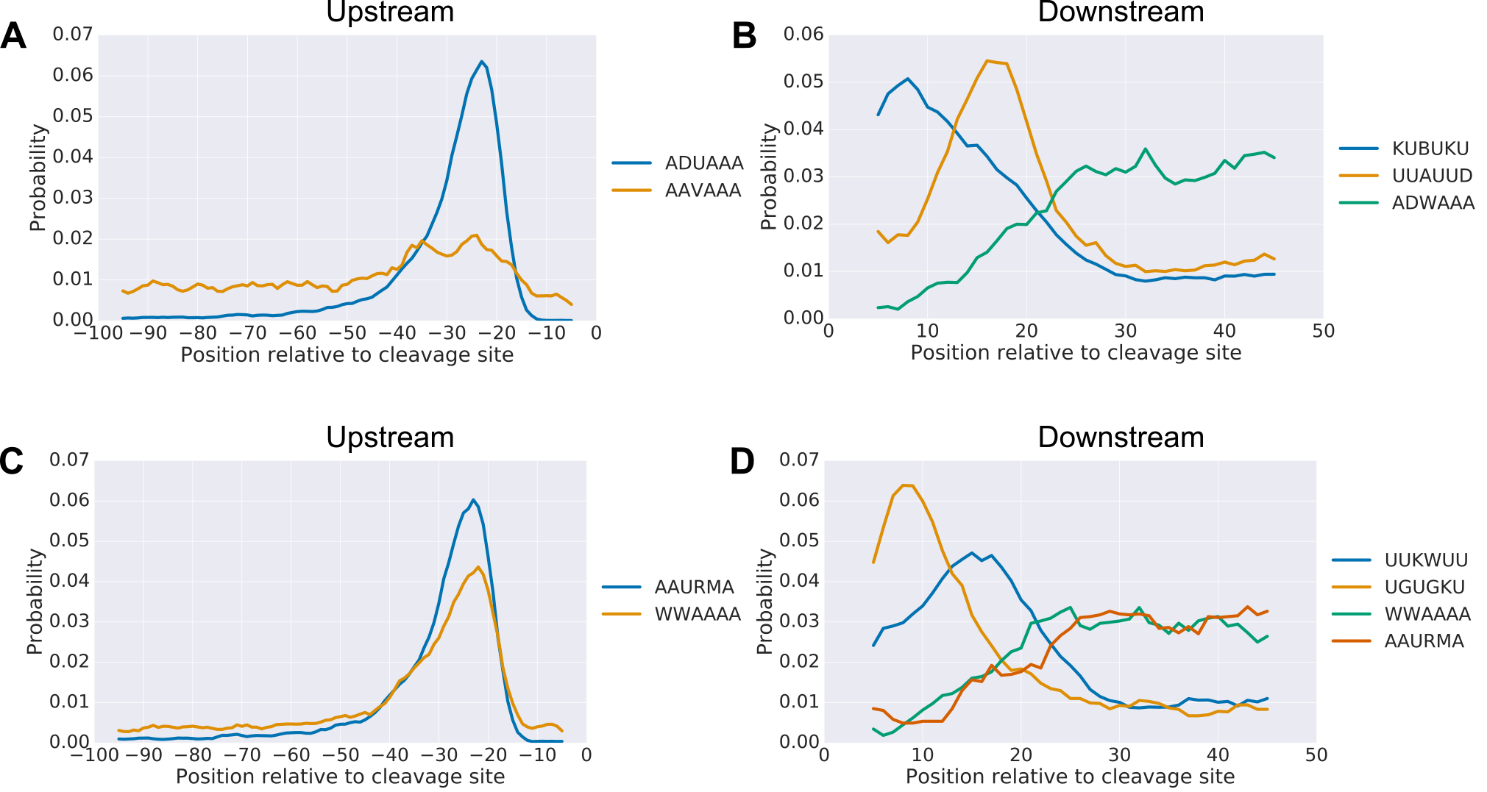
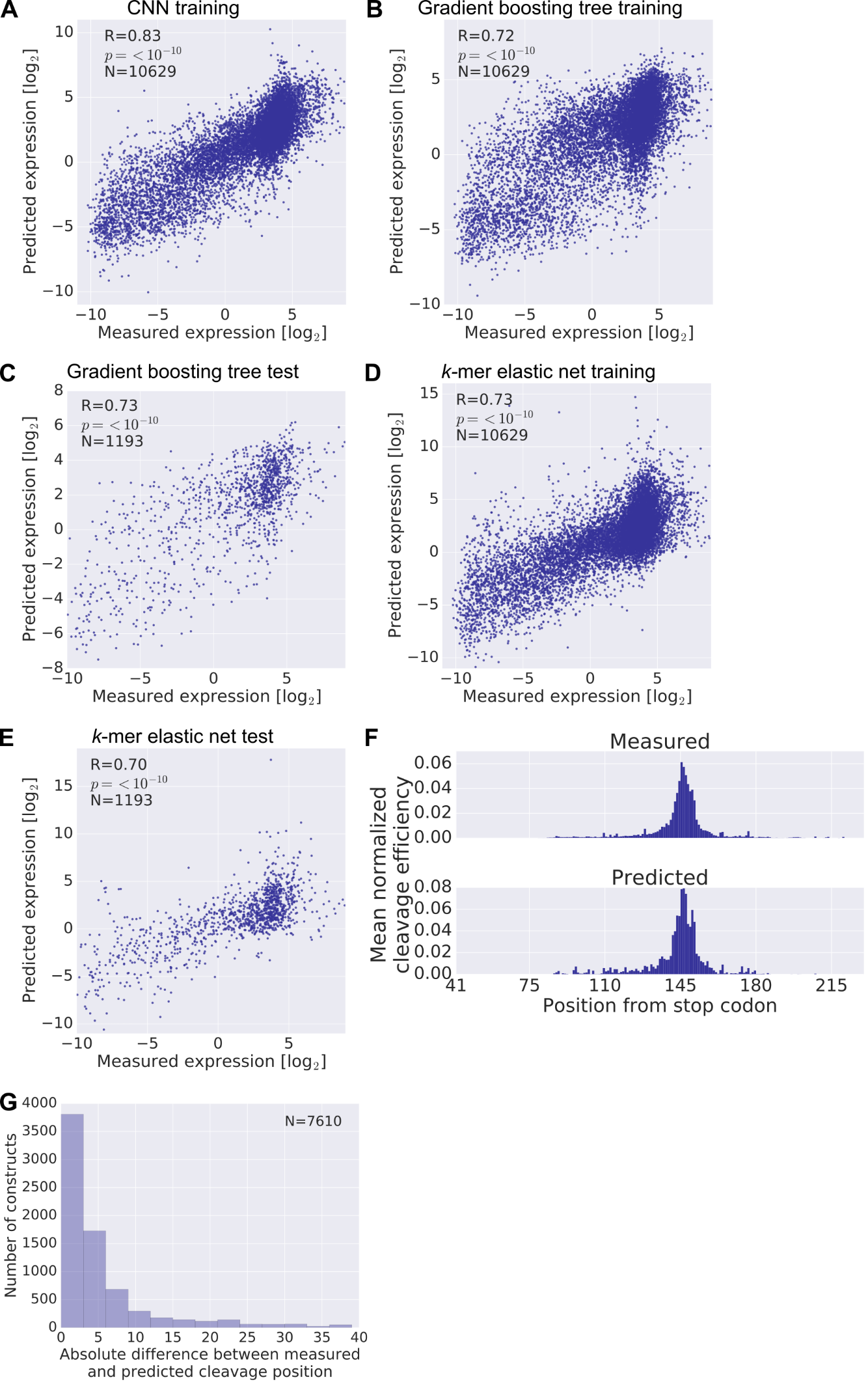
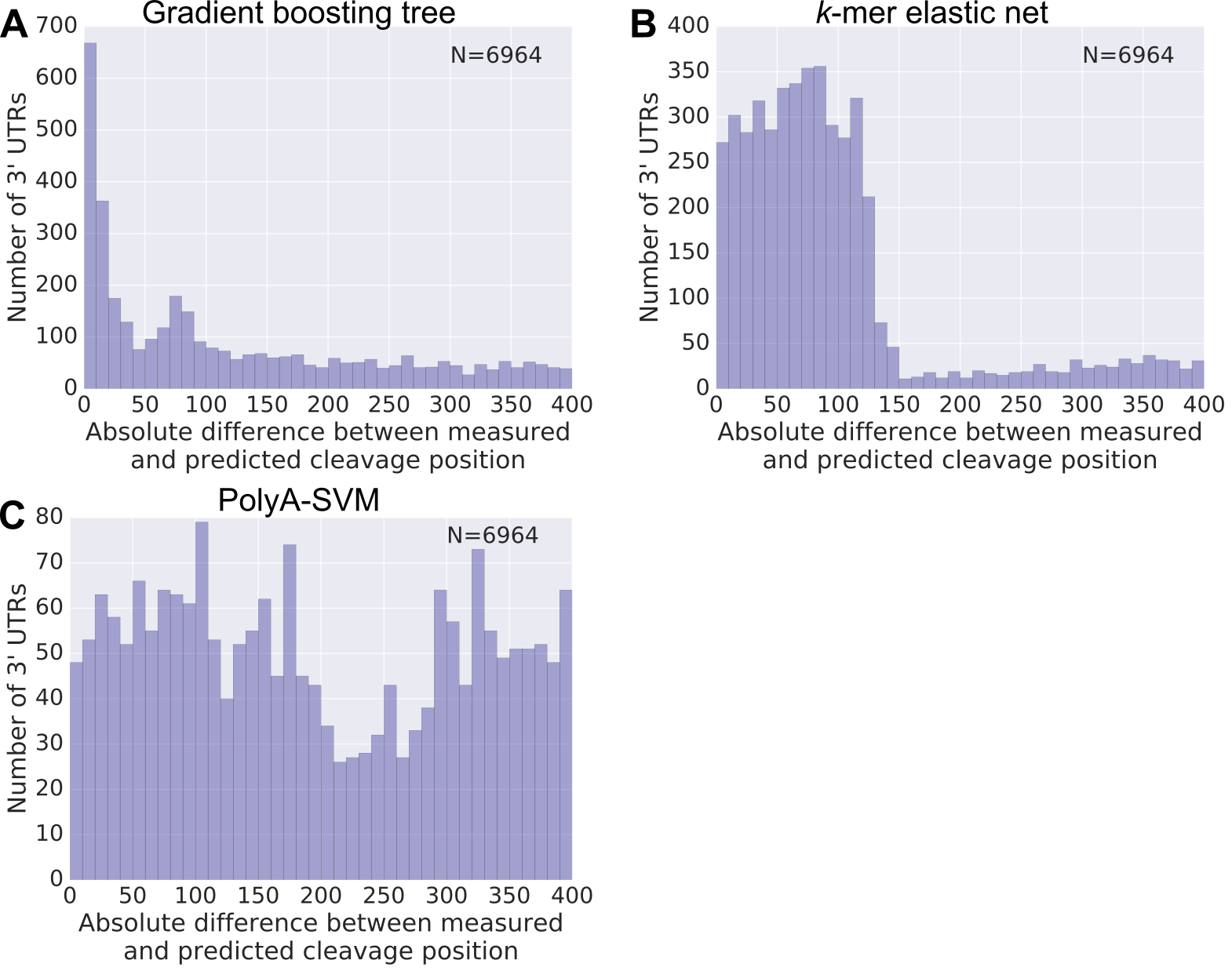


**Supplemental Figure S1 – quality control of expression and cleavage efficiency measurements.** (A) Scatter plot of expression measurements in two replicates. (B) Comparison of expression distribution across groups of constructs. Each group contains up to ten constructs differing only in the barcode sequence. Constructs span a wide expression range and show low median RSD (1.1%), estimated across all groups. (C-E) Cleavage efficiency of three representative variant groups, HIV1 (C), SPA1 (D) and SVL (E). Each group contains ten constructs differing only in the barcode sequence.



**Supplemental Figure S2 – positional preference of regulatory motifs.** (A-B) CentriMo analysis for the positional preference of upstream (A) and downstream (B) motifs found in the scanning mutagenesis data from the analysis performed in Figure 3E and Figure 3G respectively. (C-D) CentriMo analysis for the positional preference of center motifs found in the scanning mutagenesis data from the analysis performed in Figure 3F. Motifs were tested both upstream (C) and downstream (D). All of the presented results had a Fisher E-value < 10-9.



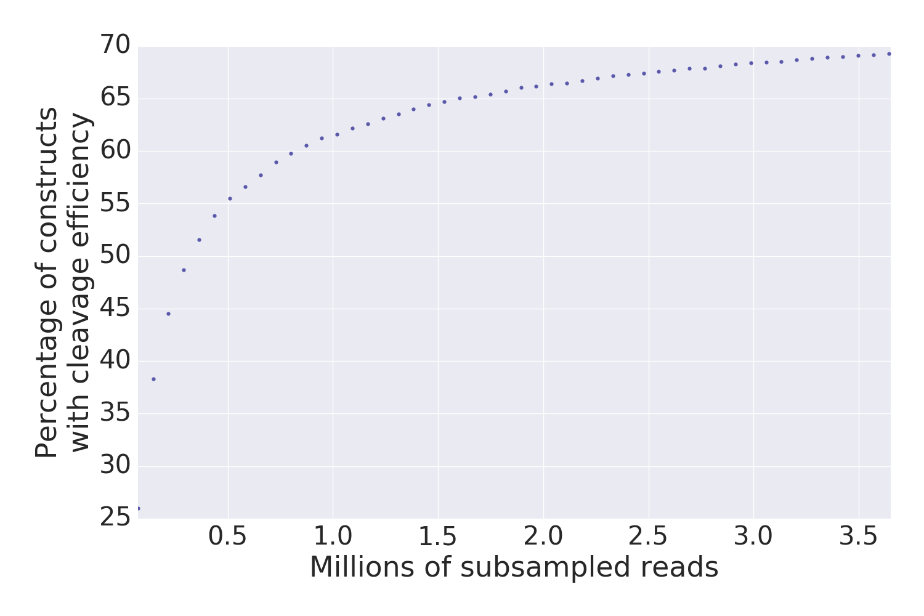
****Supplemental Note 1 – list of sequences excluded by design

**Supplemental Figure S4 – performance of alternative models for endogenous cleavage site prediction.** (A-B) Histogram of absolute differences between the measured and the predicted cleavage site on a set of endogenous 3' UTRs (Methods). Predictions were made by our gradient boosting tree (A) or kmer elastic net (B) models. The position at which the maximal expression was achieved was adjusted by 145, the most likely cleavage position within the reporter library. (C) The endogenous 3' UTR sequences were subjected to prediction using polyA-SVM. The position with the highest polyA score was selected as the predicted cleavage site and the histogram of absolute differences between the measured and predicted cleavage sites is presented.

**Supplemental Figure S3 – prediction results on training data and alternative model performance.** (A-E) Scatter plots of predicted vs measured expression levels for the CNN model training data (A), the gradient boosting tree training (B), the gradient boosting tree test (C), the kmer elastic net training (D) and kmer elastic net test (E) data. (F) Per position mean cleavage efficiency calculated over all the library members in the training set. Analysis was performed as in Fig. 4D. (G) Histogram of absolute differences between the measured and the most probable predicted cleavage site evaluated on library training data. Only constructs with measured cleavage efficiency maps were used.

The following list of short sequences was excluded by design from construct barcodes. This list was also used to avoid the introduction of unwanted sequence while performing scanning mutagenesis. The list is composed of restriction enzymes used for cloning, miRNA seeds of highly expressed miRNAs from K562 microarray data(Hershkovitz Rokah et al., 2012), the canonical hexamer and point mutants of the canonical hexamer. The list of sequences is: AAAAAA, AACAAA, AAGAAA, AATAAA, AATAAC, AATAAG, AATAAT, AATACA, AATAGA, AATATA, AATCAA, AATGAA, AATGCA, AATTAA, ACGGGT, ACGGTT, ACTAAA, ACTAGT, ACTGAC, ACTGCC, ACTGTG, ACTTGA, AGCTTT, AGTAAA, AGTCTA, ATCCAT, ATGTGA, ATTAAA, ATTATA, CACCTT, CACTAC, CACTTT, CAGGGT, CATAAA, CCAGAG, CTATGA, CTCTCC, CTGCAG, CTGTGA, CTTTAT, GAATGT, GATAAA, GCAGCT, GCATTA, GCTGCT, GGCGCGCC, GGTGCT, GTCCAG, GTTCTC, GTTTAC, TAAGCT, TACAGG, TACCTC, TATAAA, TCAGGG, TCCTCG, TGAGCC, TGCAAT, TGCACT, TGCCTT, TGCTGC, TGGGAG, TGTAGC, TGTCAT, TGTGGT, TTGCAC.

Supplemental Note 2 – effect of sequencing depth on cleavage detection

We acquire cleavage efficiency measurements for 69.3% of library members since our strict filtering criteria results in higher coverage requirement for detecting cleavage. Deeper sequencing would, in theory, allow for the calculation of cleavage efficiency for the more lowly expressed constructs in our library. However, in practice, the required amount of sequencing would be infeasible. We show this by subsampling the number of reads used for the cleavage efficiency calculation and finding only a relatively low increase in the percentage of constructs with cleavage efficiency data as a function of the number of subsampled reads at higher numbers of reads (Supplemental Fig. S5). Moreover, we note that even with only 69% of the library members with cleavage efficiency data, we have sufficient sample sizes for our hypothesis testing, motif discovery and prediction tasks.

**Supplemental Figure S5 – effect of sequencing depth on cleavage detection.** Scatter plot of the percentage of constructs with cleavage efficiency data as a function of millions of subsampled reads used for the cleavage efficiency calculation.

Supplemental Methods

**Cell culture**

K562 cells were acquired from ATCC. Cells were grown in Iscove's Modified Dulbecco Medium supplemented with 10% fetal bovine serum (Biological Industries) and 1% Penicillin-Streptomycin solution (Biological Industries). The cells were split when reaching a concentration of ~106 cells/ml. The cells were grown in an incubator at 37oC and 5% CO2. Cells were frozen in batches of 4×106 cells in growth medium supplemented with 5% DMSO. Cell lines were tested for mycoplasma contamination using EZ-PCR Mycoplasma Kit (Biological Industries).

**RNA purification and preparation for sequencing**

Each of the replicates was harvested for RNA purification, washed with PBS and split into tubes not exceeding 5×106 cells per tube. Tubes were flash frozen in liquid nitrogen and RNA was purified using NucleoSpin RNA II kit (MACHEREY-NAGEL) according to the manufacturer protocol. The purified RNA was treated with DNase (QIAGEN) in solution according to the manufacturer protocol and purified again using NucleoSpin RNA II kit (MACHEREY-NAGEL). We prepared cDNA for targeted 3' end sequencing by performing reverse transcription with the following primer: GCTCAAGCCACGACGCTCTTCCGATCTNANCNGNTNANCNGNTNANCNGNANCNANTTTTTTTTTTTTTTTTTTVN, where N is any nucleotide and V is any nucleotide except T. Reverse transcription was performed for each replicate using SuperScript® III First-Strand Synthesis System (Thermo Fisher Scientific) and 5 μg of input RNA per reaction according to the manufacturer protocol. For PCR 2 μl of the cDNA were used as template. Each amplification reaction contained 2 μl of the cDNA template, 25 μl of KAPA HiFi ready mix X2 (KAPA Biosystems), 2.5 μl of 10 μM 5′ primer, 2.5 μl of 10 μM 3′ primer and 18 μl ddw. The PCR program was 95◦C for 5 min, 20 cycles of 98◦C for 20s, 63◦C for 15s, and 72◦C for 30s, each, and one cycle of 72◦C for 5min. The forward primers used for this reaction were TATCGGATATGGGGTTCGGTATGCGC and GCAGTCGAATGGGGTTCGGTATGCGC for the first and second replicate respectively. The underlined sequenced serves as a sample barcode. The reverse primer was CACGACGCTCTTCCGATCT for both replicates. The PCR products were column purified (Qiagen).

Amplification of the library from the plasmid DNA used for the transient transfection was also performed with KAPA HiFi ready mix (KAPA Biosystems). A single amplification reaction included 1 μl of diluted plasmid template (20 ng), 12.5 μl of KAPA HiFi ready mix X2 (KAPA Biosystems), 0.75 μl of 10 μM 5′ primer, 0.75 μl of 10 μM 3′ primer and 10 μl ddw. The PCR program was 95◦C for 5 min, 14 cycles of 98◦C for 20s, 68◦C for 15s, and 72◦C for 30s, each, and one cycle of 72◦C for 2min. The primers used for the amplification were ATGGGGTTCGGTATGCGC (Fw primer) and ATCGTCTCGGGGAGCCTT (Rv primer). The PCR products were column purified (Qiagen).

The purified library amplification products from the RNA and plasmid DNA were assessed for size and purity at the TapeStation, using high sensitivity D1K screenTape (Agilent Technologies, Santa Clara, California). For each sample 20 ng library DNA were used for library preparation for second generation sequencing; specific Illumina adaptors were added, and DNA was amplified using 14 amplification cycles, protocol adopted from Blecher-Gonen *et al*.(Blecher-Gonen et al., 2013). The samples were reanalyzed using TapeStation.

Supplemental References

Blecher-Gonen, R, Barnett-Itzhaki, Z, Jaitin, D, Amann-Zalcenstein, D, Lara-Astiaso, D, and Amit, I (2013). High-throughput chromatin immunoprecipitation for genome-wide mapping of in vivo protein-DNA interactions and epigenomic states. Nat. Protoc. **8**, 539–554.

Hershkovitz Rokah, O, Granot, G, Ovcharenko, A, Modai, S, Pasmanik-Chor, M, Toren, A, Shomron, N, and Shpilberg, O (2012). Downregulation of Mir-31, Mir-155, and Mir-564 in Chronic Myeloid Leukemia Cells. PLoS ONE **7**, e35501.