Multiome Perturb-seq unlocks scalable discovery of integrated perturbation effects on the transcriptome and epigenome

Eli Metzner[1,2,*], Kaden M. Southard[1,*], and and Thomas M. Norman[1, †]

[1]Computational and Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[2]Tri-Institutional Training Program in Computational Biology and Medicine, New York, NY, USA
[*]These authors contributed equally to this work
[†]Lead contact: `normantm@mskcc.org`

SUMMARY

Single-cell CRISPR screens link genetic perturbations to transcriptional states, but high-throughput methods connecting these induced changes to their regulatory foundations are limited. Here we introduce Multiome Perturb-seq, extending single-cell CRISPR screens to simultaneously measure perturbation-induced changes in gene expression and chromatin accessibility. We apply Multiome Perturb-seq in a CRISPRi screen of 13 chromatin remodelers in human RPE-1 cells, achieving efficient assignment of sgRNA identities to single nuclei via an improved method for capturing barcode transcripts from nuclear RNA. We organize expression and accessibility measurements into coherent programs describing the integrated effects of perturbations on cell state, finding that *ARID1A* and *SUZ12* knockdowns induce programs enriched for developmental features. Pseudotime analysis of perturbations connects accessibility changes to changes in gene expression, highlighting the value of multimodal profiling. Overall, our method provides a scalable and simply implemented system to dissect the regulatory logic underpinning cell state.

KEYWORDS

Perturb-seq, multiome, single-cell genomics, functional genomics, CRISPRi, chromatin accessibility, gene regulatory networks

INTRODUCTION

Single-cell CRISPR screens have transformed the field of functional genomics by enabling simple, pooled experiments linking genetic perturbations to high-content phenotypes.[1,2] Perturb-seq, CROP-seq, and Mosaic-seq pioneered these approaches, expanding screen readouts from scalar measurements of cell fitness or reporter expression to comprehensive assessment of changes in transcriptional state.[3–6] As single-cell assays have advanced, techniques profiling multiple modalities have integrated transcriptional data with relevant covariates,[7–10] enabling new questions to be asked about the origins and regulation of cellular states. In particular, multiome assays that simultaneously measure gene expression and chromatin accessibility in the same cells have the potential to help explain how chromatin state influences and directs the transitions between transcriptional states.[11,12] During development, changes in accessibility have been shown to "foreshadow" gene expression changes and ultimately specify cell lineages.[11,13–15] Conversely, bulk co-profiling of expression and accessibility changes following chemical perturbations has suggested that certain pathways induce expression changes within open chromatin independent of remodeling, while others trigger global rewiring of chromatin state.[12,16–18] The unclear interplay between regulatory layers calls for a method to simultaneously characterize many perturbation-induced effects on the transcriptome and epigenome in parallel. These connections would have diverse applications because of their mechanistic interpretability, from elucidating regulators of cellular plasticity to gathering data to constrain causal modeling of gene regulatory networks.

To this end, we introduce Multiome Perturb-seq, an extension of the Perturb-seq platform that captures simultaneous scRNA-seq and scATAC-seq measurements of perturbation effects. Our protocol adapts the CROP-seq vector[5] to be compatible with commercially available 10x Single Cell Multiome ATAC + Gene Expression reagents ("10x Multiome"). The core challenge is robust sgRNA assignment from nuclear cDNA, as isolating nuclei is a requirement of scATAC-seq protocols. To overcome this, we modified the CROP-seq vector and developed a new amplification protocol to enhance recovery of CROP-seq sgRNA identity transcripts ("IDs"), increasing perturbation assignment to 68%. We apply this protocol to a library of 13 chromatin remodelers from three regulatory complexes, demonstrate how Multiome Perturb-seq can profile and organize expression and accessibility changes into coherent effects on cell state, and interpret the dynamic regulatory processes connecting these two layers of gene regulation.

RESULTS

An optimized sgRNA assignment workflow recovers perturbation identities of single nuclei

We chose 10x Multiome as the target platform due to its ubiquity, quality, and the commercial availability of reagents, but this introduces a challenge for sgRNA capture: the corresponding beads contain oligo-dT primers to capture mRNA, and ATAC primers to capture transposed genomic DNA, but critically lack primers for capture of sgRNAs. This limitation restricted our ID expression and assignment strategy to the polyadenylated transcript-dependent CROP-seq approach, which, in the case of nuclei isolation, is hindered by active nuclear export of polyadenylated mRNAs.[19] To mitigate this effect, we engineered a modified CROP-seq vector and library preparation protocol to increase capture and assignment of sgRNA ID transcripts **(Fig. 1a)**. We incorporated an internal high-activity T7 promoter[20] upstream of the sgRNA cassette (similarly to methods used to boost sgRNA signal in spatial contexts[21]), enabling efficient linear amplification of IDs from cDNA before index PCR **(Methods, Document S2)**. We also flanked the ID transcript with short sequence elements ("SIRLOINs") that increase nuclear localization of RNA,[22] added a partial 10x TSO binding sequence to

assist amplification of IDs in case of incomplete template switching during reverse transcription,[23] and removed the highly structured woodchuck hepatitis virus post-transcriptional regulatory element (WPRE), which we reasoned might impair efficiency of reverse transcription. We observed that SIRLOIN elements drove a modest increase in nuclear abundance of IDs coupled with a modest decrease in overall expression, consistent with their original characterization **(Fig. S1a)**, and did not alter knockdown efficiency of cell surface markers **(Fig. S1b)**. Finally, we included a cassette for inserting auxiliary barcodes (e.g. as lineage markers[24, 25]), which can be read out alongside IDs via altered sequencing parameters in our protocol if desired.
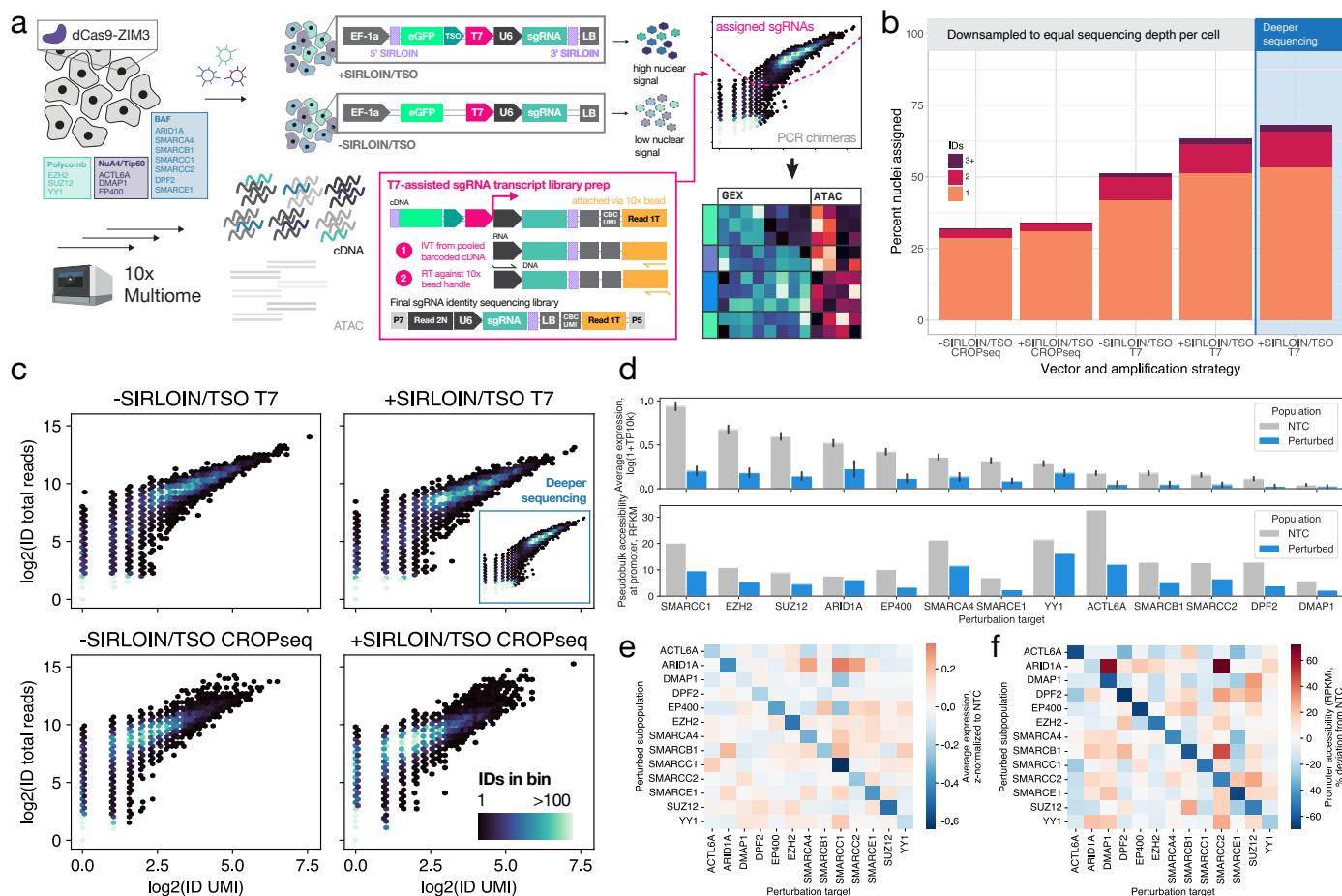

Fig. 1: Improved perturbation assignment to single nuclei with T7-assisted sgRNA library prep and an engineered CROP-seq vector. **a)** Schematic of Multiome Perturb-seq experimental design, library, and guide assignment strategy. sgRNA identity transcripts are biased toward nuclear retention by addition of dual SIRLOIN elements flanking the sgRNA cassette (purple), and a partial 10x TSO capture sequence (green) assists recovery. Transcripts are linearly amplified via an internal high-activity T7 promoter from barcoded cDNA. U6: modified U6 Pol III promoter, LB: lineage barcode cassette. **b)** Percentage of single nuclei assigned to 1, 2, or 3+ sgRNA identities using traditional CROP-seq or T7-assisted library prep, in a standard CROP-seq vector with T7 promoter (-SIRLOIN/TSO) or a CROP-seq vector with T7 promoter, dual SIRLOINs and partial TSO capture sequence (+SIRLOIN/TSO), at indicated sequencing depths (downsampled for comparison, deeper sequencing for downstream analysis). **c)** Density of UMI and total sequencing reads of sgRNA identity transcripts amplified by traditional CROP-seq or T7-assisted library prep in nuclei-containing droplets, in each vector condition, downsampled to equal sequencing depth per nucleus. **d)** Average single-cell expression and pseudobulk promoter accessibility of perturbation target genes in NTC-assigned populations vs. populations assigned to each perturbation. Error bars for single-cell expression show 95% confidence intervals. **e)** Average normalized expression of perturbation target genes in perturbed subpopulations. **f)** Pseudobulk normalized accessibility at promoters of perturbation target genes in perturbed subpopulations.

We applied Multiome Perturb-seq in a CRISPRi screen of RPE-1 cells stably expressing ZIM3-dCas9[26] using both the fully engineered Multiome Perturb-seq vector and a vector lacking SIRLOIN and TSO elements. Importantly, these hTERT-immortalized but untransformed cells allow us to investigate gene regulation in

2

a more physiological context than cancer cell lines. Our library included three non-targeting control (NTC) sgRNAs and 13 sgRNAs targeting chromatin remodelers belonging to three regulatory complexes[27–30] **(Fig. 1a, Supplementary Table 1)**. We isolated nuclei from RPE-1-ZIM3 populations expressing each version of the library, encapsulated them in two 10x Multiome lanes targeting 10,000 nuclei each, and recovered 9,318 (+SIRLOIN/TSO) and 9,504 (-SIRLOIN/TSO) nuclei after cellranger-arc cell calling **(Methods)**.

At equal sequencing depth of 850 reads per nucleus, traditional CROP-seq PCR-based library amplification yielded 31.8% assignment of single nuclei to at least one ID, improving to 33.8% in the +SIRLOIN/TSO condition **(Fig. 1b, Fig. S2)**. In contrast, our T7 amplification protocol yielded 51.2% assignment in the -SIRLOIN/TSO condition, increasing to 63.4% in the +SIRLOIN/TSO condition. Deeper sequencing of the T7 +SIRLOIN/TSO library to ∼4,000 reads per nucleus modestly increased assignment to 68.0%. We traced these efficiency differences to shifts in library composition. Linear amplification of IDs before PCR shifted final ID UMI and total read density toward the high-UMI/high-read mode representing assignable transcripts, with a more pronounced shift in the +SIRLOIN/TSO condition **(Fig. 1c, Methods)**. Overall, our ID assignment workflow represents a greater than 40% improvement over the current state-of-the-art single-cell ATAC-seq screening approach on the 10x platform (Spear-ATAC: 48% assignment),[31] while additionally providing gene expression phenotypes.

Analysis of the +SIRLOIN/TSO lane, after filtering for single ID assignments and performing basic quality control **(Fig. S3, Methods)**, yielded a final population of 4,724 nuclei with an average of 2,739 genes and 19,202 unique ATAC fragments detected. We observed robust on-target knockdown of expression and decreased accessibility at target gene promoters in perturbed populations **(Fig. 1d)**. Notably, we observed these effects across a range of native chromatin contexts, with both high-accessibility (e.g., *ACTL6A*) and low-accessibility (e.g., *DMAP1*) promoters displaying decreased accessibility upon perturbation. This result is consistent with the mechanism of the ZIM3 CRISPRi effector, whose KRAB domain recruits repressive chromatin-modifying complexes to induce heterochromatin formation.[26, 32] While the on-target effect of each perturbation was strongest as expected, we additionally observed modest up-regulation of *SMARCA4*, *SMARCC1*, and *SMARCC2* in the sgARID1A-containing population, suggesting crosstalk between these BAF complex members **(Fig. 1e)**. Increased expression of *SMARCC2* was concordant with increased accessibility of the *SMARCC2* promoter in sgARID1A-containing cells **(Fig. 1f)**. Overall, our optimized assignment workflow efficiently associates perturbations with single-cell transcriptomes and accessibility profiles.

Induced accessibility and expression changes provide complementary and largely non-overlapping descriptions of perturbation effects

We next analyzed the global phenotypes induced by perturbations. Identification of differentially expressed genes and differentially accessible peaks within each perturbation **(Supplementary Table 2, Methods)** revealed highly variable behaviors among individual perturbations: some mainly affected the chromatin landscape (*ARID1A*, *SMARCC2*), while others mostly modified gene expression (*DMAP1*, *EP400*), and some altered both equally (*SMARCE1*, *YY1*) **(Fig. 2a)**. In line with previous work,[16] we found that while examples of concordant changes in accessibility and expression at loci other than perturbation targets were present in 7 of 13 perturbations, these instances were relatively rare, and most differential expression and accessibility events across perturbations were independent **(Supplementary Table 3)**.

To compare how the distinct readouts organized perturbations, we performed hierarchical clustering on both expression and accessibility profiles **(Methods)**. Both modalities separated targets into known complexes, with tighter clustering mirroring the relative impact of different complexes on either the transcriptome or chromatin state: NuA4 members clustered distinctly in transcriptome space, while BAF complex members

were clearly delineated in accessibility space **(Fig. 2b, c)**. These differences underscore that expression and accessibility profiles offer complementary information, enabling nuanced characterization of perturbation effects across related targets.
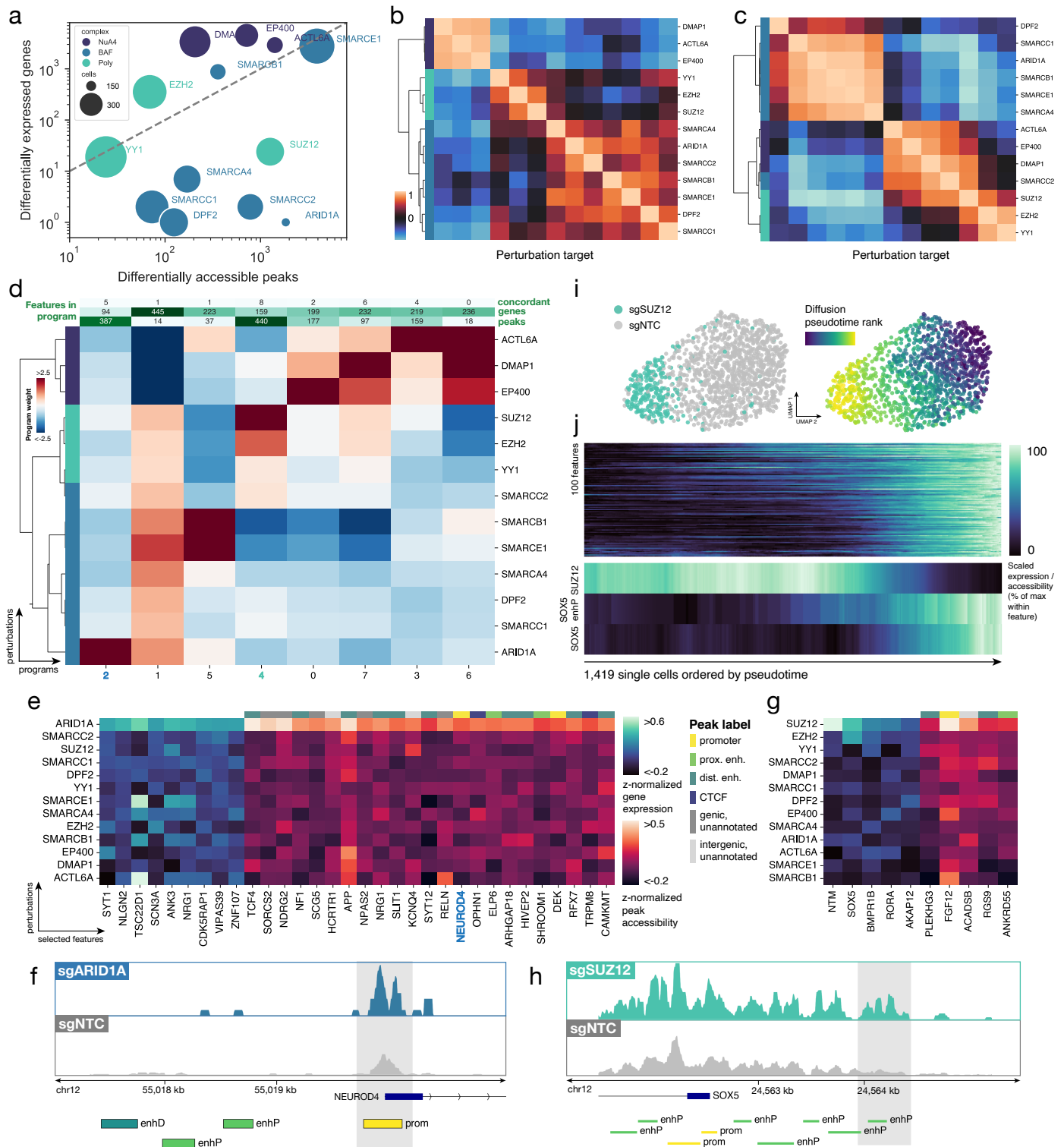

Fig. 2: Multiome Perturb-seq characterizes transcriptional and epigenetic state to elucidate regulatory relationships. **a)** Number of differentially expressed genes (q < 0.1, Kolmogorov-Smirnov test on normalized expression, Benjamini-Hochberg FDR correction) vs. number of differentially accessible peaks (q < 0.1, Mann-Whitney test on paired insertion counts within peaks, BH FDR correction) for each perturbation target. Point sizes represent number of cells assigned each sgRNA identity. **b)** Hierarchical clustering of average PCA embeddings of normalized expression in perturbed subpopulations. **c)** Hierarchical clustering of average spectral embeddings of ATAC tile matrix in perturbed subpopulations. (legend continued on next page)

4

Fig. 2: (cont.) **d)** Program weights on perturbed subpopulations and number of features with nonzero loadings per program from sparse PCA embedding of 7,885 genes and 5,060 peaks differentially expressed or accessible in at least one perturbation condition. **e)** Normalized expression and accessibility of select genes related to nervous system development in Program 2; peaks labeled with nearest gene. **f)** Pseudobulk ATAC-seq track at the *NEUROD4* promoter in sgARID1A-containing and sgNTC-containing cells, with ENCODE cCRE annotations (prom, promoter; enhP, proximal enhancer; enhD, distal enhancer). Differential peak selected as feature in Program 2 highlighted in grey. **g)** Normalized expression and accessibility of *SOX5* and *SOX5* target genes in Program 4. **h)** Pseudobulk ATAC-seq track at the *SOX5* promoter for sgSUZ12-containing and sgNTC-containing populations, with ENCODE cCRE annotations (prom, promoter; enhP, proximal enhancer; enhD, distal enhancer). Differential peak in proximal enhancer highlighted in grey. **i)** UMAP representations of 1,419 sgSUZ12-containing and sgNTC-containing single cells, colored by sgRNA identity and diffusion pseudotime rank. **j)** Top) Expression and accessibility of top 100 Program 4 features by weight in the population shown in Fig. 2i. (Bottom) Expression of *SUZ12*, accessibility of the *SOX5* proximal enhancer shown in Fig. 2h, and expression of *SOX5* in the same population. Single cells are ordered along the x-axis by their diffusion pseudotime values, with each feature separately scaled to a range of 0-1.

Co-embedding genes and peaks into sparse programs organizes effects on cell state

To organize the effects of perturbations into coherent programs, we concatenated filtered expression and accessibility matrices and co-embedded genes and peaks via non-negative sparse PCA **(Methods)**. This approach, applied to 7,885 genes and 5,060 peaks differentially expressed or accessible in at least one perturbation (and for peaks, limited to those within 50 kb of a gene), yielded 8 sparse programs, each containing hundreds of individual features **(Fig. 2d, Supplementary Table 4)**. Notably, programs varied in composition: some programs (2, 4) were largely peak-driven, others (1, 5, 6) gene-driven, and others (0, 3, 7) mixed. Labeling each peak with its nearest gene to facilitate interpretation, we focused on Programs 2 and 4, reasoning that these represented a "slow" and more consequential rewiring of cell state, in contrast to possibly noisier, "faster" gene-driven programs operating within open chromatin.

Program 2, which was highly specific to the sgARID1A-containing population, featured multiple regulators of neural development and function as top hits **(Fig. 2e)**, intriguing given the neuroectodermal ancestry of the retinal pigment epithelium.[33] Further suggesting cellular plasticity induced by this program, we observed increased accessibility at the promoter of *NEUROD4*, a transcription factor normally expressed in retinal progenitors throughout embryonic development[34] that specifies photoreceptor differentiation and function[35] **(Fig. 2f)**. While *ARID1A* knockdown did not directly induce NEUROD4 expression, the increased promoter accessibility suggests it may be primed for activation by additional stimuli, hinting at a synergistic role for *ARID1A* and *NEUROD4* in cell fate determination.

Ordering accessibility and expression changes along pseudotime trajectories elucidates regulatory relationships

Finally, the single-cell resolution of Multiome Perturb-seq has the potential to yield deeper insights into the relationship between chromatin accessibility and gene expression. In Program 4, specific to Polycomb members *EZH2* and *SUZ12*, we found high feature weights on *SOX5* (a TF implicated in nervous system and skeletal system development[36,37]). The program contained several SOX5 target genes **(Fig. 2g)**, and we observed a global increase in accessibility around the *SOX5* TSS as well as a differentially accessible peak corresponding to a *SOX5* proximal enhancer in sgSUZ12-containing cells **(Fig. 2h)**.

Given these concordant changes, we reasoned that sgSUZ12-containing cells could anchor a pseudotime trajectory ordering heterogeneity within both control and perturbed populations, allowing us to visualize sequential chromatin opening preceding increased gene expression.[38–40] As expected, a diffusion pseudotime metric constructed on a combined population of 1,419 sgSUZ12-containing and sgNTC-containing cells (Methods) mapped a trajectory from control cells to perturbed cells **(Fig. 2i)**. We visualized the expression

5

and accessibility of the top 100 features by weight within Program 4 in single cells ordered by pseudotime, revealing broad upregulation aligned with this principal axis. **(Fig. 2j, top)** Some features "fired" earlier along the trajectory than others, suggesting candidate regulatory relationships between these "early" features and "late" targets. In the case of *SOX5*, accessibility of the proximal enhancer precedes increased *SOX5* expression and mirrors *SUZ12* knockdown, suggesting a causal role for *SUZ12* in repression of this locus and for this proximal enhancer in modulating the expression level of *SOX5* **(Fig. 2j, bottom)**. Overall, Multiome Perturb-seq's unique combination of single-cell resolution, CRISPR perturbations, and multi-modal profiling underscores its utility as a tool for assembling regulatory networks and elucidating mechanistic relationships connecting transcriptional states to their epigenomic foundations.

DISCUSSION

We developed Multiome Perturb-seq, a single-cell CRISPR screening method that efficiently assigns perturbation identities to simultaneous transcriptional and epigenomic phenotypes. Though we target use with robust, commercially available reagents, the protocol is likely simple to adapt to other workflows like combinatorial indexing[12, 41–43]or to in vivo applications that require nuclei isolation. Our T7-assisted amplification protocol and engineered Multiome Perturb-seq vector recovered perturbation identities for 68% of input nuclei, a >40% improvement over comparable techniques for scATAC-seq.[31] This method revealed that different perturbations distinctly alter chromatin accessibility and gene expression, underscoring the value of our multiomic approach in decoupling these layers of gene regulation. By leveraging single-cell heterogeneity, we could disentangle cis-regulatory dynamics and visualize sequential changes to chromatin state that foreshadowed changes in gene expression. Future screens employing our vector's lineage barcodes and measuring changes over time could be used to study epigenetic heterogeneity and memory within clonal lineages.[44]

Looking forward, we envision that Multiome Perturb-seq will not only enable experiments that advance basic biological discovery but also facilitate rational engineering of cell state. By simultaneously profiling the transcriptome and epigenome, this method provides a comprehensive view of both a cell's current state and potential future avenues for perturbation, as the epigenomic data can reveal genes newly primed for activation. Our observation of increased *NEUROD4* promoter accessibility in response to *ARID1A* knockdown exemplifies this concept, suggesting *NEUROD4* may be primed for activation and represent the next step in a defined regulatory path. Such insights could guide the design of synergistic perturbations, where an initial change in chromatin state sensitizes specific genes to subsequent interventions. This approach could direct programmed combinatorial genetic perturbations towards previously unreachable cell states, with relevance for both basic research and translational applications.

Acknowledgements

Author Contributions

Conceptualization, E.M., K.M.S. and T.M.N. Methodology, E.M., K.M.S., and T.M.N. Formal analysis, E.M. Investigation, E.M. and K.M.S. Writing – Original Draft, E.M. Writing – Review and Editing, E.M., K.M.S., and T.M.N. Visualization, E.M. Supervision, T.M.N. Funding Acquisition, T.M.N.

Declarations of Interest

T.M.N. is an author on U.S. Patent No. 11,214,797B2, related to Perturb-seq. The authors otherwise declare no competing interests.

Supplemental Information

Document S1: Figures S1-S3 and Supplemental Figure legends.

Document S2: Protocol for T7-assisted CROP-seq library prep, related to Figure 1.

Table S1: sgRNA protospacers used in this study, related to Figure 1 and Figure 2.

Table S2: Differentially expressed genes and ATAC-seq peaks in each perturbation condition, related to Figure 2.

Table S3: Quantification of concordant changes in differential expression and accessibility, related to Figure 2.

Table S4: Annotated features and weights for sparse programs, related to Figure 2.

Table S5: Oligonucleotide sequences used in this study.

METHODS

Experimental

Construction of engineered CROP-seq vectors

Vector pEM040 (+SIRLOIN/TSO) was constructed sequentially from vector pKS010 (available upon request from lead contact). A high-activity T7 promoter with appropriate overlaps was ordered from IDT as a duplexed oligo (Table S5) and cloned into PpuMI-linearized pKS010 via HiFi assembly (NEB #E2621). Ultramer oligos containing the SIRLOIN consensus sequence[22] and appropriate HiFi assembly overlaps were ordered from IDT (Table S5) and sequentially cloned into the 5' and 3' UTRs of the CROP-seq transcript, first via linearization with BsiWI (5' UTR) to create an intermediate vector, then via double digestion with SalI and BamHI (3' UTR), removing an internal Nextera Read 2 from the lineage barcode cassette in the parent vector. The partial TSO sequence was ordered from IDT as an Ultramer oligo and similarly cloned by HiFi assembly into an intermediate vector linearized with MluI. Vector pEM047 (-SIRLOIN/TSO) was constructed by adding the T7 promoter as described above and digesting with SalI / BamHI to remove the Nextera Read 2, but the 3' SIRLOIN oligo was replaced with a duplex oligo ligated into the digested product with Hi-T4 ligase (NEB #M2622) at RT for 1 hr. For each cloning reaction, assembled vectors were transformed into NEB Stable E. coli, (NEB #C3040H) and grown at 32˚C overnight. Single colonies were picked and miniprepped (Qiagen #27104), and sequences were verified by Plasmidsaurus whole plasmid sequencing.

sgRNA library construction and quality control

Vectors pEM040 and pEM047 were sequentially digested with BstXI and BlpI, and the desired 8119 bp / 7973 bp products were gel purified from an 0.8% agarose gel. Each vector contains a 1133-bp LacZ fragment that is removed by double digestion with BstXI and BlpI, allowing visualization of complete digestion during gel purification. sgRNAs were chosen based on previous published and unpublished data demonstrating on-target activity, and NTC guides were selected that minimized differential expression in previous screens. sgRNA libraries were ordered as IDT oPools with 5' overhang 5'- agtatcccttggagaaccaccttgttg-3' and 3' overhang 5'-gtttaagagctaagctggaaacagcatagcaag- 3'. For library cloning, oPools were resuspended at 100 ng/μL, and 10 μL (1 μg) was used as input to a second strand synthesis reaction with 2x KAPA HiFi Master Mix (Roche, #07958935001) and an equimolar amount of primer P276R (Table S5). The reaction was run in a thermal cycler with the following parameters: 95˚ 3', 64˚ 15" (ramp down 0.1˚C/sec), 72˚ 10', 4˚ hold. After purification of the double-stranded product (NEB #T1030), the library was cloned into digested pEM040 and pEM047 via HiFi assembly (NEB #E2621) for 1 hr at 50˚C, using a 5:1 molar ratio of insert library to vector. Electroporation was performed with an Eppendorf Eporator into Endura Electrocompetent Cells (LGC 60242-2) per manufacturer protocol. After electroporation, each reaction was incubated in 1 mL Recovery Media at 32˚C for 1.5 hr, then expanded into 100 mL LB + carbenicillin for overnight culture. Plasmid libraries were produced via midiprep (Qiagen #12941).

For QC of the constructed libraries, sgRNA protospacers were amplified via 2x KAPA HiFi Master Mix (Roche, #07958935001) from midipreps with primers containing Illumina P5 and P7 adapters and Nextera i7 indices (oJR232, oJR233, Table S5). Following 2x ProNex purification (Promega NG2001), amplicons were sequenced on an Illumina MiSeq with PE70 read structure. Resulting fastq files were processed with a custom Python script to quantify representation of each sgRNA in the library (Fig. S3).

Cell lines and lentiviral packaging

RPE-1-ZIM3 cells were obtained from Replogle and Saunders et al., 2022[45] and maintained in DMEM:F12 media (Gibco, 11320033), 10% Fetal Bovine Serum (FBS) (VWR, 97068-085), 100U penicillin and 100 mg/mL streptomycin (Gibco, 15140122), and 0.01 mg/mL Hygromycin B (Thermo Fisher, 10687-010). LentiX 293T cells were purchased from Takara Bio (632180) and maintained in DMEM (Gibco #11965092), 10% Fetal Bovine Serum (FBS) (VWR, 97068-085), 100U penicillin and 100 mg/mL streptomycin (Gibco, 15140122). For lentiviral packaging, LentiX cells were grown to 70-80% confluence and co-transfected with library vector, transfer, and packaging plasmids psPAX2 and pMD2G using TransIT-LTI Transfection Reagent (Mirus, MIR 2300). 24 hr post-transfection, media was changed to DMEM + 10% BSA (Sigma #A9418). Virus was harvested by filtering collected media through a 0.45μm filter 48 hr post-transfection.

Multiome Perturb-seq: experimental procedure

For the Multiome Perturb-seq experiment, populations of 1.2×10^7 RPE-1-ZIM3 cells were infected with lentivirus containing each library at a concentration of 400 μL harvested lentivirus per million cells (targeting MOI 0.1) and plated into 3 T175 flasks each in the presence of 8 μg/mL polybrene (Day 0). On Day 3, cells were sorted for GFP fluorescence by FACS (FacsSymphony S6, BD Biosciences). Infection titers of 6.6% and 9.4% were obtained for the +SIRLOIN/TSO vector and -SIRLOIN/TSO vector, respectively. 1.2×10^6 and 2.0×10^6 cells were recovered from sorting, and each population was plated in a 10cm dish. On Day 6, 1×10^6 cells from each population were prepared for encapsulation via nuclei isolation, following the 10x protocol "Nuclei Isolation for Single Cell Multiome ATAC + Gene Expression Sequencing" (CG000365, Rev C) with 4 min lysis time (Step 2d) and verification of complete lysis by Trypan Blue staining and visualization on the Countess II FL (Life Technologies). We observed that digitonin quality is extremely important to high-quality nuclei extraction in RPE-1 cells; digitonin that is expired or not thawed with rapid mixing at 65°C before use will result in incomplete lysis and affect the quality of extracted nuclei.

Isolated nuclei were resuspended in 500 μL PBS + 1 U/μL Sigma Protector RNase inhibitor (Sigma #3335402001), and 25 μL 7-AAD (Thermo #00-6993-50) was added to stain intact nuclei. Samples were immediately sorted by FACS (FacsSymphony S6, BD) to purify 7-AAD+ nuclei and remove debris, resulting in yields of 409,591 nuclei (+SIRLOIN/TSO) and 404,204 nuclei (-SIRLOIN/TSO). While sorting nuclei is an optional step of the 10x Multiome protocol, we observed that sorting for intact nuclei reduced ambient RNA which could impede ID assignment. Sorted nuclei were resuspended in 10x Nuclei Buffer targeting a concentration of 4,000 nuclei/μL, assuming 30% loss of the sorted yield during resuspension. Targeting 10,000 nuclei per reaction and accounting for the 10x-indicated recovery factor of 1.61, we loaded 4 μL $((10,000 \text{ nuclei} \times 1.61)/(4,000 \text{ nuclei}/\mu L))$ of each nuclei stock into transposition reactions per 10x protocols (Chromium Next GEM Single Cell Multiome ATAC + Gene Expression User Guide, Rev F), and proceeded with GEM generation, barcoding, and cleanup using the Chromium Controller and Chromium Next Gen Chip J (10x Genomics #1000283) with no protocol modifications. ATAC libraries were constructed with no protocol modifications and 7 PCR cycles for ATAC library amplification (Step 5.1d). cDNA amplification was performed with no protocol modifications and 6 PCR cycles (Step 6.1d). Gene expression libraries were constructed with no protocol modifications, using 12 PCR cycles for library amplification (Step 7.5d). After PicoGreen quantification and quality control by Agilent TapeStation, libraries were pooled to equimolar concentrations and sequenced to a target depth of 20,000 reads per cell on an Illumina NovaSeq X in a PE28/88 run using the NovaSeq X 10B or 25B Reagent Kit (100 or 300 Cycles) (Gene Expression) or to a target depth of 25,000 reads per cell on a NovaSeq 6000 in a PE50 run using the NovaSeq 6000 S1, S2, or S4 Reagent Kit (100 or 200 Cycles) (ATAC). The loading concentration was 1 nM (NovaSeq X) or 0.6-1.2 nM (NovaSeq 6000) and a

1% spike-in of PhiX was added to each run for quality control purposes.

Multiome Perturb-seq: T7-assisted library prep

A full step-by-step protocol is available in Document S2. To construct perturbation identity transcript libraries from pooled cDNA, 10 μL (25% of final yield by volume) of amplified cDNA from each library (Chromium Next GEM Single Cell Multiome ATAC + Gene Expression User Guide Rev F, Step 6.2n product) was in-vitro transcribed (IVT) with the NEB Quick T7 HiScribe Kit (NEB #E2050) in 25 μL reactions per manufacturer protocol. IVT reactions were incubated at 37°C for 2 hr. 3.8 μL dH2O and 1.2 μL DNase I were then added directly to each reaction (30 μL total reaction volume), and the reactions were incubated at 37°C for 15 min. IVT RNA product was purified with 1.8x RNAClean (Beckman Coulter #A63987) and eluted in 16.5 μL RNase-free dH2O. 1 μL product was used to quantify RNA yield with the Qubit RNA BR kit. The entire purified IVT product was used as template for reverse transcription with the Maxima H Minus First Strand cDNA Synthesis Kit (Thermo # K1652) as follows. On ice, 2 μL 10 μM RT primer (T7-TruSeqRead1, Table S5) and 1.25 μL 10 mM dNTPs were added to 15.5 μL IVT product. The reactions were incubated at 65°C for 5 min and immediately returned to ice. To each reaction on ice, 5 μL 5x RT buffer and 1.25 μL RT Enzyme Mix were added. Reactions were incubated at 50°C for 30 min and heat-inactivated at 85°C for 5 min. Next, the entire RT product was used directly in index PCR across 5 50 μL reactions to ensure that the volume of unpurified RT product did not exceed 10% of each individual PCR reaction. A 5x PCR master mix for each sample was built on ice as follows: to 25 μL unpurified RT product, we added 97 μL dH2O, 1.5 μL 100 μM P065-P5 (P5 reverse primer), 1.5 μL 100 uM P065-N7xx (indexed P7 forward primer, captures U6 promoter directly 5' of protospacer), and 125 μL 2x KAPA HiFi Master Mix (Roche #07958935001) (Table S5). The master mix was divided among 5 50-μL reactions in a PCR strip, and PCR was performed with the following parameters: 95° 3', 8 cycles of [98° 15", 70° 20"], 72° 1', 4° hold. PCR reactions were pooled and purified via 1.15x ProNex (+SIRLOIN/TSO) or 1.2x ProNex (-SIRLOIN/TSO). Yield was measured with the Qubit dsDNA HS kit (Thermo #Q32851) and amplicons were visualized as single peaks at 584 and 516 bp respectively on a Bioanalyzer (Agilent) using the High Sensitivity DNA Chip. After PicoGreen quantification and quality control by Agilent TapeStation, libraries were pooled to equimolar concentrations and sequenced to a target depth of 3,000 reads per cell on an Illumina NovaSeq X in a PE28/88 run using the NovaSeq X 10B or 25B Reagent Kit (100 or 300 Cycles). The loading concentration was 1 nM and a 1% spike-in of PhiX was added to the run for quality control purposes.

Multiome Perturb-seq: PCR enrichment CROP-seq library prep

ID transcripts were isolated from Gene Expression library (Chromium Next GEM Single Cell Multiome ATAC + Gene Expression User Guide Rev F, Step 7.6q product) for comparison with T7-assisted library prep by traditional CROP-seq. Transcripts were amplified from 200 ng library product across 5 50-μL PCR reactions with 2x KAPA HiFi Master Mix (Roche, #07958935001), P065-P5, and P065-N7xx (Table S5). PCR was performed with the following parameters: 95° 3', 20 cycles of [98° 15", 70° 10"], 72° 1', 4° hold. Reactions were pooled and purified with 1.1x (+SIRLOIN/TSO) or 1.15x (-SIRLOIN/TSO) ProNex. We reduced ProNex ratios slightly vs. T7 library prep to reduce the presence of starting material in final libraries, a known issue with CROP-seq. Yield was measured with the Qubit dsDNA HS kit (Thermo #Q32851) and amplicons were visualized as single peaks at 584 and 516 bp respectively on a Bioanalyzer (Agilent) using the High Sensitivity DNA Chip. After PicoGreen quantification and quality control by Agilent TapeStation, libraries were pooled to equimolar concentrations and sequenced on a NovaSeq X in a PE28/88 run using the NovaSeq X 10B or 25B Reagent Kit (100 or 300 Cycles). The loading concentration was 1nM and a 1% spike-in of PhiX was added to the run for quality control purposes. While we targeted the same sequencing depth as the T7 libraries, the

presence of gene expression library starting material in the final CROP-seq library resulted in fewer ID reads at the same total sequencing depth. Of 30 million target reads, we recovered 8.33 million (+SIRLOIN/TSO) and 13.8 million (-SIRLOIN/TSO).

RT-qPCR characterization of CROP-seq vectors

Leftover sorted nuclei and whole cells from the Multiome Perturb-seq experiment were used to quantify ID transcript expression. Total RNA was extracted from each sample (+/- SIRLOIN/TSO, nuclei and whole cells, 4 samples total) with the NEB Total RNA Miniprep kit (NEB #T2010). 10 ng nuclear RNA and 100 ng whole-cell RNA was used as template for RT-qPCR using the NEB Luna Universal One-Step RT-qPCR kit (NEB #E3005). PCR was performed per manufacturer recommendations, though the RT temperature was raised to 60°C. We targeted two 250 bp amplicons in the ID transcript for consistency, one within the coding sequence of GFP and the other within the U6 promoter. The *MALAT1* nuclear RNA was used as a normalization control, similarly to the original characterization of SIRLOIN elements.[22] C_t values were measured on a CFX96 Real-Time System (Bio-Rad), and ΔC_t was calculated as the difference between amplicon and *MALAT1* C_t values in each condition. $\Delta\Delta C_t$ was calculated by subtracting the C_t values of the +SIRLOIN/TSO samples from that of the -SIRLOIN/TSO samples, and expression fold change for each amplicon in the +SIRLOIN/TSO condition was calculated as $2^{-\Delta\Delta C_t}$. Standard errors were propagated through C_t calculations as previously described.[46]

Flow cytometry validation of on-target knockdown

1.5×10^5 RPE-1-ZIM3 cells were transduced with lentivirus containing CD81 or NTC sgRNAs in +SIRLOIN/TSO and -SIRLOIN/TSO CROP-seq vectors to a target MOI of 0.1 via plating with 8 μg/ml polybrene (Day 0). Transduced cells were sorted for GFP positivity on Day 3. On Day 6, cells were harvested, washed with PBS, and resuspended in 180 μL staining buffer (BD #554657). 20 μL CD81-APCH7 antibody (BD #656154) was added, and cells were incubated protected from light at 4°C for 15 min. Cells were then washed with 1 ml PBS and resuspended in 200 μL staining buffer, and CD81 signal was analyzed on an Attune CytPix flow cytometer (Thermo).

Computational

sgRNA identity calling

For analysis, sgRNA assignments were called on the T7 +SIRLOIN/TSO library with deeper sequencing using a custom Python script, `cropseq_call_guides_multiome.py`. Briefly, cellranger-7.1.0 was run on ID library reads, associating sgRNA-containing reads with 10x cell barcodes and UMIs (`cellranger count --chemistry=ARC-v1`). Based on cells called by cellranger-arc (below), reads were filtered to those associated with cell-containing droplets and aligned to the protospacer library with fuzzy string matching (rapidfuzz). Reads were collapsed by cell barcode and UMI, creating a dataframe of the ID transcripts in each cell (one of 16 options based on alignment to the protospacer library) labeled with UMI and total read counts. The relationship between UMI and total read counts was used to separate true ID transcripts from PCR artifacts as follows. First, UMI and total read counts were log2-transformed, and possible "Top IDs" representing true assignments with log2(UMI) >= 2.75 were selected. An HDBSCAN model (`min_samples = 10, metric = 'l1', cluster_selection_epsilon = 0.5`) was fit to Top IDs to further filter these putative true assignments. Minority clusters and IDs categorized as noise (label -1) were removed, and the cluster with higher average UMI was selected as representative of true assignments. A logistic regression model was trained on all Top

IDs, then applied to the entire dataset of ID transcripts to infer membership in the true assignment cluster. IDs for which the model predicted membership in this cluster were kept as true assignments, and other IDs were removed. This procedure results in a clear plane of separation between called and non-called transcripts (Fig. S2), and removes likely PCR chimeras (e.g., transcripts for which the UMI-total read relationship deviates significantly from confident assignments, such as transcripts with a single UMI and read representing a likely template-switching event during index PCR). We additionally filtered the model predictions such that no ID transcript with less than 3 UMI would be called, regardless of its cluster prediction.

For comparisons of assignment between library prep and vector conditions, we downsampled individual reads from the above dataframe of ID transcripts (`pd.DataFrame.sample()`) before collapsing to UMI and total read counts, such that the total number of reads in this dataframe divided by the number of cells called by cellranger was constant (850 reads per cell, based on the minimum common sequencing depth available across conditions). We ran the ID assignment script as above, but changed the "Top ID" log2(UMI) threshold to 2 (T7) or 1.75 (CROPseq).

Single-cell data processing and QC filtering

cellranger-arc-2.0.2 (10x Genomics) was used for alignment of Gene Expression and ATAC libraries to the transcriptome and genome respectively (reference: cellranger-arc-GRCh38-2020-A-2.0.0), using cellranger-arc count with default parameters. For gene expression processing, sgRNA ID assignments were merged with cellranger gene expression outputs in Scanpy,[47] and quality control metrics were calculated with `sc.pp.calculate_qc_metrics()`. We inspected these metrics by sgRNA assignment status (Fig. S3) and filtered low-quality cells based on total UMI and library complexity (number of unique genes). Mitochondrial genes were removed, and cells were filtered to those with single sgRNA assignments. The resulting AnnData object was further filtered using the perturbseq codebase[3] to remove genes expressed in fewer than 5% of cells with `strip_low_expression(threshold = 0.05)`, and counts were then z-normalized to control by normalizing total counts in all cells to median UMI, subtracting the NTC mean of each gene, and dividing by the NTC standard deviation. The resulting normalized matrix was read back into Scanpy to regress out cell cycle genes (obtained from Harmonizome.[48])

For ATAC processing, 10x fragment files were converted to tile matrices with SnapATAC2.[49] QC metrics were calculated, and cells with ≤1000 unique fragments were removed. From the genome-wide tile matrix, 2.5×10^5 features were selected for downstream analyses based on accessibility in all cells. The combined quality control process applied to both modalities left 4,724 high-quality nuclei with single sgRNA assignments and complete expression and accessibility profiles.

Identification of differentially expressed genes and differentially accessible peaks

Differentially expressed genes were called on the filtered z-normalized matrix described above with a Kolmogorov-Smirnov test (`perturbseq.ks_de()`). P-values were corrected for false discovery rate via the Benjamini-Hochberg procedure, and genes with FDR-corrected q < 0.1 were considered differentially expressed. Differentially accessible ATAC peaks were identified using SnapATAC2's suggested workflow. From the filtered tile matrix described above, peaks were called with MACS3[50] (`snap.tl.macs3()`) on pseudobulk accessibility profiles quantified with paired-insertion counting[51,52] for each perturbation target and NTC, and a cell-by-peak matrix with fixed-width, non-overlapping features was created with `snap.tl.merge_peaks()` followed by `snap.tl.make_peak_matrix(counting_strategy = 'paired-insertion')`. For each perturbation, differentially accessible peaks were identified with a Mann-Whitney test (`scipy.stats`) between counts in NTC and perturbed cells, performed on each feature of

the cell-by-peak matrix for which any perturbed or NTC cells had at least one count. P-values were corrected for false discovery rate via the Benjamini-Hochberg procedure, and peaks with FDR-corrected q < 0.1 were considered differentially accessible. Log fold change values were calculated as the log2-transformed ratio of mean counts in perturbed cells to control cells. A pseudocount of 1×10^{-3} was added to the numerator and denominator during this calculation to avoid infinite values. To identify concordant differential changes, each peak was labeled with its nearest gene, and matches were considered concordant if both the peak and its nearest gene were differentially expressed or accessible with q < 0.1 and the sign of the gene expression z-score was equal to the sign of the peak log fold change.

Annotation of ATAC-seq peaks with ENCODE cis-regulatory elements

A .bed file of cis-regulatory elements was downloaded from ENCODE SCREEN.[53] To label ATAC differential features, PyRanges was used to intersect this bed file with the coordinates of fixed-width merged peaks composing the features of the cell-by-peak matrix described above. Peaks were labeled with CREs based on overlap of at least 1bp. To label total peaks for Table S3, unique peaks from MACS3 output were intersected and labeled with the regulatory element for which the peak had the largest overlap. For a finer-grained characterization of promoters, we supplemented promoter annotations with a database of human promoters downloaded from the Eukaryotic Promoter Database,[54] and we extended coordinates by 950 bp upstream to represent promoter regions. Features annotated as "prom" overlap a promoter region. The SOX5 enhancer peak shown in Figs. 2i and 2j was labeled as overlapping a promoter region, but contained a proximal enhancer in the ENCODE dataset. We characterize the peak as an enhancer based on its ENCODE annotation to maintain consistency, but the feature remains labeled as overlapping a promoter region in the raw data (e.g., in Table S4).

Clustering of expression and accessibility profiles

Gene expression profiles were clustered by decomposing the processed single-cell expression matrix with PCA into 10 components. Component values were grouped by perturbation identity and averaged, and these pseudobulk embeddings were hierarchically clustered using 1 - Pearson correlation as the distance metric and Ward's linkage. ATAC profiles were clustered by decomposing the processed single-cell tile matrix with SnapATAC2's spectral embedding algorithm into 11 components. We removed the first component of this embedding, finding that it correlated strongly with the number of unique fragments recovered in each single cell. The remaining 10 components were similarly grouped by perturbation identity and averaged, and the pseudobulk spectral embeddings were clustered by hierarchical clustering with 1 - cosine similarity as the distance metric and Ward's linkage. For clustermap generation in each case, we used `hc.linkage(method = 'ward', optimal_ordering = True)` from `scipy.cluster.hierarchy`.

Sparse PCA co-embedding of expression and accessibility features

To organize perturbation effects and decompose peak-driven vs. gene-driven changes, we adapted a sparse PCA approach used by our group to interpret Perturb-seq data to the Multiome setting. For this analysis, we first z-normalized paired insertion counts within the peak matrix as follows: from the cell-peak matrix described above, features were pre-filtered to remove the lower 25% based on count values in all cells and any peaks more than 50 kb away from the nearest gene. Means and standard deviations of counts in NTC cells for each feature were then generated, and the filtered peak matrix was z-normalized by subtracting the NTC mean and dividing by the NTC standard deviation. From the z-normalized peak matrix, we then

removed peaks that were not differentially accessible in any perturbations, resulting in 5,060 peaks. We similarly z-normalized gene expression matrices as described above, and removed genes not differentially expressed in any perturbations. This resulted in a concatenated expression/accessibility matrix of 5,060 peaks and 7,885 genes for downstream analysis.

Our sparse PCA model was constructed by modifying sklearn's SparsePCA object to constrain feature weights to be nonnegative, using `dict_learning(positive_code = True)`. This procedure promotes interpretability by 1) limiting the number of gene and peak features within each resulting program, as the model is fit with a regularization term penalizing nonzero feature weights; and 2) ensuring that features within a program are co-regulated in the same direction by each perturbation, via the nonnegativity constraint. We fit a sparse PCA model to the above concatenated matrix, deciding empirically on the number of programs to fit (limited by the low-rank structure of the matrix, with only 13 perturbations and 3 underlying complexes) and the value of the regularization parameter alpha. Fitting 8 programs with alpha = 0.25 resulted in expected clustering of perturbation weight vectors into known complexes, with a sparsity level that allowed for substantial gene-peak heterogeneity between programs while not sacrificing the value of the method as a feature selection tool. Peak features in each program were then labeled with their nearest gene and ENCODE cis-regulatory element annotations based on overlap, if applicable. Features, weights, and annotations for each program are available in Table S4. Nervous system genes highlighted in Fig. 2 were manually curated from Program 2 features. *SOX5* target genes were obtained from Harmonizome.[48]

Generation and visualization of ATAC-seq coverage tracks

Bigwig coverage tracks were generated at pseudobulk level for each perturbation and NTC cells as follows. The .bam output of cellranger-arc was filtered with pysam to include only reads with the 'CB' tag corresponding to perturbation-assigned cell barcodes, creating a pseudobulk .bam file for each perturbed subpopulation and NTC. These files were indexed with samtools index and converted to bigwig coverage tracks with deepTools `bamCoverage --normalizeUsing RPGC --effectiveGenomeSize 2913022398 --binSize 20`. Tracks were visualized with pyGenomeTracks.[55]

Construction of pseudotime trajectories

We applied a diffusion pseudotime approach as previously described[40] to create an ordered embedding of control and perturbed single cells. We constructed a new gene expression matrix in scanpy by normalizing each cell to equal total UMI and log1p-transforming normalized counts, avoiding z-normalization to maintain meaningful values for NTC cells. Similarly, starting with the cell-by-peak matrix as above, we filtered peaks more than 50 kb away from the nearest gene and tf-idf normalized raw paired insertion count values. We then constructed an individual pseudotime trajectory for each perturbation as follows. We selected only features that were differentially expressed or accessible in the perturbation and concatenated these features into a combined expression and accessibility matrix. For many perturbations, the target gene and promoter were differentially expressed and accessible (Fig. 1); we removed these genes and peaks from the calculation to maintain on-target expression and accessibility as a "positive control" to evaluate the fidelity of our embedding in separating perturbed and control cells. We then scaled each feature to zero mean and unit variance and decomposed the matrix into 10 principal components. We constructed the nearest-neighbors graph on this PCA embedding using `sc.pp.neighbors(metric = 'correlation')`, and constructed a diffusion map with `sc.tl.diffmap()`. To set a root cell for diffusion pseudotime construction (required by scanpy's implementation), we identified the diffusion component with the highest absolute-value Pearson correlation with a binary perturbed/control variable for each single cell, and chose the cell at the appropriate extreme of

14

this component (lowest value if correlation was negative, highest value if positive) as the root cell.

We then constructed the pseudotime with `sc.tl.dpt()`, and visualized cells ordered by their pseudotime rank via a UMAP embedding fit on the above concatenated differential expression / accessibility matrix. Pseudotime ranks aligned well with perturbed/control identities for most perturbations. Notably, alternative trajectories that do not require the a priori choice of a root cell (single-component spectral embedding, linear discriminant analysis) produced indistinguishable results when applied to the combined population. Visualizations (Fig. 2j) were constructed by smoothing ordered single-cell expression and accessibility values with a 500-cell rolling average, then scaling each feature separately to a range of 0-1.
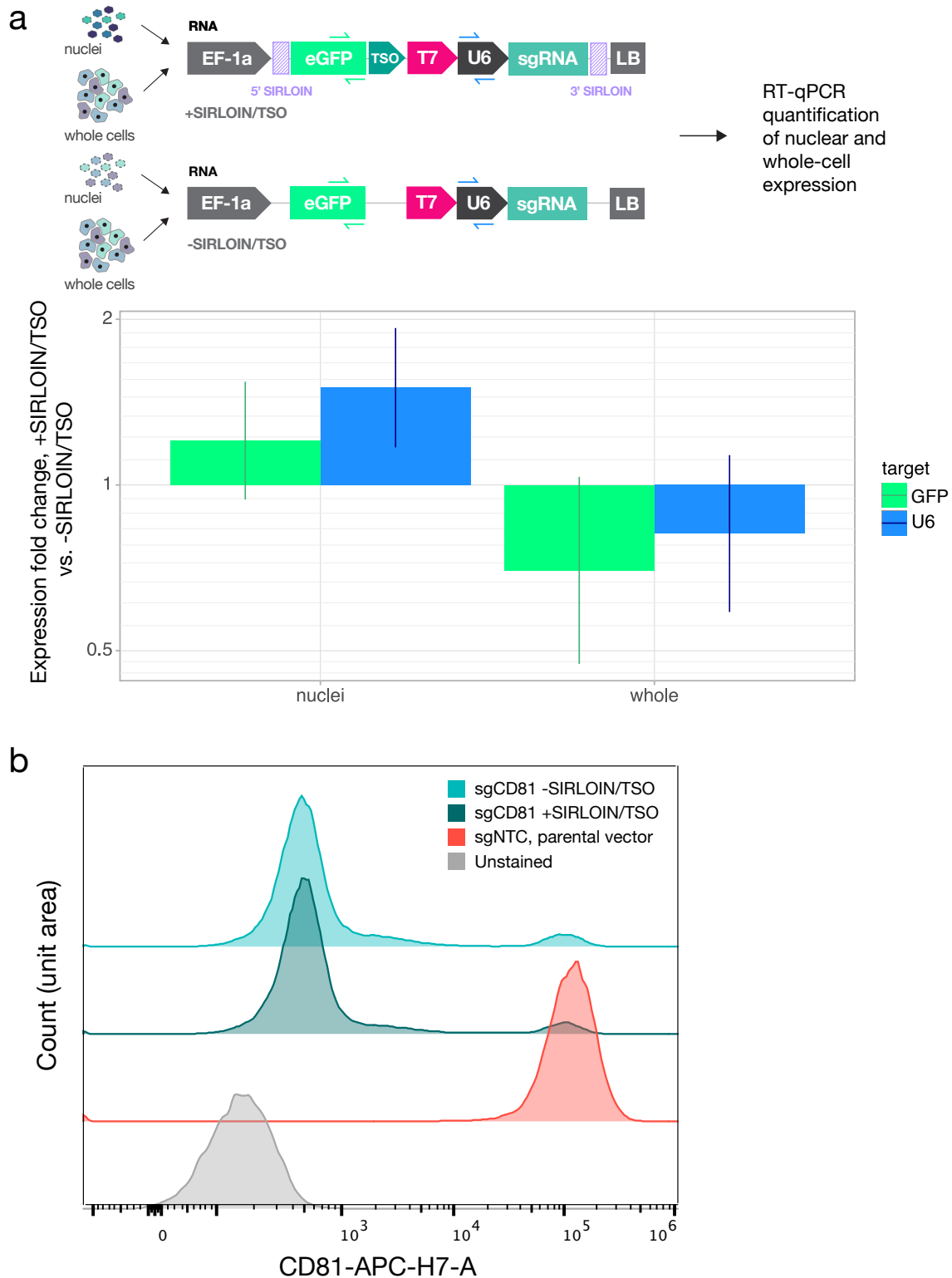
Fig. S1: Characterization of engineered CROP-seq vectors. **a)** RT-qPCR quantification of nuclear and whole-cell expression of CROP-seq vectors with and without SIRLOIN elements (purple). Error bars show 95% confidence intervals across 3 technical replicates of each condition targeting a unique amplicon within the ID transcript, each of length 250 bp. **b)** Flow cytometry validation of on-target knockdown of cell surface marker CD81 by sgRNAs expressed from each vector in RPE-1-ZIM3 cells.
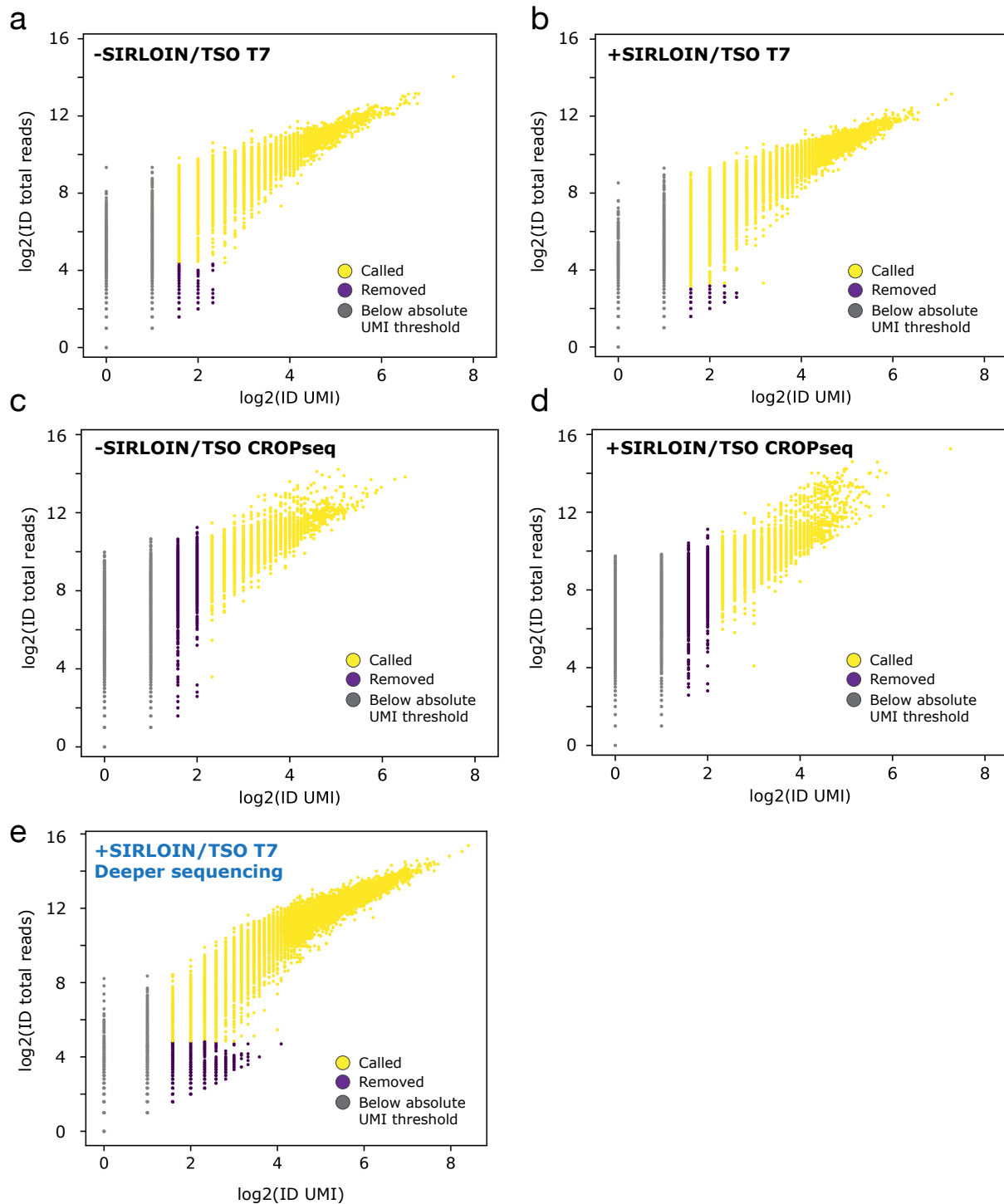
Fig. S2: Outputs of CROP-seq ID calling pipeline. **a)** Called vs. non-called ID transcripts in nuclei-containing droplets, -SIRLOIN/TSO vector, T7-assisted library prep, downsampled to 850 reads per nucleus. **b)** Called vs. non-called ID transcripts in nuclei-containing droplets, +SIRLOIN/TSO vector, T7-assisted library prep, downsampled to 850 reads per nucleus. **c)** Called vs. non-called ID transcripts in nuclei-containing droplets, -SIRLOIN/TSO vector, CROPseq library prep, downsampled to 850 reads per nucleus. **d)** Called vs. non-called ID transcripts in nuclei-containing droplets, +SIRLOIN/TSO vector, CROPseq library prep, downsampled to 850 reads per nucleus. **e)** Called vs. non-called ID transcripts in nuclei-containing droplets, +SIRLOIN/TSO vector, T7-assisted library prep, full sequencing depth (4,034 reads per nucleus).
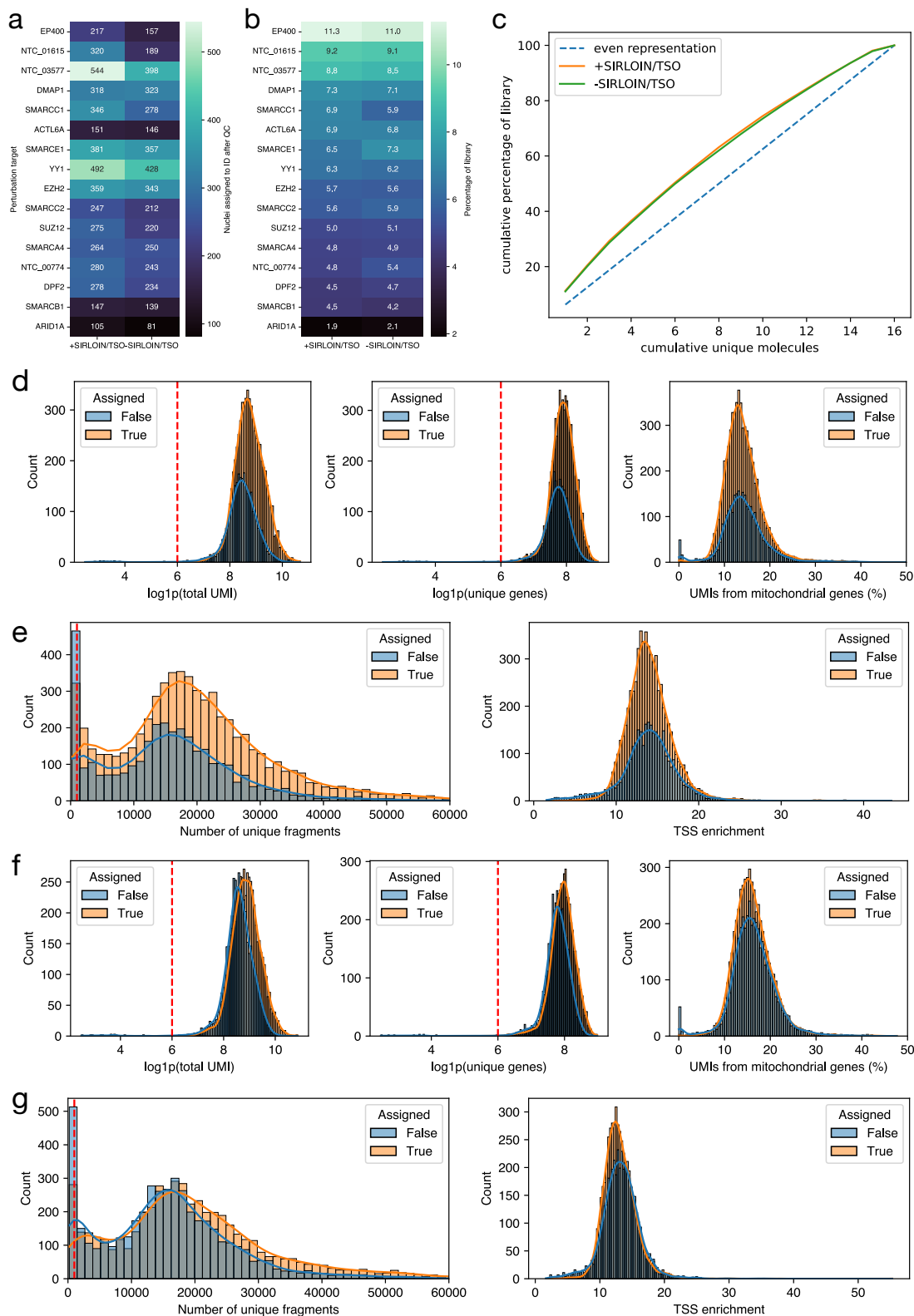
Fig. S3: Quality control metrics related to the Multiome Perturb-seq screen. **a)** Number of cells assigned to each perturbation in each vector condition after QC. **b)** Representation of sgRNAs in cloned libraries. **c)** Gini plot showing skew of sgRNA representation in each library. **d)** Gene expression UMI, library complexity (number of unique genes), and mitochondrial transcript percentage in all cellranger-called cells, stratified by ID assignment, in +SIRLOIN/TSO condition. Dashed lines show QC filters applied. **e)** Unique ATAC fragments recovered and TSS enrichment in all cellranger-called cells, stratified by ID assignment, in +SIRLOIN/TSO condition. Dashed lines show QC filters. **f)** Gene expression UMI, library complexity (number of genes), and mitochondrial transcript percentage in all cellranger-called cells, stratified by ID assignment, in -SIRLOIN/TSO condition. **g)** Unique ATAC fragments recovered and TSS enrichment in all cellranger-called cells, stratified by ID assignment, in -SIRLOIN/TSO condition.

REFERENCES

[1] Christoph Bock, Paul Datlinger, Florence Chardon, Matthew A. Coelho, Matthew B. Dong, Keith A. Lawson, Tian Lu, Laetitia Maroc, Thomas M. Norman, Bicna Song, Geoff Stanley, Sidi Chen, Mathew Garnett, Wei Li, Jason Moffat, Lei S. Qi, Rebecca S. Shapiro, Jay Shendure, Jonathan S. Weissman, and Xiaowei Zhuang. High-content CRISPR screening. *Nature Reviews Methods Primers*, 2(1):8, February 2022.

[2] Stefan Peidli, Tessa D. Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J. Schumacher, Jake P. Taylor-King, Debora S. Marks, Augustin Luna, Nils Blüthgen, and Chris Sander. scPerturb: harmonized single-cell perturbation data. *Nature Methods*, 21(3):531–540, March 2024.

[3] Britt Adamson, Thomas M. Norman, Marco Jost, Min Y. Cho, James K. Nuñez, Yuwen Chen, Jacqueline E. Villalta, Luke A. Gilbert, Max A. Horlbeck, Marco Y. Hein, Ryan A. Pak, Andrew N. Gray, Carol A. Gross, Atray Dixit, Oren Parnas, Aviv Regev, and Jonathan S. Weissman. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell*, 167(7):1867–1882.e21, December 2016.

[4] Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M. Norman, Eric S. Lander, Jonathan S. Weissman, Nir Friedman, and Aviv Regev. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, 167(7):1853–1866.e17, December 2016.

[5] Paul Datlinger, André F Rendeiro, Christian Schmidl, Thomas Krausgruber, Peter Traxler, Johanna Klughammer, Linda C Schuster, Amelie Kuchler, Donat Alpar, and Christoph Bock. Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods*, 14(3):297–301, March 2017.

[6] Shiqi Xie, Jialei Duan, Boxun Li, Pei Zhou, and Gary C. Hon. Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Molecular Cell*, 66(2):285–299.e5, April 2017.

[7] Patrik L. Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O. Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson, Simone Codeluppi, Åke Borg, Fredrik Pontén, Paul Igor Costea, Pelin Sahlén, Jan Mulder, Olaf Bergmann, Joakim Lundeberg, and Jonas Frisén. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, July 2016.

[8] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868, September 2017.

[9] Eleni P. Mimitou, Anthony Cheng, Antonino Montalbano, Stephanie Hao, Marlon Stoeckius, Mateusz Legut, Timothy Roush, Alberto Herrera, Efthymia Papalexi, Zhengqing Ouyang, Rahul Satija, Neville E. Sanjana, Sergei B. Koralov, and Peter Smibert. Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nature Methods*, 16(5):409–412, May 2019.

[10] Franco Izzo, Robert M. Myers, Saravanan Ganesan, Levan Mekerishvili, Sanjay Kottapalli, Tamara Prieto, Elliot O. Eton, Theo Botella, Andrew J. Dunbar, Robert L. Bowman, Jesus Sotelo, Catherine Potenski, Eleni P. Mimitou, Maximilian Stahl, Sebastian El Ghaity-Beckley, JoAnn Arandela, Ramya Raviram, Daniel C. Choi, Ronald Hoffman, Ronan Chaligné, Omar Abdel-Wahab, Peter Smibert, Irene M. Ghobrial, Joseph M.

Scandura, Bridget Marcellino, Ross L. Levine, and Dan A. Landau. Mapping genotypes to chromatin accessibility profiles in single cells. *Nature*, 629(8014):1149–1157, May 2024.

[11] Sai Ma, Bing Zhang, Lindsay M. LaFave, Andrew S. Earl, Zachary Chiang, Yan Hu, Jiarui Ding, Alison Brack, Vinay K. Kartha, Tristan Tay, Travis Law, Caleb Lareau, Ya-Chieh Hsu, Aviv Regev, and Jason D. Buenrostro. Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell*, 183(4):1103–1116.e20, November 2020.

[12] Junyue Cao, Darren A. Cusanovich, Vijay Ramani, Delasa Aghamirzaie, Hannah A. Pliner, Andrew J. Hill, Riza M. Daza, Jose L. McFaline-Figueroa, Jonathan S. Packer, Lena Christiansen, Frank J. Steemers, Andrew C. Adey, Cole Trapnell, and Jay Shendure. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, 361(6409):1380–1385, September 2018.

[13] Bradley E. Bernstein, Tarjei S. Mikkelsen, Xiaohui Xie, Michael Kamal, Dana J. Huebert, James Cuff, Ben Fry, Alex Meissner, Marius Wernig, Kathrin Plath, Rudolf Jaenisch, Alexandre Wagschal, Robert Feil, Stuart L. Schreiber, and Eric S. Lander. A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell*, 125(2):315–326, April 2006.

[14] Alvaro Rada-Iglesias, Ruchi Bajpai, Tomek Swigut, Samantha A. Brugmann, Ryan A. Flynn, and Joanna Wysocka. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, 470(7333):279–283, February 2011.

[15] David Lara-Astiaso, Assaf Weiner, Erika Lorenzo-Vivas, Irina Zaretsky, Diego Adhemar Jaitin, Eyal David, Hadas Keren-Shaul, Alexander Mildner, Deborah Winter, Steffen Jung, Nir Friedman, and Ido Amit. Chromatin state dynamics during blood formation. *Science*, 345(6199):943–949, August 2014.

[16] Karun Kiani, Eric M Sanford, Yogesh Goyal, and Arjun Raj. Changes in chromatin accessibility are not concordant with transcriptional changes for single-factor perturbations. *Molecular Systems Biology*, 18(9):e10979, September 2022.

[17] Sam John, Peter J Sabo, Robert E Thurman, Myong-Hee Sung, Simon C Biddie, Thomas A Johnson, Gordon L Hager, and John A Stamatoyannopoulos. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature Genetics*, 43(3):264–268, March 2011.

[18] Marta Moreno-Oñate, Lourdes Gallardo-Fuentes, Pedro M Martínez-García, Silvia Naranjo, Sandra Jiménez-Gancedo, Juan J Tena, and José M Santos-Pereira. Rewiring of the epigenome and chromatin architecture by exogenously induced retinoic acid signaling during zebrafish embryonic development. *Nucleic Acids Research*, 52(7):3682–3701, April 2024.

[19] Hiroyuki Fuke and Mutsuhito Ohno. Role of poly (A) tail as an identity element for mRNA nuclear export. *Nucleic Acids Research*, 36(3):1037–1049, February 2008.

[20] Thomas Conrad, Izabela Plumbom, Maria Alcobendas, Ramon Vidal, and Sascha Sauer. Maximizing transcription of nucleic acids with efficient T7 promoters. *Communications Biology*, 3(1):439, August 2020.

[21] Takamasa Kudo, Ana M. Meireles, Reuben Moncada, Yushu Chen, Ping Wu, Joshua Gould, Xiaoyu Hu, Opher Kornfeld, Rajiv Jesudason, Conrad Foo, Burkhard Höckendorf, Hector Corrada Bravo, Jason P. Town, Runmin Wei, Antonio Rios, Vineethkrishna Chandrasekar, Melanie Heinlein, Shuangyi Cai, Cherry Sakura Lu, Cemre Celen, Noelyn Kljavin, Jian Jiang, Jose Sergio Hleap, Nobuhiko Kayagaki, Felipe De Sousa E Melo, Lisa McGinnis, Bo Li, Avtar Singh, Levi Garraway, Orit Rozenblatt-Rosen, Aviv Regev, and Eric

Lubeck. Highly multiplexed, image-based pooled screens in primary cells and tissues with PerturbView, December 2023.

[22] Yoav Lubelsky and Igor Ulitsky. Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. *Nature*, 555(7694):107–111, March 2018.

[23] Catherine Gutierrez, Aziz M. Al'Khafaji, Eric Brenner, Kaitlyn E. Johnson, Satyen H. Gohil, Ziao Lin, Binyamin A. Knisbacher, Russell E. Durrett, Shuqiang Li, Salma Parvin, Anat Biran, Wandi Zhang, Laura Rassenti, Thomas J. Kipps, Kenneth J. Livak, Donna Neuberg, Anthony Letai, Gad Getz, Catherine J. Wu, and Amy Brock. Multifunctional barcoding with ClonMapper enables high-resolution study of clonal dynamics during tumor evolution and treatment. *Nature Cancer*, 2(7):758–772, July 2021.

[24] Alejo E. Rodriguez-Fraticelli, Caleb Weinreb, Shou-Wen Wang, Rosa P. Migueles, Maja Jankovic, Marc Usart, Allon M. Klein, Sally Lowell, and Fernando D. Camargo. Single-cell lineage tracing unveils a role for TCF15 in haematopoiesis. *Nature*, 583(7817):585–589, July 2020.

[25] Amjad Askary, Luis Sanchez-Guardado, James M. Linton, Duncan M. Chadly, Mark W. Budde, Long Cai, Carlos Lois, and Michael B. Elowitz. In situ readout of DNA barcodes and single base edits facilitated by in vitro transcription. *Nature Biotechnology*, 38(1):66–75, January 2020.

[26] Nader Alerasool, Dmitri Segal, Hunsang Lee, and Mikko Taipale. An efficient KRAB domain for CRISPRi applications in human cells. *Nature Methods*, 17(11):1093–1096, November 2020.

[27] Shuang He, Zihan Wu, Yuan Tian, Zishuo Yu, Jiali Yu, Xinxin Wang, Jie Li, Bijun Liu, and Yanhui Xu. Structure of nucleosome-bound human BAF complex. *Science*, 367(6480):875–881, February 2020.

[28] Weisi Lu, Lekun Fang, Bin Ouyang, Xiya Zhang, Shaoquan Zhan, Xuyang Feng, Yaofu Bai, Xin Han, Hyeung Kim, Quanyuan He, Ma Wan, Feng-Tao Shi, Xin-Hua Feng, Dan Liu, Junjiu Huang, and Zhou Songyang. Actl6a Protects Embryonic Stem Cells From Differentiating Into Primitive Endoderm. *Stem Cells*, 33(6):1782–1793, June 2015.

[29] Zhanping Lu, Courtney C. Hong, Guangyao Kong, Anna L.F.V. Assumpção, Irene M. Ong, Emery H. Bresnick, Jing Zhang, and Xuan Pan. Polycomb Group Protein YY1 Is an Essential Regulator of Hematopoietic Stem Cell Quiescence. *Cell Reports*, 22(6):1545–1559, February 2018.

[30] Anne Laugesen, Jonas Westergaard Højfeldt, and Kristian Helin. Role of the Polycomb Repressive Complex 2 (PRC2) in Transcriptional Regulation and Cancer. *Cold Spring Harbor Perspectives in Medicine*, 6(9):a026575, September 2016.

[31] Sarah E. Pierce, Jeffrey M. Granja, and William J. Greenleaf. High-throughput single-cell chromatin accessibility CRISPR screens enable unbiased identification of regulatory networks in cancer. *Nature Communications*, 12(1):2969, May 2021.

[32] Luke A. Gilbert, Matthew H. Larson, Leonardo Morsut, Zairan Liu, Gloria A. Brar, Sandra E. Torres, Noam Stern-Ginossar, Onn Brandman, Evan H. Whitehead, Jennifer A. Doudna, Wendell A. Lim, Jonathan S. Weissman, and Lei S. Qi. CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. *Cell*, 154(2):442–451, July 2013.

[33] Eberhard Buse, Thomas Eichmann, Heink deGroot, and Anja Leker. Differentiation of the mammalian retinal pigment epithelium in vitro: Influence of presumptive retinal neuroepithelium and head mesenchyme. *Anatomy and Embryology*, 187(3), March 1993.

34 Timothy J. Cherry, Sui Wang, Ingo Bormuth, Markus Schwab, James Olson, and Constance L. Cepko. NeuroD Factors Regulate Cell Fate and Neurite Stratification in the Developing Retina. *The Journal of Neuroscience*, 31(20):7365–7379, May 2011.

35 Yichen Ge, Xushen Chen, Nan Nan, Jonathan Bard, Fuguo Wu, Donald Yergeau, Tao Liu, Jie Wang, and Xiuqian Mu. Key transcription factors influence the epigenetic landscape to regulate retinal cell differentiation. *Nucleic Acids Research*, 51(5):2151–2176, March 2023.

36 Kenneth Y. Kwan, Mandy M. S. Lam, Željka Krsnik, Yuka Imamura Kawasawa, Veronique Lefebvre, and Nenad Šestan. SOX5 postmitotically regulates migration, postmigratory differentiation, and projections of subplate and deep-layer neocortical neurons. *Proceedings of the National Academy of Sciences*, 105(41):16021–16026, October 2008.

37 Patrick Smits, Peter Dy, Srijeet Mitra, and Véronique Lefebvre. *Sox5* and *Sox6* are needed to develop and maintain source, columnar, and hypertrophic chondrocytes in the cartilage growth plate. *The Journal of Cell Biology*, 164(5):747–758, March 2004.

38 Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386, April 2014.

39 Manu Setty, Vaidotas Kiseliovas, Jacob Levine, Adam Gayoso, Linas Mazutis, and Dana Pe'er. Characterization of cell fate probabilities in single-cell data with Palantir. *Nature Biotechnology*, 37(4):451–460, April 2019.

40 Laleh Haghverdi, Maren Büttner, F Alexander Wolf, Florian Buettner, and Fabian J Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13(10):845–848, October 2016.

41 Noa Liscovitch-Brauer, Antonino Montalbano, Jiale Deng, Alejandro Méndez-Mancilla, Hans-Hermann Wessels, Nicholas G. Moss, Chia-Yu Kung, Akash Sookdeo, Xinyi Guo, Evan Geller, Suma Jaini, Peter Smibert, and Neville E. Sanjana. Profiling the genetic determinants of chromatin accessibility with scalable single-cell CRISPR screens. *Nature Biotechnology*, 39(10):1270–1277, October 2021.

42 Alexander B. Rosenberg, Charles M. Roco, Richard A. Muscat, Anna Kuchina, Paul Sample, Zizhen Yao, Lucas T. Graybuck, David J. Peeler, Sumit Mukherjee, Wei Chen, Suzie H. Pun, Drew L. Sellers, Bosiljka Tasic, and Georg Seelig. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, 360(6385):176–182, April 2018.

43 Junyue Cao, Jonathan S. Packer, Vijay Ramani, Darren A. Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N. Furlan, Frank J. Steemers, Andrew Adey, Robert H. Waterston, Cole Trapnell, and Jay Shendure. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352):661–667, August 2017.

44 Sydney M. Shaffer, Benjamin L. Emert, Raúl A. Reyes Hueros, Christopher Cote, Guillaume Harmange, Dylan L. Schaff, Ann E. Sizemore, Rohit Gupte, Eduardo Torre, Abhyudai Singh, Danielle S. Bassett, and Arjun Raj. Memory Sequencing Reveals Heritable Single-Cell Gene Expression Programs Associated with Distinct Cellular Behaviors. *Cell*, 182(4):947–959.e17, August 2020.

[45] Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, Jeffrey A. Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J. Wagner, Karen Adelman, Gila Lithwick-Yanai, Nika Iremadze, Florian Oberstrass, Doron Lipson, Jessica L. Bonnar, Marco Jost, Thomas M. Norman, and Jonathan S. Weissman. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575.e28, July 2022.

[46] Arun Kumar and Daniel Lorand. Robust ΔΔct estimate. *Genomics*, 113(1):420–427, January 2021.

[47] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, December 2018.

[48] Andrew D. Rouillard, Gregory W. Gundersen, Nicolas F. Fernandez, Zichen Wang, Caroline D. Monteiro, Michael G. McDermott, and Avi Ma'ayan. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database*, 2016:baw100, 2016.

[49] Kai Zhang, Nathan R. Zemke, Ethan J. Armand, and Bing Ren. A fast, scalable and versatile tool for analysis of single-cell omics data. *Nature Methods*, 21(2):217–227, February 2024.

[50] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137, November 2008.

[51] Zhen Miao and Junhyong Kim. Uniform quantification of single-nucleus ATAC-seq data with Paired-Insertion Counting (PIC) and a model-based insertion rate estimator. *Nature Methods*, 21(1):32–36, January 2024.

[52] Laura D. Martens, David S. Fischer, Vicente A. Yépez, Fabian J. Theis, and Julien Gagneur. Modeling fragment counts improves single-cell ATAC-seq analysis. *Nature Methods*, 21(1):28–31, January 2024.

[53] The ENCODE Project Consortium, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, 583(7818):699–710, July 2020.

[54] René Dreos, Giovanna Ambrosini, Romain Groux, Rouaïda Cavin Périer, and Philipp Bucher. The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. *Nucleic Acids Research*, 45(D1):D51–D55, January 2017.

[55] Lucille Lopez-Delisle, Leily Rabbani, Joachim Wolff, Vivek Bhardwaj, Rolf Backofen, Björn Grüning, Fidel Ramírez, and Thomas Manke. pyGenomeTracks: reproducible plots for multivariate genomic datasets. *Bioinformatics*, 37(3):422–423, April 2021.