Discovery and engineering of hypercompact epigenetic modulators for durable gene activation

Giovanni A. Carosso[1], Robin W. Yeo[1], T. Blair Gainous[1], M. Zaki Jawaid[1], Xiao Yang[1], Vincent Cutillas[1], Lei Stanley Qi[2,3,4], Timothy P. Daley[1], Daniel O. Hart[1]

[1] Epicrispr Biotechnologies, South San Francisco, CA 94080, USA;
[2] Department of Bioengineering, Stanford University, Stanford, CA 94305, USA;
[3]Sarafan ChEM-H, Stanford University, Stanford, CA 94305, USA;
[4]Chan Zuckerberg Biohub-San Francisco, San Francisco, CA 94158, USA

Abstract

Transcriptional and epigenetic regulators shape the chromatin microenvironment and corresponding gene expression during cellular differentiation and homeostasis. Programmable modulators of transcription provide a powerful toolkit for controlling gene dosage in therapeutic applications, but a limited catalog of functional domains constrains their robustness and durability profiles, and large cargo sizes impede clinical delivery. To address these limitations, here we perform high-throughput screening to discover novel classes of transcriptional modulators among human, viral, and archaeal proteomes and characterize their functions in a multitude of endogenous human contexts. We identify compact, potent activators from viral proteomes with exceptional robustness across silent and expressed genes in varied cell types using distinct dCas systems. Insights from predicted 3-dimensional structures and machine learning models enabled us to rationally engineer improved activators, both in potency and persistence. Notably, engineered activators achieved mitotically durable gene activation following transient delivery. Our discovery pipeline provides a predictive rubric for the systematic development of hypercompact activators from unannotated proteomes, yielding superior efficiency and kinetics profiles that broadly expand the epigenetic editing toolkit for therapeutic applications.

Introduction

RNA-guided epigenetic control of gene expression via nuclease-dead CRISPR/Cas (dCas) systems offers potential therapeutic avenues for treating a wide range of disease indications without creating double strand breaks or changing the DNA sequence[1,2,3]. However, existing tools are limited in clinical utility by large coding DNA cargo sizes, variable efficacies at diverse targets, and temporally transient windows of activity[4,5]. Classical approaches to programmable target gene activation have relied on a limited catalog of peptide domains sourced from eukaryotic and viral proteomes, or fusions thereof[6,7,8]. Such domains can act by recruiting transcriptional cofactors in part through acidic and hydrophobic contact interfaces[9,10,11], or by encoding enzymatic domains that alter DNA methylation status or histone post-translational modifications[12,13,14]. Emerging literature suggests highly disparate activity levels among these modulatory mechanisms in terms of potency, context-dependent robustness, and durability of effect[15,16]. Expansion of the transcriptional toolkit is therefore critical to achieve control of target gene expression in distinct chromatin contexts. Unbiased tiling and testing of diverse proteomes with high-throughput screening is a powerful method for identifying compact peptide fragments capable of transcriptional modulation[17,18,19]. The most potent peptide fragments, in turn, can provide a basis for iterative refinement of desirable modulator properties by subsequent rounds of rational engineering[20,21,22].

While the human genome encodes over 2,000 transcription factors and chromatin regulators[23], virological surveys calculate up to 320,000 mammalian-tropic viral species among the estimated 10^{31} viral particles comprising the global virome[24,25]. Key to viral evolution is the ability of their proteins to manipulate host gene expression, suggesting that these genomes represent a vast reservoir of rapidly evolving molecular tools optimized for transcriptional modulation with compact coding size[26]. Previous studies have separately evaluated sets of either human- or viral-encoded domains[27,28], but direct comparisons of their group-wise transcription-modulatory functions at multiple endogenous human targets remain to be performed. The restrictive environmental constraints of other biological classes, such as extremophilic archaea, could result in peptide domains with unique biochemical properties and thereby functionalities[29].

Unmodified histone tails carry a positive charge that drives electrostatic interactions with DNA and neighboring nucleosomes contributing to a compacted chromatin environment that restricts engagement by the transcriptional apparatus, preventing gene expression in the default state[30]. While repressive epigenetic marks such as DNA methylation and H3K27me3/H3K9me3 marks on histone tails can be propagated heritably during cellular replication, epigenetic persistence of permissive chromatin marks remains to be discovered[31,32,33,34,35,36]. Accordingly, while others have demonstrated robust and mitotically durable gene silencing via dCas-mediated transcriptional modulation, robust gene activation dissipates in the absence of the initiator[37,38,8,39]. Prolonged activation windows have been reported via co-deposition of active marks H3K4me3 and H3K79me3 by dCas9 fusions to PRDM9 and DOT1L histone methyltransferases, respectively, or via DNA demethylation by dCas9-Tet1, suggesting that opportunities exist for mitotically stable gene activation[13,14]. Intrinsic programs exist for cells to durably activate gene expression programs, as in early embryonic development,

that persist heritably against the challenges of rapid cell division and differentiation[40,41,42]. Moreover, cellular environmental responses require sustained gene activation, as in H3K4me1-mediated p65 induction following transient hyperglycemic spike[43], or during innate immune responses with H3K27ac-mediated immune gene activation which precedes DNA demethylation[44,45]. Mitotically stable activation memory, if mediated by H3K27ac, would likely require cognate epigenetic readers, such as CBP/P300-associated BET family bromodomains, to recognize and redeposit the permissive acetyl marks on nucleosomes via 'read-write' mechanisms[46,47].

Here, we performed high-throughput dCas-modulator recruitment screens to systematically interrogate the transcriptional potencies of tens of thousands of human, viral, and archaeal-derived peptide sequences in opposing promoter contexts; we discovered known and novel peptide domains with strong biases toward activation by viral domains, suppression by human domains, and context-dependent dual-activity profiles for archaeal domains. Sequence-based biochemical composition analysis, paired with extensive validation testing at multiple target promoters in diverse chromatin contexts, revealed a predictive biochemical rubric for engineering functional improvements to minimal core activator domains. Importantly, a subset of the resultant activators displays a novel ability to maintain target activation through dozens of cell divisions after a single transient delivery. This pipeline for modulator discovery and engineering yielded hypercompact transcriptional activators as compact as 64 amino acids that outperform previous benchmarks with respect to potency, context-independent robustness, and mitotic durability of effect, despite occupying ~12-20% of their protein-coding cargo size.

Results

Identification of transcriptional modulator domains by high-throughput screening

To interrogate the regulatory abilities of naturally occurring peptide fragments, we first generated libraries by tiling across candidate human, viral, and archaeal genomic coding sequences (**Fig. 1a** and **Extended Data Fig. 1a**). For the human set, we identified a shortlist of 549 human nuclear-localized proteins including DNA- and histone-modifying enzymes, chromatin remodelers, and transcription factors. We then curated a custom set of viral full-genome coding sequences (n=3,548) distributed across 188 viral families. Selections were weighted to enrich for viruses that evolved predominantly in mammalian host reservoirs or have undergone zoonotic transfer, reasoning that such domains may be adapted to modulate human gene expression[48,49]. We further sought to ask whether domains from acidophilic, thermophilic, or otherwise extremophilic viruses and host archaea might confer more robust or heritable transcription-modulatory functions, owing to evolutionary pressures that favor vertical versus horizontal host transmission, so we included viral peptides from metagenomics surveys (n=129)[50,51] and the full proteome of volcanic archaeon *Acidianus infernus*[52]. Lastly, we added a set of annotated human virus transcriptional regulators (vTRs)[26] (n=419). As positive controls, we added human proteins with known transcription-modulatory function, and GC-matched random-sequence negative controls. We partitioned the sequences into overlapping DNA tiles encoding 85 amino acid peptide fragments and pooled these into a merged screening library of uniquely barcoded oligonucleotides (n=43,938) to enable group-wise comparisons of modulator potency.

To eliminate noise from variability in sgRNA efficiencies, we engineered a pair of fluorescent reporter systems with seven identical sgRNA binding sites upstream of either a minimal CMV (mCMV) promoter-GFP (default OFF) or EF1α promoter-GFP (default ON) to screen for activators and suppressors, respectively (**Fig. 1b,top** and **Extended Data Fig. 1b**). As expected, the mCMV-GFP cell line could be activated by dCas9-VPR and the EF1α-GFP cell line could be suppressed by dCas9-KRAB (**Extended Data Fig. 1c**). mCMV-GFP and EF1α-GFP reporter cells stably expressing a targeting Cas9 sgRNA were lentivirally transduced with dCas9-modulator library fusions, then sampled by FACS-based separation of GFP-ON and GFP-OFF cells (**Fig. 1b,bottom**). We computed barcode enrichments in GFP-ON versus GFP-OFF cells, identifying 560 activators and 1,330 suppressors of mCMV-GFP and EF1α-GFP, respectively (**Fig. 1c** and **Extended Data Fig. 1d,e**) with taxon-dependent potencies (**Extended Data Fig. 1f**). Comparisons between both screens provided early insights to context-robustness; among the 560 mCMV-GFP activators, 303 were depleted among EF1α-GFP suppressors (**Fig. 1c,d** and **Extended Data Fig. 2a**). Similarly, among 1,330 EF1α-GFP suppressors, 622 were depleted among mCMV-GFP activators. Random negative control tiles showed a uniform distribution about $\log_2$FC=0 in both screens (**Extended Data Fig. 2b**).

While activator hits were distributed proportionately across viral (n=177), human (n=209), and archaeal (n=154) tiles, the most potent were viral in origin and segregated into distinct sequence homology clusters (**Fig. 1e,f**) of both known and novel classes: E1A (hAdv), VP16 (HHV), VLTF3 (Pox), IE1/IE2 (HCMV), and vIRF2 (HHV) (**Fig. 1g**). Human activator clusters largely consisted of homologous tiles from different isoforms of major families including E2F5, ARNTL2, ZFX/ZFY, NHSL1, TET3, HSF1, MESP1, and TOX4 (**Fig. 1e-g**). We next trained a regularized regression model to predict activation strength based on peptide sequence (**Fig. 1h**) and extracted scores for amino acid importance (**Fig. 1i** and **Extended Data Fig. 2c-h**). While enrichment of acidic residues is indeed predictive of activation[11], we find that depletion of basic residues is a far stronger biochemical predictor (**Fig. 1i-l**). Results from a subsequent validation screen (**Extended Data Fig. 3**) were consistent with these biochemical predictions (**Fig 1k,l**). Interestingly, activators display taxonomy-dependent biochemical trends; viral activators are strongly enriched for acidic residues while human activators are primarily depleted of basic residues (**Fig. 1j**). We generated 3-dimensional structure predictions using ESMFold and computed quantitative structural

features using Yasara (see Methods)[53,54]. We found that activator sequences are enriched for small amino acids, intolerant of β-sheets, and enriched for turns (**Fig. 1l, Extended Data Fig. 2)**. In contrast to previous reports[10,11], we do not observe enrichment of hydrophobic residues, suggesting that they are dispensable for activation of this locus (**Extended Data Fig. 2)**.

Suppressor hits were dominated by human tiles in potency, followed by archaeal tiles, although overall hit frequencies were distributed proportionately across human (n=502), viral (n=388), and archaeal (n=389) origins (**Fig. 1c**, **Extended Data Fig. 1f**). Human ZNF-derived tiles were the most potent suppressors, clustering based on KRAB-like domains or ZF C2H2 domains (**Extended Data Fig. 4a-c)**. Surprisingly, hydrophobic residue enrichment was greater among suppressors than activators (**Extended Data Fig. 4d)**. Archaeal suppressors, while poorly annotated, displayed higher potencies with increasing content of aliphatic residues and β-sheets (**Extended Data Fig. 4e**). Both activators and suppressors required a smaller solvent-accessible surface area (SASA) and presence of charged, acidic residues conferring negative electrostatic potential (**Extended Data Fig. 4f**).

Biochemical features predict robust and context-independent activator domains

We next focused on the less characterized or unannotated viral and archaeal elements of our modulator library. In-silico analysis of amino acid composition among the pooled screen hits suggested that relative to null peptides, robust viral activators require depletion of basic and aromatic residues and β-sheets, and that they tend to have fewer aliphatic, polar, and hydrophobic residues and lower helix and flexibility (B-factor) scores (**Extended Data Fig. 2c-g**). Further, they require highly negative net charge and electrostatic potential, low mass, prefer higher content of small, tiny, acidic, and coil residues, and exhibit a non-zero degree of turn residues (**Fig. 1i-l** and **Extended Data Fig. 2c-g**). Robust viral suppressors display fewer biochemical biases beyond hydrophobicity but are weakly enriched for charged and acidic residues and more tolerant to β-sheets, while robust archaeal suppressor potency increases with aliphatic residues and β-sheets (**Extended Data Fig. 4c-f**).

To test these predictions by context, we designed sgRNAs targeting a panel of twenty human and synthetic promoters and evaluated modulator potencies when fused C-terminally to the DNA-binding effector dCasMINI[55] as well as dCas9. We selected a set of 85-amino acid viral and archaeal modulators (n=95) with variable potency predictions in our K562 cell screens (**Extended Data Fig. 1e**) and diverse biochemical features (**Extended Data Fig. 2c,d**), and compared their potencies across promoter contexts in HEK293T cells relative to canonical benchmarks p65, Rta, VP64, tripartite VP64-p65-Rta (VPR), and ZNF10 KRAB (**Fig. 2a** and **Extended Data Fig. 5**). Overall concordance of target effects was greatest for the predicted strong viral activators as measured by protein expression (**Fig. 2b**) and mRNA measurements (**Extended Data Fig. 6**), independently of dCas recruitment system (**Extended Fig. 5a,b**). We observe hypercompact viral activators that robustly match or outperform larger benchmarks at all targets except TRE3G-GFP (**Fig. 2a,c**, **Extended Data Fig. 5**, **Extended Data Fig. 6**). Viral and archaeal suppressors outperformed human ZNF10 KRAB in some cases (**Extended Data Fig. 6**), but overall lacked the context-robustness of the benchmark. Consistently, the most robust viral activators were those most depleted of basic residues (**Fig. 2d,e**), lacked β-sheets, and were less enriched for hydrophobic residues than suppressors (**Fig. 2f,g**). Robust archaeal suppressors displayed an aliphatic dependence not seen for viral suppressors (**Fig. 2g**). Global comparison of target effects across each gene revealed the exceptional robustness of two high-potency peptides originating from human herpesvirus 8 (KSHV) protein (Q2HR71), both fulfilling our activator rubric of basic residue depletion, low negative charge, low hydrophobicity scores, and other features (**Fig. 2h** and **Extended Data Fig. 7**). We therefore hypothesized that exploiting this set of properties by domain engineering could further enhance transcriptional activator functions.

Engineering of vCD-containing modulators that outperform benchmarks

Positional mapping of our screen tiles revealed an overlap of two exceptionally high-potency and context-robust activators (**Fig. 2c,h**) from the viral interferon regulatory factor 2 (vIRF2) KSHV protein (Q2HR71), homologous to human interferon regulatory factors[56] (**Fig. 3a,b**). Among our 95 characterized viral and archaeal tiles, both tiles were biochemical outliers with respect to basic residue depletion, negative net charge, and low mass (**Fig. 2d,h; Extended Data Fig. 7**). We applied a convolutional neural network (ADPred)[57] to identify predicted transactivation domains and identified a 32-amino acid vIRF2 core domain (vCD) that was located within the overlapping region, but in differing positions (**Fig. 3c**). Despite sharing 68.2% sequence identity, these peptides differed notably in electrostatic potential and turn scores (**Extended Data Fig. 7**), suggesting a role for domain positional effects on secondary structure. We therefore rationally designed a set of vCD-based variants (**Extended Data Fig. 8**) to test configurations of 1) the isolated core domain by itself in various N-to-C distal positions and with varied flanking sequence compositions (n=24); 2) double-vCD tandem repeats, adjusting positions and inter-vCD linkers (n=17); 3) triple-vCD tandem repeats (n=4); 4) vCD fusions to VP16[6] in various orientations; and 5) vCD fusions to VP7 or VP28 (4xVP7)[58]. To each class we added mutational and partial sequence inversion variants to alter the vCD structure and gain mechanistic insights into vCD's mode of activation.

We subjected the engineered vCD variant set (n=101) to arrayed activator testing at CD45, IFNG, and CXCR4 in HEK293T cells and observed a wide range of potency enhancement and abrogation effects. At each target, enhanced vCD-based modulators as

compact as 64 amino acids surpassed the unmodified vCD and matched or outperformed VPR in potency (**Fig. 3d** and **Extended Data Fig. 5,8**). Potency increased with more negative electrostatic potential (ESP), but was dampened by amino acid substitutions in the most negative ESP variants (**Fig 3b**). We calculated a normalized robustness score based on activity at these four targets, finding positive correlations with higher helix score and lower coil score (**Fig. 3e**). Among single-core variants, stepwise increases in activation strength accompany the N-to-C distal shuffling of vCD, however substitutions of native vCD-flanking sequence resulted in total loss of activation strength (**Fig. 3f**). Duplicate and triplicate fusions led to stepwise increases in activation potency, and partial vCD inversions or single amino acid substitutions were largely tolerated, while multiple substitutions ablated activation (**Fig. 3g**). vCD-VP16 fusions gave the highest potencies, comparable to VPR, VP64, and Rta, whether oriented as vCD plus VP64 (4xVP16) or as alternately interspersed single vCD and VP16 domains, and vCD-VP28 fusions gave similarly high potency ranges and mutational tolerance patterns (**Fig. 3g**). Examination of 3-dimensional structure predictions revealed that the strongest activator fusions shared an exceptionally close alignment of vCD cores (RMSD<1.5Å) in a stable vCD α-helix, with a postulated exposed interface of charged residues (**Fig. 3i,j**). In contrast, structures of activators weakened by multiple substitutions (RMSD=2.25 Å) suggests these residues collectively, though not singly, are critical to maintain helical or inter helix stability (**Fig. 3k**).

We then tested the robustness of vCD-VP64 fusions (85-98 total amino acids) at seventeen additional promoter contexts (**Fig. 3l**; **Extended Data Figs. 5-8**). Here, these activators surpassed all unmodified tiles and that of VP64 or single-vCD, frequently matching VPR (523 amino acids), independent of dCas-effector system and cell type. Analyzing amino acid composition, we find that vCD-VP64 fusions yield emergent activator-predictive properties in a super-additive manner beyond the ranges of either VP64 or single-vCD, namely in basic residue depletion, enrichment of acidic, turn percent, B-factor values, helix percentage, and a greater negative net charge but not electrostatic potential (**Fig. 3m,n**). Thus, engineered vCD activators achieve functionally equivalent or greater potency and context-robustness compared to larger benchmarks by enhancing and exploiting the biochemical properties intrinsic to unmodified vCD (**Fig. 3i-n**).

Context-dependent durability of engineered activators

Having optimized the biochemical and biophysical properties for activator potency and robustness, we next set out to evaluate the temporal activation kinetics of engineered variants relative to gold-standard benchmarks (VPR, VP64, Rta, and P300) following transient transfection. We selected four human genes with diverse expression profiles and chromatin contexts: CD45, IFNG, CXCR4, and CD81 (**Fig. 4a**) and performed time-series analyses in HEK293T cells with protein measurements by flow cytometry or ELISA and mRNA detection by RT-qPCR. Following transfection, we first confirmed that dCasMINI-modulator plasmid expression became undetectable between 6 and 9 days post-transfection (d.p.t.) (**Fig. 4b,c**).

Benchmark activators VPR, VP64, and Rta consistently achieved high initial activation potencies at IFNG, CD45, and CXCR4 (3 d.p.t.), followed by an acute drop in activity between 3 and 9 d.p.t., concurrent with loss of mCherry and 3xFLAG detection, down to baseline levels indistinguishable from negative controls (**Fig. 4d-f, Extended Data Fig. 9a-e**). Using this threshold, we calculated potency scores based on 3- and 6-d.p.t. values, and durability scores based on values at 9 d.p.t. and later values. In contrast to benchmarks, subsets of vCD-based modulators maintained durability (target elevation >2 SD above benchmarks) in a target-dependent manner (26.1% at IFNG, 43.5% at CD45, 44.6% at CXCR4, 17.4% at CD81) (**Fig. 4e**). At CD45 and CXCR4, we find that single-vCD variants are able to sustain a ~2-fold activation plateau to 27 d.p.t. and 40 d.p.t., respectively (**Fig. 4f-h, Extended Data Fig. 9c,f**). Reasoning that potential artifacts contributing to observations of durable target gene activation could include modulator-dependent changes in cell doubling times, we measured cell division rates following S-phase cell labeling at the 8 d.p.t mark and observed no significant changes to mitotic cycle or cell division rates (**Fig. 4i**). Target genes showed variable permissiveness to durable activation in terms of the fraction of cells responding; CXCR4 activation by single-vCD domains occurred in a population-wide manner through 40 d.p.t. (**Extended Data Fig. 9d**), while the number of single-vCD activated CD45+ cells dropped precipitously after 6 d.p.t. to low but consistently stable fractions in a modulator-dependent manner (**Extended Data Fig. 9e**). In contrast to CD45, IFNG, and CXCR4, the highly expressed CD81 showed trends of activator responses largely uncorrelated to the lower-expressed genes (**Fig. 4e** and **Extended Data Fig. 9a**); benchmarks and most of the otherwise robust multi-vCD domains in fact resulted in negligible or negative expression effects, yet single-vCD variants, most often vCD_p3 and vCD_p6, mildly activated CD81 with a slower onset (**Fig. 4d**). At another highly expressed target, synthetic EF1α-GFP, vCD_p6 stably maintained GFP upregulation through 29 d.p.t., in contrast to all other variants, benchmarks, and all other viral or archaeal modulators tested (n=142) (**Extended Data Fig. 9g**). We found that activation durability is not a direct function of initial potency. Single-vCD domains, for example, display moderate potency at early time points, but most consistently maintain target elevation more durably than high-potency activators (**Extended Data Fig. 9f**).

Bromodomain dependence of vCD activator durability

We next set out to investigate potential mechanisms of action underlying sustained transcriptional activation over many successive rounds of cellular division. Given that the role of the KSHV vIRF-2 protein is in part to mediate recruitment of CBP/P300 acetyltransferase complexes to target genes[56,59], we reasoned that a mechanism by which the vCD peptide produces heritable activation through multiple mitotic cycles might involve the detection and re-deposition of instructive histone acetyl marks in a bromodomain-dependent manner. We therefore asked whether chemical inhibition of bromodomains such as BRD4 and related BET family proteins (BRDi)[46,47,60] alters the target activation response following transient activator delivery and loss in a modulator-dependent fashion.

As before, we performed a time series experiment for CD45 activation and observed loss of modulator plasmid expression and benchmark activity between 6 and 9 d.p.t. (**Extended Data Fig. 10a**). Starting at 5 d.p.t., we applied an arrayed panel of small molecule inhibitors to disrupt endogenous epigenetic enzymes including CBP/P300-associated bromodomains, WDR5/MLL histone methyltransferase, EZH2/PRC2 demethylase, histone deacetylase, and DNA methyltransferase, then monitored CD45 expression at subsequent time points. In DMSO vehicle control conditions, we again observed that single-vCD domains led to sustained increases in CD45+ cell fractions relative to other modulators and controls (**Extended Data Fig. 10a,b**). Here, we observed that BRDi drugs JQ1 and GNE049 resulted in uniformly negative effects on CD45 activation in a vCD-dependent manner, while cells that received VPR or P300 showed less drug effect (**Fig. 4j, Extended Data Fig. 10b-d, Extended Data Fig. 11**). For VPR- and P300-transfected cells, only the cocktail of both BRDi drugs appreciably reduced CD45 levels, and conditions treated with MLLi drug OICR-9429 or a cocktail of MLLi plus BRDi drugs resulted in similar depressive effects for all activators. HDACi and DNMTi drugs broadly produced additive effects on CD45 levels independently of modulator (**Extended Data Fig. 10** and **Extended Data Fig. 11**). Finally, we tested the effects of BRDi drugs JQ1 and GNE049 on dCas-mediated activation of a different gene, CXCR4, finding a vCD-specific reduction in CXCR4 expression of over 50% (**Extended Data Fig. 10e**). Thus, selective inhibition of CBP/P300-associated BRD bromodomains abrogated the activation durability of vCD, implicating a histone acetylation-mediated mechanism for transmission of mitotically stable CD45 activation by vIRF-2 based modulators[60].

Discussion

In this study we performed the first direct comparison of transcription-modulatory potencies among tens of thousands of peptides sourced from divergent taxonomic lineages of extremophilic and mesophilic environments. We uncovered a marked activator potency bias for viral peptides and suppressor potency biases for human peptides and for thermophilic and acidophilic archaeal peptides. Analysis of amino acid composition revealed distinct biochemical strategies by which these taxa have evolved unique transcriptional capabilities, forming the basis for predictive activator selection criteria previously unappreciated in the field. Using machine learning and insights from structural predictions, we identified biochemical and sequence-level peptide features of activation allowing us to develop a rational engineering strategy that enhanced the magnitude and durability of transcriptional activation at diverse human gene targets. Importantly, we report the discovery of activators that induce the most durable and mitotically stable gene activation reported to date. Our discovery pipeline provides a predictive rubric for the systematic engineering of hypercompact activators, as small as 64 amino acids, from natural proteome sources, yielding superior potency and kinetics profiles that broadly expand the potential therapeutic landscape.

CRISPR-based regulation of gene expression holds great promise for biotechnological application in human cell engineering and gene therapy. The ability to modulate specific genomic loci in a programmable manner with discrete activation or suppression domains is a powerful tool in the arsenal of genetic medicine. As such, there has been a recent explosion in studies describing the identification and characterization of domains capable of eliciting desired changes in gene expression[17,18,19,27]. Typically, these studies have constrained the search space for such activities to known regulatory domains, screening with reporter genes at synthetically controlled but uncharacterized epigenomic contexts. This approach, while productive, risks limiting the utility of such domains as it does not account for the heterogeneity of chromatin contexts among the hundreds or thousands of disease-relevant target genes in human cells.

Programmable induction of heritable gene silencing, but not activation, has typically been achieved with large multi-partite enzyme complexes. Here, we have taken the first steps to challenge this paradigm by identifying and exploiting a previously unknown viral transactivation domain capable of durable activation. We functionally implicate a bromodomain-dependent mechanism for mitotically durable target gene activation by vIRF2-based peptides, stimulating future investigations. The vIRF2 core domain may act as an “initiator”, similar to endogenous factors that establish cellular identity, modifying the chromatin structure to confer accurate transmission of gene expression programs. Viral activators may circumvent the obstacle of human transcriptional cofactor scarcity, negating the rate-limiting competition for co-activating proteins in human cells[61].

We have shown that the organismal provenance of tiled peptide fragments confers categorical differences in transcription-modulatory function as defined by 1) potency, 2) context-independent robustness, and 3) durability of activity, facilitating future screening efforts for accelerated discovery and de novo design of epigenetic engineering tools[62]. We and others have begun to characterize the extensive and complex interplays of chromatin state, cell type, and delivery modality contexts in producing a desired transcriptional effect, making the discovery of rare context-robust core modulator domains critically valuable to the field. Here we

have laid out a framework for, and functionally validated, the sequence-based *a priori* prediction of modulatory functions based on key biochemical parameters, an advance that may accelerate tool development from unannotated proteomes with targeted screening methodologies.

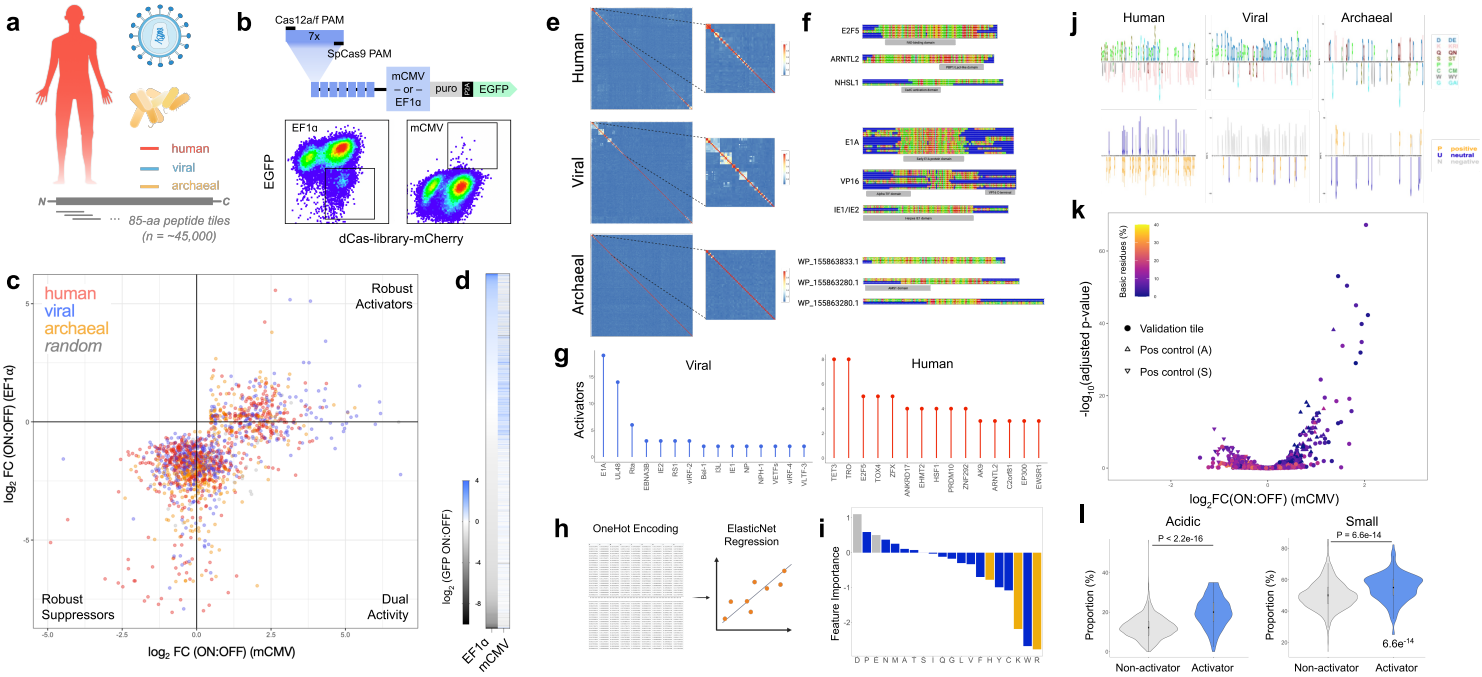


Fig. 1: Identification of transcriptional modulator domains by high-throughput screening

a, Schematic depiction of pooled human, viral, and archaeal coding libraries and peptide tiling strategy. **b,** Schematic depiction of custom mCMV-GFP and EF1α-GFP promoters targeted for pooled modulator screening. Representative FACS plots of EF1α-GFP (left) and mCMV-GFP (right) K562 cells following lentiviral transduction of the pooled dCas9-library fusions. Dotted boxes represent OFF and ON gates for EF1α-GFP and mCMV-GFP cell lines respectively. **c,** Scatterplot comparing activation strengths ($\log_2$ fold-change enrichment of barcodes in ON vs OFF bins) of top-significance modulators (n=560 activators, n=1,330 suppressors) from mCMV-GFP and EF1α-GFP screens respectively . Robust activators, suppressors, and dual activity hit quadrants are indicated. **d,** Heatmap of activation strengths ($\log_2$ fold-change enrichment of barcodes in ON vs OFF bins) for statistically significant modulators indicating activation versus suppression activities at mCMV-GFP and EF1α-GFP promoters. **e,** Sequence homology clustering of human (top, n=209), viral (middle, n=177), and archaeal (bottom, n=154) activators identified in the mCMV-GFP high-throughput screen identifies clusters of modulators with sequence similarity. **f,** Selected sequence alignments for human (top), viral (middle), and archaeal (bottom) sequence homology clusters illustrating common functional domains within activator clusters. **g,** Ranked lists of viral (left) and human (right) genes that activation hits originated from. **h,** Schematic depicting featurization of peptide sequences by OneHot encoding and subsequent training of a generalized linear regression model to predict activation strength based on peptide sequence alone. **i,** Extracted feature importance from top generalized linear regression model detailing which residue types were predictive of activation strength (gray: acidic, blue: neutral, gold: basic). **j,** Enrichment of amino acid residues composing human (n=209), viral (n=177), and archaeal (n=154) activators colored by residue type (top) and charge (bottom) calculated using Fisher's exact test. **k,** Volcano plot illustrating screening results from the mCMV-GFP validation screen colored by composition of basic residues. Validation tiles are represented by circles, positive controls (activation) by triangles, and positive controls (suppression) by inverted triangles. **l,** Violin plots comparing proportion of acidic residues (left) and small residues (right) in activation hits (blue) and non-activator peptides (gray). P-values reported are based on Wilcoxon rank-sum testing.

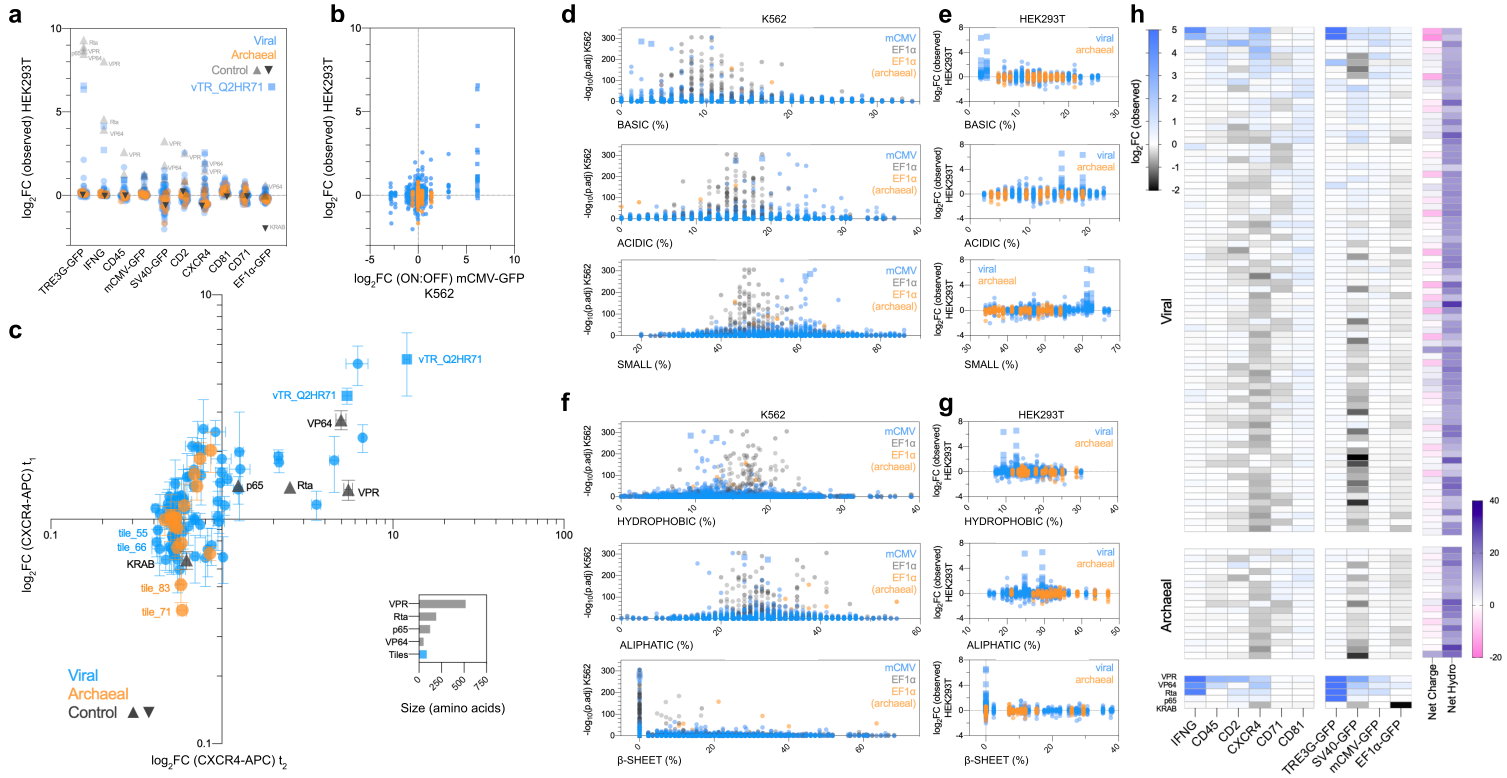


Fig. 2: Biochemical features predict robust and context-independent activator domains

a, Functional testing by arrayed dCasMINI-mediated recruitment of selected modulator screen tiles (n=95) at ten target genes in HEK293T cells. Dots represent observed modulator activity per target (mean fold-change of 3 or more replicates) in observed protein expression, relative to non-targeting sgRNA (sgNT) and dCasMINI recruited without modulator fusion (dCasMINI) conditions. Viral (blue), archaeal (orange), benchmark controls (triangle), and top viral hit tiles (square) are indicated. **b,** Comparison of observed mean activities in HEK293T cells (y-axis) against computed barcode enrichment scores (mCMV-GFP ON:OFF) from pooled high-throughput screens in K562 cells (x-axis). **c,** Representative functional testing experiment by flow cytometry at one of the ten targeted contexts shown in **(a)**. Fold-changes in CXCR4-APC fluorescence at 3 days post-transfection (d.p.t.) (t_1, x-axis) versus 7 d.p.t. (t_2, y-axis) by each of 95 viral and archaeal modulators and benchmark controls VPR, VP64, Rta, p65, KRAB (mean±SEM). Relative modulator peptide sizes (amino acids) plotted for reference (inlaid). **d,e,** Biochemical feature scores (x-axis) plotted against hit significance (ON:OFF, adj. p-val) for the full modulator libraries in mCMV-GFP and EF1α-GFP K562 screens **(d)** or against observed activities in HEK293T cells (mean fold-change per target) **(e)**. Features predicting activators are shown. **f,g,** Biochemical feature scores (x-axis) plotted against hit significance (ON:OFF, adj. p-val) for the full modulator libraries in mCMV-GFP and EF1α-GFP K562 screens **(f)** or against observed activities in HEK293T cells (mean fold-change per target) **(g)**. Features predicting suppressors are shown. **h,** Ranked heatmap summary of context robustness in modulator activities in arrayed HEK293T cell experiments, with observed protein fold-changes per target (left) and scores for net charge and hydrophobicity (right). Gradients of mean observed activation (blue), suppression (black), and no effect (white) by each modulator at each locus are indicated.

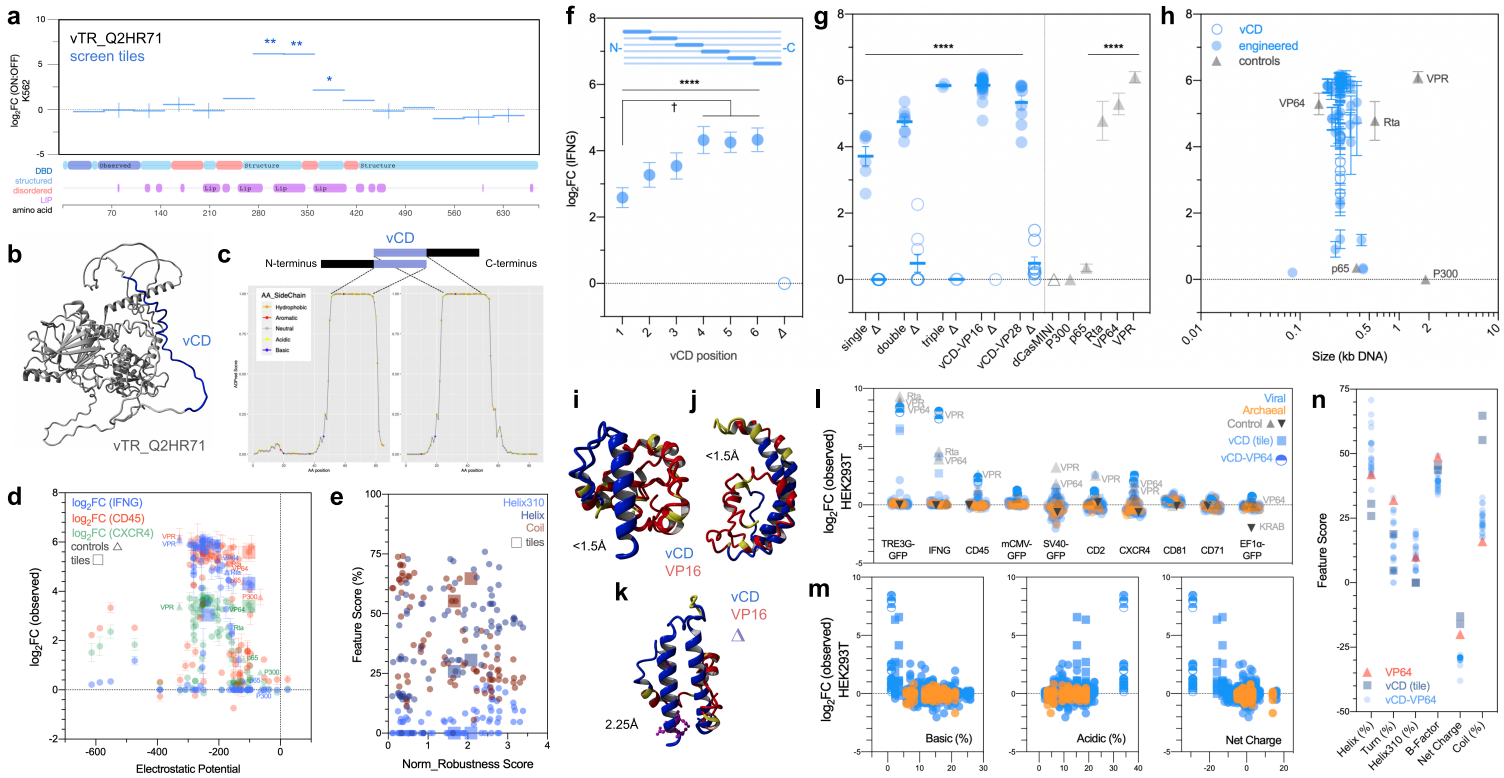


Fig. 3: Engineering of vCD-containing modulators that outperform benchmarks

a, Positional mapping of vIRF2 peptide tiles (x-axis) and respective activation strengths (log_2 fold-change of ON:OFF barcode enrichment scores, mean±SEM, **$P<1.72\times10^{-61}$, **$P<1.72\times10^{-61}$) in the initial mCMV-GFP K562 pooled screen (top). MobiDB schematic (below) indicates vIRF2 (Q2HR71) native protein features including structured (blue) versus disordered (red) regions, DNA-binding domain (DBD, dark blue), and linear interacting peptides (LIP, purple). **b,** 3-dimensional structure prediction of the full-length native vIRF2 (vTR_Q2HR71) and predicted vIRF2 core domain (vCD) (blue). **c,** Minimal vIRF2 core domain (vCD) prediction by ADPred. The 32aa vCD shared by the top two screened vIRF2 tiles, shown in **(a)**, is enriched for acidic and hydrophobic residues. **d,** Activation potencies (mean±SEM) of engineered vCD-based modulators (n=101) at IFNG, CD45, and CXCR4 at 3 d.p.t., relative to electrostatic potential of each modulator. **e,** Normalized robustness score at IFNG, CD45, and CXCR4 (y-axis), relative to percent residue enrichments for helix and coil structures. **f,g,** Activation potencies at IFNG (mean±SEM; *P<0.001; †P<0.05, one-way ANOVA) of single-vCD modulators with the vCD at six positions of increasing distance from the dCasMINI and respective mutants **(f)**, and multi-vCD fusions and respective mutants with benchmark activators (gray) **(g)**. **h,** Activation potencies at IFNG (mean±SEM) of all engineered vCD-based modulators (n=101) (blue) and benchmarks (gray) relative to peptide coding size (DNA kb) (x-axis), with high-potency domains as compact as 64 amino acids. **i,j,** 3-dimensional structure predictions of high-potency vCD modulators with close alignment of vCD cores (RMSD <1Å) contributing to form a stable vCD helix, with the postulated interface of charged residues exposed to the solute. **k,** 3-dimensional structure predictions of weak activator variants with mutations C4L, L5D, M7D, and L19D (RMSD 2.25Å). **l,** Mean activation potencies of two vCD-VP64 modulators at ten target contexts, relative to non-engineered viral (blue) and archaeal (orange) screen tiles (n=95). **m,** Biochemical feature scores (x-axis) plotted against mean activation potencies at all ten target contexts (y-axis) of vCD-VP64 fusions and non-engineered modulators. **n,** Changes in biochemical feature scores of vCD-VP64 fusions relative to scores of either VP64 or single-vCD modulators alone.

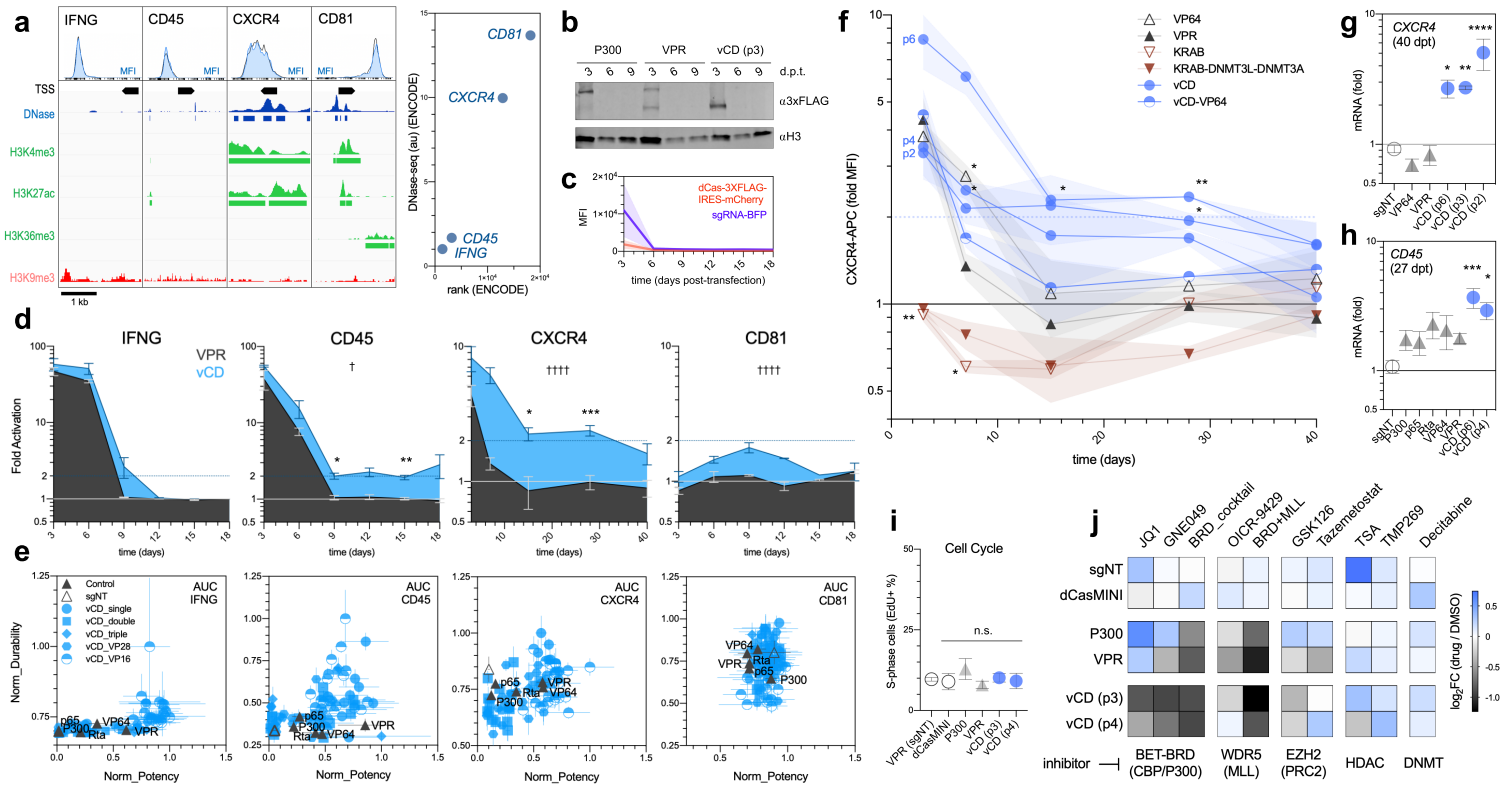
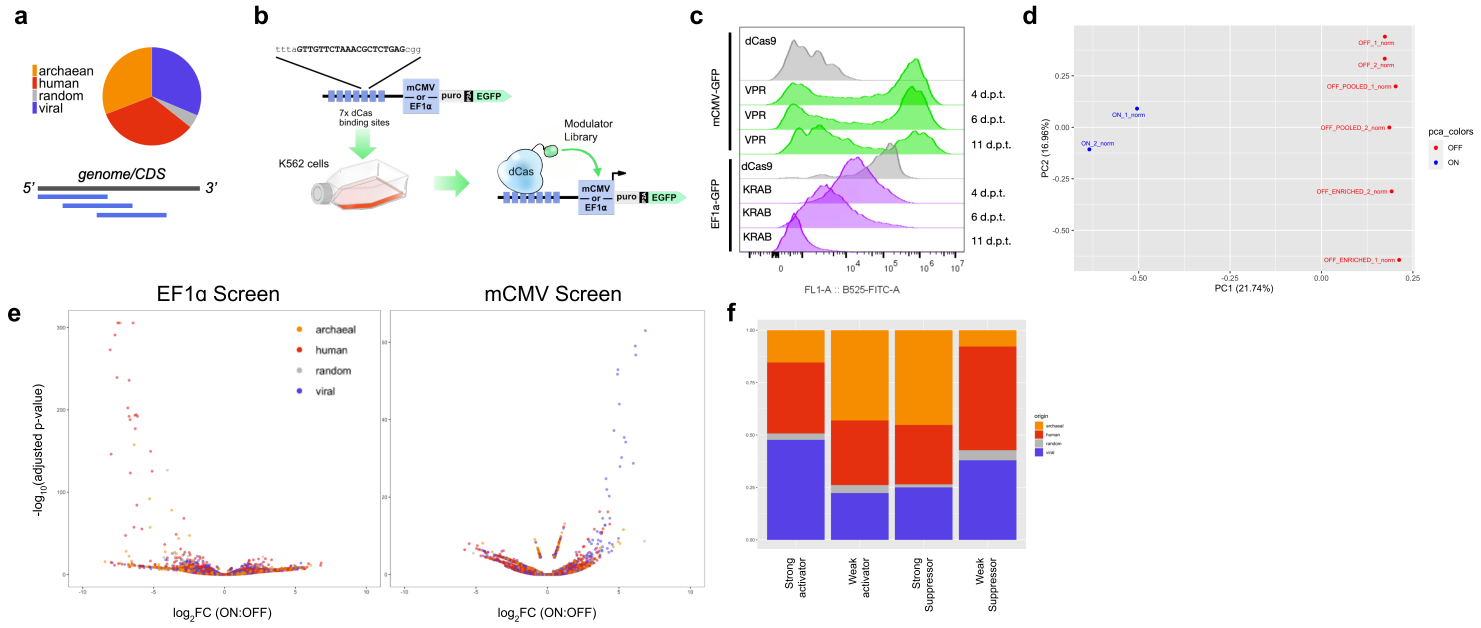


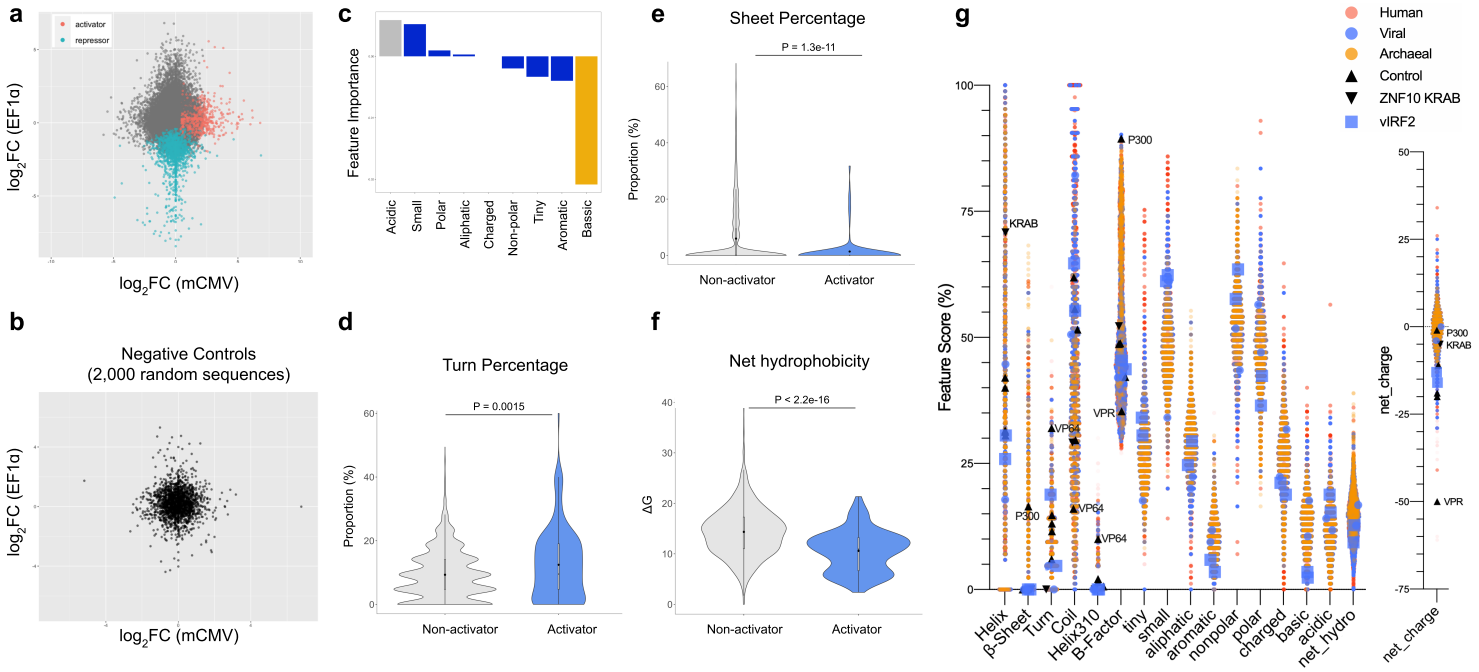
Fig. 4: Context-dependent durability of engineered activators

a, Chromatin contexts and HEK293T flow cytometric baselines of surface protein expression (left) and ENCODE DNAse-seq rankings (right) for CD45, IFNG, CXCR4, and CD81 genes targeted for activation time series measurements. **b,** Immunostaining to detect expression of 3XFLAG-tagged dCasMINI-modulator fusions following transient plasmid lipofection in HEK293T cells at 3- and 9-d.p.t. **c,** Detection of plasmid loss between 6- and 9-d.p.t. by flow cytometry for dCas-modulator-IRES-mCherry and sgRNA-BFP plasmids. **d,** Target fold changes in activation (mean±SEM) at each locus following transient recruitment of an individual vCD-based modulator versus VPR benchmark for the full time series: 3-to-29 d.p.t. for CXCR4 and 3- to 18-d.p.t. for IFNG, CD45, and CD81 (significance from VPR for the series, †P<0.05, ††††P<0.0001, and individual time points *P<0.05, **P<0.01, ***P<0.01, two-way ANOVA). Time series AUC indicated for normalized activator score calculations. **e,** Normalized durability scores (9 d.p.t. forward) at all four targets (±SD) (y-axis) versus normalized potency scores (3 and 6 d.p.t.) (±SD) (x-axis) for benchmarks VPR, VP64, Rta, p65, and P300 (black) and vCD modulators (n=101) (blue) with engineering classes indicated by shape. **f,** CXCR4 full time series comparing fold-change activations (CXCR4-APC geometric MFI, mean±SEM, significance from VPR, *P<0.05, **P<0.01, two-way ANOVA) at each time point for single-vCD variants and controls. **g,** *CXCR4* mRNA detection (mean±SEM, significance from sgNT, ****P<0.0001, **P<0.01, *P<0.05, one-way ANOVA) at 40 d.p.t. by indicated modulators following transient plasmid lipofection in HEK293T cells. **h,** *CD45* mRNA detection (mean±SEM, significance from sgNT, ***P<0.0001 *P<0.01, one-way ANOVA) at 27 d.p.t. by indicated modulators following transient plasmid lipofection in HEK293T cells. **i,** Cell cycle analysis of S-phase labeled cells following pulse-chase of EdU and flow cytometric EdU detection at 8 d.p.t. of indicated modulator plasmids. **j,** Heatmap summary of normalized drug effects on CD45 activation at 9 d.p.t. of indicated vCD and benchmark modulators in the presence of selective epigenetic inhibitor drugs. Observed mean activation levels in DMSO vehicle-treated control served as normalization denominators to quantify drug effects on CD45-APC fluorescence intensity. Modulator-dependent abrogation of CD45 activation in vCD-transfected cells in the presence of either BET bromodomain inhibitor alone, JQ1 or GNE049, and non-modulator-dependent abrogation of CD45 activation in vCD-, VPR-, and P300-transfected cells in the presence of equimolar cocktail of JQ1, GNE049, and histone H3K4 methyltransferase (MLL) inhibitor OICR-9429.



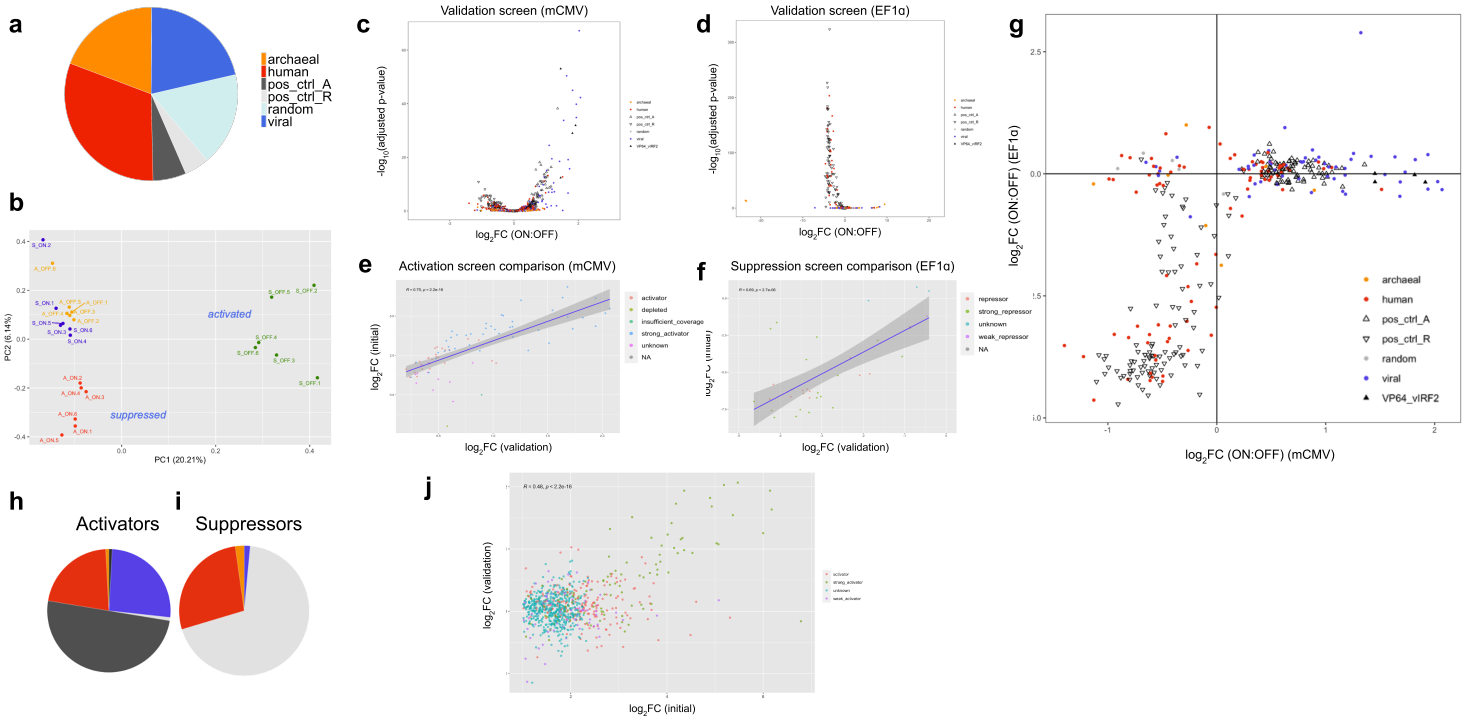
Extended Data Fig. 1: Pooled screening for human, viral, and archaeal modulators

a, Illustrations of initial high-throughput library composition and tiling strategy. 85 amino acid fragments were encoded as 300-mer DNA oligonucleotides, each labeled by a unique 12-mer DNA barcode. 549 human nuclear proteins were tiled with 50% sequence overlap to generate 16,139 oligos. Viral full genomes, human viral transcriptional regulators (hVTRs)[26], metagenomic viral peptide predictions, and the archaeon full genome were tiled at 66.6% overlap to generate 27,799 oligos. GC content-matched scrambled sequences were also included as random negative controls. **b,** Schematic of modulator recruitment strategy and custom transcriptional reporter constructs for screening activators (mCMV promoter with default GFP-OFF activity) and suppressors (EF1α promoter with default GFP-ON activity). Each reporter construct contains seven identical sgRNA landing sites with dual-PAM sites for targeting via dCas9 or dCasMINI. **c,** Reporter validations in pilot studies monitoring flow cytometric GFP fluorescence in mCMV-GFP and EF1α-GFP K562 cells bearing lentiviral integration of sgRNA-BFP vectors, sampled at indicated time points post-lentiviral transduction (d.p.t.) of dCas9-VPR, dCas9-KRAB, or dCas9 without modulator fusion. **d,** PCA plot of barcodes detected in pooled screen samples with separation of ON and OFF samples. **e,** Volcano plots of screen hits in EF1α-GFP suppression screens (left) and mCMV-GFP activation screens **(**right**)**, colored to illustrate library sources as human (red), viral (blue), archaeal (orange), or random (gray). **f,** Distributions of human, viral, archaeal, and random origins of strong/weak activators/suppressors defined by the highest and lowest quartiles of activators/suppressors respectively.



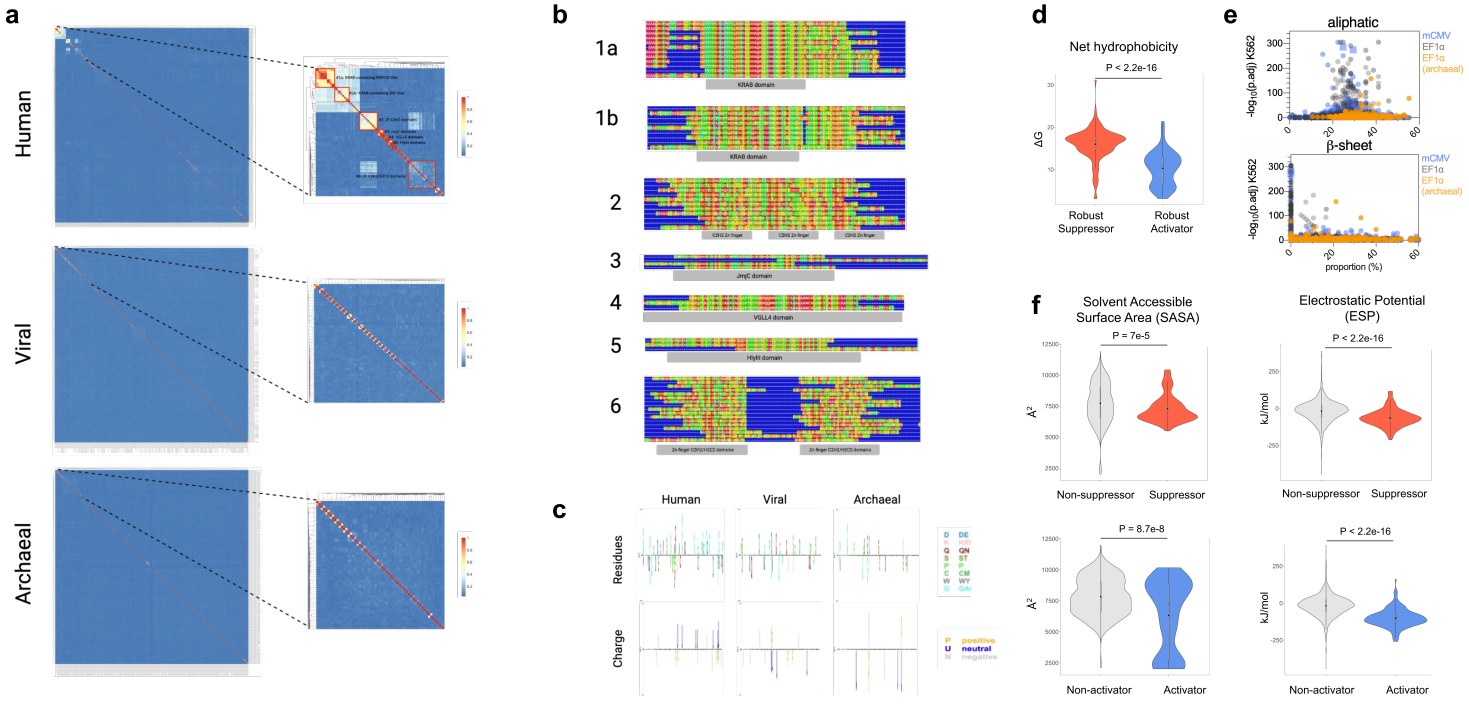
Extended Data Fig. 2: Modulator context robustness in pooled screens

a,b, Transcriptional effects (fold-change in ON:OFF barcode enrichment) of each modulator tile in both mCMV-GFP (x-axis) and EF1α-GFP (y-axis) screens. Colors distinguish robust activators and suppressors (**a**), or random negative control tiles **(b)**. **c,** Extracted feature importance from top generalized linear regression model detailing which residue types were predictive of activation strength (gray: acidic, gold: basic, blue: other). **d-f,** Violin plots comparing turn percentage (**d**), sheet percentage (**e**), and net hydrophobicity (**f**) in activation hits (blue) and non-activator peptides (gray). P-values reported are based on Wilcoxon rank-sum testing. **g,** Biochemical feature distributions among screened human (red), viral (blue), and archaeal (orange) peptides. Benchmark modulators (triangle), vIRF2 (vTR_Q2RH71) high-potency tiles (square), and other potent viral activators (larger blue circle) are indicated.



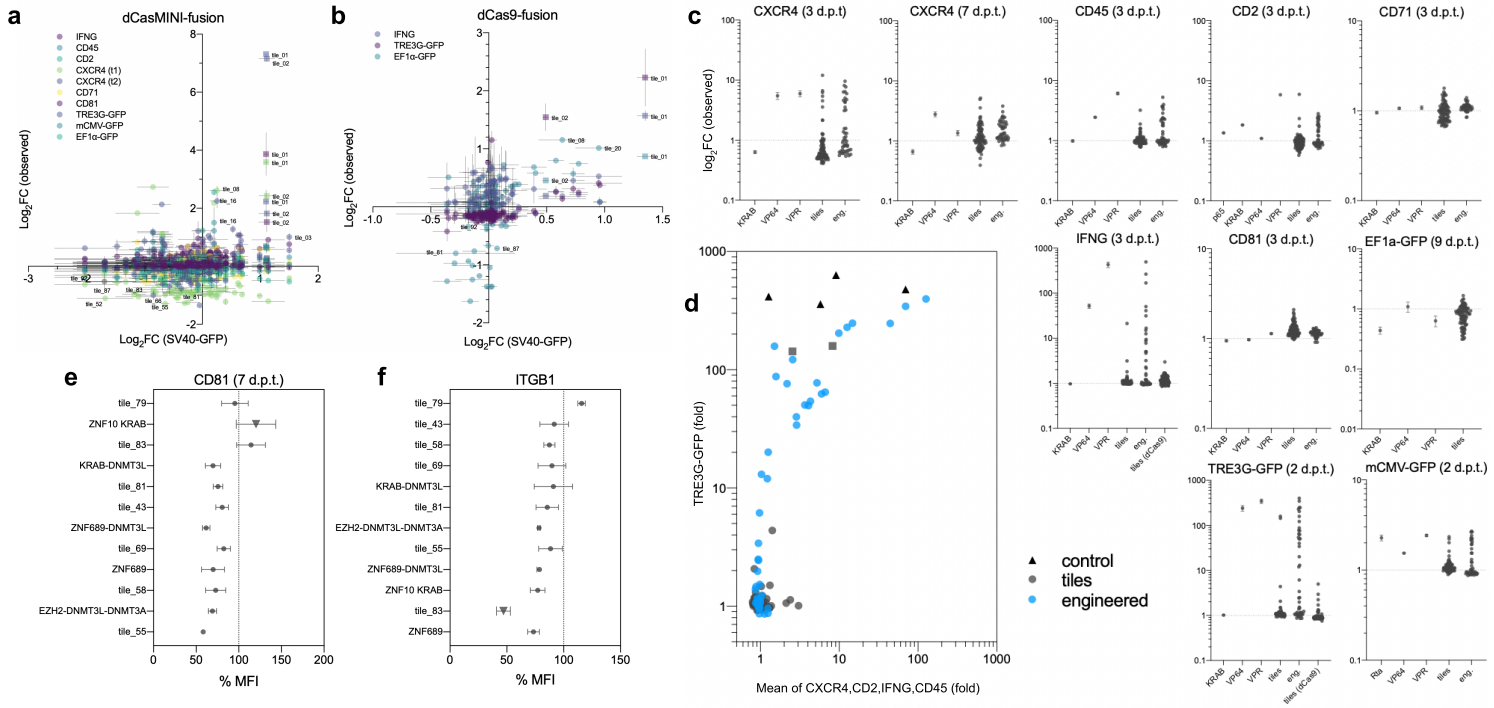
Extended Data Fig. 3: Modulator validation by pooled sub-library validation screening

a, Sub-library composition for the validation screens of putative modulators from the full-library screen (954 activators, 1,228 suppressors, and 22 dual-activity tiles) and selected engineered activators containing vIRF2-VP16 fusions. 949 negative controls included scrambled sequence tiles, tiles depleted in both full-library mCMV-GFP and EF1α-GFP screens, and tiles initiated with stop codons. Positive control tiles were taken from published data[10,18]. **b,** Principal component analysis (PCA) of modulator validation libraries. 6 replicate samples were analyzed for each of four conditions: mCMV-GFP-ON, mCMV-GFP-OFF, EF1α-GFP-ON, and EF1α-GFP-OFF. **c,d,** Correlation of screen hits detected in sub-library validation (x-axis) and full-library (y-axis) for both mCMV-GFP **(c)** and EF1α-GFP **(d)** screens. **e,f,** Volcano plots indicate screen hits for mCMV-GFP **(e)** and EF1α-GFP **(f)** screens. Colors differentiate tile sources. Shapes differentiate test tiles (circle) from published positive controls (clear triangle) or engineered vIRF2-VP16 fusions (filled triangle). **g,** Context robustness of validation screen hits at both mCMV-GFP and EF1α-GFP promoters. Dots are colored by taxonomic origin (red: human, blue: viral, orange: archaeal, gray: random) and positive controls for activators and suppressors are represented by triangles and inverted triangles respectively. **h,i,** Distributions of tile provenance of validated mCMV-GFP activators **(h)** and EF1α-GFP suppressors **(i)**. **j,** Correlation of mCMV-GFP activators in full-library (x-axis) and sub-library validation (y-axis), colored by classification as strong or weak activators by barcode enrichment.



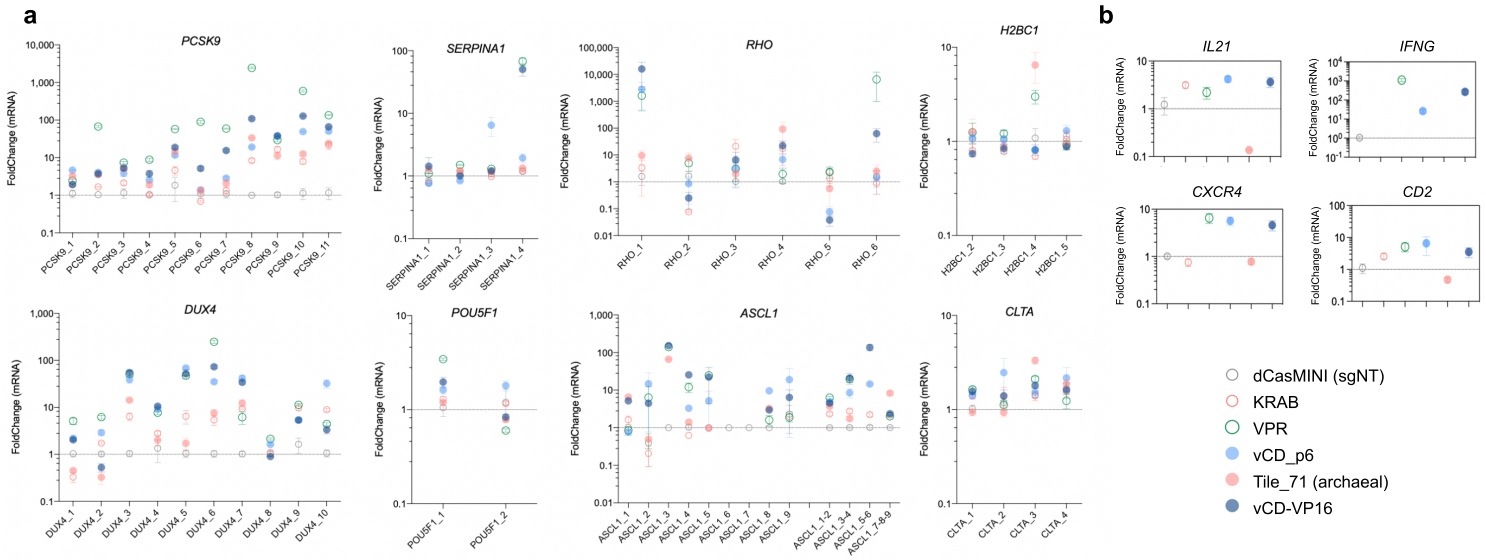
Extended Data Fig. 4: Clustering, alignment, and biochemical enrichment of suppression tiles

a, Sequence homology clustering of human (top), viral (middle), and archaeal (bottom) suppressors identified in the EF1α-GFP high-throughput screen. **b,** Selected sequence alignments for human (top), viral (middle), and archaeal (bottom) sequence homology clusters illustrating common functional domains within suppressor clusters. **c,** Enrichment of amino acid residues composing human, viral, and archaeal suppressors colored by residue type (top) and charge (bottom) calculated using Fisher's exact test. **d,** Violin plot comparing net hydrophobicity in robust activators (blue) and robust suppressors (red). P-values reported are based on Wilcoxon rank-sum testing. **e,** Biochemical feature scores (x-axis) plotted against hit significance (ON:OFF, adj. p-val) for the full modulator libraries in mCMV-GFP and EF1α-GFP K562 screens. Scores in mCMV-GFP (blue) and EF1α-GFP (gray). vIRF2 activator hits at mCMV-GFP (triangles), and archaeal suppressor hits at EF1α-GFP (orange), are indicated. **f,** Violin plots comparing solvent accessible surface area (SASA, left) and electrostatic potential (ESP, right) in suppressors (red) vs. non-suppressors (gray) (top) and activators (blue) vs. non-activators (gray) (bottom). P-values reported are based on Wilcoxon rank-sum testing.



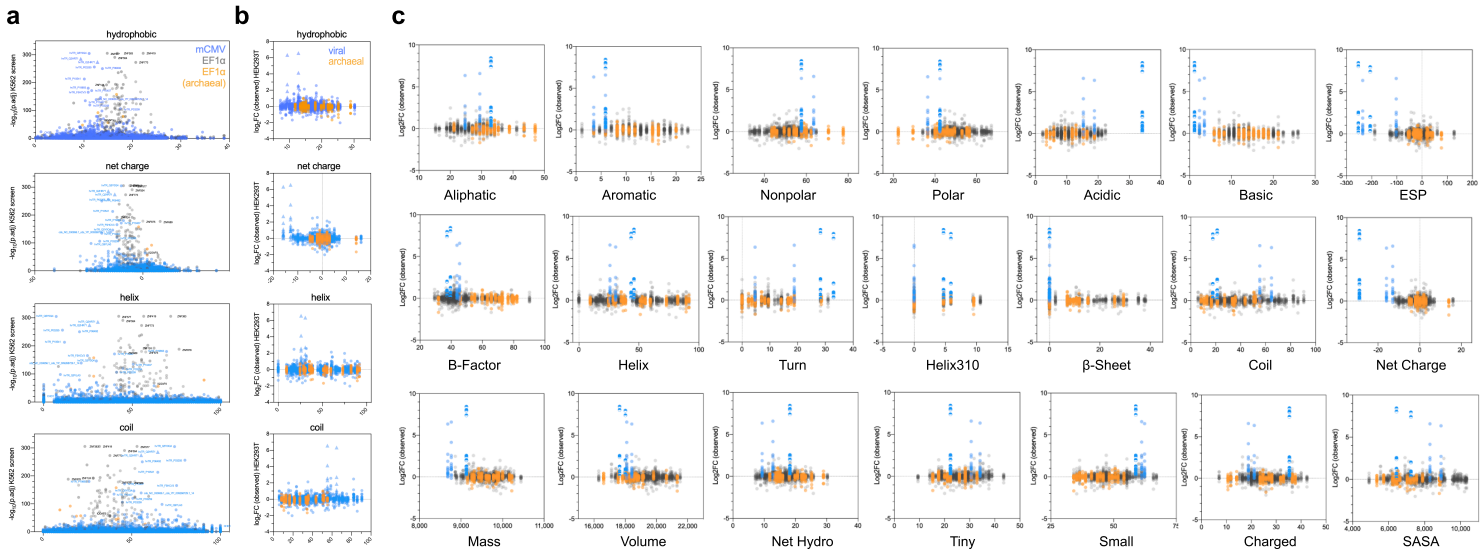
Extended Data Fig. 5: Modulation of protein expression in individual recruitment assays

a,b, Observed transcriptional effects in HEK293T cells **(Fig. 2)** of selected viral and archaeal modulator screen tiles (n=95) targeting synthetic and endogenous genes, comparing modulation of SV40-GFP (x-axis) against modulation at indicated targets (y-axis) when fused to dCasMINI **(a)** or dCas9 **(b)**. Dots represent observed modulator activity per target (mean±SEM of 2 or more replicates) in observed protein expression, relative to non-targeting sgRNA (sgNT) and dCasMINI recruited without modulator fusion (dCasMINI) conditions. **c,** Individual experiments at various targets shown in **(a)** testing viral and archaeal modulator screen tiles (n=95), benchmark modulators VPR, VP64, Rta, p65, KRAB, and engineered variants based on viral activator tiles (tile_1 and tile_2) from vIRF2 (vTR_Q2RH71). **d,** Correlation of TRE3G-GFP activation against the mean activation of four endogenous genes indicated, per modulator. **e,f,** Suppression of endogenous genes CD81 **(e)** and ITGB1 **(f)** at 7 d.p.t. by selected dCasMINI fusions to human, viral, and archaeal tiles, and KRAB, KRAB-DNMT3L, ZNF689, ZNF689-DNMT3L, EZH2-DNMT3L-DNMT3A fusions. Each modulator was co-transfected with multiplexed sgRNA plasmids as indicated.



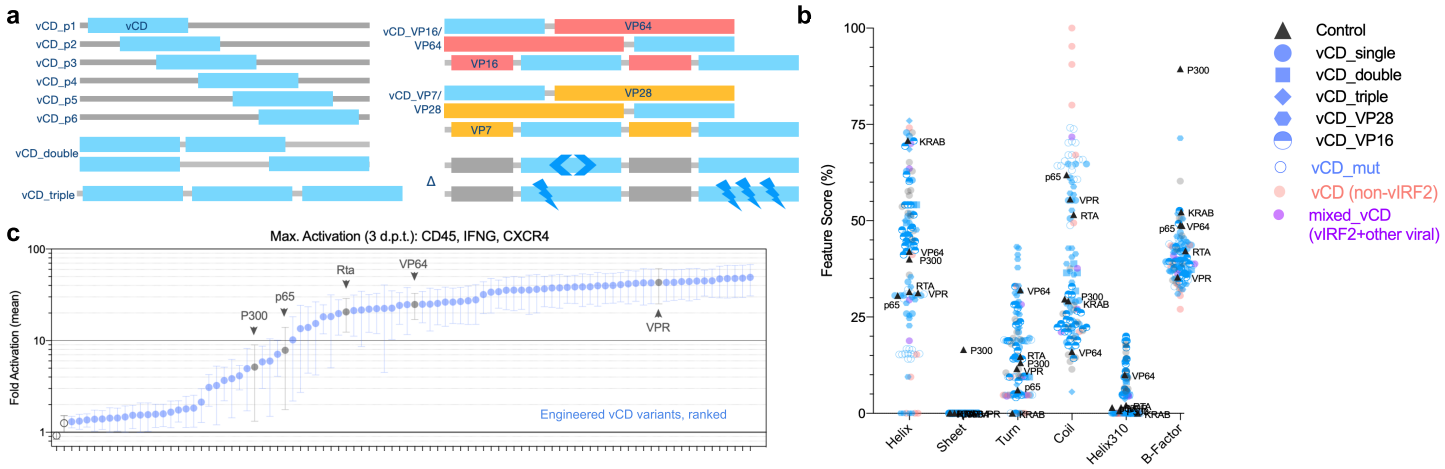
Extended Data Fig. 6: Modulation of mRNA expression in individual recruitment assays

a,b, Transcriptional effects at indicated genes HEK293T cells sampled at 3 d.p.t. following plasmid transfections of dCasMINI without modulator (empty circle), or dCasMINI fusions to KRAB (red, empty), VPR (green, empty), vIRF2 screen tile (light blue), archaeal suppressor tile_71 (red), or engineered vCD-VP16 fusion (dark blue). dCasMINI modulator plasmids were co-transfected with either individual sgRNA plasmids **(a)** or multiplexed sgRNA plasmid pools **(b)**.



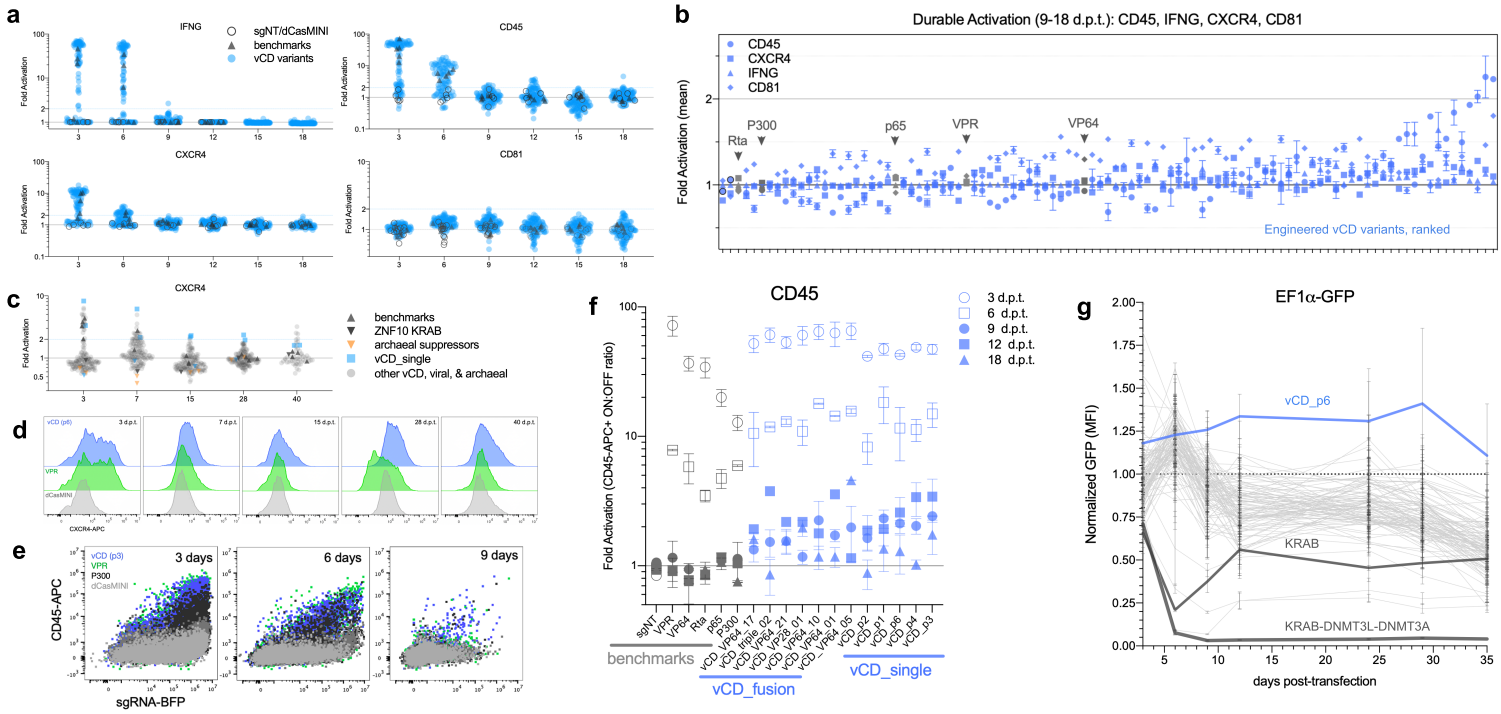
Extended Data Fig. 7: Feature correlations with robust modulator performance

a,b, Side-by-side comparison of K562 screen predictions for feature importance on modulator activity and observed functional testing in HEK293T cells. **(a)** Biochemical feature scores (x-axis) plotted against hit significance (ON:OFF, adj. p-val) for the full modulator libraries in mCMV-GFP and EF1α-GFP K562 screens. Scores in mCMV-GFP (blue) and EF1α-GFP (gray). vIRF2 activator hits at mCMV-GFP (triangles), and archaeal suppressor hits at EF1α-GFP (orange), are indicated. **(b)** Indicated biochemical feature scores (x-axis) are plotted against observed protein activation fold-changes (mean of technical triplicates) for 95 selected viral (blue) and archaeal (orange) modulators in ten experiments of dCasMINI-modulator fusions co-transfected in HEK293T cells with targeting sgRNAs for either IFNG, CD45, CXCR4, CD2, CD71, CD81, TRE3G-GFP, SV40-GFP, mCMV-GFP, or EF1α-GFP. vIRF2 activator hits at mCMV-GFP are indicated (triangles). **c,** Biochemical feature differences between vIRF2 activator tiles and engineered vCD-VP16 fusions. Feature scores (x-axis) are plotted against observed protein activation fold-changes, adding the activity and scores of two engineered vCD-VP16 fusions (blue semi-circles) and highlighting the original vIRF2 screen tiles (blue). Archaeal suppressor hits (orange) are indicated.



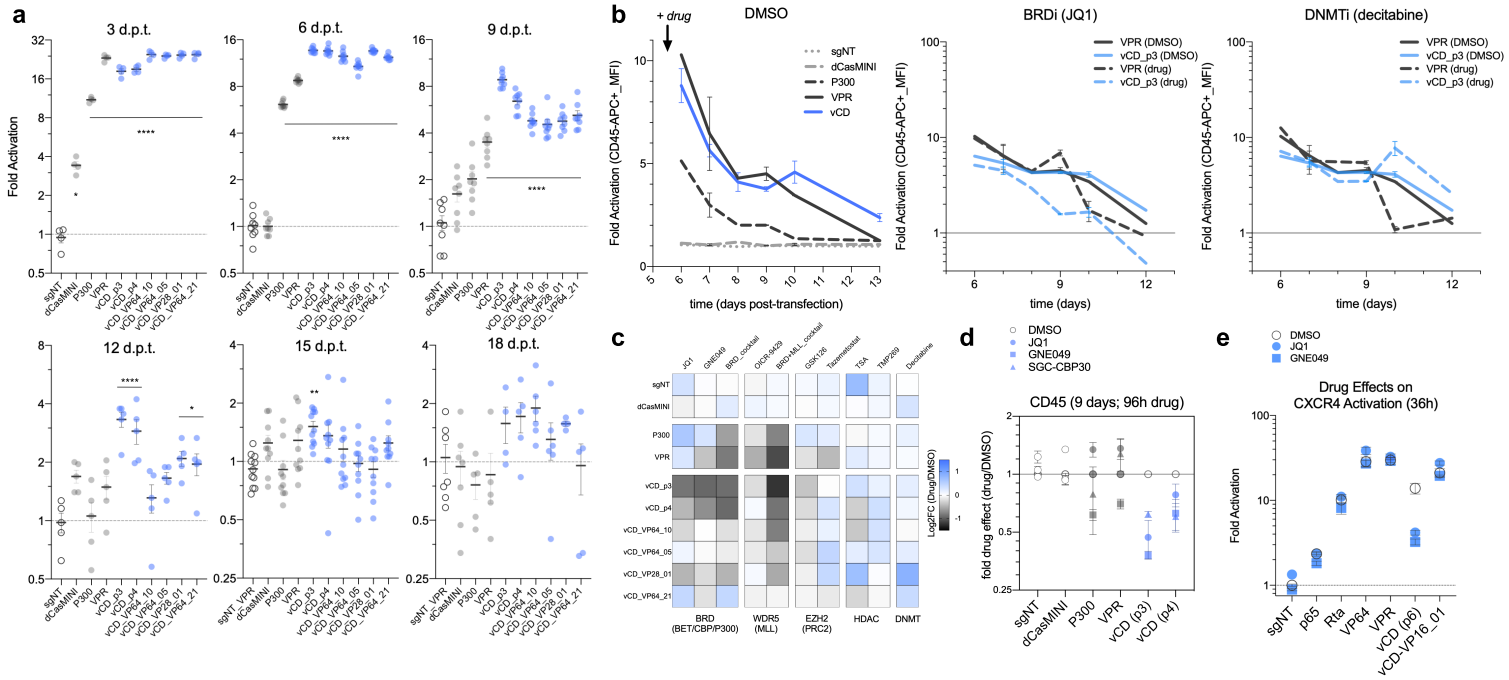
Extended Data Fig. 8: Domain-based engineering of hypercompact activators

a, Schematics illustrate engineered variants of the 32 amino acid vIRF2 core domain (vCD) testing various configurations of N-to-C position, fusion partners, linker sequences, vCD inversions, and vCD mutations. **b,** Biochemical properties altered by engineering the vCD domain. **c,** Summarized activation potencies following co-transfection in HEK293T cells with targeting sgRNA plasmids and dCasMINI fusions to all engineered vCD variants (n=101) relative to positive controls VPR, VP64, Rta, p65, and P300. dCasMINI without modulator and paired with targeting sgRNAs, and dCasMINI-VPR paired with non-targeting sgRNA served as negative controls for normalization and fold-change calculations. Y-axis values are an averaged protein activation score for three experiments targeting CD45, CXCR4, and IFNG (mean±SEM).



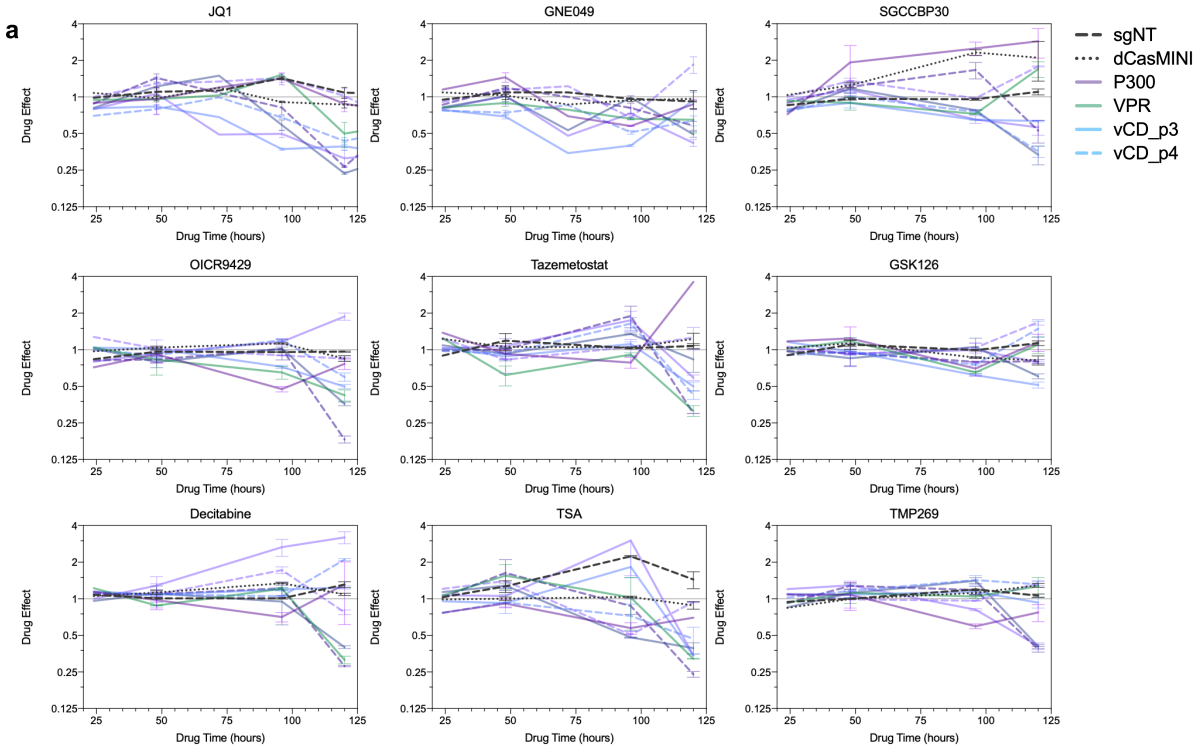
Extended Data Fig. 9: Prolonged activation kinetics of hypercompact activators

a, Arrayed durability screening in HEK293T cells following transient lipofection of dCasMINI-modulator fusions and targeting sgRNA plasmids. Observed fold-changes per modulator in secreted protein by ELISA (IFNG), APC MFI (CXCR4, CD81), and normalized fraction of CD45-APC+ cells (CD45) in HEK293T cells at indicated time points (x-axis). Dots represent means per modulator. Positive controls (VPR, VP64, Rta, p65, P300) were transfected in 5-10 technical replicates. Normalization (negative) controls (n=29 replicates per experiment) were dCasMINI without modulator fusion with (n=5) and without (n=2) targeting sgRNAs (sgT), and non-targeting sgRNA (sgNT) with dCasMINI-modulator fusions to VPR (n=4), VP64 (n=2), Rta (n=2), p65 (n=2), P300 (n=2), vCD_single (n=4), vCD-VP16 (n=4), and vCD-VP64 (n=2). Test modulators were vCD-based variants (n=101) transfected in 2-3 technical replicates. **b,** Summarized durabilities per target for all modulators at post-plasmid time points (mean±SEM of fold changes from 9-18 d.p.t.) at IFNG (triangle), CD45 (circle), CXCR4 (square), and CD81 (diamond). Ranked by the total mean of values across four targets, i.e.context-robustness of durability. Benchmarks indicated (gray, arrows). **c,d,** Independently repeated CXCR4 experiment **(Fig. 4d-g)** testing engineered modulators (n=48), viral and archaeal screen tiles (n=95), and benchmarks. Observed mean fold-changes **(c)** in CXCR4-APC shown per modulator **(d)** Representative CXCR4 flow cytometry data at 3-40 d.p.t. for VPR (green), vCD_p6 (blue), dCasMINI without modulator fusion and targeting sgRNAs (gray). **e,** Representative CD45 flow cytometry data at 9 d.p.t. for VPR (green), P300 (black), vCD_p3 (blue), dCasMINI without modulator fusion and targeting sgRNAs (dark gray), and VPR with non-targeting sgRNA (light gray). **f,** CD45 activation data (normalized fraction of CD45-APC+ cells) at indicated time points for selected modulators (blue) and benchmarks (gray) following transient lipofection. Early time points where modulator and sgRNA plasmids remain expressed (empty symbols) distinguished from later time points with undetectable plasmid expression (filled symbols). **g,** Observed fold-changes in EF1α-GFP fluorescence in HEK293T cells (mean±SEM) measured by flow cytometry from 3-to-35 d.p.t. following transient lipofection of dCasMINI-modulator fusions and targeting sgRNA plasmids. Super-activation of EF1α-GFP by vIRF2 screen tile (blue) is sustained for 29 days, while suppression is sustained to varying degrees by KRAB, KRAB-DNMT3L-DNMT3A, and a subset of viral and archaeal modulators (colors). Relative dCas-modulator-mCherry expression is shown (broken red line) to indicate presence of modulator plasmid expression at each time point.



Extended Data Fig. 10: Bromodomain dependence of vCD activator durability

a, Independently repeated CD45 experiment **(Fig. 4j)** testing selected modulators from 3-18 d.p.t. in the presence of selective epigenetic inhibitor drugs starting from 5 d.p.t. Observed activation levels in DMSO vehicle-treated control conditions for each modulator (mean±SEM, significance from sgNT, ****P<0.0001, *P<0.05, one-way ANOVA). **b,** Observed CD45 activation levels in DMSO vehicle-treated control conditions (solid lines) for VPR (gray) and vCD_p3 (blue) (mean±SEM), or in the presence of drugs (dashed lines) BET bromodomain inhibitor JQ1 or decitabine. **c,** Heatmap summary of normalized drug effects on CD45 activation at 9 d.p.t. of indicated vCD and benchmark modulators in the presence of selective epigenetic inhibitor drugs. Observed mean activation levels in DMSO vehicle-treated control served as normalization denominators to quantify drug effects on CD45-APC fluorescence intensity. **d,** Normalized drug effects on CD45 activation at 9 d.p.t. (mean±SEM) of indicated vCD and benchmark modulators in the presence of selective epigenetic inhibitor drugs. **e,** CXCR4 activation by vCD_p6, vCD-VP16 fusion, and controls at 3-d.p.t. (mean±SEM) in the presence of BET bromodomain inhibitors JQ1 or GNE049 (blue) or DMSO-treated control (empty).



Extended Data Fig. 11: Modulator-dependent effects of epigenetic inhibitors

a, Normalized drug effects (mean±SEM) on CD45 activation from 6 to 10 d.p.t. of indicated vCD and benchmark modulators and negative controls, with selective epigenetic inhibitor drugs applied at 5 d.p.t. and re-dosed in fresh media each day. Observed mean activation levels in DMSO vehicle-treated control served as normalization denominators to quantify drug effects on CD45-APC fluorescence intensity (y-axis). Small molecule inhibitor drugs were chosen to selectively target BET/BRD4i/CBP/P300-associated bromodomains (JQ1, GNE049, SGC-CBP30), MLL/WDR5 and EZH2 histone methyltransferases (OICR-9229, tazemetostat), histone deacetylases (TSA, TMP269, RG2833), and DNA methyltransferase (5-aza-cytidine analog decitabine).

Supplementary Information

Supplementary Table 1: Full library screen tiles, data, and biochemical features
Supplementary Table 2: Validation sub-library screen tiles, data, and biochemical features
Supplementary Table 3: Sequences and individual recruitment data for viral and archaeal tiles in HEK293T cells
Supplementary Table 4: Sequences and individual recruitment data for engineered vCD modulators in HEK293T cells

Methods

Cell lines

All experiments were carried out in K562 cells (ATCC; CCL-243) or HEK293T cells (ATCC; CRL-3216). Cells were cultured in a humidified incubator at 37°C and 5% CO_2, in either RPMI 1640 (Gibco, 61870036) media (K562 cells) or DMEM (Gibco, 10569010) (HEK293T cells), supplemented with 10% FBS (Takara 632180). HEK293T-LentiX cells (Takara 632180) were used to produce lentivirus. The custom GFP reporter construct was generated by cloning 7 copies of a synthetic guide RNA recognition sequence (5'-TTTA GTTGTTCTAAACGCTCTGAG CGG-3'), with CasMINI and Cas9 PAM sequences, upstream of a minimal CMV promoter (miniCMV) or constitutive human EF1α promoter followed by a Puromycin resistance cassette and EGFP with an intervening P2A self-cleaving peptide (Puro-P2A-EGFP) in a lentiviral transfer plasmid. These constructs were packaged into lentivirus and used to transduce K562 cells. After recovery, transduced activation reporter cells were enriched by puromycin selection (1μg mL^{-1}) following nucleofection of dCas9-VPR mRNA (TriLink) and a Cas9 sgRNA targeting the ESR protospacer (IDT). Suppression reporter cells were enriched by puromycin selection directly. Single cells of each reporter cell line were isolated by serial dilution in 96-well plates. After expansion, individual clones were validated by co-transduction of dCas9-VPR or dCas9-KRAB and ESR sgRNA lentivirus and analyzed by flow cytometry to determine the dynamic range of GFP expression for each clone following activation or suppression by dCas9-VPR or -KRAB. Top performing clones were selected for further expansion and used in downstream experiments.

Tiled library design and cloning

Human nuclear factors were determined by the searching Human Protein Atlas (.org) for "subcell_location:Nucleoplasm,Nuclear speckles,Nuclear bodies AND subcell_location:Nuclear membrane,Nucleoli,Nucleoli fibrillar center NOT subcell_location:Actin filaments,Aggresome,Cell Junctions,Centriolar satellite,Centrosome,Cleavage furrow,Cytokinetic bridge,Cytoplasmic bodies,Cytosol,Endoplasmic reticulum,Endosomes,Focal adhesion sites,Golgi apparatus,Intermediate filaments,Lipid droplets,Lysosomes,Microtubule ends,Microtubules,Midbody,Midbody ring,Mitochondria,Mitotic spindle,Peroxisomes,Plasma membrane,Rods". Sequences encoding human viral transcriptional regulators (hvTRs) were obtained from published sources[26]. Sequences encoding viruses of the families *Adenoviridae, Arenaviridae, Bornaviridae, Coronaviridae, Filoviridae, Flaviviridae, Hepadnaviridae, Herpesviridae, Orthomyxoviridae, Papillomaviridae, Paramyxoviridae, Parvoviridae, Peribunyaviridae, Phenuiviridae, Pneumoviridae, Polyomaviridae, Poxviridae, Retroviridae,* and *Rhabdoviridae,* sequences encoding viruses with known zoonotic transmission, *Flaviviridae, Lyssaviridae, Filoviridae, Paramyxoviridae, Orthomyxoviridae, Coronaviridae, Reoviridae, Togaviridae, Phenuviridae, Hantaviridae*, *Adenoviridae, and Poxviridae,* metagenomic viruses of families *Siphoviridae, podoviridae, and myoviridae,* and archaeal genome *Acidianus infernus* were manually obtained from dBatVir and NCBI. Screened peptide tiles were encoded by DNA oligos (Twist) 300 nucleotides in length, of which 255 nucleotides were target specific. The 5' and 3' ends of each oligo consisted of sequences complementary to the destination vector, with the 3' 15 nucleotide overlap being composed of part of the Illumina Read 1 Primer, where [vector overlap 1-target sequence-stop-barcode- vector overlap 2 (illumina Read 1 partial)]. This design allowed for convenient cloning of the library using either NEB HiFi or In-Fusion cloning approaches, and for the convenient downstream generation of Illumina-compatible NGS libraries. For the validation sub-library: 954 predicted activators (FDR<0.25; $log_2FC>1$), 1,228 predicted suppressors (FDR<0.001; $log_2FC<-1$), and 22 predicted dual-activity modulators (enriched in both activation and suppression lists) were tested alongside literature-based positive control tiles, namely published activator tiles from yeast and human transcription factors[10], plus activator and suppressor tiles from Pfam-annotated human proteins[18]. As negative controls, we included a set of 949 peptides predicted to be inactive: 314 tiles depleted from ON bins in initial mCMV and EF1α screens, 563 scrambled sequence tiles with opposing activities in initial mCMV and EF1α screens, and 72 tiles with early stop codons and opposing activities in initial mCMV and EF1α screens.

Pooled library screening

The custom DNA oligo library (each 300 bp, Twist) of 43,938 putative modulator elements (original screen) and ~3,750 elements (sub-library validation screen) paired to unique 12mer DNA barcodes was cloned into dCas9 lentiviral expression plasmid at high coverage (1,000x), packaged into lentivirus, and transduced into the mCMV and EF1α GFP reporter cell lines at MOI=0.3. Cells were treated with blasticidin (10μg mL^{-1}) to enrich for positively transduced cells, followed by fluorescence-activated cell sorting (BD FACSAria) to separate populations of interest: mCMV-GFP-ON cells for activation and EF1α-GFP-OFF cells for suppression, at 6- and 10-days post-transduction, respectively. Sorted populations of interest were further enriched by culturing for 6 additional days and subjected to 4-way gated FACS separation into discrete bins based on GFP fluorescence intensity. Genomic DNA was extracted from sorted cells in each discrete bin, and from bulk mCMV-GFP-ON and EF1α-GFP-OFF cells. Barcoded modulator sequences were PCR amplified from these gDNA samples with primers containing Illumina adapter sequences and a unique i7 index for each sample. Pooled libraries were sequenced on an Illumina NextSeq 550 (Gladstone Genomics Core) to identify barcodes present in each sample.

Pooled library screen analysis

Read count matrices for each library were generated based on alignment of sequenced modulator barcodes using a custom Python script. All subsequent data analysis was performed using R version 4.1.0. For the activation screen, technical replicates for GFP-OFF libraries were collapsed (counts per barcode were summed) resulting in two GFP-OFF replicates. For the GFP-ON conditions, we used one GFP-ON library collected at 6 days post-transduction and another that we built *in silico* by taking the weighted sum of binned GFP-ON gates P7, P8, P9, and P10 (collected after a further 6 days of enrichment). DESEQ2 (version 1.32.0) was used to identify statistically significant activator sequences that were enriched in the two GFP-ON

libraries compared to the two GFP-OFF libraries (FDR<0.05, $log_2FC>0$). For the suppression screen, technical replicates for GFP-ON libraries were collapsed (counts per barcode were summed) resulting in two GFP-ON replicates. For the GFP-OFF conditions, we used one GFP-OFF library collected at 10 days post-transduction and another that we built *in silico* by taking the weighted sum of GFP-OFF gates P6, P7, P8, and P9 (collected after a further 6 days of enrichment). DESEQ2 (version 1.32.0) was used to identify statistically significant suppressor sequences that were enriched in the two GFP-OFF libraries compared to the two GFP-ON libraries (FDR<0.001, $log_2FC<0$). The following packages were used for downstream annotation and analysis: parSeqSim for all-by-all sequence homology to identify clusters of tiles; DECIPHER for multiple sequence alignment to identify conserved domains; DagLogo for amino acid-level enrichment of biochemical properties in tiles; MOTIF Search for enrichment of amino acid motifs and domains.

Generalized linear regression for activator strength prediction

To identify biochemical features that were predictive of activation sequences, we trained generalized linear regression models based on the proportion of amino acids in the 85 amino acid peptide tiles (OneHot encoding) using “caret” (v6.0) and “glmnet” (v4.1) in R (v4.1.0). The top model was selected with 10-fold cross-validation and feature importance was extracted and plotted for visualization. We additionally trained generalized linear regression models based on sets of biochemical features ("biochem_tiny","biochem_small","biochem_aliphatic","biochem_aromatic","biochem_nonpolar","biochem_polar","biochem_charged","biochem_basic","biochem_acidic") generated using the “Peptides” package (v2.4.4).

Statistical metrics for activation time series analysis

Normalized potency, robustness, and durability scores are defined as follows. We first calculated the average of the mean fold change per modulator, relative to baselines observed in non-targeting sgRNA (sgNT) and dCasMINI recruited without modulator fusion (dCasMINI) conditions, across replicates for the gene of interest (CXCR4, CD45, IFNG, CD81) at days 3 and 6 to define early activation potency. These averages were then plotted as mean_fold_change vs. time (days). For simplicity, we assume that the behavior of the change between time points was linear. The area under the curve (AUC) for this plot was then calculated numerically, and the subsequent errors were propagated. The AUC was then normalized to the maximum AUC for that particular gene of interest. For example, if a modulator had the highest AUC measured for CXCR4, then the normalized CXCR4 potency score for this modulator will have a value of 1. Normalized robustness score is calculated similarly, factoring modulator activity across multiple targets. Normalized durability score is calculated similarly, but the area under the curve is calculated at post-plasmid time points, i.e. all time points after day 9 post-transfection.

Protein structure predictions

Protein structures were predicted using ESMFold[53] within the ColabFold suite[64]. The YASARA software package[54] was used to calculate structural features such as solvent accessible surface area and electrostatic potential. The electrostatic potential was calculated using the AMBER14 forcefield as implemented in YASARA[63]. Structural alignments were calculated using the SHEBA algorithm further checked with MUSTANG[65]. For presentation of the modulator sequence, the modulator was spaced from the Cas protein within YASARA. The modulator linker was then used to link the modulator region to the Cas protein using YASARA’s BuildLoop protocol.

Arrayed testing of individual modulators

Oligos encoding putative human, viral, archaeal, and engineered modulator domains, and target-specific duplexed sgRNA spacer sequences, were synthesized as eBlocks (IDT) and cloned as direct fusions into dCasMINI or dCas9 vector plasmids, or sgRNA vector plasmids, respectively. For transient plasmid delivery experiments, wild-type HEK293T cells were seeded in 96-well plates at a density of 20,000 cells per well. The same day, cells were co-transfected (Mirus X2) with uniform masses of dCasMINI-modulator- and sgRNA-expressing plasmids such that each well received a single modulator construct to be tested, but the same targeting sgRNA across all wells. Each well received 100ng of modulator plasmid and 33.3ng sgRNA plasmid and experiments were performed in technical triplicate at a minimum.

Flow cytometry, ELISA, RT-qPCR

Cells were immunostained for surface protein detection with APC direct-conjugated primary antibodies against CD45, CXCR4, CD81, CD2, CD71 (BioLegend; 1:100 dilution) or for GFP reporter expression. Cells were analyzed by flow cytometry (Cytoflex LX) with analysis gates (FlowJo) to ensure measurements of live, singlet, and double-transfected cells to verify both dCas-modulator and sgRNA plasmid expression via mCherry and BFP fluorescence, respectively. Geometric mean of APC fluorescence or GFP, or percent CD45-APC+ frequency of parent for each condition were normalized against those of negative controls and reported as fold-change relative to negative controls. For IFNG assay, cell supernatants were collected to monitor IFNG protein expression by ELISA according to manufacturer protocols (BioLegend). For cell cycle assays, cells were pulsed with 1μM 5-ethynyl-2'-deoxyuridine (EdU) for 1 hour to label S-phase cells, followed by PBS wash and replacement of fresh DMEM/FBS media. EdU detection was by Click chemistry-based AlexaFluor-488 flow cytometry assay (ThermoFisher). For mRNA quantifications, cell pellets were flash-frozen and stored in -80°C before processing for cell lysis and analysis by Cells-to-CT 1-Step TaqMan Kit with Taqman probes (ThermoFisher).

Data Availability

The illumina sequencing datasets generated in this study will be made available in the NCBI Sequencing Read Archive.

Code Availability

All custom codes used for data analysis are available upon request.

Acknowledgements

We thank Christopher Still, T. Danny Ko, Ian Lam, Kavita Jadhav, Aayushma Gautam, Mohamed Ghazal, Tabitha Tcheau, and Xiaoshu Xu for helpful conversations and assistance, and Mylinh Bernardi of the Gladstone Genomics core for assistance with illumina sequencing. L.S.Q. is supported by Sarafan ChEM-H, Stanford University and is a Chan-Zuckerberg BioHub-San Francisco Investigator.

Author Contributions

D.O.H., L.S.Q., and T.P.D. designed the study. G.A.C., R.W.Y., D.O.H., and T.P.D. designed screening libraries. G.A.C., R.W.Y., and V.C. designed engineered activators. T.B.G. designed reporter cells and library cloning strategies. G.A.C., T.B.G., and V.C. performed experiments. R.W.Y., G.A.C., M.Z.J., and X.Y. analyzed data. G.A.C. and R.W.Y. wrote the manuscript with significant contributions from all authors. D.O.H. supervised the study.

Competing Interests

L.S.Q. is founder and shareholder of Epicrispr Biotechnologies. G.A.C., R.W.Y., T.B.G., M.Z.J., X.Y., V.C., L.S.Q., T.P.D., and D.O.H. are inventors on provisional patents relating to this work, are employees of and acknowledge outside interest in Epicrispr Biotechnologies.

References

1. Qi, L.S., Larson, M.H., Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P. and Lim, W.A., 2013. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, *152*(5), pp.1173-1183.
2. Qian, J., Guan, X., Xie, B., Xu, C., Niu, J., Tang, X., Li, C.H., Colecraft, H.M., Jaenisch, R. and Liu, X.S., 2023. Multiplex epigenome editing of MECP2 to rescue Rett syndrome neurons. *Science Translational Medicine*, *15*(679), p.eadd4666.
3. Matharu, N., Rattanasopha, S., Tamura, S., Maliskova, L., Wang, Y., Bernard, A., Hardin, A., Eckalbar, W.L., Vaisse, C. and Ahituv, N., 2019. CRISPR-mediated activation of a promoter or enhancer rescues obesity caused by haploinsufficiency. Science, 363(6424), p.eaau0629.
4. Doudna, J.A., 2020. The promise and challenge of therapeutic genome editing. *Nature*, *578*(7794), pp.229-236.
5. Jensen, T.I., Mikkelsen, N.S., Gao, Z., Foßelteder, J., Pabst, G., Axelgaard, E., Laustsen, A., König, S., Reinisch, A. and Bak, R.O., 2021. Targeted regulation of transcription in primary cells using CRISPRa and CRISPRi. *Genome Research*, *31*(11), pp.2120-2130.
6. Sadowski, I., Ma, J., Triezenberg, S. and Ptashne, M., 1988. GAL4-VP16 is an unusually potent transcriptional activator. *Nature*, *335*(6190), pp.563-564.
7. Beerli, R.R., Segal, D.J., Dreier, B. and Barbas III, C.F., 1998. Toward controlling gene expression at will: specific regulation of the erbB-2/HER-2 promoter by using polydactyl zinc finger proteins constructed from modular building blocks. *Proceedings of the National Academy of Sciences*, *95*(25), pp.14628-14633.
8. Chavez, A., Scheiman, J., Vora, S., Pruitt, B.W., Tuttle, M., PR Iyer, E., Lin, S., Kiani, S., Guzman, C.D., Wiegand, D.J. and Ter-Ovanesyan, D., 2015. Highly efficient Cas9-mediated transcriptional programming. *Nature methods*, *12*(4), pp.326-328.
9. Kotha, S.R. and Staller, M.V., 2023. The balance of acidic and hydrophobic residues predicts acidic transcriptional activation domains from protein sequence. bioRxiv, pp.2023-02.
10. Sanborn, A.L., Yeh, B.T., Feigerle, J.T., Hao, C.V., Townshend, R.J., Lieberman Aiden, E., Dror, R.O. and Kornberg, R.D., 2021. Simple biochemical features underlie transcriptional activation domain diversity and dynamic, fuzzy binding to Mediator. *Elife*, *10*, p.e68068.
11. Staller, M.V., Ramirez, E., Kotha, S.R., Holehouse, A.S., Pappu, R.V. and Cohen, B.A., 2022. Directed mutational scanning reveals a balance between acidic and hydrophobic residues in strong human activation domains. Cell systems, 13(4), pp.334-345.
12. Hilton, I.B., D'ippolito, A.M., Vockley, C.M., Thakore, P.I., Crawford, G.E., Reddy, T.E. and Gersbach, C.A., 2015. Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nature biotechnology*, *33*(5), pp.510-517.
13. Liu, X.S., Wu, H., Krzisch, M., Wu, X., Graef, J., Muffat, J., Hnisz, D., Li, C.H., Yuan, B., Xu, C. and Li, Y., 2018. Rescue of fragile X syndrome neurons by DNA methylation editing of the FMR1 gene. Cell, 172(5), pp.979-992.
14. Cano-Rodriguez, D., Gjaltema, R.A.F., Jilderda, L.J., Jellema, P., Dokter-Fokkens, J., Ruiters, M.H.J. and Rots, M.G., 2016. Writing of H3K4Me3 overcomes epigenetic silencing in a sustained but context-dependent manner. *Nature communications*, *7*(1), p.12284.
15. Wu, Q., Wu, J., Karim, K., Chen, X., Wang, T., Iwama, S., Carobbio, S., Keen, P., Vidal-Puig, A., Kotter, M.R. and Bassett, A., 2023. Massively parallel characterization of CRISPR activator efficacy in human induced pluripotent stem cells and neurons. *Molecular Cell*, *83*(7), pp.1125-1139.
16. Wang, K., Escobar, M., Li, J., Mahata, B., Goell, J., Shah, S., Cluck, M. and Hilton, I.B., 2022. Systematic comparison of CRISPR-based transcriptional activators uncovers gene-regulatory features of enhancer–promoter interactions. *Nucleic Acids Research*, *50*(14), pp.7842-7855.
17. Alerasool, N., Leng, H., Lin, Z.Y., Gingras, A.C. and Taipale, M., 2022. Identification and functional characterization of transcriptional activators in human cells. *Molecular cell*, *82*(3), pp.677-695.
18. Tycko, J., DelRosso, N., Hess, G.T., Banerjee, A., Mukund, A., Van, M.V., Ego, B.K., Yao, D., Spees, K., Suzuki, P. and Marinov, G.K., 2020. High-throughput discovery and characterization of human transcriptional effectors. *Cell*, *183*(7).
19. Klann, T.S., Black, J.B., Chellappan, M., Safi, A., Song, L., Hilton, I.B., Crawford, G.E., Reddy, T.E. and Gersbach, C.A., 2017. CRISPR–Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nature biotechnology*, *35*(6), pp.561-568.
20. Mahata, B., Li, J., Cabrera, A., Brenner, D.A., Guerra-Resendez, R.S., Goell, J., Wang, K., Escobar, M., Guo, Y., Parthasarathy, A.K. and Hilton, I.B., 2022. Compact engineered human transactivation modules enable potent and versatile synthetic transcriptional control. *bioRxiv*, pp.2022-03.
21. Omachi, K. and Miner, J.H., 2022. Comparative analysis of dCas9-VP64 variants and multiplexed guide RNAs mediating CRISPR activation. *Plos one*, *17*(6), p.e0270008.
22. Lebar, T., Lainšček, D., Merljak, E., Aupič, J. and Jerala, R., 2020. A tunable orthogonal coiled-coil interaction toolbox for engineering mammalian cells. *Nature chemical biology*, *16*(5), pp.513-519.
23. Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R. and Weirauch, M.T., 2018. The human transcription factors. *Cell*, *172*(4), pp.650-665.
24. Anthony, S.J., Epstein, J.H., Murray, K.A., Navarrete-Macias, I., Zambrana-Torrelio, C.M., Solovyov, A., Ojeda-Flores, R., Arrigo, N.C., Islam, A., Ali Khan, S. and Hosseini, P., 2013. A strategy to estimate unknown viral diversity in mammals. *MBio*, *4*(5), pp.e00598-13.

25. Carlson, C.J., Zipfel, C.M., Garnier, R. and Bansal, S., 2019. Global estimates of mammalian viral diversity accounting for host sharing. *Nature ecology & evolution*, *3*(7), pp.1070-1075.
26. Liu, X., Hong, T., Parameswaran, S., Ernst, K., Marazzi, I., Weirauch, M.T. and Bass, J.I.F., 2020. Human virus transcriptional regulators. *Cell*, *182*(1), pp.24-37.
27. DelRosso, N., Tycko, J., Suzuki, P., Andrews, C., Mukund, A., Liongson, I., Ludwig, C., Spees, K., Fordyce, P., Bassik, M.C. and Bintu, L., 2023. Large-scale mapping and mutagenesis of human transcriptional effector domains. *Nature*, pp.1-8.
28. Ludwig, C.H., Thurm, A.R., Morgens, D.W., Yang, K.J., Tycko, J., Bassik, M.C., Glaunsinger, B.A. and Bintu, L., 2022. High-Throughput Discovery and Characterization of Viral Transcriptional Effectors in Human Cells. *bioRxiv*, pp.2022-12.
29. Straub, C.T., Counts, J.A., Nguyen, D.M., Wu, C.H., Zeldes, B.M., Crosby, J.R., Conway, J.M., Otten, J.K., Lipscomb, G.L., Schut, G.J. and Adams, M.W., 2018. Biotechnology of extremely thermophilic archaea. *FEMS Microbiology Reviews*, *42*(5), pp.543-578.
30. Reinberg, D. and Vales, L.D., 2018. Chromatin domains rich in inheritance. Science, 361(6397), pp.33-34.
31. Bird, A., 2002. DNA methylation patterns and epigenetic memory. Genes & development, 16(1), pp.6-21.
32. Margueron, R. and Reinberg, D., 2011. The Polycomb complex PRC2 and its mark in life. *Nature*, *469*(7330), pp.343-349.
33. Smith, Z.D. and Meissner, A., 2013. DNA methylation: roles in mammalian development. Nature Reviews Genetics, 14(3), pp.204-220.
34. Trojer, P. and Reinberg, D., 2007. Facultative heterochromatin: is there a distinctive molecular signature?. Molecular cell, 28(1), pp.1-13.
35. Sump, B., Brickner, D.G., D'urso, A., Kim, S.H. and Brickner, J.H., 2022. Mitotically heritable, RNA polymerase II-independent H3K4 dimethylation stimulates INO1 transcriptional memory. Elife, 11, p.e77646.
36. Harvey, Z.H., Chakravarty, A.K., Futia, R.A. and Jarosz, D.F., 2020. A prion epigenetic switch establishes an active chromatin state. Cell, 180(5), pp.928-940.
37. Nuñez, J.K., Chen, J., Pommier, G.C., Cogan, J.Z., Replogle, J.M., Adriaens, C., Ramadoss, G.N., Shi, Q., Hung, K.L., Samelson, A.J. and Pogson, A.N., 2021. Genome-wide programmable transcriptional memory by CRISPR-based epigenome editing. *Cell*, *184*(9), pp.2503-2519.
38. O'Geen, H., Bates, S.L., Carter, S.S., Nisson, K.A., Halmai, J., Fink, K.D., Rhie, S.K., Farnham, P.J. and Segal, D.J., 2019. Ezh2-dCas9 and KRAB-dCas9 enable engineering of epigenetic memory in a context-dependent manner. *Epigenetics & chromatin*, *12*(1), pp.1-20.
39. Beyersdorf, J.P., Bawage, S., Iglesias, N., Peck, H.E., Hobbs, R.A., Wroe, J.A., Zurla, C., Gersbach, C.A. and Santangelo, P.J., 2022. Robust, Durable Gene Activation In Vivo via mRNA-Encoded Activators. *ACS nano*, *16*(4), pp.5660-5671.
40. Ruthenburg, A.J., Allis, C.D. and Wysocka, J., 2007. Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark. Molecular cell, 25(1), pp.15-30.
41. Cirillo, L.A., Lin, F.R., Cuesta, I., Friedman, D., Jarnik, M. and Zaret, K.S., 2002. Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Molecular cell*, *9*(2), pp.279-289.
42. Iwafuchi-Doi, M. and Zaret, K.S., 2014. Pioneer transcription factors in cell reprogramming. *Genes & development*, *28*(24), pp.2679-2692.
43. El-Osta, A., Brasacchio, D., Yao, D., Pocai, A., Jones, P.L., Roeder, R.G., Cooper, M.E. and Brownlee, M., 2008. Transient high glucose causes persistent epigenetic changes and altered gene expression during subsequent normoglycemia. *The Journal of experimental medicine*, *205*(10), pp.2409-2417.
44. Pacis, A., Mailhot-Léonard, F., Tailleux, L., Randolph, H.E., Yotova, V., Dumaine, A., Grenier, J.C. and Barreiro, L.B., 2019. Gene activation precedes DNA demethylation in response to infection in human dendritic cells. *Proceedings of the National Academy of Sciences*, *116*(14), pp.6938-6943.
45. Lio, C.W.J. and Rao, A., 2019. TET enzymes and 5hmC in adaptive and innate immune systems. *Frontiers in Immunology*, *10*, p.210.
46. Dey, A., Nishiyama, A., Karpova, T., McNally, J. and Ozato, K., 2009. Brd4 marks select genes on mitotic chromatin and directs postmitotic transcription. *Molecular biology of the cell*, *20*(23), pp.4899-4909.
47. Filippakopoulos, P., Qi, J., Picaud, S., Shen, Y., Smith, W.B., Fedorov, O., Morse, E.M., Keates, T., Hickman, T.T., Felletar, I. and Philpott, M., 2010. Selective inhibition of BET bromodomains. *Nature*, *468*(7327), pp.1067-1073.
48. Brierley, L., Vonhof, M.J., Olival, K.J., Daszak, P. and Jones, K.E., 2016. Quantifying global drivers of zoonotic bat viruses: a process-based perspective. *The American Naturalist*, *187*(2), pp.E53-E64.
49. Chen, L., Liu, B., Yang, J. and Jin, Q., 2014. DBatVir: the database of bat-associated viruses. *Database*, *2014*.
50. Munson-McGee, J.H., Snyder, J.C. and Young, M.J., 2018. Archaeal viruses from high-temperature environments. *Genes*, *9*(3), p.128.
51. Dávila-Ramos, S., Castelán-Sánchez, H.G., Martínez-Ávila, L., Sánchez-Carbente, M.D.R., Peralta, R., Hernández-Mendoza, A., Dobson, A.D., Gonzalez, R.A., Pastor, N. and Batista-García, R.A., 2019. A review on viral metagenomics in extreme environments. *Frontiers in microbiology*, *10*, p.2403.
52. Segerer, A., Neuner, A., Kristjansson, J.K. and Stetter, K.O., 1986. Acidianus infernus gen. nov., sp. nov., and Acidianus brierleyi comb. nov.: facultatively aerobic, extremely acidophilic thermophilic sulfur-metabolizing archaebacteria. *International Journal of Systematic Bacteriology*, *36*(4), pp.559-564.
53. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., ... & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637), 1123-1130.
54. Ozvoldik, K., Stockner, T., Rammner, B. and Krieger, E., 2021. Assembly of biomolecular gigastructures and visualization with the vulkan graphics API. *Journal of Chemical Information and Modeling*, *61*(10), pp.5293-5303.
55. Xu, X., Chemparathy, A., Zeng, L., Kempton, H.R., Shang, S., Nakamura, M. and Qi, L.S., 2021. Engineered miniature CRISPR-Cas system for mammalian genome regulation and editing. *Molecular Cell*, *81*(20), pp.4333-4345.
56. Burysek, L., Yeow, W.S. and Pitha, P.M., 1999. Unique properties of a second human herpesvirus 8-encoded interferon regulatory factor (vIRF-2). *Journal of human virology*, *2*(1), pp.19-32.
57. Erijman, A., Kozlowski, L., Sohrabi-Jahromi, S., Fishburn, J., Warfield, L., Schreiber, J., Noble, W.S., Söding, J. and Hahn, S., 2020. A high-throughput screen for transcription activation domains reveals their sequence features and permits prediction by deep learning. *Molecular cell*, *78*(5), pp.890-902.
58. Seipel, K., Georgiev, O. and Schaffner, W., 1994. A minimal transcription activation domain consisting of a specific array of aspartic acid and leucine residues. *Biological Chemistry Hoppe-Seyler*, *375*(7), pp.463-470.
59. Davis, Z.H., Verschueren, E., Jang, G.M., Kleffman, K., Johnson, J.R., Park, J., Von Dollen, J., Maher, M.C., Johnson, T., Newton, W. and Jäger, S., 2015. Global mapping of herpesvirus-host protein complexes reveals a transcription strategy for late genes. *Molecular cell*, *57*(2), pp.349-360.
60. Raisner, R., Kharbanda, S., Jin, L., Jeng, E., Chan, E., Merchant, M., Haverty, P.M., Bainer, R., Cheung, T., Arnott, D. and Flynn, E.M., 2018. Enhancer activity requires CBP/P300 bromodomain-dependent histone H3K27 acetylation. *Cell reports*, *24*(7), pp.1722-1729.

61. Gillespie, M.A., Palii, C.G., Sanchez-Taltavull, D., Shannon, P., Longabaugh, W.J., Downes, D.J., Sivaraman, K., Espinoza, H.M., Hughes, J.R., Price, N.D. and Perkins, T.J., 2020. Absolute quantification of transcription factors reveals principles of gene regulation in erythropoiesis. *Molecular cell*, *78*(5), pp.960-974.
62. Jawaid, M.Z., Yeo, R.W., Gautam, A., Gainous, T.B., Hart, D.O. and Daley, T.P., 2023. Improving few-shot learning-based protein engineering with evolutionary sampling. *bioRxiv*, pp.2023-05.
63. Maier, J.A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K.E. and Simmerling, C., 2015. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of chemical theory and computation*, *11*(8), pp.3696-3713.
64. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S. and Steinegger, M., 2022. ColabFold: making protein folding accessible to all. *Nature methods*, *19*(6), pp.679-682.
65. Konagurthu, A.S., Whisstock, J.C., Stuckey, P.J. and Lesk, A.M., 2006. MUSTANG: a multiple structural alignment algorithm. *Proteins: Structure, Function, and Bioinformatics*, *64*(3), pp.559-574.