# Integrative dissection of gene regulatory elements at base resolution

Zeyu Chen[1,2,3,6], Nauman Javed[1,2,3,6], Molly Moore[2], Jingyi Wu[1,2,3], Michael Vinyard[2,4,5], Luca Pinello[2,4], Fadi J. Najm[2,*], Bradley E. Bernstein[1,2,3,7,*]

[1]Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA.
[2]Gene Regulation Observatory, Broad Institute, Cambridge, MA, USA.
[3]Department of Cell Biology and Pathology, Harvard Medical School, Boston, MA, USA.
[4]Department of Pathology, Massachusetts General Hospital and Harvard Medical School.
[5]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA.
[6]These authors contributed equally
[7]Lead Contact
#Correspondence: fadinajm@broadinstitute.org; bradley_bernstein@dfci.harvard.edu

**Key words:**
Regulatory elements, CRISPRi, base editing, transcription factor, immune response

**Highlights:**

- Base editing screens and deep learning pinpoint sequences and single bases affecting immune gene expression
- An artificial C-to-T variant in a regulatory element suppresses CD69 expression by altering the balance of transcription factor binding
- Competition between GATA3 and BHLHE40 regulates inducible immune genes and T cell states

## Summary

Although vast numbers of putative gene regulatory elements have been cataloged, the sequence motifs and individual bases that underlie their functions remain largely unknown. Here we combine epigenetic perturbations, base editing, and deep learning models to dissect regulatory sequences within the exemplar immune locus encoding CD69. Focusing on a differentially accessible and acetylated upstream enhancer, we find that the complementary strategies converge on a ~170 base interval as critical for CD69 induction in stimulated Jurkat T cells. We pinpoint individual cytosine to thymine base edits that markedly reduce element accessibility and acetylation, with corresponding reduction of CD69 expression. The most potent base edits may be explained by their effect on binding competition between the transcriptional activator GATA3 and the repressor BHLHE40. Systematic analysis of GATA and bHLH/Ebox motifs suggests that interplay between these factors plays a general role in rapid T cell transcriptional responses. Our study provides a framework for parsing gene regulatory elements in their endogenous chromatin contexts and identifying operative artificial variants.

## Introduction

Genome-wide maps of chromatin state and transcription factor (TF) binding have nominated more than a million cell type-specific regulatory elements (REs) in the human genome as potential context-specific regulators of gene expression(Andersson and Sandelin, 2020; ENCODE Project Consortium et al., 2020; Stunnenberg et al., 2016). A critical next step is to determine their functions and sequence determinants.  Computational tools that predict functional bases and/or gene targets are rapidly evolving, but require systematic benchmarking against perturbational data (Avsec et al., 2021; Nasser et al., 2021). Massively parallel reporter assays (MPRA) enable high-throughput analysis of sequence determinants within REs, but are based on exogenously introduced constructs that do not recapitulate the native chromatin contexts (Kheradpour et al., 2013; Klein et al., 2020; Maricque et al., 2018; Melnikov et al., 2012). CRISPR interference (CRISPRi) with fusions between dCas9 and the KRAB repressor provides a means to suppress a regulatory element in its native context and evaluate consequent transcriptional changes (Canver et al., 2015; Fulco et al., 2016; Gilbert et al., 2013; Korkmaz et al., 2016; Sanjana et al., 2016). Traditional CRISPR-based genetic perturbations offer increased resolution(Diao et al., 2017; Rajagopal et al., 2016), but may incur variable sequence changes due to heterogeneity of indels after DNA repair. Base editors can incur single base variants without frame-shifts or indels. They have been used to systematically characterize coding variants(Cuella-Martin et al., 2021; Gaudelli et al., 2017; Hanna et al., 2021; Kim et al., 2017; Komor et al., 2016), but have yet to be applied to noncoding REs.

In this study, we integrated CRISPRi, dCas9 and base editing with computational predictions to parse non-coding regulatory sequences in the CD69 locus. We found a ~170bp interval within a ~1500bp enhancer proximal to the CD69 promoter which plays a key role in regulating gene  expression. Within this interval, base editing and deep learning converge upon a critical cytosine at chr12:9764948 (hg38), where a C-to-T transition significantly reduces element accessibility and CD69 expression. We show that this C-to-T base edit ablates a GATA3 binding site, thereby exposing a nearby E-box/bHLH site for BHLHE40 binding. Systematic analysis of chromatin accessibility and TF binding during T-cell activation supports a global role for binding competition between GATA3 and BHLHE40 in immune gene responses and T cell polarization.

## Results

### Resolving functional bases within immune regulatory elements

To dissect functional sequences within  regulatory elements, we established a workflow combining chromatin profiling, deep learning, CRISPRi, dCas9 and base editing (**Figure 1A**). We combined ATAC-Seq accessibility maps with deep learning models to predict REs and functional sequences that regulate inducible gene expression in T cells. We then incorporated CRISPRi, dCas9 interference and base editing to directly test the regulatory functions of sequences and individual bases (Figure 1A).

We focused on the CD69 locus, which encodes a key molecule for T cell signal transduction and tissue residency(Cibrián and Sánchez-Madrid, 2017; Sathaliyawala et al., 2013). CD69 expression is rapidly induced upon stimulation by T cell receptor cross linking or PMA/ionomycin in both CD4+ T cells and the Jurkat T cell line (**Figure S1A-S1B**). Chromatin accessibility maps nominated putative regulatory sites that gain accessibility upon stimulation in primary T cells and Jurkat cells (**Figure 1B**). We refined these predictions using the Enformer model (Avsec et al., 2021) trained on chromatin maps and CAGE-seq data (FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al., 2014)(ENCODE Project Consortium et al., 2020). Genomic intervals corresponding to the promoter, 3’ UTR and an RE located ~4 kb upstream of the TSS were predicted to impact CD69 transcriptional induction (**Figure 1B**).

We used CRISPRi to test the functional impact of the promoter (RE-3), the putative upstream RE (RE-4) and two other sites in the locus that also gained accessibility upon T cell activation (**Figure 1B and Table S1**). We infected Jurkat cells with lentiviral constructs containing KRAB-dCas9 and sgRNAs, selected positive cells, applied PMA/ionomycin stimulation, and measured CD69 surface protein expression by flow cytometry. We found that sgRNAs targeting RE-4 had the strongest suppressive effect on CD69 induction (**Figure 1C and S1C**), while sgRNAs targeting the TSS-proximal RE-3 had a weaker suppressive effect (**Figure 1C and S1C**). RE-4 corresponds to a DNase hypersensitive site bound by multiple TFs that has scored in a luciferase reporter assay and a CRISPR activation screen (ENCODE Project Consortium et al., 2020; Laguna et al., 2015; Mumbach et al., 2017). Whereas chromatin accessibility over RE-4 spans ~1.4 kb, the Enformer model predicted that a specific ~170 bp sequence interval within RE-4 is most critical for CD69 regulation (**Figure 1D and S1D**).

To resolve the functional sequences within these elements and test the Enformer prediction, we designed a library of 101 sgRNAs that tile sequences spanning RE-3 and RE-4 (**Figure S2A-S2B and Table S2**). We reasoned that dCas9 without the repressive KRAB domain would specifically occlude TFs overlapping its target site, and thus affect a narrower interval than KRAB-dCas9 (Dominguez et al., 2015). We infected Jurkat cells with a pooled lentiviral CRISPR library composed of dCas9 and the 101 sgRNAs, selected for puromycin resistance and stimulated with PMA/ionomycin for 5 hours. We then isolated genomic DNA from pre-sorted and sorted CD69- and CD69+ subsets (**Figure S2C**), and amplified the sgRNA-cassettes for sequencing. The relative effect of each sgRNA on CD69 expression was calculated based on its enrichment/depletion in CD69+ relative to CD69- libraries. Multiple sgRNAs within the ~1.7 kb tiled region suppressed CD69 activation (**Figure 1E and S2D**).

To pinpoint individual functional bases in these REs, we complemented the dCas9 tiling with Cytidine Base Editor (CBE) and Adenine Base Editor(ABE) screens. We infected Jurkat cells with lentiviral constructs containing CBE or ABE and the same pool of 101 sgRNAs (**Figure. S2A-S2B, and Table S2**). We stimulated and sorted the cells, and then sequenced the sgRNA-cassettes from pre-sorted, CD69- and CD69+ subsets (**Figure S2C**). Multiple sgRNAs scored in these screens as reducing CD69 activation (**Figure 1F and S2E-S2G**). Notably, the CBE and dCas9 perturbations both pinpointed a ~150 bp interval within RE4 centered at sg#70 as critical for CD69 expression (**Figure 2A**; Chr12:9764860-9765010). This experimentally identified interval closely coincided with the region identified by the deep learning model (**Figure 1D**). Several ABE hits in or near this interval also suppressed CD69 induction, but with lower fold-enrichment, potentially due to reduced effect sizes (**Figure S2G**).

The implicated interval in RE-4 is over-represented for multiple TF motifs relevant to immune function, including GATA, bHLH/Ebox, TCF, ETS and STAT (**Figure 2B**). Notably, a second top scoring interval from the CBE and dCas9 screens, centered at sg#48, showed similar TF motif enrichments (**Figure 2A**; Chr12:9765200-9765310). We scanned the locus for annotated expression quantitative trait loci (eQTLs). However, the implicated RE-4 intervals are highly conserved evolutionarily, devoid of natural variation in the human population, and thus invisible to eQTL analysis (**Figure 2C**)(Võsa et al., 2021). These findings highlight the importance of engineered variants for parsing highly conserved regulatory sequences.

### A single nucleotide artificial variant suppresses CD69 expression by affecting TF competition

We next sought to validate individual base edits and their transcriptional consequences. The top scoring CBE screen sgRNA, sg#70, is predicted to incur C->T transitions at positions 948 and/or 952 within RE-4 (chr12: 9,764,948 and 9,764,952). We infected Jurkat cells with a CBE vector containing either sg#70 or a control sgRNA (sgCtrl) and measured CD69 by flow cytometry (gating strategy in **Figure S3A**). This confirmed that CBE-sg#70 strongly reduced CD69 induction upon stimulation (**Figure 3A and S3B**). We next amplified and sequenced the target region from genomic DNA isolated from Jurkat cells infected with CBE-sg#70(Clement et al., 2019). In unsorted cells, C-948 was replaced by T on ~57.0% of alleles. The proportion of C-948 edited alleles was higher in sorted CD69- Jurkat cells (67.0%) and lower in the CD69+

population (53.6%), consistent with a suppressive effect on CD69 induction (**Figure 3B**). In contrast, edits to the other candidate site, C-952, were less frequent (14.4% at baseline, 16.6% in CD69-, 13.9% in CD69+, **Figure 3B**). These results indicate that the single C-948->T edit strongly impacts transcriptional induction of CD69 in response to stimulation.

We also examined the impact of the C-948 edit on chromatin accessibility. ATAC-seq profiles revealed reduced RE-4 accessibility in cells harboring the CBE-sg#70 construct, relative to CBE controls (**Figure 3C and S3C**). The reduced accessibility was specific to RE-4, as we did not observe any other accessibility changes in the CD69 locus or neighboring genomic regions (**Figure S3D**), nor in the vicinity of other activation associated genes such as CD28 and NR4A1 (**Figure S3E**). Hence, the single base substitution at position C-948 reduces RE-4 accessibility and suppresses CD69 induction in stimulated Jurkat cells.

We next considered the mechanism that underlies the potent effect of this single base mutation. Scanning the region for motifs showed that C-948 directly overlaps a GATA site predicted by the optimized Enformer model  to impact both CD69 expression and element accessibility in Jurkat cells (**Methods; Figure 3D and S3F**). Importantly, the C-948->T edit disrupts a critical position in the GATA motif. The GATA motif is adjacent to a bHLH/Ebox motif that also scores in the Enformer model. We sought to identify specific TFs that are dynamically expressed and likely to bind these respective motifs (**Figure 3E**). GATA3 is highly expressed in Jurkat cells, up-regulated upon stimulation, and broadly implicated in T-cell lineage commitment (Ho et al., 2009). Among bHLH factors, BHLHE40 and BHLHE22 are both highly expressed and strongly induced upon stimulation. BHLHE40 in particular has established roles in T cell differentiation, inflammation and autoimmunity (Cook et al., 2020).

Whereas GATA3 is generally associated with transcriptional activation in T cells, BHLHE40 is a transcriptional repressor (Asanoma et al., 2015; Cook et al., 2020; Emming et al., 2020; Honma et al., 2002; Huynh et al., 2018; Zawel et al.). The close juxtaposition of their cognate motifs suggests that only one factor - either activator or repressor - can bind the implicated site at a given time. We therefore hypothesized that the paired motifs constitute a dynamic regulatory switch that contributes to CD69 induction. The potent suppressive effect of the base edit could then be explained by its ability to displace the GATA3 activator by disrupting its motif, allowing in turn BHLHE40 repressor binding due to relief of steric hindrance.

To investigate this hypothesis, we used ChIP-seq to map GATA3, BHLHE40 and the enhancer-associated histone acetylation mark H3K27Ac. Whereas a strong GATA3 binding peak is evident over RE-4 in stimulated Jurkat cells, binding is lost in CBE-sg#70 infected cells (**Figure 3F**). Remarkably, GATA3 loss in the edited cells is accompanied by broader BHLHE40 binding over RE-4, consistent with a switch in TF binding at the edited site (**Figure 3F**). H3K27ac signal over RE-4 is also reduced in the CBE-sg#70 edited cells, providing further support for the model that a switch from activator to repressor binding suppresses element activity (**Figure 3F**).

### BHLHE40 suppresses gene expression during Jurkat T cell activation via invade regulatory motifs near GATA binding sites

We further tested our model with GATA3 and BHLHE40 loss- and gain-of-function experiments. First, we confirmed that GATA3 knockout suppressed CD69 induction in stimulated Jurkat cells (**Figure S4A**). Next, we investigated GATA3-BHLHE40 antagonism by lentiviral BHLHE40 overexpression. We found that BHLHE40 overexpression suppressed CD69 induction in both control and CBE-sg#70 edited Jurkat cells (**Figure 4A**). However, the magnitude of suppression was greater in the edited cells, potentially due to relief of GATA factor competition. In contrast, overexpression of BHLHE41, the homolog of BHLHE40, had no effect on CD69 expression (**Figure S4B**). We also assessed the impact of BHLHE40 overexpression on chromatin accessibility using ATAC-seq. We found that overexpression reduced RE-4 accessibility, consistent with a direct repressive impact on the element and with our proposed TF switch model (**Figure 4B**).

Further evidence for the importance of interplay between these TFs emerged in our examination of the second interval identified in our dCas9 and CBE screens. Remarkably, the top base edit hit in this interval (sg#48) also incurs a C->T edit that disrupts a GATA motif flanked by a bHLH/Ebox motif (**Figure 2A-2B**). Here again, the respective motifs are too close to permit concurrent binding. Hence, this second hit may also be explained by its impact on competitive binding dynamics between the GATA3 and BHLHE40. This result prompted us to examine whether interplay between these factors plays a more general role in T cell transcriptional responses. We collated all GATA3 bound sites in Jurkat cells that contain a GATA motif and a bHLH/Ebox motif within the corresponding accessible site. We found that BHLHE40 overexpression reduced the aggregate accessibility of these sites (**Figure 4C-4D**), consistent with a global repressive role. Overall, 909 (95.4%) of these GATA3 sites with proximate bHLH/Ebox motifs became less accessible upon BHLHE40 overexpression, while just 44 (4.6%) became more accessible (FDR < 0.2). We also examined the edge-edge distance between the GATA and bHLH/Ebox motifs (**Figure 4E**). Sites that were repressed by BHLHE40 overexpression showed a strong enrichment for motif spacing of 0 to 3 bp, consistent with steric hindrance and competition between factors (FDR < 0.05). In contrast, a similar analysis of sites that were not repressed revealed a preferential spacing of 6 to 9 bp between motifs, consistent with previously reported sites of coordinate GATA and Ebox factor (e.g., TAL1) binding(Sanda et al., 2012). These findings suggest that the precise spacing of competitive or collaborative TF binding motifs is a critical determinant of regulatory element dynamics.

Finally, we considered the influence of the competitive TFs on T cell phenotypes. GATA3 is an established T cell regulator that promotes Th2 over Th1 differentiation(Wan, 2014), and is also implicated in the maintenance of naive T cells (Singer et al., 2017). Although BHLHE40 has also been associated with Th2 responses, we found that BHLHE40 overexpression downregulated multiple immune gene targets involved in Th2 differentiation (IL4R, IL21R, EGR2, etc.) or naive T cell maintenance (CD248, BACH2, etc.)(**Figure 4F**). Gene Set Enrichment Analysis confirmed that BHLHE40 overexpression upregulated Th1 response genes and effector T cell signatures, while down regulating genes associated with Th2 responses or naive T cells(Godec et al., 2016) (**Figure S5**). These results are consistent with a general role of BHLHE40 in restraining GATA3 mediated activation at immune loci. Notably, many of the immune loci subject to opposing regulation contain elements with closely spaced GATA and bHLH/Ebox motifs (0-3 bp), consistent with a general role for competitive TF binding on T cell transcriptional programs and phenotypes (**Figure 4G-4H**). We suggest that competition between repressor and activator poises key immune genes for rapid transcriptional responses, potentially explaining the association of both activator and repressor with T cell stimulation and Th2 phenotypes.

## Discussion

Resolving functional sequences within the vast numbers of putative regulatory elements in the human genome is a critical challenge with exciting potential to unlock an underlying regulatory code. Here we integrate chromatin maps, deep learning, epigenetic editing and base editing to parse sequences that control an exemplar inducible gene in Jurkat T cells. Regulatory base edits clustered in an evolutionarily conserved interval within a CD69 enhancer that was also highlighted by the deep learning model, but more precisely pinpointed critical regulatory sequences. Further characterization of top scoring edits revealed a role for competition between the GATA3 activator and the BHLHE40 repressor in the activation of CD69 upon T cell stimulation. Genomewide analysis suggests that dynamic interplay between these factors plays a general role in immune gene responsiveness and T cell phenotypes. Our study and results emphasize the importance of epigenetic perturbations and artificial sequence variants for characterizing regulatory sequences, which tend to be highly conserved and may be invisible to methods that rely on natural genetic variation.

We also note limitations of our study and approach. Although our pooled screen tested thousands of perturbations, it was limited to one inducible gene locus in one cell model. Extension of the approach to additional immune loci and in primary T cells is an exciting future opportunity. Furthermore, our base editing screen could target only ~12% of nucleotide positions due to the requirement for nearby PAM sites. Critical bases and functional motifs will be missed as a consequence. Base editors with less restrictive PAM site requirements (Rosello et al., 2022) could improve the resolution of future screens. While base editor approaches are mainly focused on C-to-T or A-to-G transitions, prime editors could enable more systematic base changes if they could be applied at scale (Anzalone et al., 2019).

There remains a considerable gap between the throughput of current approaches and the eventual goal of deciphering the regulatory code of the entire human genome. Functional perturbations will need to be combined with computational approaches, such as the deep learning model incorporated here. While the model predictions and experimental data both highlighted similar genomic intervals in our study, the computational approach did not distinguish individual bases or motifs identified by the base editing. Nonetheless, algorithmic improvements, ideally trained in iterative cycles with experimental tests of artificial variants, may ultimately yield sufficiently accurate predictive models to resolve regulatory sequences across the vast noncoding genome.

Our study also highlights competitive interplay between GATA3 and BHLHE40 in the rapid induction of CD69 upon T cell stimulation. Two top scoring C-T base edits that suppress the CD69 response both appear to act by shifting the balance of TF binding from the GATA3 activator to the BHLHE40 repressor. Both cytosine bases and their surrounding regions are highly conserved, invariant in the human population, and hence invisible to QTL mapping studies. Hence, the artificial variants were essential to uncover functional bases, motifs and TF interactions. Interplay between GATA3 and BHLHE40 appears to play a much broader role in poising immune genes for T cell stimulation, with BHLHE40 repressing hundreds of GATA3-bound elements. The precise spacing between the GATA and BHLH/E-box motifs appears critical, with the antagonistic pairs tending to be very closely spaced, consistent with steric hindrance between TFs. Other adjacent motif pairs with wider spacing conducive to concurrent binding may have distinct biochemical properties and regulatory impacts. Thus our study links the well established principles of competitive and cooperative TF binding to specific motifs, functional elements, transcriptional responses and T cell phenotypes.

In conclusion, we have benchmarked emerging experimental and computational strategies to resolve regulatory genomic sequences with increasing precision. Our study demonstrates in particular the potential of base editing screens to identify critical regulatory motifs and TF interactions that underlie rapid and robust transcriptional responses. Further computational and experimental innovations will be needed to scale these approaches and address the daunting challenge of human regulatory genomics.

## Figure Legends

**Figure 1. Integrative analysis of the CD69 regulatory landscape.**
A) Gene regulatory landscape characterization by successive functional assays and deep learning.
B) Genomic tracks depict accessibility of the CD69 locus in primary CD4+ T cells and Jurkat cells, without or with stimulation (PMA/ionomycin). Enformer signal track shows the predicted contribution of underlying sequence to CD69 expression (magnitude of the model gradient at each position with respect to CD69 promoter signal, summed over 128 bp bins)in Jurkat. Grey bars depict regions with differential accessibility in stimulated Jurkat cells, relative to resting (FDR=0.2). CRISPRi sgRNA positions are also indicated. ATAC signal corresponds to reads per genomic content (RPGC).
C) Flow cytometry of CD69 expression in Jurkat cells targeted with the indicated CRISPRi sgRNA following a stimulation time course. Samples gated from the lentiviral transduced population (mCherry+).
D) Expanded view of Enformer signal at single base resolution over RE-4, as denoted in panel b.
E) Enrichment/depletion plot of dCas9 sgRNAs in CD69+ Jurkat cells, relative to CD69- cells (y-axis; $Log_2$ Odds Ratio of normalized sgRNA reads). sgRNAs along the x-axis according to their 5' starting position on the positive strand. Each data point represents mean±s.e.m.
F) Enrichment/depletion plot of Cytidine Base Editor (CBE) sgRNAs in CD69+ Jurkat cells, relative to CD69- cells (as in panel e).
For C,E,F, data represent 2-3 biological independent experiments. A 170 bp region critical for CD69 activation is denoted (D-F, light red).

**Figure 2. A critical sequence interval within RE-4 influences CD69 expression.**
A) Enrichment/depletion plot of sgRNAs in dCas9 and CBE tiling screens as in Fig 1e/f, limited to the central portion of RE-4 with sgRNAs shown to scale. Expected C->T edit positions highlighted for CBE-sgRNAs sg#70 and sg#48 (dashed grey lines).
B) Transcription factor motif locations (grouped by broad motif class) for key immune regulators shown across the same interval as in panel a (FDR<0.05). Dark grey areas represent overlapping motifs.
C) Zoomed out view of the CD69 locus shows CBE sgRNA depletion (red boxes indicate significantly depleted sgRNAs), common SNPs (black vertical stripes), eQTLs (blue vertical stripes)(Võsa et al., 2021) and PhastCon100 conservation score (green stripes).

**Figure 3. Top scoring base edits target competitive TF binding sites.**
A) Flow cytometry plots of CD69 signal for CBE-sgCtrl and CBE-sg#70 Jurkat cells under resting or stimulated conditions. Bar plot depicts the proportion of CD69+ cells in CBE-sgCtrl (grey) and CBE-sg#70 (red) after stimulation. P-value based on unpaired t test, **P<0.01. Data are from 4 independent experiments each with 2-3 technical replicates, mean±s.e.m.
B) Table depicts frequency of incurred base edits in CBE-sg#70 infected Jurkat cells. PCR amplicons from unsorted, CD69- and CD69+ populations were sequenced by Illumina Nextseq500. Consensus sequence is shown along with stacked bars that depict the proportions of cytosine and thymine bases in the sequencing data (numbers indicate percent of alleles with C->T edit). Shaded boxes indicate the sg#70 target sequence.
C) Chromatin accessibility shown over the CD69 locus for stimulated CBE-sg#70 (red) and CBE-sgCtrl (grey) Jurkat cells. Bar plot depicts the mean ATAC-seq signal over RE-4 (TMM normalized counts per million; CPM). P-value based on unpaired t test, *P<0.05. Data are from 3 replicates, mean±s.e.m.

D) Enformer signal (letter height) for the sg#70 target region indicates the predicted impact of each base on RE-4 accessibility. The sgRNA directly coincides with a GATA motif and a bHLH/E-box motif, and incurs an edit that disrupts the former (vertical dashed line).
E) Volcano plot depicts gene expression fold-change (x-axis) and significance (y-axis) for TF genes in stimulated Jurkat cells, relative to resting cells. Labels identify differential GATA (red) and bHLH/Ebox (blue) family members.
F) Genomic tracks for the CD69 locus depict chromatin accessibility (ATAC), H3K27 acetylation (H3K27ac), GATA3 binding and BHLHE40 binding in CBE-sgCtrl (grey) and CBE-sg#70 (red) Jurkat cells. Y-axis represents the -$\log_{10}$(p-value) to input controls.
Jurkat cells in a, b, c and f were stimulated with PMA/ionomycin for 2 hours.

**Figure 4. GATA3-BHLHE40 competition impacts global T cell transcriptional responses.**
A) Flow cytometry plots of CD69 signal for stimulated Jurkat cells transduced with CBE-sg#70 and a BHLHE40 overexpression construct (BHLHE40-OE), or with corresponding controls (sgCtrl and Ctrl-LV, respectively). Bar plot depicts the proportion of CD69+ cells in each condition. P-value based on unpaired t test, ****$P<0.0001$. Data are from 3 independent experiments with 2-3 technical replicates, mean±s.e.m.
B) Chromatin accessibility in the CD69 locus for CBE-sg#70 Jurkat cells transduced with either BHLHE40 overexpression lentivirus (light blue) or control (grey). Cells were stimulated with PMA/ionomycin. P-value based on unpaired t test without multiple testing correction, *$P<0.05$. Bar plot data are from 2 replicates, mean±s.e.m ATAC-seq signal over RE-4 (TMM normalized CPM).
C) Plot depicts aggregate accessibility (y-axis) for GATA3 bound sites that also harbor bHLH/E-box motifs (centered on the motifs). Data shown for stimulated Jurkat cells transduced with either BHLHE40 overexpression lentivirus (light blue) or control (grey).
D) For the set of GATA3 bound sites with bHLH/E-box motifs in c, volcano plot depicts fold-change (x-axis) and significance (y-axis) of chromatin accessibility in Jurkat cells transduced with BHLHE40 overexpression lentivirus, relative to control. Differentially accessible sites (FDR < 0.1) are indicated in red.
E) For differentially accessible sites in d, histogram shows the number of sites (y-axis) with the indicated spacing (x-axis) between GATA and bHLH/E-box motifs. Sites are stratified by whether their accessibility is reduced (red) or increased (grey) in the BHLHE40 overexpressing cells. Sites with significant peak differential between reduced and increased accessibility (FDR < 0.05) are denoted (*).
F) Heatmap shows differentially expressed genes with BHLHE40 binding in their REs in BHLHE40 overexpressing Jurkat cells, relative to control. Cells were stimulated with PMA/ionomycin.
G-H) Genomic views of the IL21R(F) and CD248(G) loci show ChIP-seq data for H3K27Ac, BHLHE40, and GATA3 in stimulated Jurkat cells (signal corresponds to P-value enrichment over input). Accessibility (ATAC) and expression (RNA-seq) are also shown for BHLHE40 overexpressing Jurkat cells and controls. Sites with combined GATA and bHLH/Ebox motifs are indicated (pink shade).
Jurkat cells in b-f were stimulated with PMA/ionomycin for 2 hours.

## Star Methods

### KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| **Antibodies and Dyes** | | |
| Brilliant Violet 510™ anti-human CD69 Antibody | Biolegend | Cat#310936<br>RRID:AB_2563834 |
| APC anti-human CD69 Antibody | Biolegend | Cat# 310910<br>RRID:AB_314845 |
| Zombie NIR™ Fixable Viability Kit | Biolegend | Cat# 423106 |
| H3K27Ac antibody | Active Motif | Cat# 39133<br>RRID:AB_2561016 |
| GATA-3 (D13C9) XP Rabbit mAb | Cell Signaling Technology | Cat# 5852,<br>RRID:AB_10835690 |
| Dec1 Antibody | Novus Biologicals | Cat#NB100-1800 |
| Cell lines and primary cells | | |
| Jurkat cell line, Clone E6.1 | ATCC | Cat#TIB152,<br>RRID:CVCL_0367 |
| CD4+ T cells | AllCells | NA |
| **Chemicals and Buffers** | | |
| Phorbol 12-myristate 13-acetate | Sigma-Alrich | Cat#P8139 |
| Ionomycin calcium salt | Sigma-Alrich | Cat# I0634 |
| Chloroquine diphosphate | Millipore Sigma | Cat#C6628 |
| Polybrene | Sigma-Alrich | Cat#107689 |
| Brilliant Staining Buffer | BD | Cat#566349 |
| **Recombinant DNA** | | |
| KRAB-dCas9-sgRNA-Puro | Broad GPP | pXPR_066 |
| LentiCRISPR v2-dCas9 | Addgene | Cat#112233<br>RRID:Addgene_112233 |
| rApobec-nCas9-UGI-Puro | Broad GPP | pRDA_256 |
| EFS-ABE8e-V106W-nCas9-puro | Broad GPP | pRDA_426 |
| BHLHE40-Overexpression-GFP | In this study | |
| BHLHE41-Overexpression-GFP | OriGene | CAT#:<br>RC206882L2 |
| GATA3-KO CRISPR-Cas9-GFP | In this study | |

| | | |
|---|---|---|
| psPAX2 | Addgene | Cat#12260<br>RRID:Addgene_12260 |
| pMD2.G | Addgene | Cat#12259<br>RRID:Addgene_12259 |
| Lentivirus packing | | |
| Lipofectamine 3000 Transfection Reagent | ThermoFisher | Cat#L3000001 |
| OptiMEM | ThermoFisher | Cat# 31985070 |
| **ATAC-Seq reagents** | | |
| Illumina tagmentation kit | Illumina | Cat#20034197 |
| Nextera XT Index Kit | Illumina | Cat# FC-131-1001 |
| MinElute Reaction Purification Kit | QIAGEN | Cat#28003 |
| MinElute PCR Purification Kit | QIAGEN | Cat#28004 |
| NEBNext High-Fidelity 2X PCR Master Mix | NEB | Cat# M0541 |
| **RNA-Seq reagents** | | |
| QIAGEN RNeasy Micro kit | QIAGEN | Cat# 74004 |
| Dynabeads mRNA Direct Kit | ThermoFisher | Cat# 610.12 |
| RNA Fragmentation Reagents | ThermoFisher | Cat# AM8740 |
| Turbo DNase | ThermoFisher | Cat#AM2238 |
| FastAP enzyme | ThermoFisher | Cat# EF0651 |
| Dynabeads MyOne Silane | ThermoFisher | Cat# 37002D |
| T4 RNA ligase | NEB | Cat#M0204L |
| AffinityScript RT Enzyme | Agilent | Cat#600107 |
| Phusion Master Mix | NEB | Cat# M0531L |
| AMPure XP Beads | Beckman Coulter | Cat# B23318 |
| IDT indexes | IDT | NA |
| **ChIP-Seq reagents** | | |
| Protein G beads | ThermoFisher | Cat#10003D |
| RNAse | Roche | Cat#11119915001 |
| Proteinase K | Invitrogen | Cat# 25530-015 |
| DNA end-repair kit | Epicenter Biotech | Cat# ER0720 |
| Klenow Fragment | NEB | Cat# M0212L |
| Quick Ligation kit | NEB | Cat# M2200S |
| PFU Ultra II HS 2x Master Mix | Agilent | Cat# 600850-51 |
| **Amplicon-Seq Reagents and Primers** | | |
| QIAamp DNA Micro Kit | QIAGEN | Cat#6304 |
| DNeasy Blood and Tissue Kit | QIAGEN | Cat#69504 |
| Titanium® Taq DNA Polymerase | Takara | Cat# 639208 |
| Agencourt AMPure XP SPRI beads | Beckman Coulter | Cat# A63880 |
| P5 Primer for tiling:<br>AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCT TCCGATCT<br><br>TTGTGGAAAGGACGAAACACCG | IDT | NA |

| | | |
|---|---|---|
| P7 Primer for tiling:<br>CAAGCAGAAGACGGCATACGAGATNNNNNNNNGTGACTGGAGTTCAGAC<br>GTGTGCTCTTCCGATCTCCAATTCCCACTCCTTTCAAGACCT | IDT | NA |
| sg#70 amplicon F-primer:<br>GGTGAGACGTCAGAAAGGAAGT | IDT | NA |
| sg#70 amplicon R-primer:<br>GGTGAGACGTCAGAAAGGAAGT | IDT | NA |
| **Software and algorithms** | | |
| CRISPResso2 | (Clement et al., 2019) | http://crispresso.pinellolab.org/submission |
| eQTLGEN | (Võsa et al., 2021) | https://eqtlgen.org/cis-eqtls.html |
| Python v3.9 | | https://www.python.org/downloads/release/python-390/ |
| R v4.2 | | https://www.r-project.org/ |
| Bioconductor v3.15 | | https://www.bioconductor.org/ |
| DESeq v2 | Anders and Huber 2010 | https://bioconductor.org/packages/release/bioc/html/DESeq2.html |
| CSAW | Lun et al. 2016 | https://bioconductor.org/packages/release/bioc/html/csaw.html |
| ComplexHeatmap | Gu et al. 2016 | https://bioconductor.org/packages/release/bioc/html/ComplexHeatmap.html |
| Enformer v1 | Avsec et al. 2021 | https://tfhub.dev/deepmind/enformer/1 |
| Fine-tuning code for this paper | | https://github.com/BernsteinLab/BE_CD69_paper_2022 |
| DeepTools 3.5.0 | Ramirez et al. 2016 | https://github.com/deeptools/deepTools |
| TFModisco 0.4.2.3 | Shrikumar et al. 2018 | https://github.com/kundajelab/tfmodisco |
| MEME suite v5.4.1 | Bailey et al. 2015 | https://meme-suite.org/meme/ |

| | | |
|---|---|---|
| ENCODE ATAC-seq pipeline v 2.1.3 | | https://github.com/ENCODE-DCC/atac-seq-pipeline |
| ENCODE ChIP-seq pipeline v 2.1.6 | | https://github.com/ENCODE-DCC/chip-seq-pipeline2 |
| STAR v2.7.9a | Dobin et al. 2013 | https://github.com/alexdobin/STAR |
| Salmon v1.6 | Patro et. al 2017 | https://github.com/COMBINE-lab/salmon |
| Bedtools v2.30.0 | Quinlan et al. 2010 | https://github.com/arq5x/bedtools2 |
| Samtools v1.12 | Li et al. 2009 | https://github.com/samtools/samtools |
| **Datasets** | | |
| Jurkat ATAC-seq, wild-type | Nasser et al. 2021 | GSE155555 |
| CD4+ T-cell ATAC-seq | | GSE124867 |
| Jurkat RNA-seq, wild-type | Brignall et al. 2017 | GSE90718 |
| Jurkat RNA-seq, ATAC-seq, edited | this paper | GSE206377 |

## Resource Availability

### Lead contact

**Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Bradley E. Bernstein (bradley_bernstein@dfci.harvard.edu).**

### Material Availability

Base-editor construct will be available on addgene upon publication. sgRNA library request will be directed to Fadi J. Najm(fadinajm@broadinstitute.org)

### Data and code accessibility

For Jurkat and CD4+ T cell ATAC-seq datasets, we adapted GSE155555(Nasser et al., 2021) and GSE124867 for accessibility analysis on the CD69 loci. For the wild-type Jurkat RNA-Seq, we adapted GSE90718. ATAC-Seq data for 1) CBE+sgCtrl and CBE+sg#70, 2) Ctrl-GFP and BHLHE40-GFP, 3) CBE+sg#70+Ctrl-GFP and CBE+sg#70+BHLHE40-GFP, as well as RNA-Seq data for Ctrl-GFP and BHLHE40-GFP were generated for this study and are available at GSE206377.
Analysis code and custom scripts are available on github at https://github.com/BernsteinLab/BE_CD69_paper_2022.git.

## Method Details

### Guide library design and cloning

Pooled libraries for expression of sgRNAs were generated as detailed previously (Joung et al Nature Protocol 2017). Briefly, DNA oligos were annealed into double stranded fragments with compatible overhangs and ligated into BsmBI sites into vectors. Vector backbones were

CRISPRi+guide puro (pXPR_066, Broad GPP), lentiCRISPR v2-dCas9 (gift of Thomas Gilmore, Addgene 112233), rApobec-nCas9-UGI-puro (pRDA_256, Broad GPP) and EFS-ABE8e-V106W-nCas9-puro(pRDA_426, Broad GPP). Libraries were then transformed by electroporation into electrocompetent coli (Invitrogen) and spread onto bioassay plates. Bacterial colonies were harvested and isolated using the Plasmid Plus Midi Kit (Qiagen). Four putative regulatory, ATACseq accessible regions were identified near the CD69 locus. Peak proximity and acceptable on-target efficacy scores(Doench et al., 2016) determined sgRNA selection for the CRISPRi tests. After RE3 and RE4 were identified, all sgRNAs possible in these peak regions were selected and included for screening with dCas9 and base editors and can be found in Table S2.

**Cell culture and stimulation**

The Jurkat cell line (ATCC, Clone E6.1, TIB152) was cultured in complete RPMI (RPMI Medium 1640,Gibco, 11875085, 1% Penicillin-Streptomycin, Gibco, 15140122, 10% Heat Inactivate Fetal Bovine Serum, Peak Serum, 20mM HEPES,Gibco,15630080,1% Sodium Pyruvate, Gibco, 11360070, and 1% NEAA,Gibco, 11140050) at a maximum density of 2 X $10^6$ cells/ml in 25 cm or 75 cm cell culture dishes. Stimulation of Jurkat cells for 2-7 hour experiments was achieved with 50ng/ml Phorbol 12-myristate 13-acetate (PMA, Sigma-Alrich, P8139) and 500ng/ml ionomycin calcium salt from Streptomyces conglobatus (ionomycin, Sigma-Alrich, I0634).

Cryopreserved CD4$^+$ T cells isolated from healthy donors were obtained from AllCells. On the day of stimulation, cells were thawed in RPMI 1640 medium supplemented with 2mM L-glutamine and 50% FBS, counted and resuspended in TexMACS medium (Miltenyi Biotec) supplemented with 20 IU/mL human Interleukin-2 (IL-2) and 1% penicillin-streptomycin. Cells were seeded at 1 million cells per well in a 48-well plate. Cells were either left untreated or stimulated with 10 µL T Cell TransAct™, human (Miltenyi Biotec) via CD3 and CD28 for 24hrs.

**Lentivirus production**

293T cells approaching 70-80% confluency in 10 cm cell culture dishes were used for packaging. Cells were pre-treated with 25 uM chloroquine diphosphate (Millipore Sigma, C6628) in 3 ml of complete DMEM (Gibco DMEM with 1% Penicillin-Streptomycin and 10% Heat Inactivate Fetal Bovine Serum) and incubate in the 37°C and 5% CO2 incubator for more than 30 minutes. Lipofectamine 3000 Transfection Reagent (ThermoFisher, L3000001) was used to deliver plasmids into 293T cells. Briefly, 15 ug lentiviral vector plasmid, 15 ug of psPAX2 and 5 ug pMD.G plasmid were vortexed with 40 ul P3000 reagent in 1.5 ml OptiMEM (ThermoFisher, 31985070). Then 40 ul Lipofectamine was added to 1.5 ml OptiMEM and briefly vortexed. The two OptiMEM solutions were combined and mixed well by vortexing for 30s and incubated at room temperature for at least 20 minutes. Carefully, the OptiMEM mixture was added dropwise to 293T cells and incubated in a 37°C and 5% $CO_2$ incubator for 6 hours. Media were aspirated and replaced with 5ml of fresh complete RPMI. Lentiviral supernatant was harvested between 24 hours and 48 hours after transfection.

**Lentivirus tranduction**

Jurkat cells were resuspended in 1ml media and seeded at a density of 2-5 x $10^5$ cells per well of a 12-well plate. Lentiviral supernatant was supplemented with 8 ug/ml polybrene (Sigma-Alrich) added to the Jurkat cells. The plate was then centrifuged at 2000xg, 32°C for 60mins. Cells were then incubated at 37°C and 5% $CO_2$ overnight and changed into complete RPMI on the next day. For GFP+ or mCherry+ marked lentivirus, cells were sorted or analyzed 5 days after transfection via flow cytometry. For blasticidin selection, 5 ug/ml of blasticidin was added to the transduced cells and selected for 14 days. For puromycin selection, 5 ug/ml of puromycin was added to the transduced cells and selected for 3 days.

## Flow cytometry and sorting

Suspended cells were centrifuged down at 300xg, room temperature for 5 minutes. The cells were stained with the antibody cocktail in the staining buffer of a 1:1 mix of PBS and Brilliant Staining Buffer (BD, 566349), at room temperature for 20 mins or at 4°C for 30-40 mins. Cells were washed once in PBS with 1% FBS and then resuspended in the same buffer. Flow cytometry or FACS was processed on either BD LSRFortessa X-20 or SONY SH800 following the manufacturing instructions. Antibodies and dyes used from Biolegend: Brilliant Violet 510™ anti-human CD69 Antibody (310936)；APC anti-human CD69 Antibody (310910); Zombie NIR™ Fixable Viability Kit (423106).

At least 2 x $10^5$ CRISPR library infected Jurkat cells were collected as a pre-sorted baseline. 2-4 x $10^6$ CRISPR library infected Jurkat cells were resuspended in 2 ml of complete RPMI and stimulated with 50 ng/ml PMA and 500 ng/ml ionomycin for 5 hours, and then processed for FACS as described above. Sorted CD69- and CD69+ populations were collected for genomic DNA isolation.

## Genomic DNA isolation and sequencing

Genomic DNA (gDNA) was isolated using QIAamp DNA Micro Kit (QIAGEN, 6304) or DNeasy Blood and Tissue Kit (QIAGEN, 69504) according to the manufacturer's protocol. The gDNA concentrations were quantified by Qubit. For PCR amplification, at least 330 ng of gDNA was used per reaction for greater than 500-fold library coverage. Each reaction contained 1.5 ul Titanium Taq (Takara), 10 µl of 10× Titanium Taq buffer, 8 µl deoxyribonucleotide triphosphate provided with the enzyme, 5 µl DMSO, 0.5 µl P5 stagger primer mix (stock at 100 µM concentration), 10 µl of a uniquely barcoded P7 primer (stock at 5 µM concentration), and water up to 100ul.
P5 Primer: AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCT
TCCGATCT
TTGTGGAAAGGACGAAACACCG
P7 Primer: CAAGCAGAAGACGGCATACGAGATNNNNNNNNGTGACTGGAGTTCAGAC
GTGTGCTCTTCCGATCTCCAATTCCCACTCCTTTCAAGACCT

PCR cycling conditions included: an initial 5 min at 95°C; followed by 30 s at 54°C, 30 s at 53°C, 20 s at 72°C, for 28 cycles; and a final 10-min extension at 72°C. PCR primers were synthesized at Integrated DNA Technologies. PCR products were purified with Agencourt AMPure XP SPRI beads according to the manufacturer's instructions (Beckman Coulter, A63880). Samples were sequenced on a MiSeq (Illumina). Reads were counted by alignment to a reference file of all possible guide RNAs present in the library. The read was then assigned to a condition on the basis of the 8-nt index included in the P7 primer.

## Amplicon sequencing

To assess base editing frequency of the sg#70 locus, we designed primers flanking this region resulting in a 214bp product. Forward primer: GGTGAGACGTCAGAAAGGAAGT and reverse primer: AATTCACCCACTGAAAGGAAAA. Amplicons were next ligated with Illumina Truseq adaptors, cleaned and size selected with AMPure XP SPRI beads, and sequenced on a MiSeq paired end run. FASTQ files were processed with CRISPResso2 v2 with standard settings for base editor(Clement et al., 2019).

## ATAC-Seq experimental processing

ATAC-Seq lysis buffer contains 10mM Tris-HCl (pH=7.4), 10mM NaCl, 3mM MgCl2, 0.1% Tween-20, 0.1% NP40, 0.1% Digitonin, 1% BSA and topped up with ddH2O. ATAC-Seq

washing buffer contains 10mM Tris-HCl (pH=7.4), 10mM NaCl, 3mM MgCl2, 1% BSA and topped up with ddH2O.
5 x 10^4 cells were centrifuged down with the resuspension buffer (PBS with 1%BSA) in a low-binding eppendorf tube at 4℃, 500xg for 5 mins. Each pellet is resuspended with 50 ul of lysis buffer and incubated on ice for 5 minutes. 50 ul of wash buffer was added to the lysis buffer containing nuclei and centrifuged down at 4℃, 500xg for 5 minutes. The supernatant is then removed and 50ul resuspension buffer is added to the tube without disturbing the pellet. Nucleus are then centrifuged down at 4℃, 500xg for 5minutes. Tagmentation of the genome DNA is processed using the Illumina tagmentation kit (20034197) for 30 mins in 37℃. Fragmented products are then isolated via MinElute Reaction Purification Kit (QIAGEN, 28003) according to the manufacturer's instructions. Illumina Nextera XT SetA indexes and NEBNext High-Fidelity 2X PCR Master mix (NEB, M0541) are used to amplify the fragmented products of each sample, with 12 PCR cycles of 98°C-10s, 63°C-30s and 72°C-1min. PCR products are then isolated via MinElute PCR Purification Kit (QIAGEN, 28004) following manufacturer's instructions.

## RNA-Seq experimental processing

Whole RNA was extracted from over 1 × 10^5 cells using the QIAGEN RNeasy Micro kit (QIAGEN, 74004) according to the manufacturer's instructions. 1ug RNA was then used to prepare the RNA-Seq library. Poly-A+ RNA is enriched using Dynabeads mRNA Direct Kit (ThermoFisher, 610.12) according to the manufacturer's instructions and eluted in 18ul Tris-HCl buffer(pH=7.4). Zinc fragmentation are processed using RNA Fragmentation Reagents(ThermoFisher, AM8740), followed by Turbo DNase (ThermoFisher, AM2238) and FastAP enzyme (EF0651) treatment. Then the fragmented RNA are cleaned-up using Dynabeads MyOne Silane (ThermoFisher, 37002D) and eluted in 7ul of nuclease-free water. Next, RNA-adaptors are ligased to eluted RNA using T4 RNA ligase (NEB, M0204L) at 23℃ for 1 hour and adaptor-ligated RNA was cleaned-up using Dynabeads MyOne Silane and eluted in 13.5ul of nuclease-free water. First strand of cDNA is synthesized using AffinityScript RT Enzyme (Agilent, 600107) according to the manufacturer's instructions at 54℃ for 1 hour. First-strand cDNA was cleaned-up using Dynabeads MyOne Silane and eluted in 5.5ul of nuclease-free water, followed by cDNA adaptor ligation. After another round of clean-up, the adaptor-ligated cDNA was processed to library PCR amplification using Phusion Master Mix (NEB, M0531L) with IDT adaptor indexes. The final library was cleaned-up with AMPure XP Beads (Beckman Coulter, B23318) to a final size around 280bps.

## ChIP-seq experimental processing

Jurkat cells were pelleted (2.5X10^7 per sample) and fixed using 1% formaldehyde at 37℃ for 10 mins then quenched by glycine. Samples were next washed with cold PBS+proteinase inhibitor (ThermoFisher, 78429), resuspended in lysis buffer (1% SDS, 0.25% DOC, 50mM Tris-HCl, pH=7.4), and incubated on ice for 10 mins. Samples were diluted up to 1ml in eppendorf using ChIP dilution buffer (0.01% SDS, 150mM NaCl, 0.25% Triton, 50mM Tris-HCl, pH=7.4) and sonicated using a Covaris E220, with the following settings: 24 mins with 5% duty factor, 140W max power and 200 cycles/burst. Each sample was then split into 4 eppendorf tubes: 1) 20ul, top up to 200ul for input; 2)180ul, top up to 1ml for H3K27Ac ChIP (2.5ul, Active Motif, 39133); 3) 400ul, top up to 1ml for GATA3 ChIP (10ul, CST-D13C9, 5852); 4) 400ul, top up to 1ml for BHLHE40 ChIP (10ul, Novus Biological, NB100-1800). The tubes were incubated overnight at 4℃ on a rotator.
On the next day, Protein G beads (ThermoFisher, 10003D) were washed and added to the antibody-containing suspension and rotated at 4℃ for 2 hours. The beads were then washed with ice-cold RIPA wash buffer: RIPA-500, LiCl, and 10mM Tris-HCl buffer (pH=8.5). The beads were eluted in a wash buffer (10mM Tris-HCl, pH=8.0, 0.1% SDS, 150mM NaCl, 5mM DTT) and

incubated at 65℃ on a shaker for 1 hour. Samples were then treated with RNAse (Roche, 11119915001) at 37℃ for 30 mins and then with proteinase K (Invitrogen, 25530-015) at 63℃ for 3 hours. AMPure XP Beads (Beckman Coulter, B23318) were used to purify the DNA fragments from the samples. Eluted fragments were then processed for DNA end-repair (Epicenter Biotech, ER0720), Klenow A base adding (Klenow from NEB, M0212L), adaptor ligation (Ligase from NEB, M2200S) and PCR amplification (PFU Ultra II HS 2x master mix from Agilent, 600850-51) according to manufacturer's protocols. Index primers were ordered from Integrative DNA Technology. PCR was set up with the following conditions: 2 mins for 95℃; 30 sec at 95℃, 30 sec at 55℃,30 sec at 72℃ for 16 cycles; 1 min at 72℃. PCR products were purified using AMPure XP Beads with a final size of around 300 bps.

**Enformer predictions and fine-tuning**

The published Enformer model without any modifications was downloaded from https://tfhub.dev/deepmind/enformer/1. For model fine-tuning, we loaded the model checkpoint made available by the authors at https://github.com/deepmind/deepmind-research/blob/master/enformer/enformer-training.ipynb. The cell-type/organism specific heads in the original model were then replaced with two untrained dense layers, corresponding to ATAC-seq from resting and stimulated Jurkat T-cells not in the original training data. These data were first converted to RPGC normalized bigwigs using DeepTools and then converted the required model input format using the scripts publicly available at https://github.com/calico/basenji. The modified model was then trained on a Google Cloud TPU-VM v3-64 pod-slice using a multi-learning rate scheme. The original model trunk, consisting of all convolutional and transformer layers shared for all organisms/tracks was trained using the AdamW optimizer from the tensorflow addons library at a learning rate of 1.0e-05 and weight decay of 1.0e-05. The two added output heads were trained at a higher learning rate of 5.0e-03 and weight decay of 5.0e-02. The model was trained for 52 epochs, with checkpointing every 8 epochs, and training was stopped when the validation loss did not decrease by 1.0e-03 relative to the lowest recorded validation loss for 30 epochs. The best checkpointed model was chosen at epoch 24 which reached a validation pearson's correlation of 0.8021 and 0.7859 for stimulated and resting Jurkat T respectively.

Model interpretation was conducted as described at https://github.com/deepmind/deepmind-research/blob/master/enformer/enformer-usage.ipynb. For CAGE-seq interpretation, we calculated the gradient of the model for unstimulated Jurkat T-cells with respect to the predicted CAGE-seq signal at the CD69 promoter. This was achieved by centering a 393216 bp genomic window within the CD69 promoter(chr12:9760820-9760903) and computing the gradient for human output head # 4831 with respect to output bins 446-450.The absolute value of the gradients were then summed in 128bp bins for coarse grain resolution (**Fig 1D**). A similar approach to nominate bases contributing to RE-4 accessibility was adopted to obtain the base resolution contribution scores for the fine-tuned model corresponding to Figure S2 and 3. For this analysis, the window was centered around RE-4(chr12:9764300-9765900) and the gradient was computed with respect to output bins 442-454(**Fig S1D, 3D).**

For identifying TF motifs using Enformer base importance scores (**Fig S3**), we used the TFModisco suite (Shrikumar et al., 2018). This tool clusters short stretches of bases using base importance scores to discover motifs that can then be matched to known databases. First we centered 393216 bp genomic windows as above at the promoter of each of 2195 genes that were differentially expressed(FDR=0.01, see RNA-seq processing below) between resting and stimulated Jurkat cells. Then, we computed the gradient of the model at each base within the window for output head 4831 as above with respect to the CAGE-seq signal at the promoter, corresponding to bins 446-450. For each window, we also computed the model gradient on a dinucleotide shuffled version of the sequence which was averaged across all genes in order to obtain an empirical null distribution of gradients. In order to reduce computing time, we extracted model gradients, sequence, and null gradients for the 750 centered bp window centered at each ATAC-seq peak detected from unstimulated Jurkat cells. Predictions were run in parallel across all genes simultaneously using a custom WDL/Google Cloud script. Finally, hypothetical contribution scores at each position within the 750 bp input window were computed as the model gradient corresponding to each non-reference base. TFmodisco was then used with default settings in order to identify putative regulatory motifs. Candidate seqlets were then matched to HOCOMOCO v11 motifs(Kulakovskiy et al., 2018) using Tomtom from the MEME-suite V5.4.1(Bailey et al., 2015).

**ATAC-seq data processing, differential accessibility analysis, and coverage tracks**
All ATAC-seq data were aligned and processed using the ENCODE uniform ATAC-seq processing pipeline v2.1.3 available at https://github.com/ENCODE-DCC/atac-seq-pipeline. The pipeline was configured to use default parameters, adapter auto-detection, the bowtie2 aligner (Langmead and Salzberg, 2012), and MACS2 (Zhang et al., 2008) for peak calling. GRCh38 V29 and associated mitochondrial genomes and blacklists were obtained from https://www.encodeproject.org/references/ENCSR938RZZ/. Differential accessibility analysis was conducted with the CSAW package v 1.28(Lun and Smyth, 2016). Briefly, a consensus set of peaks was obtained from the union of peaks across all input samples/replicates. Peaks lying within blacklist regions and with low signal(less than -3 $\log_{10}$CPM) were removed. Reads were counted in 300 bp windows genomewide and merged to a maximum width of 5 kb. Finally, counts were TMM normalized and significant differentially accessible peaks were identified based on a genome-wide (for global analysis) or local window ( <= 1 Mb) FDR correction.

ATAC-seq tracks in all figures were computed by pooling replicates where applicable using samtools (Li et al., 2009) and creating signal tracks using DeepTools bamcoverage (Ramírez et al., 2016). Signal tracks were normalized using the reads per genomic bin normalization (RPGC) options in DeepTools with the pre-computed effective genome size for GRCh38 in order to create coverage bigwigs. Average ATAC-seq coverage profiles over bHLH/E-box sites were obtained by taking genome-wide motif-scans (see motif analysis below) and filtering to keep bHLH(Ebox) motif sites overlapping the union of ATAC-seq peaks between the BHLHE40-OE and BHLHE40-WT conditions. DeepTools plotProfile function with a 4kb window centered at each bHLH(Ebox) motif was then used to compute the mean coverage profile for each condition.

**RNA-seq data processing, differential expression analysis, and gene-set enrichment**
RNA-seq datasets were processed using a custom pipeline utilizing fastp v0.23.2((Chen et al., 2018) for automatic adapter trimming with default settings for paired-end datasets, and the STAR aligner (Dobin et al., 2013) v2.7.9a with default GTEx (GTEx Consortium, 2020) settings obtained from https://github.com/broadinstitute/gtex-pipeline. Gene quantifications were obtained using Salmon v1.6 (Patro et al., 2017) and the GENCODE V38 annotation (Frankish et al., 2021) with the seqBias, gcBias, posBias, and validateMappings flags enabled.

Differential expression analysis between conditions was conducted using DESeq V2 (Anders and Huber, 2010) with default settings. Log fold change values were corrected with the lfcShrink option using the apeglm method. For BHLHE40-OE gene expression analysis, BHLHE40-OE was compared to wild-type only in the stimulated case. Unless otherwise stated, significance is based on an FDR=0.05 cutoff.

Heatmap was constructed using the Complex heatmap package v 2.12.0 (Gu et al., 2016). Genes were subsetted to only keep those differentially expressed between BHLHE40-OE and BHLHE40-WT at FDR=0.05, and those with a nearby(< 50 kb peak-promoter distance) BHLHE40 ChIP-seq peak(see ChIP-seq processing below).

Gene-set enrichment analysis was conducted using the FGsea package v 1.22.0 (Korotkevich et al., 2021) and the ImmuneSigDB gene sets (Godec et al., 2016). The product of -$\log_{10}$(p-value) and $\log_2$FoldChange was used as the ranking metric for input to gene set enrichment. Significant pathways were collapsed using the collapsePathways function from FGsea with default settings. Normalized enrichment scores for enriched, down-regulated gene sets (identified as those containing the suffix _DN) were multiplied by -1.

## ChIP-seq data processing

ChIP-seq read alignment, quality filtering, duplicate marking and removal, peak calling, signal generation, and quality-control was conducted using the ENCODE ChIP-seq pipeline v2.1.6 available at https://github.com/ENCODE-DCC/chip-seq-pipeline2. GRCh38 V29 and blacklists were obtained from https://www.encodeproject.org/references/ENCSR938RZZ/. In brief, reads were aligned to the GRCh38 genome using bowtie2(-X2000), filtered to remove poor quality reads(Samtools) and de-duplicated(Picard MarkDuplicates). Both histone and TF peaks were then called using MACS2(Zhang et al., 2008). All datasets were assessed for enrichment quality and replicate concordance using the included ENCODE ChIP-seq quality control pipeline. H3K27ac ChIP-seq datasets passed all ENCODE QC standards and were not further processed beyond obtaining signal tracks as described below. For GATA3 ChIP-seq in the unedited cells (sgCtrl-P258), we noted a small number of peaks and therefore increased the MACS2 q-value cutoff to 0.05 and max number of peaks to 500000 to improve sensitivity. Consensus peak sets were then obtained using IDR analysis between replicates with an FDR cutoff of 0.05. De-novo motif discovery using the XSTREME (Grant and Bailey, 2021) program from the MEME-suite(Bailey et al., 2015) yielded the expected GATA motif among the top discovered motifs. A final peak set for sgCtrl-P258 was then obtained by retaining peaks that contained the expected GATA motif, and that overlapped open chromatin regions as defined by ATAC-seq. For BHLHE40-P258(sgCtrl), replicates showed poor concordance and the replicate(replicate 1) with the greater number of peaks was selected for further analysis. A peak set was obtained replicate 1 was obtained by taking the intersection of peaks between pseudo-replicates for this dataset. Motif discovery for this peak set yielded the expected e-box motif(CANNTG) among the top recovered motifs. As with GATA3, only peaks overlapping with ATAC-seq peaks and which contained an E-box/bHLH motif were retained. Target genes for each factor were then nominated by computing the set of genes with an annotated TSS within 50kb of a called peak using bedtools closest.

## Motif Analysis

To avoid motif redundancy, we obtained non-redundant motif cluster definitions and the corresponding PWMs from https://resources.altius.org/~jvierstra/projects/motif-clustering-v2.0beta/ (Vierstra et al., 2020). Motif scan of the RE-4 region corresponding to chr12: 9764556 - 9765505 was conducted by extracting the regions genomic sequence using bedtools getfasta (Quinlan and Hall, 2010) and scanned using the MOODs motif scanner v1.9.4.1 (Korhonen et al., 2009) with a p-value cutoff of 0.0001 and background base probabilities of 2.977e-01 2.023e-01 2.023e-01 2.977e-01. We filtered to keep motifs matched with MOODs score > 4, and further clustered motifs based on whether the motif cluster name contained GATA, bHLH, TCF, ETS, NFKB, NFAT, CREB, or STAT.

For motif spacing analysis, genome-wide motif scans were obtained from https://www.vierstra.org/resources/motif_clustering. Each GATA3 ChIP-seq peak(see ChIP-seq analysis) was intersected with the set of consensus ATAC-seq peaks for both the BHLHE40-OE and WT stimulated Jurkat samples. For the remaining peaks, the highest MOODs score GATA motif located within the middle ⅓ of the peak was chosen as the representative GATA motif. The closest bHLH/E-box motif was then located using the bedtools closest tool with the -t all and -d options enabled. The motif-motif distance was then computed as the edge-edge distance between the central GATA motif and the core CANNTG E-box motif(i.e. the sequence GATAGGCACCTG would yield a distance of 2 bp). Motif pairs were then separated based on whether the ATAC-seq peak showed reduced or increased accessibility with BHLHE40-OE(FDR=0.2). Motif spacing enrichment was then computed by obtaining the number of motif-spacings at each distance from 1 to 50 bp for each group separately.Enrichment significance was tested based on the SpaMo approach (Whitington et al., 2011). For a given motif-pair with a maximum motif-motif distance of r, we assume that the number of sequences exhibiting a motif-motif distance 0 < x < r is binomially distributed, with all spacings being equally likely. This means that the probability of a given pair having a specific motif-spacing, ignoring strand and motif-orientation, is 1 / r. We choose restrict r <= 50, reasoning that motif pairs exceeding this edge-edge distance are unlikely to represent meaningful TF interactions, and to reduce the burden of multiple testing. The probability that we will observe a specific number of motif-pairs *n* out of a total number of motif-pairs *m* exhibiting a specific motif-spacing *x* by chance alone is then computed as 1.0 - CDF_binomial(n,m,1.0 / r). This p-value is then corrected for multiple testing using the BH procedure.

## Common SNP, eQTL and Conservation Score Analysis

Common SNPs are adapted from Ensembl GRCh38(Cunningham et al., 2022), with a cut-off of more than 1% minor allele frequency. Expression quantitative trait loci(eQTL) are adapted from eQTLGEN(Võsa et al., 2021), only cis-eQTLs are considered here with FDR< 0.05 (data from https://eqtlgen.org/cis-eqtls.html, gene locus for CD69). Conservation score are adapted via phastCons100way score(0-1, clear to dark green)(Siepel et al., 2005), sg#70 regions are marked specifically in the plot.

## Acknowledgment

We thank S. M. Bevil, S. Battaglia, G. Rahme, K. Macias and J. Verga for technical assistance. We thank the Google TPU Research Cloud for providing TPU access and support. This work was supported by funds from the NCI/NIH Director’s Fund (DP1CA216873 to B.E.B.), the Gene Regulation Observatory and the Variant-to-Function Initiative at the Broad Institute. Z.C. is supported by NCI-CA-234842. L.P. is supported by the National Human Genome Research Institute (NHGRI) (R35HG010717, UM1HG012010). J.W. is supported by a postdoctoral fellowship from the Damon-Runyon Cancer Research Foundation. M.E.V. is supported by the National Cancer Institute (NCI) of the National Institutes of Health under the Ruth L. Kirschstein National Research Service Award (F31CA257625). B.E.B. is the Richard and Nancy Lubin Family Endowed Chair at the Dana-Farber Cancer Institute and an American Cancer Society Research Professor.

## Author information

Z.C., N.J., F.J.N., and B.E.B. conceived the study. Z.C., F.J.N., and B.E.B. designed the experiments. Z.C. and M.M. performed the experiments. M.E.V. and L.P. provided computational assistance. Z.C., N.J., J.W. analyzed the data. Z.C., N.J., F.J.N., and B.E.B. interpreted the data and wrote the manuscript.

## Declare of Interests

B.E.B. declares outside interests in Fulcrum Therapeutics, HiFiBio, Arsenal Biosciences, Design Pharmaceuticals, Cell Signaling Technologies, and Chroma Medicine.

## References

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106. https://doi.org/10.1186/gb-2010-11-10-r106.

Andersson, R., and Sandelin, A. (2020). Determinants of enhancer and promoter activities of regulatory elements. Nat. Rev. Genet. *21*, 71–87. https://doi.org/10.1038/s41576-019-0173-8.

Anzalone, A.V., Randolph, P.B., Davis, J.R., Sousa, A.A., Koblan, L.W., Levy, J.M., Chen, P.J., Wilson, C., Newby, G.A., Raguram, A., et al. (2019). Search-and-replace genome editing without double-strand breaks or donor DNA. Nature *576*, 149–157. https://doi.org/10.1038/s41586-019-1711-4.

Asanoma, K., Liu, G., Yamane, T., Miyanari, Y., Takao, T., Yagi, H., Ohgami, T., Ichinoe, A., Sonoda, K., Wake, N., et al. (2015). Regulation of the Mechanism of TWIST1 Transcription by BHLHE40 and BHLHE41 in Cancer Cells. Mol. Cell. Biol. *35*, 4096–4109. https://doi.org/10.1128/MCB.00678-15.

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Taylor, K.R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D.R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. Nat. Methods *18*, 1196–1203. https://doi.org/10.1038/s41592-021-01252-x.

Bailey, T.L., Johnson, J., Grant, C.E., and Noble, W.S. (2015). The MEME Suite. Nucleic Acids Res. *43*, W39–W49. https://doi.org/10.1093/nar/gkv416.

Brignall, R., Cauchy, P., Bevington, S.L., Gorman, B., Pisco, A.O., Bagnall, J., Boddington, C., Rowe, W., England, H., Rich, K., et al. (2017). Integration of Kinase and Calcium Signaling at the Level of Chromatin Underlies Inducible Gene Activation in T Cells. J. Immunol. *199*, 2652–2667. https://doi.org/10.4049/jimmunol.1602033.

Canver, M.C., Smith, E.C., Sher, F., Pinello, L., Sanjana, N.E., Shalem, O., Chen, D.D., Schupp, P.G., Vinjamur, D.S., Garcia, S.P., et al. (2015). BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature *527*, 192–197. https://doi.org/10.1038/nature15521.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

Cibrián, D., and Sánchez-Madrid, F. (2017). CD69: from activation marker to metabolic gatekeeper. Eur. J. Immunol. *47*, 946–953. https://doi.org/10.1002/eji.201646837.

Clement, K., Rees, H., Canver, M.C., Gehrke, J.M., Farouni, R., Hsu, J.Y., Cole, M.A., Liu, D.R., Joung, J.K., Bauer, D.E., et al. (2019). CRISPResso2 provides accurate and rapid genome editing sequence analysis. Nat. Biotechnol. *37*, 224–226. https://doi.org/10.1038/s41587-019-0032-3.

Cook, M.E., Jarjour, N.N., Lin, C.-C., and Edelson, B.T. (2020). Transcription Factor Bhlhe40 in Immunity and Autoimmunity. Trends Immunol. *41*, 1023–1036. https://doi.org/10.1016/j.it.2020.09.002.

Cuella-Martin, R., Hayward, S.B., Fan, X., Chen, X., Huang, J.-W., Taglialatela, A., Leuzzi, G., Zhao, J., Rabadan, R., Lu, C., et al. (2021). Functional interrogation of DNA damage response variants with base editing screens. Cell *184*, 1081–1097.e19. https://doi.org/10.1016/j.cell.2021.01.041.

Cunningham, F., Allen, J.E., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Austine-Orimoloye, O., Azov, A.G., Barnes, I., Bennett, R., et al. (2022). Ensembl 2022. Nucleic Acids Res. *50*, D988–D995. https://doi.org/10.1093/nar/gkab1049.

Diao, Y., Fang, R., Li, B., Meng, Z., Yu, J., Qiu, Y., Lin, K.C., Huang, H., Liu, T., Marina, R.J., et al. (2017). A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. Nat. Methods *14*, 629–635. https://doi.org/10.1038/nmeth.4264.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

Doench, J.G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E.W., Donovan, K.F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., et al. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat. Biotechnol. *34*, 184–191. https://doi.org/10.1038/nbt.3437.

Dominguez, A.A., Lim, W.A., and Qi, L.S. (2015). Beyond editing: repurposing CRISPR–Cas9 for precision genome regulation and interrogation. Nat. Rev. Mol. Cell Biol. *17*, 5–15. https://doi.org/10.1038/nrm.2015.2.

Emming, S., Bianchi, N., Polletti, S., Balestrieri, C., Leoni, C., Montagner, S., Chirichella, M., Delaleu, N., Natoli, G., and Monticelli, S. (2020). A molecular network regulating the proinflammatory phenotype of human memory T lymphocytes. Nat. Immunol. *21*, 388–399. https://doi.org/10.1038/s41590-020-0622-8.

ENCODE Project Consortium, Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699–710. https://doi.org/10.1038/s41586-020-2493-4.

FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest, A.R.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J.L., Haberle, V., Lassmann, T., Kulakovskiy, I.V., Lizio, M., et al. (2014). A promoter-level mammalian expression atlas. Nature *507*, 462–470. https://doi.org/10.1038/nature13182.

Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J.E., Mudge, J.M., Sisu, C., Wright, J.C., Armstrong, J., Barnes, I., et al. (2021). GENCODE 2021. Nucleic Acids Res. *49*, D916–D923. https://doi.org/10.1093/nar/gkaa1087.

Fulco, C.P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S.R., Perez, E.M., Kane, M., Cleary, B., Lander, E.S., and Engreitz, J.M. (2016). Systematic mapping of functional enhancer–

promoter connections with CRISPR interference. Science *354*, 769–773. https://doi.org/10.1126/science.aag2445.

Gaudelli, N.M., Komor, A.C., Rees, H.A., Packer, M.S., Badran, A.H., Bryson, D.I., and Liu, D.R. (2017). Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. Nature *551*, 464–471. https://doi.org/10.1038/nature24644.

Gilbert, L.A., Larson, M.H., Morsut, L., Liu, Z., Brar, G.A., Torres, S.E., Stern-Ginossar, N., Brandman, O., Whitehead, E.H., Doudna, J.A., et al. (2013). CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell *154*, 442–451. https://doi.org/10.1016/j.cell.2013.06.044.

Godec, J., Tan, Y., Liberzon, A., Tamayo, P., Bhattacharya, S., Butte, A.J., Mesirov, J.P., and Haining, W.N. (2016). Compendium of Immune Signatures Identifies Conserved and Species-Specific Biology in Response to Inflammation. Immunity *44*, 194–206. https://doi.org/10.1016/j.immuni.2015.12.006.

Grant, C.E., and Bailey, T.L. (2021). XSTREME: Comprehensive motif analysis of biological sequence datasets. bioRxiv https://doi.org/10.1101/2021.09.02.458722.

GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330. https://doi.org/10.1126/science.aaz1776.

GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, et al. (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213. https://doi.org/10.1038/nature24277.

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics *32*, 2847–2849. https://doi.org/10.1093/bioinformatics/btw313.

Hanna, R.E., Hegde, M., Fagre, C.R., DeWeirdt, P.C., Sangree, A.K., Szegletes, Z., Griffith, A., Feeley, M.N., Sanson, K.R., Baidi, Y., et al. (2021). Massively parallel assessment of human variants with base editor screens. Cell *184*, 1064–1080.e20. https://doi.org/10.1016/j.cell.2021.01.012.

Ho, I.-C., Tai, T.-S., and Pai, S.-Y. (2009). GATA3 and the T-cell lineage: essential functions before and after T-helper-2-cell differentiation. Nat. Rev. Immunol. *9*, 125–135. https://doi.org/10.1038/nri2476.

Honma, S., Kawamoto, T., Takagi, Y., Fujimoto, K., Sato, F., Noshiro, M., Kato, Y., and Honma, K.-I. (2002). Dec1 and Dec2 are regulators of the mammalian molecular clock. Nature *419*, 841–844. https://doi.org/10.1038/nature01123.

Huynh, J.P., Lin, C.-C., Kimmey, J.M., Jarjour, N.N., Schwarzkopf, E.A., Bradstreet, T.R., Shchukina, I., Shpynov, O., Weaver, C.T., Taneja, R., et al. (2018). Bhlhe40 is an essential repressor of IL-10 during Mycobacterium tuberculosis infection. J. Exp. Med. *215*, 1823–1838. https://doi.org/10.1084/jem.20171704.

Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., Alston, J., Mikkelsen, T.S., and Kellis, M. (2013). Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome Res. *23*, 800–811. https://doi.org/10.1101/gr.144899.112.

Kim, Y.B., Komor, A.C., Levy, J.M., Packer, M.S., Zhao, K.T., and Liu, D.R. (2017). Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions. Nat. Biotechnol. *35*, 371–376. https://doi.org/10.1038/nbt.3803.

Klein, J.C., Agarwal, V., Inoue, F., Keith, A., Martin, B., Kircher, M., Ahituv, N., and Shendure, J. (2020). A systematic evaluation of the design and context dependencies of massively parallel reporter assays. Nat. Methods *17*, 1083–1091. https://doi.org/10.1038/s41592-020-0965-y.

Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A., and Liu, D.R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature *533*, 420–424. https://doi.org/10.1038/nature17946.

Korhonen, J., Martinmäki, P., Pizzi, C., Rastas, P., and Ukkonen, E. (2009). MOODS: fast search for position weight matrix matches in DNA sequences. Bioinformatics *25*, 3181–3182. https://doi.org/10.1093/bioinformatics/btp554.

Korkmaz, G., Lopes, R., Ugalde, A.P., Nevedomskaya, E., Han, R., Myacheva, K., Zwart, W., Elkon, R., and Agami, R. (2016). Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. Nat. Biotechnol. *34*, 192–198. https://doi.org/10.1038/nbt.3450.

Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M.N., and Sergushichev, A. (2021). Fast gene set enrichment analysis.

Kulakovskiy, I.V., Vorontsov, I.E., Yevshin, I.S., Sharipov, R.N., Fedorova, A.D., Rumynskiy, E.I., Medvedeva, Y.A., Magana-Mora, A., Bajic, V.B., Papatsenko, D.A., et al. (2018). HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. *46*, D252–D259. https://doi.org/10.1093/nar/gkx1106.

Laguna, T., Notario, L., Pippa, R., Fontela, M.G., Vázquez, B.N., Maicas, M., Aguilera-Montilla, N., Corbí, Á.L., Odero, M.D., and Lauzurica, P. (2015). New insights on the transcriptional regulation of CD69 gene through a potent enhancer located in the conserved non-coding sequence 2. Mol. Immunol. *66*, 171–179. https://doi.org/10.1016/j.molimm.2015.02.031.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Lun, A.T.L., and Smyth, G.K. (2016). csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. Nucleic Acids Res. *44*, e45. https://doi.org/10.1093/nar/gkv1191.

Maricque, B.B., Chaudhari, H.G., and Cohen, B.A. (2018). A massively parallel reporter assay dissects the influence of chromatin structure on cis-regulatory activity. Nat. Biotechnol. https://doi.org/10.1038/nbt.4285.

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C.G., Jr, Kinney, J.B., et al. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat. Biotechnol. *30*, 271–277. https://doi.org/10.1038/nbt.2137.

Mumbach, M.R., Satpathy, A.T., Boyle, E.A., Dai, C., Gowen, B.G., Cho, S.W., Nguyen, M.L., Rubin, A.J., Granja, J.M., Kazane, K.R., et al. (2017). Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. Nature Genetics *49*, 1602–1612. https://doi.org/10.1038/ng.3963.

Nasser, J., Bergman, D.T., Fulco, C.P., Guckelberger, P., Doughty, B.R., Patwardhan, T.A., Jones, T.R., Nguyen, T.H., Ulirsch, J.C., Lekschas, F., et al. (2021). Genome-wide enhancer maps link risk variants to disease genes. Nature *593*, 238–243. https://doi.org/10.1038/s41586-021-03446-x.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods *14*, 417–419. https://doi.org/10.1038/nmeth.4197.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

Rajagopal, N., Srinivasan, S., Kooshesh, K., Guo, Y., Edwards, M.D., Banerjee, B., Syed, T., Emons, B.J.M., Gifford, D.K., and Sherwood, R.I. (2016). High-throughput mapping of regulatory DNA. Nat. Biotechnol. *34*, 167–174. https://doi.org/10.1038/nbt.3468.

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44*, W160–W165. https://doi.org/10.1093/nar/gkw257.

Rosello, M., Serafini, M., Mignani, L., Finazzi, D., Giovannangeli, C., Mione, M.C., Concordet, J.-P., and Del Bene, F. (2022). Disease modeling by efficient genome editing using a near PAM-less base editor in vivo. Nat. Commun. *13*, 3435. https://doi.org/10.1038/s41467-022-31172-z.

Sanda, T., Lawton, L.N., Barrasa, M.I., Fan, Z.P., Kohlhammer, H., Gutierrez, A., Ma, W., Tatarek, J., Ahn, Y., Kelliher, M.A., et al. (2012). Core transcriptional regulatory circuit controlled by the TAL1 complex in human T cell acute lymphoblastic leukemia. Cancer Cell *22*, 209–221. https://doi.org/10.1016/j.ccr.2012.06.007.

Sanjana, N.E., Wright, J., Zheng, K., Shalem, O., Fontanillas, P., Joung, J., Cheng, C., Regev, A., and Zhang, F. (2016). High-resolution interrogation of functional elements in the noncoding genome. Science *353*, 1545–1549. https://doi.org/10.1126/science.aaf7613.

Sathaliyawala, T., Kubota, M., Yudanin, N., Turner, D., Camp, P., Thome, J.J.C., Bickham, K.L., Lerner, H., Goldstein, M., Sykes, M., et al. (2013). Distribution and compartmentalization of human circulating and tissue-resident memory T cell subsets. Immunity *38*, 187–197. https://doi.org/10.1016/j.immuni.2012.09.020.

Shrikumar, A., Tian, K., and Avsec, Ž. (2018). Technical Note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5.6.5. arXiv https://doi.org/ https://doi.org/10.48550/arXiv.1811.00416.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050. https://doi.org/10.1101/gr.3715005.

Singer, M., Wang, C., Cong, L., Marjanovic, N.D., Kowalczyk, M.S., Zhang, H., Nyman, J., Sakuishi, K., Kurtulus, S., Gennert, D., et al. (2017). A Distinct Gene Module for Dysfunction Uncoupled from Activation in Tumor-Infiltrating T Cells. Cell *171*, 1221–1223. https://doi.org/10.1016/j.cell.2017.11.006.

Stunnenberg, H.G., International Human Epigenome Consortium, and Hirst, M. (2016). The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. Cell *167*, 1897. https://doi.org/10.1016/j.cell.2016.12.002.

Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Haugen, E., et al. (2020). Global reference mapping of human transcription factor footprints. Nature *583*, 729–736. https://doi.org/10.1038/s41586-020-2528-x.

Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Yazar, S., et al. (2021). Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat. Genet. *53*, 1300–1310. https://doi.org/10.1038/s41588-021-00913-z.

Wan, Y.Y. (2014). GATA3: a master of many trades in immune regulation. Trends Immunol. *35*, 233–242. https://doi.org/10.1016/j.it.2014.04.002.

Whitington, T., Frith, M.C., Johnson, J., and Bailey, T.L. (2011). Inferring transcription factor complexes from ChIP-seq data. Nucleic Acids Res. *39*, e98. https://doi.org/10.1093/nar/gkr341.

Zawel, Yu, and Torrance DEC1 is a downstream target of TGF-β with sequence-specific transcriptional repressor activities. Proc. Estonian Acad. Sci. Biol. Ecol.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137. https://doi.org/10.1186/gb-2008-9-9-r137.

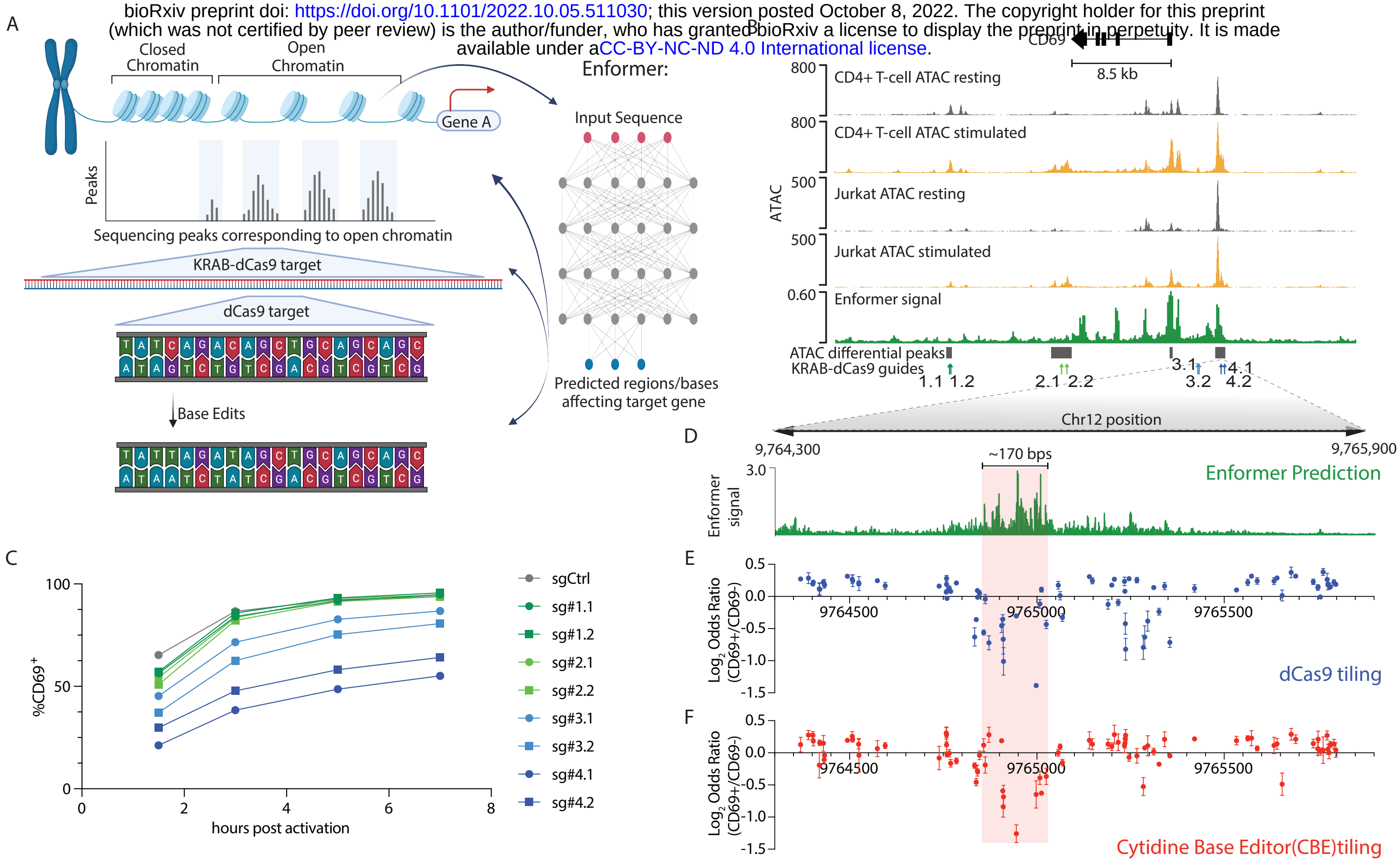

Figure 1. Integrative analysis of the CD69 regulatory landscape.
A) Gene regulatory landscape characterization by successive functional assays and deep learning.
B) Genomic tracks depict accessibility of the CD69 locus in primary CD4+ T cells and Jurkat cells, without or with stimulation (PMA/ionomycin). Enformer signal track shows the predicted contribution of underlying sequence to CD69 expression (magnitude of the model gradient at each position with respect to CD69 promoter signal, summed over 128 bp bins)in Jurkat cells. Grey bars depict regions with differential accessibility in stimulated Jurkat cells, relative to resting (FDR=0.2). CRISPRi sgRNA positions are also indicated. ATAC signal corresponds to reads per genomic content (RPGC).
C) Flow cytometry of CD69 expression in Jurkat cells targeted with the indicated CRISPRi sgRNA following a stimulation time course. Samples gated from the lentiviral transduced population (mCherry+).
D) Expanded view of Enformer signal at single base resolution over RE-4, as denoted in panel b.
E) Enrichment/depletion plot of dCas9 sgRNAs in CD69+ Jurkat cells, relative to CD69- cells (y-axis; Log$_2$ Odds Ratio of normalized sgRNA reads). sgRNAs along the x-axis according to their 5' starting position on the positive strand. Each data point represents mean±s.e.m.
F) Enrichment/depletion plot of Cytidine Base Editor (CBE) sgRNAs in CD69+ Jurkat cells, relative to CD69- cells (as in panel e).
For c,e,f, data represent 2-3 biological independent experiments. A 170 bp region critical for CD69 activation is denoted (d-f, light red).

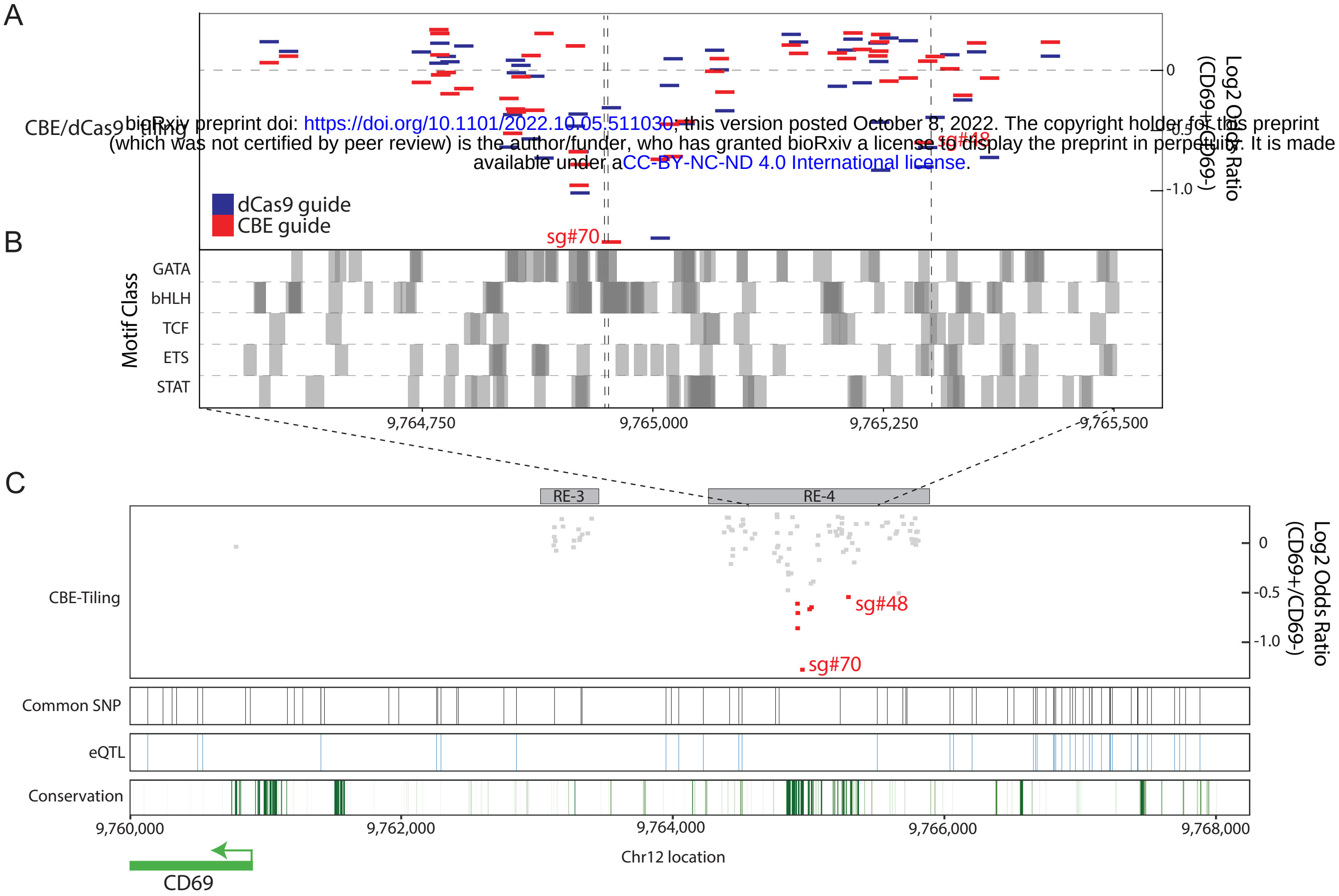

Figure 2. A critical sequence interval within RE-4 influences CD69 expression.

A) Enrichment/depletion plot of sgRNAs in dCas9 and CBE tiling screens as in Fig 1e/f, limited to the central portion of RE-4 with sgRNAs shown to scale. Expected C->T edit positions highlighted for CBE-sgRNAs sg#70 and sg#48 (dashed grey lines).

B) Transcription factor motif locations (grouped by broad motif class) for key immune regulators shown across the same interval as in panel a (FDR<0.05). Dark grey areas represent overlapping motifs.

C) Zoomed out view of the CD69 locus shows CBE sgRNA depletion (red boxes indicate significantly depleted sgRNAs), common SNPs (black vertical stripes), eQTLs (blue vertical stripes)27 and PhastCon100 conservation score (green stripes).

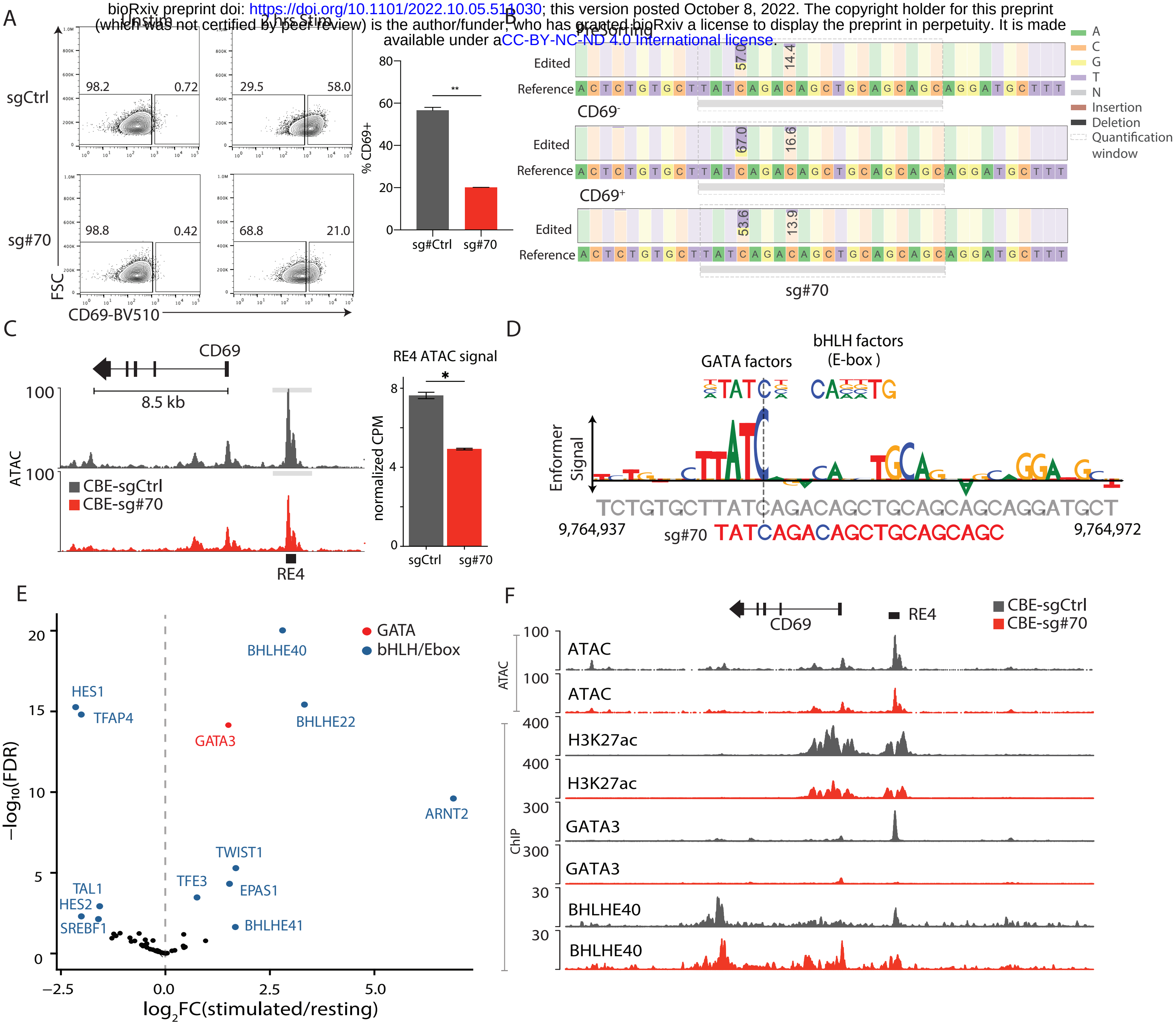

Figure 3. Top scoring base edits target competitive TF binding sites.

A) Flow cytometry plots of CD69 signal for CBE-sgCtrl and CBE-sg#70 Jurkat cells under resting or stimulated conditions. Bar plot depicts the proportion of CD69+ cells in CBE-sgCtrl (grey) and CBE-sg#70 (red) after stimulation. P-value based on unpaired t test, **P<0.01. Data are from 4 independent experiments each with 2-3 technical replicates, mean±s.e.m.

B) Table depicts frequency of incurred base edits in CBE-sg#70 infected Jurkat cells. PCR amplicons from unsorted, CD69- and CD69+ populations were sequenced by Illumina Nextseq500. Consensus sequence is shown along with stacked bars that depict the proportions of cytosine and thymine bases in the sequencing data (numbers indicate percent of alleles with C->T edit). Shaded boxes indicate the sg#70 target sequence.

C) Chromatin accessibility shown over the CD69 locus for stimulated CBE-sg#70 (red) and CBE-sgCtrl (grey) Jurkat cells. Bar plot depicts the mean ATAC-seq signal over RE-4 (TMM normalized counts per million; CPM). P-value based on unpaired t test, *P<0.05. Data are from 3 replicates, mean±s.e.m.

D) Enformer signal (letter height) for the sg#70 target region indicates the predicted impact of each base on RE-4 accessibility. The sgRNA directly coincides with a GATA motif and a bHLH/E-box motif, and incurs an edit that disrupts the former (vertical dashed line).

E) Volcano plot depicts gene expression fold-change (x-axis) and significance (y-axis) for TF genes in stimulated Jurkat cells, relative to resting cells. Labels identify differential GATA (red) and bHLH/Ebox (blue) family members.

F) Genomic tracks for the CD69 locus depict chromatin accessibility (ATAC), H3K27 acetylation (H3K27ac), GATA3 binding and BHLHE40 binding in CBE-sgCtrl (grey) and CBE-sg#70 (red) Jurkat cells. Y-axis represents the -log10(p-value) to input controls.

Jurkat cells in A, B, C and F were stimulated with PMA/ionomycin for 2 hours.

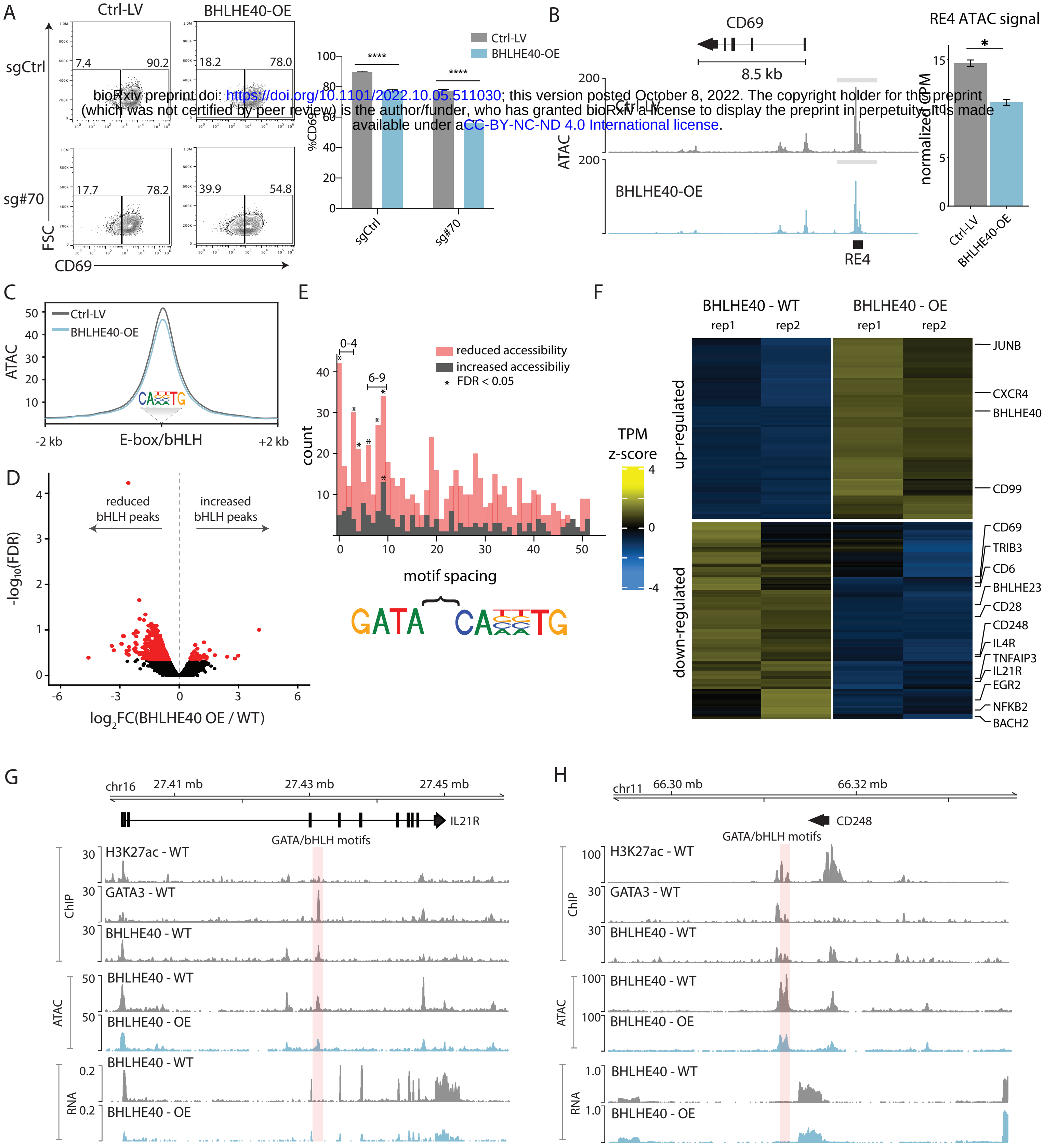

Figure 4. GATA3-BHLHE40 competition impacts global T cell transcriptional responses.
A) Flow cytometry plots of CD69 signal for stimulated Jurkat cells transduced with CBE-sg#70 and a BHLHE40 overexpression construct (BHLHE40-OE), or with corresponding controls (sgCtrl and Ctrl-LV, respectively). Bar plot depicts the proportion of CD69+ cells in each condition. P-value based on unpaired t test, ****P<0.0001. Data are from 3 independent experiments with 2-3 technical replicates, mean±s.e.m.
B) Chromatin accessibility in the CD69 locus for CBE-sg#70 Jurkat cells transduced with either BHLHE40 overexpression lentivirus (light blue) or control (grey). Cells were stimulated with PMA/ionomycin. P-value based on unpaired t test without multiple testing correction, *P<0.05. Bar plot data are from 2 replicates, mean±s.e.m ATAC-seq signal over RE-4 (TMM normalized CPM).
C) Plot depicts aggregate accessibility (y-axis) for GATA3 bound sites that also harbor bHLH/E-box motifs (centered on the motifs). Data shown for stimulated Jurkat cells transduced with either BHLHE40 overexpression lentivirus (light blue) or control (grey).
D) For the set of GATA3 bound sites with bHLH/E-box motifs in c, volcano plot depicts fold-change (x-axis) and significance (y-axis) of chromatin accessibility in Jurkat cells transduced with BHLHE40 overexpression lentivirus, relative to control. Differentially accessible sites (FDR < 0.1) are indicated in red.
E) For differentially accessible sites in d, histogram shows the number of sites (y-axis) with the indicated spacing (x-axis) between GATA and bHLH/E-box motifs. Sites are stratified by whether their accessibility is reduced (red) or increased (grey) in the BHLHE40 overexpressing cells. Sites with significant peak differential between reduced and increased accessibility (FDR < 0.05) are denoted (*).
F) Heatmap shows differentially expressed genes with BHLHE40 binding in their REs, in BHLHE40 overexpressing Jurkat cells relative to control. Cells were stimulated with PMA/ionomycin.
G-H) Genomic views of the IL21R(F) and CD248(G) loci show ChIP-seq data for H3K27Ac, BHLHE40, and GATA3 in stimulated Jurkat cells (signal corresponds to P-value enrichment over input). Accessibility (ATAC) and expression (RNA-seq) are also shown for BHLHE40 overexpressing Jurkat cells and controls. Sites with combined GATA and bHLH/Ebox motifs are indicated (pink shade).
Jurkat cells in B-F were stimulated with PMA/ionomycin for 2 hours.