nature methods

SUPPLEMENTARY INFORMATION

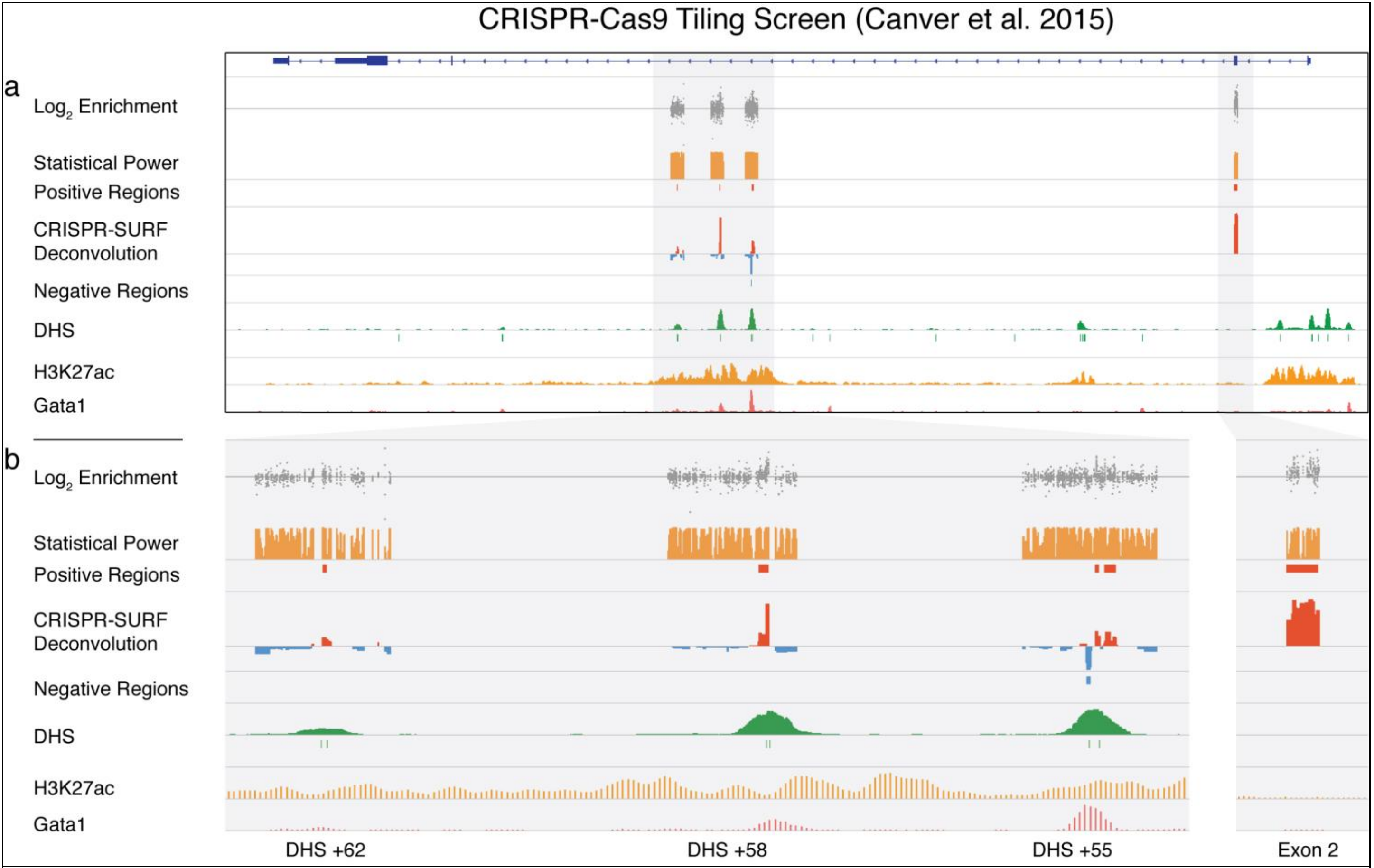
https://doi.org/10.1038/s41592-018-0238-1

In the format provided by the authors and unedited.

CRISPR-SURF: discovering regulatory elements by deconvolution of CRISPR tiling screen data

Jonathan Y. Hsu[1,2], Charles P. Fulco[3,4], Mitchel A. Cole[5,6,7], Matthew C. Canver[2,5,6,7], Danilo Pellin[8,9], Falak Sher[5,6,7], Rick Farouni[2,3,10], Kendell Clement[2,3,10], Jimmy A. Guo[2], Luca Biasco[8,9], Stuart H. Orkin[5,6,7,11], Jesse M. Engreitz[3,12], Eric S. Lander[3,4,13], J. Keith Joung[2,10], Daniel E. Bauer[5,6,7]* and Luca Pinello[2,3,10]*

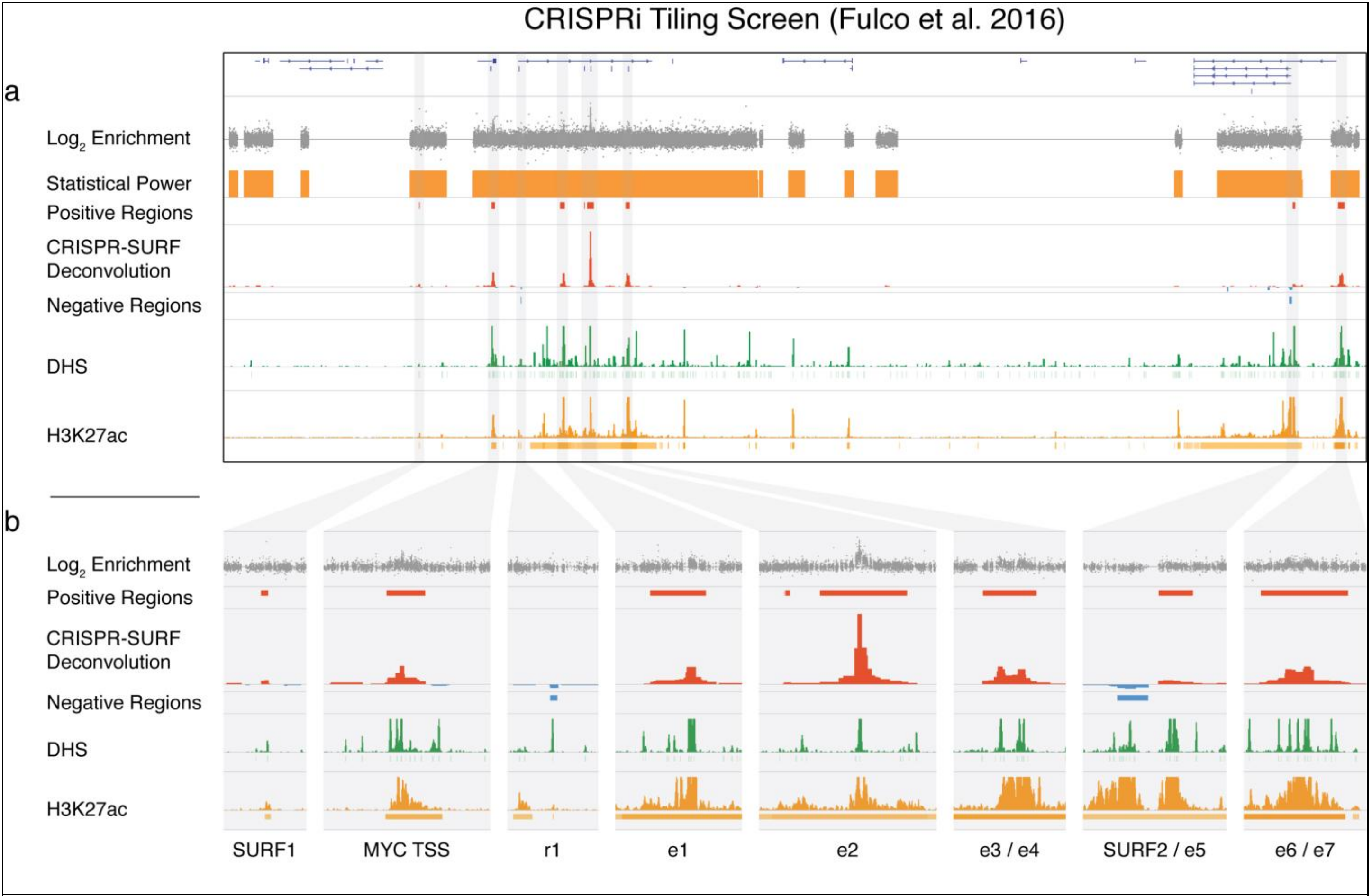
[1]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Molecular Pathology Unit, Center for Cancer Research, Center for Computational and Integrative Biology, Massachusetts General Hospital, Charlestown, MA, USA. [3]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [4]Department of Systems Biology, Harvard Medical School, Boston, MA, USA. [5]Division of Hematology/Oncology, Boston Children's Hospital, Boston, MA, USA. [6]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [7]Harvard Stem Cell Institute, Department of Pediatrics, Harvard Medical School, Boston, MA, USA. [8]Gene Therapy Program, Harvard Medical School, Boston, MA, USA. [9]Dana-Farber/Boston Children's Cancer and Blood Disorders Center, Boston, MA, USA. [10]Department of Pathology, Harvard Medical School, Boston, MA, USA. [11]Howard Hughes Medical Institute, Boston, MA, USA. [12]Harvard Society of Fellows, Harvard University, Cambridge, MA, USA. [13]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. *e-mail: bauer@bloodgroup.tch.harvard.edu; lpinello@mgh.harvard.edu



Supplementary Figure 1

Reanalysis of the CRISPR–Cas9 tiling screen from Canver et al.[1].

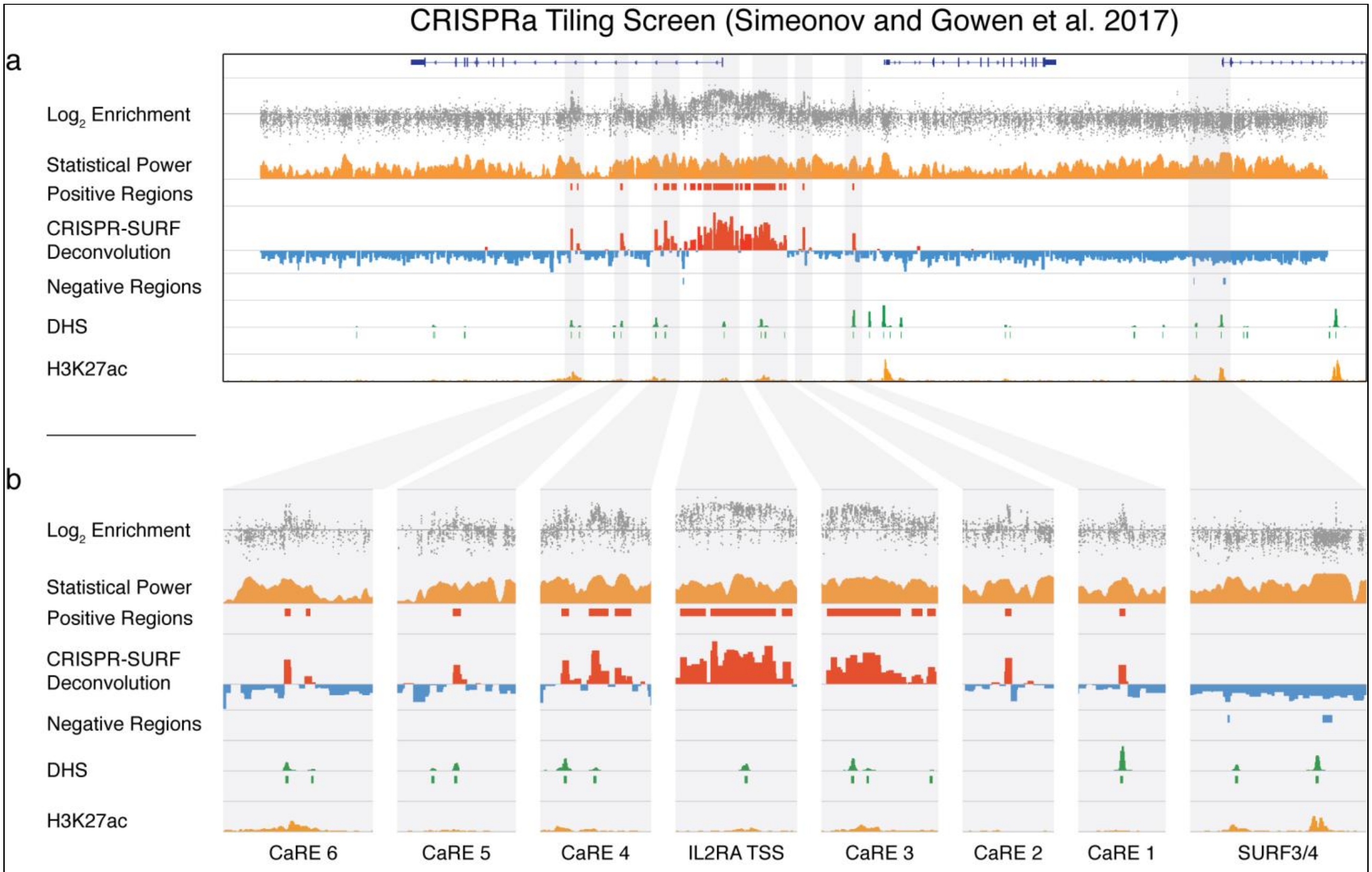
(a) An overview of the *BCL11A* CRISPR–Cas9 enhancer dissection tiling screen. (b) Zoom-in panels of DHS +55, +58, +62, and *BCL11A* exon 2 to highlight critical regions identified by CRISPR-SURF. All significant regions identified with FDR < 0.05. All panels are shown at same scale.



Supplementary Figure 2

Reanalysis of a CRISPRi tiling screen from Fulco et al.[2].

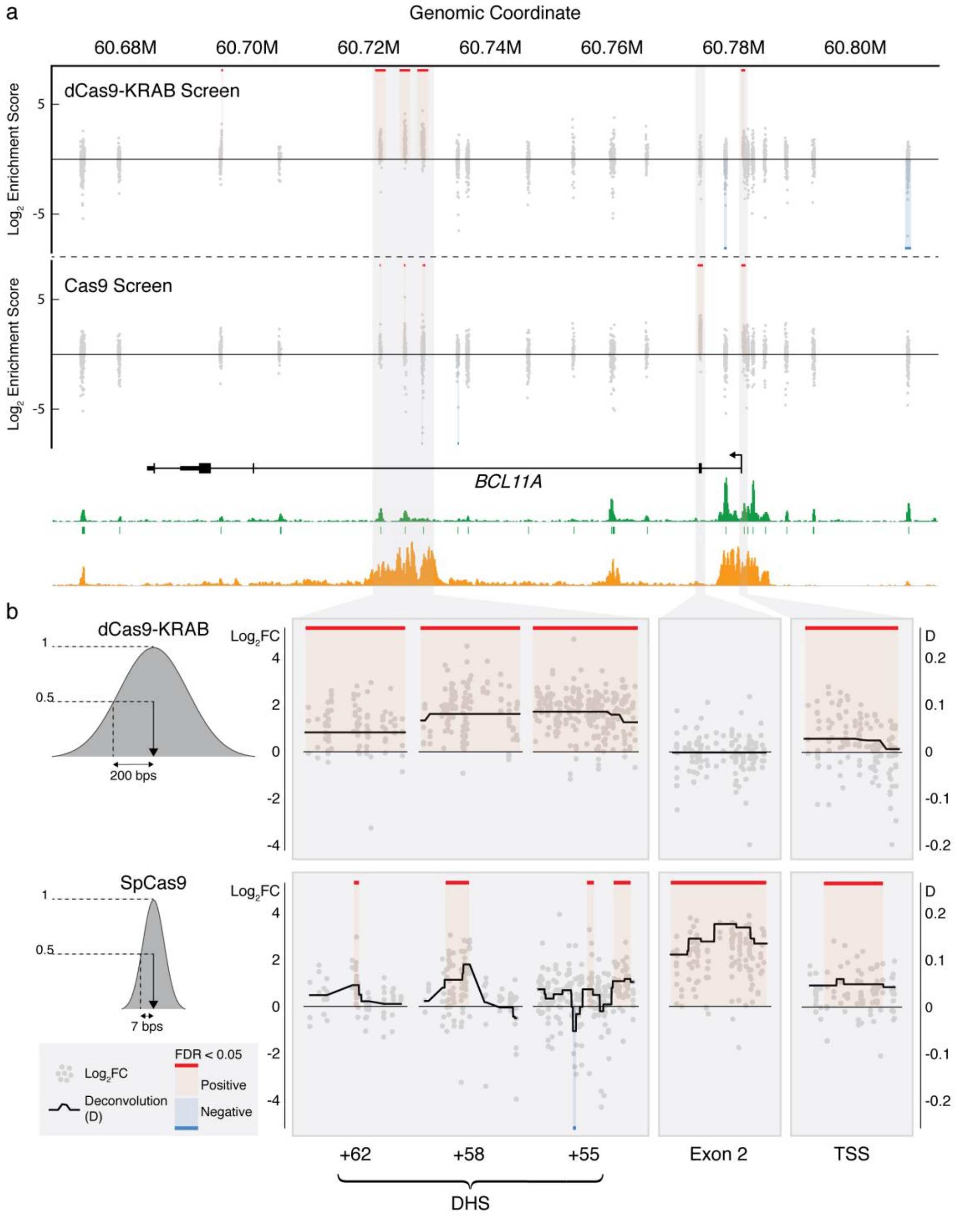
(a) An overview of the *MYC* CRISPRi enhancer discovery tiling screen. (b) Zoom-in panels of *MYC* TSS, e1–e7, and r1 (regions identified in Fulco et al.[2]) along with newly identified regions by CRISPR-SURF (SURF1 and SURF2). All significant regions identified with FDR < 0.05. All panels are shown at same scale.



Supplementary Figure 3

Reanalysis of a CRISPRa tiling screen from Simeonov et al.[3].

(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. 3) 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.



Supplementary Figure 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.

Supplementary Notes

Supplementary Note 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)
```

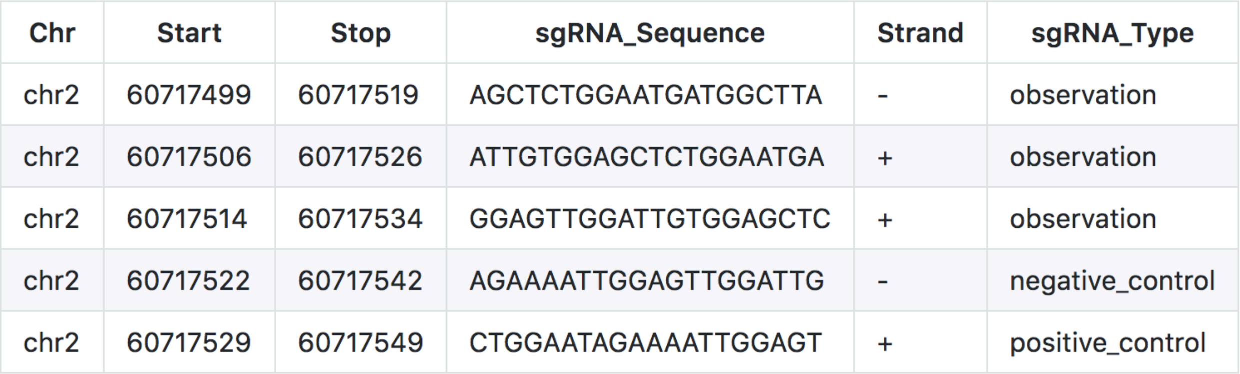
```
-dropout, --dropout_penalty
      The dropout penalty removes sgRNAs that have a 0 count in any of the
control/sample replicates. (Default: True)
-TTTT, --TTTT_penalty
      The TTTT penalty removes sgRNAs that have a homopolymer stretch of Ts >= 4.
(Default: True)
-sgRNA_length, --sgRNA_length
      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 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:

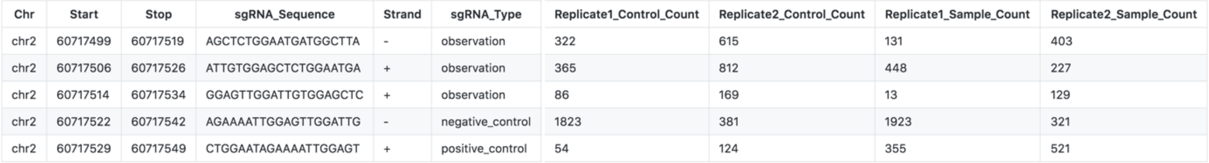
```
docker run -v ${PWD}/:/DATA -w /DATA pinellolab/crisprsurf SURF_count -f
sgRNA_library_file.csv -control_fastqs rep1_control.fastq rep2_control.fastq -
sample_fastqs rep1_sample.fastq rep2_sample.fastq -nuclease cas9 -pert indel
```

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 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

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[1] 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

Supplementary Note 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 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 $log_2 FC$ 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[2]:

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

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:

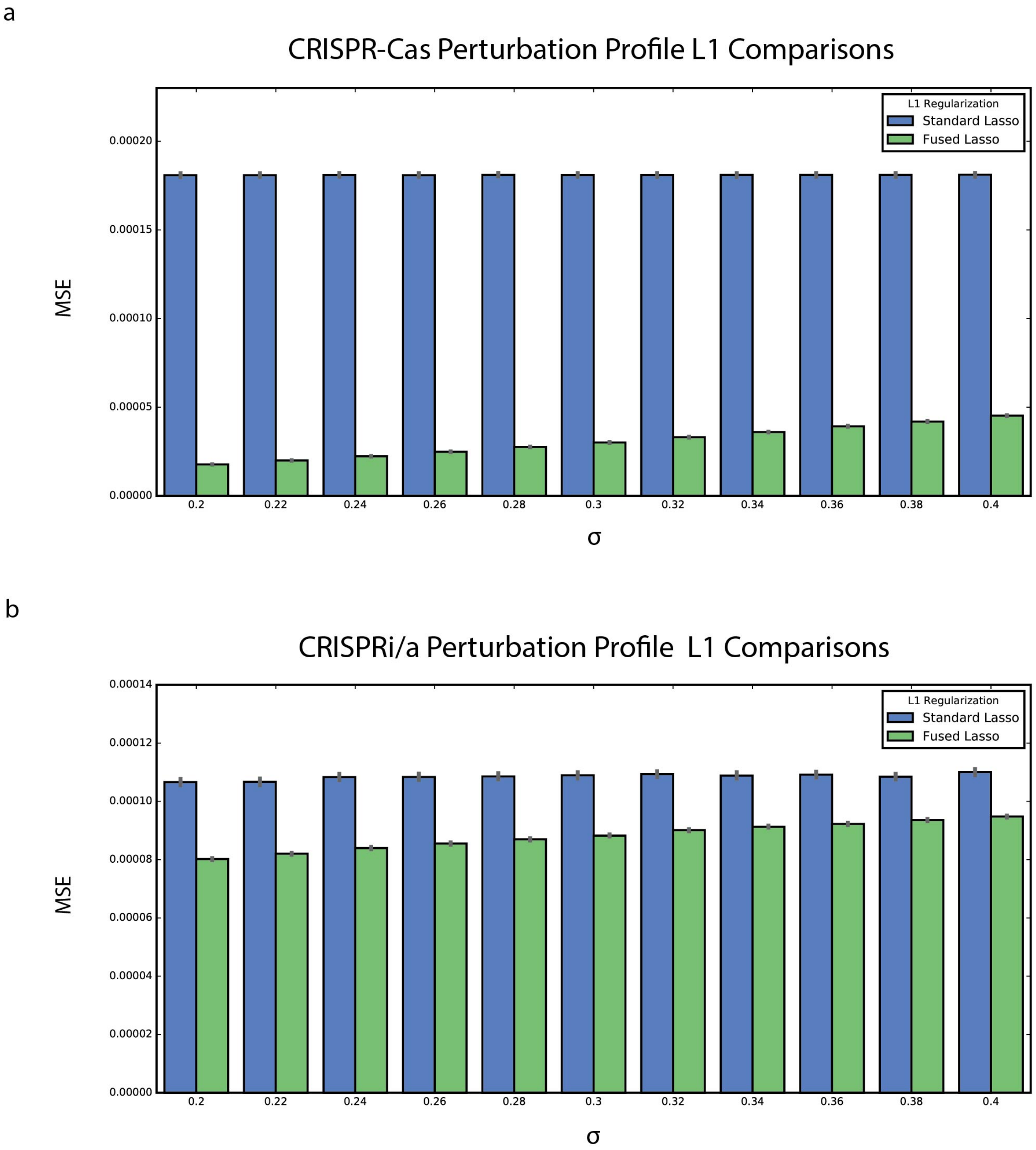
$$X = MC$$

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

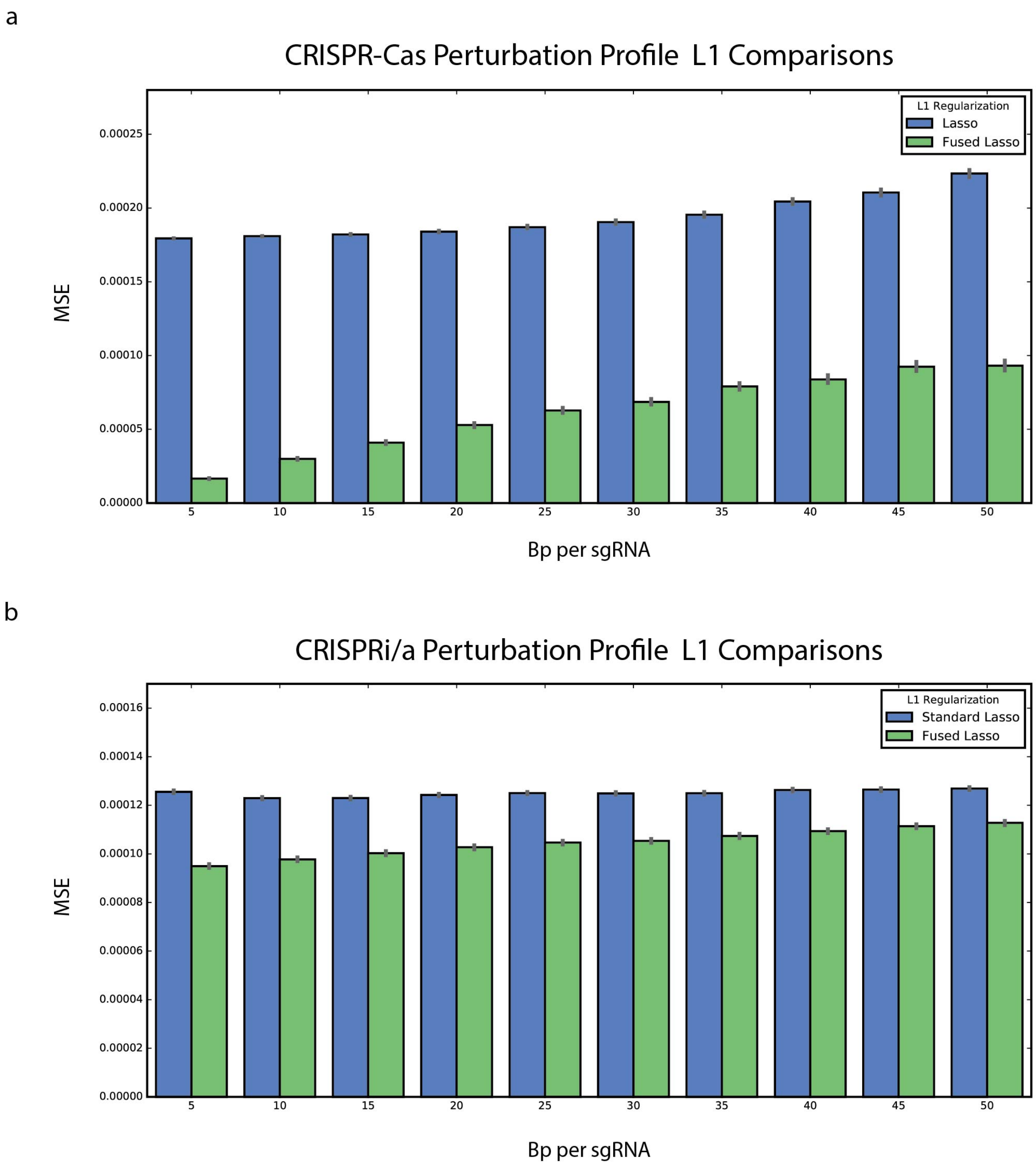
where $M \in \mathbb{R}^{nxp}$ and $C \in \mathbb{R}^{pxp}$.

$\hat{\beta}$ 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.



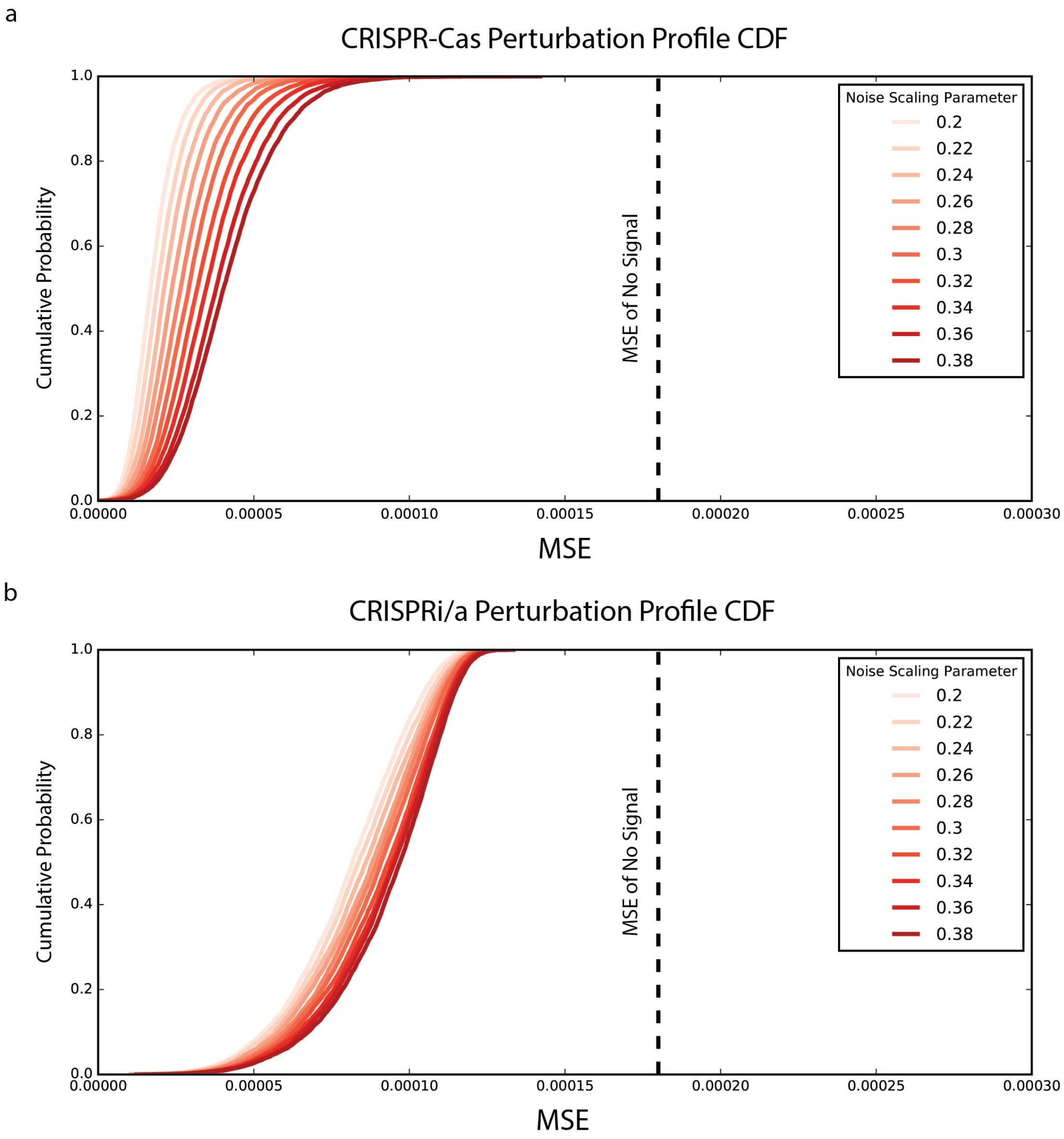
Supplementary Figure SN2.1: Comparison of L1 regularization methods with varying noise
(a) Comparison of deconvolution mean-squared error (MSE) for the lasso and fused lasso L1 regularization methods with varying Gaussian noise for the CRISPR-Cas nuclease perturbation profile. A total of 1000 simulations 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.



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.

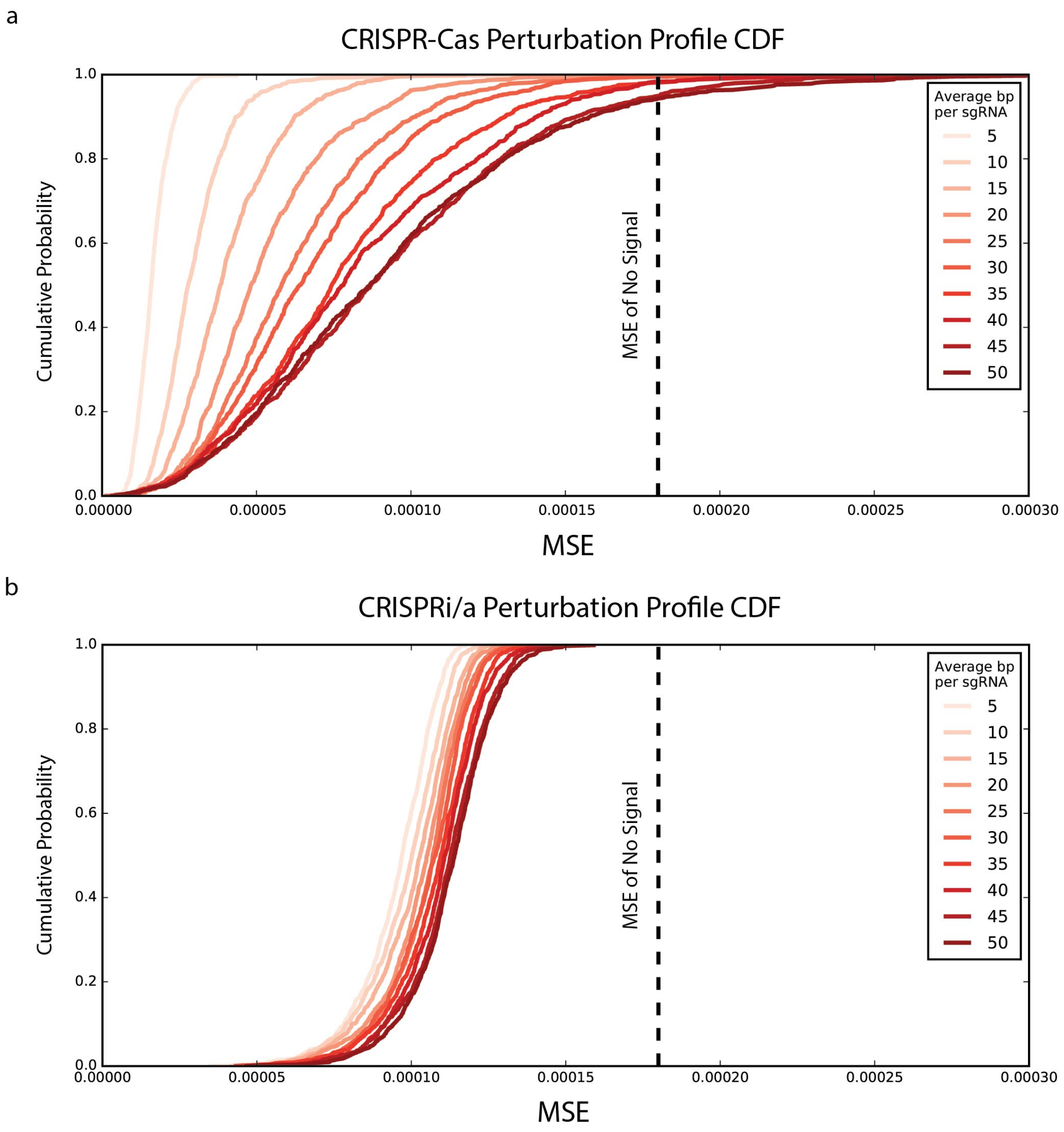
(b) Comparison of deconvolution mean-squared error (MSE) for the lasso and fused lasso L1 regularization methods with varying sgRNA targeting density for the CRISPRi/a perturbation profile. A total of 1000 simulations were performed for each comparison for both the lasso and fused lasso. Grey bars represent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

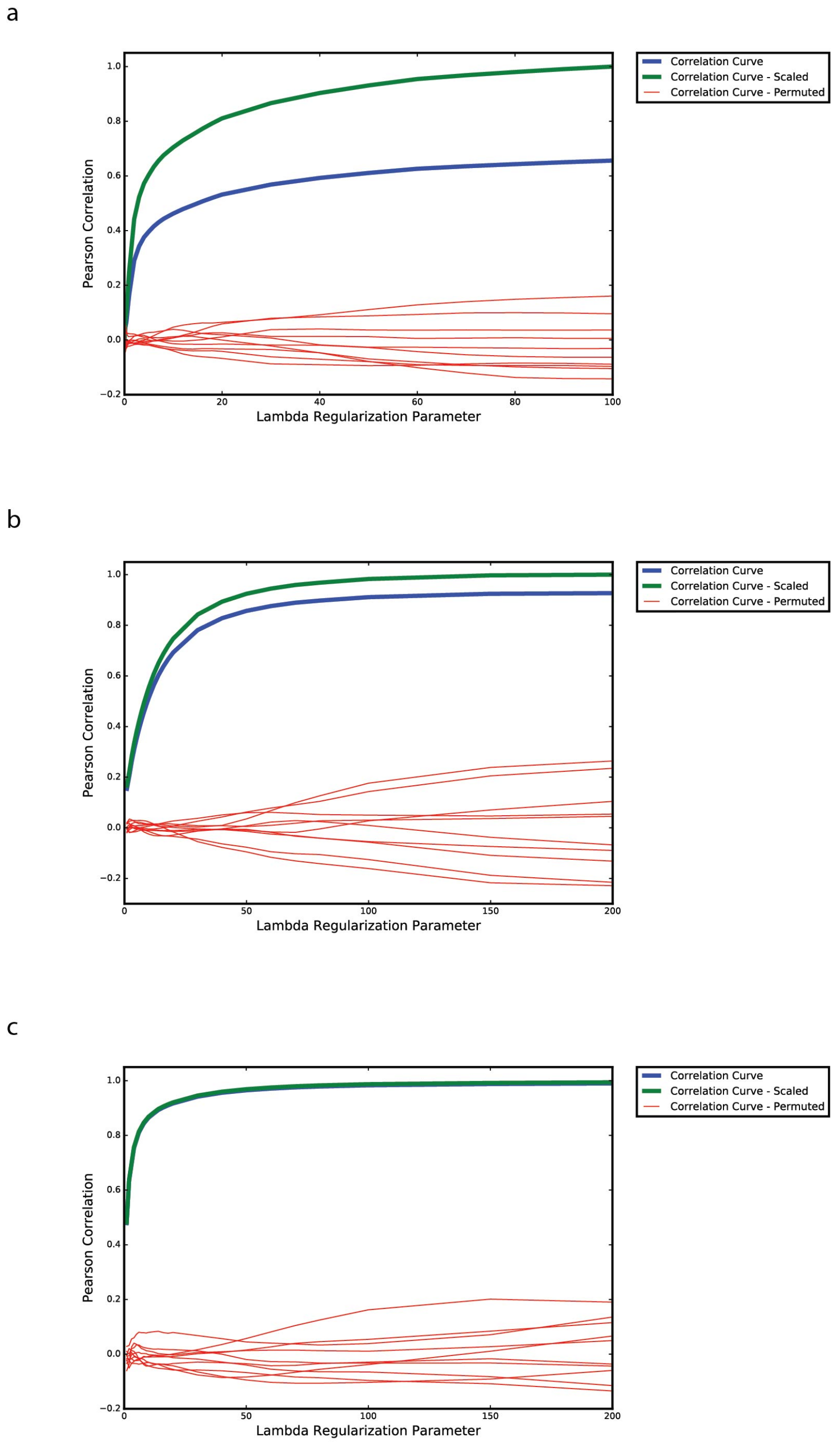
The convolution matrix C and regularization parameter λ need to be specified to perform the deconvolution algorithm for CRISPR-SURF analysis. The perturbation profile, encoded within C, represents the perturbation range of the CRISPR screening modality employed. The perturbation profile is represented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

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

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

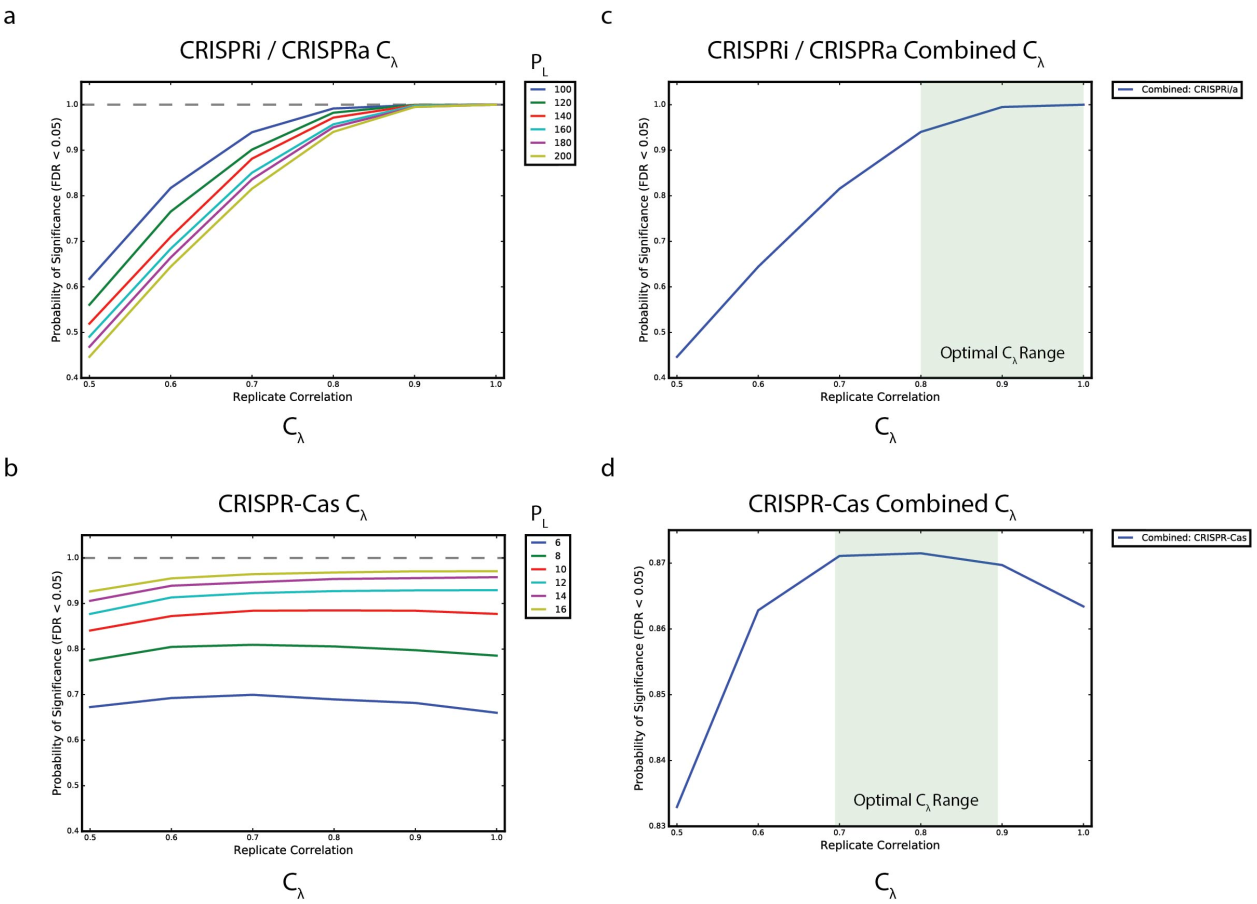
The selection of a characteristic perturbation length P_L is different for varying CRISPR screening modalities. CRISPR-Cas nucleases introduce indel mutations to the DNA sequence and provide a much narrower perturbation profile compared to CRISPRi and CRISPRa strategies which can epigenetically modify 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[3]. 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[4]. 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[5]. We further elaborate on perturbation profiles for different CRISPR technologies in **Supplementary Note 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[6] (n = 6), **(b)** Fulco et al. 2016[7] (n = 2) and **(c)** Simeonov and Gowen et al. 2017[8] (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[6,7,8]. 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:

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

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[9].

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 $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \dots, \hat{\beta}_p)$ 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 $\frac{1}{N}\mathrm{sum}\left(H_{a_i} \geq \alpha\right)$

Supplementary Note 3: Re-Analysis of Published Datasets

Canver et al. 2015[6]: 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 1**).

Fulco et al. 2016[7]: 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**). 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[10].

Simeonov and Gowen et al 2017[8]: 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 3**).

Supplementary Tables for CRISPR-SURF Re-Analyses

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 SN3.1: Replicate correlations in Canver et al. 2015[6] re-analysis
The correlations (Pearson's r) across experimental replicates (n = 6) pre- and post-deconvolution for the Canver et al. 2015 study.

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

Supplementary Table SN3.2: Replicate correlations in Fulco et al. 2016[7] re-analysis
The correlations (Pearson's r) across experimental replicates (n = 2) pre- and post-deconvolution for the Fulco et al. 2016 study.

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

Supplementary Table SN3.3: Replicate correlations in Simeonov and Gowen et al. 2017[8] 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.

P_L (bp)	Overlap against Reference (max 6)
5	6
7	6

10	6
15	6
20	6
25	6
30	5

Supplementary Table SN3.4: Assessment of characteristic perturbation length for Canver et al. 2015[6]

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 9)
100	9
150	9
200	9
250	9
300	9
350	9
400	8

Supplementary Table SN3.5: Assessment of characteristic perturbation length for Fulco et al. 2016[7]

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 7)
100	7
150	7
200	7
250	7
300	7
350	7
400	7

Supplementary Table SN3.6: Assessment of characteristic perturbation length for Simeonov and Gowen et al. 2017[8]

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.

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 SN3.7: Assessment of $C_\lambda - \lambda$ for Canver et al. 2015[6]

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 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 SN3.8: Assessment of $C_\lambda - \lambda$ for Fulco et al. 2016[7]

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 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

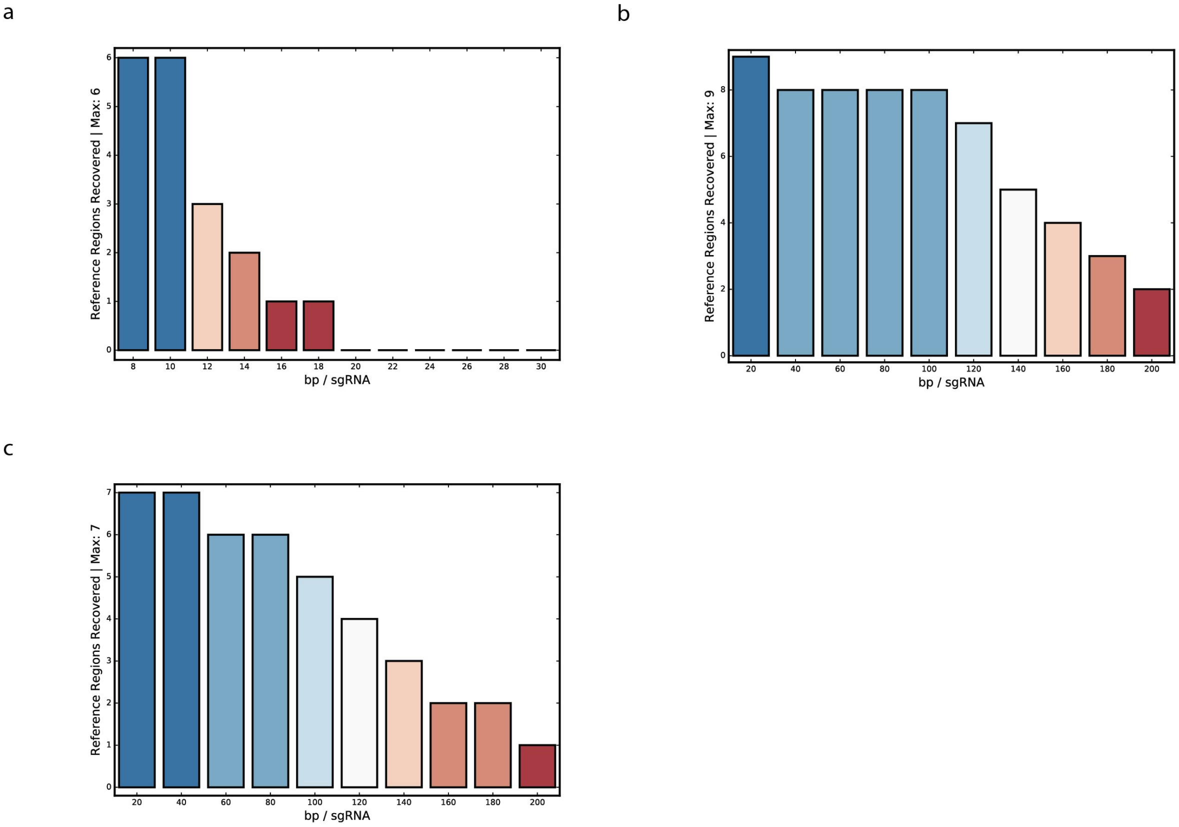
Supplementary Table SN3.9: Assessment of $C_\lambda - \lambda$ for Simeonov and Gowen et al. 2017[8]

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.

Supplementary Note 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[6], **(b)** Fulco et al. 2016[7] and **(c)** Simeonov and Gowen et al. 2017[8]. 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 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

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 SN4.1: Downsampling sgRNAs assessment for Canver et al. 2015[6]
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 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 SN4.2: Downsampling sgRNAs assessment for Fulco et al. 2016[7]
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 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

Supplementary Table SN4.3: Downsampling sgRNAs assessment for Simeonov and Gowen et al. 2017[8]

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.

Supplementary Note 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[6] (CRISPR-Cas9), Fulco et al. 2016[7] (CRISPRi), and Simeonov and Gowen et al. 2017[8] (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, 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, 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 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.

Supplementary Note 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 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.

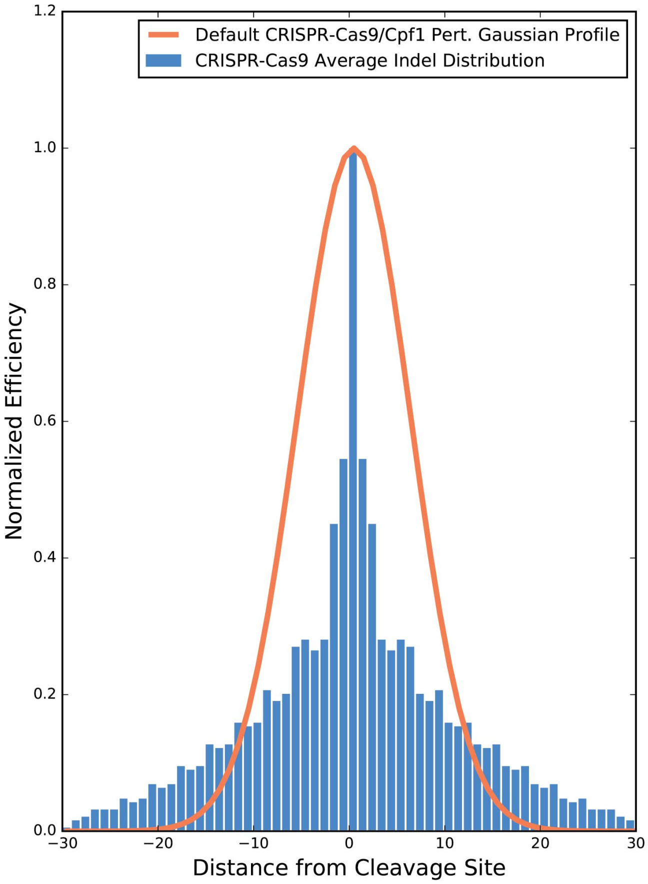
Next, the convolution operation readily-adapts to varying sgRNA targeting densities (non-uniform spacing) intrinsic in CRISPR tiling screen data due to sgRNA design limitations (PAM constraints). This is important because the degree of shared information used for the reconstruction of the genomic regulatory 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 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 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[11]. 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[11] (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[5], 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)[4]. 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.

Supplementary Note 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[6,12]. 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[13]. 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[6]. Using the *DNA Striker* tool[12], 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[14]. 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[6]. 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.

References

1. Robinson, J. T. et al. Integrative Genomics Viewer. *Nature Biotechnology*, **29**, 24-26. (2011).

2. Tibshirani, R. and Taylor, J. The solution path of the generalized lasso. *The Annals of Statistics*, **39**, 1335-1371 (2011).

3. van Overbeek, M. et al. DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks. *Molecular Cell*, **63**, 633-646 (2016).

4. Gilbert, L. A. *et al.* Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell,* **159**, 647–661 (2014).

5. Thakore, P. et al. Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nature Methods*, **12**, 1143-1149 (2015).

6. Canver, M. C. et al. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature,* **527**, 192–197 (2015).

7. Fulco, C. et al. Systematic mapping of functional enhancer–promoter connections with CRISPR interference. *Science*, **354**, 769-773 (2016).

8. Simeonov, D. R. and Gowen, B. G. et al. Discovery of stimulation-responsive immune enhancers with CRISPR activation. *Nature*, **549**, 111-115 (2017).

9. Benjamini, Y. and Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, **29**, 1165-1188 (2001).

10. Cho, S. W. and Xu, J. et al. Promoter of lncRNA gene PVT1 is a tumor-suppressor DNA boundary element. *Cell*, (2018).

11. Allen, F. et al. Mutations generated by repair of Cas9-induced double strand breaks are predictable from surrounding sequence. *bioRxiv*. (2018).

12. Canver, M. C., Lessard, S., Pinello, L. et al. Variant-aware saturating mutagenesis using multiple Cas9 nucleases identifies regulatory elements at trait-associated loci. *Nature Genetics*, **49**, 625-634 (2017).

13. Funnell, A. P. 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).

14. Chen, B. et al. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell*, **155**, 1479-1491 (2013).