.502386825	0.62627306	1
311	F	1.46125194	0.17464461	1
311	L	-0.78357654	0.45144832	1
311	C	-0.801912203	0.44124333	1
311	W	-0.285499225	0.78108785	1
311	H	-2.65138408	0.02425536	1
311	Q	0.622693572	0.54742232	1
311	P	-0.816995101	0.43296405	1
311	D	2.16815514	0.05534215	1
311	E	-1.035431748	0.3248574	1
311	V	-0.189634358	0.85338954	1

311	A	-0.550908122	0.59379012	1
311	G	-1.685446958	0.12280236	1
312	N	-0.500995449	0.62721749	1
312	K	-0.020003793	0.98443385	1
312	I	1.66555752	0.12677168	1
312	M	0.385203069	0.70815342	1
312	T	0.581724592	0.57363112	1
312	S	-0.316312575	0.75826905	1
312	R	0.265073899	0.7963364	1
312	STOP	-0.21441151	0.83453519	1
312	F	1.00041663	0.34070122	1
312	L	-0.354778533	0.73012356	1
312	C	-0.598044219	0.56310894	1
312	W	-0.499232073	0.62841544	1
312	H	-0.666309877	0.52028804	1
312	Q	-1.984797875	0.07526949	1
312	P	1.931772193	0.08219301	1
312	D	-1.595399353	0.14170712	1
312	E	0.101318222	0.92130055	1
312	V	-0.048076917	0.96260158	1
312	A	-0.53771981	0.60253082	1
312	G	-0.445816163	0.66522708	1
313	N	-0.372865318	0.71703033	1
313	K	-0.226962669	0.82502528	1
313	I	-1.896444314	0.08713178	1
313	M	-0.947866104	0.36554028	1
313	T	0.517748243	0.6158935	1
313	S	0.500936506	0.62725751	1
313	R	0.165848132	0.87158157	1
313	Y	-1.007440393	0.33747791	1
313	STOP	-0.994109917	0.34361483	1
313	F	-5.079789453	0.00047793	0.078858968
313	L	0.396256568	0.70023922	1
313	C	1.26008631	0.23624953	1
313	W	-0.691632567	0.50490819	1
313	H	-1.339343125	0.21010367	1

313	Q	0.69725927	0.50152868	1
313	P	0.451884586	0.66099522	1
313	D	-1.101462382	0.2965001	1
313	E	0.233185678	0.82032104	1
313	V	0.291703622	0.77647481	1
313	A	0.142926519	0.8891871	1
313	G	-0.290853594	0.77710628	1
314	N	-0.359227981	0.72689391	1
314	K	-0.154193352	0.88052478	1
314	I	-0.510991606	0.62044815	1
314	M	-0.966614315	0.35653195	1
314	T	-1.160123522	0.27294654	1
314	S	0.032412155	0.97478117	1
314	R	-0.125834728	0.90235671	1
314	Y	0.304735418	0.76681535	1
314	STOP	3.248515334	0.00874244	1
314	F	2.30968195	0.04353097	1
314	L	-0.761823856	0.46375356	1
314	C	0.504913871	0.62455958	1
314	W	-0.514650895	0.61797931	1
314	H	0.582237841	0.57329857	1
314	Q	-0.371265303	0.71818481	1
314	P	-1.011738733	0.33551654	1
314	D	0.220858198	0.82964698	1
314	E	-0.046253818	0.96401862	1
314	V	-0.037370846	0.97092483	1
314	A	-0.02763629	0.97849599	1
314	G	-0.578636687	0.57563414	1
315	N	-0.158126772	0.87750446	1
315	K	0.410823518	0.68986676	1
315	I	6.980226469	3.81E-05	0.006278843
315	M	10.65240545	8.88E-07	0.000146524
315	T	1.008713461	0.33689611	1
315	S	0.276624928	0.78770145	1
315	R	-0.573321226	0.57909102	1
315	Y	1.946609375	0.08019785	1

315	STOP	0.782016614	0.4523236	1
315	F	0.051000161	0.96032974	1
315	L	8.53935972	6.62E-06	0.001092261
315	C	0.316267597	0.75830219	1
315	W	-0.358706267	0.72727231	1
315	H	0.349089534	0.73426101	1
315	Q	0.106505273	0.9172878	1
315	P	-0.530394538	0.60741453	1
315	D	0.212403518	0.83605927	1
315	E	6.380549134	8.03E-05	0.013249912
315	V	2.122858409	0.05973544	1
315	A	-0.086752756	0.93258057	1
315	G	-0.502981972	0.62586931	1
316	N	1.463210441	0.17411957	1
316	K	-0.118707329	0.90785797	1
316	I	-0.744175313	0.47389434	1
316	M	0.021320512	0.9834094	1
316	T	-0.269590948	0.79295618	1
316	S	1.46490931	0.17366523	1
316	R	-0.601782916	0.5607136	1
316	Y	-0.116188168	0.90980361	1
316	STOP	1.939428188	0.08115778	1
316	F	-1.661266955	0.12764292	1
316	L	0.203578488	0.8427659	1
316	C	-0.336009818	0.74380722	1
316	W	-0.223719472	0.82747985	1
316	H	-0.007697629	0.99400964	1
316	Q	-0.360710667	0.72581894	1
316	P	-0.4095552	0.69076724	1
316	D	-0.795764906	0.44464759	1
316	V	0.101975659	0.92079182	1
316	A	-0.815697283	0.43367235	1
316	G	-0.245062545	0.81136366	1
317	N	-1.911918813	0.08493535	1
317	K	-0.863234193	0.40823539	1
317	I	-0.465471287	0.6515659	1

317	M	0.282331576	0.78344648	1
317	T	-0.478538324	0.64255758	1
317	S	-0.803718771	0.44024617	1
317	R	-0.528634001	0.60859131	1
317	Y	-0.970222234	0.35481707	1
317	STOP	-0.789215928	0.44829329	1
317	F	-0.041190931	0.96795449	1
317	L	-0.041224941	0.96792805	1
317	C	-1.069076094	0.31016187	1
317	W	-0.884944314	0.39696602	1
317	H	-0.942245722	0.36827257	1
317	Q	0.088268341	0.93140607	1
317	P	-0.870485518	0.40444704	1
317	D	-0.73935987	0.47668559	1
317	E	-0.215729036	0.83353557	1
317	V	-0.51210358	0.6196974	1
317	A	-0.465315044	0.65167398	1
317	G	-0.91758749	0.38043318	1
318	N	-0.638306696	0.53761677	1
318	K	-0.845746559	0.4174717	1
318	I	0.979285868	0.35053559	1
318	T	-0.116936966	0.90922522	1
318	S	0.587485638	0.56990436	1
318	R	-0.332836198	0.74613044	1
318	Y	-1.15521962	0.27485739	1
318	STOP	-0.172245664	0.86668044	1
318	F	-0.871837317	0.40374351	1
318	L	-0.688312814	0.50690859	1
318	C	-0.468304069	0.64960794	1
318	W	-0.18424623	0.85750298	1
318	H	-1.34012395	0.20985857	1
318	Q	0.050846932	0.96044881	1
318	P	-1.119352186	0.28915529	1
318	D	-0.435272546	0.67260909	1
318	E	-0.259998775	0.80013952	1
318	V	0.055772662	0.95662152	1

318	A	-0.877112746	0.40100606	1
318	G	-0.985688576	0.34753397	1

Supplementary Table 4.4: Oligonucleotide sequences to construct PCR pegRNAs and ngRNAs for saturation mutagenesis at the IRF1 5' UTR element.

Oligo sequence	Description
ATATCTTGTGGAAAGGACGAAACACCGCCTCGGTTCGGCGGG GCTCGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	IRF1 5' UTR peg1/ng1 spacer oligo
ATATCTTGTGGAAAGGACGAAACACCGGCCTCGGTTCGGCGG GGCTGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	IRF1 5' UTR peg2/ng2 spacer oligo 2
ATATCTTGTGGAAAGGACGAAACACCGCGGGTGGCCTCGGTT CGGCGGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	IRF1 5' UTR peg3/ng3 spacer oligo 3
ATATCTTGTGGAAAGGACGAAACACCGTCCGGGTGGCCTCGG TTCGGGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	IRF1 5' UTR peg4/ng4 spacer oligo 4
ATATCTTGTGGAAAGGACGAAACACCGGCACGGCTCCGGGTG GCCTGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	IRF1 5' UTR peg5/ng5 spacer oligo 5
ATATCTTGTGGAAAGGACGAAACACCGGACTGGGCACGGCTC CGGGGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	IRF1 5' UTR peg6/ng6 spacer oligo 6
ATATCTTGTGGAAAGGACGAAACACCGCGTGGACTGGGCACG GCTCGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	IRF1 5' UTR peg7/ng7 spacer oligo 7
ATATCTTGTGGAAAGGACGAAACACCGAGCTCGCCACTCCTT AGTCGGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	IRF1 5' UTR peg8/ng8 spacer oligo 8
ATATCTTGTGGAAAGGACGAAACACCGCTCGGTGGCGCCGCT GCCCGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	IRF1 5' UTR peg9/ng9 spacer oligo 9
ATATCTTGTGGAAAGGACGAAACACCGTAAGTGTTTGGATTG CTCGGGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	IRF1 5' UTR peg10/ng10 spacer oligo 10
ATATCTTGTGGAAAGGACGAAACACCGCGCTAAGTGTTTGGA TTGCTGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAG	IRF1 5' UTR peg11/ng11 spacer oligo 11
GGCACCGAGTCGGTGCAGGCANGACGTNCGCCCGAGCCCCGC CGAACTTTTTTTGGTGTATCACGTGCTGC	IRF1 5' UTR pegRNA 1 w/o PAM mutation extension oligo 1
GGCACCGAGTCGGTGCAGGCANGACGTNCGGGCGAGCCCCGC CGAACTTTTTTTGGTGTATCACGTGCTGC	IRF1 5' UTR pegRNA 1 w/ PAM mutation extension oligo 1

GGCACCGAGTCGGTGCAGGCANGACGTNCGCCCGAGCCCCGC CGAACCTTTTTTTGGTGTATCACGTGCTGC	IRF1 5' UTR pegRNA 2 w/o PAM mutation extension oligo 1
GGCACCGAGTCGGTGCAGGCANGACGTNCGCGGGAGCCCCGC CGAACCTTTTTTTGGTGTATCACGTGCTGC	IRF1 5' UTR pegRNA 2 w/ PAM mutation extension oligo 1
GGCACCGAGTCGGTGCAGGCANGACGTNCGCCCGAGCCCCGC CGAACCGAGGCTTTTTTTGGTGTATCACGTGCTGC	IRF1 5' UTR pegRNA 3 w/o PAM mutation extension oligo 1
GGCACCGAGTCGGTGCAGGCANGACGTNCGCCCGAGGGCCGC CGAACCGAGGCTTTTTTTGGTGTATCACGTGCTGC	IRF1 5' UTR pegRNA 3 w/ PAM mutation extension oligo 1
GGCACCGAGTCGGTGCAGGCANGACGTNCGCCCGAGCCCCGC CGAACCGAGGCCACTTTTTTTGGTGTATCACGTGCTGC	IRF1 5' UTR pegRNA 4 w/o PAM mutation extension oligo 1
GGCACCGAGTCGGTGCAGGCANGACGTNCGCCCGAGCCGGGC CGAACCGAGGCCACTTTTTTTGGTGTATCACGTGCTGC	IRF1 5' UTR pegRNA 4 w/ PAM mutation extension oligo 1
GGCACCGAGTCGGTGCAGGCANGACGTNCGCCCGAGCCCCGC C	IRF1 5' UTR pegRNA 5 w/o PAM mutation extension oligo 1
GGCACCGAGTCGGTGCAGGCANGACGTNCGCCCGAGCCCCGC C	IRF1 5' UTR pegRNA 5 w/ PAM mutation extension oligo 1
GGCACCGAGTCGGTGCAGGCANGACGTNCGCCCGAGCCCCGC CG	IRF1 5' UTR pegRNA 6 w/o PAM mutation extension oligo 1
GGCACCGAGTCGGTGCAGGCANGACGTNCGCCCGAGCCCCGC CG	IRF1 5' UTR pegRNA 6 w/ PAM mutation extension oligo 1
GGCACCGAGTCGGTGCAGGCANGACGTNCGCCCGAGCCCCGC CGA	IRF1 5' UTR pegRNA 7 w/o PAM mutation extension oligo 1
GGCACCGAGTCGGTGCAGGCANGACGTNCGCCCGAGCCCCGC CGA	IRF1 5' UTR pegRNA 7 w/ PAM mutation extension oligo 1

GGCACCGAGTCGGTGCGGGCGNACGTCNTGCCTCGACTAAGG AGTGGCTTTTTTTGGTGTATCACGTGCTGC	IRF1 5' UTR pegRNA 8 w/o PAM mutation extension oligo 1
GGCACCGAGTCGGTGCGGGCGNACGTCNTGGGTCGACTAAGG AGTGGCTTTTTTTGGTGTATCACGTGCTGC	IRF1 5' UTR pegRNA 8 w/ PAM mutation extension oligo 1
GGCACCGAGTCGGTGCGGGCGNACGTCNTGCCTCGACTAAGG AGTGGCGAG	IRF1 5' UTR pegRNA 9 w/o PAM mutation extension oligo 1
GGCACCGAGTCGGTGCGGGCGNACGTCNTGCCTCGACTAAGG AGTGGCGAG	IRF1 5' UTR pegRNA 9 w/ PAM mutation extension oligo 1
GGCACCGAGTCGGTGCGGGCGNACGTCNTGCCTCGACTAAGG AGTGGCGAGCTCTGCCAGGG	IRF1 5' UTR pegRNA 10 w/o PAM mutation extension oligo 1
GGCACCGAGTCGGTGCGGGCGNACGTCNTGCCTCGACTAAGG AGTGGCGAGCTCTGCCAGGG	IRF1 5' UTR pegRNA 10 w/ PAM mutation extension oligo 1
GGCACCGAGTCGGTGCGGGCGNACGTCNTGCCTCGACTAAGG AGTGGCGAGCTCTGCCAGGGC	IRF1 5' UTR pegRNA 11 w/o PAM mutation extension oligo 1
GGCACCGAGTCGGTGCGGGCGNACGTCNTGCCTCGACTAAGG AGTGGCGAGCTCTGCCAGGGC	IRF1 5' UTR pegRNA 11 w/ PAM mutation extension oligo 1
GCAGCACGTGATACACCAAAAAAAGCTCCGGGTGGCCTCGGT TCGGCGGGGCTCGGGCG	IRF1 5' UTR pegRNA 5 w/o PAM mutation extension oligo 2
GCAGCACGTGATACACCAAAAAAAGCTCCGGGTGGCCTCCCT TCGGCGGGGCTCGGGCG	IRF1 5' UTR pegRNA 5 w/ PAM mutation extension oligo 2
GCAGCACGTGATACACCAAAAAAAGGCACGGCTCCGGGTGGC CTCGGTTCGGCGGGGCTCGGGCG	IRF1 5' UTR pegRNA 6 w/o PAM mutation extension oligo 2
GCAGCACGTGATACACCAAAAAAAGGCACGGCTCCGGGTCCC CTCGGTTCGGCGGGGCTCGGGCG	IRF1 5' UTR pegRNA 6 w/ PAM mutation extension oligo 2

GCAGCACGTGATACACCAAAAAAAACTGGGCACGGCTCCGGGTGGCCTCGGTTCGGCGGGGCTCGGGCG	IRF1 5' UTR pegRNA 7 w/o PAM mutation extension oligo 2
GCAGCACGTGATACACCAAAAAAAACTGGGCACGGCTCCCCGTGGCCTCGGTTCGGCGGGGCTCGGGCG	IRF1 5' UTR pegRNA 7 w/ PAM mutation extension oligo 2
GCAGCACGTGATACACCAAAAAAATGGCGCCGCTGCCCTGGCAGAGCTCGCCACTCCTTAGTCGAGGCA	IRF1 5' UTR pegRNA 9 w/o PAM mutation extension oligo 2
GCAGCACGTGATACACCAAAAAAATGGCGCCGCTGCCCTCCCAGAGCTCGCCACTCCTTAGTCGAGGCA	IRF1 5' UTR pegRNA 9 w/ PAM mutation extension oligo 2
GCAGCACGTGATACACCAAAAAAATGTTTGGATTGCTCGGTGGCGCCGCTGCCCTGGCAGAGCTCGCCAC	IRF1 5' UTR pegRNA 10 w/o PAM mutation extension oligo 2
GCAGCACGTGATACACCAAAAAAATGTTTGGATTGCTCGGTCCCGCCGCTGCCCTGGCAGAGCTCGCCACT	IRF1 5' UTR pegRNA 10 w/ PAM mutation extension oligo 2
GCAGCACGTGATACACCAAAAAAACTAAGTGTTTGGATTGCTCGGTGGCGCCGCTGCCCTGGCAGAGCTCGCC	IRF1 5' UTR pegRNA 11 w/o PAM mutation extension oligo 2
GCAGCACGTGATACACCAAAAAAACTAAGTGTTTGGATTGCTCCCTGGCGCCGCTGCCCTGGCAGAGCTCGCCA	IRF1 5' UTR pegRNA 11 w/ PAM mutation extension oligo 2
GGCACCGAGTCGGTGCAGGCNNNACGTGCGCCCGAGCCCCGCCGAACCGAGGCCACCCGGAGCCGTGCCCAGTTTTTTTT	IRF1 5' UTR pegRNA 7 extension oligo 3bp Tile 1
GGCACCGAGTCGGTGCAGGCAAGNNNTGCGCCCGAGCCCCGCCGAACCGAGGCCACCCGGAGCCGTGCCCAGTTTTTTTT	IRF1 5' UTR pegRNA 7 extension oligo 3bp Tile 2
GGCACCGAGTCGGTGCAGGCAAGACGNNNGCCCGAGCCCCGCCGAACCGAGGCCACCCGGAGCCGTGCCCAGTTTTTTTT	IRF1 5' UTR pegRNA 7 extension oligo 3bp Tile 3
GGCACCGAGTCGGTGCAGGCAAGACGTGCNNNCGAGCCCCGCCGAACCGAGGCCACCCGGAGCCGTGCCCAGTTTTTTTT	IRF1 5' UTR pegRNA 7 extension oligo 3bp Tile 4
GGCACCGAGTCGGTGCAGGCAAGACGTGCGCCNNNGCCCCGCCGAACCGAGGCCACCCGGAGCCGTGCCCAGTTTTTTTT	IRF1 5' UTR pegRNA 7 extension oligo 3bp Tile 5
GGCACCGAGTCGGTGCAGGCAAGACGTGCGCCCGANNNCCGCCGAACCGAGGCCACCCGGAGCCGTGCCCAGTTTTTTTT	IRF1 5' UTR pegRNA 7 extension oligo 3bp Tile 6

GGCACCGAGTCGGTGCAGGCAAGACGTGCGCCCGAGCCNNNCCGAACCGAGGCCACCCGGAGCCGTGCCCAGTTTTTTTT	IRF1 5' UTR pegRNA 7 extension oligo 3bp Tile 7
GGCACCGAGTCGGTGCAGGCAAGACGTGCGCCCGAGCCCCGNNNAACCGAGGCCACCCGGAGCCGTGCCCAGTTTTTTTT	IRF1 5' UTR pegRNA 7 extension oligo 3bp Tile 8
GGCACCGAGTCGGTGCAGGCAAGACGTGCGCCCGAGCCCCGCCGNNNCGAGGCCACCCGGAGCCGTGCCCAGTTTTTTTT	IRF1 5' UTR pegRNA 7 extension oligo 3bp Tile 9
GGCACCGAGTCGGTGCAGGCAAGACGTGCGCCCGAGCCCCGCCGAACNNNGGCCACCCGGAGCCGTGCCCAGTTTTTTTT	IRF1 5' UTR pegRNA 7 extension oligo 3bp Tile 10
GGCACCGAGTCGGTGCAGGCAAGACGTGCGCCCGAGCCCCGCCGAACCGANNNCACCCGGAGCCGTGCCCAGTTTTTTTT	IRF1 5' UTR pegRNA 7 extension oligo 3bp Tile 11
GGCACCGAGTCGGTGCAGGCAAGACGTGCGCCCGAGCCCCGCCGAACCGAGGCNNNCCGGAGCCGTGCCCAGTTTTTTTT	IRF1 5' UTR pegRNA 7 extension oligo 3bp Tile 12
GCAGCACGTGATACACCAAAAAAAACTGGGCACGGCTCCGG	IRF1 5' UTR pegRNA 7 extension oligo 2

Supplementary Table 4.5: Oligonucleotide sequences to construct PCR pegRNAs and ngRNAs to assess barcoding effects for pooled pegRNA optimization.

Oligo sequence	PBS length	RTT length	Site	Edit
CAGCACGTGATACACCAAAAAAACTGAGCACGNNNNTGATGGCAGAGCACCGACTCGGTGCCACTT	9	10	HEK site 3	Barcode
CAGCACGTGATACACCAAAAAAACTGAGCACGNNNNTGATGGCAGAGGGCACCGACTCGGTGCCACTT	9	12	HEK site 3	Barcode
CAGCACGTGATACACCAAAAAAACTGAGCACGNNNNTGATGGCAGAGGAAGCACCGACTCGGTGCCACTT	9	14	HEK site 3	Barcode
CAGCACGTGATACACCAAAAAAACTGAGCACGNNNNTGATGGCAGAGGAAAGGCACCGACTCGGTGCCACTT	9	16	HEK site 3	Barcode
CAGCACGTGATACACCAAAAAAACTGAGCACGNNNNTGATGGCAGAGGAAAGGAGCACCGACTCGGTGCCACTT	9	18	HEK site 3	Barcode

CAGCACGTGATACACCAAAAAAACTGAGCACGNNNNTGATGGCAGAGGAAAGGAAGGCACCGACTCGGTGCCACTT	9	20	HEK site 3	Barcode
CAGCACGTGATACACCAAAAAAAGACTGAGCACGNNNNTGATGGCAGAGCACCGACTCGGTGCCACTT	11	10	HEK site 3	Barcode
CAGCACGTGATACACCAAAAAAAGACTGAGCACGNNNNTGATGGCAGAGGGCACCGACTCGGTGCCACTT	11	12	HEK site 3	Barcode
CAGCACGTGATACACCAAAAAAAGACTGAGCACGNNNNTGATGGCAGAGGAAGCACCGACTCGGTGCCACTT	11	14	HEK site 3	Barcode
CAGCACGTGATACACCAAAAAAAGACTGAGCACGNNNNTGATGGCAGAGGAAAGGCACCGACTCGGTGCCACTT	11	16	HEK site 3	Barcode
CAGCACGTGATACACCAAAAAAAGACTGAGCACGNNNNTGATGGCAGAGGAAAGGAGCACCGACTCGGTGCCACTT	11	18	HEK site 3	Barcode
CAGCACGTGATACACCAAAAAAAGACTGAGCACGNNNNTGATGGCAGAGGAAAGGAAGGCACCGACTCGGTGCCACTT	11	20	HEK site 3	Barcode
CAGCACGTGATACACCAAAAAAACAGACTGAGCACGNNNNTGATGGCAGAGCACCGACTCGGTGCCACTT	13	10	HEK site 3	Barcode
CAGCACGTGATACACCAAAAAAACAGACTGAGCACGNNNNTGATGGCAGAGGGCACCGACTCGGTGCCACTT	13	12	HEK site 3	Barcode
CAGCACGTGATACACCAAAAAAACAGACTGAGCACGNNNNTGATGGCAGAGGAAGCACCGACTCGGTGCCACTT	13	14	HEK site 3	Barcode
CAGCACGTGATACACCAAAAAAACAGACTGAGCACGNNNNTGATGGCAGAGGAAAGGCACCGACTCGGTGCCACTT	13	16	HEK site 3	Barcode
CAGCACGTGATACACCAAAAAAACAGACTGAGCACGNNNNTGATGGCAGAGGAAAGGAGCACCGACTCGGTGCCACTT	13	18	HEK site 3	Barcode

CAGCACGTGATACACCAAAAAAACAG ACTGAGCACGNNNNTGATGGCAGAGG AAAGGAAGGCACCGACTCGGTGCCACT T	13	20	HEK site 3	Barcode
CAGCACGTGATACACCAAAAAAACCCA GACTGAGCACGNNNNTGATGGCAGAG CACCGACTCGGTGCCACTT	15	10	HEK site 3	Barcode
CAGCACGTGATACACCAAAAAAACCCA GACTGAGCACGNNNNTGATGGCAGAG GGCACCGACTCGGTGCCACTT	15	12	HEK site 3	Barcode
CAGCACGTGATACACCAAAAAAACCCA GACTGAGCACGNNNNTGATGGCAGAG GAAGCACCGACTCGGTGCCACTT	15	14	HEK site 3	Barcode
CAGCACGTGATACACCAAAAAAACCCA GACTGAGCACGNNNNTGATGGCAGAG GAAAGGCACCGACTCGGTGCCACTT	15	16	HEK site 3	Barcode
CAGCACGTGATACACCAAAAAAACCCA GACTGAGCACGNNNNTGATGGCAGAG GAAAGGAGCACCGACTCGGTGCCACTT	15	18	HEK site 3	Barcode
GCACGTGATACACCAAAAAAACCCAG ACTGAGCACGNNNNTGATGGCAGAGG AAAGGAAGGCACCGACTCGGTGCCACT T	15	20	HEK site 3	Barcode
CAGCACGTGATACACCAAAAAAAGGC CCAGACTGAGCACGNNNNTGATGGCA GAGCACCGACTCGGTGCCACTT	17	10	HEK site 3	Barcode
CAGCACGTGATACACCAAAAAAAGGC CCAGACTGAGCACGNNNNTGATGGCA GAGGGCACCGACTCGGTGCCACTT	17	12	HEK site 3	Barcode
CAGCACGTGATACACCAAAAAAAGGC CCAGACTGAGCACGNNNNTGATGGCA GAGGAAGCACCGACTCGGTGCCACTT	17	14	HEK site 3	Barcode
CAGCACGTGATACACCAAAAAAAGGC CCAGACTGAGCACGNNNNTGATGGCA GAGGAAAGGCACCGACTCGGTGCCACT T	17	16	HEK site 3	Barcode

GCACGTGATACACCAAAAAAAGGCCCAGACTGAGCACGNNNNTGATGGCAGAGGAAAGGAGCACCGACTCGGTGCCACTT	17	18	HEK site 3	Barcode
ACGTGATACACCAAAAAAAGGCCCAGACTGAGCACGNNNNTGATGGCAGAGGAAAGGAAGGCACCGACTCGGTGCCACTT	17	20	HEK site 3	Barcode
CAGCACGTGATACACCAAAAAAACTGAGCACGAGATGGCAGAGCACCGACTCGGTGCCACTT	9	10	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAACTGAGCACGAGATGGCAGAGGGCACCGACTCGGTGCCACTT	9	12	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAACTGAGCACGAGATGGCAGAGGAAGCACCGACTCGGTGCCACTT	9	14	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAACTGAGCACGAGATGGCAGAGGAAAGGCACCGACTCGGTGCCACTT	9	16	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAACTGAGCACGAGATGGCAGAGGAAAGGAGCACCGACTCGGTGCCACTT	9	18	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAACTGAGCACGAGATGGCAGAGGAAAGGAAGGCACCGACTCGGTGCCACTT	9	20	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAAGACTGAGCACGAGATGGCAGAGCACCGACTCGGTGCCACTT	11	10	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAAGACTGAGCACGAGATGGCAGAGGGCACCGACTCGGTGCCACTT	11	12	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAAGACTGAGCACGAGATGGCAGAGGAAGCACCGACTCGGTGCCACTT	11	14	HEK site 3	Substitution

CAGCACGTGATACACCAAAAAAAGACT GAGCACGAGATGGCAGAGGAAAGGCA CCGACTCGGTGCCACTT	11	16	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAAGACT GAGCACGAGATGGCAGAGGAAAGGAG CACCGACTCGGTGCCACTT	11	18	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAAGACT GAGCACGAGATGGCAGAGGAAAGGAA GGCACCGACTCGGTGCCACTT	11	20	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAACAG ACTGAGCACGAGATGGCAGAGCACCG ACTCGGTGCCACTT	13	10	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAACAG ACTGAGCACGAGATGGCAGAGGGCAC CGACTCGGTGCCACTT	13	12	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAACAG ACTGAGCACGAGATGGCAGAGGAAGC ACCGACTCGGTGCCACTT	13	14	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAACAG ACTGAGCACGAGATGGCAGAGGAAAG GCACCGACTCGGTGCCACTT	13	16	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAACAG ACTGAGCACGAGATGGCAGAGGAAAG GAGCACCGACTCGGTGCCACTT	13	18	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAACAG ACTGAGCACGAGATGGCAGAGGAAAG GAAGGCACCGACTCGGTGCCACTT	13	20	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAACCCA GACTGAGCACGAGATGGCAGAGCACC GACTCGGTGCCACTT	15	10	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAACCCA GACTGAGCACGAGATGGCAGAGGGCA CCGACTCGGTGCCACTT	15	12	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAACCCA GACTGAGCACGAGATGGCAGAGGAAG CACCGACTCGGTGCCACTT	15	14	HEK site 3	Substitution

CAGCACGTGATACACCAAAAAAACCCA GACTGAGCACGAGATGGCAGAGGAAA GGCACCGACTCGGTGCCACTT	15	16	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAACCCA GACTGAGCACGAGATGGCAGAGGAAA GGAGCACCGACTCGGTGCCACTT	15	18	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAACCCA GACTGAGCACGAGATGGCAGAGGAAA GGAAGGCACCGACTCGGTGCCACTT	15	20	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAAGGC CCAGACTGAGCACGAGATGGCAGAGC ACCGACTCGGTGCCACTT	17	10	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAAGGC CCAGACTGAGCACGAGATGGCAGAGG GCACCGACTCGGTGCCACTT	17	12	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAAGGC CCAGACTGAGCACGAGATGGCAGAGG AAGCACCGACTCGGTGCCACTT	17	14	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAAGGC CCAGACTGAGCACGAGATGGCAGAGG AAAGGCACCGACTCGGTGCCACTT	17	16	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAAGGC CCAGACTGAGCACGAGATGGCAGAGG AAAGGAGCACCGACTCGGTGCCACTT	17	18	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAAGGC CCAGACTGAGCACGAGATGGCAGAGG AAAGGAAGGCACCGACTCGGTGCCACT T	17	20	HEK site 3	Substitution
CAGCACGTGATACACCAAAAAAAGCA GAAGAANNNNGAAGGGCTCCGCACCG ACTCGGTGCCACTT	9	10	EMX1 site 1	Barcode
CAGCACGTGATACACCAAAAAAAGCA GAAGAANNNNGAAGGGCTCCCAGCAC CGACTCGGTGCCACTT	9	12	EMX1 site 1	Barcode
CAGCACGTGATACACCAAAAAAAGCA GAAGAANNNNGAAGGGCTCCCATCGC ACCGACTCGGTGCCACTT	9	14	EMX1 site 1	Barcode

CAGCACGTGATACACCAAAAAAAGCA GAAGAANNNNGAAGGGCTCCCATCAC GCACCGACTCGGTGCCACTT	9	16	EMX1 site 1	Barcode
CAGCACGTGATACACCAAAAAAAGCA GAAGAANNNNGAAGGGCTCCCATCAC ATGCACCGACTCGGTGCCACTT	9	18	EMX1 site 1	Barcode
CAGCACGTGATACACCAAAAAAAGCA GAAGAANNNNGAAGGGCTCCCATCAC ATCAGCACCGACTCGGTGCCACTT	9	20	EMX1 site 1	Barcode
CAGCACGTGATACACCAAAAAAAGAG CAGAAGAANNNNGAAGGGCTCCGCAC CGACTCGGTGCCACTT	11	10	EMX1 site 1	Barcode
CAGCACGTGATACACCAAAAAAAGAG CAGAAGAANNNNGAAGGGCTCCCAGC ACCGACTCGGTGCCACTT	11	12	EMX1 site 1	Barcode
CAGCACGTGATACACCAAAAAAAGAG CAGAAGAANNNNGAAGGGCTCCCATC GCACCGACTCGGTGCCACTT	11	14	EMX1 site 1	Barcode
CAGCACGTGATACACCAAAAAAAGAG CAGAAGAANNNNGAAGGGCTCCCATC ACGCACCGACTCGGTGCCACTT	11	16	EMX1 site 1	Barcode
CAGCACGTGATACACCAAAAAAAGAG CAGAAGAANNNNGAAGGGCTCCCATC ACATGCACCGACTCGGTGCCACTT	11	18	EMX1 site 1	Barcode
CAGCACGTGATACACCAAAAAAAGAG CAGAAGAANNNNGAAGGGCTCCCATC ACATCAGCACCGACTCGGTGCCACTT	11	20	EMX1 site 1	Barcode
CAGCACGTGATACACCAAAAAAACCG AGCAGAAGAANNNNGAAGGGCTCCGC ACCGACTCGGTGCCACTT	13	10	EMX1 site 1	Barcode
CAGCACGTGATACACCAAAAAAACCG AGCAGAAGAANNNNGAAGGGCTCCCA GCACCGACTCGGTGCCACTT	13	12	EMX1 site 1	Barcode
CAGCACGTGATACACCAAAAAAACCG AGCAGAAGAANNNNGAAGGGCTCCCA TCGCACCGACTCGGTGCCACTT	13	14	EMX1 site 1	Barcode

CAGCACGTGATACACCAAAAAAACCG AGCAGAAGAANNNNGAAGGGCTCCCA TCACGCACCGACTCGGTGCCACTT	13	16	EMX1 site 1	Barcode
CAGCACGTGATACACCAAAAAAACCG AGCAGAAGAANNNNGAAGGGCTCCCA TCACATGCACCGACTCGGTGCCACTT	13	18	EMX1 site 1	Barcode
CAGCACGTGATACACCAAAAAAACCG AGCAGAAGAANNNNGAAGGGCTCCCA TCACATCAGCACCGACTCGGTGCCACT T	13	20	EMX1 site 1	Barcode
CAGCACGTGATACACCAAAAAAAGTCC GAGCAGAAGAANNNNGAAGGGCTCCG CACCGACTCGGTGCCACTT	15	10	EMX1 site 1	Barcode
CAGCACGTGATACACCAAAAAAAGTCC GAGCAGAAGAANNNNGAAGGGCTCCC AGCACCGACTCGGTGCCACTT	15	12	EMX1 site 1	Barcode
CAGCACGTGATACACCAAAAAAAGTCC GAGCAGAAGAANNNNGAAGGGCTCCC ATCGCACCGACTCGGTGCCACTT	15	14	EMX1 site 1	Barcode
CAGCACGTGATACACCAAAAAAAGTCC GAGCAGAAGAANNNNGAAGGGCTCCC ATCACGCACCGACTCGGTGCCACTT	15	16	EMX1 site 1	Barcode
CAGCACGTGATACACCAAAAAAAGTCC GAGCAGAAGAANNNNGAAGGGCTCCC ATCACATGCACCGACTCGGTGCCACTT	15	18	EMX1 site 1	Barcode
CAGCACGTGATACACCAAAAAAAGTCC GAGCAGAAGAANNNNGAAGGGCTCCC ATCACATCAGCACCGACTCGGTGCCAC TT	15	20	EMX1 site 1	Barcode
CAGCACGTGATACACCAAAAAAAGAG TCCGAGCAGAAGAANNNNGAAGGGCT CCGCACCGACTCGGTGCCACTT	17	10	EMX1 site 1	Barcode
CAGCACGTGATACACCAAAAAAAGAG TCCGAGCAGAAGAANNNNGAAGGGCT CCCAGCACCGACTCGGTGCCACTT	17	12	EMX1 site 1	Barcode

CAGCACGTGATACACCAAAAAAAGAG TCCGAGCAGAAGAANNNNGAAGGGCT CCCATCGCACCGACTCGGTGCCACTT	17	14	EMX1 site 1	Barcode
CAGCACGTGATACACCAAAAAAAGAG TCCGAGCAGAAGAANNNNGAAGGGCT CCCATCACGCACCGACTCGGTGCCACT T	17	16	EMX1 site 1	Barcode
CAGCACGTGATACACCAAAAAAAGAG TCCGAGCAGAAGAANNNNGAAGGGCT CCCATCACATGCACCGACTCGGTGCCA CTT	17	18	EMX1 site 1	Barcode
CAGCACGTGATACACCAAAAAAAGAG TCCGAGCAGAAGAANNNNGAAGGGCT CCCATCACATCAGCACCGACTCGGTGC CACTT	17	20	EMX1 site 1	Barcode
CAGCACGTGATACACCAAAAAAAGCA GAAGAAGAAGTGCTCCGCACCGACTCG GTGCCACTT	9	10	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAAGCA GAAGAAGAAGTGCTCCCAGCACCGACT CGGTGCCACTT	9	12	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAAGCA GAAGAAGAAGTGCTCCCATCGCACCGA CTCGGTGCCACTT	9	14	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAAGCA GAAGAAGAAGTGCTCCCATCACGCACC GACTCGGTGCCACTT	9	16	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAAGCA GAAGAAGAAGTGCTCCCATCACATGCA CCGACTCGGTGCCACTT	9	18	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAAGCA GAAGAAGAAGTGCTCCCATCACATCAG CACCGACTCGGTGCCACTT	9	20	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAAGAG CAGAAGAAGAAGTGCTCCGCACCGACT CGGTGCCACTT	11	10	EMX1 site 1	Substitution

CAGCACGTGATACACCAAAAAAAGAG CAGAAGAAGAAGTGCTCCCAGCACCG ACTCGGTGCCACTT	11	12	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAAGAG CAGAAGAAGAAGTGCTCCCATCGCACC GACTCGGTGCCACTT	11	14	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAAGAG CAGAAGAAGAAGTGCTCCCATCACGCA CCGACTCGGTGCCACTT	11	16	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAAGAG CAGAAGAAGAAGTGCTCCCATCACATG CACCGACTCGGTGCCACTT	11	18	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAAGAG CAGAAGAAGAAGTGCTCCCATCACATC AGCACCGACTCGGTGCCACTT	11	20	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAACCG AGCAGAAGAAGAAGTGCTCCGCACCG ACTCGGTGCCACTT	13	10	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAACCG AGCAGAAGAAGAAGTGCTCCCAGCAC CGACTCGGTGCCACTT	13	12	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAACCG AGCAGAAGAAGAAGTGCTCCCATCGCA CCGACTCGGTGCCACTT	13	14	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAACCG AGCAGAAGAAGAAGTGCTCCCATCACG CACCGACTCGGTGCCACTT	13	16	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAACCG AGCAGAAGAAGAAGTGCTCCCATCACA TGCACCGACTCGGTGCCACTT	13	18	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAACCG AGCAGAAGAAGAAGTGCTCCCATCACA TCAGCACCGACTCGGTGCCACTT	13	20	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAAGTCC GAGCAGAAGAAGAAGTGCTCCGCACC GACTCGGTGCCACTT	15	10	EMX1 site 1	Substitution

CAGCACGTGATACACCAAAAAAAGTCC GAGCAGAAGAAGAAGTGCTCCCAGCA CCGACTCGGTGCCACTT	15	12	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAAGTCC GAGCAGAAGAAGAAGTGCTCCCATCGC ACCGACTCGGTGCCACTT	15	14	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAAGTCC GAGCAGAAGAAGAAGTGCTCCCATCAC GCACCGACTCGGTGCCACTT	15	16	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAAGTCC GAGCAGAAGAAGAAGTGCTCCCATCAC ATGCACCGACTCGGTGCCACTT	15	18	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAAGTCC GAGCAGAAGAAGAAGTGCTCCCATCAC ATCAGCACCGACTCGGTGCCACTT	15	20	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAAGAG TCCGAGCAGAAGAAGAAGTGCTCCGCA CCGACTCGGTGCCACTT	17	10	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAAGAG TCCGAGCAGAAGAAGAAGTGCTCCCAG CACCGACTCGGTGCCACTT	17	12	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAAGAG TCCGAGCAGAAGAAGAAGTGCTCCCAT CGCACCGACTCGGTGCCACTT	17	14	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAAGAG TCCGAGCAGAAGAAGAAGTGCTCCCAT CACGCACCGACTCGGTGCCACTT	17	16	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAAGAG TCCGAGCAGAAGAAGAAGTGCTCCCAT CACATGCACCGACTCGGTGCCACTT	17	18	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAAGAG TCCGAGCAGAAGAAGAAGTGCTCCCAT CACATCAGCACCGACTCGGTGCCACTT	17	20	EMX1 site 1	Substitution
CAGCACGTGATACACCAAAAAAAAGTC ATTACNNNNCTGAGGTGTTGCACCGAC TCGGTGCCACTT	9	10	RNF2 site 1	Barcode

CAGCACGTGATACACCAAAAAAAAGTC ATTACNNNNCTGAGGTGTTCGGCACCG ACTCGGTGCCACTT	9	12	RNF2 site 1	Barcode
CAGCACGTGATACACCAAAAAAAAGTC ATTACNNNNCTGAGGTGTTCGTTGCAC CGACTCGGTGCCACTT	9	14	RNF2 site 1	Barcode
CAGCACGTGATACACCAAAAAAAAGTC ATTACNNNNCTGAGGTGTTCGTTGTGC ACCGACTCGGTGCCACTT	9	16	RNF2 site 1	Barcode
CAGCACGTGATACACCAAAAAAAAGTC ATTACNNNNCTGAGGTGTTCGTTGTAA GCACCGACTCGGTGCCACTT	9	18	RNF2 site 1	Barcode
CAGCACGTGATACACCAAAAAAAAGTC ATTACNNNNCTGAGGTGTTCGTTGTAA CTGCACCGACTCGGTGCCACTT	9	20	RNF2 site 1	Barcode
CAGCACGTGATACACCAAAAAAATTAG TCATTACNNNNCTGAGGTGTTGCACCG ACTCGGTGCCACTT	11	10	RNF2 site 1	Barcode
CAGCACGTGATACACCAAAAAAATTAG TCATTACNNNNCTGAGGTGTTCGGCAC CGACTCGGTGCCACTT	11	12	RNF2 site 1	Barcode
CAGCACGTGATACACCAAAAAAATTAG TCATTACNNNNCTGAGGTGTTCGTTGC ACCGACTCGGTGCCACTT	11	14	RNF2 site 1	Barcode
CAGCACGTGATACACCAAAAAAATTAG TCATTACNNNNCTGAGGTGTTCGTTGT GCACCGACTCGGTGCCACTT	11	16	RNF2 site 1	Barcode
CAGCACGTGATACACCAAAAAAATTAG TCATTACNNNNCTGAGGTGTTCGTTGT AAGCACCGACTCGGTGCCACTT	11	18	RNF2 site 1	Barcode
CAGCACGTGATACACCAAAAAAATTAG TCATTACNNNNCTGAGGTGTTCGTTGT AACTGCACCGACTCGGTGCCACTT	11	20	RNF2 site 1	Barcode
CAGCACGTGATACACCAAAAAAATCTT AGTCATTACNNNNCTGAGGTGTTGCAC CGACTCGGTGCCACTT	13	10	RNF2 site 1	Barcode

CAGCACGTGATACACCAAAAAAATCTT AGTCATTACNNNNCTGAGGTGTTCGGC ACCGACTCGGTGCCACTT	13	12	RNF2 site 1	Barcode
CAGCACGTGATACACCAAAAAAATCTT AGTCATTACNNNNCTGAGGTGTTCGTT GCACCGACTCGGTGCCACTT	13	14	RNF2 site 1	Barcode
CAGCACGTGATACACCAAAAAAATCTT AGTCATTACNNNNCTGAGGTGTTCGTT GTGCACCGACTCGGTGCCACTT	13	16	RNF2 site 1	Barcode
CAGCACGTGATACACCAAAAAAATCTT AGTCATTACNNNNCTGAGGTGTTCGTT GTAAGCACCGACTCGGTGCCACTT	13	18	RNF2 site 1	Barcode
CAGCACGTGATACACCAAAAAAATCTT AGTCATTACNNNNCTGAGGTGTTCGTT GTAACTGCACCGACTCGGTGCCACTT	13	20	RNF2 site 1	Barcode
CAGCACGTGATACACCAAAAAAACATC TTAGTCATTACNNNNCTGAGGTGTTGC ACCGACTCGGTGCCACTT	15	10	RNF2 site 1	Barcode
CAGCACGTGATACACCAAAAAAACATC TTAGTCATTACNNNNCTGAGGTGTTCG GCACCGACTCGGTGCCACTT	15	12	RNF2 site 1	Barcode
CAGCACGTGATACACCAAAAAAACATC TTAGTCATTACNNNNCTGAGGTGTTCG TTGCACCGACTCGGTGCCACTT	15	14	RNF2 site 1	Barcode
CAGCACGTGATACACCAAAAAAACATC TTAGTCATTACNNNNCTGAGGTGTTCG TTGTGCACCGACTCGGTGCCACTT	15	16	RNF2 site 1	Barcode
CAGCACGTGATACACCAAAAAAACATC TTAGTCATTACNNNNCTGAGGTGTTCG TTGTAAGCACCGACTCGGTGCCACTT	15	18	RNF2 site 1	Barcode
CAGCACGTGATACACCAAAAAAACATC TTAGTCATTACNNNNCTGAGGTGTTCG TTGTAACTGCACCGACTCGGTGCCACT T	15	20	RNF2 site 1	Barcode
CAGCACGTGATACACCAAAAAAAGTCA TCTTAGTCATTACNNNNCTGAGGTGTT GCACCGACTCGGTGCCACTT	17	10	RNF2 site 1	Barcode

CAGCACGTGATACACCAAAAAAAGTCA TCTTAGTCATTACNNNNCTGAGGTGTT CGGCACCGACTCGGTGCCACTT	17	12	RNF2 site 1	Barcode
CAGCACGTGATACACCAAAAAAAGTCA TCTTAGTCATTACNNNNCTGAGGTGTT CGTTGCACCGACTCGGTGCCACTT	17	14	RNF2 site 1	Barcode
CAGCACGTGATACACCAAAAAAAGTCA TCTTAGTCATTACNNNNCTGAGGTGTT CGTTGTGCACCGACTCGGTGCCACTT	17	16	RNF2 site 1	Barcode
CAGCACGTGATACACCAAAAAAAGTCA TCTTAGTCATTACNNNNCTGAGGTGTT CGTTGTAAGCACCGACTCGGTGCCACT T	17	18	RNF2 site 1	Barcode
CAGCACGTGATACACCAAAAAAAGTCA TCTTAGTCATTACNNNNCTGAGGTGTT CGTTGTAACTGCACCGACTCGGTGCCA CTT	17	20	RNF2 site 1	Barcode
CAGCACGTGATACACCAAAAAAAAGTC ATTACATGAGGTGTTGCACCGACTCGG TGCCACTT	9	10	RNF2 site 1	Substitution
CAGCACGTGATACACCAAAAAAAAGTC ATTACATGAGGTGTTCGGCACCGACTC GGTGCCACTT	9	12	RNF2 site 1	Substitution
CAGCACGTGATACACCAAAAAAAAGTC ATTACATGAGGTGTTCGTTGCACCGAC TCGGTGCCACTT	9	14	RNF2 site 1	Substitution
CAGCACGTGATACACCAAAAAAAAGTC ATTACATGAGGTGTTCGTTGTGCACCG ACTCGGTGCCACTT	9	16	RNF2 site 1	Substitution
CAGCACGTGATACACCAAAAAAAAGTC ATTACATGAGGTGTTCGTTGTAAGCAC CGACTCGGTGCCACTT	9	18	RNF2 site 1	Substitution
CAGCACGTGATACACCAAAAAAAAGTC ATTACATGAGGTGTTCGTTGTAACTGC ACCGACTCGGTGCCACTT	9	20	RNF2 site 1	Substitution

CAGCACGTGATACACCAAAAAAATTAG TCATTACATGAGGTGTTGCACCGACTC GGTGCCACTT	11	10	RNF2 site 1	Substitution
CAGCACGTGATACACCAAAAAAATTAG TCATTACATGAGGTGTTCGGCACCGAC TCGGTGCCACTT	11	12	RNF2 site 1	Substitution
CAGCACGTGATACACCAAAAAAATTAG TCATTACATGAGGTGTTCGTTGCACCG ACTCGGTGCCACTT	11	14	RNF2 site 1	Substitution
CAGCACGTGATACACCAAAAAAATTAG TCATTACATGAGGTGTTCGTTGTGCACC GACTCGGTGCCACTT	11	16	RNF2 site 1	Substitution
CAGCACGTGATACACCAAAAAAATTAG TCATTACATGAGGTGTTCGTTGTAAGC ACCGACTCGGTGCCACTT	11	18	RNF2 site 1	Substitution
CAGCACGTGATACACCAAAAAAATTAG TCATTACATGAGGTGTTCGTTGTAACTG CACCGACTCGGTGCCACTT	11	20	RNF2 site 1	Substitution
CAGCACGTGATACACCAAAAAAATCTT AGTCATTACATGAGGTGTTGCACCGAC TCGGTGCCACTT	13	10	RNF2 site 1	Substitution
CAGCACGTGATACACCAAAAAAATCTT AGTCATTACATGAGGTGTTCGGCACCG ACTCGGTGCCACTT	13	12	RNF2 site 1	Substitution
CAGCACGTGATACACCAAAAAAATCTT AGTCATTACATGAGGTGTTCGTTGCAC CGACTCGGTGCCACTT	13	14	RNF2 site 1	Substitution
CAGCACGTGATACACCAAAAAAATCTT AGTCATTACATGAGGTGTTCGTTGTGC ACCGACTCGGTGCCACTT	13	16	RNF2 site 1	Substitution
CAGCACGTGATACACCAAAAAAATCTT AGTCATTACATGAGGTGTTCGTTGTAA GCACCGACTCGGTGCCACTT	13	18	RNF2 site 1	Substitution
CAGCACGTGATACACCAAAAAAATCTT AGTCATTACATGAGGTGTTCGTTGTAA CTGCACCGACTCGGTGCCACTT	13	20	RNF2 site 1	Substitution

CAGCACGTGATACACCAAAAAAACATC TTAGTCATTACATGAGGTGTTGCACCG ACTCGGTGCCACTT	15	10	RNF2 site 1	Substitution
CAGCACGTGATACACCAAAAAAACATC TTAGTCATTACATGAGGTGTTCGGCAC CGACTCGGTGCCACTT	15	12	RNF2 site 1	Substitution
CAGCACGTGATACACCAAAAAAACATC TTAGTCATTACATGAGGTGTTCGTTGCA CCGACTCGGTGCCACTT	15	14	RNF2 site 1	Substitution
CAGCACGTGATACACCAAAAAAACATC TTAGTCATTACATGAGGTGTTCGTTGTG CACCGACTCGGTGCCACTT	15	16	RNF2 site 1	Substitution
CAGCACGTGATACACCAAAAAAACATC TTAGTCATTACATGAGGTGTTCGTTGTA AGCACCGACTCGGTGCCACTT	15	18	RNF2 site 1	Substitution
CAGCACGTGATACACCAAAAAAACATC TTAGTCATTACATGAGGTGTTCGTTGTA ACTGCACCGACTCGGTGCCACTT	15	20	RNF2 site 1	Substitution
CAGCACGTGATACACCAAAAAAAGTCA TCTTAGTCATTACATGAGGTGTTGCACC GACTCGGTGCCACTT	17	10	RNF2 site 1	Substitution
CAGCACGTGATACACCAAAAAAAGTCA TCTTAGTCATTACATGAGGTGTTCGGC ACCGACTCGGTGCCACTT	17	12	RNF2 site 1	Substitution
CAGCACGTGATACACCAAAAAAAGTCA TCTTAGTCATTACATGAGGTGTTCGTTG CACCGACTCGGTGCCACTT	17	14	RNF2 site 1	Substitution
CAGCACGTGATACACCAAAAAAAGTCA TCTTAGTCATTACATGAGGTGTTCGTTG TGCACCGACTCGGTGCCACTT	17	16	RNF2 site 1	Substitution
CAGCACGTGATACACCAAAAAAAGTCA TCTTAGTCATTACATGAGGTGTTCGTTG TAAGCACCGACTCGGTGCCACTT	17	18	RNF2 site 1	Substitution
CAGCACGTGATACACCAAAAAAAGTCA TCTTAGTCATTACATGAGGTGTTCGTTG TAACTGCACCGACTCGGTGCCACTT	17	20	RNF2 site 1	Substitution

Supplementary Table 4.6: Oligonucleotide sequences to construct PCR pegRNAs and ngRNAs to assess efficiencies of pooled and single construct prime editing.

Oligo Order	Barcode	PBS length	RTT length	Site
CGCACCAAAAAAAGGCCCAGACTGAGC ACGACCATGATGGCAGAGGAAAGGAAG CCCTGCTTCCGCACCGACTCGGTGCC	ACCA	17	30	HEK site 3
CGCACCAAAAAAAGGCCCAGACTGAGC ACGAACGTGATGGCAGAGGAAAGGAAG CCCTGCACCGACTCGGTGCC	AACG	17	24	HEK site 3
CGCACCAAAAAAAGGCCCAGACTGAGC ACGGAGATGATGGCAGAGGAAAGGAAG CACCGACTCGGTGCC	GAGA	17	19	HEK site 3
CGCACCAAAAAAAGGCCCAGACTGAGC ACGGTCATGATGGCAGAGGAAAGCACC GACTCGGTGCC	GTCA	17	15	HEK site 3
CGCACCAAAAAAAGGCCCAGACTGAGC ACGAGCTTGATGGCAGAGCACCGACTCG GTGCC	AGCT	17	10	HEK site 3
CGCACCAAAAAAAGGCCCAGACTGAGC ACGGATCTGATGGCGCACCGACTCGGTG CC	GATC	17	7	HEK site 3
CGCACCAAAAAAAGCCCAGACTGAGCA CGACGTTGATGGCAGAGGAAAGGAAGC CCTGCTTCCGCACCGACTCGGTGCC	ACGT	16	30	HEK site 3
CGCACCAAAAAAAGCCCAGACTGAGCA CGGTGTTGATGGCAGAGGAAAGGAAGC CCTGCACCGACTCGGTGCC	GTGT	16	24	HEK site 3
CGCACCAAAAAAAGCCCAGACTGAGCA CGATCGTGATGGCAGAGGAAAGGAAGC ACCGACTCGGTGCC	ATCG	16	19	HEK site 3
CGCACCAAAAAAAGCCCAGACTGAGCA CGCGTATGATGGCAGAGGAAAGCACCG ACTCGGTGCC	CGTA	16	15	HEK site 3

CGCACCAAAAAAAGCCCAGACTGAGCACGACCTTGATGGCAGAGCACCGACTCGGTGCC	ACCT	16	10	HEK site 3
CGCACCAAAAAAAGCCCAGACTGAGCACGGAGTTGATGGCGCACCGACTCGGTGCC	GAGT	16	7	HEK site 3
CGCACCAAAAAAACCCAGACTGAGCACGATCCTGATGGCAGAGGAAAGGAAGCCCTGCTTCCGCACCGACTCGGTGCC	ATCC	15	30	HEK site 3
CGCACCAAAAAAACCCAGACTGAGCACGGTTCTGATGGCAGAGGAAAGGAAGCCCTGCACCGACTCGGTGCC	GTTC	15	24	HEK site 3
CGCACCAAAAAAACCCAGACTGAGCACGGTCTTGATGGCAGAGGAAAGGAAGCACCGACTCGGTGCC	GTCT	15	19	HEK site 3
CGCACCAAAAAAACCCAGACTGAGCACGCTCTTGATGGCAGAGGAAAGCACCGACTCGGTGCC	CTCT	15	15	HEK site 3
CGCACCAAAAAAACCCAGACTGAGCACGCTACTGATGGCAGAGCACCGACTCGGTGCC	CTAC	15	10	HEK site 3
CGCACCAAAAAAACCCAGACTGAGCACGATGCTGATGGCGCACCGACTCGGTGCC	ATGC	15	7	HEK site 3
CGCACCAAAAAAACCAGACTGAGCACGAGTGTGATGGCAGAGGAAAGGAAGCCCTGCTTCCGCACCGACTCGGTGCC	AGTG	14	30	HEK site 3
CGCACCAAAAAAACCAGACTGAGCACGCAGATGATGGCAGAGGAAAGGAAGCCCTGCACCGACTCGGTGCC	CAGA	14	24	HEK site 3
CGCACCAAAAAAACCAGACTGAGCACGGATGTGATGGCAGAGGAAAGGAAGCACCGACTCGGTGCC	GATG	14	19	HEK site 3
CGCACCAAAAAAACCAGACTGAGCACGTTCCTGATGGCAGAGGAAAGCACCGACTCGGTGCC	TTCC	14	15	HEK site 3

CGCACCAAAAAAACCAGACTGAGCACG GACATGATGGCAGAGCACCGACTCGGT GCC	GACA	14	10	HEK site 3
CGCACCAAAAAAACCAGACTGAGCACG GTAGTGATGGCGCACCGACTCGGTGCC	GTAG	14	7	HEK site 3
CGCACCAAAAAAACAGACTGAGCACGA CTCTGATGGCAGAGGAAAGGAAGCCCT GCTTCCGCACCGACTCGGTGCC	ACTC	13	30	HEK site 3
CGCACCAAAAAAACAGACTGAGCACGT ACGTGATGGCAGAGGAAAGGAAGCCCT GCACCGACTCGGTGCC	TACG	13	24	HEK site 3
CGCACCAAAAAAACAGACTGAGCACGT AGCTGATGGCAGAGGAAAGGAAGCACC GACTCGGTGCC	TAGC	13	19	HEK site 3
CGCACCAAAAAAACAGACTGAGCACGT TGCTGATGGCAGAGGAAAGCACCGACT CGGTGCC	TTGC	13	15	HEK site 3
CGCACCAAAAAAACAGACTGAGCACGC GTTTGATGGCAGAGCACCGACTCGGTGC C	CGTT	13	10	HEK site 3
CGCACCAAAAAAACAGACTGAGCACGC AAGTGATGGCGCACCGACTCGGTGCC	CAAG	13	7	HEK site 3
CGCACCAAAAAAAAGACTGAGCACGCA ACTGATGGCAGAGGAAAGGAAGCCCTG CTTCCGCACCGACTCGGTGCC	CAAC	12	30	HEK site 3
CGCACCAAAAAAAAGACTGAGCACGGC TATGATGGCAGAGGAAAGGAAGCCCTG CACCGACTCGGTGCC	GCTA	12	24	HEK site 3
CGCACCAAAAAAAAGACTGAGCACGGA CTTGATGGCAGAGGAAAGGAAGCACCG ACTCGGTGCC	GACT	12	19	HEK site 3
CGCACCAAAAAAAAGACTGAGCACGCT GTTGATGGCAGAGGAAAGCACCGACTC GGTGCC	CTGT	12	15	HEK site 3
CGCACCAAAAAAAAGACTGAGCACGTC TCTGATGGCAGAGCACCGACTCGGTGCC	TCTC	12	10	HEK site 3

CGCACCAAAAAAAAGACTGAGCACGAC ACTGATGGCGCACCGACTCGGTGCC	ACAC	12	7	HEK site 3
CGCACCAAAAAAAGACTGAGCACGAAG CTGATGGCAGAGGAAAGGAAGCCCTGC TTCCGCACCGACTCGGTGCC	AAGC	11	30	HEK site 3
CGCACCAAAAAAAGACTGAGCACGTTC GTGATGGCAGAGGAAAGGAAGCCCTGC ACCGACTCGGTGCC	TTCG	11	24	HEK site 3
CGCACCAAAAAAAGACTGAGCACGTCT GTGATGGCAGAGGAAAGGAAGCACCGA CTCGGTGCC	TCTG	11	19	HEK site 3
CGCACCAAAAAAAGACTGAGCACGTCG ATGATGGCAGAGGAAAGCACCGACTCG GTGCC	TCGA	11	15	HEK site 3
CGCACCAAAAAAAGACTGAGCACGCAG TTGATGGCAGAGCACCGACTCGGTGCC	CAGT	11	10	HEK site 3
CGCACCAAAAAAAGACTGAGCACGAAC CTGATGGCGCACCGACTCGGTGCC	AACC	11	7	HEK site 3
CGCACCAAAAAAAAACTGAGCACGCTTGT GATGGCAGAGGAAAGGAAGCCCTGCTT CCGCACCGACTCGGTGCC	CTTG	10	30	HEK site 3
CGCACCAAAAAAAAACTGAGCACGGTAC TGATGGCAGAGGAAAGGAAGCCCTGCA CCGACTCGGTGCC	GTAC	10	24	HEK site 3
CGCACCAAAAAAAAACTGAGCACGCATG TGATGGCAGAGGAAAGGAAGCACCGAC TCGGTGCC	CATG	10	19	HEK site 3
CGCACCAAAAAAAAACTGAGCACGCTCA TGATGGCAGAGGAAAGCACCGACTCGG TGCC	CTCA	10	15	HEK site 3
CGCACCAAAAAAAAACTGAGCACGCGAA TGATGGCAGAGCACCGACTCGGTGCC	CGAA	10	10	HEK site 3
CGCACCAAAAAAAAACTGAGCACGCTGA TGATGGCGCACCGACTCGGTGCC	CTGA	10	7	HEK site 3
CGCACCAAAAAAACTGAGCACGGCTTTG ATGGCAGAGGAAAGGAAGCCCTGCTTC CGCACCGACTCGGTGCC	GCTT	9	30	HEK site 3

CGCACCAAAAAAACTGAGCACGTCCTTG ATGGCAGAGGAAAGGAAGCCCTGCACC GACTCGGTGCC	TCCT	9	24	HEK site 3
CGCACCAAAAAAACTGAGCACGCCAAT GATGGCAGAGGAAAGGAAGCACCGACT CGGTGCC	CCAA	9	19	HEK site 3
CGCACCAAAAAAACTGAGCACGACGAT GATGGCAGAGGAAAGCACCGACTCGGT GCC	ACGA	9	15	HEK site 3
CGCACCAAAAAAACTGAGCACGGCAAT GATGGCAGAGCACCGACTCGGTGCC	GCAA	9	10	HEK site 3
CGCACCAAAAAAACTGAGCACGCACTT GATGGCGCACCGACTCGGTGCC	CACT	9	7	HEK site 3
CGCACCAAAAAAATGAGCACGGAACTG ATGGCAGAGGAAAGGAAGCCCTGCTTC CGCACCGACTCGGTGCC	GAAC	8	30	HEK site 3
CGCACCAAAAAAATGAGCACGAGTCTG ATGGCAGAGGAAAGGAAGCCCTGCACC GACTCGGTGCC	AGTC	8	24	HEK site 3
CGCACCAAAAAAATGAGCACGAGCATG ATGGCAGAGGAAAGGAAGCACCGACTC GGTGCC	AGCA	8	19	HEK site 3
CGCACCAAAAAAATGAGCACGGTTGTG ATGGCAGAGGAAAGCACCGACTCGGTG CC	GTTG	8	15	HEK site 3
CGCACCAAAAAAATGAGCACGAGAGTG ATGGCAGAGCACCGACTCGGTGCC	AGAG	8	10	HEK site 3
CGCACCAAAAAAATGAGCACGTCCATG ATGGCGCACCGACTCGGTGCC	TCCA	8	7	HEK site 3
CGCACCAAAAAAAGAGCACGGTGATGA TGGCAGAGGAAAGGAAGCCCTGCTTCC GCACCGACTCGGTGCC	GTGA	7	30	HEK site 3
CGCACCAAAAAAAGAGCACGACAGTGA TGGCAGAGGAAAGGAAGCCCTGCACCG ACTCGGTGCC	ACAG	7	24	HEK site 3

CGCACCAAAAAAAGAGCACGTCACTGATGGCAGAGGAAAGGAAGCACCGACTCGGTGCC	TCAC	7	19	HEK site 3
CGCACCAAAAAAAGAGCACGACTGTGATGGCAGAGGAAAGCACCGACTCGGTGCCC	ACTG	7	15	HEK site 3
CGCACCAAAAAAAGAGCACGGAAGTGATGGCAGAGCACCGACTCGGTGCC	GAAG	7	10	HEK site 3
CGCACCAAAAAAAGAGCACGCCTATGATGGCGCACCGACTCGGTGCC	CCTA	7	7	HEK site 3
CGCACCAAAAAAAAAGCACGCCTTTGATGGCAGAGGAAAGGAAGCCCTGCTTCCGCACCGACTCGGTGCC	CCTT	6	30	HEK site 3
CGCACCAAAAAAAAAGCACGTCAGTGATGGCAGAGGAAAGGAAGCCCTGCACCGACTCGGTGCC	TCAG	6	24	HEK site 3
CGCACCAAAAAAAAAGCACGTCGTTGATGGCAGAGGAAAGGAAGCACCGACTCGGTGCC	TCGT	6	19	HEK site 3
CGCACCAAAAAAAAAGCACGCATCTGATGGCAGAGGAAAGCACCGACTCGGTGCC	CATC	6	15	HEK site 3
CGCACCAAAAAAAAAGCACGAGACTGATGGCAGAGCACCGACTCGGTGCC	AGAC	6	10	HEK site 3
CGCACCAAAAAAAAAGCACGCTAGTGATGGCGCACCGACTCGGTGCC	CTAG	6	7	HEK site 3
CAAAAAAAGAGTCCGAGCAGAAGAAACCAGGCTCCCATCACATCAACCGGTGGCGCATTGCCACGCACCGACTCGGTGCC	ACCA	17	39	EMX1 site 1
CAAAAAAAGAGTCCGAGCAGAAGAAAACGGGCTCCCATCACATCAACCGGTGGCGCATTGCGCACCGACTCGGTGCC	AACG	17	36	EMX1 site 1
CAAAAAAAGAGTCCGAGCAGAAGAAGAGAGGCTCCCATCACATCAACCGGTGGCGCGCACCGACTCGGTGCC	GAGA	17	31	EMX1 site 1
CAAAAAAAGAGTCCGAGCAGAAGAAGTCAGGCTCCCATCACATCAACCGGTGCACCGACTCGGTGCC	GTCA	17	26	EMX1 site 1

CAAAAAAAGAGTCCGAGCAGAAGAAAG CTGGCTCCCATCACATCAGCACCGACTC GGTGCC	AGCT	17	20	EMX1 site 1
CAAAAAAAGAGTCCGAGCAGAAGAAGA TCGGCTCCCATCAGCACCGACTCGGTGC C	GATC	17	15	EMX1 site 1
CAAAAAAAGAGTCCGAGCAGAAGAAAC GTGGCTCCGCACCGACTCGGTGCC	ACGT	17	10	EMX1 site 1
CAAAAAAAAGTCCGAGCAGAAGAAGTG TGGCTCCCATCACATCAACCGGTGGCGC ATTGCCACGCACCGACTCGGTGCC	GTGT	16	39	EMX1 site 1
CAAAAAAAAGTCCGAGCAGAAGAAATC GGGCTCCCATCACATCAACCGGTGGCGC ATTGCGCACCGACTCGGTGCC	ATCG	16	36	EMX1 site 1
CAAAAAAAAGTCCGAGCAGAAGAACGT AGGCTCCCATCACATCAACCGGTGGCGC GCACCGACTCGGTGCC	CGTA	16	31	EMX1 site 1
CAAAAAAAAGTCCGAGCAGAAGAAACC TGGCTCCCATCACATCAACCGGTGCACC GACTCGGTGCC	ACCT	16	26	EMX1 site 1
CAAAAAAAAGTCCGAGCAGAAGAAGAG TGGCTCCCATCACATCAGCACCGACTCG GTGCC	GAGT	16	20	EMX1 site 1
CAAAAAAAAGTCCGAGCAGAAGAAATC CGGCTCCCATCAGCACCGACTCGGTGCC	ATCC	16	15	EMX1 site 1
CAAAAAAAAGTCCGAGCAGAAGAACTA GGGCTCCGCACCGACTCGGTGCC	CTAG	16	10	EMX1 site 1
CAAAAAAAGTCCGAGCAGAAGAAGTCT GGCTCCCATCACATCAACCGGTGGCGCA TTGCCACGCACCGACTCGGTGCC	GTCT	15	39	EMX1 site 1
CAAAAAAAGTCCGAGCAGAAGAACTCT GGCTCCCATCACATCAACCGGTGGCGCA TTGCGCACCGACTCGGTGCC	CTCT	15	36	EMX1 site 1
CAAAAAAAGTCCGAGCAGAAGAACTAC GGCTCCCATCACATCAACCGGTGGCGCG CACCGACTCGGTGCC	CTAC	15	31	EMX1 site 1

CAAAAAAAGTCCGAGCAGAAGAAATGC GGCTCCCATCACATCAACCGGTGCACCG ACTCGGTGCC	ATGC	15	26	EMX1 site 1
CAAAAAAAGTCCGAGCAGAAGAAAGTG GGCTCCCATCACATCAGCACCGACTCGG TGCC	AGTG	15	20	EMX1 site 1
CAAAAAAAGTCCGAGCAGAAGAACAGA GGCTCCCATCAGCACCGACTCGGTGCC	CAGA	15	15	EMX1 site 1
CAAAAAAAGTCCGAGCAGAAGAAGATG GGCTCCGCACCGACTCGGTGCC	GATG	15	10	EMX1 site 1
CAAAAAAATCCGAGCAGAAGAATTCCG GCTCCCATCACATCAACCGGTGGCGCAT TGCCACGCACCGACTCGGTGCC	TTCC	14	39	EMX1 site 1
CAAAAAAATCCGAGCAGAAGAAGACAG GCTCCCATCACATCAACCGGTGGCGCAT TGCGCACCGACTCGGTGCC	GACA	14	36	EMX1 site 1
CAAAAAAATCCGAGCAGAAGAAGTAGG GCTCCCATCACATCAACCGGTGGCGCGC ACCGACTCGGTGCC	GTAG	14	31	EMX1 site 1
CAAAAAAATCCGAGCAGAAGAAACTCG GCTCCCATCACATCAACCGGTGCACCGA CTCGGTGCC	ACTC	14	26	EMX1 site 1
CAAAAAAATCCGAGCAGAAGAATACGG GCTCCCATCACATCAGCACCGACTCGGT GCC	TACG	14	20	EMX1 site 1
CAAAAAAATCCGAGCAGAAGAATAGCG GCTCCCATCAGCACCGACTCGGTGCC	TAGC	14	15	EMX1 site 1
CAAAAAAATCCGAGCAGAAGAATTGCG GCTCCGCACCGACTCGGTGCC	TTGC	14	10	EMX1 site 1
CAAAAAAACCGAGCAGAAGAACGTTGG CTCCCATCACATCAACCGGTGGCGCATT GCCACGCACCGACTCGGTGCC	CGTT	13	39	EMX1 site 1
CAAAAAAACCGAGCAGAAGAACAAGGG CTCCCATCACATCAACCGGTGGCGCATT GCGCACCGACTCGGTGCC	CAAG	13	36	EMX1 site 1

CAAAAAAACCGAGCAGAAGAACAACGGCTCCCATCACATCAACCGGTGGCGCGCACCGACTCGGTGCC	CAAC	13	31	EMX1 site 1
CAAAAAAACCGAGCAGAAGAAGCTAGGCTCCCATCACATCAACCGGTGCACCGACTCGGTGCC	GCTA	13	26	EMX1 site 1
CAAAAAAACCGAGCAGAAGAAGACTGGCTCCCATCACATCAGCACCGACTCGGTGCC	GACT	13	20	EMX1 site 1
CAAAAAAACCGAGCAGAAGAACTGTGGCTCCCATCAGCACCGACTCGGTGCC	CTGT	13	15	EMX1 site 1
CAAAAAAACCGAGCAGAAGAATCTCGGCTCCGCACCGACTCGGTGCC	TCTC	13	10	EMX1 site 1
CAAAAAAACGAGCAGAAGAAACACGGCTCCCATCACATCAACCGGTGGCGCATTGCCACGCACCGACTCGGTGCC	ACAC	12	39	EMX1 site 1
CAAAAAAACGAGCAGAAGAAAAGCGGCTCCCATCACATCAACCGGTGGCGCATTGCGCACCGACTCGGTGCC	AAGC	12	36	EMX1 site 1
CAAAAAAACGAGCAGAAGAATTCGGGCTCCCATCACATCAACCGGTGGCGCGCACCGACTCGGTGCC	TTCG	12	31	EMX1 site 1
CAAAAAAACGAGCAGAAGAATCTGGGCTCCCATCACATCAACCGGTGCACCGACTCGGTGCC	TCTG	12	26	EMX1 site 1
CAAAAAAACGAGCAGAAGAATCGAGGCTCCCATCACATCAGCACCGACTCGGTGCC	TCGA	12	20	EMX1 site 1
CAAAAAAACGAGCAGAAGAACAGTGGCTCCCATCAGCACCGACTCGGTGCC	CAGT	12	15	EMX1 site 1
CAAAAAAACGAGCAGAAGAAAACCGGCTCCGCACCGACTCGGTGCC	AACC	12	10	EMX1 site 1
CAAAAAAAGAGCAGAAGAAAGACGGCTCCCATCACATCAACCGGTGGCGCATTGCCACGCACCGACTCGGTGCC	AGAC	11	39	EMX1 site 1

CAAAAAAAGAGCAGAAGAAGTACGGCTCCCATCACATCAACCGGTGGCGCATTGCGCACCGACTCGGTGCC	GTAC	11	36	EMX1 site 1
CAAAAAAAGAGCAGAAGAACATGGGCTCCCATCACATCAACCGGTGGCGCGCACCGACTCGGTGCC	CATG	11	31	EMX1 site 1
CAAAAAAAGAGCAGAAGAACTCAGGCTCCCATCACATCAACCGGTGCACCGACTCGGTGCC	CTCA	11	26	EMX1 site 1
CAAAAAAAGAGCAGAAGAACGAAGGCTCCCATCACATCAGCACCGACTCGGTGCC	CGAA	11	20	EMX1 site 1
CAAAAAAAGAGCAGAAGAACTGAGGCTCCCATCAGCACCGACTCGGTGCC	CTGA	11	15	EMX1 site 1
CAAAAAAAGAGCAGAAGAAGCTTGGCTCCGCACCGACTCGGTGCC	GCTT	11	10	EMX1 site 1
CAAAAAAAAGCAGAAGAATCCTGGCTCCCATCACATCAACCGGTGGCGCATTGCCACGCACCGACTCGGTGCC	TCCT	10	39	EMX1 site 1
CAAAAAAAAGCAGAAGAACCAAGGCTCCCATCACATCAACCGGTGGCGCATTGCGCACCGACTCGGTGCC	CCAA	10	36	EMX1 site 1
CAAAAAAAAGCAGAAGAAACGAGGCTCCCATCACATCAACCGGTGGCGCGCACCGACTCGGTGCC	ACGA	10	31	EMX1 site 1
CAAAAAAAAGCAGAAGAAGCAAGGCTCCCATCACATCAACCGGTGCACCGACTCGGTGCC	GCAA	10	26	EMX1 site 1
CAAAAAAAAGCAGAAGAACACTGGCTCCCATCACATCAGCACCGACTCGGTGCC	CACT	10	20	EMX1 site 1
CAAAAAAAAGCAGAAGAAGAACGGCTCCCATCAGCACCGACTCGGTGCC	GAAC	10	15	EMX1 site 1
CAAAAAAAAGCAGAAGAAAGTCGGCTCCGCACCGACTCGGTGCC	AGTC	10	10	EMX1 site 1
CAAAAAAAGCAGAAGAAAGCAGGCTCCCATCACATCAACCGGTGGCGCATTGCCACGCACCGACTCGGTGCC	AGCA	9	39	EMX1 site 1

CAAAAAAAGCAGAAGAAGTTGGGCTCC CATCACATCAACCGGTGGCGCATTGCGC ACCGACTCGGTGCC	GTTG	9	36	EMX1 site 1
CAAAAAAAGCAGAAGAAAGAGGGCTCC CATCACATCAACCGGTGGCGCGCACCGA CTCGGTGCC	AGAG	9	31	EMX1 site 1
CAAAAAAAGCAGAAGAATCCAGGCTCC CATCACATCAACCGGTGCACCGACTCGG TGCC	TCCA	9	26	EMX1 site 1
CAAAAAAAGCAGAAGAAGTGAGGCTCC CATCACATCAGCACCGACTCGGTGCC	GTGA	9	20	EMX1 site 1
CAAAAAAAGCAGAAGAAACAGGGCTCC CATCAGCACCGACTCGGTGCC	ACAG	9	15	EMX1 site 1
CAAAAAAAGCAGAAGAATCACGGCTCC GCACCGACTCGGTGCC	TCAC	9	10	EMX1 site 1
CAAAAAAACAGAAGAAACTGGGCTCCC ATCACATCAACCGGTGGCGCATTGCCAC GCACCGACTCGGTGCC	ACTG	8	39	EMX1 site 1
CAAAAAAACAGAAGAAGAAGGGCTCCC ATCACATCAACCGGTGGCGCATTGCGCA CCGACTCGGTGCC	GAAG	8	36	EMX1 site 1
CAAAAAAACAGAAGAACCTAGGCTCCC ATCACATCAACCGGTGGCGCGCACCGAC TCGGTGCC	CCTA	8	31	EMX1 site 1
CAAAAAAACAGAAGAACCTTGGCTCCC ATCACATCAACCGGTGCACCGACTCGGT GCC	CCTT	8	26	EMX1 site 1
CAAAAAAACAGAAGAATCAGGGCTCCC ATCACATCAGCACCGACTCGGTGCC	TCAG	8	20	EMX1 site 1
CAAAAAAACAGAAGAATCGTGGCTCCC ATCAGCACCGACTCGGTGCC	TCGT	8	15	EMX1 site 1
CAAAAAAACAGAAGAACATCGGCTCCG CACCGACTCGGTGCC	CATC	8	10	EMX1 site 1
CGCACCAAAAAAAGTCATCTTAGTCATT ACACTCCTGAGGTGTTCGTTGTAACTCA TATAAACTGCACCGACTCGGTGCC	ACTC	17	30	RNF2 site 1

CGCACCAAAAAAAGTCATCTTAGTCATTACGTGTCTGAGGTGTTCGTTGTAACTCATATGCACCGACTCGGTGCC	GTGT	17	25	RNF2 site 1
CGCACCAAAAAAAGTCATCTTAGTCATTACTCACCTGAGGTGTTCGTTGTAACTGCACCGACTCGGTGCC	TCAC	17	20	RNF2 site 1
CGCACCAAAAAAAGTCATCTTAGTCATTACTACGCTGAGGTGTTCGTTGCACCGACTCGGTGCC	TACG	17	14	RNF2 site 1
CGCACCAAAAAAAGTCATCTTAGTCATTACGACACTGAGGTGTTGCACCGACTCGGTGCC	GACA	17	10	RNF2 site 1
CGCACCAAAAAAAGTCATCTTAGTCATTACAGAGCTGAGGTGCACCGACTCGGTGCC	AGAG	17	7	RNF2 site 1
CGCACCAAAAAAATCATCTTAGTCATTACTAGGCTGAGGTGTTCGTTGTAACTCATATAAACTGCACCGACTCGGTGCC	TAGG	16	30	RNF2 site 1
CGCACCAAAAAAATCATCTTAGTCATTACGTTGCTGAGGTGTTCGTTGTAACTCATATGCACCGACTCGGTGCC	GTTG	16	25	RNF2 site 1
CGCACCAAAAAAATCATCTTAGTCATTACTCGACTGAGGTGTTCGTTGTAACTGCACCGACTCGGTGCC	TCGA	16	20	RNF2 site 1
CGCACCAAAAAAATCATCTTAGTCATTACCACACTGAGGTGTTCGTTGCACCGACTCGGTGCC	CACA	16	14	RNF2 site 1
CGCACCAAAAAAATCATCTTAGTCATTACACAGCTGAGGTGTTGCACCGACTCGGTGCC	ACAG	16	10	RNF2 site 1
CGCACCAAAAAAATCATCTTAGTCATTACGGAACTGAGGTGCACCGACTCGGTGCC	GGAA	16	7	RNF2 site 1
CGCACCAAAAAAACATCTTAGTCATTACAGTCCTGAGGTGTTCGTTGTAACTCATATAAACTGCACCGACTCGGTGCC	AGTC	15	30	RNF2 site 1

CGCACCAAAAAAACATCTTAGTCATTACTGCACTGAGGTGTTCGTTGTAACTCATATGCACCGACTCGGTGCC	TGCA	15	25	RNF2 site 1
CGCACCAAAAAAACATCTTAGTCATTACTCGTCTGAGGTGTTCGTTGTAACTGCACCGACTCGGTGCC	TCGT	15	20	RNF2 site 1
CGCACCAAAAAAACATCTTAGTCATTACGACTCTGAGGTGTTCGTTGCACCGACTCGGTGCC	GACT	15	14	RNF2 site 1
CGCACCAAAAAAACATCTTAGTCATTACATCCCTGAGGTGTTGCACCGACTCGGTGCC	ATCC	15	10	RNF2 site 1
CGCACCAAAAAAACATCTTAGTCATTACACCACTGAGGTGCACCGACTCGGTGCC	ACCA	15	7	RNF2 site 1
CGCACCAAAAAAAATCTTAGTCATTACAAGGCTGAGGTGTTCGTTGTAACTCATATAAACTGCACCGACTCGGTGCC	AAGG	14	30	RNF2 site 1
CGCACCAAAAAAAATCTTAGTCATTACAGACCTGAGGTGTTCGTTGTAACTCATATGCACCGACTCGGTGCC	AGAC	14	25	RNF2 site 1
CGCACCAAAAAAAATCTTAGTCATTACGATCCTGAGGTGTTCGTTGTAACTGCACCGACTCGGTGCC	GATC	14	20	RNF2 site 1
CGCACCAAAAAAAATCTTAGTCATTACCTCTCTGAGGTGTTCGTTGCACCGACTCGGTGCC	CTCT	14	14	RNF2 site 1
CGCACCAAAAAAAATCTTAGTCATTACGCAACTGAGGTGTTGCACCGACTCGGTGCC	GCAA	14	10	RNF2 site 1
CGCACCAAAAAAAATCTTAGTCATTACAGCTCTGAGGTGCACCGACTCGGTGCC	AGCT	14	7	RNF2 site 1
CGCACCAAAAAAATCTTAGTCATTACTGGACTGAGGTGTTCGTTGTAACTCATATAAACTGCACCGACTCGGTGCC	TGGA	13	30	RNF2 site 1
CGCACCAAAAAAATCTTAGTCATTACAACCCTGAGGTGTTCGTTGTAACTCATATGCACCGACTCGGTGCC	AACC	13	25	RNF2 site 1

CGCACCAAAAAAATCTTAGTCATTACTG TGCTGAGGTGTTCGTTGTAACTGCACCG ACTCGGTGCC	TGTG	13	20	RNF2 site 1
CGCACCAAAAAAATCTTAGTCATTACCT TCCTGAGGTGTTCGTTGCACCGACTCGG TGCC	CTTC	13	14	RNF2 site 1
CGCACCAAAAAAATCTTAGTCATTACCA GACTGAGGTGTTGCACCGACTCGGTGCC	CAGA	13	10	RNF2 site 1
CGCACCAAAAAAATCTTAGTCATTACGA GACTGAGGTGCACCGACTCGGTGCC	GAGA	13	7	RNF2 site 1
CGCACCAAAAAAACTTAGTCATTACTGC TCTGAGGTGTTCGTTGTAACTCATATAA ACTGCACCGACTCGGTGCC	TGCT	12	30	RNF2 site 1
CGCACCAAAAAAACTTAGTCATTACCAA CCTGAGGTGTTCGTTGTAACTCATATGC ACCGACTCGGTGCC	CAAC	12	25	RNF2 site 1
CGCACCAAAAAAACTTAGTCATTACGAT GCTGAGGTGTTCGTTGTAACTGCACCGA CTCGGTGCC	GATG	12	20	RNF2 site 1
CGCACCAAAAAAACTTAGTCATTACGTG ACTGAGGTGTTCGTTGCACCGACTCGGT GCC	GTGA	12	14	RNF2 site 1
CGCACCAAAAAAACTTAGTCATTACTTG CCTGAGGTGTTGCACCGACTCGGTGCC	TTGC	12	10	RNF2 site 1
CGCACCAAAAAAACTTAGTCATTACTTC CCTGAGGTGCACCGACTCGGTGCC	TTCC	12	7	RNF2 site 1
CGCACCAAAAAAATTAGTCATTACTGAC CTGAGGTGTTCGTTGTAACTCATATAAA CTGCACCGACTCGGTGCC	TGAC	11	30	RNF2 site 1
CGCACCAAAAAAATTAGTCATTACCAGT CTGAGGTGTTCGTTGTAACTCATATGCA CCGACTCGGTGCC	CAGT	11	25	RNF2 site 1
CGCACCAAAAAAATTAGTCATTACCTGT CTGAGGTGTTCGTTGTAACTGCACCGAC TCGGTGCC	CTGT	11	20	RNF2 site 1

CGCACCAAAAAAATTAGTCATTACCTTG CTGAGGTGTTCGTTGCACCGACTCGGTG CC	CTTG	11	14	RNF2 site 1
CGCACCAAAAAAATTAGTCATTACTGTC CTGAGGTGTTGCACCGACTCGGTGCC	TGTC	11	10	RNF2 site 1
CGCACCAAAAAAATTAGTCATTACGTAC CTGAGGTGCACCGACTCGGTGCC	GTAC	11	7	RNF2 site 1
CGCACCAAAAAAATAGTCATTACCGTTC TGAGGTGTTCGTTGTAACTCATATAAAC TGCACCGACTCGGTGCC	CGTT	10	30	RNF2 site 1
CGCACCAAAAAAATAGTCATTACGCATC TGAGGTGTTCGTTGTAACTCATATGCAC CGACTCGGTGCC	GCAT	10	25	RNF2 site 1
CGCACCAAAAAAATAGTCATTACGAGTC TGAGGTGTTCGTTGTAACTGCACCGACT CGGTGCC	GAGT	10	20	RNF2 site 1
CGCACCAAAAAAATAGTCATTACCCTTC TGAGGTGTTCGTTGCACCGACTCGGTGC C	CCTT	10	14	RNF2 site 1
CGCACCAAAAAAATAGTCATTACAGTGC TGAGGTGTTGCACCGACTCGGTGCC	AGTG	10	10	RNF2 site 1
CGCACCAAAAAAATAGTCATTACGAACC TGAGGTGCACCGACTCGGTGCC	GAAC	10	7	RNF2 site 1
CGCACCAAAAAAAAGTCATTACGGTTCT GAGGTGTTCGTTGTAACTCATATAAACT GCACCGACTCGGTGCC	GGTT	9	30	RNF2 site 1
CGCACCAAAAAAAAGTCATTACATGCCT GAGGTGTTCGTTGTAACTCATATGCACC GACTCGGTGCC	ATGC	9	25	RNF2 site 1
CGCACCAAAAAAAAGTCATTACATCGCT GAGGTGTTCGTTGTAACTGCACCGACTC GGTGCC	ATCG	9	20	RNF2 site 1
CGCACCAAAAAAAAGTCATTACGTCTCT GAGGTGTTCGTTGCACCGACTCGGTGCC	GTCT	9	14	RNF2 site 1
CGCACCAAAAAAAAGTCATTACACTGCT GAGGTGTTGCACCGACTCGGTGCC	ACTG	9	10	RNF2 site 1

CGCACCAAAAAAAAGTCATTACTCCACT GAGGTGCACCGACTCGGTGCC	TCCA	9	7	RNF2 site 1
CGCACCAAAAAAAGTCATTACTTCGCTG AGGTGTTCGTTGTAACTCATATAAACTG CACCGACTCGGTGCC	TTCG	8	30	RNF2 site 1
CGCACCAAAAAAAGTCATTACGTAGCTG AGGTGTTCGTTGTAACTCATATGCACCG ACTCGGTGCC	GTAG	8	25	RNF2 site 1
CGCACCAAAAAAAGTCATTACAACGCTG AGGTGTTCGTTGTAACTGCACCGACTCG GTGCC	AACG	8	20	RNF2 site 1
CGCACCAAAAAAAGTCATTACCGATCTG AGGTGTTCGTTGCACCGACTCGGTGCC	CGAT	8	14	RNF2 site 1
CGCACCAAAAAAAGTCATTACACGACTG AGGTGTTGCACCGACTCGGTGCC	ACGA	8	10	RNF2 site 1
CGCACCAAAAAAAGTCATTACACACCTG AGGTGCACCGACTCGGTGCC	ACAC	8	7	RNF2 site 1
CGCACCAAAAAAATCATTACAGGACTG AGGTGTTCGTTGTAACTCATATAAACTG CACCGACTCGGTGCC	AGGA	7	30	RNF2 site 1
CGCACCAAAAAAATCATTACGTCACTGA GGTGTTCGTTGTAACTCATATGCACCGA CTCGGTGCC	GTCA	7	25	RNF2 site 1
CGCACCAAAAAAATCATTACCACTCTGA GGTGTTCGTTGTAACTGCACCGACTCGG TGCC	CACT	7	20	RNF2 site 1
CGCACCAAAAAAATCATTACAGCACTGA GGTGTTCGTTGCACCGACTCGGTGCC	AGCA	7	14	RNF2 site 1
CGCACCAAAAAAATCATTACATGGCTGA GGTGTTGCACCGACTCGGTGCC	ATGG	7	10	RNF2 site 1
CGCACCAAAAAAATCATTACCTACCTGA GGTGCACCGACTCGGTGCC	CTAC	7	7	RNF2 site 1
CGCACCAAAAAAACATTACGTTCCTGAG GTGTTCGTTGTAACTCATATAAACTGCA CCGACTCGGTGCC	GTTC	6	30	RNF2 site 1

CGCACCAAAAAAACATTACCATCCTGAG GTGTTCGTTGTAACTCATATGCACCGAC TCGGTGCC	CATC	6	25	RNF2 site 1
CGCACCAAAAAAACATTACCGAACTGA GGTGTTCGTTGTAACTGCACCGACTCGG TGCC	CGAA	6	20	RNF2 site 1
CGCACCAAAAAAACATTACCCAACTGA GGTGTTCGTTGCACCGACTCGGTGCC	CCAA	6	14	RNF2 site 1
CGCACCAAAAAAACATTACCTAGCTGAG GTGTTGCACCGACTCGGTGCC	CTAG	6	10	RNF2 site 1
CGCACCAAAAAAACATTACTAGCCTGAG GTGCACCGACTCGGTGCC	TAGC	6	7	RNF2 site 1

Supplementary Table 4.7: Target sites for high-throughput pooled pegRNA optimization.

Target name	Spacer sequence	PBS max	RTT max
HEK site 1	GGGAAAGACCCA GCATCCGT	GATGCTGGGTCTT TCCC	CGATTTCCCACA GCTTTTCAGCGAC CCACG
HEK site 2	GAACACAAAGCA TAGACTGC	GTCTATGCTTTGT GTTC	TTGCAGCTATTCA GGCTGGCCCGCC CCGCA
HEK site 3	GGCCCAGACTGA GCACGTGA	CGTGCTCAGTCTG GGCC	GGAAGCAGGGCT TCCTTTCCTCTGC CATCA
HEK site 4	GGCACTGCGGCT GGAGGTGG	CCTCCAGCCGCA GTGCC	ACAGCACCAGAG TCTCCGCTTTAAC CCCCA
VEGFA site 1	GGGTGGGGGGAG TTTGCTCC	GCAAACTCCCCC CACCC	TTTGGGAGGTCA GAAATAGGGGGT CCAGGA
VEGFA site 2	GACCCCCTCCAC CCCGCCTC	GCGGGGTGGAGG GGGTC	GCGGGCAGGGGC CGGAGCCCGCGC CCGGAG
VEGFA site 3	GGTGAGTGAGTG TGTGCGTG	GCACACACTCAC TCACC	CTCCCCGCTCCAA CGCCCTCAACCC CACAC

VEGFA site 4	GATGTCTGCAGG CCAGATGA	TCTGGCCTGCAG ACATC	CCTCTGACAATGT GCCATCTGGAGC CCTCA
VEGFA site 5	GCTGAGGTGCAG AATCCAGG	GGATTCTGCACCT CAGC	TTACCACTGCGG CTCCTGCAGGGA CCCCCT
EMX1 site 1	GAGTCCGAGCAG AAGAAGAA	TTCTTCTGCTCGG ACTC	CGCCACCGGTTG ATGTGATGGGAG CCCTTC
EMX1 site 2	GTCACCTCCAAT GACTAGGG	TAGTCATTGGAG GTGAC	CTGCCCTCGTGG GTTTGTGGTTGCC CACCC
EMX1 site 3	GGGAAGACTGAG GCTACATA	GTAGCCTCAGTCT TCCC	GGGACCCCGGCC TGGGGCCCCTAA CCCTAT
EMX1 site 4	GGCCCCAGTGGC TGCTCTGG	GAGCAGCCACTG GGGCC	GGCACAGATGAG AAACTCAGGAGG CCCCCA
FANCF site 1	GGAATCCCTTCTG CAGCACC	GCTGCAGAAGGG ATTCC	CCGCCAGAAGCT CGGAAAAGCGAT CCAGGT
FANCF site 2	GCTGCAGAAGGG ATTCCATG	GGAATCCCTTCTG CAGC	AAGCGGAAGTAG GGCCTTCGCGCA CCTCAT
FANCF site 3	GGCGGCTGCACA ACCAGTGG	CTGGTTGTGCAG CCGCC	ACCCCGCCCAAA GCCGCCCTCTTGC CTCCA
FANCF site 4	GCTCCAGAGCCG TGCGAATG	TCGCACGGCTCT GGAGC	CCATCGGCGCTTT GGTCGGCATGGC CCCAT
FANCF site 5	GGACTCTCTGAT GAAGACCC	TCTTCATCAGAG AGTCC	GCAGACGCTCCA GCAGCAGCTCCG CCTGGG
RNF2 site 1	GTCATCTTAGTCA TTACCTG	GTAATGACTAAG ATGAC	AGTTTATATGAGT TACAACGAACAC CTCAG

RUNX1 site 1	GCATTTTCAGGAGGAAGCGA	CTTCCTCCTGAAAATGC	ATGACTCAAATATGCTGTCTGAAGCCATCG
RUNX1 site 2	GGGAGAAGAAAGAGAGATGT	TCTCTCTTTCTTCTCCC	CTGTTTCAGCCTCACCCCTCTAGCCCTACA
ZSCAN2 site 1	GTGCGGCAAGAGCTTCAGCC	TGAAGCTCTTGCCGCAC	TCTGGTGCATGACCAGAATGGAGCCCCGGC
DNMT1 site 1	GTCACTCTGGGGAACACGCC	GTGTTCCCCAGAGTGAC	ACTCGCCTGTCAAGTGGCGTGACACCGGGC
DNMT1 site 2	GAGTGCTAAGGGAACGTTCA	ACGTTCCCTTAGCACTC	GACCGTTTGAGGAGTGTTCAGTCTCCGTGA
DNMT1 site 3	GAGACTGAACACTCCTCAAA	GAGGAGTGTTCAGTCTC	TCTGGGTCTAGAACCCTCTGGGGACCGTTT
DNMT1 site 4	GGAGTGAGGGAAACGGCCCC	GCCGTTTCCCTCACTCC	GGTCAGGTTGGCTGCTGGGCTGGCCCTGGG
matched site 1	GATTGAAGGAAAAGTTACAA	TAACTTTTCCTTCAATC	GTTATAAAGAAACTCTTTGTGCTACCTTTG
matched site 2	GGCAAATAGGAATGGCAAGA	TGCCATTCCTATTTGCC	GCGTCATTTATGCAGAAAATGCATCCCTCT
matched site 3	GAGCTGCTTAAGCATTTCAA	AAATGCTTAAGCAGCTC	AATTTCAGAGTTTCAGGGTTTCTTCCCTTG
matched site 4	GCCTAAAAACATCAATAGAA	TATTGATGTTTTTAGGC	GGACAGATTTGCAGATGACACTTGCCTTTC
matched site 5	GGATGCCACTAAAAGGGAAA	CCCTTTTAGTGGCATCC	GACCCCAGAATCAGTAAAGTAATCCCCTTT

matched site 6	GGGTGATCAGAC CCAACAGC	GTTGGGTCTGATC ACCC	CACTGGACTTCGT CGCCCCCATGAC CTGCT
matched site 7	GAGAGGGTTCCT GGGTTTAA	AACCCAGGAACC CTCTC	GCTTTGTATCTCA AGAGTTGTCACC CCTTA
matched site 8	GGGACGGGGAGA AGGAAAAG	TTCCTTCTCCCCG TCCC	TCCCTCCCTCCCT CCTTTCCTCCCCC TCTT
matched site 9	GGCATGAATTAT AATGCTGT	GCATTATAATTCA TGCC	ATAGTTGTTTACA TTGAGCTCCAGC CAACA
matched site 10	GATCGAATCTTCT AGCCCTT	GGCTAGAAGATT CGATC	ACGTAAATTGTTC CTATTAAGACAC CAAAG
matched site 11	GCTATCACTGCC ATGTCTGG	GACATGGCAGTG ATAGC	AAAGTTTTGTTTA GTTTTGTTTTTCC CCCA
matched site 12	GGTGCCTTGTTTA GGGGTAG	CCCCTAAACAAG GCACC	GAATGAGGCAAA GCTCAGACTTCA CCTCTA
matched site 18	GGGGAGGTGACA CCACTGAA	AGTGGTGTCACC TCCCC	CTGCTGCCTCCCA CTTGGGCAGCCC CATTC
matched site 20	GAGAAAATATGG GTTGAGGT	TCAACCCATATTT TCTC	TTGTCTTTCCTCT CCTCATCCCTCCC CACC
DYRK1A site 1	GTTGCCCTCATTA TTCAGCA	TGAATAATGAGG GCAAC	CTCTTAGGCTGCA AACCTGTTCAGC CGTGC
GRIN2B site 1	GCAGGAAGAACA GTTCAAGA	TGAACTGTTCTTC CTGC	CCAAGGCCAGGC TTCAACCTGCTAC CGTCT
MECP2 site 1	GCAGAGCTAGGG GTTCAGAG	TGAACCCCTAGC TCTGC	GCAGAACTGAAA CATGCTTCTTCAC CCCTC

MECP2 site 2	GCCTGCATATTACAACCAAG	GGTTGTAATATGCAGGC	GCACACTCGTGTTTCTGCGAAAAGCCTCTT
PTEN site 1	GGAGGCTATCAACAAAGAAT	CTTTGTTGATAGCCTCC	TGTCAATTTTTATAATGTTTCAAGCCCATT
SHANK3 site 1	GCCCGCTGCCGTATCCCGAG	GGGATACGGCAGCGGGC	GATCATGGAGCGCGCGCGCTTCTGCCGCTC
SHANK3 site 2	GAGCCCTCCCCGACCCACCG	TGGGTCGGGGAGGGCTC	TAGTCGAGGCCACCCGGGCGCGGACCGCGG
CUL3 site 1	GTAAACCTGGAATAACACGA	TGTTATTCCAGGTTTAC	CACTGGGGGCACGTCGAATGGGTTCCATCG
CUL3 site 2	GCCCAGTTCGAACTTCCTGG	GGAAGTTCGAACTGGGC	TTTCCCCTCACAGTGTAAACAAGGCCTCCA
UBE3A site 1	GTACAGTTAGTACTCAGCAG	CTGAGTACTAACTGTAC	TTCAGAAATCTGCCATTTGAGAGTCCACTG
UBE3A site 2	GCAGTTGTAGGGAAATAGGG	TATTTCCCTACAACTGC	TGTTAAACTCAACTCTTTCCTCAACCTCCC
HBB02	CTTGCCCCACAGGGCAGTAA	CTGCCCTGTGGGGCAAG	ATCTGACTCCTGAGGAGAAGTCTGCCGTTA
HBB04	CCACGTTCACCTTGCCCCAC	GGGCAAGGTGAACGTGG	TGAGGAGAAGTCTGCCGTTACTGCCCTGTG
HBG1	GTGGGGAAGGGGCCCCCAAG	GGGGGCCCCTTCCCCAC	GGCTATAAAAAAAATTAGCAGTATCCTCTT
BCL11A	TTTATCACAGGCTCCAGGAA	CTGGAGCCTGTGATAAA	CCCCCACCCTAATCAGAGGCCAAACCCTTC

CLTA	GCAGATGTAGTG TTTCCACA	GGAAACACTACA TCTGC	CCCCGCTGGTGC ACTGAAGAGCCA CCCTGT
DMD	CTTTCTACCTACT GAGTCTG	ACTCAGTAGGTA GAAAG	CCCCCCATCATAT CCCTATAAAGAC CCCAG
AAVS1	CTCCCTCCCAGG ATCCTCTC	AGGATCCTGGGA GGGAG	CTCTAAGGTTTGC TTACGATGGAGC CAGAG
HPRT	TCGAGATGTGAT GAAGGAGA	CCTTCATCACATC TCGA	AGAGGGCTACAA TGTGATGGCCTCC CATCT
CFTR site 1	ATTAAAGAAAAT ATCATCTT	ATGATATTTTCTT TAAT	TATCTATATTCAT CATAGGAAACAC CAAAG
CFTR site 2	TCTGTATCTATAT TCATCAT	ATGAATATAGAT ACAGA	ACCATTAAAGAA AATATCATCTTTG GTGTTTCCTATG
F8 site 1	GGTGACTCACCA TGCCCGGG	GGGCATGGTGAG TCACC	GGGCTGTGTCTCT GGCTGCCTCTGCC CCCC
F8 site 2	TCTGGCTGCCTCT GCCCCCC	GGGCAGAGGCAG CCAGA	CTCTGTGCCTGGT GACTCACCATGC CCGGG
PAH site 1	TAGCGAACTGAG AAGGGCCG	CCCTTCTCAGTTC GCTA	CTTAGGAACTTTG CTGCCACAATAC CTCGG
PAH site 2	ACTTTGCTGCCAC AATACCT	TATTGTGGCAGC AAAGT	TGGGTCGTAGCG AACTGAGAAGGG CCGAGG
PAH site 3	GGGTCGTAGCGA ACTGAGAA	TCAGTTCGCTACG ACCC	AACTTTGCTGCCA CAATACCTCGGC CCTTC
DMD site 1	ACTACGAGGCTG GCTCAGGG	TGAGCCAGCCTC GTAGT	ACTGCCAAAGTT TGAACCAGGACT CCCCCC

DMD site 2	TCTGGGCAGGAC TACGAGGC	TCGTAGTCCTGCC CAGA	GTTTGAACCAGG ACTCCCCCCTGA GCCAGCC
DMD site 3	TTACTGCCAAAG TTTGAACC	TCAAACTTTGGC AGTAA	AGGACTACGAGG CTGGCTCAGGGG GGAGTCCTGGT
LDLR site 1	CGCGGCGGGGAC TGCAGGTA	CTGCAGTCCCCG CCGCG	ACCTATTCTGGCG CCTGGAGCAAGC CTTAC
LDLR site 2	CTCGCCGCGGCG GGGACTGC	GTCCCCGCCGCG GCGAG	TTCTGGCGCCTGG AGCAAGCCTTAC CTGCA
LDLR site 3	TCTCAACCTATTC TGGCGCC	GCCAGAATAGGT TGAGA	GCCGCGGCGGGG ACTGCAGGTAAG GCTTGCTCCAGG C
FBN1 site 1	GGAAGGGTACTG CTTCACAG	TGAAGCAGTACC CTTCC	TGCTGGAGCCGA TCTGACACATGTT TTGTAGCACCTCT G
HEXA site 1	TACCTGAACCGT ATATCCTA	GATATACGGTTC AGGTA	ATGTAGAAATCC TTCCAGTCAGGG CCATAG
HEXA site 2	ATCCTTCCAGTCA GGGCCAT	GCCCTGACTGGA AGGAT	CCCCCTGGTACCT GAACCGTATATC CTATG
HEXA site 3	TGTAGAAATCCTT CCAGTCA	CTGGAAGGATTT CTACA	CCCCTGGTACCTG AACCGTATATCCT ATGGCCCTGA
PSEN1 site 1	AAAGAGCATGAT CACATGCT	ATGTGATCATGCT CTTT	AGGAGCTGACAT TGAAATATGGCG CCAAGC
PSEN1 site 2	GAGGAGCTGACA TTGAAATA	TTCAATGTCAGCT CCTC	AAGAGCATGATC ACATGCTTGGCG CCATAT

IDUA site 1	CCGCAGATGAGGAGCAGCTC	CTGCTCCTCATCTGCGG	GGTCCCGGCCTGCGACACTTCGGCCCAGAG
IDUA site 2	GGTCCCGGCCTGCGACACTT	TGTCGCAGGCCGGGACC	CCGCAGATGAGGAGCAGCTCTGGGCCGAAG
IDS site 1	ACTGAGGGATGTCTGAAGGC	TTCAGACATCCCTCAGT	ACTGATTGCCTATAGCCAGTATCCCCGGCC
IDS site 2	TTGCCTATAGCCAGTATCCC	ATACTGGCTATAGGCAA	ATTCCACTGAGGGATGTCTGAAGGCCGGGG
HBB site 1	GTAACGGCAGACTTCTCCTC	GAGAAGTCTGCCGTTAC	ACAGACACCATGGTGCATCTGACTCCTGAG
HBB site 2	CATGGTGCATCTGACTCCTG	GAGTCAGATGCACCATG	ACAGGGCAGTAACGGCAGACTTCTCCTCAG
TTR site 1	GGGATTGGTGACGACAGCCG	CTGTCGTCACCAATCCC	TGCTGAGCCCCTACTCCTATTCCACCACGG
TTR site 2	CCCCTACTCCTATTCCACCA	TGGAATAGGAGTAGGGG	ATTCCTTGGGATTGGTGACGACAGCCGTGG
LRRK2 site 1	ATTGCAAAGATTGCTGACTA	TCAGCAATCTTTGCAAT	ATTCTACAGCAGTACTGAGCAATGCCGTAG
HFE site 1	ATGGAGTTCGGGGCTCCACA	GGAGCCCCGAACTCCAT	AGCTGTTCGTGTTCTATGATCATGAGAGTCGCCGTGT
SOD1 site 1	ATTAGGCATGTTGGAGACTT	TCTCCAACATGCCTAAT	CATCGGCCACACCATCTTTGTCAGCAGTCACATTGCCCAAG
SOD1 site 2	AGAATCTTCAATAGACACAT	TGTCTATTGAAGATTCT	CTTGGGCAATGTGACTGCTGACAA

			AGATGGTGTGGC CGATG
ATP7B site 1	CCAGCAATACCT TTTTCTGC	GAAAAAGGTATT GCTGG	CACCTGAATGAG GCTGGCAGCCTT CCCGCA
PLOD1 site 1	CGTTGTGCAGGT GGGTGGTG	CACCCACCTGCA CAACG	CCATCTGCTCTCC CTAGACAGCTAC CGCAC
PLOD1 site 2	GAGGTCGTTGTG CAGGTGGG	ACCTGCACAACG ACCTC	CTGCTCTCCCTAG ACAGCTACCGCA CCACCC
CLCN1 site 1	ACCAGGAATGGG TCAATTCA	ATTGACCCATTCC TGGT	CCCACCAACAGC TGACCTCTCCAGC CATGA
CLCN1 site 2	TTCACCTTCCCAC CAGGAAT	CCTGGTGGGAAG GTGAA	AGCTGACCTCTCC AGCCATGAATTG ACCCATT
PEX10 site 1	CCAGGCACAGGG TGCACAGG	GTGCACCCTGTG CCTGG	GGAGGAGAGAGC CGTTTCCAGAAA CCCCCT
PEX10 site 2	CCTGCGCTCCTCC AGGCACA	GCCTGGAGGAGC GCAGG	GAGCCGTTTCCA GAAACCCCCTGT GCACCCTGT
POLG site 1	ACGTGGACATCC CTGGCTGC	GCCAGGGATGTC CACG	CACCTTGTGAGG CAGCTTGAAAAA CCAGCA
POLG site 2	TACAACGACGTG GACATCCC	ATGTCCACGTCGT TGTA	TGTGAGGCAGCT TGAAAAACCAGC AGCCAGGG
ALPL site 1	CTCAGAACAGGA CGCTCAGG	GAGCGTCCTGTTC TGAG	CCTGCTGCTCGCG CTGGCCCTCTACC CCCT
ALPL site 2	GCCCTCAGAACA GGACGCTC	CGTCCTGTTCTGA GGGC	GCTGCTCGCGCT GGCCCTCTACCCC CTGAG

Supplementary Table 4.8: Oligonucleotide sequences to perform large-scale high-throughput pooled pegRNA optimization. Table not included due to size. Please refer to the published manuscript for this table.

Supplementary Table 4.9: Top single, top composite, and 25th percentile pegRNA designs tested.

Target name	Edit type	Design 1 PBS length	Design 1 RTT length	Design 2 PBS length	Design 2 RTT length	Design 3 PBS length	Design 3 RTT length
HEK site 1	T to G	12	15	10	15	11	25
HEK site 4	T to C	9	14	12	14	15	18
VEGFA site 1	C to T	11	12	11	12	9	10
VEGFA site 3	T to A	11	12	11	10	17	10
VEGFA site 4	A to G	10	24	10	14	14	28
VEGFA site 5	A to C	14	17	14	17	10	26
EMX1 site 2	G to A	12	22	12	18	17	24
EMX1 site 4	G to T	10	12	12	12	13	30
CLTA site 1	C to G	16	10	15	10	8	30
FANCF site 2	A to T	14	20	11	11	16	10
FANCF site 3	C to A	11	11	11	11	13	26
FANCF site 4	G to C	12	20	10	20	14	23
RNF2 site 1	1bp insertion	17	13	15	13	9	29

RUNX1 site 1	1bp insertion	9	11	9	15	16	27
DNMT 1 site 1	2bp insertion	10	15	10	12	12	25
DNMT 1 site 2	2bp insertion	9	13	8	21	16	18
DNMT 1 site 4	3bp insertion	11	12	10	12	9	17
matche d site 7	3bp insertion	10	15	11	15	17	14
DYRK1 A site 1	1bp deletion	14	11	14	11	8	29
UBE3A site 1	2bp deletion	15	12	12	10	8	23

Supplementary Table 4.10: NGS primer sequences for pegRNA optimizations.

Target name	**Primer F**	**Primer R**
HEK site 1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCTGAACACCT GGAGGGCAAGTGC	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGGTTTAGTT CTCTTCCTCTTTCCCCTGTAG
HEK site 2	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCAATGATAA CAAGACCTGGCTGAGCTAACT G	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCTTCCCAAG TGAGAAGCCAGTGGAATAC
HEK site 3	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCCATGCAATT AGTCTATTTCTGCTGCAAGTA AG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCCTCCCTAG GTGCTGGCTTCC
HEK site 4	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGGCTGGGTG GAAGGAAGGGAG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGGTCAGAC GTCCAAAACCAGACTCC
VEGFA site 1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCTGTTGGCTG CCGCTCACTTTG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCGAGCGCC CCCTAGTGACTG

VEGFA site 2	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGGCAAAGTG AGTGACCTGCTTTTGG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGGACGAAA AGTTTCAGTGCGACGC
VEGFA site 3	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCACCACAGG GAAGCTGGGTGAATG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGGTTACGT GCGGACAGGGCC
VEGFA site 4	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGGACACTTCC CAAAGGACCCCAG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCCGGCTCTG GCTAAAGAGGGAAT
VEGFA site 5	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGGAGAATAT TGTCAGGGGGAAGGCAG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCCTACGGA ACACCAGACCCTGCTAG
EMX1 site 1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCTGGCCCAG GTGAAGGTGTGGT	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCAGGGAGT GGCCAGAGTCCAG
EMX1 site 2	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGGCTCCCATC ACATCAACCGGTG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCATTGCTTG TCCCTCTGTCAATGGC
EMX1 site 3	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCCTTCTCTCT GGCCCACTGTGTC	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCCGTTTGTA CTTTGTCCTCCGGTTC
EMX1 site 4	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCCCTAACCCT ATGTAGCCTCAGTCTTCC	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGATGTCCTC CCCATTGGCCTGC
FANCF site 1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCACTGGTTGT GCAGCCGCC	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCAATCAGT ACGCAGAGAGTCGCCG
FANCF site 2	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGGTCCCAGG TGCTGACGTAGGTAG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCAGAGTCA AGGAACACGGATAAAGACG C
FANCF site 3	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCCGGGAAAG AGTTGCTGCACCAG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGCAGTGGG CGCGCTACCTG
FANCF site 4	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGCCTGGAAG TTCGCTAATCCCGG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCTCATGGA ATCCCTTCTGCAGCACC

FANCF site 5	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCGACGAGAC AAAGGCGGCTG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGAGAGCCT GGCCCGCCTTG
RNF2 site 1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTTTCTCTCTTC TTTATTTCCAGCAATGTCTCAG GC	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTAAGTGTTA GCCAACATACAGAAGTCAG GAATG
RUNX1 site 1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGAATTCCTCT CACAAACAAGACAGGGAAC	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGACCTGTCT TGGTTTTCGCTCCGAAG
RUNX1 site 2	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGTGGGTACG AAGGAAATGACTCAAATATGC	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCTTTAACAA TTTGAATATTTGTTTTTACAA AGGTGC
ZSCAN2 site 1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCCCTACAAAT GCAGCGAGTGTGG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCACTCGGG GCATTTGTAGGGCTTC
DNMT1 site 1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCTAGTCCTTA GCAGCTTCCTCCTCCT	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGGGACCGT TTGAGGAGTGTTCAGTC
DNMT1 site 2	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGGCGAGTAA CAGACATGGACCATCAG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCCCTGGGG CCGTTTCCCTC
DNMT1 site 3	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCATGGACCA TCAGGAAACATTAACGTACTG ATG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCCCTGGGG CCGTTTCCCTC
DNMT1 site 4	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCTGAACACTC CTCAAACGGTCCCC	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCTTTCTCAA GGGGCTGCTGTGAGG
matched site 1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCAATAATGT ACCAATGTCAGTTTTTTAGTTT TGAC	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCCTTAAGG AAACAGAAGAGAAATCTGC GTGG
matched site 2	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGCTCTGTCTC TCACCTGGCGG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGGCAGATC TCTCTCTGACCCCCT

matched site 3	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCAAAACTTC AAACAGAATACATGTGAATTA AGTTATC	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTAATGATGA AACAATCCACATATTGATCC GTCAG
matched site 4	ACACTCTTTCCCTACACGACG CTCTTCCGATCTAGACGTGGT CTCGCTCTGATGAGAG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTACTGCTAG AACCCAGAAGGCAGA
matched site 5	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGTAACCACA GTCAAGTAGTTAATGTGTGTG AAAG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCTGTGGCAT TAATATGTACCTCACCACTG A
matched site 6	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGTAGCCATAT GTTTCTCATTCACTTGATACAC C	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCCAGAGCA TGGGCCTGGGAC
matched site 7	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGGAATTGTC AGGAAAATGGAGATAGCTTCA GG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCGCCACAC TGGGTTGGAAAGTCTTC
matched site 8	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCGTCCATAA ACGCTGCCATCATCTAATTG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCTATTCTAA TACAACCCTTGGCTTCAAAG AATGTG
matched site 9	ACACTCTTTCCCTACACGACG CTCTTCCGATCTTGCTGCCAA GGAAAACAGATGATTAGGTTC	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTTGAAATGC TATCTAATGTTCCCTAGTGC CTTAAG
matched site 10	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCAGGGGATG AGGAACTGCAGATCAC	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCCTCTGAG ACTCATCAGGAAGTGGTGTG
matched site 11	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGCTGGTGAG GACTTCTTAACAGAGCAG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCTGTCCCAT AGCCCCTGAGCC
matched site 12	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGTCATGTTAA AATTGCCAAATAAAGTACCTC TGC	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGGCCCTAA CAGAAATTCCGTGGTTTAGG

matched site 18	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCCTCATATGA GAACTGCAGTAAAGCTTTCCA	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCCCAAGGG CCTTTGTTTCTCATCCTC
matched site 20	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCCATAACGG CAGGTCTGAATTTCCCA	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCCTCATGAT CCACCCGCCTCAG
DYRK1A site 1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCCATTCCTGT AAACGCCCACACAAGTG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCGCTAGAC GGTGGAGCCTACTG
GRIN2B site 1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGGCAAGGTT GGATCCAAAACACTCC	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCAGCTTCAT CCCTGAGCCCAAAAG
MECP2 site 1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGAACTCAGG TACGGTGCTCAGTCTC	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCCCTCTATC CTCTCCACACCTCAAGTC
MECP2 site 2	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCTGCTCTCAA ACTCCTGGCCTCAAG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCACACACC TTCAATTTGTTAAAAACTCA GTATCTG
PTEN site 1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGGAATGAAC CTTCTGCAACATCTTAAGATC C	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGTTGACTG ATGTAGGTACTAACAGCATC TGA
SHANK3 site 1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGCGCTCCAG CTTCAAGCCCG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCCCAGGTT GGCGTAGGGGC
SHANK3 site 2	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCCTGGTGAA GCAGCTGCAGGTG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGTAGGGAT GGCAGATCCGCGC
CUL3 site 1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTTGGTCCTGGC TTCATTTACCTTGAAATGG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGTGTGAGC AGTTTCTCTCTGTAGTGTAGT AAAAAG
CUL3 site 2	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCTGGGACAC TCAAGCTTGGTGGG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTTTCTTCCCA GGTGCTCTGTCCCAG

UBE3A site 1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGGGAAATCC AGTGAGAAGACAGCAGTT	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCCTTTATAA GTACTGGGAATATTTCAGCT TACTGCC
UBE3A site 2	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCCATTCTACC TCAGTGACTGAAATTATTTGC TTC	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCTAGGTTTC TACTACTAGATAGATAAATG CACACGC
HBB02	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGGTCTCCTTA AACCTGTCTTGTAACCTTGAT AC	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGCCAATCT ACTCCCAGGAGCAGG
HBB04	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCTATTGGTCT CCTTAAACCTGTCTTGTAACCT TG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGAGCAGGG AGGGCAGGAGC
HBG1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGCTATTGGTC AAGGCAAGGCTGG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCTAGAGAC AAGAAGGTAAAAAACGGCT GAC
BCL11A	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGCATGGCAT ACAAATTATTTCATTCCCATTG AG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCTCTTAGAC ATAACACACCAGGGTCAATA CAAC
CLTA	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGCCTTTGTAA ATGACATTGACGAGTCGTCC	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGGTCCAAA AGAACTCAACATAATTAATC CAATGACT
DMD	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCCATCACTCC TCCAAAGTCCAGCTG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCCCCACCTT CAAGCTGAAACCCAC
AAVS1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGGTGGCCAC TGAGAACCGGG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGCCACTAG GGACAGGATTGGTGACAG
HPRT	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGACTAAGAG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTTAGCTCTTC

	GTGTTTGTTATAAAGTTTAATGTATGAAAC	AGTCTGATAAAATCTACAGTCATAG
CFTR site 1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCCTTCAGAGGGTAAAATTAAGCACAGTG	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTAGACTAACCGATTGAATATGGAGCC
CFTR site 2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCCTTCAGAGGGTAAAATTAAGCACAGTG	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTAGACTAACCGATTGAATATGGAGCC
F8 site 1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACAACCTGCGGGCTGAAGGG	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTCCATGCCCCCATTACCAGCAC
F8 site 2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACAACCTGCGGGCTGAAGGG	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTCCATGCCCCCATTACCAGCAC
PAH site 1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACTTACTGTTAATGGAATCAGCCAAAATCTTAAGC	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGCTGTTGAAGACCCTGCTCTAGG
PAH site 2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACTTACTGTTAATGGAATCAGCCAAAATCTTAAGC	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGCTGTTGAAGACCCTGCTCTAGG
PAH site 3	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACTTACTGTTAATGGAATCAGCCAAAATCTTAAGC	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGCTGTTGAAGACCCTGCTCTAGG
DMD site 1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTTCCTCAAGATCTGCTAGGATTCTCTCTAG	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTTTCTATTTTCAAATACACTCCTGAGTCCCTAACC
DMD site 2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTTCCTCAAGATCTGCTAGGATTCTCTCTAG	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTTTCTATTTTCAAATACACTCCTGAGTCCCTAACC
DMD site 3	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTTCCTCAAGATCTGCTAGGATTCTCTCTAG	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTTTCTATT

		TTCAAATACACTCCTGAGTC CCTAACC
LDLR site 1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCAGAGGCTG CGAGCATGGGG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGACCCTCG CGCTCCCCTC
LDLR site 2	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCAGAGGCTG CGAGCATGGGG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGACCCTCG CGCTCCCCTC
LDLR site 3	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCAGAGGCTG CGAGCATGGGG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGACCCTCG CGCTCCCCTC
FBN1 site 1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGGACCCCAG CCTCTCCCTCC	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCCCTCTTGC TTCTTCTCACCCAGG
HEXA site 1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCAACCCAGC CTCCTTTGGTTAGCAAG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGTGTCCTTA CTGCCATTTGACCTTTTATAA CAG
HEXA site 2	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCAACCCAGC CTCCTTTGGTTAGCAAG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGTGTCCTTA CTGCCATTTGACCTTTTATAA CAG
HEXA site 3	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCAACCCAGC CTCCTTTGGTTAGCAAG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGTGTCCTTA CTGCCATTTGACCTTTTATAA CAG
PSEN1 site 1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGAGCCTTGG CCACCCTGAGC	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGGCTTTGA GAATAATAAAACAAAACTC ATACGTAC
PSEN1 site 2	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGAGCCTTGG CCACCCTGAGC	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGGCTTTGA GAATAATAAAACAAAACTC ATACGTAC
IDUA site 1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGGAGCGAGT GGTGGGAGGC	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCGACGCTG CGGTTGGGGTG

IDUA site 2	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGGAGCGAGT GGTGGGAGGC	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCGACGCTG CGGTTGGGGTG
IDS site 1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCCACACAGT ATACCTATAGTCTATGGTGCG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCCTCGCTGC CCCGTTCCTTC
IDS site 2	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCCACACAGT ATACCTATAGTCTATGGTGCG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCCTCGCTGC CCCGTTCCTTC
HBB site 1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCTTGTAACCT TGATACCAACCTGCCCAG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCCCTAGGG TTGGCCAATCTACTCCC
HBB site 2	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCTTGTAACCT TGATACCAACCTGCCCAG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCCCTAGGG TTGGCCAATCTACTCCC
TTR site 1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGGATCTGTCT GTCTTCTCTCATAGGTGGTATT C	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGTCTCTGCC TGGACTTCTAACATAGCATA TG
TTR site 2	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGGATCTGTCT GTCTTCTCTCATAGGTGGTATT C	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGTCTCTGCC TGGACTTCTAACATAGCATA TG
LRRK2 site 1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGCCATGATTA TATACCGAGACCTGAAACCC	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCTCTGTTCA CAATGTGATAGACTCTGTTT TCC
HFE site 1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCCTGCTCCCC TCCTACTACACATGG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCACCCTTTC AGACTCTGACTCAGCTG
SOD1 site 1	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGCTCATGAA CTACCTTGATGTTTAGTGGCA T	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCCGCGACT AACAATCAAAGTGAAAAGA TACATG
SOD1 site 2	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGCTCATGAA	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTCCGCGACT

	CTACCTTGATGTTTAGTGGCAT	AACAATCAAAGTGAAAAGATACATG
ATP7B site 1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTGGTGCTCTCTGTGGTTTGACC	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTGAACTTGGAACAGAGACCTTGGGATACTG
PLOD1 site 1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTTCAGTTCGGCGGGTGGAGG	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTCCTTTGCTCCCAGCCCCTATGTG
PLOD1 site 2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTTCAGTTCGGCGGGTGGAGG	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTCCTTTGCTCCCAGCCCCTATGTG
CLCN1 site 1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGTTAACTCTGTTTCTTTTTCAGCCGCC	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTCCAGCCTCATTATTCAAGGGTCTCTTTTGC
CLCN1 site 2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGTTAACTCTGTTTCTTTTTCAGCCGCC	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTCCAGCCTCATTATTCAAGGGTCTCTTTTGC
PEX10 site 1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTGATGCACTCCCAGCAGAACAGG	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTCAGGAAGGAGTGGAGGCTGCAC
PEX10 site 2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTGATGCACTCCCAGCAGAACAGG	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTCAGGAAGGAGTGGAGGCTGCAC
POLG site 1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCAGTATGTGCCTGAAATCACACTCTGTCC	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCCCCAAGGACACCCAGCC
POLG site 2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCAGTATGTGCCTGAAATCACACTCTGTCC	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCCCCAAGGACACCCAGCC
ALPL site 1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCGTCTTCTCCAAGGGCCCC	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGGCTGCCGTGTGGGAAGTTG

ALPL site 2	ACACTCTTTCCCTACACGACG CTCTTCCGATCTCCGTCTTCTC CAAGGGCCCC	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGGGCTGCC GTGTGGGAAGTTG
BCR-ABL1 AA311-318 cDNA target	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGGAGGTGGA AGAGTTCTTGAAAGAAGCTG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGTGGATGA AGTTTTTCTTCTCCAGGTACT CC
IRF1 5' UTR gDNA target	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGCGCGAGGG TCCCGGC	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTGGAATCCC GCTAAGTGTTTGGATTGC
IRF1 5' UTR cDNA target	ACACTCTTTCCCTACACGACG CTCTTCCGATCTGCATCCGAG TGATGGGCATGTTGG	GACTGGAGTTCAGACGTGTG CTCTTCCGATCTAGAGCTCG CCACTCCTTAGTCGAG

4.9 Acknowledgements

Support for this work was provided by the National Institutes of Health (RM1 HG009490 and R35 GM118158 to J.K.J.). J.K.J. is additionally supported by the Desmond and Ann Heathwood MGH Research Scholar Award and the Robert B. Colvin, M. D. Endowed Chair in Pathology. L.P. is supported by the National Human Genome Research Institute (NHGRI) Career Development Award (R00HG008399), Genomic Innovator Award (R35HG010717) and CEGS RM1HG009490. We thank Vikram Pattanayak and Julian Grunewald for discussions and technical advice and L. Paul Pottenplackel for assistance with editing the manuscript.

4.10 Author contributions

J.Y.H., K.C.L., and J.Y.S. performed the laboratory experiments and J.Y.H. and J.Y.S. performed the computational analyses. J.Y.H., K.C.L., and J.K.J. conceived of and designed the study. J.K.J. and L.P. supervised the work. J.Y.H. and J.K.J. wrote the initial manuscript draft and all authors contributed to the writing of the final manuscript.

5 Conclusion

CRISPR-based saturation mutagenesis screens are a powerful framework for the systematic and unbiased discovery of genetic elements underlying phenotypes of interest. In this thesis, I described a generalizable computational method, called CRISPR-SURF, for the analysis of CRISPR saturation mutagenesis screen data. We conceptualized tiled CRISPR perturbations across an underlying regulatory genomic signal akin to a sensor making measurements across a discrete function (bp of the genome), and therefore modeled the screen data by means of a convolution operation. Utilizing empirical data from CRISPR experiments, we constructed specific perturbation profiles for the various CRISPR genome and epigenome editing tools to account for their different perturbation ranges. Additionally, we explicitly parameterized the non-uniform targeting of sgRNAs tiled across the genomic sequence which allowed us to estimate statistical significance and power at single-nucleotide resolution. The CRISPR-SURF framework leverages the broad CRISPRi and CRISPRa perturbation profile for efficient enhancer discovery and the narrow CRISPR–Cas perturbation profile for high-resolution mapping of critical elements within enhancers.

Future development of CRISPR-SURF could focus on incorporating sgRNA-specific predictions into the deconvolution model. Different sgRNAs exhibit varying perturbation efficiencies, and this information can play an important role in the deconvolution operation. For example, two sgRNAs targeting different regions of the genome could yield similar enrichment scores, however, their underlying perturbation efficiencies (e.g. indel frequency) could be very different. This would mean that the genomic region targeted by the less efficient sgRNA is likely more functional as it exhibits a similar enrichment score with less perturbation. Furthermore, the incorporation of sgRNA-specific convolution profiles is an attractive research direction with progress in the ability to predict indel and base editing outcomes following genetic modification by various CRISPR technologies. CRISPR-SURF currently utilizes a generalized perturbation profile for different CRISPR technologies, but the addition of sgRNA-specific convolution profiles could enhance the accuracy and precision of CRISPR-SURF to identify non-coding regulatory elements.

Next, we developed PrimeDesign software for the rapid and simplified design of prime editing gRNAs. Prime editing technology enables unprecedented precision and versatility in the installation of all point mutations and short insertion and deletion edits, but the design of its gRNAs is more complicated compared to CRISPR-Cas nucleases and base editors. The pegRNA can take on many different designs varying its PBS and RTT elements, and the empirical testing of these different designs is important to achieve maximal prime editing. To facilitate this process, PrimeDesign provides an interactive web application requiring just a single input that encodes both the original reference and the desired edited sequence, and returns a range of possible pegRNA and ngRNA combinations to install the edit of interest. Annotations for strategies known to increase editing frequency and purity (e.g. PAM mutating, PE3b) are provided. Additionally, we constructed PrimeVar (http://primedesign.pinellolab.org/primevar), a comprehensive and searchable database for pegRNA and ngRNA combinations to install or correct >68,500 pathogenic human genetic variants from ClinVar. PrimeDesign should greatly simplify the complicated process of designing candidate prime editing components and thereby increase the use of and accessibility to this powerful and important genome editing technology.

Future development of PrimeDesign could focus on the incorporation of additional important predictors of pegRNA efficiency. Attributes such as spacer sequence targeting efficiency, RNA secondary structures, and sequence elements that affect reverse transcription efficiency could provide more accurate pegRNA recommendations to achieve greater prime editing efficiencies. Furthermore, the extension of PrimeDesign for alternative Cas orthologues with different PAM requirements and prime editing strategies introducing larger genetic modifications would broaden the utility of the software for the research community.

Finally, we developed MOSAIC as a method for the high-throughput pooled optimization of pegRNAs and for non-viral *in situ* saturation mutagenesis at single-nucleotide and codon resolution. The optimization of prime editing is necessary to achieve maximal levels of prime editing, however, the individual assessment of multiple pegRNA designs is labor intensive. We developed a strategy to utilize barcode edits to uniquely map to PBS-RTT length combinations and assessed up to 210 different pegRNA designs in a single reaction. Following targeted amplicon sequencing, we utilized the frequency of each barcode edit to identify high efficiency

pegRNA designs. Using this pooled optimization strategy, we constructed pegRNA optimization profiles for 89 different spacers and assessed a total of >18,000 unique pegRNA designs. We validated top scoring pegRNA designs to install all possible point mutations and short insertion and deletion edits, achieving PE2 efficiencies upwards of 40%. Additionally, we used MOSAIC for *in situ* saturation mutagenesis across coding and non-coding sequences. To discover genetic resistance mechanisms in CML, we installed all possible amino acid variants across the imatinib binding site within the *BCR-ABL1* oncogene in K562. We identified multiple specific amino acid variants at position 315 that conferred resistance to imatinib treatment, including the T315I "gatekeeper" mutation commonly found in CML patients. Next, we utilized MOSAIC for *in situ* saturation mutagenesis of non-coding sequences across the 5' untranslated region (UTR) of the transcription factor gene *IRF1* to map regulatory elements involved in its transcriptional initiation. We observed statistically significant depletion of two contiguous 3 bp tiles across the entire mutagenesis window which directly overlapped the consensus motif of the core downstream promoter element (DPE), supporting its role in the initiation of gene transcription by RNA polymerase II. These results demonstrate the capability of MOSAIC to comprehensively screen and identify potential drug resistant variants *in situ* to profile new or existing drugs, and to finely-map non-coding regulatory elements involved in gene regulation at base-pair resolution.

The ability to empirically screen hundreds of pegRNA designs in a single reaction enables the high-throughput identification of pegRNAs to introduce maximal levels of prime editing. While the work in this thesis describes the optimization of pegRNAs for installation of small genetic modifications such as substitutions and short insertion and deletion mutations, an extension of MOSAIC could focus on the optimization of pegRNA designs for the installation of larger insertion and deletion mutations to expand the utility of prime editing technology. General pegRNA design principles for installing short genetic modifications have been discussed by the field, however, the design principles underlying pegRNAs installing larger genetic modifications is more unknown and warrants further experimental characterization. Additionally, the further characterization of pegRNA barcode placement could improve the accuracy of the MOSAIC method in identifying top pegRNA designs. For example, instead of placing the barcode in the form of an insertion edit at the nick site, the barcode placement could overlap the intended editing region to provide a more *in situ* read out of prime editing efficiency.

Utilization of MOSAIC as a high-throughput strategy for *in situ* mutagenesis addresses many limitations of related technologies. While massively parallel reporter assays (MPRAs) have enabled high-throughput characterization of sequence elements, these assays measure sequence function outside native genomic contexts which results in the inability to capture important features such as endogenous chromatin structure and interactions with other genomic loci (e.g. promoter-enhancer interactions). Saturation mutagenesis strategies with CRISPR-Cas nucleases and base editors offer an *in situ* approach for sequence characterization, however, these CRISPR technologies are limited in the types of genetic modifications they can introduce and often don't allow comprehensive profiling of all amino acid variants for coding screens and uniform fine-mapping for non-coding screens. In contrast, MOSAIC facilitates *in situ* characterization of comprehensive genetic variation by enabling the installation of all amino acid variants at a given codon and thousands of defined substitution, insertion, or deletion edits at a genomic target of interest.

MOSAIC presents a unique opportunity for the fine-mapping of critical sequence elements underlying protein function and gene regulation, and has the potential to facilitate applications in drug development, gene expression modification, and cellular barcoding. In terms of drug development, the ability to comprehensively evaluate genetic resistance susceptibility of targeted therapies in oncology could aid in the design of more effective molecules. Iterative screening of compounds against a library of potential genetic drivers of resistance could guide the design towards a more robust lead compound. Additionally, MOSAIC could be used in drug development to inform the rationale for more effective treatment regimens by discovering combinations of therapeutic agents which act orthogonally in the face of genetic resistance. Next, MOSAIC could be utilized to fine-tune endogenous gene expression by installing genetic variants that affect processes such as transcription factor binding, transcriptional initiation, and transcript stability (e.g. microRNA targeting). The ability to high-throughput screen genetic variants which modify target gene expression could provide a therapeutic strategy to treat diseases caused by haploinsufficiency. Lastly, MOSAIC could be a powerful strategy for cellular barcoding and lineage tracing through the installation of randomized insertion edits at multiple genomic loci. The use of mixed-base synthesis with MOSAIC allows for theoretically

unprecedented genetic diversity to be installed at targeted locations in the genome. Whereas lineage tracing strategies with CRISPR-Cas nucleases exhibit more limited genetic diversity inherent with indel products, MOSAIC offers a potentially differentiated strategy by being able to install orders of magnitude more genetic variation at each genomic target.

In conclusion, the work outlined in this thesis will enable researchers to perform systematic CRISPR-based saturation mutagenesis screens for the identification of genetic elements underlying a phenotype of interest. Whereas the development of CRISPR-SURF will aid in the analysis of large-scale tiling screens to identify regulatory elements at 10-100 bp resolution, PrimeDesign and MOSAIC will facilitate the use of prime editing for the high-resolution characterization of genetic variants at single-nucleotide and codon resolution.

References

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, (2001).
2. Lander, E. S. Initial impact of the sequencing of the human genome. *Nature* **470**, (2011).
3. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, (2010).
4. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, (2012).
5. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, (2015).
6. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, (2015).
7. Wong, K. H. Y. *et al.* Towards a reference genome that captures global genetic diversity. *Nature communications* **11**, (2020).
8. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *American journal of human genetics* **90**, (2012).
9. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *American journal of human genetics* **101**, (2017).
10. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic acids research* **45**, (2017).
11. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics* **20**, (2019).
12. Menzel, S. *et al.* A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nature genetics* **39**, (2007).
13. Uda, M. *et al.* Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of beta-thalassemia. *Proceedings of the National Academy of Sciences of the United States of America* **105**, (2008).
14. Bauer, D. E. *et al.* An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science (New York, N.Y.)* **342**, (2013).
15. Frangoul, H. *et al.* CRISPR-Cas9 Gene Editing for Sickle Cell Disease and β-Thalassemia. *The New England journal of medicine* **384**, (2021).

16. Mills, R. E. *et al.* An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome research* **16**, (2006).
17. de Beer, T. A. P. *et al.* Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 genomes project dataset. *PLoS computational biology* **9**, (2013).
18. Telenti, A. *et al.* Deep sequencing of 10,000 human genomes. *Proceedings of the National Academy of Sciences of the United States of America* **113**, (2016).
19. Gojobori, T., Li, W. H. & Graur, D. Patterns of nucleotide substitution in pseudogenes and functional genes. *Journal of molecular evolution* **18**, 360–9 (1982).
20. Li, W.-H., Wu, C.-I. & Luo, C.-C. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *Journal of Molecular Evolution* **21**, 58–71 (1984).
21. Blake, R. D., Hess, S. T. & Nicholson-Tuell, J. The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. *Journal of Molecular Evolution* **34**, 189–200 (1992).
22. Guo, C. *et al.* Transversions have larger regulatory effects than transitions. *BMC Genomics* **18**, (2017).
23. Holland, D. *et al.* Estimating Effect Sizes and Expected Replication Probabilities from GWAS Summary Statistics. *Frontiers in Genetics* **7**, (2016).
24. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science (New York, N.Y.)* **339**, (2013).
25. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science (New York, N.Y.)* **339**, (2013).
26. Hwang, W. Y. *et al.* Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nature Biotechnology* **31**, (2013).
27. Cho, S. W., Kim, S., Kim, J. M. & Kim, J.-S. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nature Biotechnology* **31**, (2013).
28. Jinek, M. *et al.* RNA-programmed genome editing in human cells. *eLife* **2**, (2013).
29. Zetsche, B. *et al.* Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell* **163**, 759–771 (2015).

30. Ran, F. A. *et al.* In vivo genome editing using Staphylococcus aureus Cas9. *Nature* **520**, 186–191 (2015).
31. Shen, M. W. *et al.* Predictable and precise template-free CRISPR editing of pathogenic variants. *Nature* **563**, 646–651 (2018).
32. Chen, W. *et al.* Massively parallel profiling and predictive modeling of the outcomes of CRISPR/Cas9-mediated double-strand break repair. *Nucleic Acids Research* **47**, 7989–8003 (2019).
33. Allen, F. *et al.* Predicting the mutations generated by repair of Cas9-induced double-strand breaks. *Nature Biotechnology* **37**, 64–72 (2019).
34. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic Screens in Human Cells Using the CRISPR-Cas9 System. *Science* **343**, 80–84 (2014).
35. Shalem, O. *et al.* Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. *Science* **343**, 84–87 (2014).
36. Canver, M. C. *et al.* BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527**, 192–197 (2015).
37. Canver, M. C. *et al.* Variant-aware saturating mutagenesis using multiple Cas9 nucleases identifies regulatory elements at trait-associated loci. *Nature Genetics* **49**, 625–634 (2017).
38. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, (2016).
39. Gaudelli, N. M. *et al.* Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551**, (2017).
40. Kurt, I. C. *et al.* CRISPR C-to-G base editors for inducing targeted DNA transversions in human cells. *Nature Biotechnology* **39**, (2021).
41. Kuscu, C. *et al.* CRISPR-STOP: gene silencing through base-editing-induced nonsense mutations. *Nature methods* **14**, (2017).
42. Gapinske, M. *et al.* CRISPR-SKIP: programmable gene splicing with single base editors. *Genome Biology* **19**, (2018).
43. Hanna, R. E. *et al.* Massively parallel assessment of human variants with base editor screens. *Cell* **184**, (2021).

44. Huang, C., Li, G., Wu, J., Liang, J. & Wang, X. Identification of pathogenic variants in cancer genes using base editing screens with editing efficiency correction. *Genome Biology* **22**, (2021).
45. Cuella-Martin, R. *et al.* Functional interrogation of DNA damage response variants with base editing screens. *Cell* **184**, (2021).
46. Liu, Z. *et al.* Highly efficient RNA-guided base editing in rabbit. *Nature Communications* **9**, 2717 (2018).
47. Zafra, M. P. *et al.* Optimized base editors enable efficient editing in cells, organoids and mice. *Nature Biotechnology* **36**, 888–893 (2018).
48. Villiger, L. *et al.* Treatment of a metabolic liver disease by in vivo genome base editing in adult mice. *Nature Medicine* **24**, 1519–1525 (2018).
49. Zhang, Y. *et al.* Programmable base editing of zebrafish genome using a modified CRISPR-Cas9 system. *Nature Communications* **8**, 118 (2017).
50. Li, Q. *et al.* CRISPR–Cas9-mediated base-editing screening in mice identifies DND1 amino acids that are critical for primordial germ cell development. *Nature Cell Biology* **20**, 1315–1325 (2018).
51. Yeh, W.-H., Chiang, H., Rees, H. A., Edge, A. S. B. & Liu, D. R. In vivo base editing of post-mitotic sensory cells. *Nature Communications* **9**, 2184 (2018).
52. Zeng, Y. *et al.* Correction of the Marfan Syndrome Pathogenic FBN1 Mutation by Base Editing in Human Cells and Heterozygous Embryos. *Molecular Therapy* **26**, 2631–2637 (2018).
53. Li, G. *et al.* Highly efficient and precise base editing in discarded human tripronuclear embryos. *Protein & Cell* **8**, 776–779 (2017).
54. Zhou, C. *et al.* Highly efficient base editing in human tripronuclear zygotes. *Protein & Cell* **8**, 772–775 (2017).
55. Liang, P. *et al.* Correction of β-thalassemia mutant by base editor in human embryos. *Protein & Cell* **8**, 811–822 (2017).
56. Zong, Y. *et al.* Precise base editing in rice, wheat and maize with a Cas9-cytidine deaminase fusion. *Nature Biotechnology* **35**, 438–440 (2017).
57. Shimatani, Z. *et al.* Targeted base editing in rice and tomato using a CRISPR-Cas9 cytidine deaminase fusion. *Nature Biotechnology* **35**, 441–443 (2017).

58. Levy, J. M. *et al.* Cytosine and adenine base editing of the brain, liver, retina, heart and skeletal muscle of mice via adeno-associated viruses. *Nature Biomedical Engineering* **4**, 97–110 (2020).
59. Sasaguri, H. *et al.* Introduction of pathogenic mutations into the mouse Psen1 gene by Base Editor and Target-AID. *Nature Communications* **9**, 2892 (2018).
60. Liang, P. *et al.* Effective gene editing by high-fidelity base editor 2 in mouse zygotes. *Protein & Cell* **8**, 601–611 (2017).
61. Xie, J. *et al.* Efficient base editing for multiple genes and loci in pigs using base editors. *Nature Communications* **10**, 2852 (2019).
62. Rossidis, A. C. *et al.* In utero CRISPR-mediated therapeutic editing of metabolic genes. *Nature Medicine* **24**, 1513–1518 (2018).
63. Kim, K. *et al.* Highly efficient RNA-guided base editing in mouse embryos. *Nature Biotechnology* **35**, 435–437 (2017).
64. Liu, Z. *et al.* Efficient generation of mouse models of human diseases via ABE- and BE-mediated base editing. *Nature Communications* **9**, 2338 (2018).
65. Ryu, S.-M. *et al.* Adenine base editing in mouse embryos and an adult mouse model of Duchenne muscular dystrophy. *Nature Biotechnology* **36**, 536–539 (2018).
66. Li, C. *et al.* Expanded base editing in rice and wheat using a Cas9-adenosine deaminase fusion. *Genome Biology* **19**, 59 (2018).
67. Song, C.-Q. *et al.* Adenine base editing in an adult mouse model of tyrosinaemia. *Nature Biomedical Engineering* **4**, 125–130 (2020).
68. Anzalone, A. v. *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).
69. Chen, P. J. *et al.* Enhanced prime editing systems by manipulating cellular determinants of editing outcomes. *Cell* **184**, 5635-5652.e29 (2021).
70. Maeder, M. L. *et al.* CRISPR RNA–guided activation of endogenous human genes. *Nature Methods* **10**, 977–979 (2013).
71. Perez-Pinera, P. *et al.* RNA-guided gene activation by CRISPR-Cas9–based transcription factors. *Nature Methods* **10**, 973–976 (2013).
72. Larson, M. H. *et al.* CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nature Protocols* **8**, 2180–2196 (2013).

73. Gilbert, L. A. *et al.* CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. *Cell* **154**, 442–451 (2013).
74. Chavez, A. *et al.* Highly efficient Cas9-mediated transcriptional programming. *Nature Methods* **12**, 326–328 (2015).
75. Tak, Y. E. *et al.* Inducible and multiplex gene regulation using CRISPR-Cpf1-based transcription factors. *Nature methods* **14**, 1163–1166 (2017).
76. Yeo, N. C. *et al.* An enhanced CRISPR repressor for targeted mammalian gene regulation. *Nature Methods* **15**, 611–616 (2018).
77. Alerasool, N., Segal, D., Lee, H. & Taipale, M. An efficient KRAB domain for CRISPRi applications in human cells. *Nature Methods* **17**, 1093–1096 (2020).
78. Tak, Y. E. *et al.* Augmenting and directing long-range CRISPR-mediated activation in human cells. *Nature Methods* **18**, 1075–1081 (2021).
79. Joung, J. *et al.* Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. *Nature protocols* **12**, 828–863 (2017).
80. Fulco, C. P. *et al.* Systematic mapping of functional enhancer–promoter connections with CRISPR interference. *Science* **354**, 769–773 (2016).
81. Simeonov, D. R. *et al.* Discovery of stimulation-responsive immune enhancers with CRISPR activation. *Nature* **549**, 111–115 (2017).
82. Patwardhan, R. P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nature Biotechnology* **27**, 1173–1175 (2009).
83. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nature Biotechnology* **30**, 265–270 (2012).
84. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology* **30**, 271–277 (2012).
85. Vockley, C. M. *et al.* Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. *Genome Research* **25**, 1206–1214 (2015).
86. Tewhey, R. *et al.* Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165**, 1519–1529 (2016).
87. Ulirsch, J. C. *et al.* Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell* **165**, 1530–1545 (2016).

88. Liu, S. *et al.* Systematic identification of regulatory variants associated with cancer risk. *Genome Biology* **18**, 194 (2017).
89. Arnold, C. D. *et al.* Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
90. Kwasnieski, J. C., Fiore, C., Chaudhari, H. G. & Cohen, B. A. High-throughput functional testing of ENCODE segmentation predictions. *Genome Research* **24**, 1595–1602 (2014).
91. Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Research* **27**, 38–52 (2017).
92. Klein, J. C. *et al.* Functional testing of thousands of osteoarthritis-associated variants for regulatory activity. *Nature Communications* **10**, 2434 (2019).
93. Klein, J. C. *et al.* A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nature Methods* **17**, 1083–1091 (2020).
94. Korkmaz, G. *et al.* Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nature Biotechnology* **34**, 192–198 (2016).
95. Sanjana, N. E. *et al.* High-resolution interrogation of functional elements in the noncoding genome. *Science* **353**, 1545–1549 (2016).
96. Klann, T. S. *et al.* CRISPR–Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nature Biotechnology* **35**, 561–568 (2017).
97. Bauer, D. E. *et al.* An Erythroid Enhancer of *BCL11A* Subject to Genetic Variation Determines Fetal Hemoglobin Level. *Science* **342**, 253–257 (2013).
98. Vierstra, J. *et al.* Functional footprinting of regulatory DNA. *Nature Methods* **12**, 927–930 (2015).
99. Robinson, J. T. *et al.* Integrative genomics viewer. *Nature Biotechnology* **29**, 24–26 (2011).
100. Tibshirani, R. J. & Taylor, J. The solution path of the generalized lasso. *The Annals of Statistics* **39**, (2011).
101. van Overbeek, M. *et al.* DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks. *Molecular Cell* **63**, 633–646 (2016).
102. Gilbert, L. A. *et al.* Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159**, 647–661 (2014).

103. Thakore, P. I. *et al.* Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nature Methods* **12**, 1143–1149 (2015).
104. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29**, (2001).
105. Cho, S. W. *et al.* Promoter of lncRNA Gene PVT1 Is a Tumor-Suppressor DNA Boundary Element. *Cell* **173**, 1398-1412.e22 (2018).
106. Funnell, A. P. W. *et al.* 2p15-p16.1 microdeletions encompassing and proximal to BCL11A are associated with elevated HbF in addition to neurologic impairment. *Blood* **126**, 89–93 (2015).
107. Chen, B. *et al.* Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System. *Cell* **155**, 1479–1491 (2013).
108. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology* **34**, 184–191 (2016).
109. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research* **44**, D862–D868 (2016).
110. Lin, Q. *et al.* Prime genome editing in rice and wheat. *Nature Biotechnology* **38**, 582–585 (2020).
111. Liu, Y. *et al.* Efficient generation of mouse models with the prime editing system. *Cell Discovery* **6**, 27 (2020).
112. Kim, H. K. *et al.* Predicting the efficiency of prime editing guide RNAs in human cells. *Nature Biotechnology* **39**, 198–206 (2021).
113. Hsu, J. Y. *et al.* PrimeDesign software for rapid and simplified design of prime editing guide RNAs. *Nature Communications* **12**, 1034 (2021).