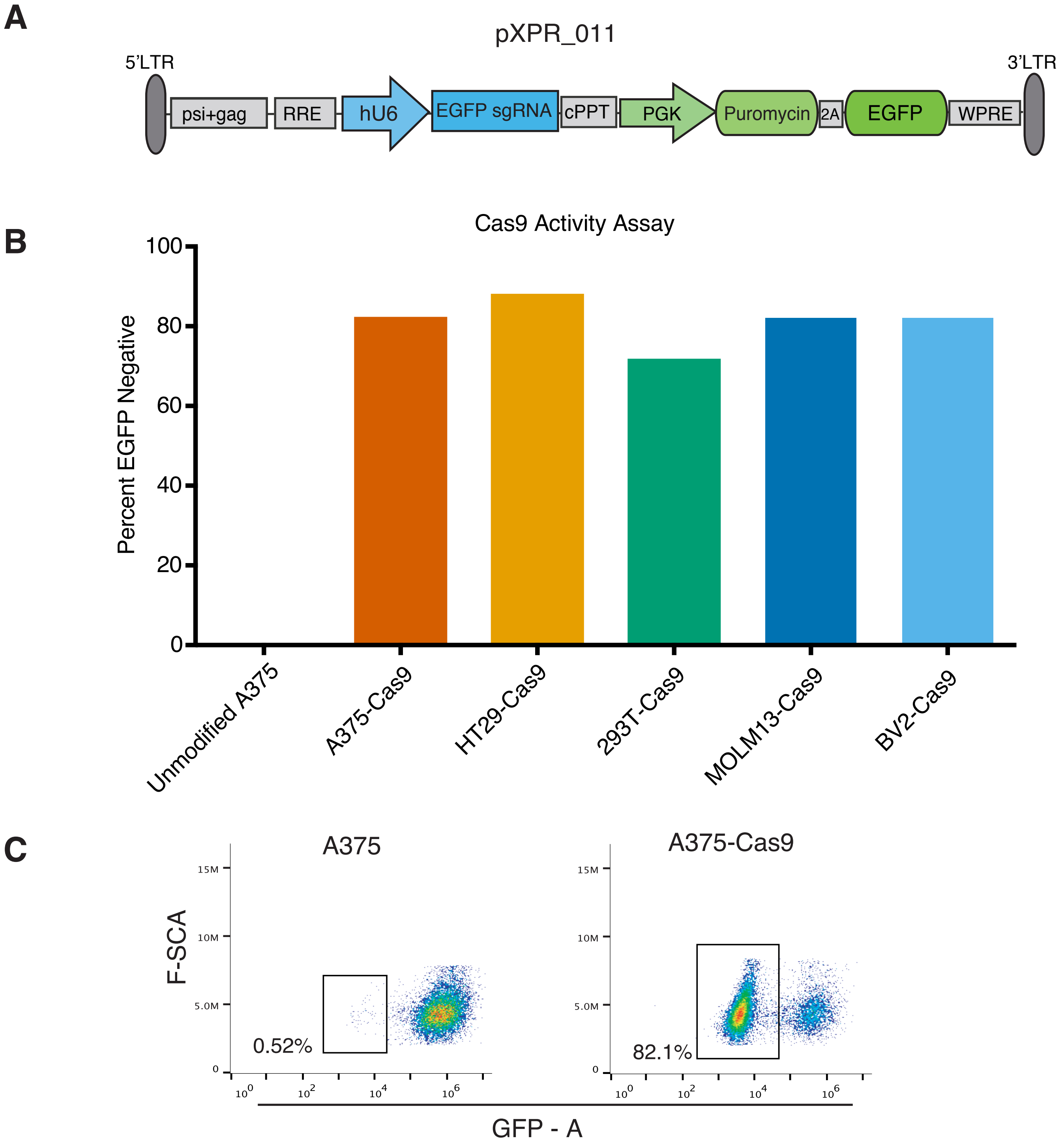
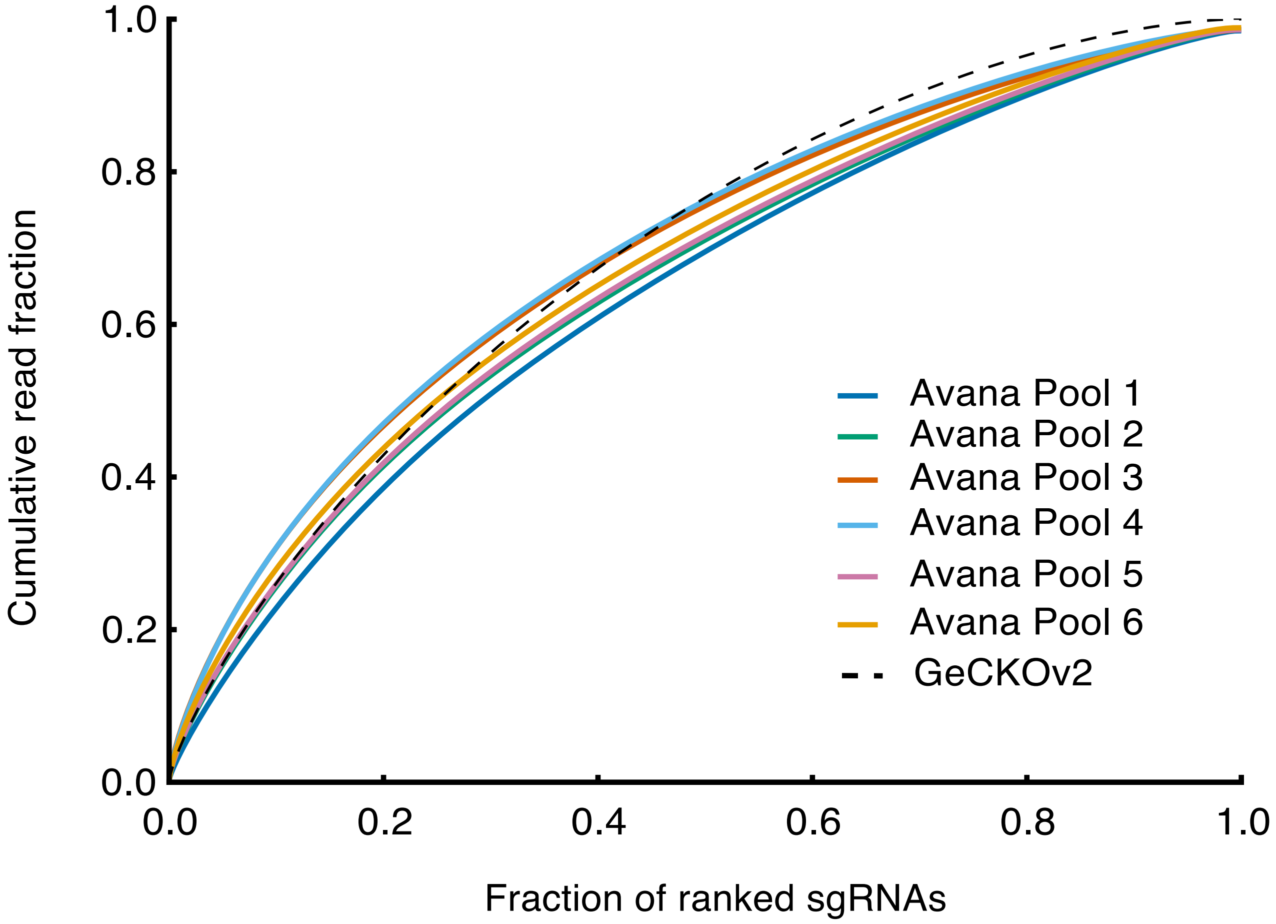
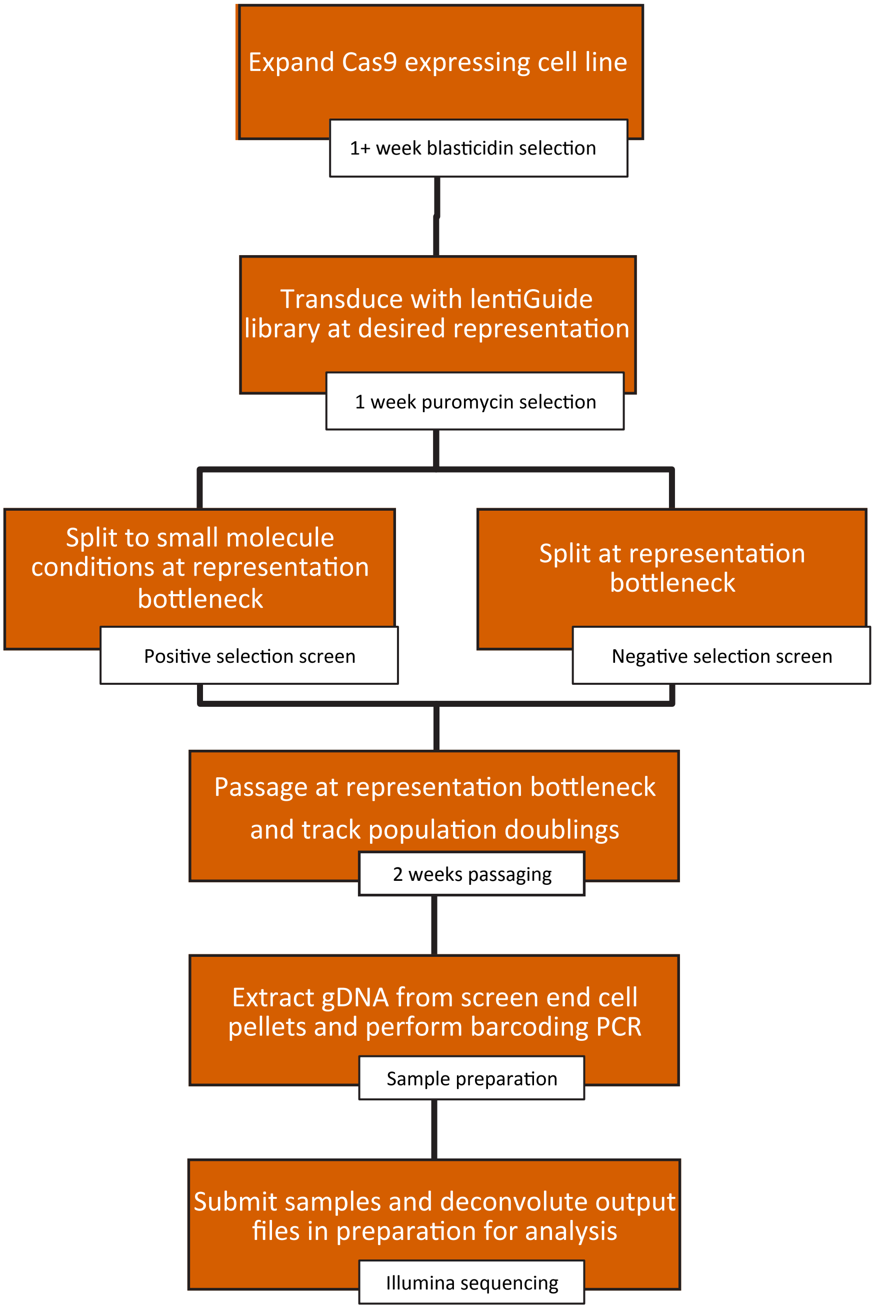


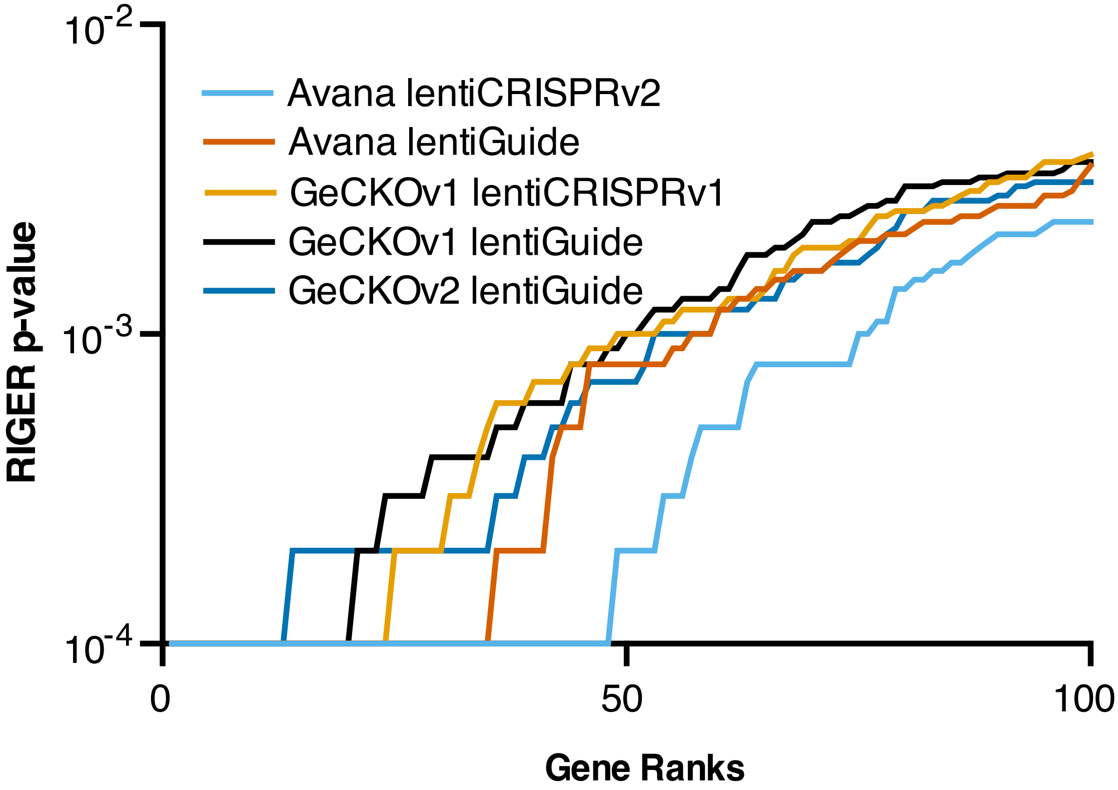
Supplementary Figure 1. Cas9 activity assays for cell lines used in this study. (**a**) Schematic of pXPR_011 (Addgene #59702). Cells that do not express Cas9 will express EGFP and puromycin resistance, while cells with sufficient levels of active Cas9 will utilize the sgRNA targeting EGFP. Placement of EGFP downstream of a 2A site allows for continued puromycin resistance after indel formation. (**b**) Quantitation of the fraction of cells with Cas9 activity. Samples were analyzed ten to fourteen days post-infection.(**c**) Example flow cytometey plots.



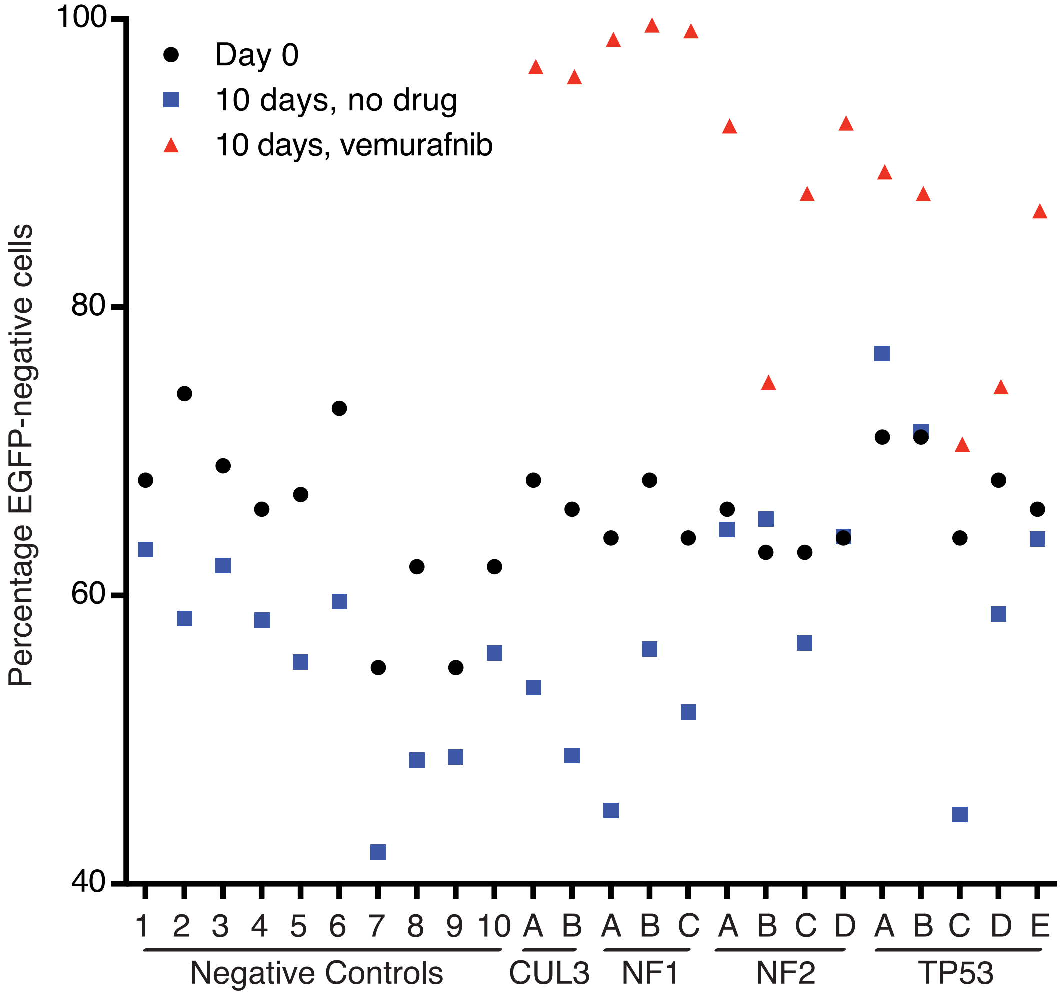
Supplementary Figure 2 Distribution of sgRNA abundance in the plasmid DNA pool of each Avana subpool in lentiGuide (solid lines) and GeCKOv2 library in lentiGuide (dashed line).


Expand Cas9 expressing cell line
1+ week blasticidin selection
Transduce with lentiGuide library at desired representation
1 week puromycin selection
Split to small molecule conditions at representation bottleneck
Positive selection screen
Split at representation bottleneck
Negative selection screen
Passage at representation bottleneck and track population doublings
2 weeks passaging
Extract gDNA from screen end cell pellets and perform barcoding PCR
Sample preparation
Submit samples and deconvolute output files in preparation for analysis
Illumina sequencing

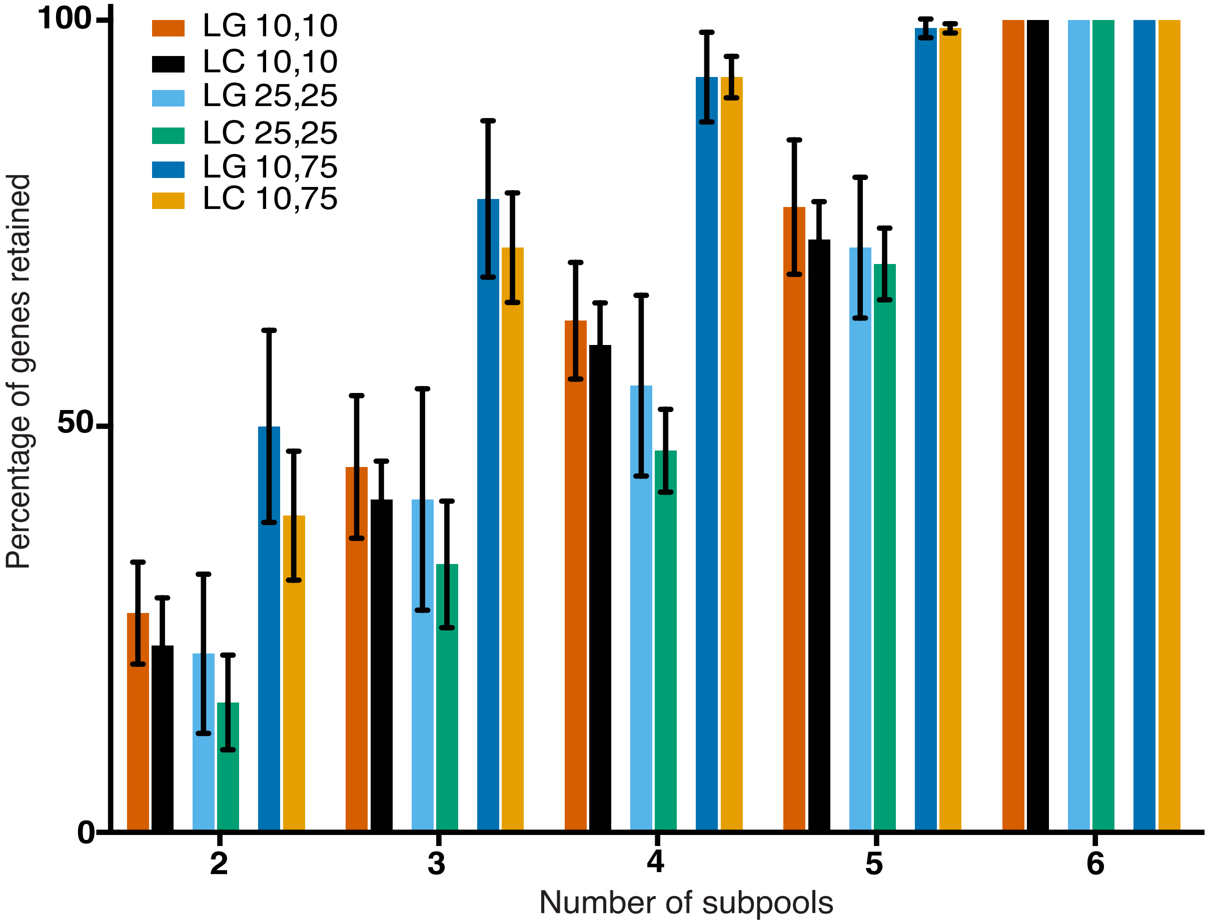
Supplementary Figure 3 Pooled Screen workflow. Cells are infected with a Cas9-expression plasmid and are subjected to one week or more of blasticidin selection. Cells are infected with library and selected with puromycin for one week before being split into small molecule conditions or continued passaging for positive and negative selection screens, respectively. At then end point, cell pellets are collected and gDNA extracted. Samples are then barcoded and sequenced by Illumina. For analysis, the $\log_2$-fold-change of each sgRNA is determined relative to the starting plasmid DNA (pDNA) pool.



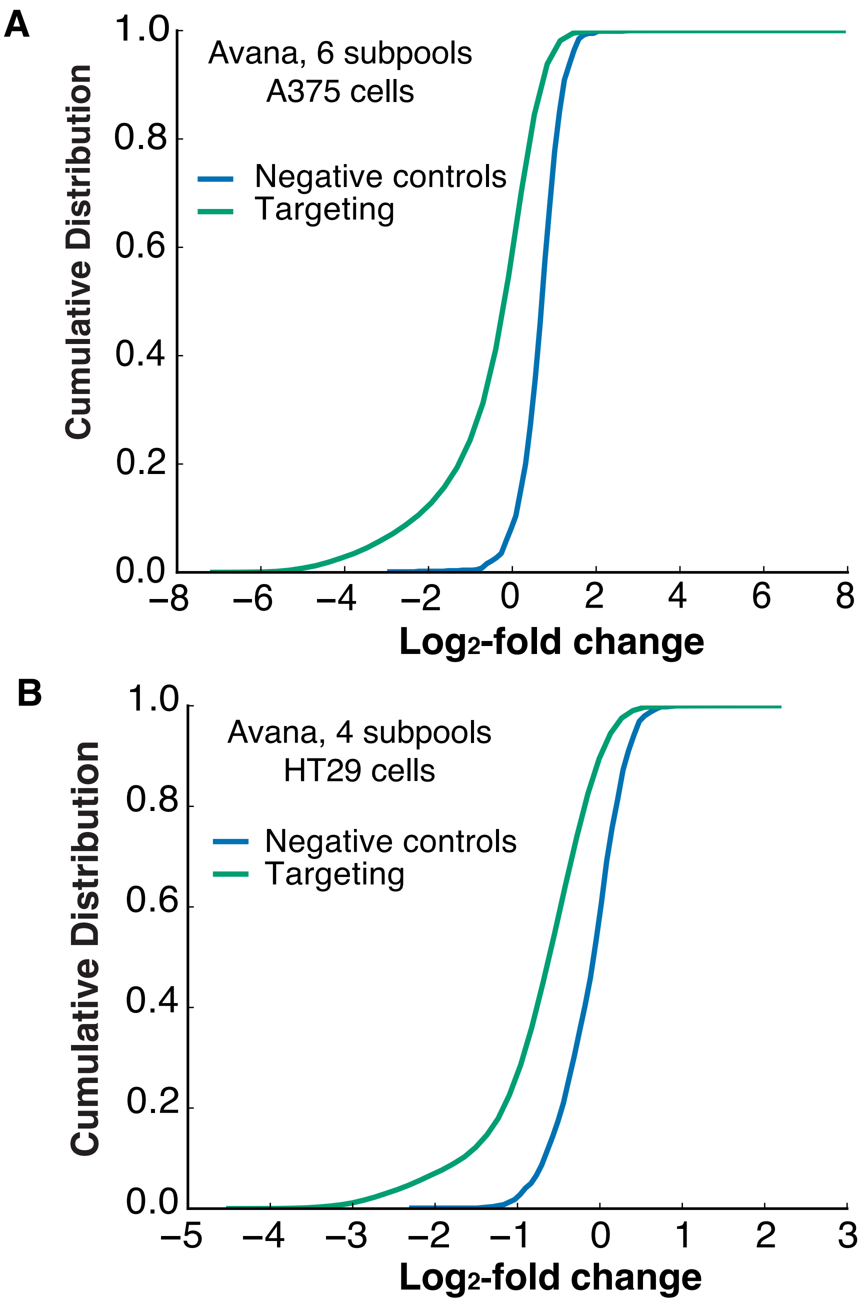
Supplementary Figure 4 Comparison of the p-values for the top 100 genes determined by weighted sum analysis in RIGER. y-axis is plotted in log10-scale.



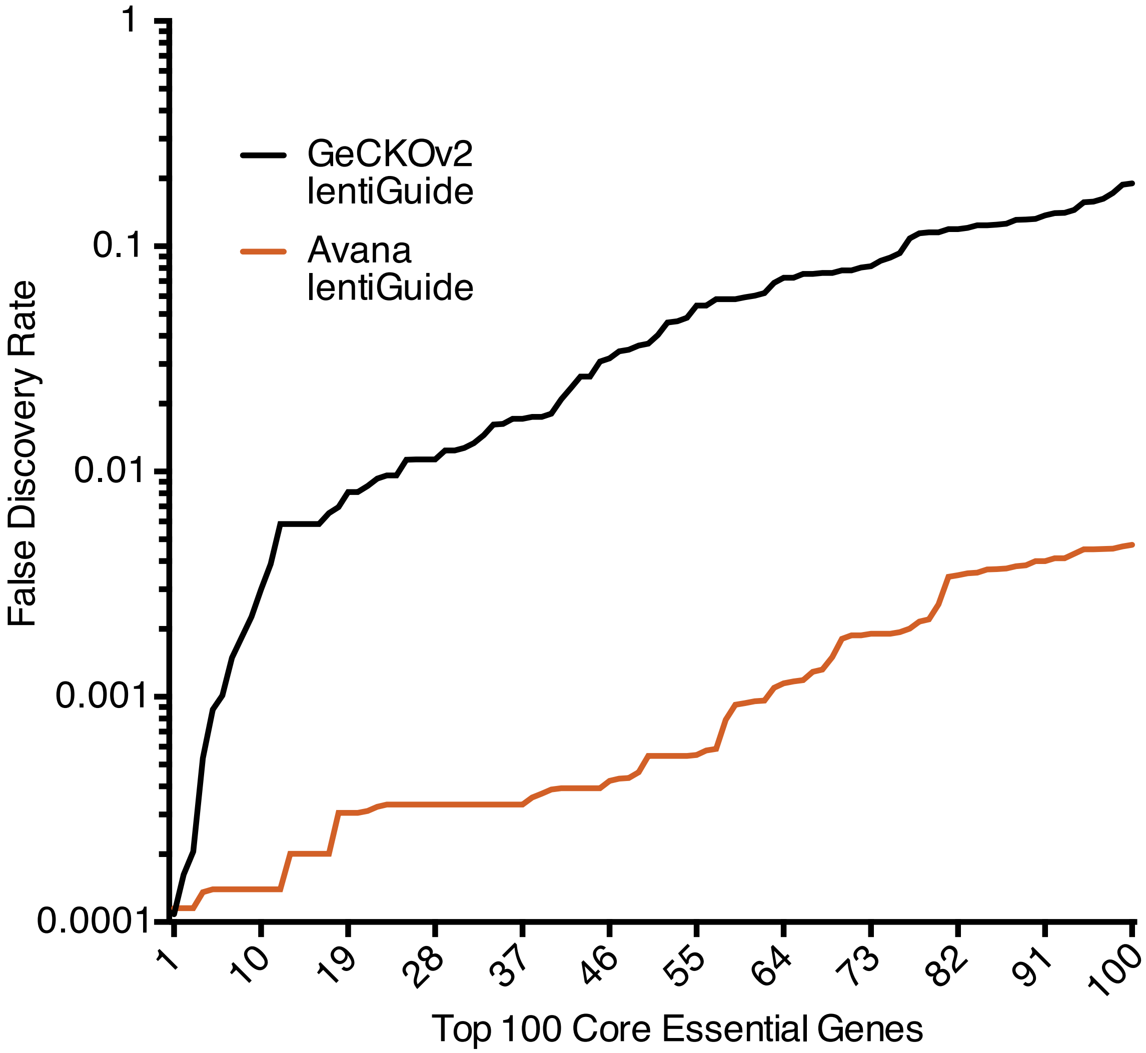
Supplementary Figure 5 Validation of EGFP competition assay to confirm individual sgRNA activity. A375-Cas9 cells were infected with individual sgRNAs, selected with puromycin for one week, and then mixed with A375-Cas9 cells also expressing EGFP and puromycin resistance. The ratio in each population was determined by flow cytometry. Populations were then maintained with vemurafenib or with no drug for 10 days, and the relative fraction of sgRNA-containing, EGFP-negative cells was determined by flow cytometry. For the 10 negative controls, all cells died after 10 days in the presence of vemurafenib and thus could not be assessed by flow cytometry.



Supplementary Figure 6 Subsampling analysis of Avana library in selumetinib resistance screening data. STARS was run on all six subpools to determine the number of genes that passed at FDR thresholds of 10% and 25% (first number in legend). Subpools were then iteratively removed and STARS was re-run to determine the average number of genes retained by using fewer subpools at FDR thresholds of 10%, 25% and 75% (second number in legend). Error bars represent one standard deviation when subsampling from different combinations of subpools.

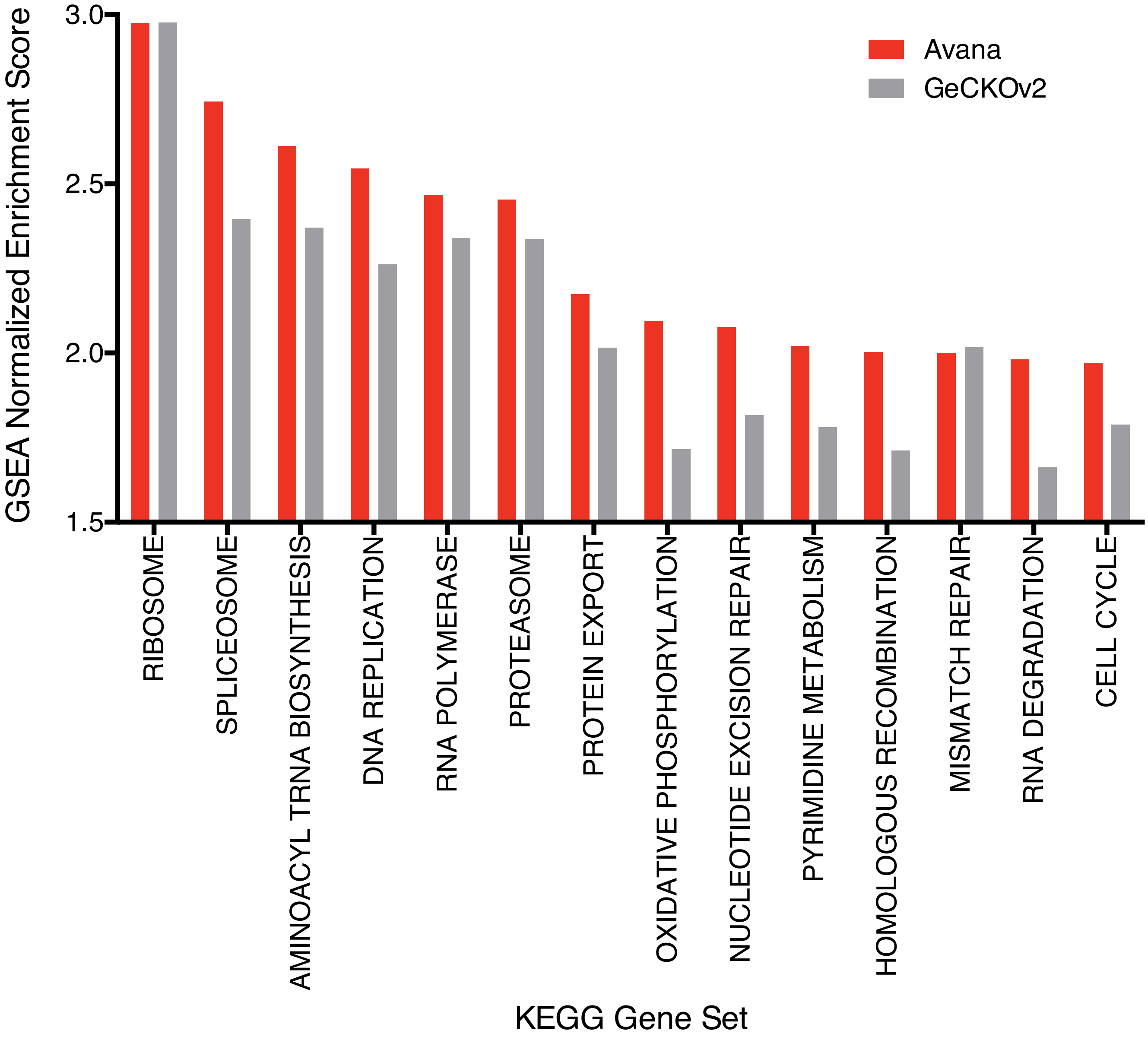


Supplementary Figure 7. Change in abundance of Avana library in negative selection screens. (**a**) Cumulative distribution of 996 negative controls and 108,467 gene targeting sgRNAs in A375 cells for the Avana library in lentiGuide. Data are from two biological replicates. (**b**) Cumulative distribution of 1000 negative controls and 73,687 gene targeting sgRNAs in HT29 cells for the Avana library screened with 4 subpools in lentiGuide. Data are from three biological replicates.

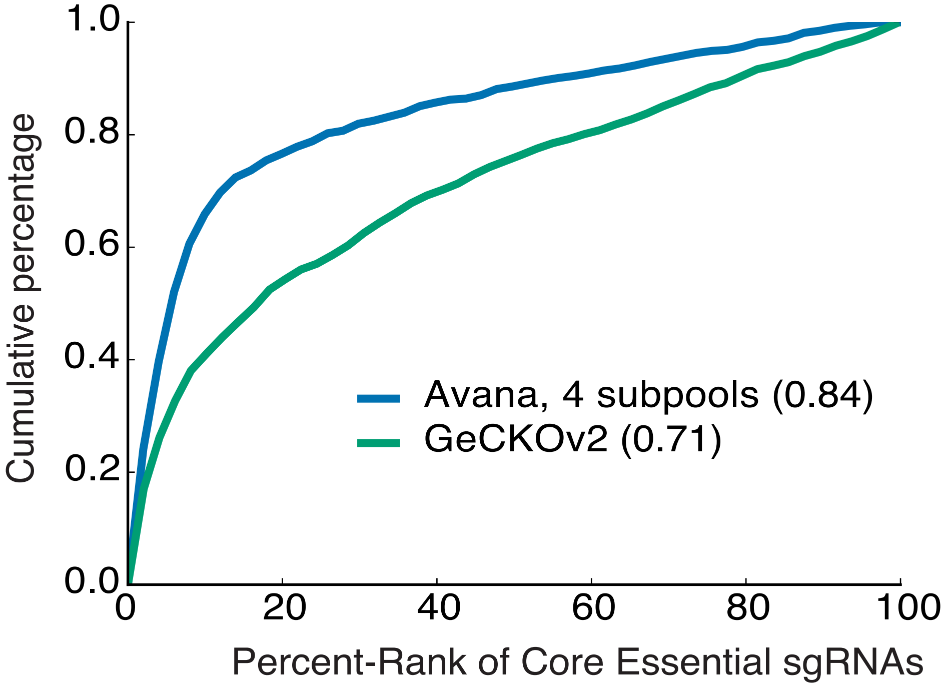


Supplementary Figure 8. Depletion of core essential genes with the Avana and GeCKOv2 libraries as assessed by STARS analysis. Data from two biological replicates were merged and STARS analysis performed for all genes.

Nature Biotechnology: doi:10.1038/nbt.3437

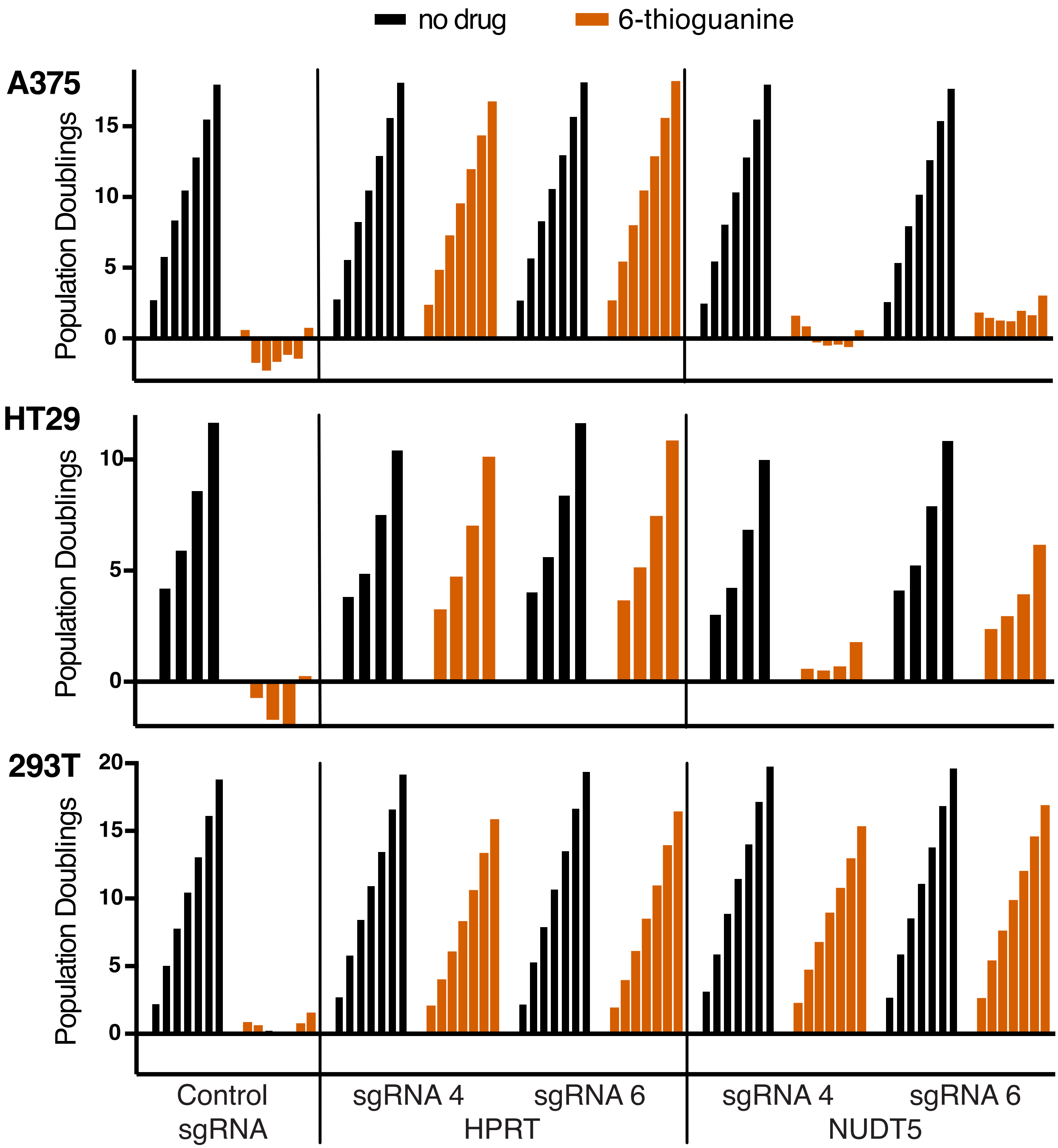


Supplementary Figure 9. Gene Set Enrichment Analysis (GSEA) of negative selection screens performed in A375 cells with the indicated libraries in lentiGuide. All KEGG gene sets with normalized enrichment scores of 2.0 or greater are shown for Avana, with the corresponding score shown for GeCKOv2. The input ranked gene list was determined by RIGER analysis using the weighted sum option.

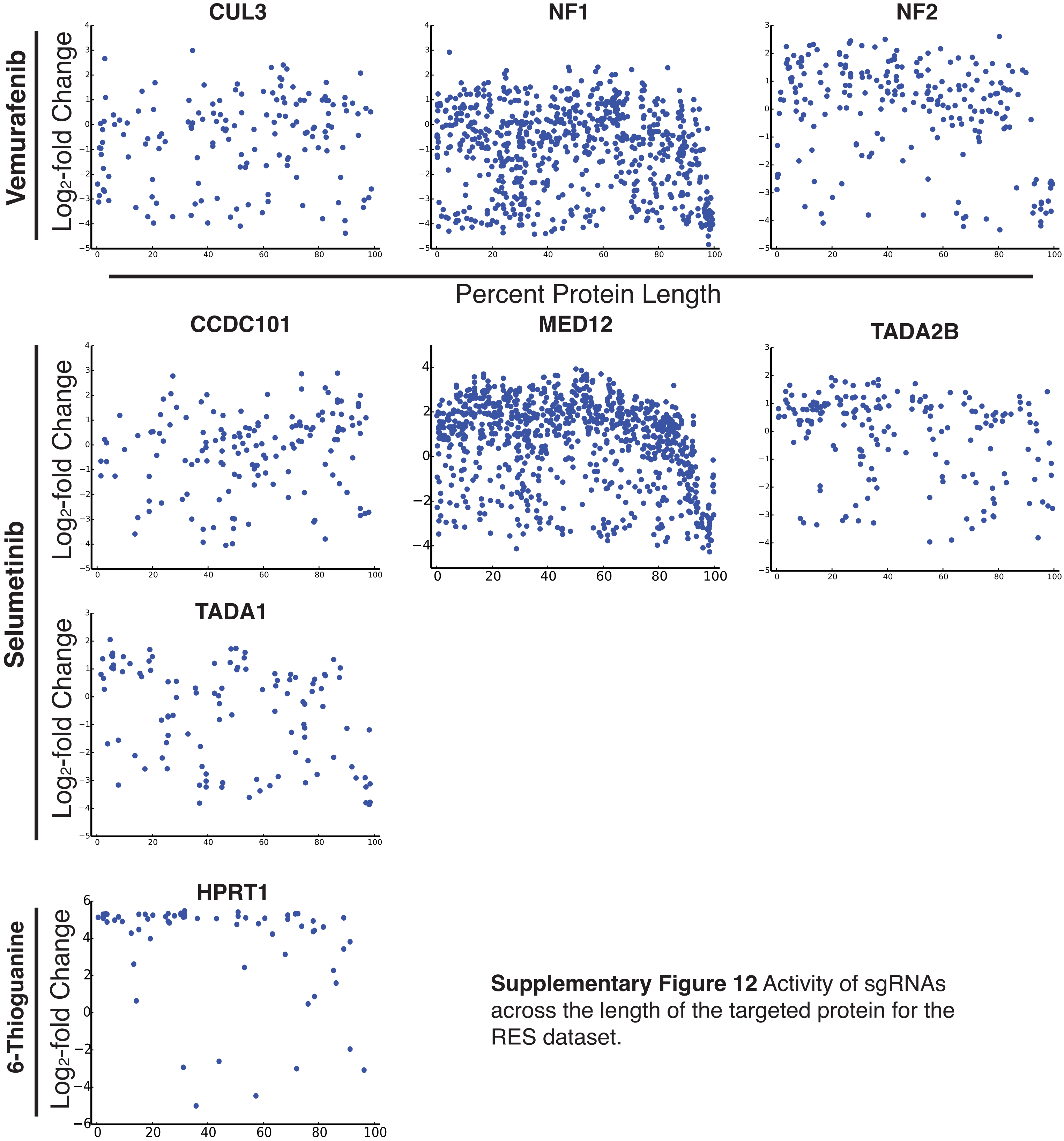


Supplementary Figure 10 ROC-AUC analysis of core essential genes in HT29 cells for Avana screened with 4 subpools and the GeCKOv2 library in lentiGuide. AUCs for each library are specified in brackets. Data are from three biological replicates.

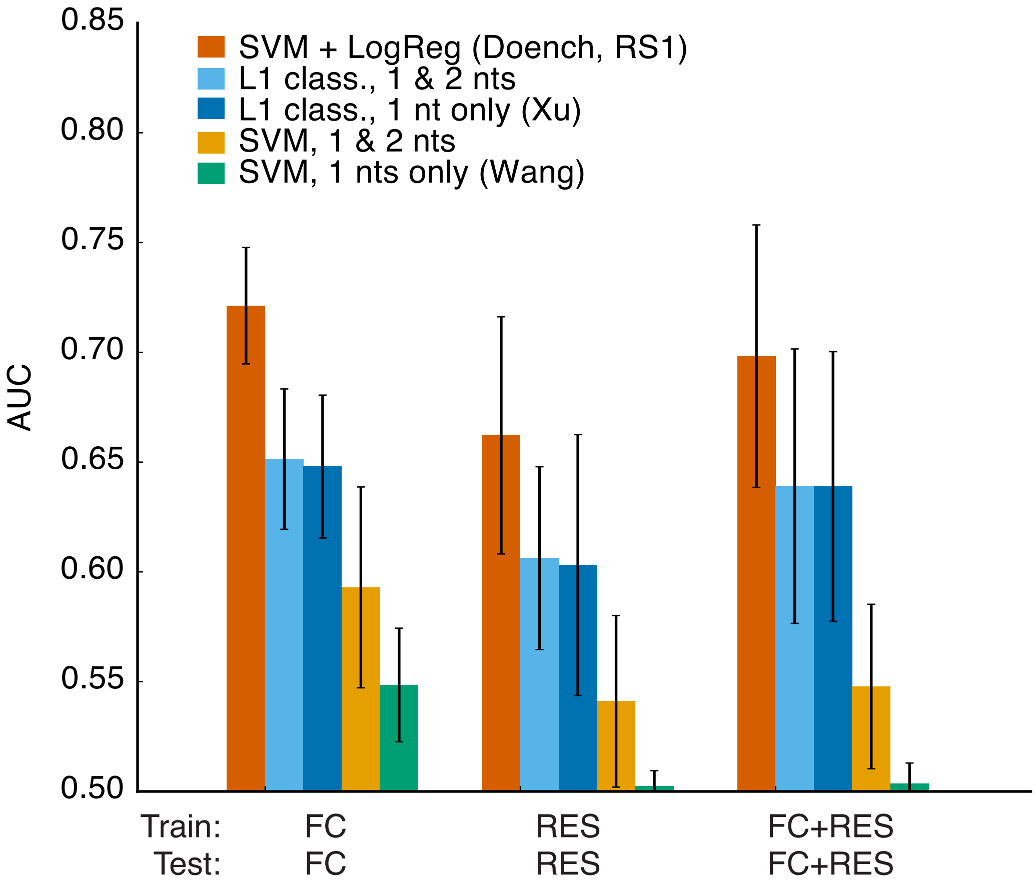
Nature Biotechnology: doi:10.1038/nbt.3437



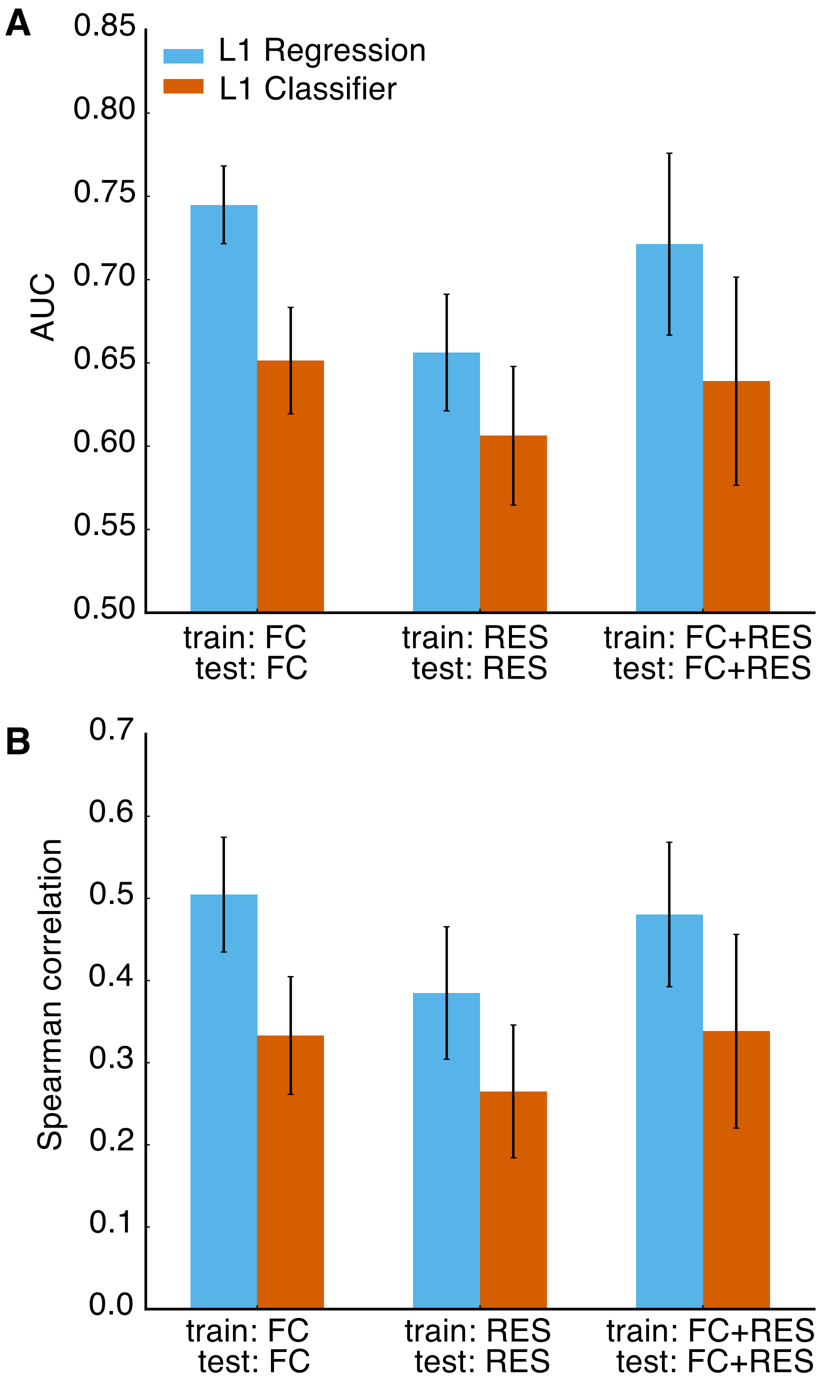
Supplementary Figure 11 Validation of 6-thioguanine resistance hits. Individual sgRNAs cloned into lentiGuide were infected into 3 cell lines, selected with puromycin, and after 1 week of selection were split into 6-thioguanine or continued to be passaged in the absence of drug (day 0). An sgRNA targeting EGFP was used as a control. The populations were counted at each passage, and the cumulative number of population doublings was determined relative to day 0. For A375 cells, the columns represent cell counts taken on days 2, 5, 7, 9, 11, 13, and 15; for HT29 cells, days 5, 7, 10, 15; for 293T cells, days 2, 5, 7, 9, 11, 13, and 15.



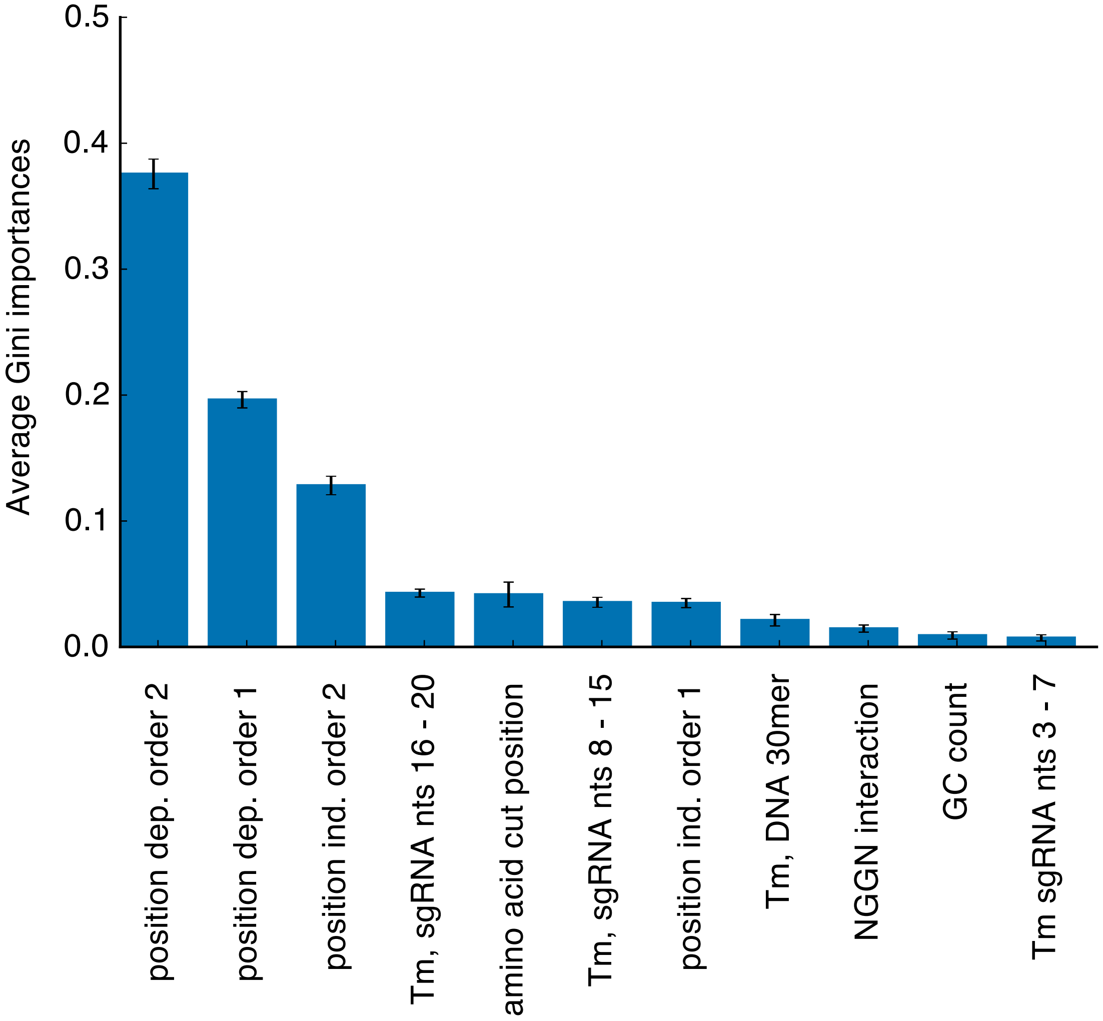
Supplementary Figure 12 Activity of sgRNAs across the length of the targeted protein for the RES dataset.



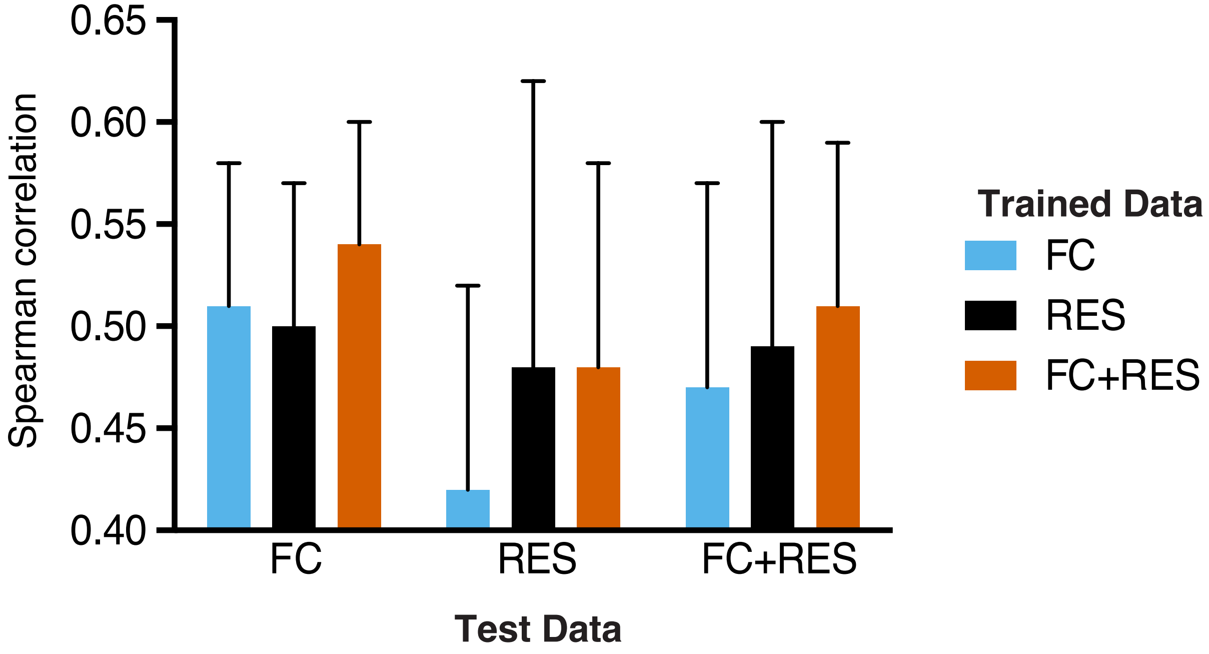
Supplementary Figure 13 Performance of various classification models assessed by area under the curve (AUC). Error bars represent one standard deviation for performance across genes using a leave-one-gene-out method for training and testing.



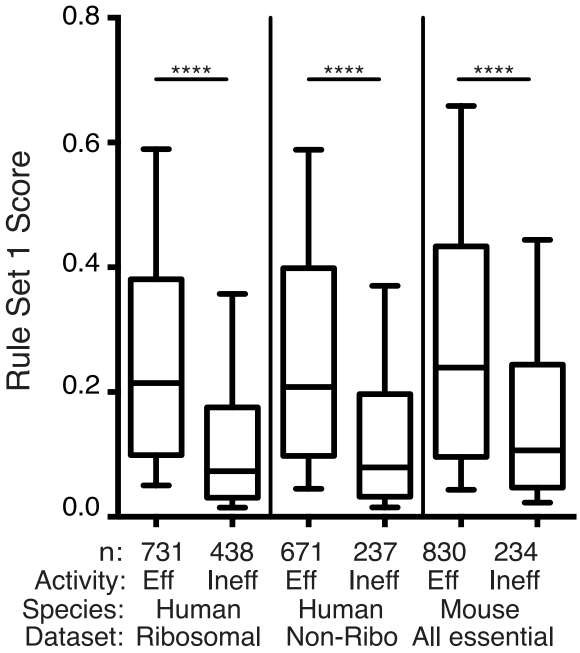
Supplementary Figure 14 Comparison of two comparable predictive models, one of which uses only a 0/1 binarization of the sgRNA scores (L1 Classifier), and the other that uses the sgRNA scores directly (L1 Regression). (**a**) uses area under the curve (AUC) as the performance metric, while (**b**) uses Spearman correlation as the metric. For all three data sets, and across both metrics, the regression approach outperforms the classification approach. Each bar shows the mean value across all genes, with the error bars denoting one standard deviation across the genes. Statistical significance of the improvement in Spearman correlation when using regression over classification is, for FC, RES and FC+RES data, respectively, $p < 1\times10^{-16}$, $p = 5.4\times10^{-13}$, $p < 1\times10^{-16}$



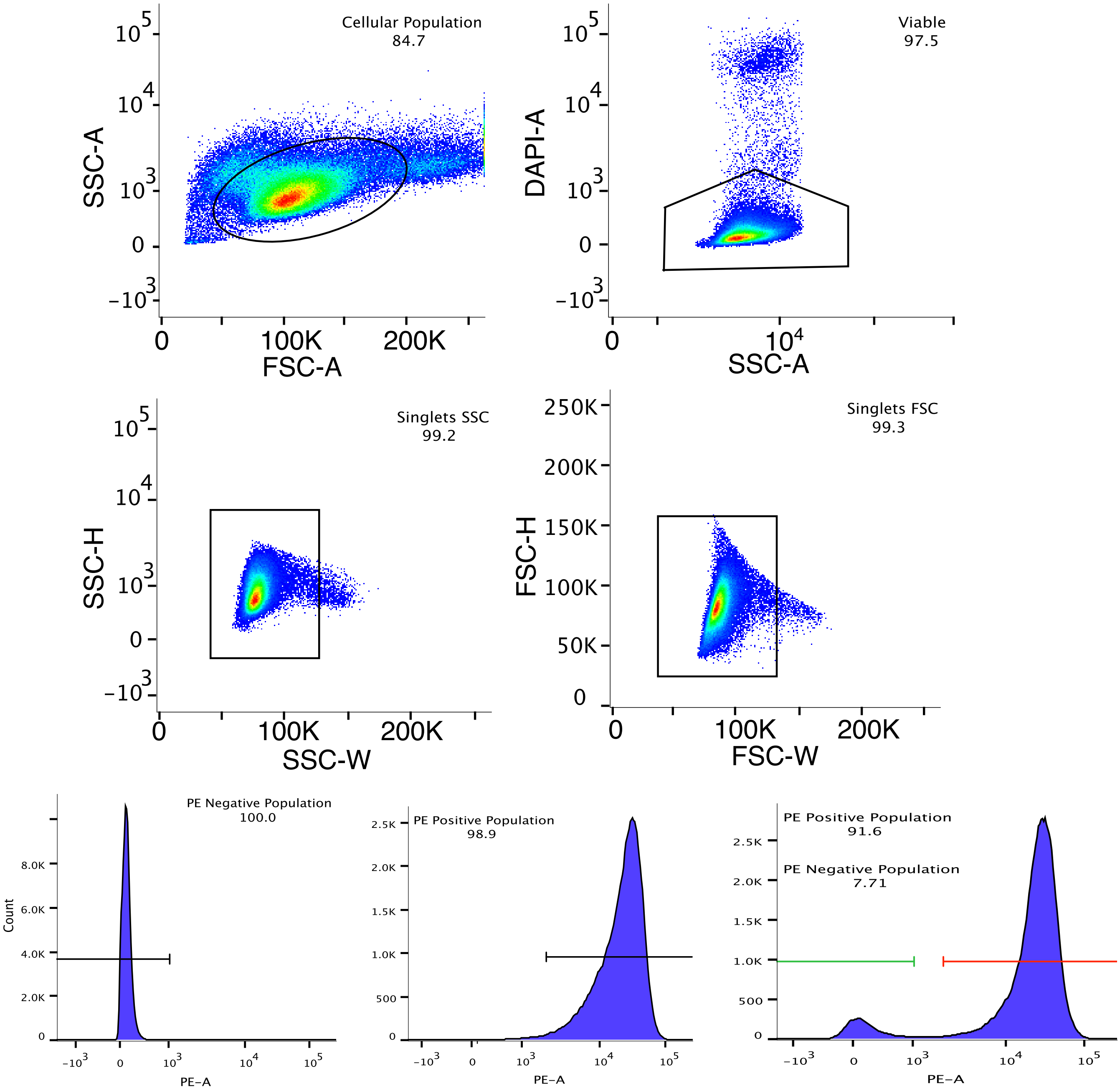
Supplementary Figure 15. Top features weights comprising Rule Set 2. The Gini importances sum to 1.0. The two most important features, postition dependent order 1 and 2, refer to the identity of a specific single and dinucleotides at a particular position, e.g. is there an A at position 3, is there an AG at positions 9 and 10. Position independent order 1 and 2 refer to the total number of particular nucleotides, e.g. how many As are there in the sequence, how many CGs. The 3 melting temperature features (Tm) for the sgRNA were discretized by the region of the sgRNA, with position 1 as the 5' end of the sgRNA and position 20 as the PAM proximal nucleotide. The NGGN interaction consisted of the two nucleotides that occur in the PAM on each side of the invariant GG.



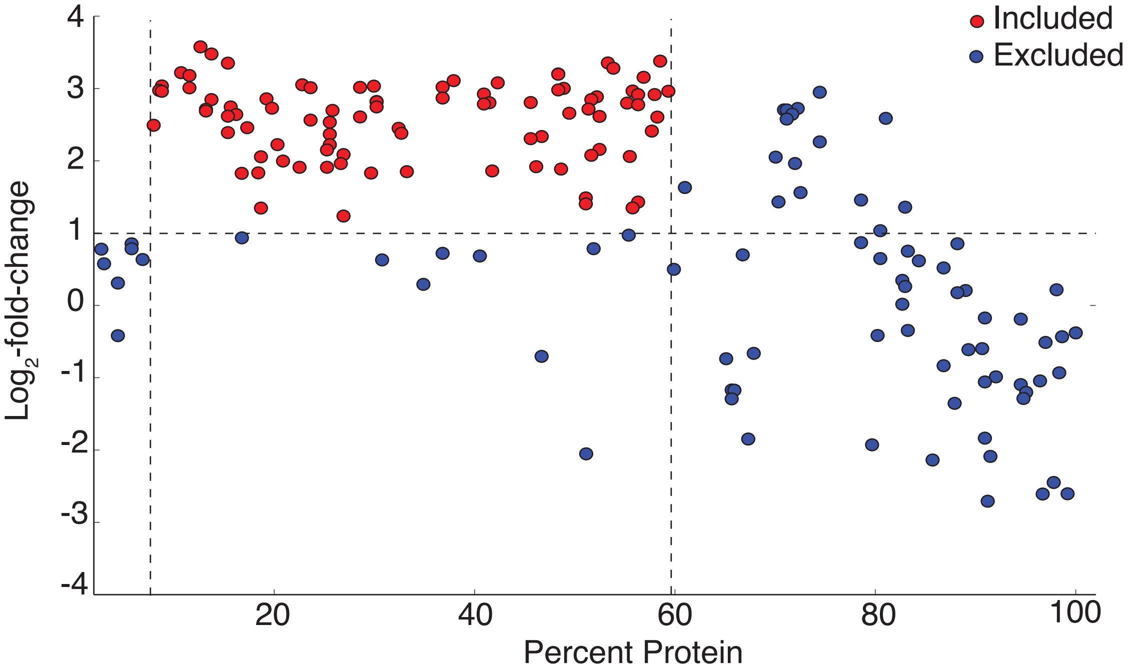
Supplementary Figure 16 Performance of the final model for Rule Set 2, gradient boosted regression trees, when trained and tested on all datasets.



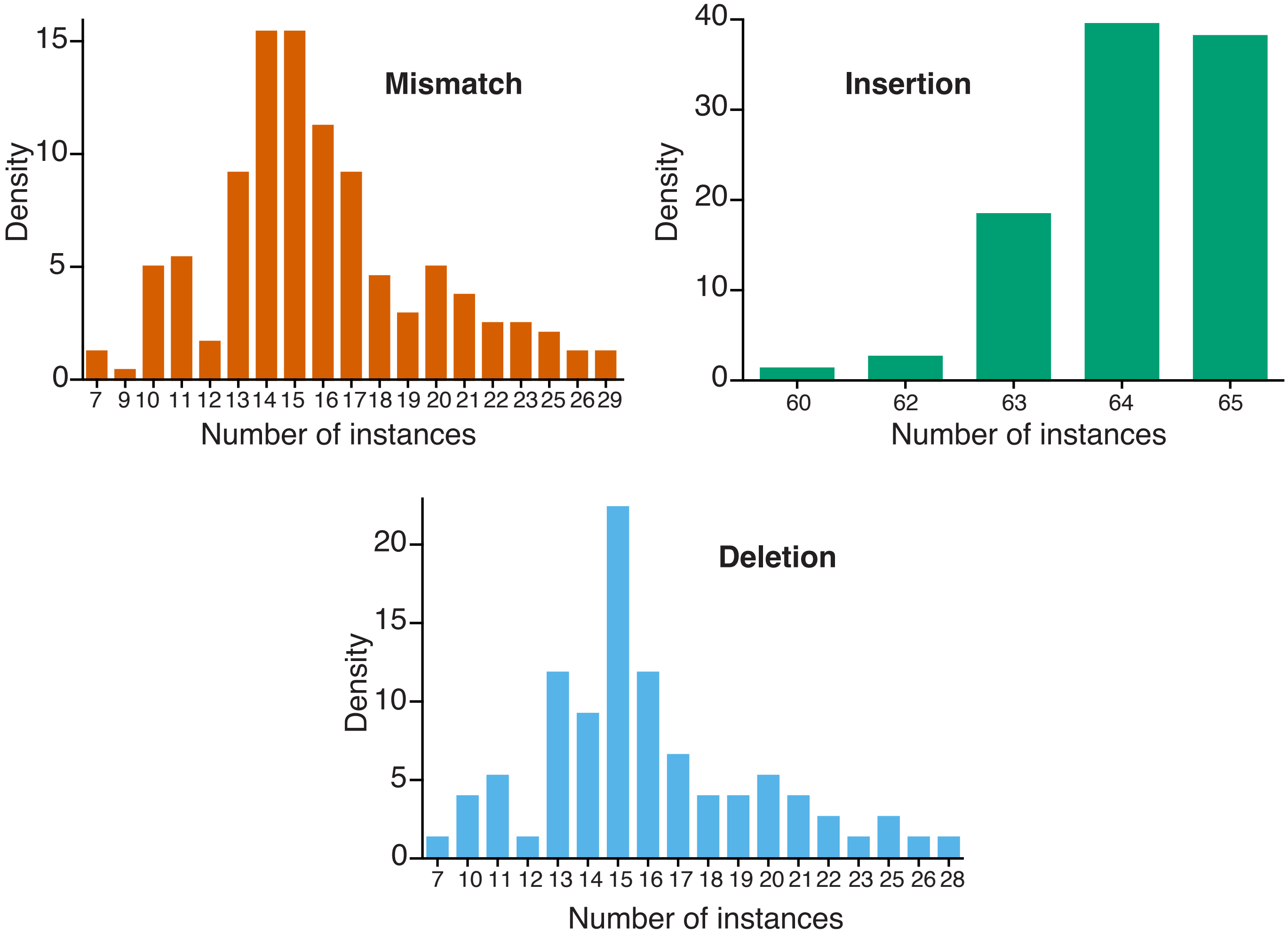
Supplementary Figure 17 Performance of Rule Set 1 on negative selection data, using negative selection data as curated by Xu et al. From Wang et al., sgRNAs were classified as either effective or ineffective for dropout screens performed in HL-60 and KBM-7 human cell lines, and split into two classes, one for ribosomal genes and the other for essential, non-ribosomal genes. The data curated from Koike-Yusa et al. represent a dropout screen performed in mouse ES cells, and all essential genes were kept as one class in the curation performed by Xu et al. Rule Set 1 effectively distinguishes effective from ineffective sgRNAs in all cases, with p-values of $1.4x10^{-32}$, $1.8x10^{-16}$, and $1.1x10^{-11}$ from left to right (two-sample Kolmogorov-Smirnov test).



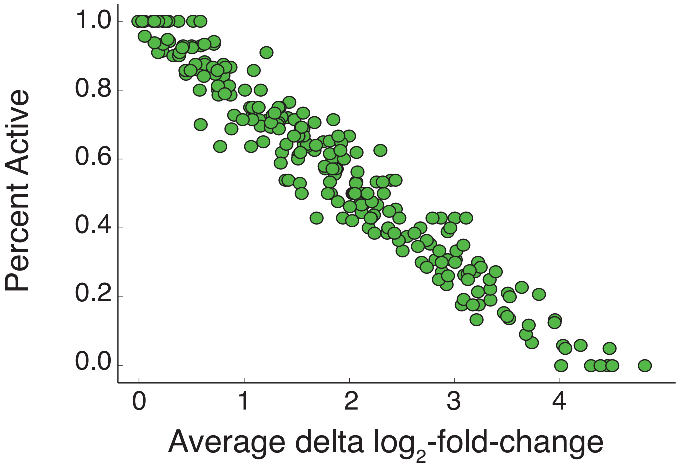
Supplementary Figure 18 CD33 flow cytometry outline. Gates were set on the BD FACSDiVa software. The sample was first gated to exclude debris; two additional gates were applied to select singlets and a fourth to select viable cells via DAPI. Finally, the population was gated, using a CD-33 DAPI positive control and a DAPI only negative control, in order to sort for PE negative cells (green bracket).



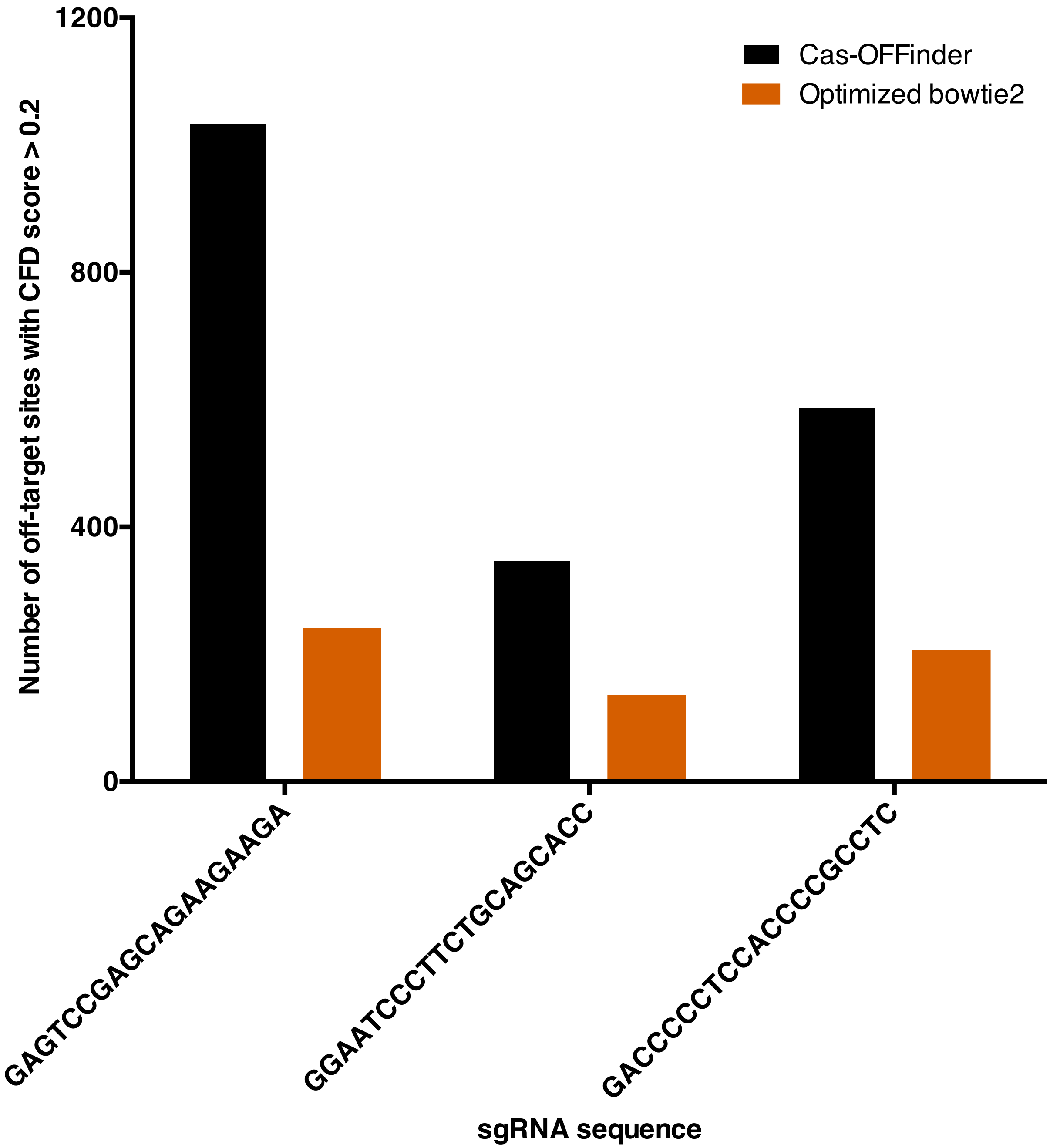
Supplementary Figure 19 Protein-activity map of CD33. Log$_2$-fold change of sgRNAs utilizing the canonical NGG PAM across the length of the target protein are shown. The sgRNAs in the region of highest activity (red) were included for further analysis, and sgRNAs targeted to regions of low activity or exhibiting low activity (blue) were excluded from subsequent analysis.



Supplementary Figure 20 Distributions of the number of unique position/nucleotide instances in the off-target data set, for the three different classes of imperfect interactions.



Supplementary Figure 21 Comparison of the average delta-$\log_2$-fold-change and percent-active calculations for variant sgRNAs for all mismatch interactions; Pearson correlation coefficient = -0.97.



Supplementary Figure 22. Optimized bowtie2 misses potential off-target sites. For three sgRNAs that had previously been assessed by Guide-Seq all possible off-target sites were identified by two search methods, optimized bowtie2 (see Methods) and Cas-OFFinder. The potential sites were then assessed by the CFD score, and the total number of off-target sites in the human genome receiving a CFD score > 0.2 are plotted.