
Engineered human Argonaute 2 for targeted C-to-U RNA base editing

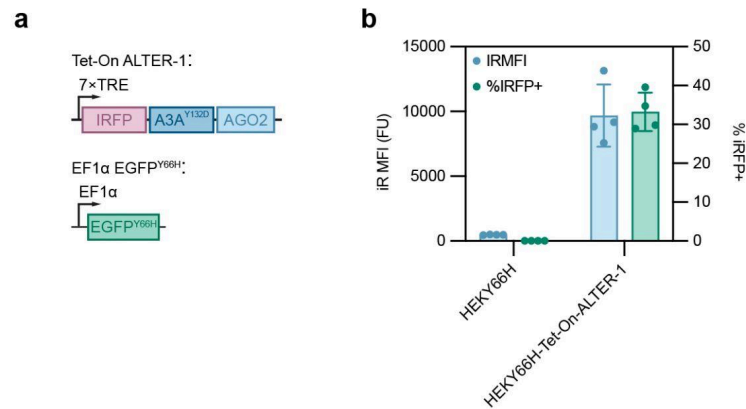
In the format provided by the
authors and unedited

Table of Contents

Table of Contents.....	2
Supplementary Figures.....	4
Supplementary Figure 1 Induction of HEKY66H-Tet-On-ALTER-1 Cells.....	4
Supplementary Figure 2 Guide SAR Mismatched shRNA Ends and GYM Score.....	5
Supplementary Figure 3 GW182 Binding Ablation Mutants.....	6
Supplementary Figure 4 Testing EGFPY66H CURE and RESCUE Guides.....	6
Supplementary Figure 5 qPCR Primer Efficiency Curves.....	7
Supplementary Figure 6 CURE and ALTER Match A3A Consensus, RESCUE Does Not.....	8
Supplementary Figure 7 Unique and Shared ALTER Off-Target Edits.....	9
Supplementary Figure 8 Alignment Frequency of miRNA Seed Binding Motifs Near All Editor Off-Targets Differs From Random UC. Appears to be due to A3A Consensus Sequence.....	10
Supplementary Figure 9 RNAseq Geometric Distance Heatmaps.....	11
Supplementary Figure 10 GLM-PCA Analysis of RNAseq Data.....	12
Supplementary Figure 11 Volcano Plot ALTER-4 PPIB.....	13
Supplementary Figure 12 Volcano Plot ALTER-4 TYMS.....	13
Supplementary Figure 13 Volcano Plot ALTER-2 TYMS.....	14
Supplementary Figure 14 Volcano Plot ALTER-3 TYMS.....	14
Supplementary Figure 15 Volcano Plot A3A Overexpression.....	15
Supplementary Figure 16 Volcano Plot CURE-C2.....	15
Supplementary Figure 17 Volcano Plot CURE-N.....	16
Supplementary Figure 18 Volcano Plot RESCUE-S.....	16
Supplementary Figure 19 RNA Expression Heat Maps.....	17
Supplementary Figure 20 RNA Expression Heat Maps.....	18
Supplementary Figure 21 AAV Titers.....	18
Supplementary Figure 22 Representative Flow Cytometry Gating.....	19
Supplementary Figure 23 Combination of A3AK30G,Y132D with AGO2Q633R Δ PAZ.....	19
Supplementary Figure 24 Several Conditions Tested for EGFPY66H Editing Quantification Were Effective.....	20
Supplementary Table 1 GSEA Hallmark Set Analysis ALTER-4 PPIB.....	20
Supplementary Table 2 GSEA Hallmark Set Analysis ALTER-4 TYMS.....	21
Supplementary Table 3 GSEA Hallmark Set Analysis ALTER-2 TYMS.....	21
Supplementary Table 4 GSEA Hallmark Set Analysis ALTER-3 TYMS.....	22
Supplementary Discussion 1 ALTER Guide Design.....	23
Supplementary Discussion 2 Editing Quantification.....	27
Software tools.....	27
1. Read QC with cutadapt.....	27
2. Editing Quantification with CRISPResso2.....	28
Supplementary Discussion 3 RNAseq Data Processing.....	29
Software Tools Used.....	29

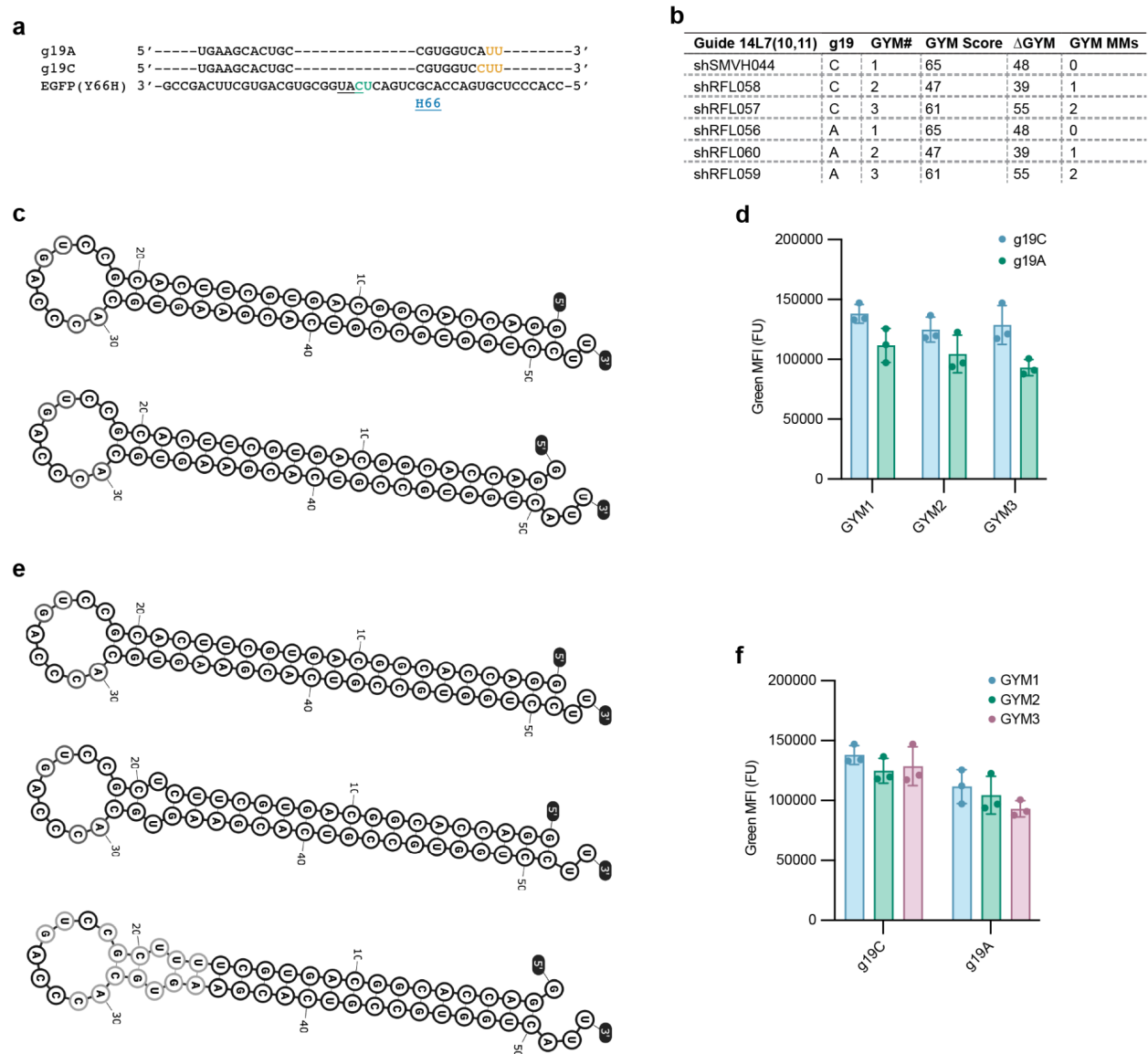
Python Libraries.....	29
R Packages.....	30
Preparing Genomic References.....	31
1. Adding Custom Contigs and Features to Reference Files.....	32
2. Building STAR Index.....	34
3. Making Genome Scatter Lists.....	35
4. Make Salmon Index.....	36
RNAseq Data Quality Control.....	37
RNAseq Variant Calling.....	38
1. Read Alignment.....	38
2. Prepare Files for HaploTypeCaller.....	38
3. HaploTypeCaller Parallelized Over Scatter Lists.....	40
4. Joint Genotyping.....	41
5. Variant Table Annotation and Final Filtering for C-to-U Off-Target Hits.....	43
RNAseq Expression Quantification and DEG Analysis.....	43
1. Quantifying Transcript-Level Raw Counts with Salmon.....	43
2. DEG Analysis, GSEA Analysis, and DEG GO Term Enrichment Analysis.....	43
Supplementary References.....	44

Supplementary Figures



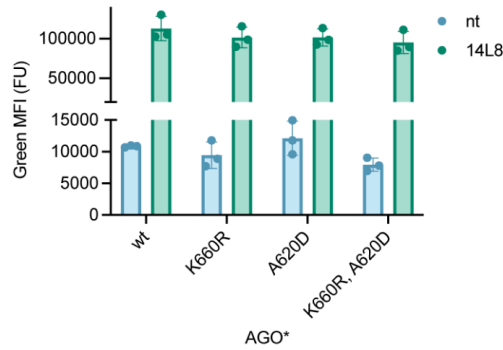
Supplementary Figure 1 | Induction of HEKY66H-Tet-On-ALTER-1 Cells

a, Diagrams of gene cassettes stably incorporated into HEKY66H-Tet-On-ALTER-1 cells. **b**, iR MFI and %iRFP positive metrics for HEKY66H and HEKY66H-Tet-On-ALTER-1 cells treated with 25 μ M doxycycline. Data are presented as mean $\pm$ s.d. of n=4 biological replicates. Data points indicate replicates; bar heights indicate means.



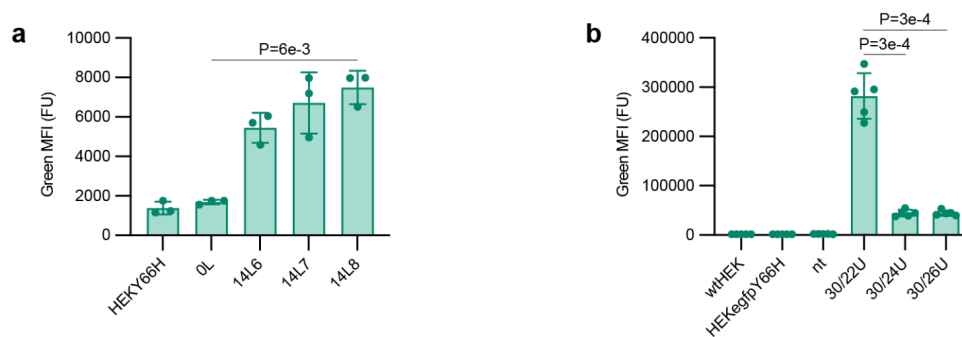
Supplementary Figure 2 | Guide SAR Mismatched shRNA Ends and GYM Score.

a, Expected base pairing of 14L7(11,10) guides with and without the g19C that results in a paired shRNA end. b, table of guide features. ΔGYM, difference in GYM score between -1-+2 and +1-+3.⁹⁹ c, Effect of g19C (top) vs g19A (bottom) on shRNA structure. d, Effect of g19C vs. g19A on fluorescence in the EGFP^{Y66H} editing assay. e, Effect of GYM1 (top), GYM2 (middle), GYM3 (bottom) of shRNA structure. GYM3 has 2 mismatches but one results in a G-U wobble pair. f, Effect of GYM motif 1 vs. 2 vs. 3 on fluorescence in the EGFP^{Y66H} editing assay. Data are presented as mean ± s.d. (d, f) of n=3 biological replicates. Data points indicate replicates; bar heights indicate means.



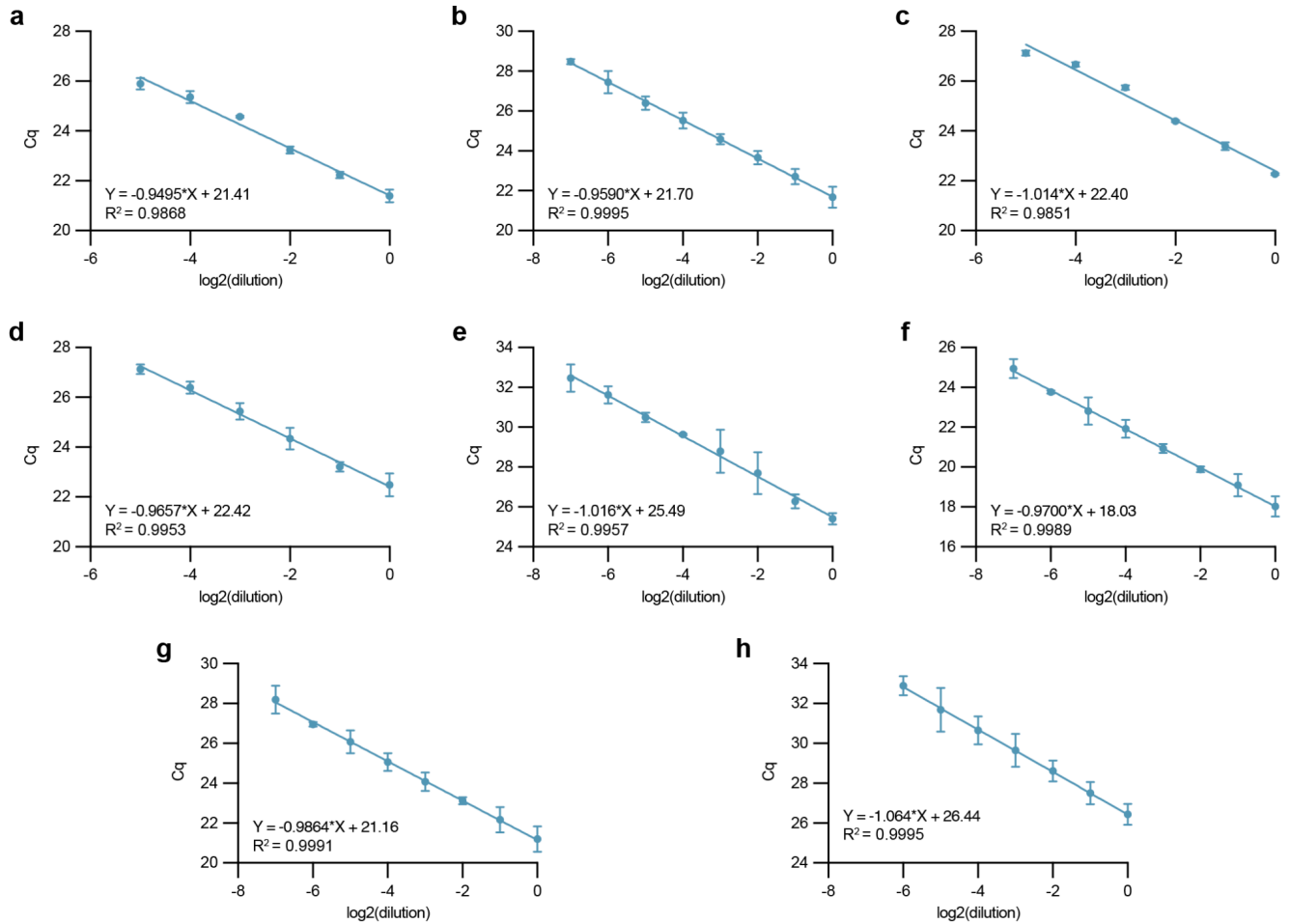
Supplementary Figure 3 | GW182 Binding Ablation Mutants.

To test whether removal of miRNA-type downregulation would improve ALTER editing, AGO2 mutants known to abolish GW182 binding fused to A3A^{Y132D} were tested in the EGFP^{Y66H} editing assay. These mutants did not lead to appreciable changes in ALTER activity or background signal. Data are presented as the mean $\pm$ s.d. of $n=3$ biological replicates. Data points indicate replicates; bar heights indicate means.



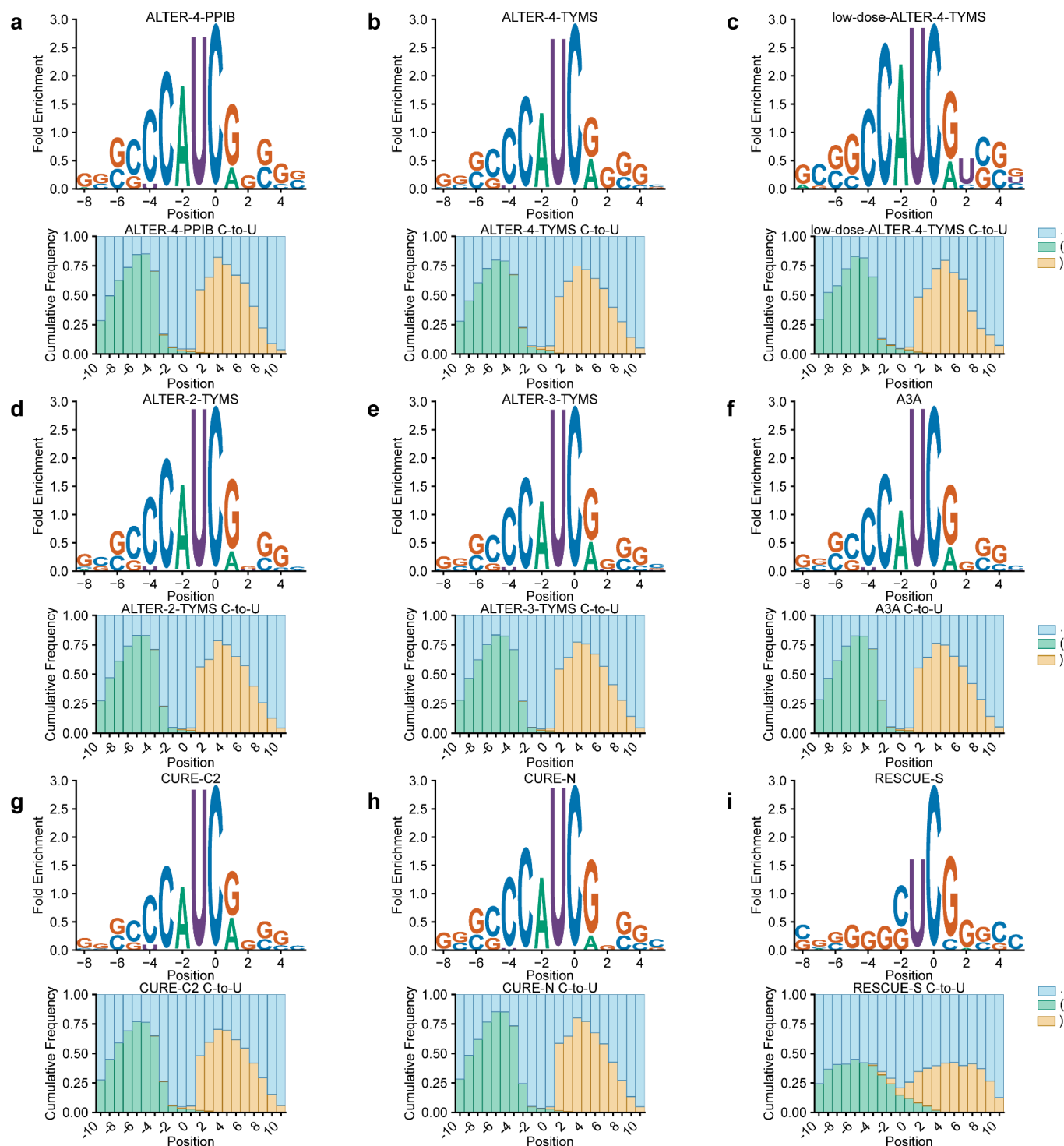
Supplementary Figure 4 | Testing EGFP^{Y66H} CURE and RESCUE Guides.

Fluorescence from HEK293T cells stably expressing EGFP^{Y66H} after transfection with **a**, CURE-N and different guides. **B**, RESCUE-S and different guides. Data are presented as the mean $\pm$ s.d. of biological replicates (**a**, $n=3$; **b**, $n=5$), with pairwise comparisons analyzed by unpaired Welch's t -test. Data points indicate replicates; bar heights indicate means.

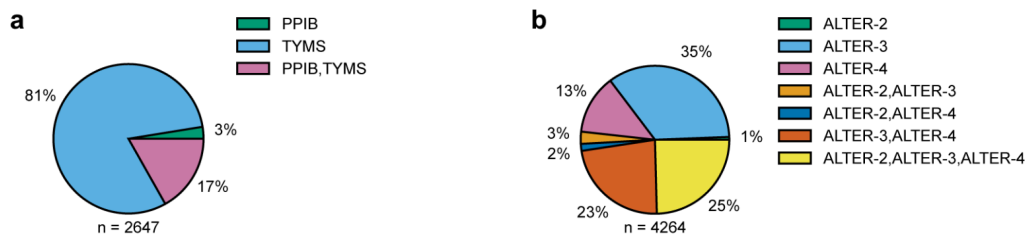


Supplementary Figure 5 | qPCR Primer Efficiency Curves

Primer Efficiency curves **a**, GAPDH. **b**, SNW1. **c**, HNRNPL. **d**, PPIB. **e**, SMARCA4. **f**, ACTB. **g**, TYMS. **h**, MALAT1. Data are presented as the mean $\pm$ s.d. of three technical replicates as is standard for qPCR primer validation. Data points indicate means.

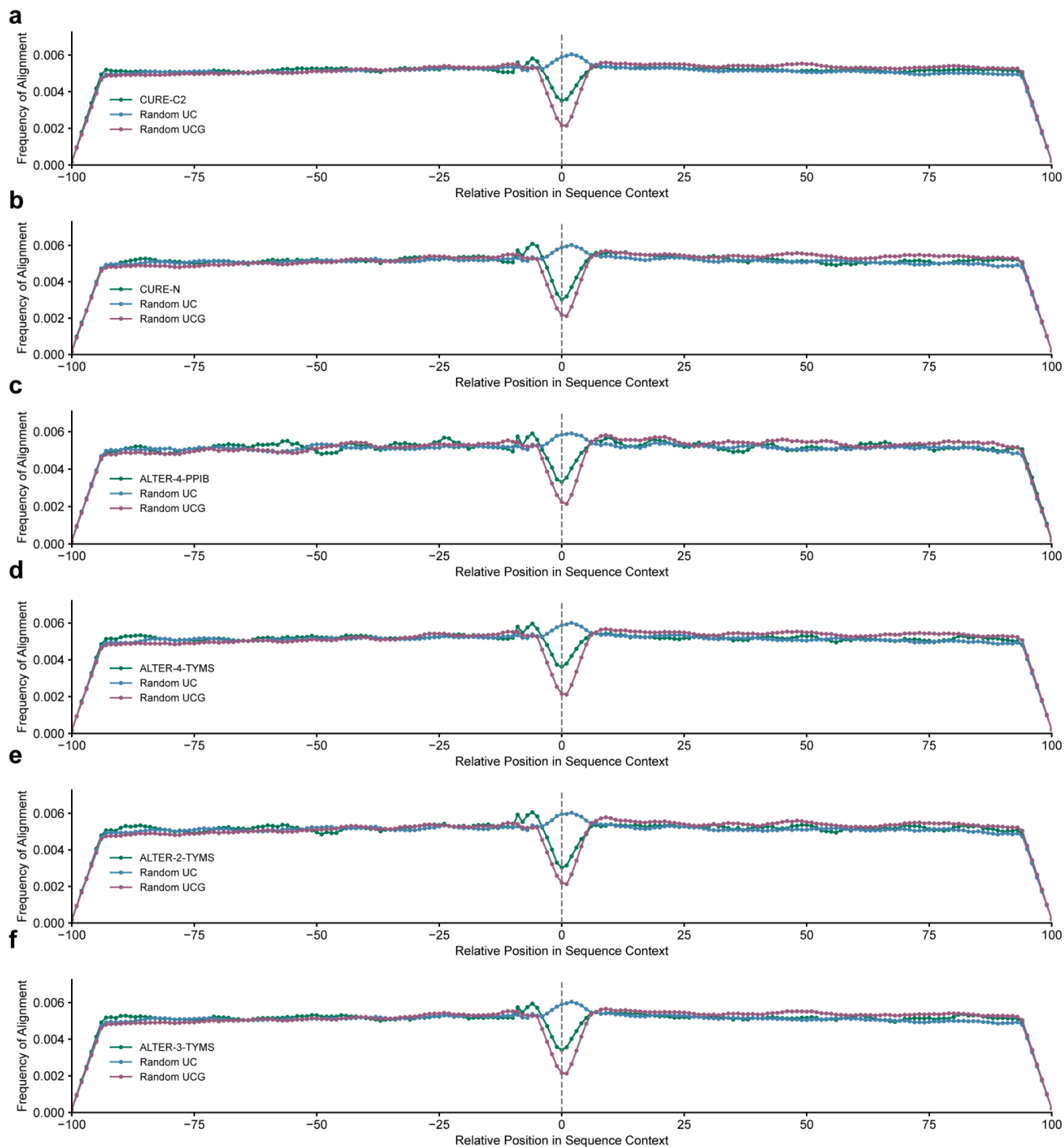


Supplementary Figure 6 | CURE and ALTER Match A3A Consensus, RESCUE Does Not
 (Top) logo of sequence consensus at C-to-U off-target edits, (bottom) secondary structure by dot bracket frequency of **a**, ALTER-4 PPIB **b**, ALTER-4-high TYMS **c**, ALTER-4-low TYMS **d**, ALTER-2 TYMS **e**, ALTER-3 TYMS **f**, A3A **g**, CURE-C2 **h**, CURE-N **i**, RESCUE-S.



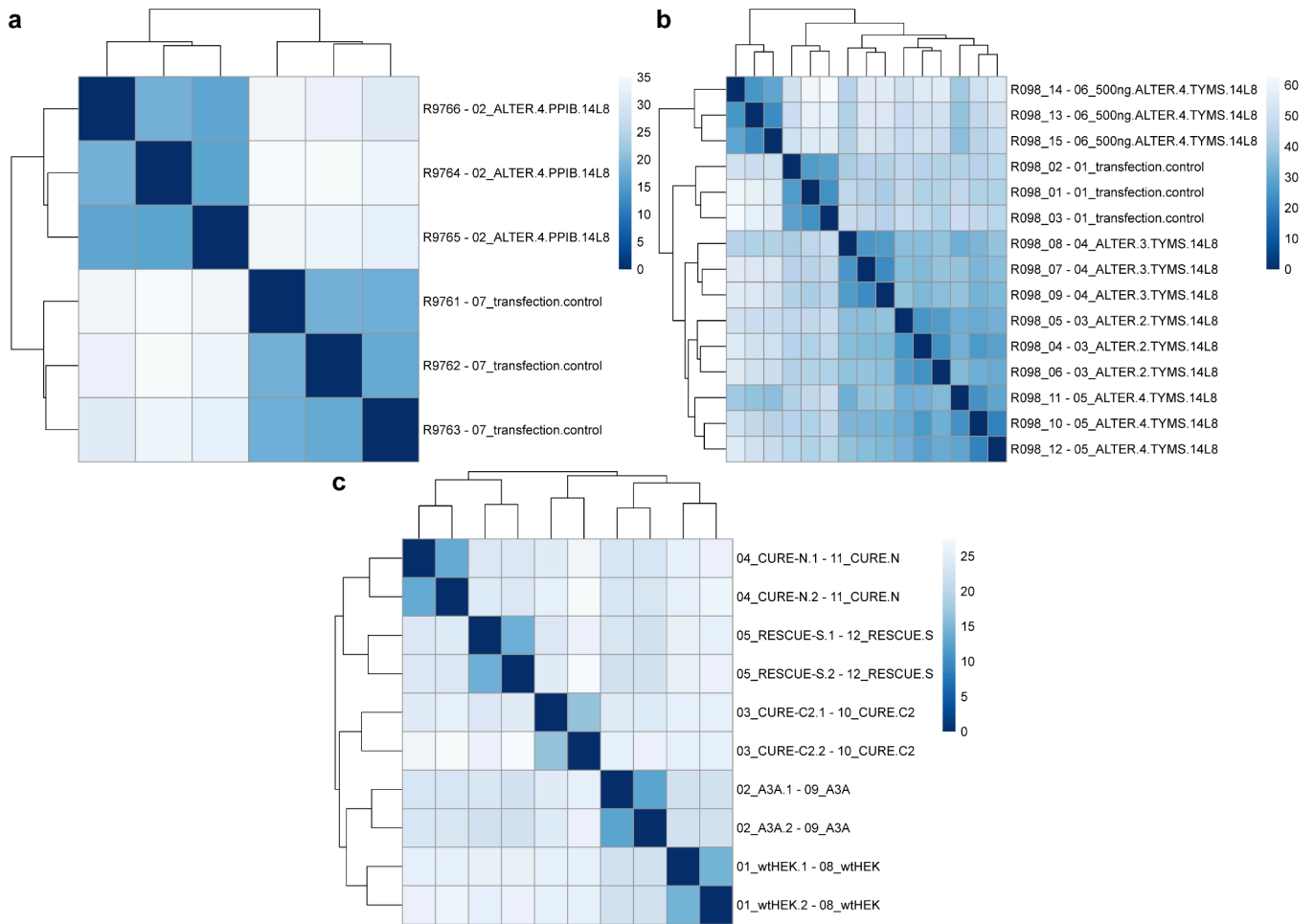
Supplementary Figure 7 | Unique and Shared ALTER Off-Target Edits

a, Between ALTER-4 PPIB and ALTER-4 TYMS. **b**, Between ALTERs 2, 3, and 4. For ALTER-4, PPIB and TYMS are grouped together



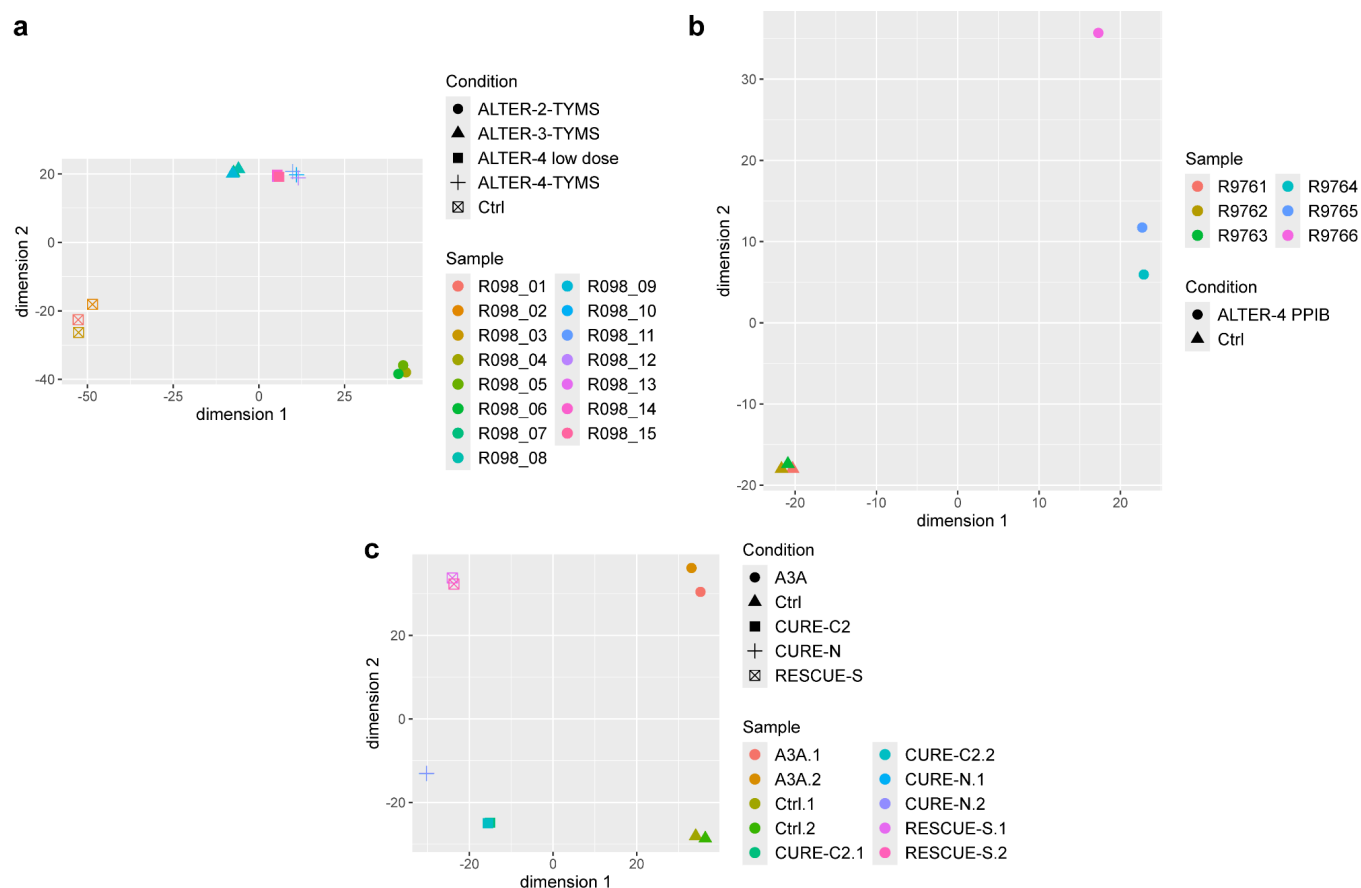
Supplementary Figure 8 | Alignment Frequency of miRNA Seed Binding Motifs Near All Editor Off-Targets Differs From Random UC. Appears to be due to A3A Consensus Sequence.

Alignment frequency of known miRNA seed motifs by position in the context window around an off target edit or a random sample of positions. **a**, CURE-C2. **b**, CURE-N. **c**, ALTER-4 PPIB **d**, ALTER-4 TYMS. **e**, ALTER-2 TYMS. **f**, ALTER-3 TYMS.



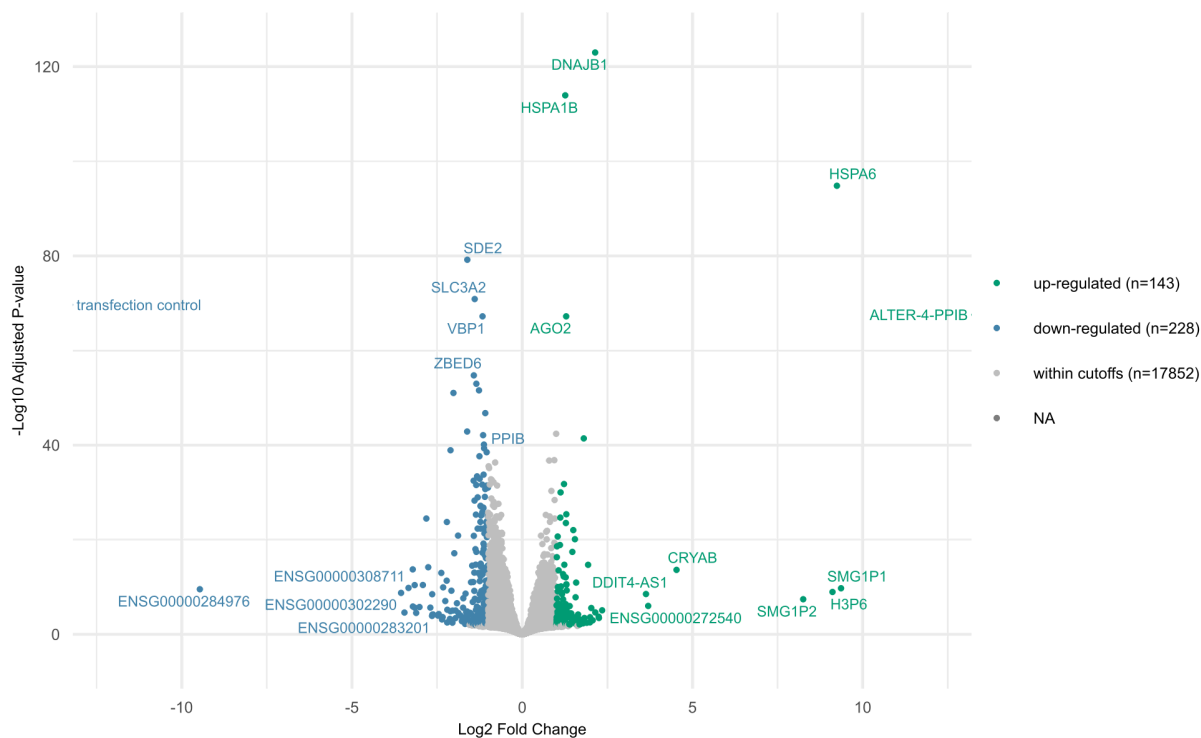
Supplementary Figure 9 | RNAseq Geometric Distance Heatmaps.

Heatmaps of the geometric distance between raw count data in RNAseq samples. **a**, **b** ALTER and transfection control. **c**, CURE and RESCUE data from BioProject PRJNA635732.



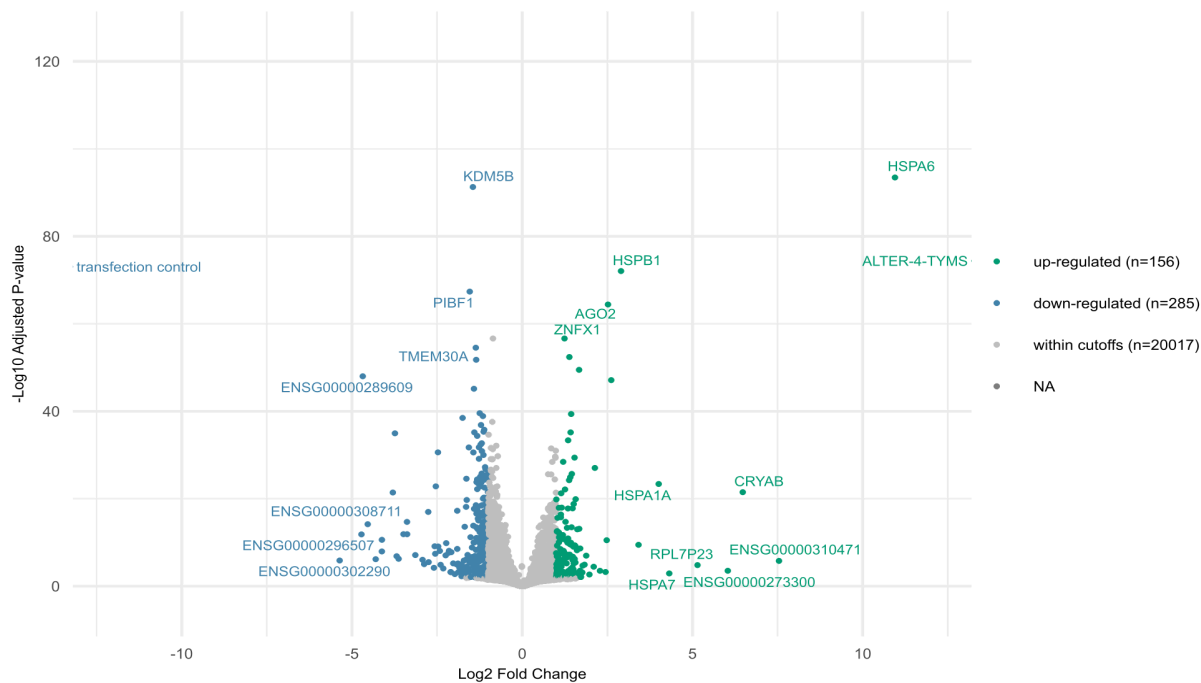
Supplementary Figure 10 | GLM-PCA Analysis of RNAseq Data

Generalized principle component analysis (GLM-PCA) of **a–b**, RNAseq raw counts for ALTER and transfection controls. **c**, RNAseq raw counts for analysis of CURE and RESCUE data from BioProject PRJNA635732.



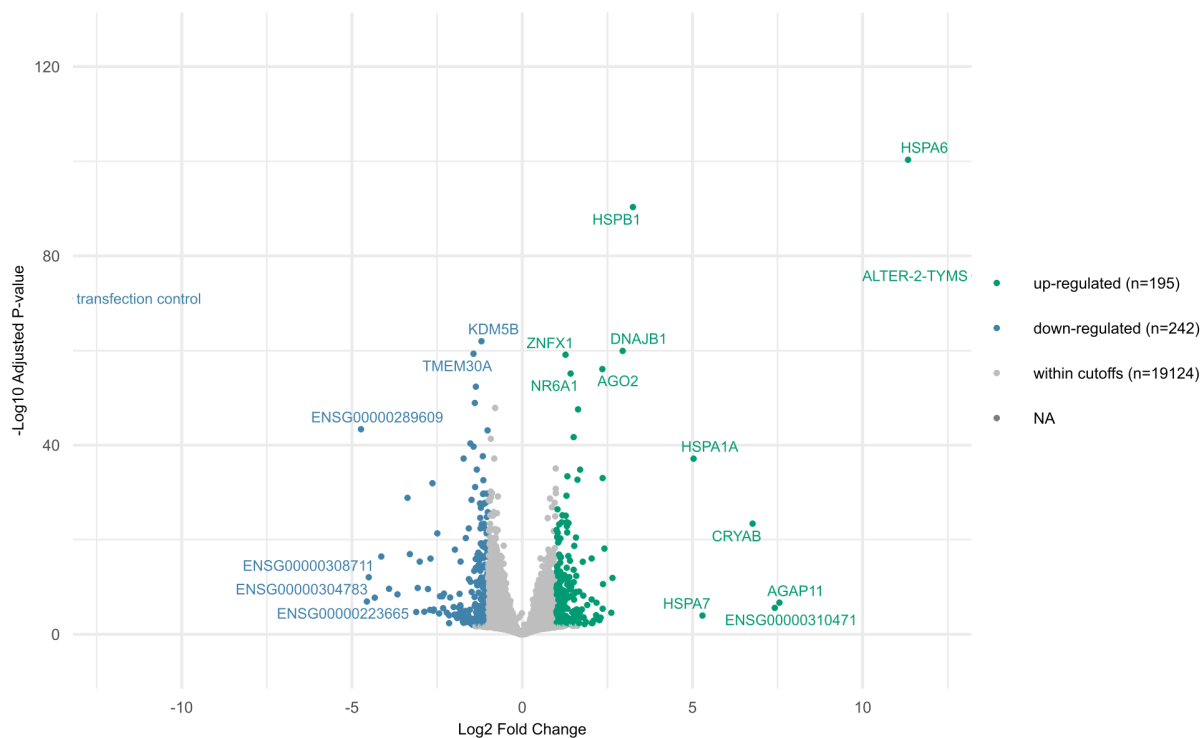
Supplementary Figure 11 | Volcano Plot ALTER-4 PPIB.

Volcano plot of transcript expression changes between HEK293T cells transfected with ALTER-4 PPIB and a transfection control.



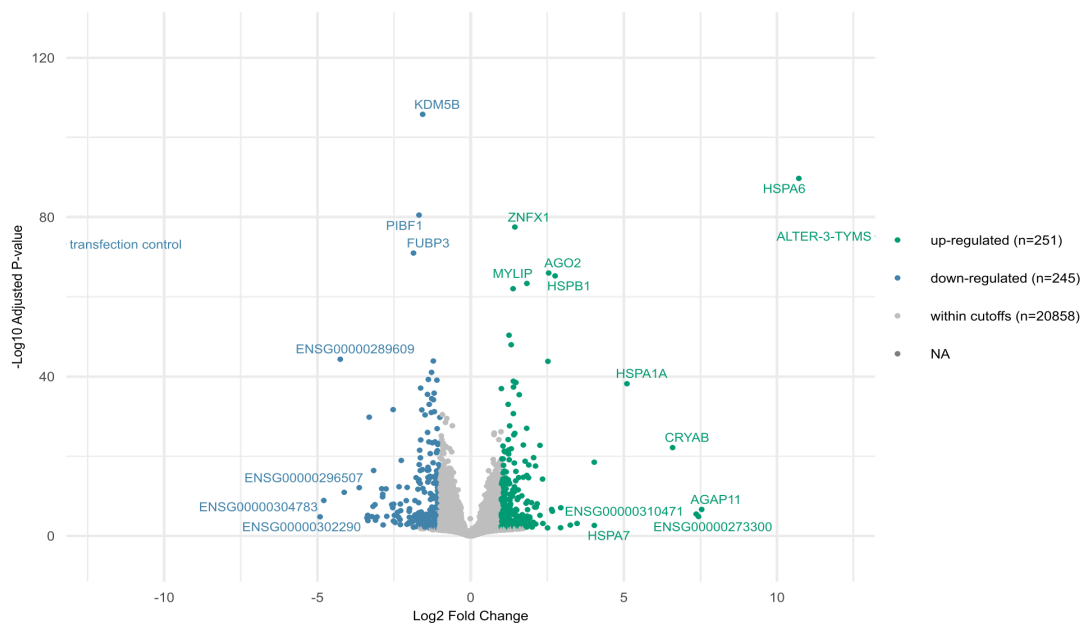
Supplementary Figure 12 | Volcano Plot ALTER-4 TYMS.

Volcano plot of transcript expression changes between HEK293T cells transfected with ALTER-4 TYMS and a transfection control.



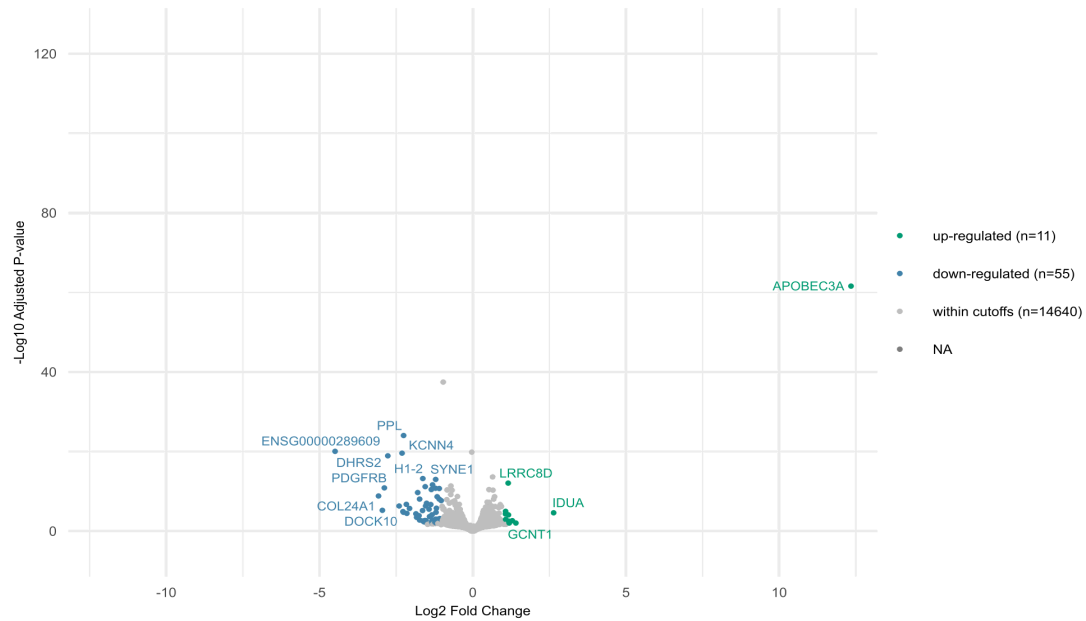
Supplementary Figure 13 | Volcano Plot ALTER-2 TYMS.

Volcano plot of transcript expression changes between HEK293T cells transfected with ALTER-4 TYMS and a transfection control.



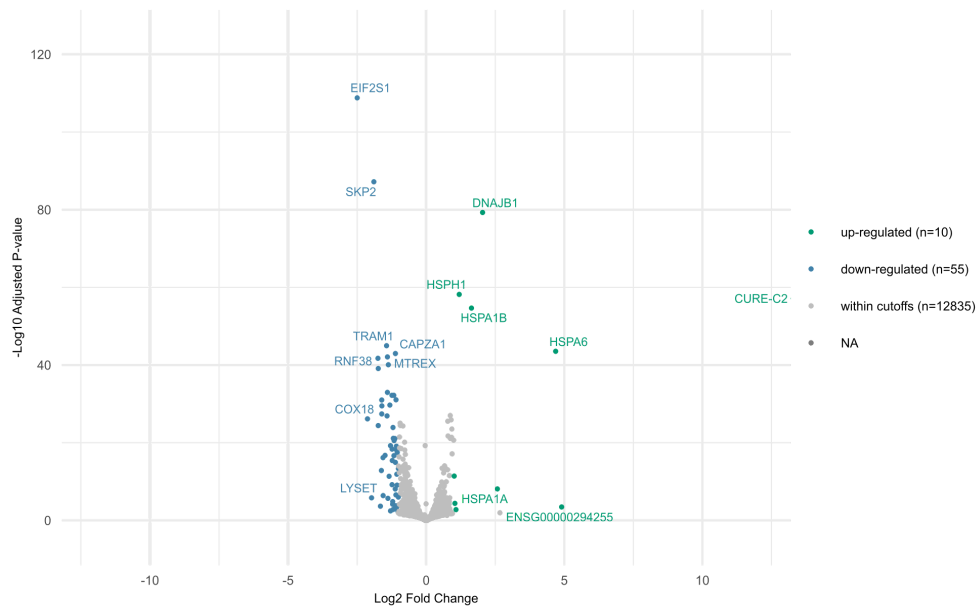
Supplementary Figure 14 | Volcano Plot ALTER-3 TYMS.

Volcano plot of transcript expression changes between HEK293T cells transfected with ALTER-4 TYMS and a transfection control.



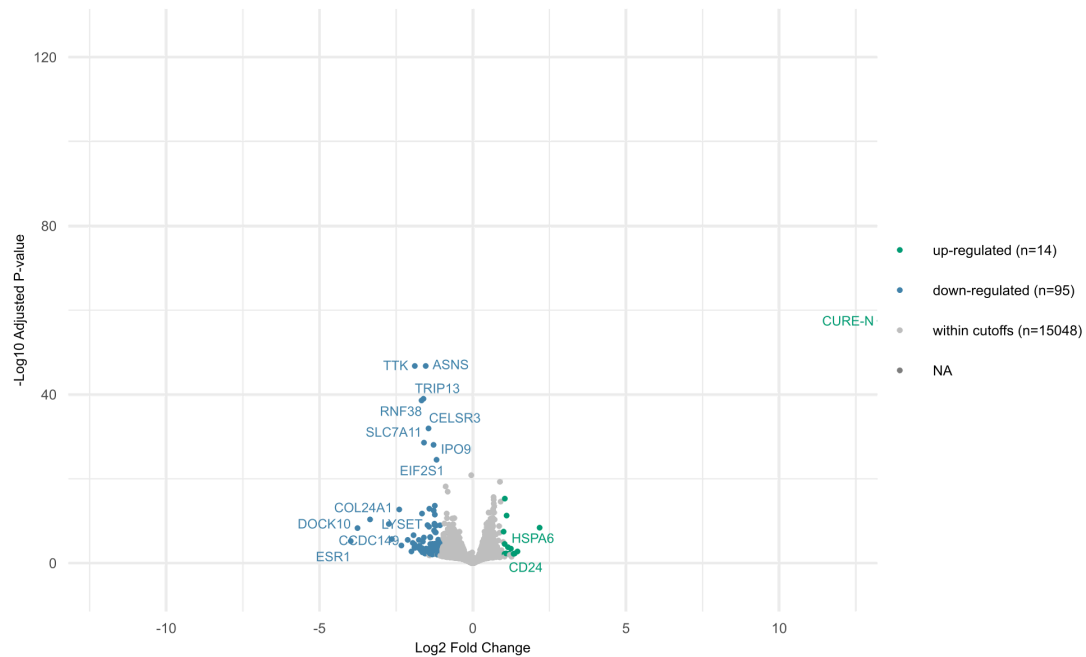
Supplementary Figure 15 | Volcano Plot A3A Overexpression.

Volcano plot of transcript expression changes between HEK293T cells transfected with a GFP editing reporter + A3A and the reporter alone. Read data from BioProject PRJNA635732.



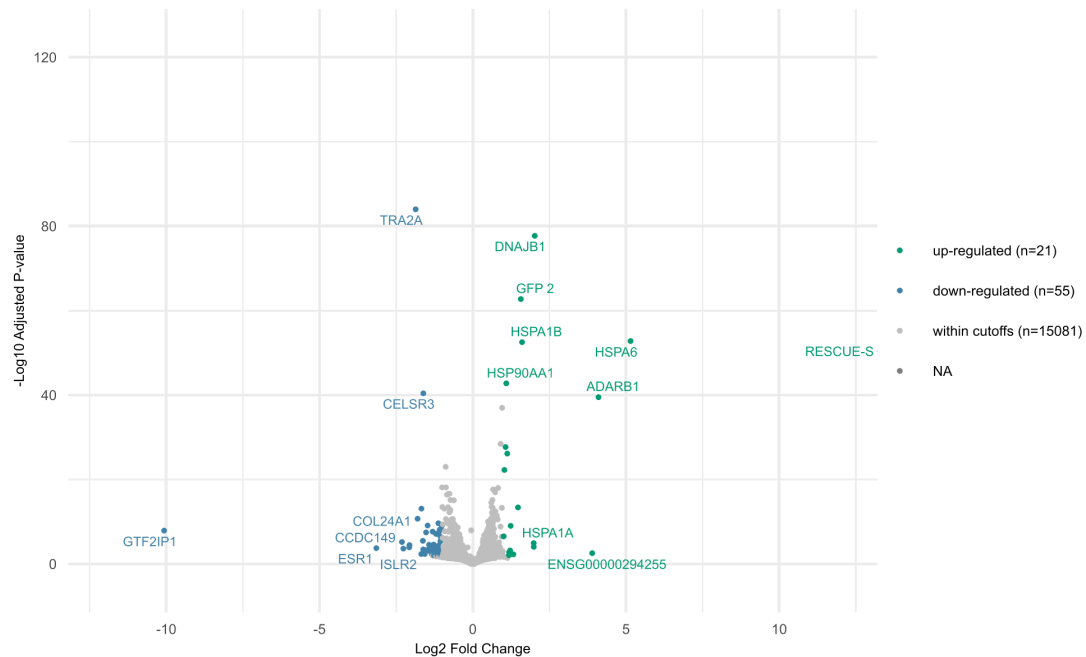
Supplementary Figure 16 | Volcano Plot CURE-C2

Volcano plot of transcript expression changes between HEK293T cells transfected with a GFP editing reporter + CURE-C2 and the reporter alone. Read data from BioProject PRJNA635732.



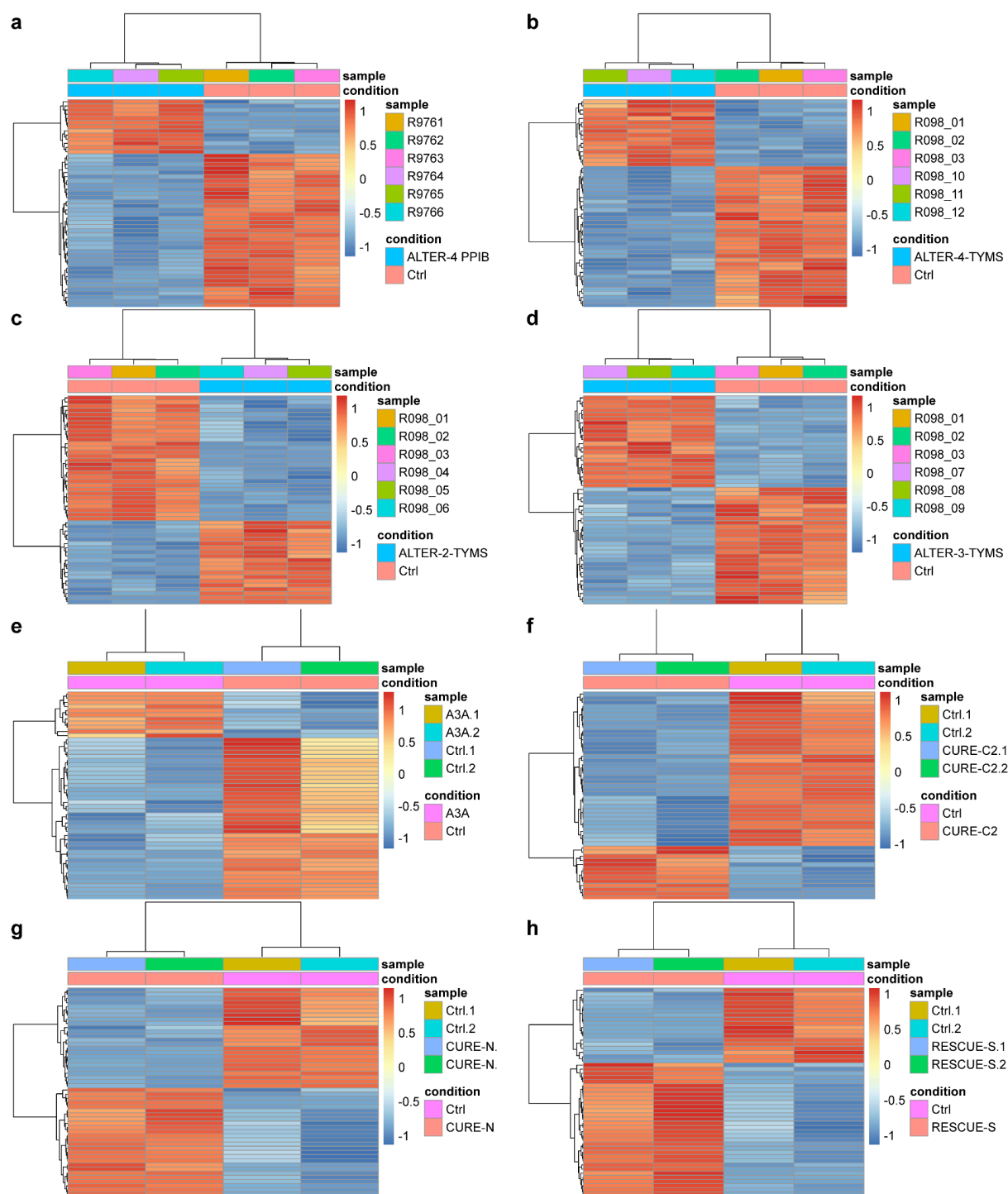
Supplementary Figure 17 | Volcano Plot CURE-N.

Volcano plot of transcript expression changes between HEK293T cells transfected with a GFP editing reporter + CURE-N and the reporter alone. Read data from BioProject PRJNA635732.



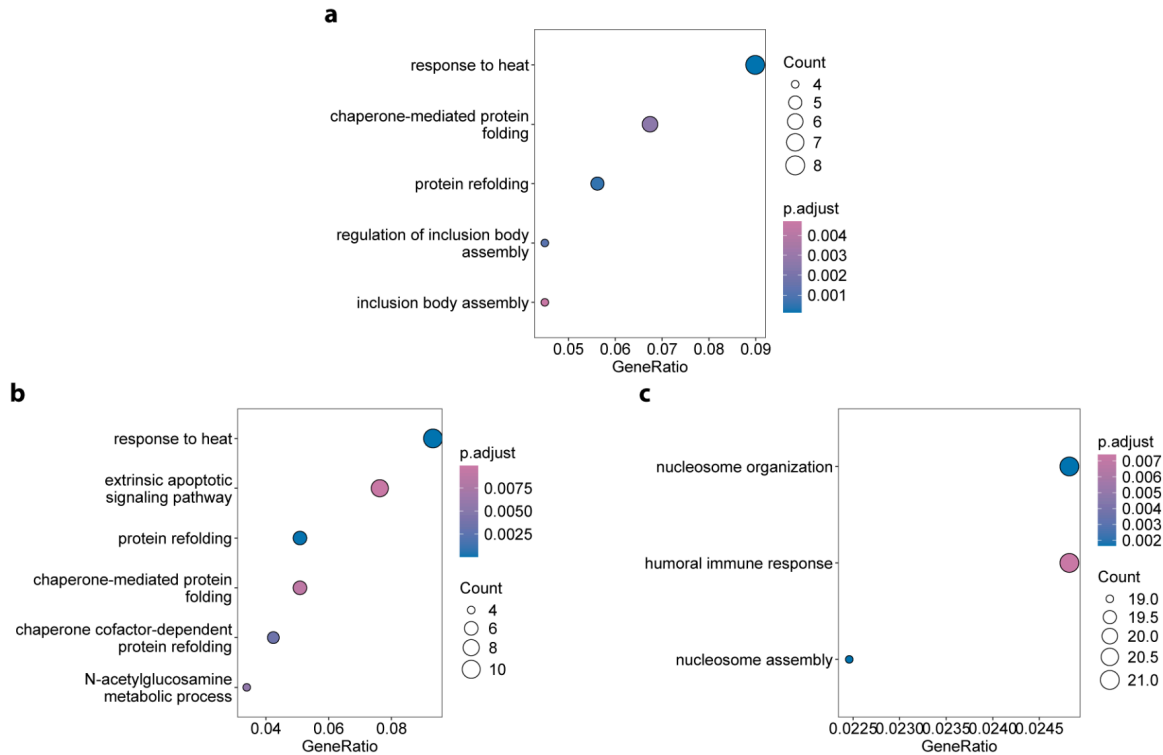
Supplementary Figure 18 | Volcano Plot RESCUE-S.

Volcano plot of transcript expression changes between HEK293T cells transfected with a GFP editing reporter + CURE-N and the reporter alone. Read data from BioProject PRJNA635732.



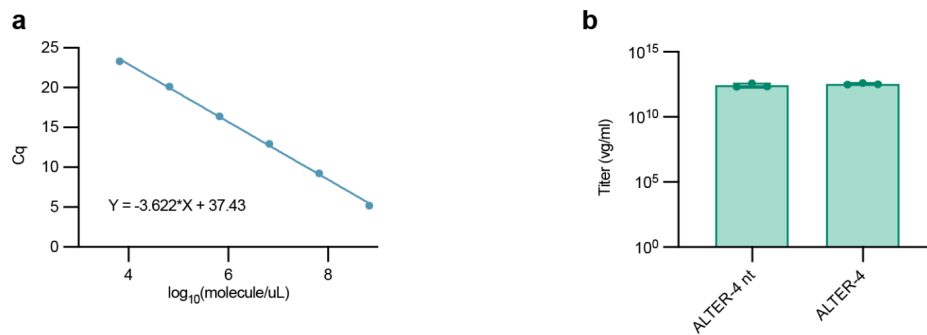
Supplementary Figure 19 | RNA Expression Heat Maps.

Clustered heat maps of RNA expression z-scores for genes with the 50 lowest padj values for the expression comparison between a given sample and the relevant control. **a**, ALTER-4 PPIB. **b**, ALTER-4 TYMS. **c**, ALTER-2 TYMS, **d**, ALTER-3 TYMS **e**, A3A overexpression. **f**, CURE-C2. **g**, CURE-N. **h**, RESCUE-S. Read data for samples other than ALTER from BioProject PRJNA635732.



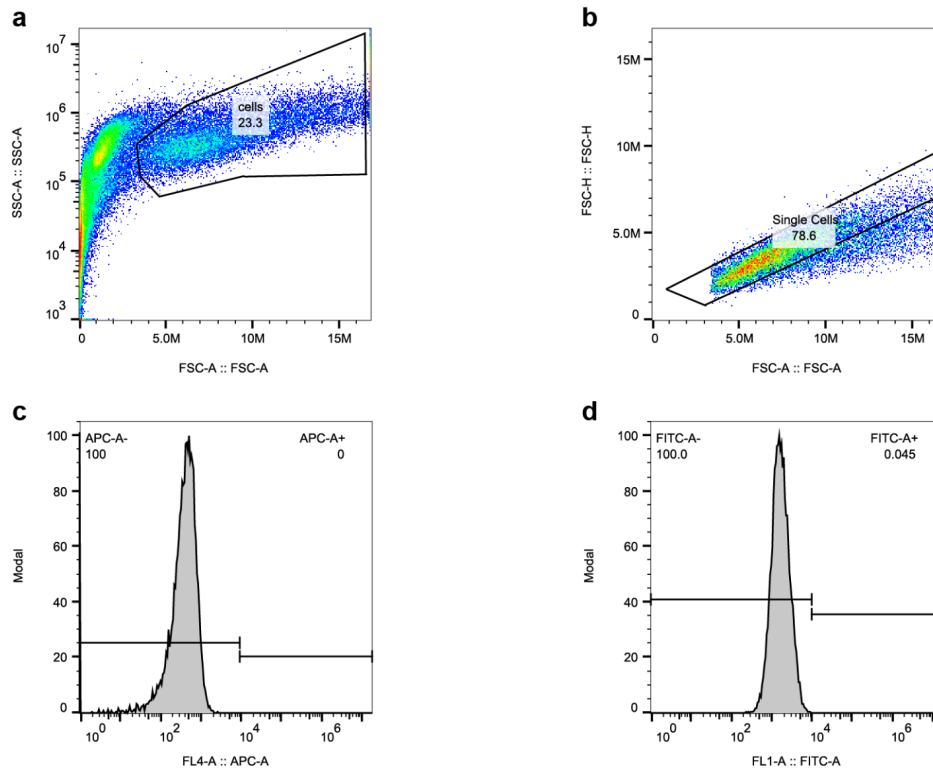
Supplementary Figure 20 | RNA Expression Heat Maps.

Biological GO term enrichment analysis results for **a**, ALTER-4 PPIB upregulated DEGs. **b**, ALTER-4 TYMS upregulated DEGs. **c**, ALTER-4 TYMS downregulated DEGs. ALTER-4 PPIB downregulated DEGs gave no hits.



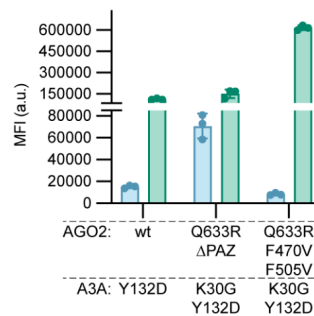
Supplementary Figure 21 | AAV Titers.

a, Standard 10-fold dilution curve. Data points are the mean of n=3 technical replicates. **b**, Titers of AAV samples used, calculated based on standard curve. Data are presented as the mean ± s.d. of n=3 biological replicates. Data points indicate replicates; bar heights indicate means.



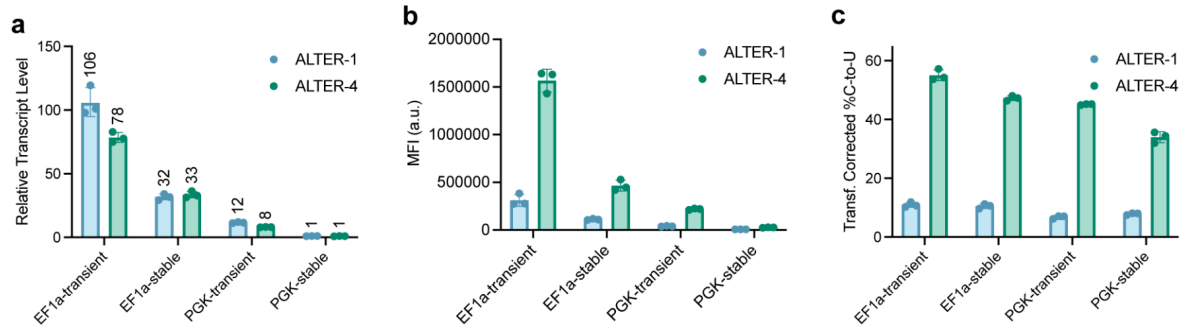
Supplementary Figure 22 | Representative Flow Cytometry Gating.

Flow cytometry gating strategy represented by data from one biological replicate of the untransfected control from the linker testing experiment shown in Fig. 5d. **a**, gating for cells based on FSC-A and SSC-A. **b**, exclusion of doublets based on FSC-A and FSC-H. **c**, iRFP +/- gate. **d**, GFP +/- gate.



Supplementary Figure 23 | Combination of A3A^{K30G,Y132D} with AGO2^{Q633RΔPAZ}

Combining A3A^{K30G,Y132D} with AGO2^{ΔPAZ,Q633R} led to no increase in editing activity and a large increase in background. Data are presented as the mean $\pm$ s.d. of n=3 biological replicates. Data points indicate replicates; bar heights indicate means.



Supplementary Figure 24 | Several Conditions Tested for EGFPY66H Editing Quantification Were Effective

To ascertain what conditions were feasible to quantify EGFP^{Y66H} editing, the following were tested with ALTER-1 and ALTER-4: (1) 7:1 cotransfection of editor and Ef1α-EGFP^{Y66H}. (2) Stably incorporated Ef1α-EGFP^{Y66H}. (3 and 4) The same for PGK-EGFP^{Y66H}. **a**, Resulting reporter transcript levels. **b**, Resulting Fluorescent signal. **c**, %C-to-U corrected for transfection rate in the stably incorporated conditions. Data are presented as the mean ± s.d. of n=3 biological replicates. Data points indicate replicates; bar heights indicate means.

Supplementary Table 1 | GSEA Hallmark Set Analysis ALTER-4 PPIB

Ten most significant results

pathway	pval	padj	log2err	ES	NES	size
HALLMARK_HYPOXIA	0.007	0.117	0.407	0.395	1.7	168
HALLMARK_MYC_TARGETS_V2	0.006	0.117	0.407	-0.500	-1.8	58
HALLMARK_MYOGENESIS	0.005	0.117	0.407	0.411	1.7	131
HALLMARK_KRAS_SIGNALING_UP	0.018	0.227	0.352	-0.357	-1.5	133
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	0.026	0.260	0.352	-0.362	-1.5	111
HALLMARK_ANGIOGENESIS	0.035	0.291	0.322	0.527	1.6	23
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	0.063	0.351	0.257	0.307	1.3	153
HALLMARK_HEME_METABOLISM	0.052	0.351	0.277	-0.307	-1.3	170
HALLMARK_GLYCOLYSIS	0.059	0.351	0.322	0.315	1.4	180
HALLMARK_MYC_TARGETS_V1	0.083	0.413	0.219	-0.284	-1.2	200

Supplementary Table 2 | GSEA Hallmark Set Analysis ALTER-4 TYMS

Ten most significant results

pathway	pval	padj	log2err	ES	NES	size
HALLMARK_TNFA_SIGNALING_VIA_NFKB	0.003	0.136	0.432	0.391	1.73	172
HALLMARK_UV_RESPONSE_UP	0.006	0.146	0.407	0.370	1.59	150
HALLMARK_MYOGENESIS	0.010	0.159	0.381	0.366	1.57	144
HALLMARK_E2F_TARGETS	0.014	0.174	0.381	-0.371	-1.58	200
HALLMARK_KRAS_SIGNALING_DN	0.020	0.201	0.352	0.353	1.47	119
HALLMARK_WNT_BETA_CATENIN_SIGNALING	0.024	0.203	0.352	0.458	1.57	39
HALLMARK_ANDROGEN_RESPONSE	0.051	0.366	0.257	-0.353	-1.35	94
HALLMARK_G2M_CHECKPOINT	0.065	0.409	0.219	-0.297	-1.27	199
HALLMARK_APICAL_SURFACE	0.082	0.412	0.234	0.415	1.38	34
HALLMARK_MYC_TARGETS_V2	0.082	0.412	0.207	-0.379	-1.34	58

Supplementary Table 3 | GSEA Hallmark Set Analysis ALTER-2 TYMS

Ten most significant results

pathway	pval	padj	log2err	ES	NES	size
HALLMARK_TNFA_SIGNALING_VIA_NFKB	2.051E-05	0.001	0.576	0.480	2.08	169
HALLMARK_E2F_TARGETS	0.027	0.236	0.352	-0.323	-1.39	200
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	0.023	0.236	0.352	0.324	1.40	163
HALLMARK_IL6_JAK_STAT3_SIGNALING	0.013	0.236	0.381	0.448	1.64	58
HALLMARK_INFLAMMATORY_RESPONSE	0.035	0.236	0.322	0.351	1.45	125

HALLMARK_KRAS_SIGNALING_DN	0.031	0.236	0.352	0.368	1.51	115
HALLMARK_MYOGENESIS	0.032	0.236	0.322	0.355	1.50	138
HALLMARK_UV_RESPONSE_UP	0.038	0.236	0.322	0.331	1.41	149
HALLMARK_HYPOXIA	0.046	0.253	0.322	0.310	1.34	172
HALLMARK_APICAL_SURFACE	0.074	0.370	0.241	0.442	1.41	33

Supplementary Table 4 | GSEA Hallmark Set Analysis ALTER-3 TYMS

Ten most significant results

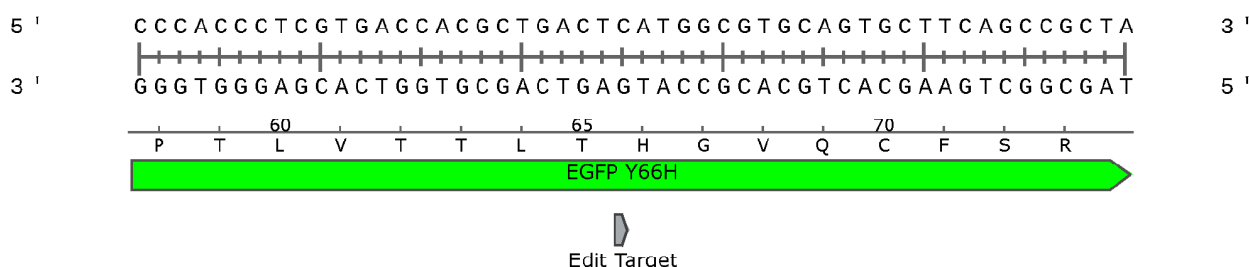
pathway	pval	padj	log2err	ES	NES	size
HALLMARK_MYC_TARGETS_V1	1.05E-05	0.0003	0.593	-0.460	-1.98	200
HALLMARK_MYOGENESIS	7.31E-05	0.0012	0.538	0.469	1.93	146
HALLMARK_INFLAMMATORY_RESPONSE	0.000	0.0040	0.498	0.449	1.81	130
HALLMARK_INTERFERON_GAMMA_RESPONSE	0.001	0.0053	0.477	0.421	1.75	157
HALLMARK_INTERFERON_ALPHA_RESPONSE	0.001	0.0069	0.477	0.494	1.86	83
HALLMARK_E2F_TARGETS	0.003	0.0160	0.432	-0.368	-1.58	200
HALLMARK_MYC_TARGETS_V2	0.003	0.0160	0.432	-0.481	-1.70	58
HALLMARK_UV_RESPONSE_UP	0.005	0.0262	0.407	0.394	1.63	152
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	0.008	0.0383	0.381	0.360	1.50	165
HALLMARK_HYPOXIA	0.008	0.0383	0.381	0.361	1.52	177

Supplementary Discussion 1 | ALTER Guide Design

Based on this study, it appears that the most optimal guide is likely to be a function of the target and specific editor used. When editing a new target we recommend trying 14L8, 12L6, 18L8 and 14L8(12-13mm). We find 14L8 to be generally effective across editors and targets tested in this study but there may be some confirmation bias. A step-by-step guide on how to design a 14L8 guide is presented here:

1. Identify target base to be edited

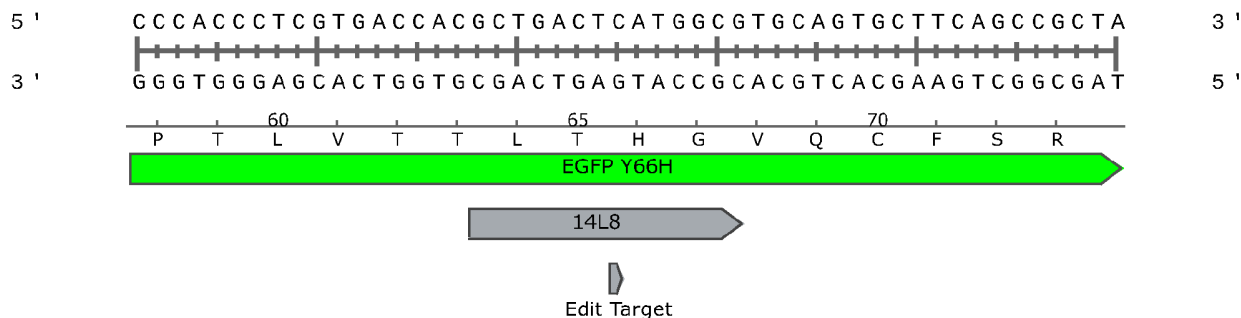
a. It will be most helpful to work on a dna sequence in a plasmid editor



2. Mark the loop to be induced

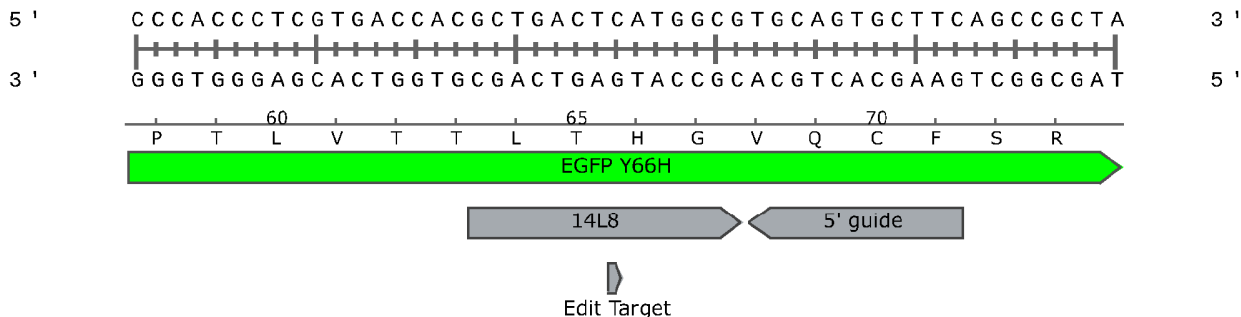
a. For example for 14L8 the loop should be 14 bases long with the target at position 8

i. Starting from position one, counting 5' > 3' in the sense strand



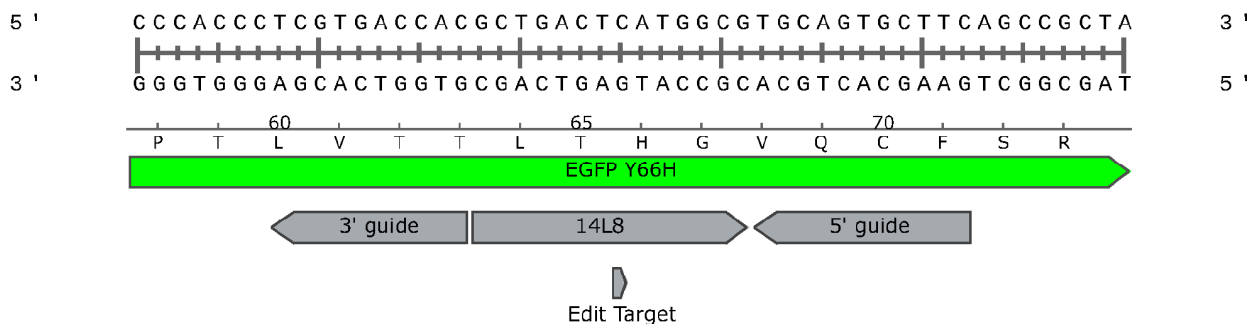
3. For the first section of the guide select 11 bases 3' of the loop in the sense strand

a. The 5' > 3' sequence of the antisense strand is the first 11 bases of your guide



4. Select 10 bases 5' of the loop in the sense strand

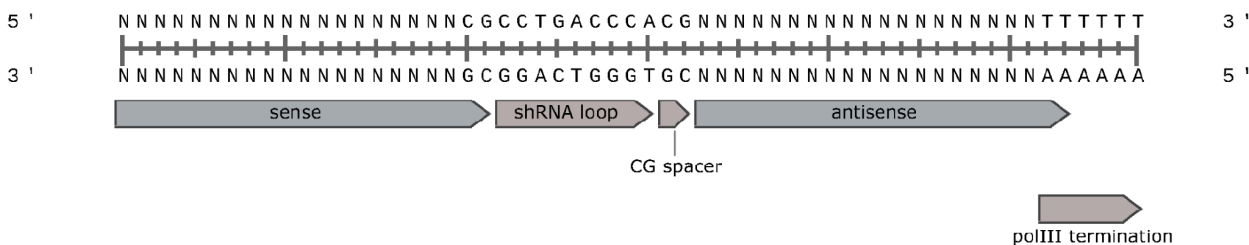
a. The 5' > 3' sequence of the antisense strand is the last 10 bases of your guide



5. Create and shRNA cassette template that looks like this

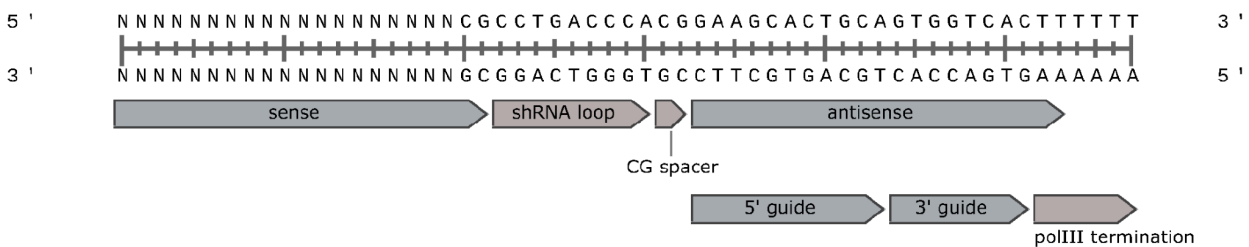
a. This will be for expression by the U6 promoter

b. The length of the cassette should always remain 57 bp



6. Use the 5' guide sequence and the first 8 bases of the 3' guide sequence that were marked on the target to fill in the variable portion of the antisense shRNA arm

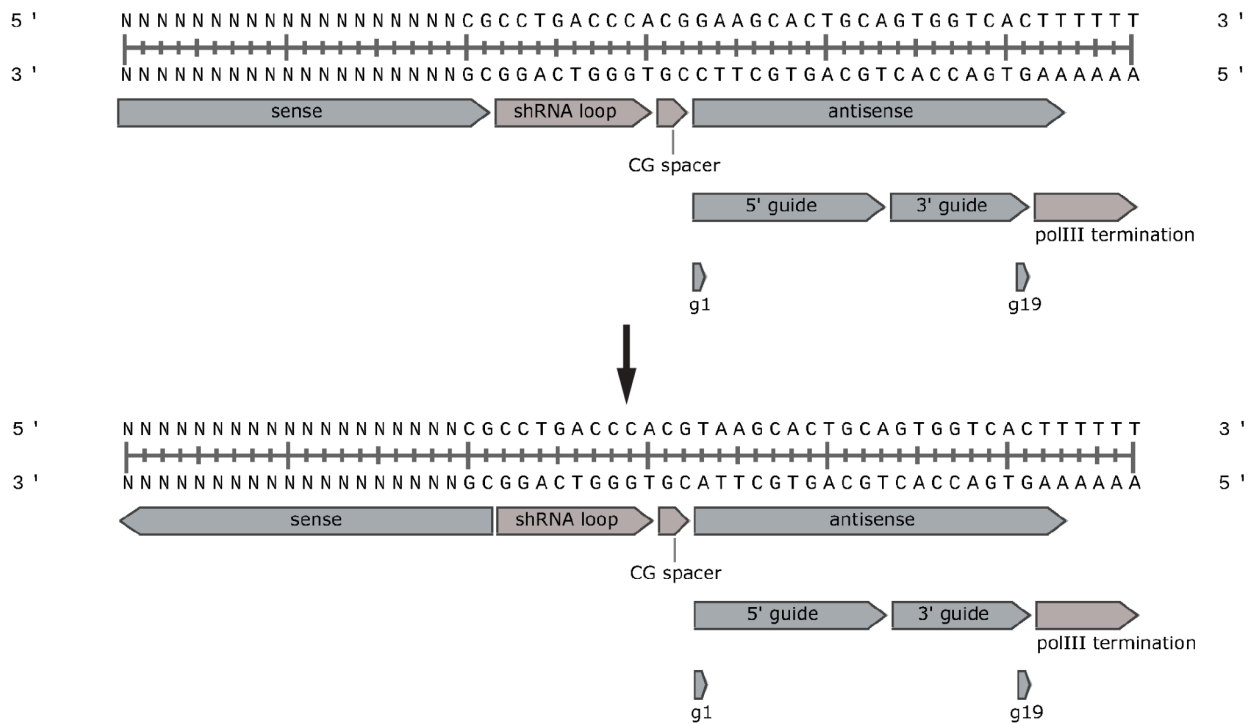
a. The sequences should come from the antisense strand of the target (5'>3') and be inserted into the sense (top) strand of the shRNA cassette



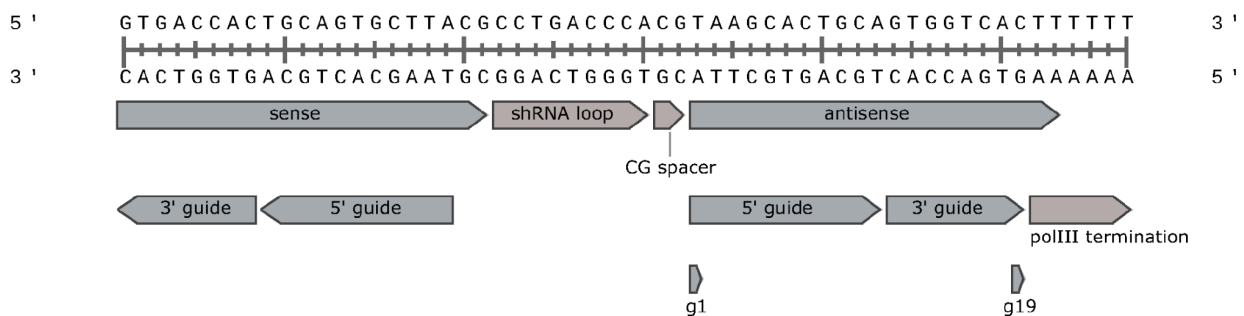
7. If necessary make the following substitution:

a. g1 (antisense arm base 1) should be a T

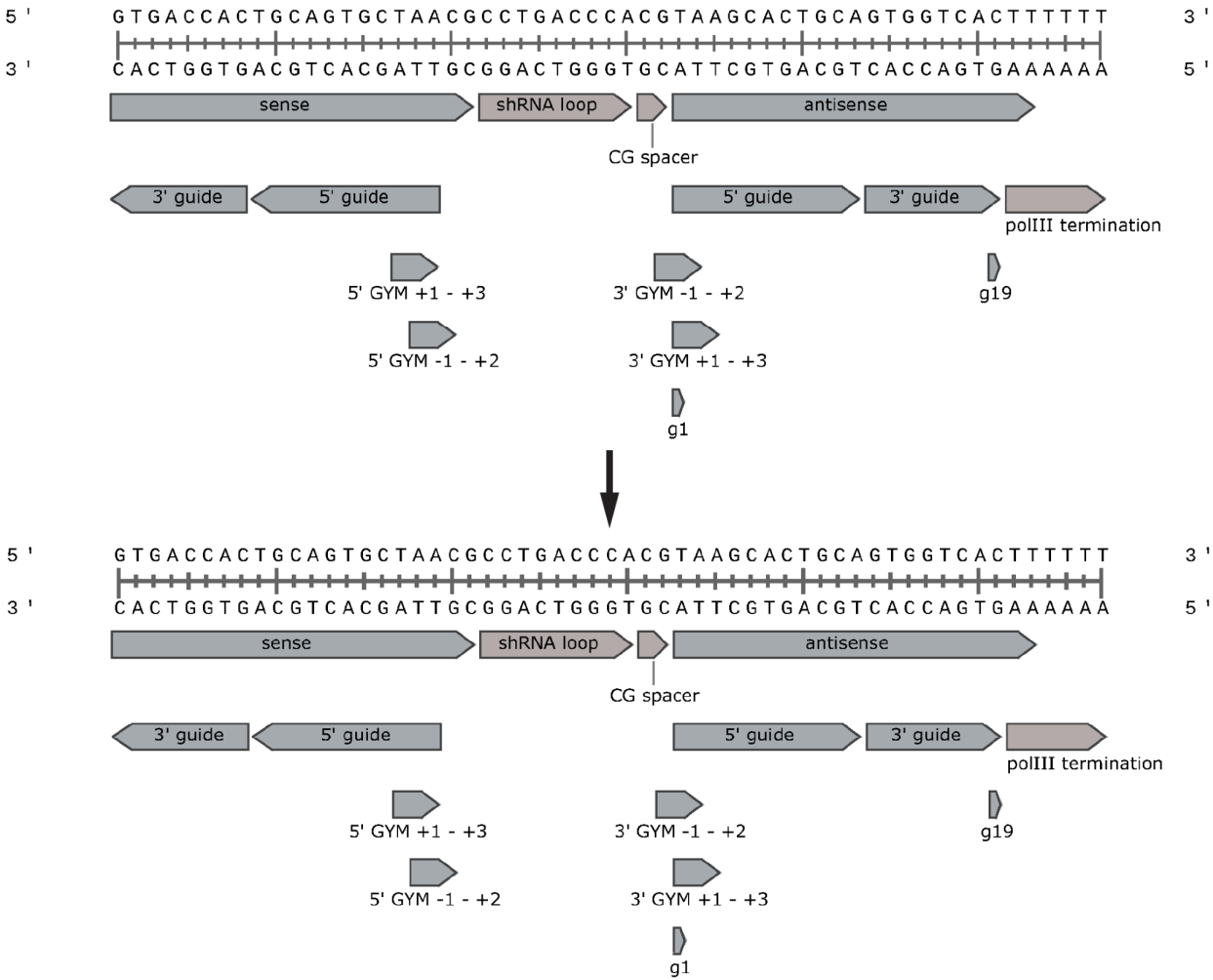
i. This base does not pair to the target and U is optimal for AGO2 loading³⁵



8. Take the first 19 bases of the antisense arm and insert their reverse complement into the variable region of the sense arm



9. If necessary, make SNPs in the sense and antisense strand to optimize the GYM scores⁹⁹



Supplementary Discussion 2 | Editing Quantification

Software tools

Name	Version
cutadapt	5.0
CRISPResso	2.3.3

1. Read QC with cutadapt

Reads sorted to contain a single barcode were processed to remove adaptors and filter for base quality with cutadapt. Cutadapt was run with the following arguments

Shell

```
cutadapt \  
  -g 'TCGTCGGCAGCGTC...CCGAGCCCACGAGAC' \  
  -g 'GTCTCGTGGGCTCGG...GACGCTGCCGACGA' \  
  --info-file '<path to info file output>' \  
  -q '30,30' \  
  -o '<path to output fastq>' \  
  --untrimmed-output '<path to output for untrimmed reads>' \  
  --minimum-length 100 \  
  
cutadapt \  
  -g 'TCGTCGGCAGCGTC' \  
  -g 'GTCTCGTGGGCTCGG' \  
  -g 'CCGAGCCCACGAGAC' \  
  -g 'GACGCTGCCGACGA' \  
  --info-file '<path to info file output>' \  
  -q '30,30' \  
  -o '<path to output fastq>' \  
  --untrimmed-output '<path to output for untrimmed reads>' \  
  --minimum-length 100 \  

```

2. Editing Quantification with CRISPResso2

Editing in the sequenced amplicon was assessed with CRISPResso2:

Shell

```
CRISPRessoBatch \  
  --batch_settings '<path to CRISPResso2 batch file>' \  
  -bo '<path to the output directory>' \  
  --amplicon_seq '<amplicon reference sequence>' \  
  --base_editor_output \  
  -g '<guide sequence for CRISPResso>' \  
  -wc -15 \  
  -w 30 \  
  --ignore_insertions \  
  --ignore_deletions \  
  --bam_output \  
  --exclude_bp_from_left 0 \  
  --exclude_bp_from_right 0 \  
  --conversion_nuc_from 'C' \  
  --conversion_nuc_to 'T' \  
  --default_min_aln_score 50 \  
  --min_single_bp_quality 0 \  
  --min_average_read_quality 0 \  
  --aln_seed_len 10 \  
  --amplicon_name '<amplicon name>' \  
  --guide_name '<guide name>'
```

Supplementary Discussion 3 | RNAseq Data Processing

Software Tools Used

NOTE: All RNAseq data analyzed for this paper were paired end reads and the software tools are applied accordingly. Due to differences in sequencing methodology, data for ALTER and data from BioProject PRJNA635732 were analyzed in separate groups. Code and files at <https://github.com/rflusi/ALTER-code>

Name	Version
samtools	1.16.1
STAR	2.7.10b
GATK	v4.6.0.0
bcftools	1.16
salmon	1.10.0
fastp	1.0.1
FastQC	v0.12.1
multiqc	1.29

Python Libraries

Scripts and jupyter notebooks were run with python 3.12.10 and used the following libraries:

Library	Version
tqdm	4.67.1
biopython	1.85
pysam	0.23.0
pandas	2.2.3
numpy	1.26.4

R Packages

Package	Version
DESeq2	1.46.0
tximport	1.34.0
dplyr	1.1.4
ggplot2	3.5.2
pheatmap	1.0.13
RColorBrewer	1.1.3
PoiClaClu	1.0.2.1
glmpca	0.2.0
apeglm	1.28.0
ggrepel	0.9.6
dplyr	1.1.4
tidyr	1.3.1
purrr	1.1.0
svglite	2.2.1
fgsea	1.32.2
data.table	1.17.8
msigdbR	25.1.1
org.Hs.eg.db	3.20.0
clusterProfiler	4.14.0
AnnotationDbi	1.68.0
enrichplot	1.26.1
biomaRt	2.62.0
ggrepel	0.9.6

Preparing Genomic References

All genomic references were based on GENCODE Release 48 which is based on GRCh38.p14. Specifically, the following files were used as a basis:

File Type	GENCODE Content	GENCODE Regions	Download Link
gtf	Comprehensive gene annotation	PRI	https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_48/gencode.v48.primary_assembly.annotation.gtf.gz
fasta	Genome sequence, primary assembly (GRCh38)	PRI	https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_48/GRCh38.primary_assembly.genome.fa.gz
fasta	Transcript sequences	ALL	https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_48/gencode.v48.transcripts.fa.gz

In addition, data from the following was also applied at different points in the pipeline

Database/Resource	File Type	Release	Download Link
dbSNP (NCBI)	vcf	138 (obtained from GATK resource bundle)	https://storage.cloud.google.com/genomics-public-data/resources/broad/hg38/v0/Homo_sapiens_assembly38.dbSNP138.vcf
ClinVar (NCBI)	vcf	2025-07-15	https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/weekly/clinvar_20250715.vcf.gz
Hallmark GSEA Gene Sets	gmt	v2025.1.Hs.symbols	https://www.gsea-msigdb.org/gsea/msigdb/download_file.jsp?filePath=/msigdb/release/2025.1.Hs/h.all.v2025.1.Hs.symbols.gmt
miRTarBase	csv	Homo sapiens	https://mirtarbase.cuhk.edu.cn/~miRTarBase/miRTarBase_2025/cache/download/10.0/hsa_MTI.csv

1. Adding Custom Contigs and Features to Reference Files

For each editor, and the CURE GFP reporter for data from BioProject PRJNA635732 sequence data for the relevant plasmids was added to the genomic reference fasta and corresponding features were added to the gtf. Data files for these additions can be found at <https://github.com/rflusi/ALTER-code>

Genome Reference	Added to Genome Fasta	Added to gtf	Added to Transcript Fasta
ALTER-transfection-control	pRFL462.fa	pRFL462.gtf	pRFL462-transcript.fa
ALTER-4	pRFL382.fa	pRFL382.gtf	pRFL382-transcript.fa
ALTER-master-ref	pRFL382-462.fa	pRFL382-462.gtf	pRFL382-462-transcript.fa
A3A-overexpression and CURE "wtHEK"	CURE-reporter.fa	CURE-reporter.gtf	CURE-reporter-transcripts.fa
CURE-C2	CURE-C2.fa	CURE-C2.gtf	CURE-C2-transcript.fa
CURE-N	CURE-N.fa	CURE-N.gtf	CURE-N-transcript.fa
RESCUE-S	RESCUE-S.fa	RESCUE-S.gtf	RESCUE-S-transcript.fa
CURE-master-ref	CURE-all.fa	CURE-all.gtf	CURE-all-transcript.fa

Additions were made as follows:

```
Shell
# --- User-defined variables ---
REF_FA="<genome fasta>"
REF_GTF="<reference gtf>"
ADD_FA="<fasta with sequences to be added to the reference genome>"
ADD_GTF="<gtf file with features to be added to the reference gtf>"

OUT_FA="<output fasta>"
OUT_GTF="<output gtf>"
SORTED_GTF="<final, sorted gtf>"

# --- Step 1:concat fastas ---
cat "$REF_FA" "$ADD_FA" > "$OUT_FA"
# print new fasta contigs to verify
samtools faidx "$OUT_FA"
cut -f1,2 "$TEMP_FAIDX"

# --- Step 2: concat gtfs ---
cat "$REF_GTF" "$ADD_GTF" > "$OUT_GTF"
```

```

# verify added records
tail "$OUT_GTF"

# --- Step 3: sort gtf ---
(grep '^#' "$OUT_GTF"; grep -v '^#' "$OUT_GTF" | sort -k1,1 -k4,4n) > "$SORTED_GTF"
# verify sorting
tail "$SORTED_GTF"

```

Corresponding transcript sequences were also added to the transcripts fasta file:

```

Shell
# --- User-defined variables ---
IN_FA="<transcripts fasta file>"
ADD_FA="<fasta file with sequences to be added>"
OUT_FA="<output fasta file>"
OUT_FASTAGZ="<compressed output fasta file>"
OUT_FAI="<compressed output fasta file index>"
REF_FA="hello"

# --- Step 1:concat fastas ---
cat "$IN_FA" "$ADD_FA" > "$OUT_FA"

# --- Step 2: Compress and Index fasta ---
bgzip \
  --threads 8 \
  -c "$OUT_FA" > "$OUT_FASTAGZ"
samtools faidx "$OUT_FASTAGZ"
# verify new contigs
cut -f1,2 "$OUT_FAI"

```

2. Building STAR Index

Alignment was carried out with STAR. This requires a STAR index which was built as follows. The remaining steps in the pipeline can use compressed fasta and gtf files as inputs, so they were compressed here:

```
Shell
# --- User-defined variables ---
IN_FA="<genome fasta>"
IN_GTF="<reference gtf>"

# set based on read length
OVERHANG=149

OUT_DICT="<STAR output dict>"
GENOME_DIR="<output directory for the STAR index>"
OUT_FASTAGZ="<output compressed genome fasta>"
OUT_GTFGZ="<output compressed gtf>"
OUT_FAI="<output fasta index>"

# --- Step 1: Build reference ---
STAR \
  --runThreadN 8 \
  --runMode genomeGenerate \
  --genomeDir "$GENOME_DIR" \
  --genomeFastaFiles "$IN_FA" \
  --sjdbGTFfile "$IN_GTF" \
  --sjdbOverhang $OVERHANG

# --- Step 2: Compress and Index fasta ---
bgzip \
  --threads 8 \
  -c "$IN_FA" > "$OUT_FASTAGZ"
samtools faidx "$OUT_FASTAGZ"

# --- Step 3: create dictionary of fa.gz ---
gatk CreateSequenceDictionary \
  -R "$OUT_FASTAGZ" \
  -O "$OUT_DICT"

# --- Step 4: Compress and Index gtf ---
bgzip \
  --threads 8 \
  -c "$IN_GTF" > "$OUT_GTFGZ"
tabix -p gff "$OUT_GTFGZ"
```

3. Making Genome Scatter Lists

Genome scatter lists were used to parallelize variant calling with GATK HaplotypeCaller (*vide infra*). They were created as follows:

Shell

```
# --- User-defined variables ---
GENOME_DICT="<STAR genome dict>"
INPUT_FAI="<genome fai>"

BED_FILE="<output bed file>"
GENOME_ILIST="<output interval_list file>"
SCATTER_DIR="<output directory for scatter lists>"

# --- Step 1: Create bed file ---
awk 'BEGIN{OFS="\t"} {print $1, 0, $2}' "$INPUT_FAI" > "$BED_FILE"

# --- Step 2: Create interval_list file ---
gatk BedToIntervalList \
  -I "$BED_FILE" \
  -O "$GENOME_ILIST" \
  -SD "$GENOME_DICT"

# --- Step 3: Create scatter lists ---
gatk IntervalListTools \
  -I "$GENOME_ILIST" \
  -O "$SCATTER_DIR" \
  --SCATTER_COUNT 4 \
  --SUBDIVISION_MODE BALANCING_WITHOUT_INTERVAL_SUBDIVISION_WITH_OVERFLOW \
  --UNIQUE true \
  --SORT true
```

4. Make Salmon Index

Transcript raw counts were quantified using Salmon. The required index was created as follows:

Shell

```
# --- User-defined variables ---
IN_FASTAGZ="<input transcript fasta.gz>"
OUT_INDEX="<output directory for salmon index>"
# adjust based on computing resources
THREADS=6

# --- Step 1: Make salmon index ---
salmon index --gencode -t "$IN_FASTAGZ" \
  -p $THREADS \
  -i "$OUT_INDEX"
```

RNAseq Data Quality Control

Sequencing adaptors and low-quality bases were trimmed from sequencing data, then reads were filtered for quality and minimum length, all with fastp. The resulting fastq files were assessed with fastqc and multiqc.

Shell

```
# --- User-defined variables ---
SAMPLES=(<list of samples corresponding to fastq file names>)
DATA_DIR="<directory containing fastq files>"
OUT_DIR="<output directory for QC fastq files>"
QC_OUT_DIR="$QC_DIR/3_final-qc"

# --- Step 1: Loop through each sample and run fastp ---
for SAMPLE in "${SAMPLES[@]"; do
    fastp \
        -i "${DATA_DIR}/${SAMPLE}_1.fq.gz" \
        -I "${DATA_DIR}/${SAMPLE}_2.fq.gz" \
        -o "${OUT_DIR}/${SAMPLE}_1.clean.fq.gz" \
        -O "${OUT_DIR}/${SAMPLE}_2.clean.fq.gz" \
        -h "${OUT_DIR}/${SAMPLE}.fastp.html" \
        -j "${OUT_DIR}/${SAMPLE}.fastp.json" \
        --detect_adapter_for_pe \
        -q 20 \
        --n_base_limit 5 \
        --length_required 35
done

# --- Step 2: generate QC reports ---
for SAMPLE in "${SAMPLES[@]"; do
    # --- Step a: Run fastqc for reads 1 ---
    fastqc "$OUT_DIR/${SAMPLE}_1.clean.fq.gz" \
        -o "$QC_OUT_DIR" \
        -t 4

    # --- Step b: Run fastqc for reads 2 ---
    fastqc "$OUT_DIR/${SAMPLE}_2.clean.fq.gz" \
        -o "$QC_OUT_DIR" \
        -t 4
done

multiqc "$QC_OUT_DIR" \
-o "$QC_OUT_DIR"
```

RNAseq Variant Calling

1. Read Alignment

Reads were aligned to the genomic reference using STAR in two-pass mode. For each sample:

Shell

```
# --- User-defined variables ---
GENOME_DIR="<directory containing STAR index>"
SAMPLE_NAME="<name of the sample being aligned>"
INPUT_1="<fastq reads 1>"
INPUT_2="<fastq reads 2>"
OUT_DIR="<directory for output files>"
OVERHANG=149
# adjust based on compute resources
BAM_RAM=60000000000
THREADS=8

# --- Step 1: STAR ---
STAR \
  --runThreadN $THREADS \
  --genomeDir "$GENOME_DIR" \
  --readFilesIn "$INPUT_1" "$INPUT_2" \
  --readFilesCommand zcat \
  --outFileNamePrefix "$OUT_DIR/" \
  --outSAMtype BAM SortedByCoordinate \
  --outSAMattrRGline ID:$SAMPLE_NAME SM:$SAMPLE_NAME PL:ILLUMINA \
  --quantMode TranscriptomeSAM GeneCounts \
  --twopassMode Basic \
  --sjdbOverhang $OVERHANG \
  --limitBAMsortRAM $BAM_RAM
```

2. Prepare Files for HaplotypCaller

To prepare data for the HaplotypCaller input, duplicate reads are marked with GATK MarkDuplicates, and reads spanning splice junctions were split with GATK SplitNCigarReads. The final bam was then reordered to match a genomic reference containing all plasmid sequences (the “master” reference) using GATK ReorderSam. This allows for joint genotyping of samples transfected with different plasmids while also allowing alignment to a reference with only the plasmid relevant to that sample. If a user wishes to avoid this they will need to use an unmodified reference or genotype each sample separately. Sample data was prepared as follows for each sample:

Shell

```
# --- User-defined variables ---
ALIGN_BAM="<bam file output from the STAR alignment>"
MARKED_BAM="<output path for bam file with duplicate reads marked>"
DUP_METRICS_OUT="<output path for a txt file with metrics on duplicate reads>"
SPLIT_BAM="<output path for bam file with split reads>"
REORDERED_BAM="<output path for bam file reordered to match the master reference>"

GENOME_DIR="<genome directory containing the STAR index for this sample>"
REF="<sample genome fasta>"
GTF="<sample genome gtf>"

MASTER_GENOME_DIR="<directory containing the STAR index for the master reference>"
MASTER_GENOME_DICT="<STAR dict file for the master reference>"
MASTER_REF="<master reference genome fasta>"
MASTER_GTF="<master reference gtf>"

# --- Step 1: MarkDuplicates ---
gatk --java-options "-Xmx6g" MarkDuplicates \
  -I "$ALIGN_BAM" \
  -O "$MARKED_BAM" \
  -M "$DUP_METRICS_OUT" \
  --VALIDATION_STRINGENCY SILENT \
  --CREATE_INDEX true

# --- Step 2: SplitNCigarReads ---
gatk --java-options "-Xmx6g" SplitNCigarReads \
  -R "$REF" \
  --create-output-bam-index true \
  -I "$MARKED_BAM" \
  -O "$SPLIT_BAM"

# --- Step 3: ReorderSam to Standardize Headers ---
gatk --java-options "-Xmx6g" ReorderSam \
  -I "$SPLIT_BAM" \
  -O "$REORDERED_BAM" \
  -SD "$MASTER_GENOME_DICT" \
  -R "$MASTER_REF" \
  --CREATE_INDEX true
```

3. Haplotypewriter Parallelized Over Scatter Lists

Variants are called for each sample using GATK Haplotypewriter. For each sample Haplotypewriter is called over four different genomic intervals concurrently, corresponding to the scatter lists made with the STAR index. For each sample-scatter list combination this is performed as follows:

Shell

```
# --- User-defined variables ---
INPUT_BAM="<sample reordered bam file from step 2>"
SCATTER_LIST="<scatter list interval_list file>"
OUT_GVCF="<path for output gvcf file>"

REF_FA="<master genome fasta>"
DBSNP_VCF="<dbsnp vcf file>"

# --- Step 1: Haplotypewriter ---
gatk --java-options -Xmx6g Haplotypewriter \
  -R "$REF_FA" \
  -I "$INPUT_BAM" \
  -O "$OUT_GVCF" \
  -L "$SCATTER_LIST" \
  --dbsnp "$DBSNP_VCF" \
  -ERC GVCF \
  --dont-use-soft-clipped-bases \
  --standard-min-confidence-threshold-for-calling 20 \
  --native-pair-hmm-threads 4
```

4. Joint Genotyping

Next, gvcf files for each sample are merged then the combined gvcf files are combined into a single gvcf containing variant data for all samples. This combined gvcf is used to make genotyping calls jointly across all samples. The final variants are passed through basic quality filters then output to a final tsv.

To merge scatter gvcfs the following is run for each sample:

Shell

```
# --- User-defined variables ---
SAMPLE_GVCFS=(
    "<scatter gvcf #1 from step 3>"
    "<scatter gvcf #2 from step 3>"
    "<scatter gvcf #3 from step 3>"
    "<scatter gvcf #4 from step 3>"
)
MERGED_GVCF="<path for the output gvcf>"

# --- Step 1: Dynamically create the --INPUT arguments for MergeVcfs ---
INPUTS_TO_MERGE=(
for GVCF in "${SAMPLE_GVCFS[@]}"; do
    INPUTS_TO_MERGE+=(--INPUT "$GVCF")
done

# --- Step 2: Run MergeVcfs for the Sample ---
gatk MergeVcfs \
    "${INPUTS_TO_MERGE[@]}" \
    --OUTPUT "$MERGED_GVCF"
```

Then the merged sample gvcfs are combined, genotyped, filtered, and converted to the final table as follows:

Shell

```
# --- User-defined variables ---
# this list should include the merged gvcfs from the merge step above for all samples 1-n
MERGED_GVCFS=(
    "<merged gvcf sample 1>"
    "<merged gvcf sample 2>"
    "<merged gvcf sample n>"
)
INPUT_ARGS_COMBINE=(
COMBINED_VCF="<path for the output combined gvcf>"
```

```

GT_VCF="<path for the output genotyped gvcf>"
FILT_VCF="<path for the output filtered gvcf>"
OUT_TSV="<path for the output variant table tsv>"

REF="<master genome reference fasta>"
DBSNP_VCF="<dbsnp vcf file>"

# --- Step 1: Combine merged gvcfs ---
# build input argument
for MERGED_GVCF in "${MERGED_GVCFS[@]}"; do
    INPUT_ARGS_COMBINE+=(-V "$MERGED_GVCF")
done

gatk CombineGVCFs \
    -R "$REF" \
    "${INPUT_ARGS_COMBINE[@]}" \
    -O "$COMBINED_VCF"

# --- Step 2: Genotyping ---
gatk --java-options "-Xmx12g" GenotypeGVCFs \
    -R "$REF" \
    -V "$COMBINED_VCF" \
    -D "$DBSNP_VCF" \
    -O "$GT_VCF"

# --- Step 3: Filter Variants ---
gatk --java-options "-Xmx6g" VariantFiltration \
    -R "$REF" \
    -V "$GT_VCF" \
    -O "$FILT_VCF" \
    --window 35 \
    --cluster 3 \
    --filter-expression "QD < 2.0" --filter-name "LowQD" \
    --filter-expression "FS > 30.0" --filter-name "StrandBias" \
    --filter-expression "DP < 10" --filter-name "LowDepth" \
    --filter-expression "QUAL < 20.0" --filter-name "LowQual"

# --- Step 5: Convert to Table ---
gatk --java-options "-Xmx6g" VariantsToTable \
    -V "$FILT_VCF" \
    -F CHROM -F POS -F ID -F REF -F ALT -F QUAL -F FILTER \

```

```
-GF GT -GF AD -GF DP -GF GQ \  
-O "$OUT_TSV"
```

5. Variant Table Annotation and Final Filtering for C-to-U Off-Target Hits

Details and scripts used in the final analysis are available at <https://github.com/rflusi/ALTER-code>

RNAseq Expression Quantification and DEG Analysis

1. Quantifying Transcript-Level Raw Counts with Salmon

Raw counts were quantified directly from the post-qc fastq files using salmon as follows:

Shell

```
# --- User-defined variables ---  
INPUT_ONE="<reads fastq 1>"  
INPUT_TWO="<reads fastq 2>"  
  
SALMON_IDX="<salmon index created when building the genome reference>"  
  
# --- Step 1: Run Salmon ---  
salmon quant \  
  -i "$SALMON_IDX" \  
  --libType A \  
  --validateMappings \  
  --gcBias \  
  --numGibbsSamples 20 \  
  --threads "$THREADS" \  
  --output "$OUT_DIR" \  
  -1 "$INPUT_ONE" \  
  -2 "$INPUT_TWO"
```

2. DEG Analysis, GSEA Analysis, and DEG GO Term Enrichment Analysis

These analyses were performed with the R scripts available at <https://github.com/rflusi/ALTER-code>

Supplementary References

99. Lee, Y.-Y., Kim, H. & Kim, V. N. Sequence determinant of small RNA production by DICER.
Nature **615**, 323–330 (2023).