The American Journal of Human Genetics, Volume *101*

Supplemental Data

CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for *HPRT1* Expression via Thousands of Large, Programmed Genomic Deletions

Molly Gasperini, Gregory M. Findlay, Aaron McKenna, Jennifer H. Milbank, Choli Lee, Melissa D. Zhang, Darren A. Cusanovich, and Jay Shendure

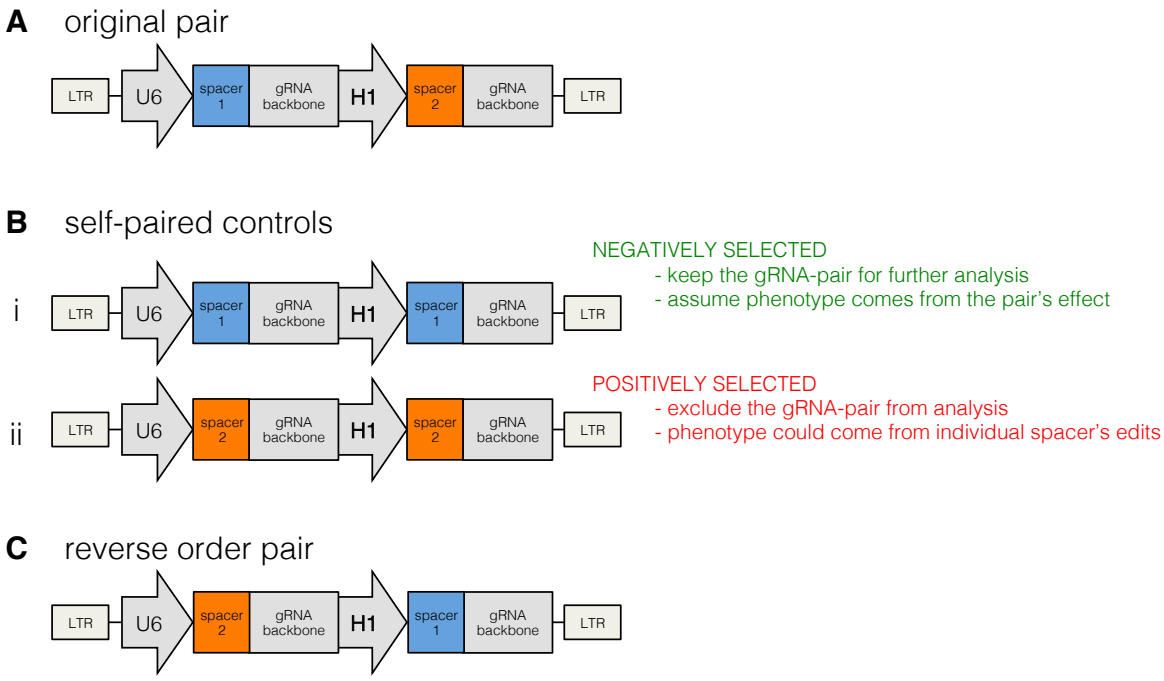


Figure S1. Self-paired spacers in the ScanDel library reveal phenotypes independently created by individual spacers.

A) The spacers used in every designed gRNA pair had their own self-paired control included in the programmed gRNA pair library.

B) The self-paired controls consisted of the exact same spacer included behind each promoter in the expression construct (two for each pair; (*i*) and (*ii*)). If a self-paired spacer was positively selected, any gRNA pairs that included that spacer were excluded from further analysis. This avoided any confounding effects of alternative repair outcomes that result from an individual gRNA's edit that could cause 6TG resistance (*e.g.* a ~10 bp indel disrupting a transcription factor binding site, or disrupting an off-target locus that affects 6TG resistance, or an individual gRNA inducing translocations of *HPRT1* at a high rate). By excluding these gRNAs, we can more confidently attribute observed phenotypes to programmed deletion induced by the gRNA pairs.

C) Each gRNA pair was included in both possible orderings on the microarray. This was intended to minimize the impact of differences between the promoters, as well as to increase the chance that each deletion will be represented in the library, as synthesizing each pair twice reduces loss due to synthesis errors and cloning bottlenecks.

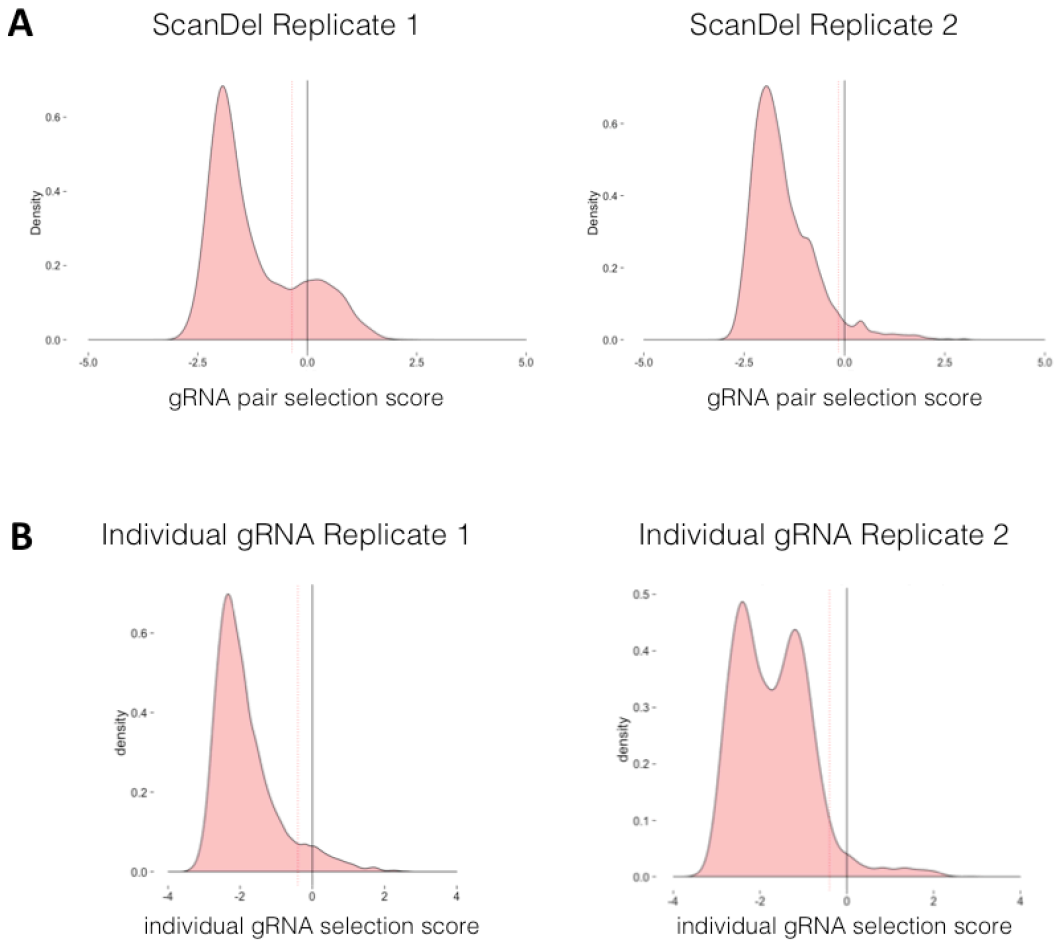


Figure S2: Distribution of selection scores across biological replicates for ScanDel gRNA pairs or individual gRNAs.

A) Each gRNA pair in the ScanDel screens was assigned a selection score (log10(after/before 6TG)). The minimum selection score threshold described in **Methods** (-0.35 for replicate 1, -0.15 for replicate 2) is drawn with a dotted red line.

B) Each gRNA in the individual gRNA screen was assigned a selection score as in **A**, for each replicate. The minimum negative selection score threshold (-0.4 for both replicates) is drawn with a dotted red line (explanation in **Methods**).

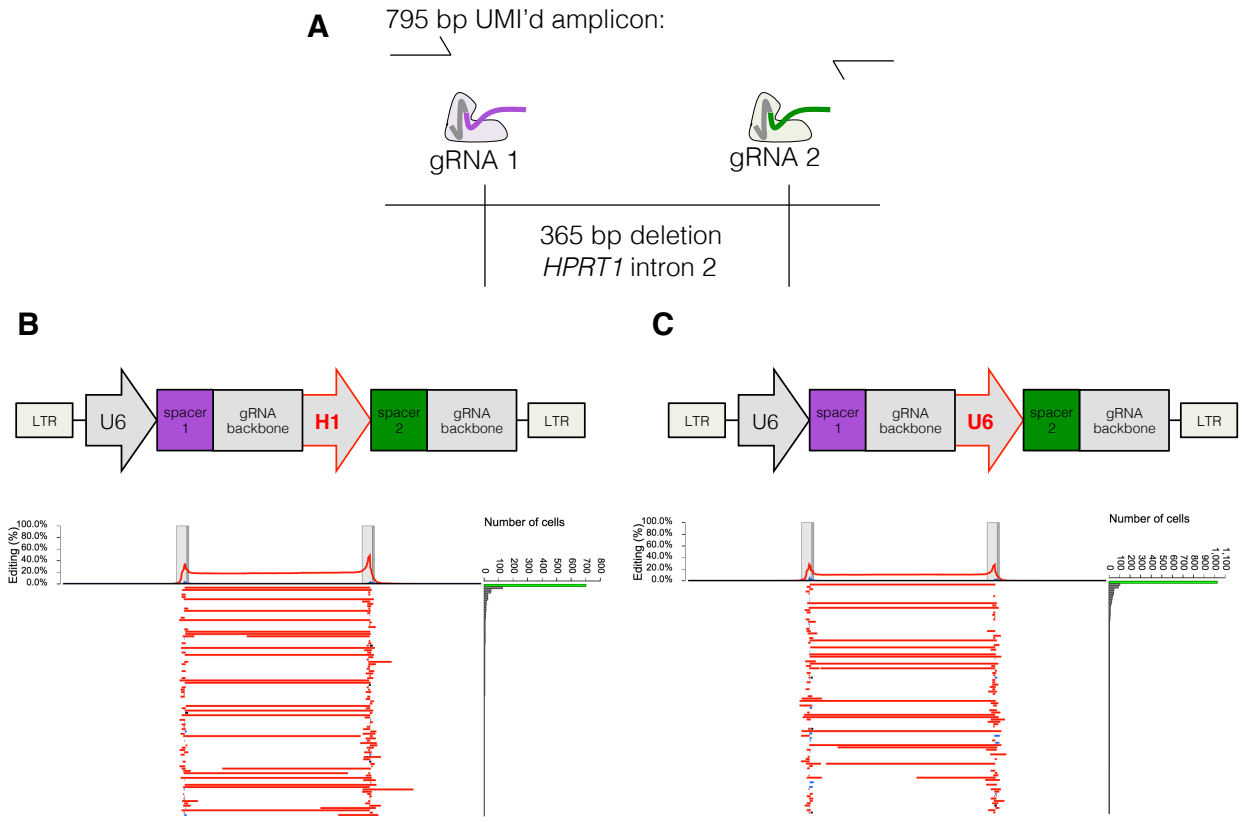


Figure S3: The U6-H1 gRNA pair expression construct induces a higher deletion rate.

A) Two spacers were chosen to program a 365 bp deletion within the second intron of *HPRT1*. To test deletion efficiency of the method as described in **Fig. 1C**, virus was made from the constructs depicted in **B** and **C**, and separately transduced into HAP1 at MOI < 0.3. Following 1 week of puromycin selection, gDNA was extracted and the targeted region amplified. The first 3 cycles of this PCR contained a forward primer with a unique molecular tag (UMI) to track reads from the same original cell. Sequencing was performed on a MiSeq. Of note, PCR bias for smaller deletion-holding amplicons was reduced by collapsing reads with the same UMI, but the potential remains for higher clustering efficiency of the shorter amplicons.

B) The spacers for the deletion in **A** were placed behind either a U6 or H1 PolIII promoter. 20% of sampled haplotypes contained the programmed deletion, but 36% of sampled haplotypes remained unedited, implying longer editing time could result in a higher deletion rate. Reads were generated as described in **A**, and aligned as described in **Methods** and **Fig. 3**. The per base-pair editing rate summed across all sampled haplotypes is charted as a percentage at top, and the top 100 most prevalent haplotypes are displayed below it. Red indicates deletions and blue insertions.

C) The spacers for the deletion in **A** were each placed behind a U6 PolIII promoter, and delivered, sampled, and visualized as above. With this expression construct, 10% of sampled haplotypes contained the programmed deletion.

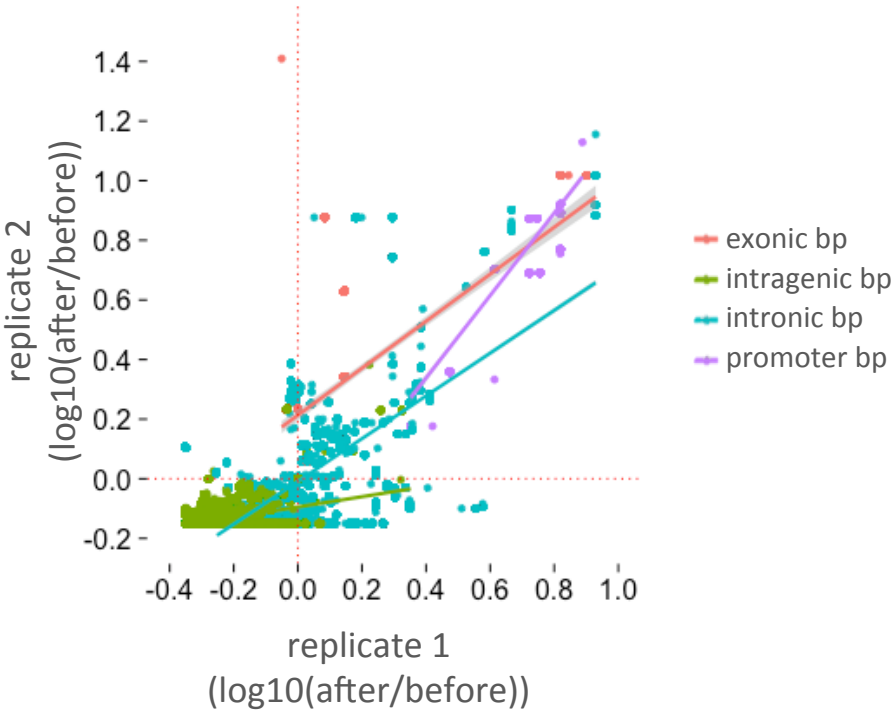
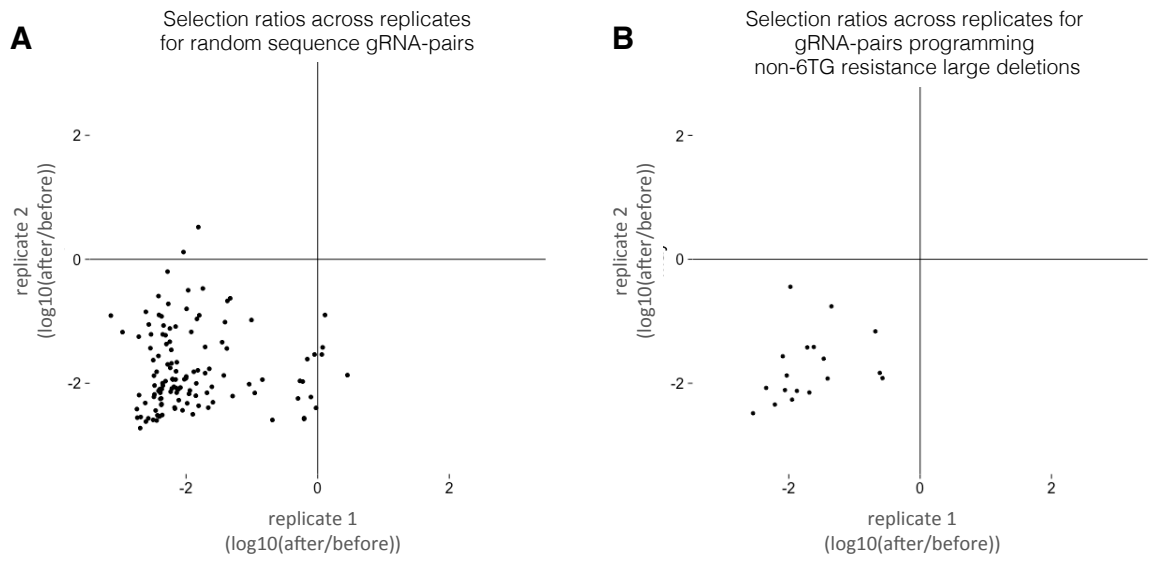


Figure S4: ScanDel scores correlate across two biological replicates.

The ScanDel selection scores for each biological replicate were calculated per base-pair by averaging the log10(after/before 6TG) for every programmed deletion that covers that base-pair. Least squares lines and points are colored by sequence content category. The stronger correlation for the 'intronic' category is driven by sequences proximal to the exons as seen in **Fig. 3**. Red corresponds to exons (Pearson: 0.736); green to intragenic regions (Pearson: 0.417); blue to intronic regions (within 2 Kb of an exon, Pearson: 0.628; deeply intronic, Pearson: -0.0194); and purple is the promoter (1 Kb upstream of the TSS, Pearson: 0.905).



C	positively selected in both	negatively selected in both	positively selected in rep 1	negatively selected in rep 1	positively selected in rep 2	negatively selected in rep 2
random sequence not found in hg19 (**A**)	0	236	8	256	4	244
deletions targeting non-6TG resistance regions (**B**)	0	18	1	31	0	18

Figure S5: None of the negative control gRNA pairs were positively selected by 6TG in both ScanDel replicates.

A) Negative control gRNA pairs targeting random sequences not found in hg19 were given a selection score of log10(after/before 6TG). Only gRNA pairs sampled in both replicates are plotted.

B) Additional negative control gRNA pairs were programmed to create 1 and 2 Kb deletions in regions not expected to cause 6TG resistance. Selection scores were calculated for each gRNA pair as in **A**, and plotted for gRNA pairs found in both replicates. These region's coordinates were randomly generated from poorly conserved sequence [1] not within 10 Kb of any gene and far from *HPRT1* (chr8:23768553-23771053, chr4:25697737-25700237, chr9:41022164-41024664, chr5:12539119-12541619, chr6:23837183-23839683, chr8:11072736-11075236).

C) Table showing counts of positively and negatively selected negative control gRNA pairs across experiments.

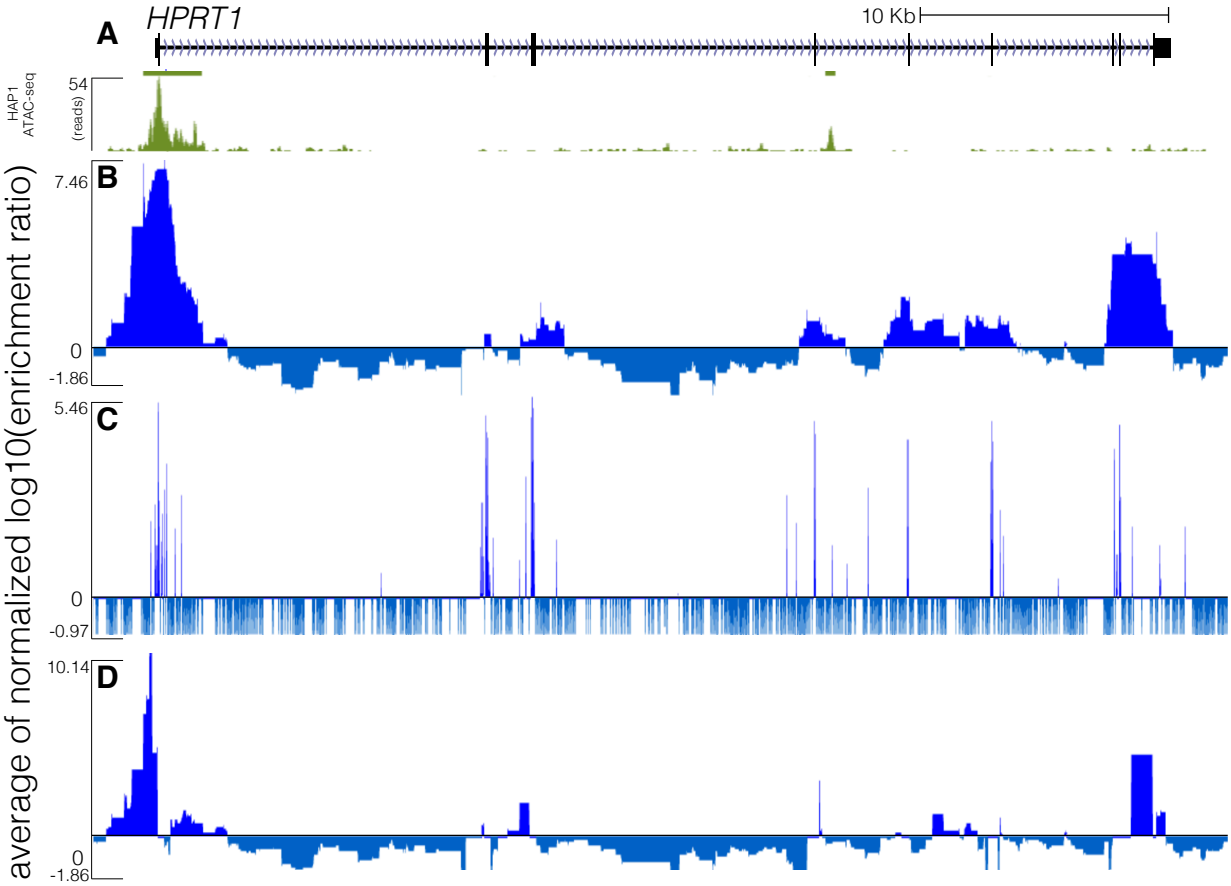


Figure S6: All exons and some exon-proximal non-coding regions score strongly in both the ScanDel gRNA pair screen and the individual gRNA screen.

A) ATAC-seq data (green) from the HAP1 cell line displayed for the *HPRT1* locus (chrX:133,591,675-133,637,198, hg19). Bars depict hotspots [2] and beneath is the pile-up representation of ATAC-seq reads.

B) The same ScanDel data is displayed as in **Fig. 2C** but zoomed-in on the *HPRT1* locus. Each base-pair's score is the mean of the log10(after/before 6TG) values for all the programmed deletions that cover that base-pair. These scores are normalized to the median positive score from the replicate. The average of the two replicates' scores for each base-pair is displayed.

C) The same individual gRNA data is displayed as in **Fig. 2D** but zoomed in on *HPRT1*. Each base-pair score is the mean of the log10(after/before 6TG) values for all the inferred ~10 bp deletions that remove that base-pair. The normalized average of the two replicates' scores for that base-pair is displayed.

D) The same ScanDel track as in **A** but with per base-pair scores calculated after excluding any deletions programmed to disrupt an exon.

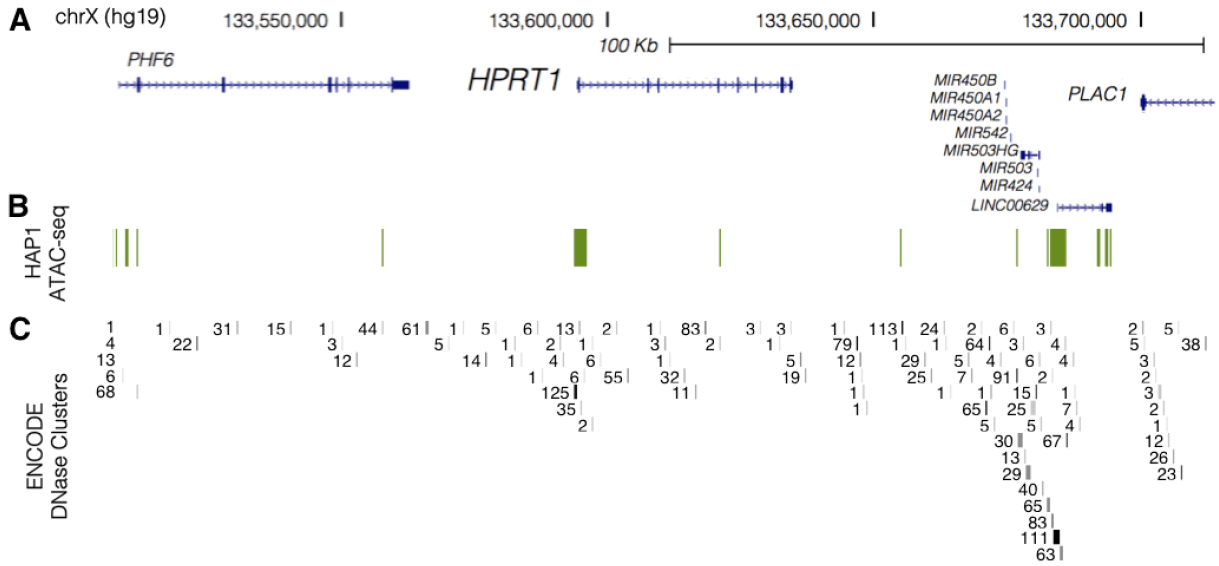
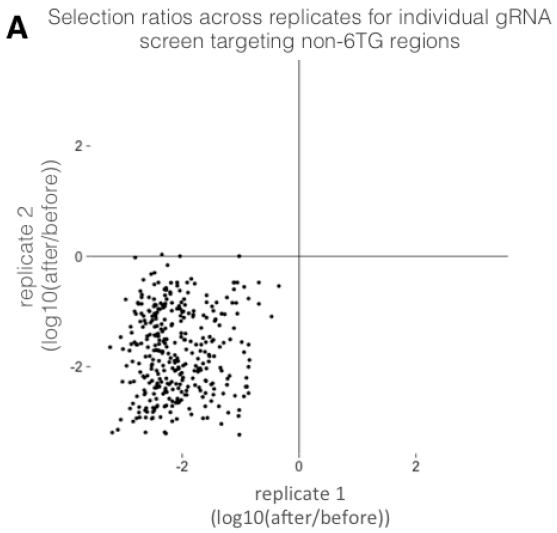


Figure S7: To confirm HAP1's suitability as a model in which to study the ubiquitously expressed *HPRT1*, regions of accessibility were compared across HAP1 and 125 ENCODE cell types.

A) The 206.1 Kb encompassing *HPRT1* and its surrounding sequence interrogated by this screen (chrX:133,507,694-133,713,798, hg19, UCSC Genes track in blue).

B) Regions of open chromatin in HAP1 cells (green) as profiled by ATAC-seq.

C) Clusters of DNase accessibility peaks across 125 cell lines assayed by the ENCODE project [3]. Each accessible region is labeled with the number of cell lines in which it is detected. Though there are many cell-type specific peaks, the HAP1 open chromatin regions match sites commonly accessible across many cell lines.



B	positively selected in both	negatively selected in both	positively selected in rep 1	negatively selected in rep 1	positively selected in rep 2	negatively selected in rep 2
gRNA targeting non-6TG resistance regions (**A**)	0	336	2	520	3	344
random sequence not found in hg19	0	9	0	12	0	9

Figure S8: None of the negative control random-sequence gRNAs were positively selected in both individual gRNA screen replicates.

A) Selection scores across replicates for individual gRNAs that target regions not expected to induce 6TG resistance (as described in **Fig. S5**). Only gRNAs sampled in both replicates are plotted.

B) Table of the negative control gRNAs selected in both, either, or neither biological replicate.

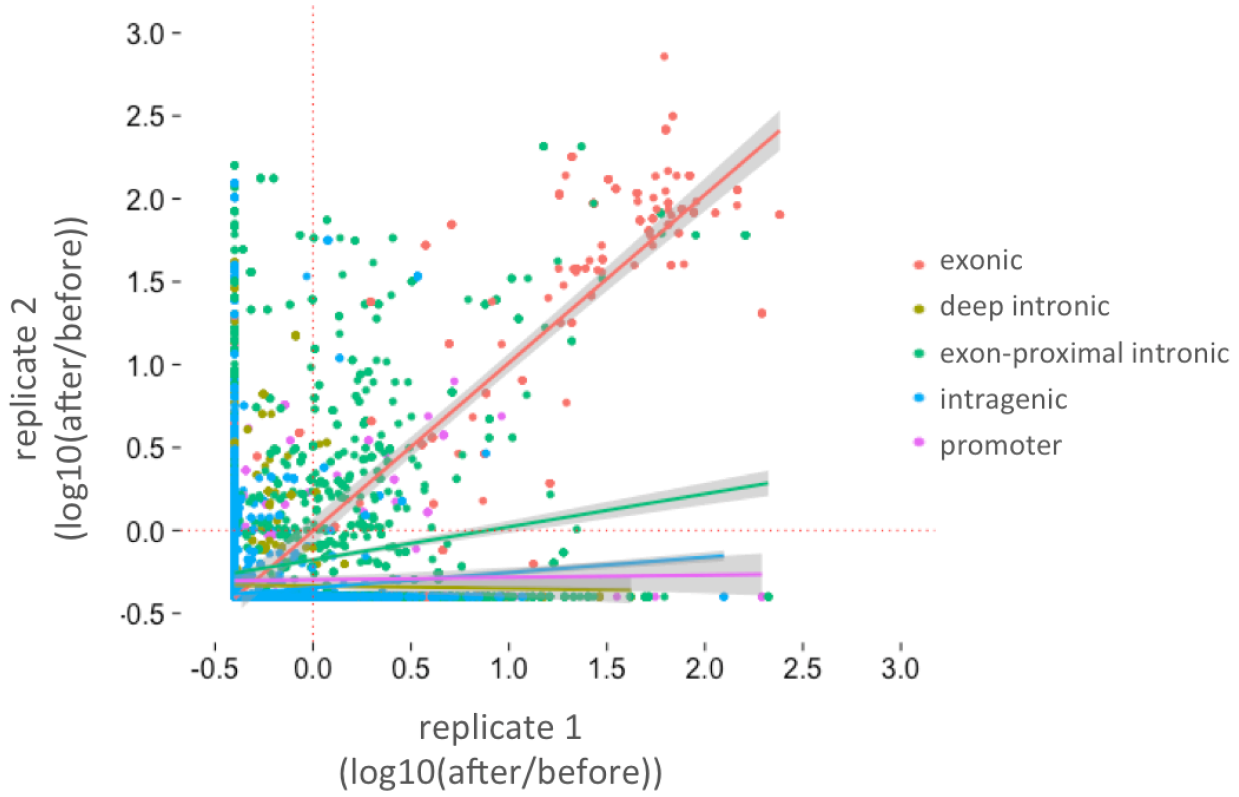


Figure S9: Correlation of the individual gRNA screen scores across two biological replicates.

The individual gRNA scores for each biological replicate were calculated per base-pair and presented as mean of log10(after/before 6TG) between replicates. Least squares lines and points are colored by sequence content category. Specifically, intronic sequence within 2 Kb of an exon is colored in green (Pearson: 0.176); exons are red (Pearson: 0.818); deep intronic is yellow (Pearson: -0.14); intragenic sequences are blue (Pearson: 0.070; and promoter sequence (2 Kb upstream of the TSS) is purple (Pearson: 0.022).

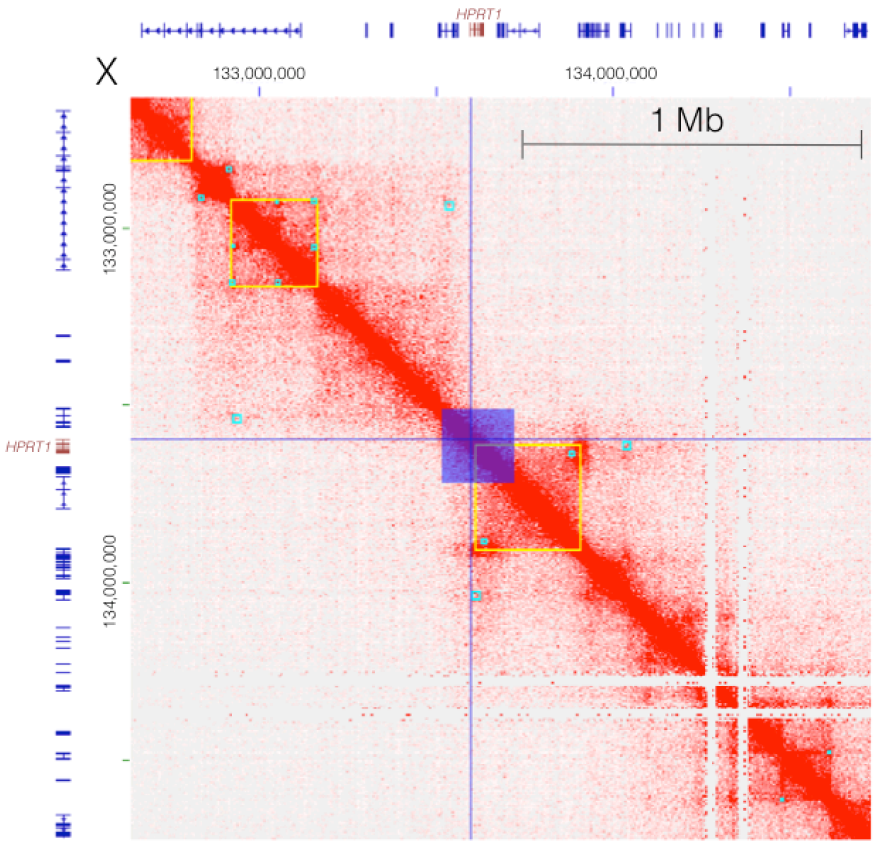


Figure S10: Region interrogated with ScanDel only partially surveys a 300 Kb topologically associated domain (TAD) found in HAP1 cells.

A heatmap of interactions between 5 Kb bins along chrX:132,669,000-134,716,000 (hg19) in HAP1 cells[4] (Juicebox 1.4 [5], balanced normalization). RefSeq gene annotations are drawn across the axes, with the *HPRT1* gene model drawn in red. Blue lines mark its TSS and the 206 Kb surveyed by ScanDel is highlighted as a dark blue box. Light blue boxes mark peaks and yellow boxes mark TADs as called by Sanborn et al.

Supplementary Tables

gRNA pair spacer 1	gRNA pair spacer 2	distance of closest protospacer to TSS (bp)	replicate 1 before 6TG raw read count (number of reads)	replicate 1 after 6TG raw read count (number of reads)	replicate 1 before 6TG normalized read count (number of reads / total reads from sample)	replicate 1 after 6TG normalized read count (number of reads / total reads from sample)	replicate 1 selection score (log 10 after / before)	replicate 2 before 6TG normalized read count (number of reads)	replicate 2 after 6TG normalized read count (number of reads)	replicate 2 before 6TG normalized read count (number of reads / total reads from sample)	replicate 2 after 6TG normalized read count (number of reads / total reads from sample)	selection score (log 10 enrichment ratio) (log 10 after / before)
CCAAGACCTTGCACTACCTG	TGGTGGATGCTGGAGCTATA	316	519	679	0.000208221	0.000510065	0.3891006	769	1	0.000273374	9.30685E-07	-2.4679549
GGACAGTACAGTCAGCAAAT	AATCAGGGAGCCCTCTGAAT	194	108	1	4.33292E-05	7.512E-07	-1.7610255	26	1	9.24282E-06	9.30685E-07	-0.9970019
TATTATGGAACACGTAACAT	CAGGCTCACTAGGTAGCCGT	105	53	511	2.12634E-05	0.000383863	1.2565433	*not sampled*				
GCGCGGGCTGACTGCTCAGG	CTTTATCCTGGAGAGGCAGC	-123	856	5716	0.000343424	0.004293857	1.0970167	1139	4260	0.000404907	0.003964716	0.9908573

Table S1. Read count data and selection scores for the 4 gRNA pairs upstream of exon 1 used for **Fig. 3A-D**. Green is positively selected and red is negatively selected.

gRNA pair spacer 1	gRNA pair spacer 2	distance of closest protospacer to 3' boundary of exon 1 (bp)	replicate 1 before 6TG raw read count (number of reads)	replicate 1 after 6TG raw read count (number of reads)	replicate 1 before 6TG normalized read count (number of reads / total reads from sample)	replicate 1 after 6TG normalized read count (number of reads / total reads from sample)	replicate 1 selection score (log 10 after / before)	replicate 2 before 6TG normalized read count (number of reads)	replicate 2 after 6TG normalized read count (number of reads)	replicate 2 before 6TG normalized read count (number of reads / total reads from sample)	replicate 2 after 6TG normalized read count (number of reads / total reads from sample)	replicate 2 selection score (log 10 after / before)
AAACTGGCCGCCCCCGCCTG	GCGTCTACCTAGGCCAGGCA	117	254	2331	0.000101904	0.001751046	1.2351068	1	1	3.55493E-07	9.30685E-07	0.4179714
ATCCGCAGTGCGGGCTCGGG	CTAAAGCATATTTAACTGGC	75	157	946	6.29879E-05	0.000710635	1.0523897	499	36	0.000177391	3.35046E-05	-0.7238266
CACGCAGTCCTCTTTTCCCA	GCGTCTACCTAGGCCAGGCA	221	100	342	4.01197E-05	0.00025691	0.8064243	145	1	5.15465E-05	9.30685E-07	-1.7433966
GCGTCTACCTAGGCCAGGCA	GTTCAGGCCCACGCGGCAGG	30	184	586	7.38202E-05	0.000440203	0.775478	35	14134	1.24423E-05	0.013154294	3.0241685
GGCACGGAAAGCGACCACCT	AGCAACCCTTGCATGGCCCC	431	1034	5822	0.000414837	0.004373485	1.0229499	2718	18435	0.00096623	0.017157168	1.2493651

Table S2. Read count data and selection scores for the 5 gRNA pairs in intron 1 used for **Fig. 3E-H**.

gRNA (spacer sequence)	distance from TSS (bp)	replicate 1 before 6TG raw read count (number of reads)	replicate 1 after 6TG raw read count (number of reads)	replicate 1 before 6TG normalized read count (number of reads / total reads from sample)	replicate 1 after 6TG normalized read count (number of reads / total reads from sample)	replicate 1 selection score (log 10 after / before)	replicate 2 before 6TG normalized read count (number of reads)	replicate 2 after 6TG normalized read count (number of reads)	replicate 2 before 6TG normalized read count (number of reads / total reads from sample)	replicate 2 after 6TG normalized read count (number of reads / total reads from sample)	replicate 2 selection score (log 10 after / before)
TATTATGGAACACGTAACAT	1910	7	1	7.35E-07	1.29E-07	-0.757	*not sampled*				
CTTTATCCTGGAGAGGCAGC	1663	540	270	5.67E-05	3.47E-05	-0.213	1437	2	0.000172975	2.42125E-07	-2.854
TGGTGGATGCTGGAGCTATA	1111	82	26	8.61E-06	3.35E-06	-0.411	1523	542	0.000183327	6.56159E-05	-0.446
CTGCTAATTAATTCTCAGAT	1033	385	1	4.04E-05	1.29E-07	-2.497	*not sampled*				
GGACAGTACAGTCAGCAAAT	931	856	1	8.99E-05	1.29E-07	-2.844	662	28	7.96864E-05	3.38975E-06	-1.371
CCAAGACCTTGCACTACCTG	308	858	4110	9.01E-05	5.29E-04	0.769	195	1	2.34726E-05	1.21063E-07	-2.288
CCAGTCATCGCGTGAATCCT	269	1731	13867	1.82E-04	1.78E-03	0.992	967	82	0.0001164	9.92714E-06	-1.069
AATCAGGGAGCCCTCTGAAT	186	216	1	2.27E-05	1.29E-07	-2.246	258	1	3.1056E-05	1.21063E-07	-2.409
CAGGCTCACTAGGTAGCCGT	97	1282	11735	1.35E-04	1.51E-03	1.050	288	42	3.46672E-05	5.08463E-06	-0.834
GCGCGGGCTGACTGCTCAGG	-131	1029	4297	1.08E-04	5.53E-04	0.709	895	50	0.000107733	6.05313E-06	-1.250

Table S3. Read count data and selection scores (for both replicates of the individual gRNA screen) for the 10 individual gRNAs targeting regions upstream of exon 1 and displayed in **Fig. 4**.

Supplementary References

1. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*. 2010;20(1):110-121. doi:10.1101/gr.097857.109.
2. John S, Sabo PJ, Thurman RE, et al. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat Genet*. 2011;43(3):264-268. doi:10.1038/ng.759.
3. Rosenbloom KR, Sloan CA, Malladi VS, et al. ENCODE Data in the UCSC Genome Browser: Year 5 update. *Nucleic Acids Res*. 2013;41(D1):56-63. doi:10.1093/nar/gks1172.
4. Sanborn AL, Rao SSP, Huang S-C, et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci*. 2015;112(47):201518552. doi:10.1073/pnas.1518552112.
5. Durand NC, Robinson JT, Shamim MS, et al. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst*. 2016;3(1):99-101. doi:10.1016/j.cels.2015.07.012.