PETRA: Prime editing of transcribed regulatory elements to assay expression

AUTHORS AND AFFILIATIONS

Magdalena Armas Reyes[1], Michael Herger[1], Laura Cubitt[1], and Gregory M. Findlay[1]

1. Genome Function Laboratory, The Francis Crick Institute, London, UK
Correspondence to Gregory M. Findlay (greg.findlay@crick.ac.uk)

ABSTRACT

Predicting how changes in human DNA sequence impact gene expression remains challenging. Here, we present PETRA (Prime Editing of Transcribed Regulatory elements to Assay expression), a multiplexed genome editing method to quantify the effects of regulatory variants at scale. PETRA leverages the delivery of variants to abundantly transcribed regions of genes such that sequence-specific effects on mRNA expression can be read out by amplicon sequencing. We demonstrate PETRA in Jurkat cells by scoring 13,935 six-nucleotide insertions delivered to the 5' untranslated regions (5' UTRs) of four genes important for T cell responses, namely *VAV1*, *IL2RA*, *CD28* and *OTUD7B*. Effects on expression are linked to the creation of new transcription factor binding sites (TFBSs), as well as to alterations in splicing and translation initiation. Combinatorial delivery of TFBSs identified using PETRA enables the discovery of alleles that increase mRNA expression more than 10-fold. Lastly, we extend PETRA to primary human T cells and compare variant effects across cell types. These experiments establish PETRA as a flexible means of dissecting the logic of gene regulation across genomic contexts and cell types.

KEY WORDS

multiplexed assays of variant effect (MAVEs), prime editing, gene regulation, transcription factor binding sites

INTRODUCTION

Identification of DNA sequence variants impacting human gene expression drives biological discovery and improves the interpretation of personal genomes. Beyond understanding which variants impact human phenotypes, knowledge of gene regulation can also be directly leveraged for therapeutic benefit. For example, genome editing of regulatory elements has emerged as a means of treating genetic disease[1]. While altering gene expression by editing the genome has a broad range of potential applications, such strategies are currently limited by our inability to accurately predict how specific changes in DNA sequence lead to changes in gene expression.

Computational models that predict expression from DNA sequence have improved substantially in recent years, but still perform poorly at predicting effects of specific variants[2–8]. One reason for this is that such models are hindered by a lack of high-quality training data. Large-scale omics studies have succeeded at associating common variants with changes in gene expression, chromatin states, and phenotypes[9–11]. However, such observational approaches are of limited benefit for inferring causal effects of new variants. Meanwhile, high-throughput experimental systems, such as massively parallel reporter assays (MPRAs), have enabled the interrogation of hundreds of thousands of regulatory elements and variants for expression effects[12,13]. Though valuable, MPRAs make use of artificial reporter cassettes that are expressed exogenously or via random genomic integration[14]. Thus, key features of variants' native loci are lost, such as surrounding DNA sequence and chromatin context. To date, few variants have been assayed for expression effects in their native genomic context, either in immortalised lines or primary cells.

While CRISPR-based methods have been established to assay expression effects of variants, all have limitations that hinder large-scale data generation. CRISPR interference (CRISPRi) and CRISPR activation (CRISPRa) technologies enable identification of noncoding elements of functional importance[15–18], but do not probe the effects of sequence variants. CRISPR and base editing screens[15,19] afford the ability to assay large libraries but are restricted to limited types of variants (i.e., indels or specific substitutions). Meanwhile, saturation genome editing (SGE) has proven clinically valuable for assaying coding variants

54 in cell lines suited to editing by homology-directed DNA repair[20], but has yet to show broad
55 utility for regulatory variants. Very recently, multiplex prime editing (PE) strategies have been
56 reported to assay variants for effects on expression[21,22]. However, thus far such methods
57 have only been applied to immortalised cell lines and scalability has been limited by the need
58 to develop a FACS-based read-out for each target gene. More broadly applicable, high-
59 throughput methods to experimentally test regulatory variants in their native context will prove
60 highly valuable.
61
62 Here, we introduce Prime Editing of Transcribed Regions to Assay expression, or PETRA.
63 We demonstrate PETRA by assaying a total of 14,888 variants across four genes with roles
64 in human T cell responses, successfully identifying individual edits, combinations of edits,
65 and transcription factor binding sites (TFBSs) that increase messenger RNA (mRNA)
66 expression by greater than 10-fold in Jurkat and primary human T cells. We anticipate
67 application of PETRA to diverse loci and cell types will refine our understanding of gene
68 expression, facilitate rapid accumulation of training data for computational models, and speed
69 the discovery of edits that confer therapeutic benefit.

RESULTS

Prime Editing of Transcribed Regulatory Elements to Assay expression

With the aim of measuring endogenously engineered variants' effects on gene expression in high-throughput, we developed PETRA. PETRA leverages prime editing[23] to install variants in non-coding regions of target genes and next-generation sequencing (NGS) to readout their effects on expression (**Figure 1A**). We reasoned that by editing abundantly transcribed regions, regulatory effects of variants could be determined in multiplex simply by performing amplicon sequencing of RNA and DNA from edited cells. In turn, this would allow individual variants and combinations of variants to be read out in an allele-specific manner, potentially across each of many target sites and cell types.

To demonstrate PETRA in Jurkat T cells, we first analysed effects of variants on *VAV1* expression. Upregulation of *VAV1* using CRISPRa has been shown to enhance T cell effector function in multiple assays[24]. Here, we tested whether increased *VAV1* mRNA expression could be achieved by introducing edits endogenously. To this end, we cloned a library of pegRNAs designed to install every possible 6-bp insertion (n = 4,096) at a single site in the gene's 5' untranslated region (5' UTR) immediately downstream of the *VAV1* transcription start site (TSS; **Figure 1B**, **Supplementary Figure 1A**). A pegRNA architecture was selected on the basis of predicted editing efficiency[25] (see **Methods**). The pegRNA library was delivered to Jurkats via nucleofection with mRNAs encoding PE7[26] and MLH1dn[27]. Seven days post nucleofection, cells were harvested and the target locus was amplified and sequenced from genomic DNA (gDNA) and mRNA. For each variant, an expression score was computed as the log2-ratio of variant frequency in mRNA over variant frequency in gDNA, normalised to the median.

In this experiment, 6-bp insertions were installed with a combined editing efficiency of 50.4% (**Supplementary Figure 1B**), allowing successful scoring of 3,902 variants (95.3% of the library) after removal of variants seen at low gDNA frequencies (less than 1.0×10^{-5}) (**Supplementary Table 1**). Across three replicate transfections, variant frequencies in gDNA and mRNA and variant expression scores were highly correlated (**Figure 1C-E**,

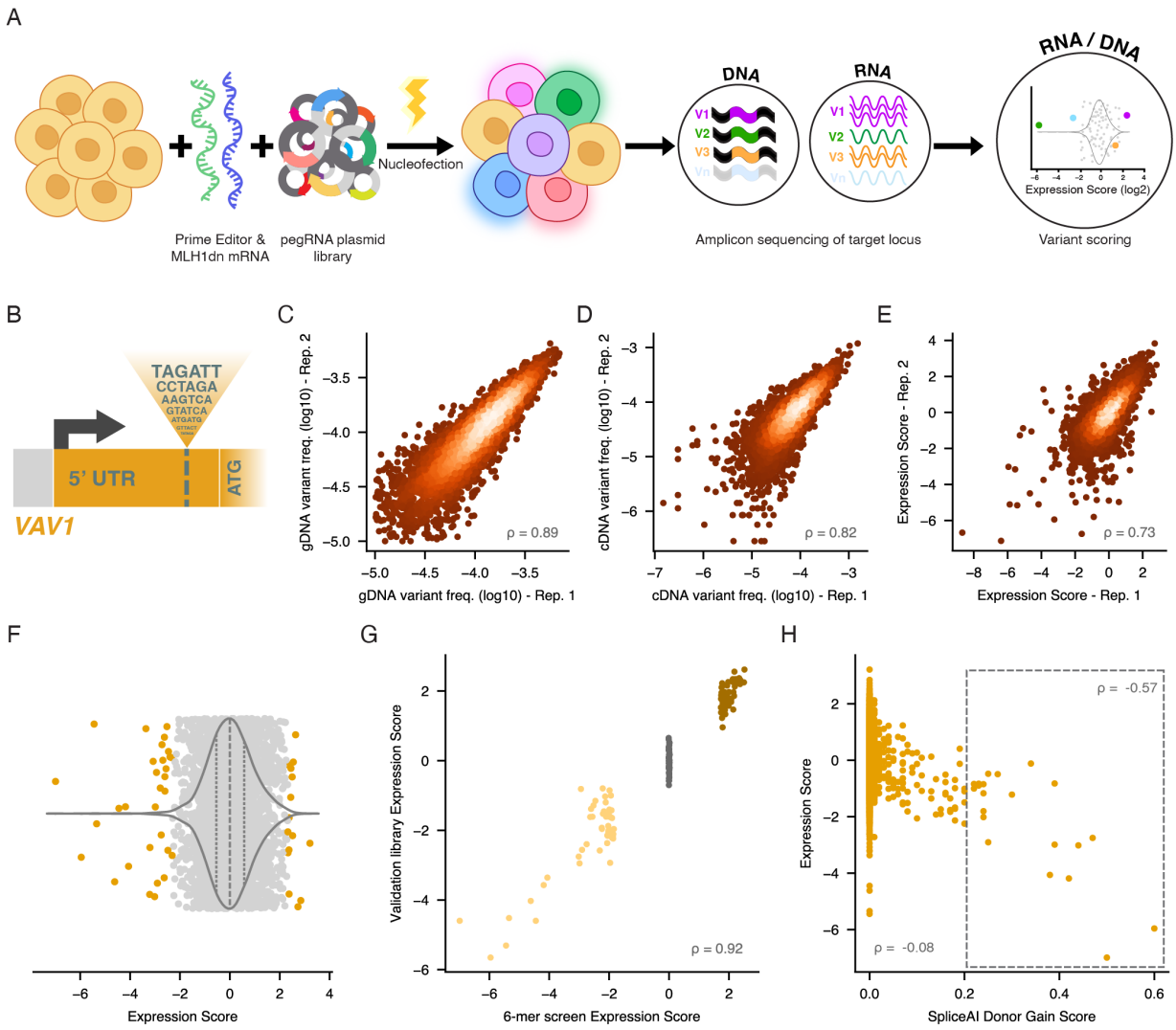


Figure 1. Testing effects of regulatory variants on *VAV1* expression with PETRA.
A. Schematic of the PETRA workflow. Prime editor and MLH1dn mRNA are delivered to cells with the pegRNA plasmid library via nucleofection. Library variants are installed in transcribed regions of the target gene. Seven days post-nucleofection, amplicon sequencing of gDNA and mRNA is performed to quantify variant frequencies. An expression score is computed for each variant as the normalised log2-ratio of mRNA variant frequency over gDNA variant frequency.
B. Schematic of the random 6-bp insertion library (n = 4,096) targeting the 5' UTR of *VAV1*.
C. Correlation of gDNA variant frequencies in replicates 1 and 2 for n = 3,902 insertions post-filtering (Spearman's ρ). Correlations across all replicates are in Supplementary Figure 1C.
D. Correlation of mRNA variant frequencies in replicates 1 and 2 (n = 3,902).
E. Correlation of variant expression scores (log2-scaled) in replicates 1 and 2 (n = 3,902).
F. Distribution of expression scores (n = 3,902) averaged across three biological replicates and normalised to the median. (Violin plot: dashed line, median; dotted lines, 25th and 75th percentiles, all points shown with orange indicate $p < 0.01$ under a normal distribution.
G. Correlation of expression scores for n = 150 selected insertions from the *VAV1* 6-mer screen re-tested using an independent validation library. (Dark brown, high-scoring in initial screen; grey, neutral-scoring; yellow, low-scoring.)
H. Expression scores versus SpliceAI donor gain scores for n = 3,902 insertions. Spearman's ρ is shown for all insertions and for only insertions with SpliceAI scores over 0.20 (dotted box).

Supplementary Figure 1C). We defined 46 insertions with significant effects on *VAV1* expression ($p < 0.01$), including insertions that decreased expression to less than 20% of baseline and insertions that increased expression by more than 5.0-fold (**Figure 1F**). A subset of 150 variants that scored either highly, lowly, or neutrally were re-tested in an independent library (**Supplementary Table 2**). Expression scores from this validation experiment were strongly correlated to those from the larger experiment (Spearman's $\rho = 0.92$; **Figure 1G, Supplementary Figure 1D**), confirming PETRA's reproducibility.

Variants introduced to the 5' UTR may impact mRNA levels through a variety of mechanisms, including sequence-specific effects on transcription initiation, transcript stability, and pre-mRNA splicing. Reasoning that variants disruptive of *VAV1* splicing may be underrepresented in mRNA, we scored insertions for their predicted splice effects using SpliceAI[3] (**Figure 1H; Supplementary Table 1**). While most insertions had no predicted impact on splicing, several variants predicted to create new donor splice sites had low expression scores, consistent with aberrant splicing leading to variant depletion in mRNA. Considering only insertions with SpliceAI scores greater than 0.20, we observed a correlation between SpliceAI scores and expression scores of $\rho = -0.57$, confirming expression scores reflect biological effects of variants on mRNA levels.

Scoring 13,935 variants in total across four genes

We next extended PETRA to three additional genes: *IL2RA*, *OTUD7B* and *CD28*. Increased expression of each of these genes has been reported to enhance T cell effector function[24,28]. We designed PETRA libraries once more comprising all possible 6-bp sequences to be inserted at a single site downstream of each gene's TSS and achieved insertion efficiencies between 31.3% and 70.3% (**Supplementary Figure 1B, E-G**). As for *VAV1*, we removed insertions not installed efficiently and assigned expression scores to the remaining 10,033 variants, increasing the total number of variants scored across the four genes to 13,935 (85.1% of library variants; **Figure 2A, Supplementary Table 1**). Once more, expression scores were correlated across replicates **(Supplementary Figure 2A**), and independently assayed subsets of variants validated highly and lowly scoring insertions from the larger

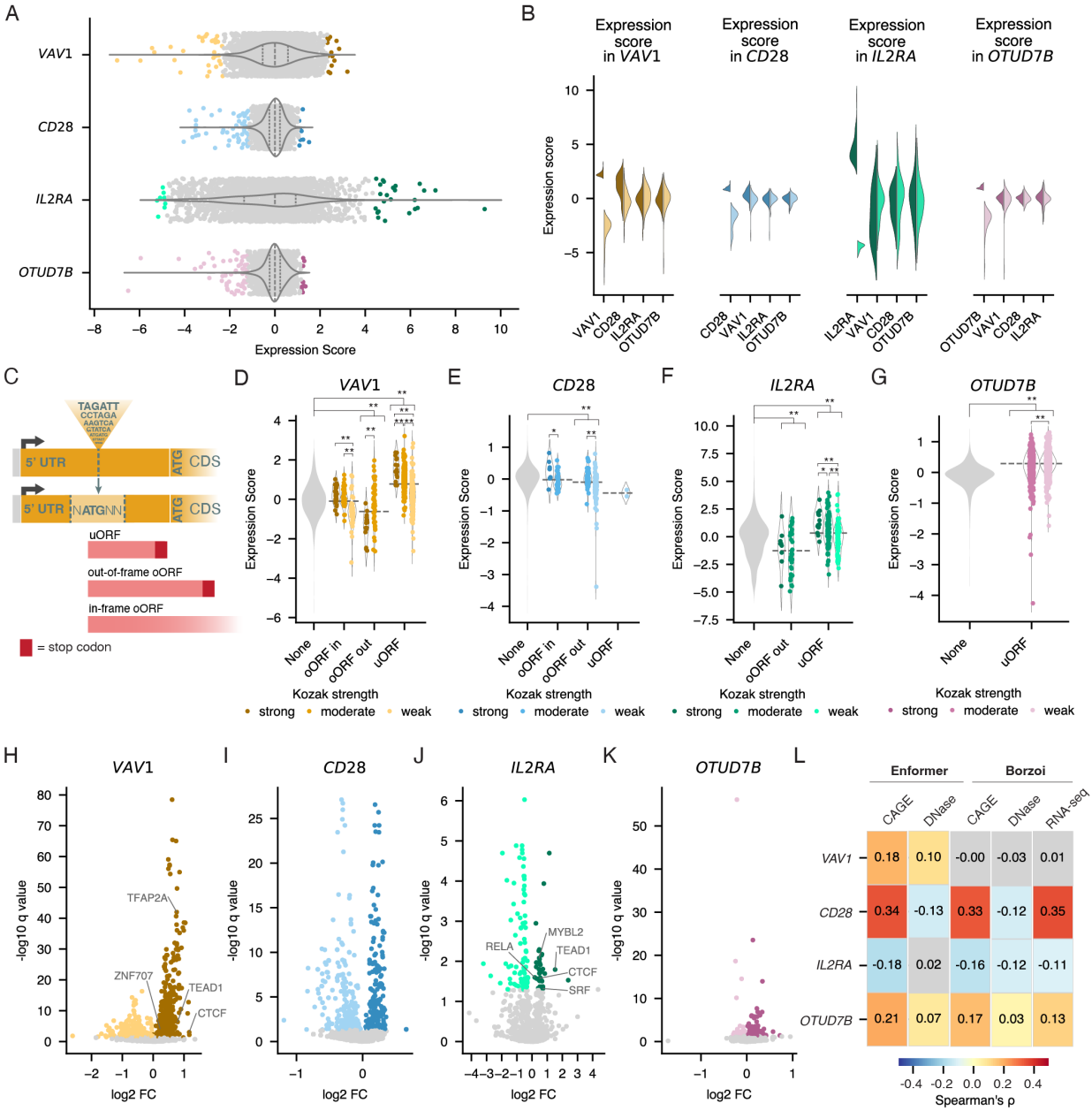


Figure 2. Exploring expression effects of 13,935 non-coding variants across four genes.
A. Distribution of expression scores for 6-bp insertions targeting the 5' UTR of *VAV1* (n = 3,902), *CD28* (n = 3,930), *IL2RA* (n = 2,419), and *OTUD7B* (n = 3,684), averaged across biological replicates (n = 3 for *VAV1* and *CD28*; n = 2 for *OTUD7B* and *IL2RA*) and median-normalised. (Violin plot: dashed line, median; dotted lines, 25th and 75th percentiles, coloured points represent standard normal distribution test $p < 0.01$).
B. For the top 50 (dark) and bottom 50 (light) insertions scored at each locus (x-axis), the split violin plots show expression scores of the same insertion sequences introduced to each gene (panels). Insertion sequences are defined in the 5' to 3' orientation of the sense strand.
C. Schematic of new ORF classification. uORF = upstream ORF, oORF = overlapping ORF.
D-G. Violin plots of expression scores per gene, grouping variants by ORF type and Kozak sequence strength. Horizontal dotted lines indicate median by ORF type. (* = $q < 0.05$ and ** = $q < 0.01$ for Mann-Whitney U-test with BH correction; oORF in = in-frame oORF, oORF out = out-of-frame oORF)

H-K. For each TF expressed in Jurkats, a log2-fold-change (log2FC) was calculated as the difference in median expression score between insertions creating a FIMO-predicted binding site for the TF and other insertions at the site. Volcano plots show the log2FC for each TF motif with q-values from a two-sample Kolmogorov-Smirnov test with BH correction. Colours indicate $q < 0.05$ for TF motifs associated with increased (dark colours) or decreased (light colours) expression. TF motifs used to design insertions in subsequent experiments are labelled.
L. Correlations (Spearman's ρ) between PETRA expression scores and Jurkat-specific predictions from Enformer and Borzoi for CAGE, DNase and RNA-seq assays. Coloured (non-grey) squares indicate significant correlations ($p < 0.05$).

experiments (**Supplementary Figure 2B-E, Supplementary Table 2**). Concordance in the validation experiment was weakest for *IL2RA*. This result is consistent with *IL2RA* transcript levels being low in Jurkats at baseline (**Supplementary Figure 2F**), and not inefficient prime editing, as *IL2RA* insertions were installed with high efficiency (**Supplementary Figure 1B**).

Across the four genes, 204 variants significantly modulated expression ($p < 0.01$), with 57 increasing expression and 147 decreasing expression (**Figure 2A**). Exploring whether the same 6-bp insertion sequences influence expression similarly across loci, we found most effects to be highly context-dependent. For instance, the 50 highest and lowest scoring 6-bp sequences inserted at each site tended to score neutrally at each of the other loci (**Figure 2B**). This high degree of locus specificity may be attributable to differences in adjacent sequence context or interactions with nearby regulatory elements.

Exploring mechanisms by which sequences impact mRNA levels

Towards explaining sequence-specific effects on expression, we first examined splicing. Splicing effects were predicted for insertions at each locus using SpliceAI (**Supplementary Figure 2G-J**). As for *VAV1*, higher SpliceAI scores predicted lower expression scores for insertions in *CD28*. Several variants strongly predicted to impact *OTUD7B* splicing had very low expression scores, but overall the correlation was insignificant ($\rho = -0.02$, $p = 0.2$). No *IL2RA* insertions had SpliceAI scores above 0.04. Across all four genes, the majority of significantly scored variants had SpliceAI scores less than 0.04, suggesting other mechanisms explain their effects.

New open reading frames (ORFs) created in the 5' UTR can lower protein expression by impacting ribosome loading and translation, yet how new ORFs impact mRNA levels is less

207 clear[29,30]. To explore this, all 2,031 insertions that created new 5'-ATG sequences were
208 classified by Kozak sequence strength and the type of ORF created. Upstream ORFs
209 (uORFs) are ORFs that end upstream of the original start codon, whereas overlapping ORFs
210 (oORFs) extend into the original ORF and can be either in-frame or out-of-frame (**Figure 2C**).
211 Excluding variants with SpliceAI score greater than 0.1, variants creating uORFs had
212 significantly higher expression scores than variants not creating new ORFs across loci (Mann-
213 Whitney U test with Benjamini-Hochberg (BH) correction for *VAV1*: $q = 2.50\times10^{-45}$; *IL2RA*: q
214 $= 2.79\times10^{-4}$; *OTUD7B*: $q = 2.18\times10^{-147}$), with the exception of *CD28*, where only three uORF
215 variants were tested (**Figure 2D-G; Supplementary Table 3**). This effect was dependent on
216 Kozak sequence strength, with insertions creating strong Kozak sequences in *VAV1* and
217 *IL2RA* scoring highest. Meanwhile, insertions creating out-of-frame oORFs scored more lowly
218 than insertions not introducing 5'-ATG sequences (Mann-Whitney U test with BH correction
219 for *VAV1*: $q = 3.10\times10^{-5}$; *CD28*: $q = 2.19\times10^{-6}$; *IL2RA*: $q = 1.23\times10^{-3}$). No difference was
220 observed for variants creating in-frame oORFs, indicating the negative effect for out-of-frame
221 oORFs is likely attributable to nonsense-mediated decay. While this analysis shows that new
222 ORFs can influence mRNA expression, only 38 of 204 (18.6%) variants scored as hits created
223 new ORFs.
224
225 We next characterised how the creation or disruption of transcription factor binding sites
226 (TFBSs) may impact expression. For this analysis, we predicted the binding of each
227 transcription factor (TF) expressed in Jurkats[31] to each variant allele using FIMO[32] and the
228 JASPAR database[33], once more excluding variants with SpliceAI scores greater than 0.10.
229 Across loci, insertions were not predicted to disrupt pre-existing TFBSs (**Supplementary**
230 **Figure 1A,E-G**).
231
232 To examine effects of newly created TFBSs, we asked whether expression scores differed
233 for variants predicted to create a binding site for a given factor compared to all other variants
234 at the site. This revealed 555 motifs that positively impacted expression of at least one gene
235 and 567 motifs that negatively impacted expression (**Figure 2H-K, Supplementary Table 4**).
236 Interestingly, no motif had a consistently significant effect across the four genes, and very
237 few motifs were consistent across three genes (**Supplementary Figure 2K-L**).
238

239 Computational models for predicting RNA expression from DNA sequence have improved
240 recently yet still perform suboptimally for predicting variant effects[7,34]. To assess whether
241 computational models accurately predict the effects observed in our experiments, we
242 compared Jurkat-specific predictions of cap analysis of gene expression (CAGE), DNase
243 hypersensitivity, and RNA-seq to PETRA expression scores. We found predictive
244 performance to vary considerably across loci and output tracks (**Figure 2L, Supplementary**
245 **Figure 2M**). CAGE predictions from Enformer and Borzoi and RNA-seq predictions from
246 Borzoi were most strongly correlated with *CD28* expression scores, with a maximum
247 correlation for Borzoi RNA-seq predictions ($\rho = 0.35$; $p = 4.8\times10^{-114}$). Weaker but significant
248 Enformer correlations were also observed for *OTUD7B* (CAGE: $\rho = 0.21$; $p = 8.4\times10^{-37}$) and
249 *VAV1* (CAGE: $\rho = 0.18$; $p = 1.9\times10^{-27}$). Meanwhile, predictions for *IL2RA* were anti-correlated
250 with expression scores (Enformer CAGE: $\rho = -0.18$; $p = 1.3\times10^{-19}$, Borzoi CAGE: $\rho = -0.16$;
251 $p = 3.5\times10^{-16}$, and Borzoi RNA-seq: $\rho = -0.11$, $p = 9.0\times10^{-8}$). Expression scores, which are
252 derived from mRNA levels, were generally more weakly correlated to DNase predictions
253 across genes. Overall, these results demonstrate the limited ability of computational models
254 to predict expression effects of human variants installed via prime editing.
255
256 **Single and combinatorial delivery of TFBSs to increase expression**
257
258 We next set out to explore positional and combinatorial effects of insertions, with the goal of
259 inducing larger changes in expression. For this purpose, we designed pegRNAs to introduce
260 the MYBL2 motif to each of three loci in the *IL2RA* 5' UTR, as 6-bp sequences that created
261 MYBL2 binding sites were associated with increased *IL2RA* expression (**Figure 2J**). We also
262 designed three additional pegRNAs to insert the EGR1 motif at the same loci, which was not
263 associated with altered *IL2RA* expression. PETRA was then performed using a pool of these
264 six pegRNAs, allowing a set of 27 alleles to be generated and scored in a single transfection
265 (**Figure 3A**).
266
267 MYBL2 binding sites individually increased expression to varying degrees (**Figure 3A,**
268 **Supplementary Table 5**). Insertion at site 2 led to the highest increase among single inserts
269 (expression score = 2.69), followed by insertion at site 1 (expression score = 1.37), validating
270 the motif's predicted effect. Intriguingly, combinations of MYBL2 insertions showed largely

271 additive effects on log2-scaled expression scores. For example, the expression score of the
272 MYBL2 motif double insertion at sites 1 and 2 was 4.57, and the score of the MYBL2 motif
273 triple-insertion was 5.48, corresponding to a 44-fold increase in mRNA expression compared
274 to unedited alleles. Meanwhile, single EGR1 insertions either had little effect or decreased
275 expression, while multiple EGR1 insertions more strongly reduced expression. For example,
276 the EGR1 triple insertion led to a 40-fold reduction in mRNA (**Figure 3A**). The additive nature
277 of expression effects was corroborated by introducing sets of MYBL2 or EGR1 motifs
278 individually (**Supplementary Figure 3A-B**).
279
280 We next explored whether multiple TF motifs inserted via single pegRNAs could be leveraged
281 to increase expression. We selected the MYBL2, SRF, and RELA motifs associated with
282 increased *IL2RA* expression and the TFAP2A and ZNF707 motifs associated with increased
283 *VAV1* expression (**Figure 2H,J**). We also selected the TEAD1 and CTCF motifs, which were
284 associated with higher expression of both genes. We then designed pegRNA libraries
285 targeting the same sites in *IL2RA* and *VAV1* to which random 6-bp sequences were inserted,
286 this time delivering up to three copies of each motif in tandem (**Figure 3B-C**). For each locus,
287 we also included 6-bp insertions predicted to bind each TF that had scored highly before, as
288 well as additional edits comprising positive and negative controls. Final libraries included 59
289 pegRNAs for *VAV1* and 77 pegRNAs for *IL2RA*. Editing rates were sufficient to score all
290 variants, though larger insertions were created less frequently, consistent with previous
291 studies[35] (**Supplementary Figure 3C-J**).
292
293 Expression scores revealed that most but not all TF motif insertions increased expression of
294 the genes for which they were selected (**Figure 3B-C; Supplementary Tables 6-7**). The
295 highest scoring insertion at the *IL2RA* locus was the double insertion of the MYBL2 motif,
296 which increased expression over 9-fold (**Figure 3B**). For the RELA and SRF motifs, *IL2RA*
297 expression was significantly increased by the 6-bp insertions identified in screening, but not
298 by the designed motif insertions, with the exception of double RELA motif insertion, which
299 increased expression 1.5-fold. All nine TFAP2A, ZNF707, and TEAD1 motif insertions
300 increased *VAV1* expression (**Figure 3C**), but not to a greater extent than the top-scoring 6-
301 bp sequences for each motif. While TEAD1 motif-containing sequences failed to increase
302 *IL2RA* expression, triple insertion of the TFAP2A motif (identified as a hit in the *VAV1*

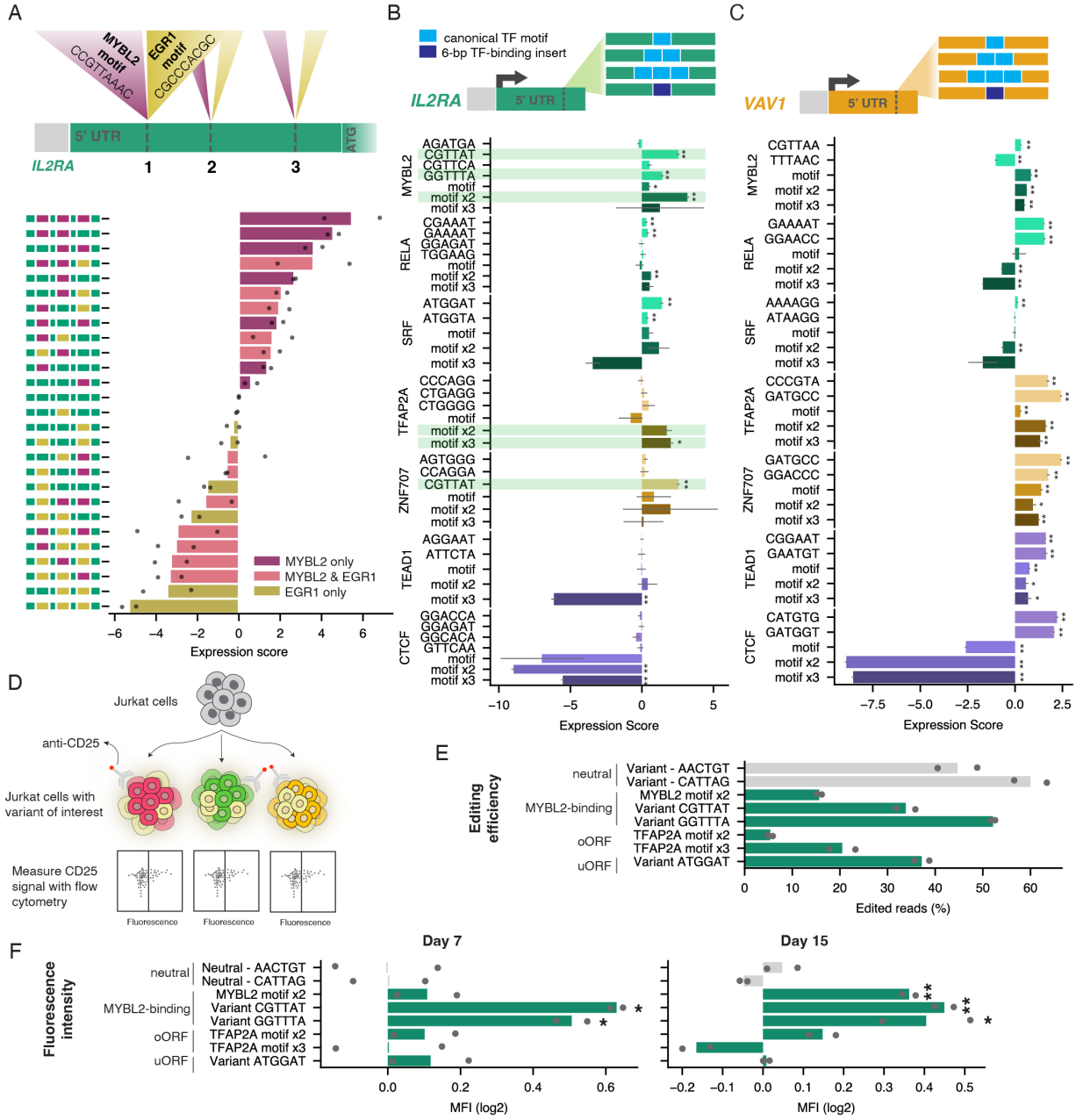


Figure 3. TF motifs identified with PETRA modulate RNA and protein expression.
A. Schematic shows insertions of either the MYBL2 or EGR1 motif at each of three sites in the 5' UTR of *IL2RA*. Six pegRNAs programming the insertions were co-delivered to cells to achieve combinatorial editing. The barplot shows PETRA expression scores for each of 27 alleles observed following editing, normalised to unedited (insertions depicted at left). Points represent independent replicates and bars represent the average across replicates.
B-C. Schematics show pegRNA libraries designed for *IL2RA* (B) and *VAV1* (C) to deliver select TF motifs individually or in tandem to a single site within each gene. High-scoring 6-bp sequences predicted to bind each TF were included, as well as controls. Expression scores, normalised to control insertions, are plotted below, grouped by TF motif. Error bars represent standard error across three replicates. Colors indicate the reason for TF motif inclusion (green shades: hit in *IL2RA* 6-bp screen; yellow shades: hit in *VAV1* 6-bp screen; purple shades: hit in both screens. Variants individually tested for protein expression are highlighted in green. Six-bp insertions predicted to create multiple TF motifs are plotted under each

corresponding motif. (* = $q < 0.05$ and ** = $q < 0.01$ for Welsch's unequal variances two-sample t-test against neutral variants with BH correction)
D. Schematic for assessing variant effects on protein expression via flow cytometry. Delivery of individual pegRNAs resulted in mixed populations of cells with one or more edited allele or no edits.
E. Average editing efficiency of insertions tested individually in Jurkats, with dots indicating individual replicates (light grey: controls; green: high expression scores).
F. Cell populations transfected with individual pegRNAs were activated with PMA for 24 h, anti-CD25 stained, and assessed by flow cytometry at day 7 (left) and day 15 (right) post-editing. Barplots represent mean fluorescence intensity (MFI) for CD25 normalised to neutral variants. (* = $q < 0.05$, ** = $q < 0.01$, two-sample Student's t-test versus neutral controls with BH correction)

screen) led to more than a 3-fold increase. In contrast to 6-bp insertions predicted to bind CTCF, insertion of either one or multiple 13-bp CTCF motifs strongly decreased expression at both loci. We repeated the experiment for *IL2RA* in cells activated with 100 ng/ml PMA and found expression scores to correlate reasonably well to those from resting Jurkats (**Supplementary Figure 3K**). In summary, these experiments demonstrate that delivery of multiple TF motifs via single pegRNAs can drive large changes in expression.

Testing effects of insertions on protein expression

Variants that alter mRNA expression may not necessarily alter protein expression. Therefore, we assessed a panel of six variants with high *IL2RA* expression scores for effects on CD25 levels using flow cytometry. Specifically, we chose three variants predicted to introduce MYBL2 binding sites as well as three variants that create new ORFs and thus may interfere with translation. We also included two 6-bp insertions that scored neutrally in our large screen as controls.

We transfected pegRNAs for each insertion to cells individually, and CD25 expression was then measured by flow cytometry on days 7 and 15 after activating cells for 24 hours with PMA (**Figure 3D**). Editing efficiencies, determined via amplicon sequencing on day 7, were less than 60% for all samples (**Figure 3E**), meaning both edited and unedited alleles contributed to total CD25 expression in each population. Nonetheless, we observed increased mean fluorescence intensities for each of the three samples to which sequences predicted to bind MYBL2 were introduced (**Figure 3F, Supplementary Table 8**). The more efficiently installed 6-bp insertions significantly increased expression at both day 7 and day 15, whereas the double MYBL2 motif insertion showed a significant increase only on day 15.

356 In contrast, the three insertions that create new ORFs did not significantly change CD25
357 protein expression despite having scored highly for mRNA expression (**Figure 3F**). The same
358 trends were observed for the percentage of IL2RA^{+} cells in each sample (**Supplementary**
359 **Figure 3L**), confirming PETRA enables identification of edits that increase expression of
360 target proteins.
361
362 **PETRA to test variants in primary human T cells**
363
364 For many potential applications, it may prove valuable to identify edits with desired effects on
365 expression in primary human cells. Towards this goal, we performed PETRA in human CD3+
366 T cells obtained from healthy donors (**Figure 4A**). Total CD3+ T cells were activated with
367 anti-CD3, anti-CD28, and anti-CD2 antibodies on day 0 and day 14 following isolation. Three
368 days post re-activation, cells were nucleofected with mRNA encoding PEMax[27] and MLH1dn,
369 as well as the pegRNA library to install all possible 6-bp insertions in the *IL2RA* 5' UTR. On
370 day 7, average editing was 12.3% across biological replicates, allowing expression scores to
371 be derived for 3,143 insertions (**Supplementary Figure 4A-C, Supplementary Table 9**).
372
373 Expression scores from primary T cells were only modestly correlated to Jurkat-derived
374 expression scores ($\rho = 0.22$, $p = 2.35 \times 10^{-25}$; **Figure 4B**), yet TF motif analysis identified
375 many motifs associated with increased expression in both primary T cells and Jurkats,
376 including MYBL2, RELA, TEAD1, and CTCF (**Figure 4C, Supplementary Table 10**). As in
377 Jurkats, uORFs were associated with higher expression scores, whereas out-of-frame
378 oORFs were associated with lower expression scores (**Supplementary Figure 4D;**
379 **Supplementary Table 3**). Once more, expression scores were anti-correlated to CAGE,
380 DNase and RNA-seq predictions from Enformer and Borzoi (**Supplementary Figure 4E**).
381
382 As several TF motifs were associated with increased *IL2RA* expression in both Jurkats and
383 primary cells, we additionally assayed the smaller pegRNA libraries designed to install
384 specific TF motifs at *IL2RA* and *VAV1* in primary T cells. For *IL2RA* insertions, expression
385 scores were modestly correlated between Jurkats and primary T cells ($\rho = 0.39$, $p = 4.84 \times$
386 10^{-4}; **Figure 4D, Supplementary Figure 4F; Supplementary Table 11**). The correlation
387

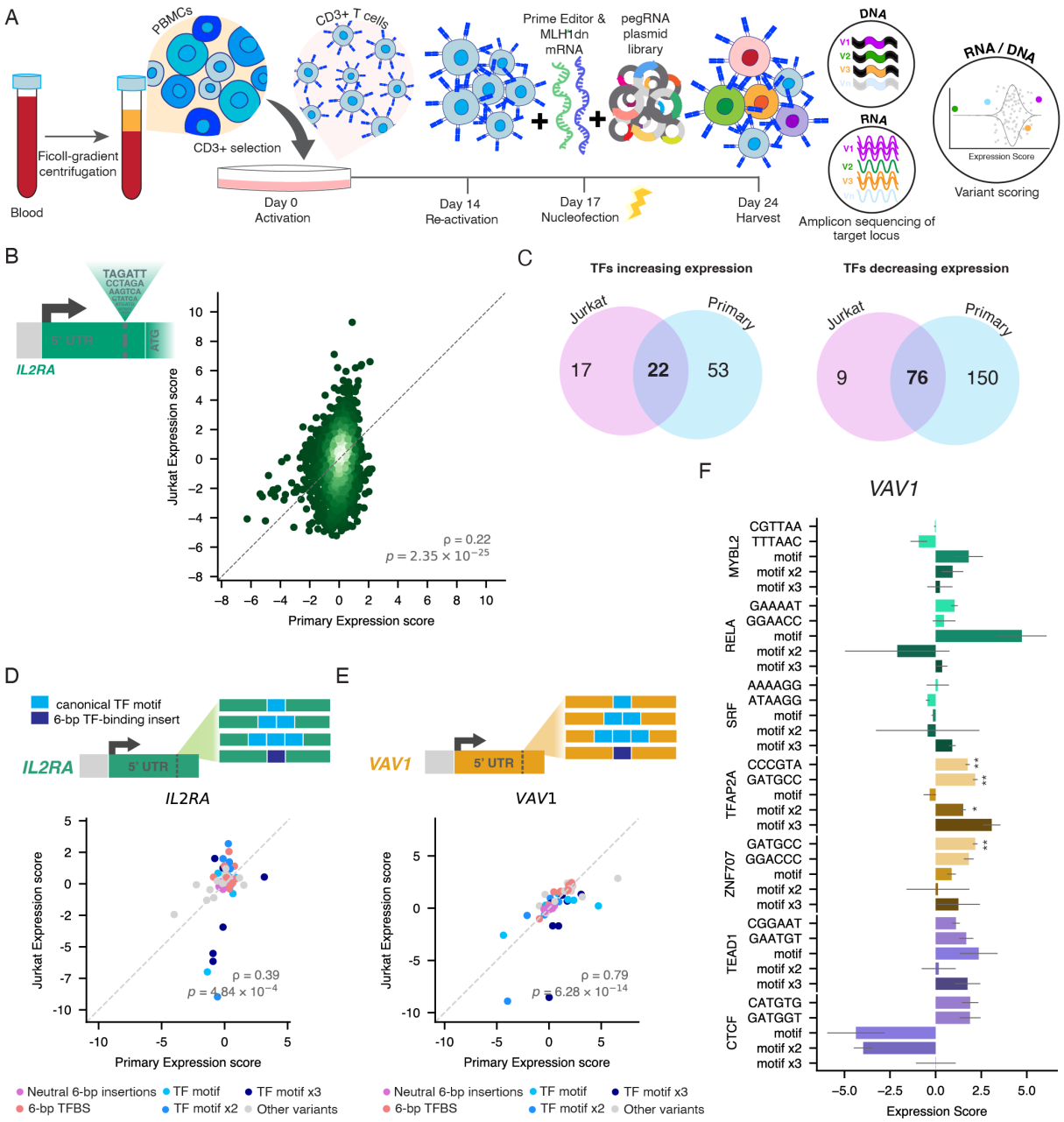


Figure 4. Testing effects of regulatory variants in human primary T cells with PETRA.
A. CD3+ primary T cells selected from donor PBMCs are activated with anti-CD2, anti-CD3, and anti-CD28 on day 0 and re-activated on day 14. Three days after re-activation, prime editing is performed to deliver a library of variants scored for effects on mRNA expression.
B. Correlation (Spearman's ρ) of expression scores for n = 2,269 *IL2RA* 6-bp insertions assayed in Jurkat and primary human T cells.
C. Venn diagrams showing the overlap of TF motifs associated with increased or decreased *IL2RA* expression scores in Jurkat and primary T cells.
D-E. The same pegRNA libraries designed to deliver TF motifs individually or in tandem to Jurkat cells (Figure 3B-C) were assayed in primary T cells. Correlations (Spearman's ρ) between expression scores

399 from Jurkat and primary T cells are plotted for *n* = 77 *IL2RA* variants (D) and *n* = 59 *VAV1* variants (E),
400 coloured by insertion type.
401 **F.** Expression scores, normalised to neutral insertions, are plotted for *VAV1* TF motif insertions assayed
402 in CD3+ primary human T cells. Error bars represent standard error across three biological replicates. (*
403 = $q < 0.05$ and ** = $q < 0.01$ for Welsch's unequal variances two-sample t-test with BH correction)
404
405
406
407 was similar when Jurkat expression scores were determined after activation with PMA (ρ =
408 0.36, p = 1.11×10^{-3}; **Supplementary Figure 4G; Supplementary Table 7**). Interestingly,
409 *VAV1* expression scores were more highly correlated (ρ = 0.79, p = 6.28×10^{-14}; **Figure 4E;**
410 **Supplementary Table 12**), with many of the same insertions that increased expression in
411 Jurkats also scoring positively in primary T cells (**Figure 4F**). Collectively, these results
412 establish the feasibility of PETRA in primary T cells.

413 **DISCUSSION**

414

415 We have demonstrated PETRA as a scalable means of accurately measuring how edits to
416 genome sequence impact mRNA levels by assaying over 14,000 variants across four genes
417 and two cell types. Further, we demonstrate that relatively large changes in expression can
418 be engineered by 6-bp insertions identified in large-scale screening experiments, as well as
419 by TF motifs programmed individually or in combination.

420

421 The scalability of PETRA is mediated by engineering variants in abundantly transcribed
422 regions, facilitating sequencing of edits in mRNA and gDNA to assay effects on expression
423 in a straightforward manner. This design enables thousands of variants per region to be
424 queried in multiplex, and allows PETRA to be applied to multiple genes with minimal
425 optimization. However, relying on variants to be observed in mRNA restricts the assay to
426 specific genomic regions, which is a key limitation.

427

428 Variants in close proximity to TSSs can impact expression through multiple mechanisms.
429 Indeed, effects on mRNA levels in PETRA experiments were associated with the creation of
430 new splice sites, ORFs and TFBSs. These effects were largely dependent on the surrounding
431 sequence context to which variants were introduced, explaining the high degree of locus-
432 specificity for insertion sequences assayed in large-scale experiments (**Figure 2B**). However,
433 the large size of PETRA libraries facilitated identification of sequence features that produce
434 consistent effects across insertion sites. For example, the MYBL2 motif was associated with
435 enhanced expression of *IL2RA* by scoring random 6-bp insertions at one site. Indeed,
436 deliberately designed MYBL2 binding sites increased expression when introduced
437 individually or in combination to additional sites in the *IL2RA* 5' UTR. However, MYBL2 motifs
438 also increased expression when delivered to the *VAV1* 5' UTR (**Figure 3C**).

439

440 While we focused on understanding how DNA sequences impact transcript levels, we also
441 found that select *IL2RA* variants predicted to create new TFBSs increased protein expression,
442 as well. Notably, this was not true for insertions that created new ORFs in the *IL2RA* 5' UTR,
443 suggesting such sequences should be avoided if increased protein expression is desired.
444 Though not established here, we anticipate it may be feasible to couple PETRA to readouts

445 directly assaying protein levels or functional phenotypes of edited cells. Indeed, two recent
446 studies in cell lines have deployed FACS-seq to link variants introduced via prime editing to
447 protein expression changes[21,22]. These studies examined effects of variants installed
448 upstream of each target gene, whereas PETRA identifies variants impacting expression
449 downstream of the TSS. The throughput afforded by targeting the 5' UTR allowed us to score
450 substantially more variants than in either previous study.
451
452 Given PETRA's scalability and adaptability to different loci, going forward we anticipate the
453 method will be well-suited to assaying variants across many genomic loci and cell types. It
454 will be informative to study, for instance, how the effects of TF motifs differ by genomic
455 position and flanking sequence context, and how the broader landscape of factor binding and
456 chromatin state at each locus influence variant effects[36,37]. Efforts to model RNA expression
457 from DNA sequence may benefit from PETRA data for training and benchmarking[38]. Towards
458 more robust prediction of variant effects across genomic contexts, we anticipate that a
459 combination of functional assays will likely prove ideal to advance computational models, for
460 instance, combining the scalability of MPRAs with the ability to test endogenously installed
461 variants across loci using PETRA.
462
463 For cell engineering applications, PETRA also offers distinct advantages. By directly testing
464 libraries of variants in their endogenous context, editing efficiencies and expression effects
465 can be measured directly for loci of therapeutic relevance. We have established that PETRA
466 identifies variants with mRNA expression effects spanning a wide range. In some potential
467 use cases, such as therapeutic editing to treat haploinsufficiency, modest expression
468 changes that remain stable over time may prove optimal[39].
469
470 In summary, we anticipate PETRA will substantially expand our ability to functionally test
471 endogenous variants for effects on RNA expression across a wide range of genomic contexts
472 and cell types. Given the scarcity of methods to reveal effects of regulatory variants across
473 genomic contexts, we anticipate this method will play a key role in improving our
474 understanding of gene regulation, building improved computational models of variant effect,
475 and identifying sequence changes of therapeutic benefit.

476 **METHODS**

477

478 **Ethics**

479

480 All experiments using primary human cells were approved by an internal review board of the
481 Francis Crick Institute (approval code: 2023 FC1; 3 April, 2023), in accordance with the UK
482 Human Tissue Act.

483

484 **Cell culture**

485

486 All cells were cultured at 37°C with 5% CO2. Viability and cell concentration were determined
487 from 10 µl of cells in media using a CellDrop Automated Cell Counter (Denovix).

488

489 *Jurkat cells*

490

491 Jurkat E6.1 cells were cultured in RPMI medium supplemented with 10% Fetal Bovine Serum
492 (FBS) and 1% penicillin/streptomycin (Gibco 15140122). For IL2RA-targeting screens, the
493 media was supplemented with 20 U/ml of interleukin-2 (IL-2) (Miltenyi 130-097-745). Cells
494 were activated by supplementing the media with 100 ng/ml phorbol 12-myristate 13-acetate
495 (PMA) (Sigma-Aldrich P1585) and 200 U/ml of IL-2 for 24 hours.

496

497 *Primary human T cells*

498

499 Human blood was obtained from the Francis Crick Institute's Healthy Volunteer Programme
500 from two anonymous donors. Peripheral blood mononuclear cells (PBMCs) were isolated by
501 Ficoll (Sigma-Aldrich H8889) density centrifugation, and stored in freezing media (90% FBS
502 + 10% DMSO) at −70°C until use. For selection of CD3+ cells, PBMCs were thawed and
503 sorted using magnetic activated cell sorting (MACS) with anti-CD3+ beads (Miltenyi 130-097-
504 043). The day of selection was deemed day 0. Primary T cells were cultured in X-VIVO media
505 (Lonza BEBP02-061Q) supplemented with 5% FBS, 1% penicillin/streptomycin and 500 U/ml
506 of IL-2. On days 0 and 14, T cells were activated using the human T Cell Activation/Expansion
507 kit (Miltenyi 130-091-441) with beads loaded with anti-CD2, anti-CD3 and anti-CD28

508 antibodies. Three days after re-activation (day 17), beads were removed by magnetic
509 separation prior to nucleofection.
510
511 ***In vitro* transcription**
512
513 *In vitro* transcription (i.v.t.) was performed as described[40] with modifications. Briefly, template
514 plasmids including pT7-PEmax for IVT (Addgene 178113), pT7-PE7 for IVT (Addgene
515 214813), and pT7-MLH1dn for IVT (Addgene 178114) were amplified with primers MA001
516 and MA002 (all primers in **Supplementary Table 13**) using Phusion U Multiplex PCR Master
517 Mix (ThermoFisher Scientific F564S). PCR products were purified and concentrated using
518 AMPure XP SPRI (Beckman Coulter A63880). I.v.t was performed using HiScribe T7 mRNA
519 Kit with CleanCap Reagent AG (New England Biolabs E2080S) with the following
520 modifications to the manufacturer's instructions per 20 µl reaction: addition of 0.5 µl RNase
521 inhibitor (New England Biolabs M0314S), replacement of UTP with 0.5 µl N1-Methyl-Pseudo-
522 UTP (New England Biolabs N0431S), reduction of GTP from 2 µl to 1 µl, use of 3 µg input
523 DNA, and addition of 0.2 µl Yeast Inorganic Pyrophosphatase (New England Biolabs
524 M2403S). Following 2 hours at 37°C, i.v.t reactions were incubated with DNaseI (New
525 England Biolabs M0303L) for 15 minutes at 37°C. RNA was purified using the Monarch RNA
526 Cleanup Kit (New England Biolabs T2050L) and quantified using Qubit RNA HS Assay Kit
527 (ThermoFisher Scientific Q32855). Single-use RNA aliquots were stored at −70°C.
528
529 **PETRA experiments**
530
531 *Nucleofections and harvesting*
532
533 All nucleofections were conducted using the Lonza Nucleofector II. Each nucleofection
534 reaction included 10 µg of MLH1dn mRNA, 1 pmol of pegRNA plasmid library and 10 µg of
535 PE7 or PEMax mRNA. A negative control nucleofected only with prime editor and MLH1dn
536 mRNA was included in each experiment.
537
538 For Jurkat cells, 1 million cells were nucleofected per reaction using programme X-001 (Lonza
539 VCA-1003). Two or three separate nucleofection reactions were performed per library and

540 treated as independent biological replicates. Jurkat cells were split every three days, and 10
541 million cells were harvested seven days post-nucleofection.
542
543 For primary T cells, 5 million cells were nucleofected per reaction using programme T-020
544 (Lonza VPA-1002). Two separate nucleofection reactions were performed for the IL2RA-6-
545 mer library and were treated as independent biological replicates. For the TF motif insertion
546 libraries, each independent biological replicate (n = 2) was obtained by pooling cells from four
547 independent nucleofection reactions. Media was replaced with fresh media three days post-
548 nucleofection, and between 2.2 million and 4.6 million (TF motif libraries) or 5 million (*IL2RA*
549 6-mer library) cells were harvested seven days post-nucleofection.
550
551 Successful nucleofection was confirmed by detecting GFP fluorescence in a sample
552 nucleofected with 2 µg of pmaxGFP plasmid (Lonza VPA-1002). Jurkat and primary T cells
553 were cultured in antibiotic-free conditions for 72 h after nucleofection. Primary T cells were
554 also cultured in antibiotic-free conditions for 72 h prior to nucleofection.
555
556 *DNA and RNA extraction*
557
558 Up to 10 million cells were centrifuged for 10 minutes at 350 × *g* (Jurkat cells) or 300 × *g*
559 (primary cells). Pellets were processed for purification of DNA and RNA using the AllPrep
560 DNA/RNA Kit (Quiagen 80204) with QIAShredder columns (Quiagen 79656). After extraction,
561 RNA concentration was measured with the Qubit RNA HS Assay Kit (ThermoFisher Scientific
562 Q32855) and DNA concentration was measured using the Qubit 1X dsDNA HS Assay Kit
563 (ThermoFisher Scientific Q33231). DNA samples were stored at −20°C and RNA samples
564 were stored at -70°C until processed for sequencing.
565
566 *Amplicon sequencing*
567
568 All PCR reactions were performed using KAPA HiFi HotStart ReadyMix (Roche 7958935001)
569 and supplemented with SYBR Safe (Invitrogen S33102) such that the number of cycles was
570 determined by qPCR to avoid over amplification. Products were verified by gel
571 electrophoresis and purified and concentrated with AMPure XP SPRI Reagent (Beckman
572 Coulter A6388). For gDNA amplification, up to 1 µg of DNA was amplified per 50 µl reaction,

573 with up to eight reactions per sample, which were pooled prior to product purification.
574 Reactions were complemented with an additional 2.5 mM $MgCl_2$.
575
576 For RNA amplification, up to 5 µg of total RNA were used as template for a reverse
577 transcription (RT) reaction using SuperScrip IV Reverse Transcriptase (ThermoFisher
578 Scientific 18090050) with gene-specific primers, following the manufacturer's instructions.
579 Each RT product was split into ten 50 µl PCR reactions, which were pooled prior to product
580 purification.
581
582 *Illumina Sequencing*
583
584 Amplicons containing Nextera adapters were dual-indexed and sequenced using the Illumina
585 Nextseq 500, with 75-, 150-, or 300-cycle kits, the Illumina Miniseq with 75-cycle kits or the
586 Illumina Novaseq 600, depending on required coverage. The length of Read 1 was sufficient
587 to include at least 10 bp of sequence downstream from the target site, assuming successful
588 editing of all variants at each locus.
589
590 *Library design and cloning*
591
592 pU6_pegRNA-tevopreQ1-polyT_BsmBI and pU6_pegRNA-polyT_BsmBI were prepared
593 from pU6_pegRNA-GG-Acceptor (AddGene 132777) through an intermediate by first
594 removing the mRFP1 cassette via PCR amplification using primers MH57 and MH58 followed
595 by Gibson self-assembly (New England Biolabs E2621L). The resulting vector was
596 subsequently amplified using either MH50 and MH52 or MH51 and MH52, prior to Gibson
597 self-assembly to generate pU6_pegRNA-polyT_BsmBI and pU6_pegRNA-tevopreQ1-
598 polyT_BsmBI, respectively.
599
600 pegRNA sequences and insertion coordinates for all libraries can be found in **Supplementary**
601 **Table 14**. Plasmids pU6_pegRNA-polyT_BsmBI and pU6_pegRNA-tevopreQ1-polyT_BsmBI
602 were linearised by BsmBI (New England Biolabs R0739L) digestion, gel-purified (Macherey-
603 Nagel, 740609.10) and used as backbones for Gibson Assembly (New England Biolabs
604 E2621L) to insert the pegRNA libraries. All pegRNAs used in this study included tevopreQ1
605 motifs[41].

606

607 *Random 6-bp insertion libraries*

608

609 The target loci for the random 6-bp insertion libraries were chosen based on high predicted
610 editing efficiency, inclusion in a cCRE, and avoidance of existing TF motifs. Detection of TF
611 motifs throughout the 5' UTR was performed by running FIMO with a detection threshold of p
612 $< 1 \times 10^{-4}$. Human TF motifs were downloaded from the JASPAR 2020 database[42] (release
613 8), and annotations for cCREs were retrieved from ENCODE[43]. PRIDICT[25] was used to
614 predict the editing efficiency of three random 6-bp insertions in every position of the 5' UTR
615 of each target gene. Of the pegRNAs designed to insert the three random insertions, the
616 highest scoring design not compromising a pre-existing TF motif was used to design the
617 pegRNA library. The 6-bp insertions used for editing prediction were replaced with a 6-bp
618 randomer (NNNNNN), such that all pegRNAs in the library differed solely by insertion
619 sequence. Each pegRNA design was extended at the 5′ end with an 18-bp fragment
620 complementary to the 3′ end of the U6 promoter (U6-comp) and at the 3′ end with the
621 tevopreQ1 sequence[41]. Oligonucleotides for each random 6-bp insertion library were
622 synthesised as PAGE-purified Ultramers with degenerate bases (Integrated DNA
623 Technologies).

624

625 Each library was amplified in six separate 50 µl reactions using primers MA059 and MA098
626 and cloned into pU6_pegRNA-polyT_BsmBI by Gibson assembly (New England Biolabs
627 E2621L). Products were purified, concentrated via AMPure XP SPRI (Beckman Coulter
628 A6388) clean-up, and used for transformation into Stable Competent E. Coli cells (New
629 England Biolabs C3040H) by mixing 25 µl of cells with 5 µl of assembled plasmid and
630 incubating on ice for 30 minutes. After a 30 second heat shock at 42°C followed by a 5-minute
631 incubation on ice, the cells were recovered for 1 hour at 30°C with shaking, and grown
632 overnight in liquid culture at 37°C. Plasmid libraries were extracted using ZymoPURE II
633 Plasmid Maxiprep Kit (Zymo Research D4203) with endotoxin removal.

634

635 *6-bp insertion validation libraries*

636

637 For the validation libraries, 50 high-scoring, 50 low-scoring and 50 neutral-scoring variants
638 were selected from the random 6-mer insertion libraries, using the same pegRNA designs.

639 Variants were selected based on expression scores prior to filtering, such that variants with
640 a range of gDNA frequencies were represented. pegRNAs were ordered as part of a
641 oligonucleotide pool (Twist Biosciences) and cloned as for the TF motif insertion libraries.
642
643 *Combinatorial MYBL2 and EGR1 motif insertion libraries in IL2RA*
644
645 Motifs for MYBL2 and EGR1 were extracted from the JASPAR 2020 database[42] (MYBL2 motif
646 MA0777.1: 5'-CCGTTAAAC; EGR1 motif MA0162.3-4: 5'-CGCCCACGC). PRIDICT was
647 used to predict the efficiency of inserting both motifs in every position of the 5' UTR of *IL2RA.*
648 Three positions were chosen based on high predicted editing efficiency for both insertions
649 while ensuring sufficient separation such that editing of each locus would not interfere with
650 pegRNA spacer and primer binding site (PBS) complementarity at another target. The highest
651 scoring pegRNA design for each insertion was chosen, and the pegRNA sequence was
652 extended at the 5′ end with the U6-comp sequence and at the 3′ end with the tevopreQ1
653 sequence followed by a stuffer sequence for length uniformity and a 20-nt library-specific
654 barcode prior to synthesis in an oligonucleotide pool (Twist Biosciences). The library-specific
655 barcode enabled the pegRNAs for each motif (MYBL2 or EGR1) to be amplified and cloned
656 separately, using MA104 and a library-specific reverse primer binding to the 20-nt library
657 barcode. Following purification of each product, a second amplification was performed with
658 MA059 and MA098. The MYBL2 motif insertion pegRNAs and the EGR1 motif insertion
659 pegRNAs were then assembled into pU6_pegRNA-polyT_BsmBI separately, and pooled in
660 equal quantities for experiments inserting both motifs.
661
662 *TF motif insertion libraries*
663
664 For single-motif insertions, the nucleotides from the JASPAR motif predicted to contribute
665 most strongly to factor binding were chosen as the core motif to be inserted. For double- and
666 triple-motif designs, tandem motifs in the same orientation were connected by adding the two
667 nucleotides flanking the core motif. Two previously assayed 6-bp insertions predicted to bind
668 each TF that had scored highly were also included in the library. Ten 6-bp insertions with high
669 expression scores not predicted to bind any of the selected TFs and ten neutral insertions
670 scoring between -0.1 and 0.1 were also included for each gene as controls, as was the 5'-
671 GGTAAG insertion which creates a canonical splice donor variant and the four possible 1-bp

672 insertions at each site. To create the *VAV1* TF motif insertion library, this process was
673 performed using data from the *VAV1* random 6-mer insertion screen. To create the *IL2RA*
674 motif insertion library, insertions were selected considering data from the *IL2RA* random 6-
675 mer screen in both Jurkat and primary T cells. Specifically, the ten top-scoring 6-bp insertions
676 chosen included five that scored highly in primary T cells and five that scored highly in Jurkats;
677 neutral variants were selected as those that showed average expression scores between cell
678 types close to zero; and an additional ten 6-bp insertions that had high expression scores in
679 both primary and Jurkat cells (but were not necessarily among the top scoring variants in
680 either) were also added to the library. All insertions were targeted to the same site used for
681 the 6-bp insertion screen. For each insertion, the pegRNA design with the highest PRIDICT2[44]
682 score using the HEK293T model was chosen, extended with cloning adapters, and
683 synthesised (Twist Biosciences). The pegRNA oligonucleotide libraries were then amplified
684 as described for the combinatorial MYBL2 and EGR1 motif insertion libraries, but replacing
685 primer MA098 with MA146. The products were then cloned into pU6_pegRNA-tevopreQ1-
686 polyT_BsmBI.
687
688 *Variant scoring and filtering*
689
690 Single-end reads were aligned to the reference amplicon using needleall (EMBOSS v6.6.0.0)
691 and reads with an alignment score lower than 300 were removed from analysis. Variants from
692 the random 6-bp, validation and TF motif insertion libraries were counted by searching for
693 exact matches to the insertion sequence plus 10 bp of flanking sequence. Variants from the
694 combinatorial MYBL2 and EGR1 insertion libraries were counted by searching for exact
695 matches of each possible allele, starting 10 bp upstream of insertion site 1 and ending 10 bp
696 downstream of insertion site 3. For gDNA samples, editing efficiencies were computed as the
697 number of reads matching an intended edit over the total number of reads.
698
699 To calculate expression scores:
700 1. Variant frequency in each sample was determined by dividing variant-specific reads
701 plus a pseudocount by the total number of reads post filtering.
702 2. Variant frequency in RNA was divided by variant frequency in gDNA from the same
703 sample and log2-transformed

3. The resulting ratios were normalised and averaged across replicates as follows for each library:
    a. For the random 6-bp insertion libraries:
        i. The log2-transformed RNA:DNA ratios were normalised to the median of all variants in the sample to produce an expression score per replicate.
        ii. Expression scores per variant were averaged across biological replicates and normalised to the median average score of all variants.
        iii. Variants for which the gDNA frequency was lower than a set threshold for any replicate were removed from further analysis and expression scores were normalised to the median of retained variants.
        iv. To determine significant effects, expression scores were standardised to z-scores, and p-values were computed assuming a standard normal distribution without formally testing for normality. A significance threshold was set to $p < 0.01$ based on q-q plot analysis.
    b. For the combinatorial MYBL2 and EGR1 motif libraries:
        i. The log2-transformed RNA:DNA ratios were normalised to the RNA:DNA ratio of the unedited allele to produce an expression score per replicate.
        ii. Expression scores per replicate for each variant were averaged across all independent biological replicates.
    c. For the TF libraries:
        i. The log2-transformed RNA:DNA ratios were normalised to the mean ratio of neutral controls to produce an expression score per replicate.
        ii. To test significance, per-replicate expression scores for each variant were compared to per-replicate expression scores for all neutral variants, using Welsch's unequal variances two-sample t-test with BH correction.
        iii. Expression scores per biological replicate were averaged, and once more normalised to the mean of all neutral variants to derive final scores.

736 For the random 6-mer insertion libraries, variants were filtered based on frequency in gDNA.
737 For 6-mer libraries targeting *CD28*, *VAV1*, and *OTUD7B*, the gDNA frequency threshold was
738 set to 1×10^{-5}. For the library targeting *IL2RA*, the gDNA threshold was set to 8×10^{-5}. A
739 higher gDNA threshold was used for *IL2RA* due to lower RNA expression in resting Jurkats
740 resulting in weaker inter-replicate correlations for variant frequency in cDNA. No variants were
741 filtered from the validation libraries, the combinatorial MYBL2 and EGR1 motif insertion
742 libraries, or the TF motif insertion libraries tested in Jurkat and primary T cells. To assess
743 reproducibility of editing and variant scoring, inter-replicate Spearman's correlations for
744 variant gDNA frequency, cDNA frequency, and expression scores were calculated.
745
746 **Analysis of sequence features in random 6-mer insertion libraries**
747
748 *uORF detection and classification*
749
750 Variants leading to a new 5'-ATG (referred to as 5'-ATG variants) were classified as uORF
751 variants if there was stop codon in-frame with the new 5'-ATG preceding the canonical ORF.
752 Otherwise, 5'-ATG variants were classified as either an in-frame oORF variant or an out-of-
753 frame oORF variant, depending on whether the new ORF was in-frame with the canonical
754 ORF. The strength of the 5'-ATG context was determined based on Kozak similarity[45]. Briefly,
755 variants were assessed for the presence of an 'A' or 'G' three nucleotides upstream of the
756 new 5'-ATG and the presence of a 'G' immediately downstream. Variants were deemed
757 "strong" if both conditions were met, "moderate" if one condition was met, and "weak" if neither
758 condition was met.
759
760 Expression scores for variants creating different types of ORFs were compared to those of
761 non-ATG variants using a Mann-Whitney U test with BH correction. Within each ORF type,
762 variants with different Kozak strengths were compared the same way. Variants with SpliceAI
763 scores over 0.1 were removed from this analysis.
764
765 *TFBS detection*
766
767 To detect potential TFBSs created by variants, FIMO[32] was run in batch mode to scan a 46-
768 bp window including each insertion and 20 bp of sequence on each side for human TF motifs

769 from JASPAR 2024 (release 10)[33]. Variants were classified as predicted to bind a given TF if
770 the TF–variant p-value was below the detection threshold parameter, which was set to 0.001.
771 FIMO output was filtered to consider only TFs expressed in Jurkat cells (log2-transformed
772 TPM+1 greater than 0 in batch-corrected public expression dataset 24Q4I, DepMap ID ACH-
773 000995)[46] and only TFBSs created by an insertion sequence. Then, for each TF, a TF motif
774 score was calculated as the difference between the median expression score for variants
775 predicted to bind and the median expression score for all other variants. Significance was
776 determined using a two-sample Kolmogorov–Smirnov test with BH correction. Variants with
777 a SpliceAI score over 0.1 were removed from this analysis.
778
779 **Prediction of variant effects with deep learning models**
780
781 *SpliceAI*
782
783 To obtain predictions for variant effects on RNA splicing, SpliceAI was run from the command
784 line, using hg38 as the reference genome, GENCODE V24 canonical files for transcript
785 annotation, and a maximum distance parameter of 300 bp. Except where donor gain scores
786 are specifically analysed, the maximum score among individual scores for donor gain,
787 acceptor gain, donor loss and acceptor loss was used as a single SpliceAI score in analysis.
788
789 *Enformer*
790
791 Enformer was run to predict the effect of each 6-bp insertion. For each variant, a reference
792 and an alternative allele were created by centring the variant in a GRCh38 window with
793 196,608 bp on either side. Enformer predictions for 5,313 output tracks were generated for
794 reference and alternative alleles partitioned into 893 bins of 128 bp. Outputs across all bins
795 were then averaged, and the predicted effect of the variant on each track was quantified as
796 the difference between the alternative and reference alleles. Predictions were normalised
797 using the EnformerScoreVariantsNormalized function. In a second approach, only predictions
798 for the three bins centred on each variant were averaged to quantify the difference between
799 alternative and reference alleles (**Supplementary Figure 2M**).
800

801 Track metadata was obtained from:
802 https://raw.githubusercontent.com/calico/basenji/master/manuscripts/cross2020/targets_hu
803 man.txt.
804 Tracks 120 and 4831 were considered as Jurkat-specific DNase and CAGE predictions,
805 respectively. To obtain primary T cell-specific DNase and CAGE predictions, predictions for
806 all primary T cell-specific tracks were averaged. A track was considered primary T cell-
807 specific if the description contained any of the terms “T cell”, “T-cell”, “Tcell”, “CD4” or “CD8”
808 (case insensitive) with the exception of tracks 4731, 5029, 1882, 3584, 3653, 3827, and 4158.
809
810 *Borzoi*
811
812 Borzoi[47] was run to predict the effect of 6-bp insertions. All model and reference files were
813 obtained from https://github.com/calico/borzoi, with params.json and model0_best.h5 used
814 as model parameters and weights. For each variant, alternative and reference alleles were
815 created by centring the variant in a window with 262,144 bp on either side (GRCh38). Borzoi
816 predictions for reference and alternative alleles were generated for 7,612 output tracks
817 partitioned into 32-bp bins. For each bin, predictions for the forward and reverse-complement
818 orientation were averaged. Using the gencode41_basic_nort.gtf file for gene annotations,
819 bins were mapped to gene exons. Only bins overlapping the target gene were considered
820 and summed to obtain a single reference and alternative prediction per track. The predicted
821 effect of each variant on each track was then computed as the difference between the log2-
822 transformed predictions for the alternative and reference alleles (logSED).
823
824 Track metadata was obtained from targets_human.txt. Tracks 1396, 312 and 6176 were
825 considered as Jurkat-specific DNase, CAGE and RNA-seq predictions, respectively. Primary
826 T cell-specific tracks were selected using the same criteria as for Enformer.
827
828 **Quantification of variant effects on CD25 protein expression**
829
830 *pegRNA cloning*
831
832 pegRNAs were selected from the TF motif insertion library targeting *IL2RA* and ordered as
833 gene fragments (Twist Biosciences). Gene fragments were amplified and cloned into

834 pU6_pegRNA-tevopreQ1-polyT_BsmBI by Gibson assembly (New England Biolabs
835 E2621L). The assembled plasmids were transformed into Stable Competent E. Coli and
836 prepared as described for pegRNA library cloning. Plasmids were sequence-verified by whole
837 plasmid sequencing (Plasmidsaurus).
838
839 *Editing*
840
841 Plasmids expressing pegRNAs were delivered individually to Jurkat cells using the PETRA
842 nucleofection protocol in two independent biological replicates. Each nucleofection resulted
843 in a mixed pool of cells with a different number of edited alleles. To determine editing
844 efficiency, 1 µg of gDNA was amplified as described for *IL2RA* gDNA samples in PETRA
845 experiments. Amplicons containing Nextera adapters were dual-indexed and sequenced
846 using a 150-cycle kit on the Illumina Nextseq 500, and editing efficiencies were calculated
847 from single-end reads as described above, following alignment with needleall (EMBOSS
848 v6.6.0.0) and removal of reads with an alignment score lower than 300.
849
850 *IL2RA staining and flow cytometry*
851
852 Jurkat cells were stained for CD25 using PE-conjugated anti-human CD25 antibody
853 (BioLegend 302606). Cells were centrifuged, washed in PBS and incubated for 30 minutes
854 at 4°C in the dark in staining solution (1 ml per 10^7 cells), comprised of a 1:20 dilution of
855 primary antibody in flow buffer (ThermoFisher Scientific 00-4222-26) complemented with 5%
856 FBS. After staining, cells were washed twice and resuspended in flow buffer (1 ml per 10^7
857 cells) for analysis. All centrifugations were performed at 350g for 10 minutes. A staining
858 control, processed with the same protocol but without antibody in the staining buffer, was
859 included in each experiment.
860
861 For each sample (n = 2 per variant), 10,000 cells were analysed in a SH800S Cell Sorter
862 (SONY) using a 100 µm microfluidic sorting chip with the following detector and threshold
863 settings: FSC: 4; BSC: 32%; FL1-FITC: 55%; FL2-PE: 48%. PE signal was used for detection
864 of anti-IL2RA antibody; FITC was used to measure autofluorescence.
865

866 Flow cytometry data was analysed using FlowJo 10.10.0 software (BD Life Sciences). Raw
867 data was filtered by gating to remove cell debris and cell doublets. The mean fluorescence
868 intensity (MFI) from the cell population was exported for analysis. Cells with a PE signal above
869 the maximum staining control signal were considered $CD25^+$. MFI and percentage of $CD25^+$
870 cells measurements were log2-transformed and normalised to the mean of neutral variant
871 samples. For each variant, the MFI and percentage of $CD25^+$ cells across replicates were
872 compared to the same values obtained from neutral variants (considered as a single group)
873 using a Student's t-test with BH correction.

DATA AVAILABILITY

All expression scores from PETRA experiments are available in **Supplementary Tables**. Fastq files from PETRA experiments will be made available prior to publication (European Nucleotide Archive accession: PRJEB102640).

CODE AVAILABILITY

Custom scripts used to analyse PETRA experiments and to generate figures are available on GitHub: https://github.com/magdaaarmas/PETRA.

AUTHOR CONTRIBUTIONS

M.A.R. performed all experiments and analyses. M.H. generated critical reagents and provided prime editing advice. L.C. assisted with flow cytometry and sequencing. M.A.R. and G.M.F. wrote the manuscript with input from all authors. G.M.F. supervised the work.

ACKNOWLEDGEMENTS AND FUNDING

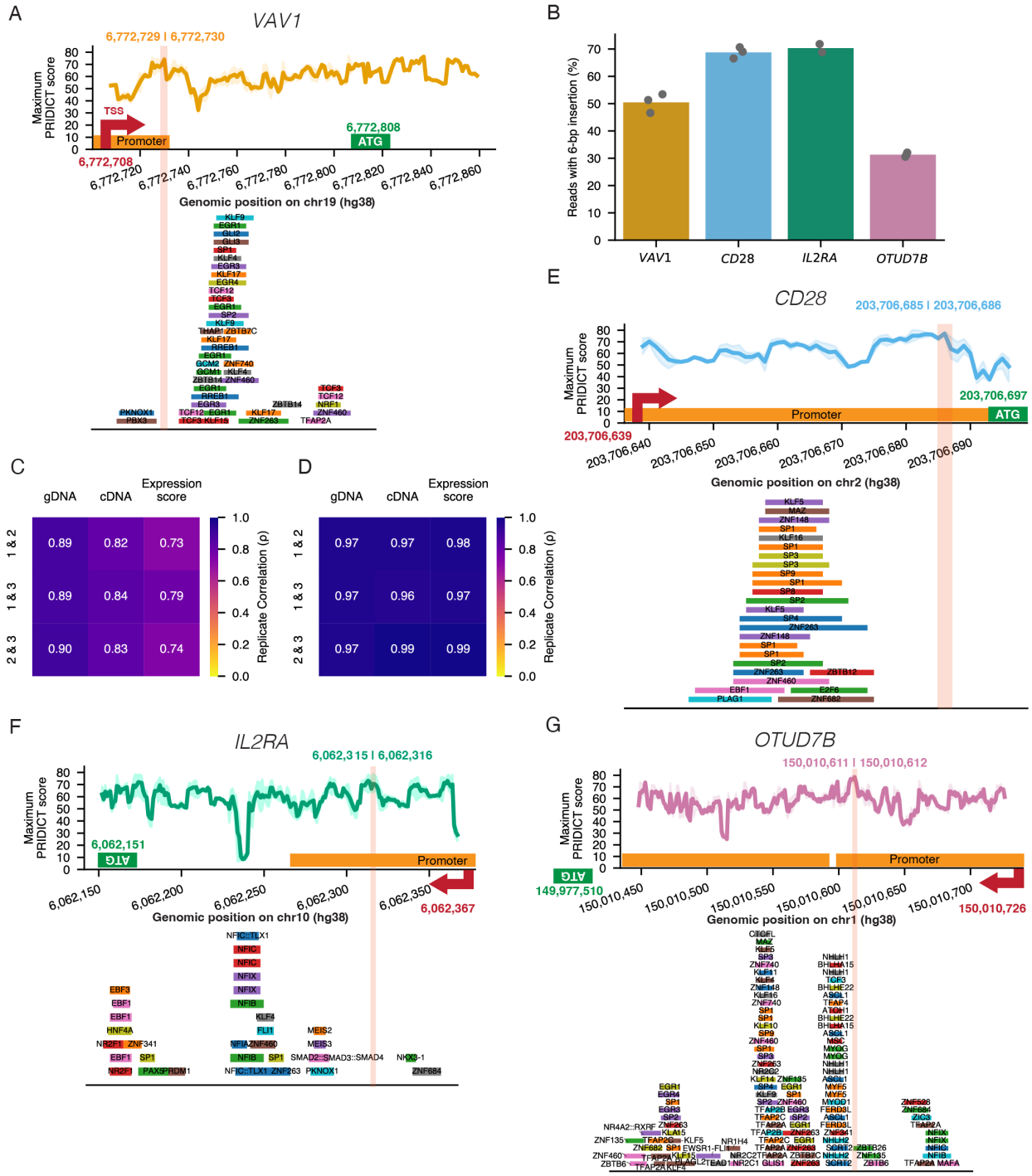
We thank the Crick's Genomics Scientific Technology Platform (STP) for performing sequencing, the laboratory of James Lee for technical advice, Gita Mistry and the Human Biology STP for facilitating access to donor blood, the Flow Cytometry STP for flow cytometry assistance, and the Cell Services STP for cell line maintenance.

G.M.F. is supported by a Starting Grant from the European Research Council (Seq2Func-NC). The Francis Crick Institute receives its core funding (G.M.F.) from Cancer Research UK (CC2190), the UK Medical Research Council (CC2190), and the Wellcome Trust (CC2190). For the purpose of Open Access, the authors have applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

COMPETING INTERESTS

The authors declare no conflicts of interest.

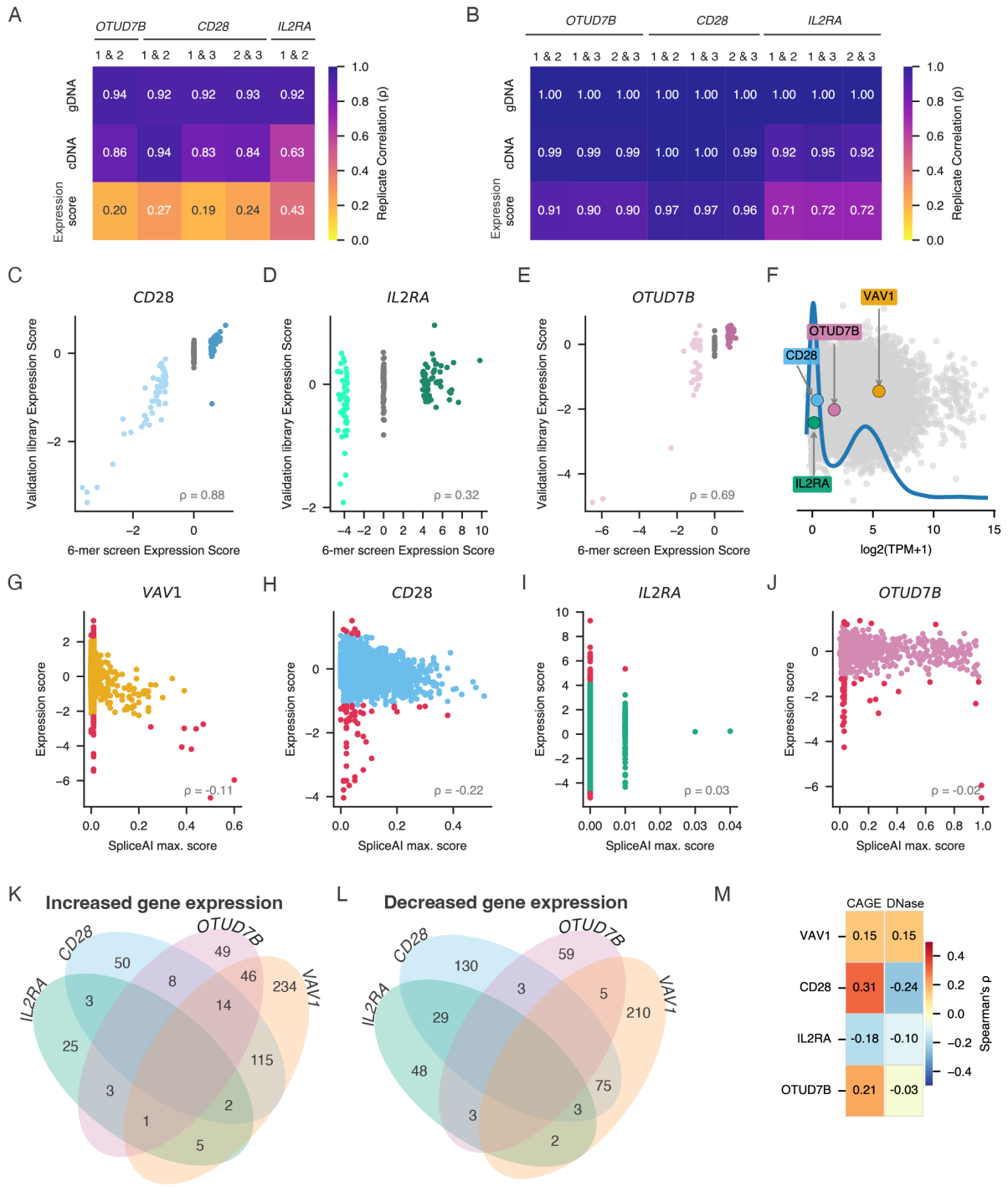
SUPPLEMENTARY FIGURES



Supplementary Figure 1. Delivery of 6-bp insertion libraries to promoter regions of four genes using prime editing.

A. Design of 6-bp insertion library targeting the 5' UTR of *VAV1*. The predicted editing efficiency of the top-scoring pegRNA designed to install each of three randomly selected 6-bp insertions is plotted by position. The dark line represents the average PRIDICT score of the three insertions, and the shaded area represents the range between maximum and minimum PRIDICT scores. Each gene's ENCODE-defined

913 promoter-like signature (PLS) candidate cis-regulatory element (cCRE) is indicated in orange. The position
914 chosen for each 6-bp library insertion is highlighted in red, with genomic coordinates provided. Predicted
915 TFBSs in each region are shown below.
916 **B.** Overall 6-bp insertion efficiency measured in day 7 gDNA sequencing of each target (bar), with individual
917 replicates as dots (*n* = 3 for *VAV1* and *CD28*, *n* = 2 for *IL2RA* and *OTUD7B*).
918 **C.** Correlations (Spearman's ρ) between biological replicates for variant frequencies in gDNA, cDNA and
919 variant expression scores, for the *VAV1* 6-bp insertion library (*n* = 3,902).
920 **D.** Correlations, as in B, between biological replicates for the *VAV1* validation library (*n* = 150).
921 **E-G.** Design of 6-bp insertion libraries targeting the 5' UTRs of *CD28* (E), *IL2RA* (F), and *OTUD7B* (G), as
922 represented in (A) for *VAV1.*
923

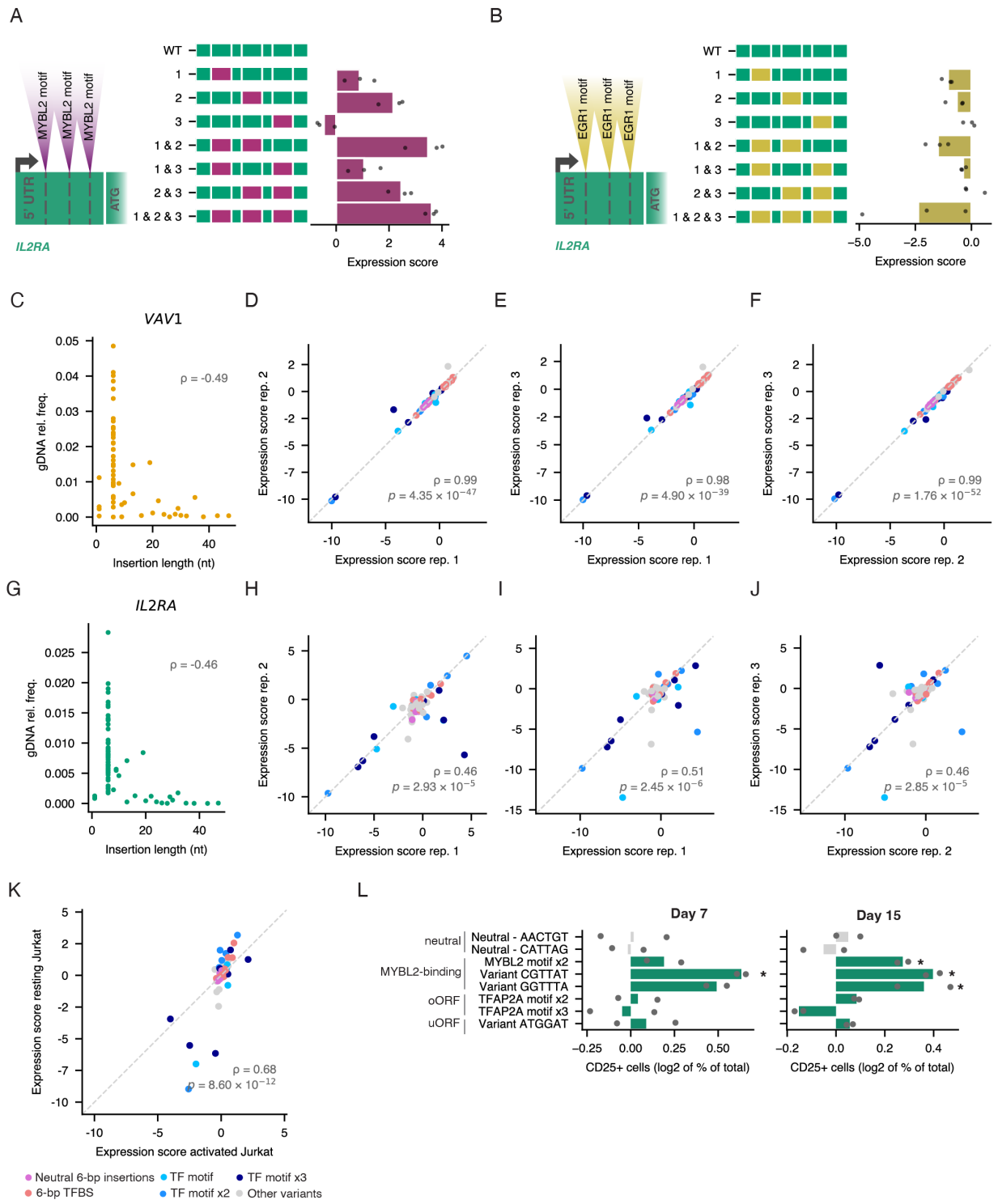


Supplementary Figure 2. Analysis of PETRA expression scores across loci.
A. Heatmap of Spearman's correlations (ρ) between biological replicates for variant frequencies in gDNA, cDNA and variant expression scores, considering variants retained post-filtering on gDNA frequency.
B. Correlations, as in A, between biological replicates for the validation libraries (*n* = 150 for each).
C-E. Correlations (ρ) between expression scores for *n* = 150 insertions scored in the 6-mer screen and re-tested in validation libraries for *CD28* (C), *IL2RA* (D), and *OTUD7B* (E). All variants included in library design are plotted. (dark shades, high scoring; grey, neutral scoring; light shades, low-scoring)
F. Log2-transformed transcripts per million (TPM)+1 values from Jurkat RNA-seq are plotted (batch corrected public expression dataset 24Q4, DepMap ID ACH-000995).

G-J. Correlation (Spearman's ρ) between expression scores and SpliceAI scores for each gene. Variants with significant expression scores are represented in red. (*VAV1 n* = 3,902, *CD28 n* = 3,930, *IL2RA n* = 2,419, *OTUD7B n* = 3,684)

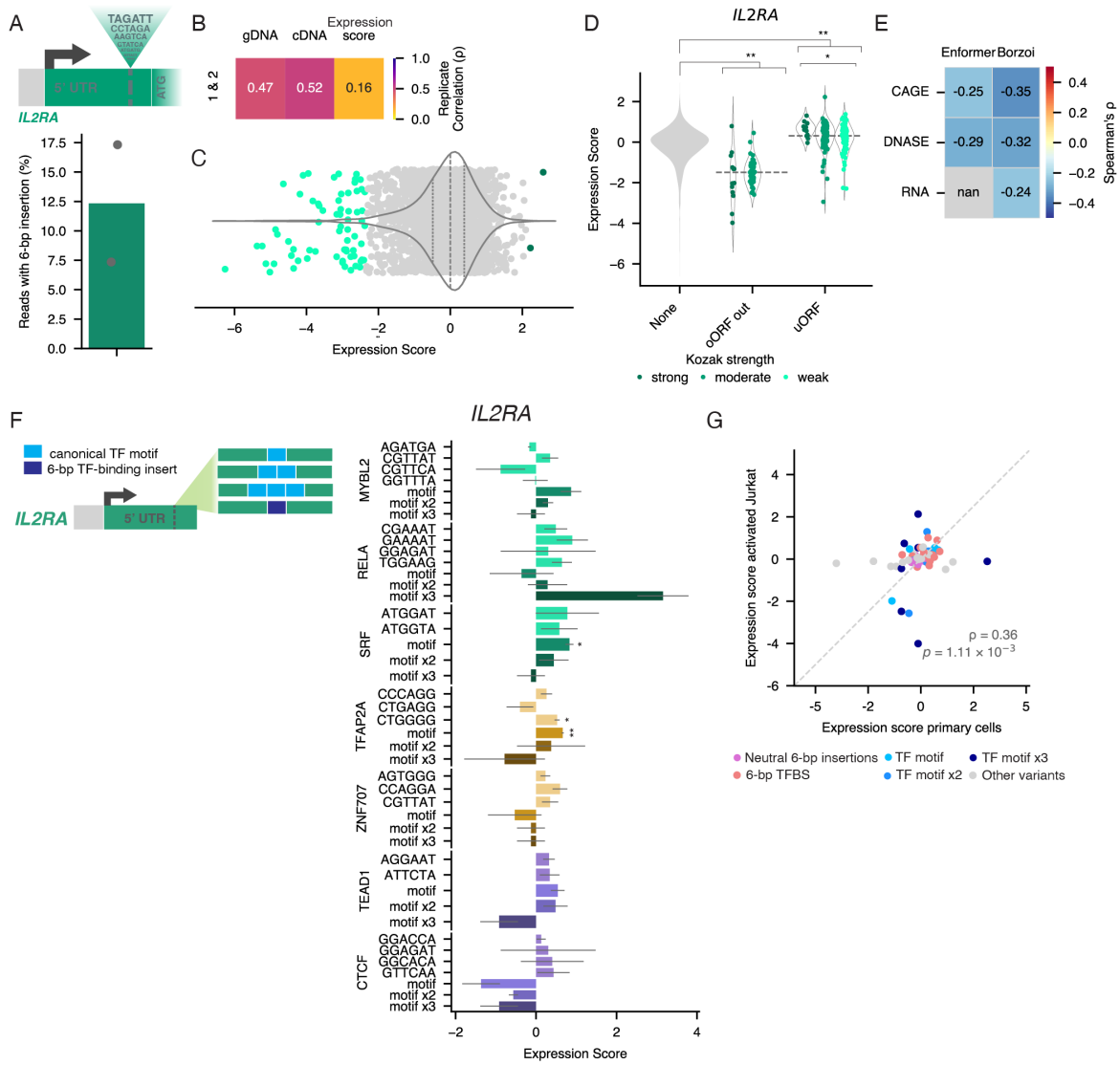
K-L. Venn diagrams showing the overlap of TF motifs associated with increased (K) or decreased (L) expression scores across loci.

M. Correlations (Spearman's ρ) between PETRA expression scores and Jurkat-specific predictions from Enformer for CAGE and DNase assays, considering only the three output bins centred on each variant to predict effects (see Methods). Coloured (non-grey) squares indicate significant correlations ($p < 0.05$).



Supplementary Figure 3. Using PETRA to assay TF motif insertions.
A-B. Schematic: three pegRNAs designed to insert either the MYBL2 (A) or EGR1 (B) motif at each of three sites in the 5' UTR of *IL2RA* were co-delivered to cells to achieve combinatorial editing. The barplot shows PETRA expression scores for each of 8 variant alleles observed following editing, normalised to the unedited allele (insertions depicted at left). Points represent independent replicates (n = 3) and bars represent the average across replicates.

950 **C.** Correlation (Spearman's ρ) between insertion length and gDNA variant frequency for the TF motif
951 insertion library targeting *VAV1*.
952 **D-F**. Scatter plots show interreplicate correlations for expression scores, with variants coloured based on
953 insertion type (as in K).
954 **G.** Same as C for *IL2RA.*
955 **H-J.** Same as D-F for *IL2RA.*
956 **K.** Correlation for expression scores for variants from the TF tandem insertion library targeting *IL2RA*
957 between resting and activated Jurkat cells. Variants are coloured based on insertion type.
958 **L.** Cell populations were assessed by flow cytometry at day 7 (left) and day 15 (right) after a 24 h
959 activation with PMA. Barplots represent the frequency of $CD25^+$ cells normalised to neutral variants. (* =
960 $q < 0.05$ and ** = $q < 0.01$ for two-sample Student's t-test versus neutral controls with BH correction)
961



Supplementary Figure 4. Analysis of PETRA expression scores in primary human T cells.
A. Schematic of the random 6-bp insertion library (n = 4,096) targeting the 5' UTR of *IL2RA*. Barplot shows average 6-bp insertion efficiency on day 7, with individual replicates as dots (n = 2).
B. Heatmap of Spearman's ρ between biological replicates for variant frequencies in gDNA, cDNA and variant expression scores for the 6-bp insertion library targeting *IL2RA* (n = 3,143 variants post-filtering).
C. Expression scores for variants from the 6-bp insertion library targeting *IL2RA* (n = 3,143) averaged across two biological replicates and median-normalised. (violin plot: dashed line, median; dotted lines, 25th and 75th percentiles, green dots indicate standard normal distribution test $p < 0.01$).
D. Violin plots of expression scores grouping insertion by ORF type and Kozak sequence strength. Horizontal dotted lines indicate median by ORF type. (* = $q < 0.05$ and ** = $q < 0.01$ for Mann-Whitney U-test with BH correction; oORF out = out-of-frame oORF)
E. Heatmap of Spearman's ρ between primary T-cell derived PETRA expression scores and human T cell-specific predictions from Enformer and Borzoi for CAGE, DNase and RNA-seq assays (n = 3,143). Non-grey squares indicate significant correlations ($p < 0.05$).
F. The *IL2RA* pegRNA library designed to deliver TF motifs individually or in tandem was assayed in CD3+ primary T cells. Expression scores, normalised to control variants, are plotted. Variants are grouped by TF motif and error bars represent standard error across n = 2 replicates. (* = $q < 0.05$ and ** = $q < 0.01$ for Welsch's unequal variances two-sample t-test against neutral variants with BH correction)

982 **G.** Correlation (Spearman's ρ) of expression scores tested in activated Jurkat and primary T cells for *n* =
983 77 variants from the library in (F), coloured by insertion type.

984 **SUPPLEMENTARY TABLES**

985 **Supplementary Table 1.** Expression scores for 13,935 6-bp insertions scored with PETRA
986 in Jurkat cells.
987 **Supplementary Table 2.** Expression scores for 6-bp insertions scored in validation
988 libraries.
989 **Supplementary Table 3**. Analysis of uORFs from the 6-bp insertion screens.
990 **Supplementary Table 4.** TF motif scores derived from expression scores of 6-bp
991 insertions in Jurkat cells.
992 **Supplementary Table 5.** Expression scores for *IL2RA* alleles edited with MYBL2 and
993 EGR1 motif insertions in Jurkat cells.
994 **Supplementary Table 6.** Expression scores for *VAV1* alleles edited with the TF motif
995 insertion library in Jurkat cells.
996 **Supplementary Table 7.** Expression scores for *IL2RA* alleles edited with the TF motif
997 insertion library in resting and activated Jurkat cells.
998 **Supplementary Table 8.** Anti-CD25 flow cytometry analysis.
999 **Supplementary Table 9.** Expression scores for 6-bp insertions in *IL2RA* assayed in CD3+
1000 primary T cells.
1001 **Supplementary Table 10.** TF motif scores derived from expression scores for *IL2RA* 6-bp
1002 insertions in CD3+ primary T cells.
1003 **Supplementary Table 11.** Expression scores for *IL2RA* alleles edited with the TF motif
1004 insertion library in CD3+ primary T cells.
1005 **Supplementary Table 12.** Expression scores for *VAV1* alleles edited with the TF motif
1006 insertion library in CD3+ primary T cells.
1007 **Supplementary Table 13.** Oligonucleotides used in this study.
1008 **Supplementary Table 14.** pegRNA designs used in this study.

REFERENCES

1. Frangoul, H. *et al.* CRISPR-Cas9 gene editing for sickle cell disease and β-thalassemia. *N. Engl. J. Med.* **384**, 252–260 (2021).
2. Lee, D. *et al.* A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* **47**, 955–961 (2015).
3. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535–548.e24 (2019).
4. Zeng, T. & Li, Y. I. Predicting RNA splicing from DNA sequence using Pangolin. *Genome Biol.* **23**, 103 (2022).
5. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
6. Sasse, A. *et al.* Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings. *Nat. Genet.* **55**, 2060–2064 (2023).
7. Sasse, A., Chikina, M. & Mostafavi, S. Unlocking gene regulation with sequence-to-function models. *Nat. Methods* **21**, 1374–1377 (2024).
8. Smith, C. & Kitzman, J. O. Benchmarking splice variant prediction algorithms using massively parallel splicing assays. *Genome Biol.* **24**, 294 (2023).
9. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
10. Kerimov, N. *et al.* A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat. Genet.* **53**, 1290–1299 (2021).
11. Abdellaoui, A., Yengo, L., Verweij, K. J. H. & Visscher, P. M. 15 years of GWAS discovery: Realizing the promise. *Am. J. Hum. Genet.* **110**, 179–194 (2023).
12. Tewhey, R. *et al.* Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165**, 1519–1529 (2016).

13. Kircher, M. *et al.* Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.* **10**, 3583 (2019).
14. Klein, J. C. *et al.* A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat. Methods* **17**, 1083–1091 (2020).
15. Canver, M. C. *et al.* BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527**, 192–197 (2015).
16. Fulco, C. P. *et al.* Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **354**, 769–773 (2016).
17. Korkmaz, G. *et al.* Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat. Biotechnol.* **34**, 192–198 (2016).
18. Gasperini, M. *et al.* A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* **176**, 377–390.e19 (2019).
19. Martin-Rufino, J. D. *et al.* Massively parallel base editing to map variant effects in human hematopoiesis. *Cell* **186**, 2456–2474.e24 (2023).
20. Findlay, G. M. *et al.* Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217–222 (2018).
21. Martyn, G. E. *et al.* Rewriting regulatory DNA to dissect and reprogram gene expression. *Cell* **188**, 3349–3366.e23 (2025).
22. Tak, Y. E. *et al.* CRISPR PERSIST-On enables heritable and fine-tunable human gene activation. *bioRxivorg* 2024.04.26.590475 (2024) doi:10.1101/2024.04.26.590475.
23. Anzalone, A. V. *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).
24. Schmidt, R. *et al.* CRISPR activation and interference screens decode stimulation responses in primary human T cells. *Science* **375**, eabj4008 (2022).
25. Mathis, N. *et al.* Predicting prime editing efficiency and product purity by deep learning. *Nat. Biotechnol.* **41**, 1151–1159 (2023).

26. Yan, J. *et al.* Improving prime editing with an endogenous small RNA-binding protein. *Nature* (2024) doi:10.1038/s41586-024-07259-6.
27. Chen, P. J. *et al.* Enhanced prime editing systems by manipulating cellular determinants of editing outcomes. *Cell* **184**, 5635–5652.e29 (2021).
28. Blaeschke, F. *et al.* Modular pooled discovery of synthetic knockin sequences to program durable cell therapies. *Cell* **186**, 4216–4234.e33 (2023).
29. Sample, P. J. *et al.* Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. *Nat. Biotechnol.* **37**, 803–809 (2018).
30. Wieder, N. *et al.* Differences in 5'untranslated regions highlight the importance of translational regulation of dosage sensitive genes. *Genome Biol.* **25**, 111 (2024).
31. DepMap: The Cancer Dependency Map Project at Broad Institute. https://depmap.org/portal/depmap/.
32. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
33. Rauluseviciute, I. *et al.* JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **52**, D174–D182 (2024).
34. Barbadilla-Martínez, L., Klaassen, N., van Steensel, B. & de Ridder, J. Predicting gene expression from DNA sequence using deep learning models. *Nat. Rev. Genet.* 1–15 (2025).
35. Koeppel, J. *et al.* Prediction of prime editing insertion efficiencies using sequence features and DNA repair determinants. *Nat. Biotechnol.* **41**, 1446–1456 (2023).
36. Zhu, F. *et al.* The interaction landscape between transcription factors and the nucleosome. *Nature* **562**, 76–81 (2018).
37. Duttke, S. H. *et al.* Position-dependent function of human sequence-specific transcription factors. *Nature* **631**, 891–898 (2024).
38. de Boer, C. G. & Taipale, J. Hold out the genome: a roadmap to solving the cis-regulatory

code. *Nature* **625**, 41–50 (2024).

39. Matharu, N. *et al.* CRISPR-mediated activation of a promoter or enhancer rescues obesity caused by haploinsufficiency. *Science* **363**, eaau0629 (2019).
40. Doman, J. L., Sousa, A. A., Randolph, P. B., Chen, P. J. & Liu, D. R. Designing and executing prime editing experiments in mammalian cells. *Nat. Protoc.* **17**, 2431–2468 (2022).
41. Nelson, J. W. *et al.* Engineered pegRNAs improve prime editing efficiency. *Nat. Biotechnol.* **40**, 402–410 (2022).
42. Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48**, D87–D92 (2020).
43. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
44. Mathis, N. *et al.* Machine learning prediction of prime editing efficiency across diverse chromatin contexts. *Nat. Biotechnol.* (2024) doi:10.1038/s41587-024-02268-2.
45. Whiffin, N. *et al.* Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals. *Nat. Commun.* **11**, 2523 (2020).
46. Arafeh, R., Shibue, T., Dempster, J. M., Hahn, W. C. & Vazquez, F. The present and future of the Cancer Dependency Map. *Nat. Rev. Cancer* **25**, 59–73 (2025).
47. Linder, J., Srivastava, D., Yuan, H., Agarwal, V. & Kelley, D. R. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. *Nat. Genet.* **57**, 949–961 (2025).