A single-cell massively parallel reporter assay detects cell type specific cis-regulatory activity

Siqi Zhao[1,2], Clarice KY Hong[1,2], David M Granas[1,2], Barak A Cohen[1,2]
1 Edison Family Center for Systems Biology and Genome Sciences, 2 Department of Genetics, Washington University School of Medicine, USA.

We developed a single-cell massively parallel reporter assay (scMPRA) to measure the activity of libraries of cis-regulatory sequences (CRSs) across multiple cell-types simultaneously. As a proof of concept, we assayed a library of core promoters in a mixture of HEK293 and K562 cells and showed that scMPRA is a reproducible, highly parallel, single-cell reporter gene assay. Our results show that housekeeping promoters and CpG island promoters have lower activity in K562 cells relative to HEK293, which likely reflects developmental differences between the cell lines. Within K562 cells, scMPRA identified a subset of developmental promoters that are upregulated in the CD34^{+}/CD38^{-} sub-state, confirming this state as more "stem-like." Finally, we deconvolved the intrinsic and extrinsic components of promoter cell-to-cell variability and found that developmental promoters have a higher proportion of extrinsic noise compared to housekeeping promoters, which may reflect the responsiveness of developmental promoters to the cellular environment. We anticipate scMPRA will be widely applicable for studying the role of CRSs across diverse cell types.

Introduction

The majority of heritable variation for human diseases maps to the non-coding portions of the genome[1–6]. This observation has led to the hypothesis that genetic variation in the *cis*-regulatory sequences (CRSs) that control gene expression underlies a large fraction of disease burden[7–10]. Because many CRSs function only in specific cell types[11], there is intense interest in high-throughput assays that can measure the effects of cell-type-specific CRSs and their genetic variants.

Massively Parallel Reporter Assays (MPRAs) are one family of techniques that allow investigators to assay libraries of CRSs and their non-coding variants *en masse*[12–18]. In an MPRA experiment, every CRS drives a reporter gene carrying a unique DNA barcode in its 3' UTR, which allows investigators to quantify the activity of each CRS by the ratio of its barcode abundances in the output RNA and input DNA. This approach allows investigators to identify new CRSs, assay the effects of non-coding variants, and discover general rules governing the functions of CRSs[12,19–23]. One limitation of MPRAs is that they are generally performed in monocultures, or as bulk assays across the cell types of a tissue. Performing cell-type specific MPRAs in tissues will require methods to simultaneously readout reporter gene activities and cell type information in heterogeneous pools of cells.

To address this problem, we developed scMPRA, a procedure that combines single-cell RNA sequencing with MPRA. scMPRA simultaneously measures the activities of reporter genes in single cells and the identities of those cells using their single-cell transcriptomes. The key

component of scMPRA is a two-level barcoding scheme that allows us to measure the copy number of all reporter genes present in a single cell from mRNA alone. A specific barcode marks each CRS of interest (CRS barcode, "cBC") and a second random barcode (rBC) acts as a proxy for DNA copy number of reporter genes in single cells (**Fig. 1a**). The critical aspect of the rBC is that it is complex enough to ensure that the probability of the same cBC-rBC appearing in the same cell more than once is vanishingly small. In this regime, the number of different cBC-rBC pairs in a single cell becomes an effective proxy for the copy number of a CRS in that cell. Even if a cell carries reporter genes for multiple different CRS, and each of those reporter genes is at a different copy number, it is still possible to normalize each reporter gene in each individual cell to its plasmid copy number. With this barcoding scheme, we can measure the activity of many CRSs with different input abundances in single cells.

Results

scMPRA enables single-cell measurement of CRS activity

As a proof of principle, we used scMPRA to test whether different classes of core promoters show different activities in different cell types. Core promoters are the non-coding sequences that surround transcription start sites, where general cofactors interact with RNA polymerase II[24,25]. Core promoters are divided into different classes by the functions of their host genes (housekeeping vs developmental), as well as by the sequence motifs they contain (TATA-box, downstream promoter element (DPE), and CpG islands). We selected 676 core promoters that we previously tested[24] and cloned them into a double-barcoded MPRA library (**Supplementary Table 1**). In the first stage of library construction each core promoter reporter gene was represented by 10 unique cBCs. We then added rBCs to the library by cloning a 25 nt random oligonucleotide (oligo) directly downstream of the cBCs. The library contains ~ $1.4X10^7$ unique cBC-rBC pairs (**Methods, Fig. 1b**). Using this complexity, we calculated that the probability of plasmids with the same cBC-rBC pair occurring in the same cell is less than $2X10^{-3}$ with our transfection protocols (**Methods**). Given this low likelihood, the number of rBC per cBC in a cell represents the copy number of a CRS in that cell. Knowing the copy number of CRSs in single cells allows us to normalize reporter gene expression from each CRS to its copy number in individual cells.

We performed a cell mixing experiment to test whether scMPRA could measure cell type specific expression of reporter genes. We transfected K562 and HEK293 cells (**Methods**), and performed scMPRA on a 1:1 mixture of those cell lines (**Fig. 1c**). We harvested cells and prepared them for sequencing using the 10X Chromium™ platform. The mRNA from single cells was captured, converted to cDNA, and pooled together. We then split the samples, with a quarter of the amplified cDNA library used for amplifying the cBC-rBC pairs and three-quarters used to amplify the transcriptome. The resulting reporter barcode abundances and transcriptome of each single cell are linked by their shared 10X cell barcode (**Methods**).

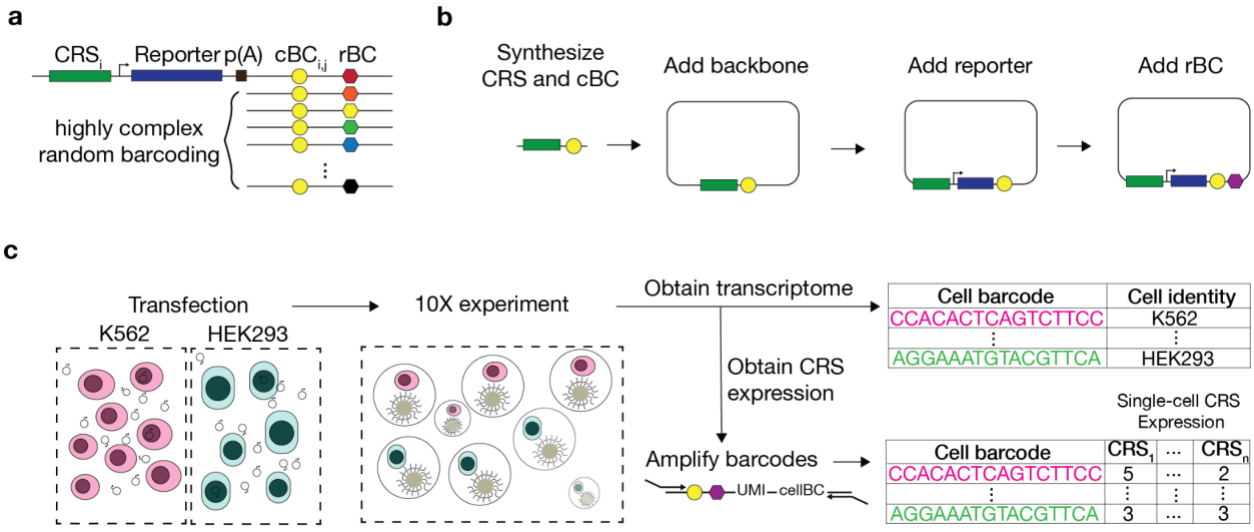


Figure 1 scMPRA measures CRS at single-cell resolution. (**a**) Each CRS reporter construct is barcoded with a cBC that encodes the identity of the CRS, as well as a highly complex rBC. The complexity of the cBC-rBC pair ensures that the probability of identical plasmids being introduced into the same cell is extremely low. (**b**) Cloning strategy for the double barcoded library. CRSs and their corresponding cBCs are synthesized together and cloned into an appropriate backbone. 25 nt rBCs are introduced to the plasmids with Hifi assembly. (**c**) Experimental overview for scMPRA using mixed cell experiment as an example. K562 cells and HEK293 cells are transfected with the double-barcoded core promoter library. After 24 hours, cells were harvested and mixed for 10X scRNA-seq. Cell identities were obtained through measuring the single transcriptome, and single-cell expression of CRSs was obtained by quantifying the barcodes. The cell identity and CRSs expression were linked by the shared 10X barcodes.

We recovered a total of 3112 cells (1524 in replicate 1 and 1588 in replicate 2) that are unambiguously assigned to one of the two cell types (**Fig. 2a, Supplementary Figs S1 a,b**). We determined the efficiency of our method by calculating the recovery rate of our input promoters. We then calculated the core promoter expression by taking the average of the cBC expression for the same promoter. We found that scMPRA recovered 99.5% (673 out of 676 core promoters) of the input library for K562 cells and 100% (676 out of 676 core promoters) for HEK293 cells, highlighting the efficiency of our method for recovering input elements.

We next calculated the number of individual cells in which each core promoter is measured. We found that the empirical distribution of the number of cells per core promoter is log normal, with a median of 76 cells per core promoter for K562 cells and 287 cells per core promoter for HEK293 cells (**Fig. 2b,c**). Given that the number of pBC-rBC pair is effectively the number of plasmids per cell, we also calculated the number of plasmid per cell, and found that fewer number of plasmids were incorporated into K562 cells compare to HEK293 cells (median plasmid number in K562 cells: 164, median plasmid number in HEK293 cells: 341. **Supplementary Fig. 1c,d**). The difference in transfection efficiency between these cell types with the same input likely reflects global cellular differences between them, and is representative of the condition when performing scMPRA in different cell types.

We calculated the biological reproducibility and found that scMPRA is highly reproducible in both cell types for measurements of mean expression (K562: Pearson R = 0.89, HEK293: Pearson R = 0.96) and cell-to-cell variance (K562: Pearson R = 0.78, HEK293: Pearson R = 0.94, **Fig 2 d-g**). To validate the measurements, we conducted bulk RNA-seq for the core promoter library in the two cell types separately, and found the bulk measurements correlate well with the aggregated single-cell measurements (**Fig. 2 h,i, Supplementary Fig. 1e,f**). This analysis shows

that single-cell measurements of library members in as few as 70 individual cells still correlate well with bulk measurements, highlighting the sensitivity of our method.

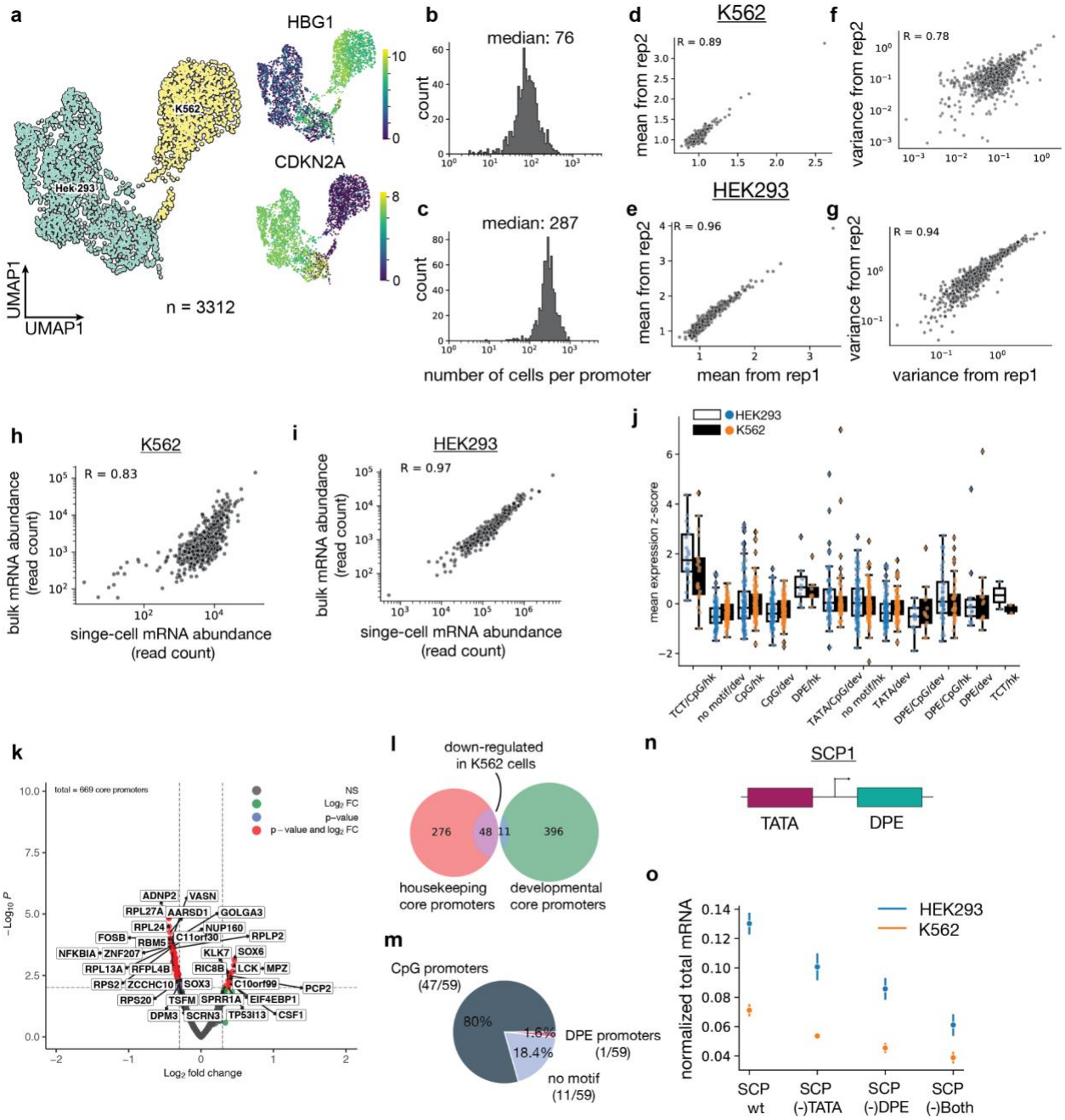


Figure 2. scMPRA detects cell type specific CRS activity. (**a**) UMAP of the transcriptome from the mixed-cell scMPRA experiment. 3312 out of 3417 cells are assigned to either K562 or HEK293 cells. Cell-type specific genes were used to identify the cell clusters (HBG1 for K562 cells and CDKN2A for HEK293 cells). Cells are labeled by their cell type. (**b,c**) Histogram of the number of cells per core promoter for HEK293 and K562 cells. (**d-g**) Reproducibility for expression mean and cell-to-cell variance for both K562 and HEK293 cells. (**h,i**) Scatterplot of reproducibility of scMPRA mean expression with bulk MPRA measurement using read count normalization. (**j**) Boxplot of mean expression from different categories of core promoters in K562 (orange) and HEK293 (blue) cells. (**k**) Volcano plot for differential expression (DE) of the core promoters in K562 and HEK293 cells (Significant DE reporters have p-value <0.01 and log-2 fold change greater than 0.3). (**l**) A Venn diagram of the functional characterization (housekeeping vs developmental) of down-regulated reporters in K562 cells. Housekeeping promoters are enriched (p-value = 1.08×10^{-11} from hypergeometric test). (**m**) Pie chart of the sequence features (CpG, DPE, TATA) of down-regulated reporter genes.

CpG promoters are enriched ($p=2.18x10^{-6}$, from hypergeometric test). (**n**) Schematic of SCP1 binding sites. (**o**) Expression of wild-type and mutated ($TATA^-$, DPE^-, and Both) versions of SCP1 core promoter (error bar: 1 s.d.)

scMPRA detects cell type specific CRS activity and non-coding variant effect

We asked whether the data allowed us to detect core promoters with differential activity between K562 and HEK293 cells. While different classes of core promoters had similar activities in both cell lines (**Fig. 2j**), our differential analysis using DEseq2[26] identified a small number of promoters (11 out of 669) that are upregulated in K562 cells, and 59 promoters that are downregulated in K562 cells (adjusted $p< 0.01$, log2 fold change > 0.3, **Fig. 2k, Supplementary Table 2**). Among the down-regulated promoters, 48 out of 59 core promoters belong to housekeeping genes ($p = 1.08x10^{-11}$**, Fig. 2l**), and 46 out of 59 core promoters are CpG-island-containing core promoters ($p=2.18x10^{-6}$**, Fig. 2m**). This down-regulation might be explained by the fact that the K562 cell line is a cancer derived cell line, and a hallmark regulatory change in cancer cells is the hypermethylation of CpG promoters[27]. These results demonstrate the ability of scMPRA to detect CRSs with cell-type specific activities.

Another application of scMPRA is to detect cell type specific effects of non-coding variants. To test whether our method can detect the effects of mutations in a given CRS, we included an artificial core promoter SCP1[28] along with mutated versions without a TATA Box or DPE motif in our library (**Fig. 2n)**. We first computed the total number of captured reporter gene transcripts, since it is the closest proxy to the bulk expression measurement. We found that deletions of the TATA motif or DPE motif both reduced expression (**Fig. 2o**) and we observed a similar trend in the bulk data (**Supplementary Fig. 1g**). When we directly calculated the mean of the single-cell expression distribution instead of total number of captured reporter gene transcripts, we found that the deletion of the DPE motif has a stronger effect in K562 cells than in HEK293 cells (40% reduction vs 20% reduction) (**Methods, Supplementary Fig. 1 h,i**). We hypothesized that the differential expression of transcription factors between K562 and HEK293 cells leads to differential sensitivity to the TATA and DPE motifs. We examined the single-cell transcriptome and found that TAF9, which recognizes the DPE motif[29], is more highly expressed in K562 cells compared to HEK293 Cells (**Supplementary Fig. 1j**, Wilcoxon $p=4.27x10^{-94}$). This observation likely explains why the deletion of the DPE motif has a stronger effect in K562 cells. Our results demonstrate that scMPRA can identify and explain cell-type specific effects of non-coding variants.

scMPRA detects cell sub-state specific CRS activity

Single-cell studies have revealed heterogeneity in cell states even within isogenic cell types[30–33]. Therefore, we asked if scMPRA can identify CRSs with cell-state specific activity. We repeated scMPRA on K562 cells alone and obtained a total of 5141 cells from two biological replicates. Measurements of the mean and variance of each library member were again highly correlated between replicates and agree well with independent bulk measurement (**Supplementary Fig. 2 a-d**).

As the phases of the cell cycle represent distinct cell-states, we asked whether scMPRA could identify reporter genes with differential activity through the cell cycle. We assigned cell cycle phases to each cell using their single cell transcriptome data (**Fig. 3a**) and then calculated the mean expression of each reporter gene in different cell cycle phases. We found that most core

promoters in our library are upregulated in the G1 phase of the cell cycle, and some housekeeping promoters are highly expressed through all cell cycle phases (**Fig. 3b**). We also identified core promoters with different expression dynamics through the cell cycle. For example, we found the core promoter for UBA52 remains highly expressed in the S phase, whereas the core promoter for CXCL10 is lowly expressed throughout (**Supplementary Fig. 2e**). This analysis illustrates the ability of scMPRA to identify CRSs whose expression naturally fluctuates with cellular dynamics.

We then asked whether scMPRA could detect reporter genes with activities that were specific to other cell-states in K562 cells, after normalizing for cell cycle effects. We focused on two specific sub-states that have been reported and experimentally validated for high proliferation rates in K562 cells[34,35]. The first is the $CD34^{+}/CD38^{-}$ sub-state that has been identified as a leukemia stem-cell subpopulation, and the second is the $CD24^{+}$ sub-state that is linked to selective activation of proliferation genes by bromodomain transcription factors[31,32]. To identify these sub-states in our single-cell transcriptome data, we first regressed out the cell cycle effects and confirmed that the single cell transcriptome data no longer clustered by cell cycle phase (**Supplementary Fig.2 f**). We then identified clusters within K562 cells that have the $CD34^{+}/CD38^{-}$ expression signature, or the $CD24^{+}$ signature (**Fig. 3 c,d**). Although the $CD34^{+}/CD38^{-}$ cells represent only 9.3% of the cells in a K562 culture, scMPRA revealed two distinct classes of core promoters that are upregulated and downregulated in these cells respectively (**Fig 3e**). Conversely, the expression patterns of promoters are similar between the $CD24^{+}$ cluster and cells in the "differentiated" cluster (**Fig. 3e, f**). Motif analysis of the up/down regulated classes of promoters in $CD34^{+}/CD38^{-}$ cells showed that different core promoter motifs are enriched in each class, with the TATA box and Motif 5 being enriched in the upregulated class and MTE and TCT motifs being enriched in downregulated class (**Fig. 3g, Methods**). This result suggests that differences in core promoter usage might be driving the differences between $CD34^{+}/CD38^{-}$ and the other clusters. Because the TATA box is mostly found in developmental core promoters, the $CD34^{+}/CD38^{-}$ subpopulation likely reflects a more "stem-like" cellular environment in these cells. Our analysis highlights the ability of scMPRA to identify CRSs with differential activity in rare cell populations.

With the single-cell expression data, we asked how certain promoters achieve higher expression in the $CD34^{+}/CD38^{-}$ state. We asked whether the single-cell expression distribution for the $CD34^{+}/CD38^{-}$ state is shifted higher than for the other states, or if the range of expression is the same for each sub-state, with only the proportion of cells with high expression changing in each state. To answer this question, we calculated the proportion of cells in each sub-state belonging to the 90th percentile of the total single cell expression distribution. For the majority of promoters, the $CD34^{+}/CD38^{-}$ cluster has a much higher proportion of cells in the 90th percentile (**Supplementary Fig 3a**). At the same time, there is no difference in the maximum expression of cells in different sub-states, and this maximum level is mainly set by the promoter identity (**Supplementary Fig 3b**). Even for the most differentially expressed promoter in the $CD34^{+}/CD38^{-}$ subpopulation, TIA1, the expression distributions for cells in the three sub-states cover the same range, but the proportion of cells in the right-tail of the distribution is higher for $CD34^{+}/CD38^{-}$ cells (**Fig. 3h**). This result suggests that the "stem-like" cellular environment of the $CD34^{+}/CD38^{-}$ subpopulation increases the probability of certain promoters having higher expression, without shifting the maximum expression those promoters achieve. Taken together, these analyses

highlight how the joint transcriptome and CRS measurements in scMPRA can be used to understand differences in behavior in cellular sub-states.

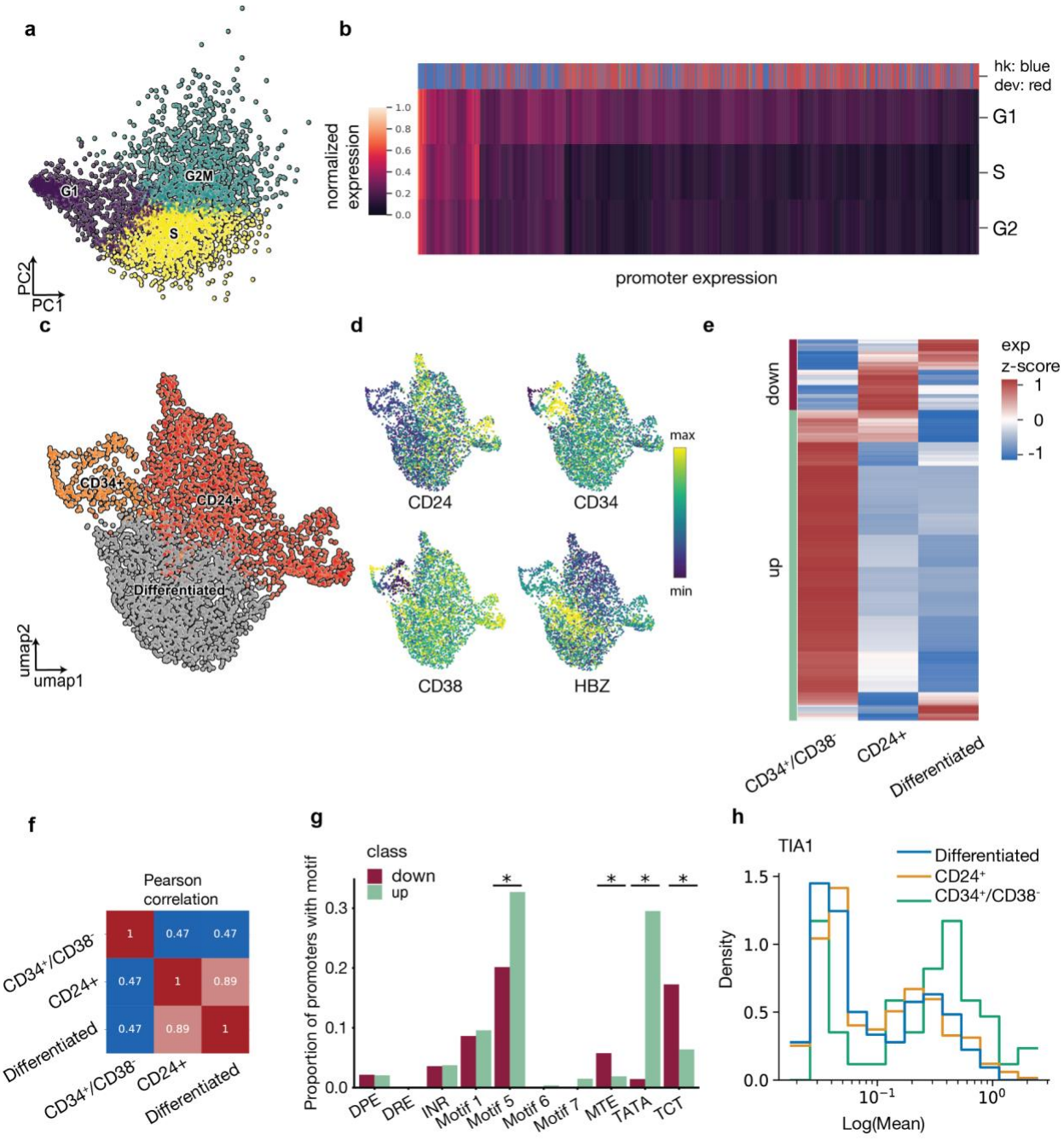


Figure 3. scMPRA detects cell sub-state-specific CRS activity. (**a**) PCA plot of K562 cells classified based on the cell cycle score. (**b**) Heatmap of reporter expression in different cell cycle phases (Color bar indicates housekeeping (blue) vs developmental (red) promoters). (**c**) Representative expression dynamics of reporter genes through cell cycle for UBA52, CSF1, and CXCL10. (**d**) UMAP embedding of K562 cells with high proliferation sub-states ($CD34^+/CD38^-$ and $CD24^+$). (**e**) Marker gene expression signifies different cell sub-states in K562 cells. CD34, CD38 marks the "leukemia stem cell" sub-state; CD24 marks a high proliferation sub-state, and HBZ marks the differentiated leukemia sub-state; left color bar: hierarchical clustering showing 2 clusters based on expression pattern in the three substates. (**f**) Heatmap showing the correlation matrix of core promoter expression in three substates ($CD34^+/CD38^-$, $CD24^+$, and Differentiated). (**g**) Proportion of promoters in each cluster that contains the indicated core promoter

motif. * represents significant enrichment in one cluster over the other ($p < 0.05$, Fisher's exact test). (**h**) Histogram of single-cell expression of TIA1 promoter in three substates.

scMPRA enables decomposition of intrinsic and extrinsic noise

Finally, we analyzed the cell-to-cell variability of reporter genes across K562 cells. Cell-to-cell variability, or expression noise, is the phenomenon where gene expression varies among the cells of an isogenic population. Expression noise has important roles in development[36], rare-cell cancer resistance[30,37], and its origin is a central question in single-cell biology. A common framework for studying expression noise is to decompose it into its intrinsic component, which arises from the thermal fluctuations of macromolecular interactions, and its extrinsic component, which results from fluctuations in the global cellular environment[38–42]. Intrinsic and extrinsic noise can be decomposed using dual-reporter experiments, where two identical reporter genes are measured across the same single-cells[39]. High covariance of the two reporter genes indicates high extrinsic noise and low intrinsic noise, while independent variation of the two reporters suggests high intrinsic noise and low extrinsic noise. In scMPRA, plasmids with the same CRS but different barcodes are sometimes incorporated into the same cells, effectively serving as a dual-reporter experiment. We extracted pair-wise expression for the same core promoter labeled with different cBCs from our scMPRA data, and computed intrinsic noise and extrinsic noise using a previously developed statistical framework[43] (**Methods**). We found that different core promoters have distinct intrinsic and extrinsic noise profiles (**Fig 4 a,b**). Globally, we found that intrinsic noise correlates with mean expression levels (Pearson $\rho = 0.455$), while extrinsic noise is not correlated with mean expression (Pearson $\rho = -0.172$, **Fig. 4 c,d**). This result agrees with the notion that intrinsic noise arises from the thermodynamics of transcription at different promoters, whereas many sources for extrinsic noise are independent of the specific promoters. We also found that developmental promoters have a higher proportion of noise that is extrinsic, reflecting their role in driving developmental promoters that respond to extrinsic cues during development (**Fig. 4 e,f**). This analysis suggests that scMPRA could be a powerful tool to study the mechanistic origin of cell-to-cell variability in a high throughput manner.

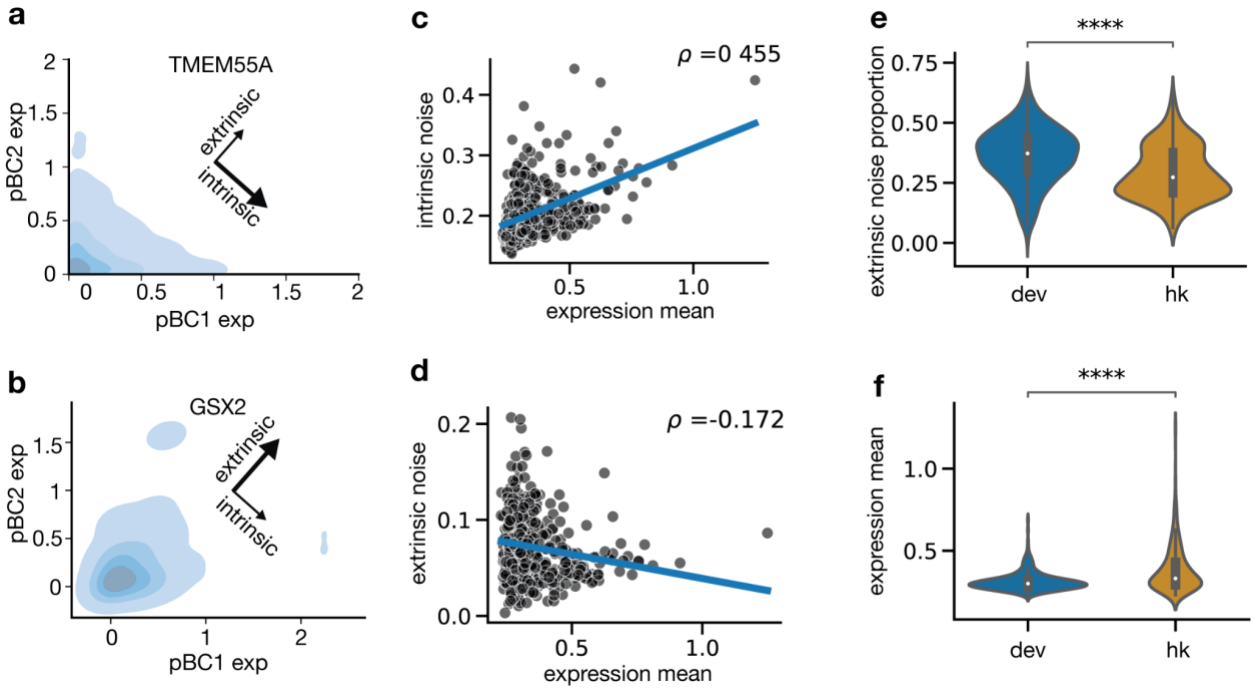


Figure 4. scMPRA deconvolves intrinsic and extrinsic cell-to-cell variability. (**a, b**) Density plots for single-cell expression of paired cBC expression for the same promoter. TMEM55A has high intrinsic noise, and GSX2 has high extrinsic noise. (**c**) Scatterplot of expression against intrinsic noise. Blue line shows the linear regression (Pearson ρ = 0.455) (**d**) Scatterplot of expression against extrinsic noise. Blue line shows the linear regression (Pearson ρ = -0.172) (**e**) Violin plot of extrinsic noise proportion for housekeeping and developmental promoters (Mann-Whitney U test. Starts indicate significance: **** : $p < 1X10^{-4}$) (**f**) Violin plot of expression mean for housekeeping and developmental promoters (Mann-Whitney U test. Stars indicate significance: **** : $p < 1X10^{-4}$)

Conclusions

We have presented a method to measure the cell-type and cell-state specific effects of CRSs by devising a barcoding scheme to read out input copy number with mRNA. We demonstrated that scMPRA detects cell-type specific reporter gene activity in a mixed population of cells, and cell-state specific activity in an isogenic population. We also demonstrated that scMPRA can be a powerful tool to study how different CRS control cell-to-cell variability. The assay is reproducible and reports accurate mean levels of reporter gene activity in as few as 70 cells. The primary limitation of scMPRA is that it relies on mRNA counts of the rBC to estimate plasmid DNA abundance, and therefore it cannot accurately measure CRSs that are truly silent in a given cell type. The inclusion of a separate constitutive promoter on each plasmid driving expression of the rBCs would allow us to quantify plasmid copy number independent of the expression of the reporter gene.

A future direction is to perform scMPRA in complex tissues to measure the cell type specific effects of genetic variation in CRSs. With the burgeoning of Adeno-associated viral delivery systems with distinct tropisms[44–47], we anticipate that scMPRA will be widely used to study cis-regulatory effects in a variety of complex tissues.

Methods

Cell culture

K562 cells were cultured using a medium consisting of Iscove's Modified Dulbecco's Medium (IMDM) + 10% Fetal Bovine Serum (FBS) + 1% non-essential amino acids + 1% pen/strep at 37 C with 5% of CO_2. HEK293 cells were cultured using a medium consisting of Eagle's Minimum Essential Medium (EMEM) + 10% Fetal Bovine Serum (FBS) + 1% pen/strep at 37 C with 5% of CO_2.

Cloning Strategy

We developed a two-level barcoding technology to enable single-cell normalization for plasmid copy number. We applied this strategy to a promoter library we previously tested in bulk assays[24]. The original library contains 676 core promoters with a length of 133bp. Each core promoter has 10 promoter barcodes to provide redundancy in the measurements. We then synthesized a single-stranded 90 bp DNA oligonucleotide containing a 25 bp random sequence, a restriction site, and 30 bp homology on each side of the barcode region.

We used Hifi Assembly™ to add the random barcodes to the plasmid library. 4 µg of the plasmid library were split into 4 reactions and digested with 2µl of SalI for 1.5 hours at 37℃. The digested products were run at 100V for 2 hours on a 0.7% agarose gel. The correct size band was cut and purified with the Monarch Gel Extraction Kit (New England BioLabs T1020L). The insert single-stranded DNA was diluted in TE to a stock concentration of 100 uM. The insert was then further diluted to 1 uM with ddH2O. Three assembly reactions were pooled together, each reaction containing 100 ng of digested library backbone, 1 uM of insert DNA, 1µl of NEBuffer 2, 10 µl of 2X Hifi assembly mix, and H2O up to 20 ul, The reaction was incubated at 50°C for 1 hour. The assembled product was purified with the Monarch PCR&DNA Cleanup kit (New England BioLabs T1030L) and eluted in 12 µl of H2O.

The assembled plasmid was transformed using Gene Pulser Xcell Electroporation Systems by electroporation (BIO-RAD 1652661), 50 µl of ElectroMax DH10B electrocompetent cells (Invitrogen 18290015) with 1 µl of hifi assembled product at 2 kV, 2000 Ω, 25 nF, with 1 mm gap. 950 µl of SOC medium (Invitrogen 15544034) was added to the cuvette and then transferred to a 15 ml Falcon tube. Two transformations were performed, and each tube was incubated at 37 ℃ for 1 hour on a rotator with 300 rpm. The culture was then added to pre-warmed 150 µl LB/Amp medium and grown overnight at 37 ℃. 1 µl of the culture was also diluted 1:100 and 50 µl of the diluted cultured was plated on a LB agar plate to check the transformation efficiency. For the core promoter library, we obtained more than $4X10^8$ colonies, large enough to cover a complex library.

Estimating Library Complexity

To estimate the library complexity, we sequenced the DNA library using a nested PCR-based Illumina library preparation protocol. Briefly, we first used Q5 polymerase (New England BioLabs M0515) to amplify the region containing the two barcodes with SCARED P17 (5'-GACGAGCTCTATAAGTAATCTAGA-3') and SCARED P18 (5'-TTTTCTAGGTCTCTGGTCGA-3'). The total reaction volume is 50µl with 50ng of plasmids with 2.5 µl of 10uM primer each. The annealing temperature is 61°C with an extension time of 10s. 25 cycles of amplification were done. The product was then purified with the Monarch PCR&DNA Cleanup kit (New England BioLabs T1030L), and eluted with 20 µl of ddH2O. For the second PCR (SCARED P19: 5'-GGACGAGCTCTATAAGTAATCTAGA-3', SCARED P20: 5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3'), a 25 µl reaction was set up with 0.25 µl product from the previous step, the annealing temperature is 61°C, and the extension time is 10s, a total of 10 cycles was done. The PCR product was cleaned up using the Monarch PCR&DNA Cleanup kit. For the last PCR to add the P5 and P7 Illumina adapters (P5: 5'-AATGATACGGCGACCACCGAGATCTACACACCCGCACACTCTTTCCCTACACGACGCT-3', P7:5'-CAAGCAGAAGACGGCATACGAGATAAGTTGACAGTGACTGGAGTTCAGACGTG-3'), a reaction with 25 µl of total volume was set up with 2 µl of cleaned product from PCR2, a total of 10 cycles of PCR was done.

The constructed Illumina library was sequenced on an Illumina MiSeq. A total of 1,693,933 reads was generated for this library. A filtering strategy was applied to the raw reads, where reads that do not have matching promoter barcodes and wrong-length random barcodes were filtered out. We obtained a total of 1,359,176 reads (80% of the total reads) that contain the correct promoter barcode and correct length random barcode.

The shallow sequencing of the input plasmid library enabled us to estimate the library complexity and the probability of two identical copies of the plasmid being transfected into the same cell. We first calculated that each random barcode is attached to 1.9 promoter barcodes on average. For a total of 6760 input promoter barcodes, this suggests that a given random barcode is being reused by 3200 different promoters. The reuse of random barcodes is the effective labeling complexity for the double-barcoding. For the Hifi assembly experiment, we used 300 ng input backbone plasmids containing only the promoter barcode ($4.5X10^9$ total copies and on average $6.65X10^6$ copies of plasmids per promoter barcode). Given the effective labeling complexity, the average copy number of the plasmid containing the same promoter barcode-andom barcode pair is at most $2.08X10^3$. For the transfection experiment done in this study, with 2 µg ($6X10^9$ copy of plasmids) for cell mixing experiment and 10 µg ($3X10^{10}$ copy of plasmids) for K562 along experiment, the estimation of the average copy number for an identical plasmid is $4.4X10^2$ and $2.2X10^3$ respectively.

After obtaining the average copy number for identical plasmids, we estimate the probability of an identical plasmid being transfected into the same cell. We first define this probability as the collision rate. We note that the transfection of the identical copies of different plasmids are independent, so we could only calculate the collision rate for only one of such plasmids. The

calculation of the collision rate for a given library member can be formulated as such: given the number of the identical copies of a plasmid, what is the probability of two or more of the copies being transfected into the same cell? We first write the expectation:

$$n^{-m}\sum_{k=0}^{n}\binom{n}{k}\sum_{q=0}^{(n-k)}\binom{n-k}{q}\binom{m}{q}q!\left\{\begin{matrix} m-q \\ n-k-q \end{matrix}\right\}_{n\geq 2}(n-k-q)!(m-q)$$

where n denotes the total number of cells, m denotes the total number of identical plasmids, k denotes the number of cells with no plasmid, q denotes the cells with exactly 1 plasmid, parentheses denote binomial coefficient, and brackets denote partition function.

The above equation was simplified by substituting with the bivariate generating function, and the expected number is:

$$m(1-(\frac{n-1}{n})^{m-1})$$

For a given transfection experiment, we can estimate the effective percentage of plasmid that is successfully transfected into the cell. Given the estimated copy number for identical plasmids is $4.4X10^2$ and $2.2X10^3$ for mixed cell experiment and K562 alone experiment respectively, the expected number of cells having more than 1 identical plasmid can be calculated with the aforementioned equation, and the probability of two copies of an identical plasmid appearing in the same cell is 0.0004 and 0.002 respectively. On a practical note, researchers have suggested that the effective number of the plasmid that are incorporated into the nucleus is about 0.01 - 0.1 of the input amount[48], hence a library containing around $2.5X10^5$ different members transfected to 1 million cells has a theoretical collision rate around 1%.

Transfection

K562 cells were transfected using electroporation with the Neon transfection system (Invitrogen MPK5000). 1 million cells were transfected with 2 µg of plasmid DNA (mixed-cell experiment) or 10 µg of plasmid DNA (K562 sub-state experiment), with 3 pulses of 1450 V for 10 ms. The cells were then plated to pre-warmed K562 medium.

HEK293 cells were transfected using the Lipofectamine3000 protocol. 4 µl of p3000 reagent, 4µl of Lipofectamine, and OptiMEM were mixed with 2 µg of plasmid DNA to a volume of 250 µl. The lipofectamine reagents and plasmid were mixed and incubated at room temp for 15 minutes and then added dropwise to the cells.

Bulk RNA extraction and sequencing

We determined the optimal harvest time based on plasmid dilution and protein maturation and found the optimal harvest time is between 22 - 28 hours after transfection. The rationale behind the choice of time is to balance the transcription rate and the plasmid dilution during cell replication.

For both K562 cells and HEK293 cells, we harvested the cells after transfection at 24 hours, and proceeded to extract total mRNA with Qiagen RNeasy kit for K562 cells and Monarch Total RNA miniprep kit for HEK293 cells. The reverse transcription was done with Superscript IV Reverse Transcriptase (Invitrogen 18090010). The final sequencing library was constructed using a nested PCR strategy. Briefly, we first used Q5 (New England BioLabs M0515) polymerase to amplify the region containing the 2 barcodes with SCARED P17 and SCARED P18. The total reaction volume is 50µl with 50ng of backbone with 2.5 µl of 10uM primer each. The annealing temperature is 61°C with an extension time of 10s. 25 cycles of amplification was done. The product was then purified with the Monarch PCR&DNA Cleanup kit (New England BioLabs T1030L), and eluted with 20 µl of ddH2O. For the second PCR using primers SCARED P19 and SCARED p20, a 25 µl reaction was set up with 0.25 µl product from the previous step, the annealing temperature is 61°C, and the extension time is 10s, a total of 10 cycles was done. The PCR product was cleaned up using the Monarch PCR&DNA Cleanup kit (New England BioLabs T1030L). For the last PCR to add the P5 and P7 Illumina adapters, a reaction with 25 µl of total volume was set up with 2 µl of cleaned product from PCR2, a total of 10 cycles of PCR was done. The sequencing library was sequenced on an Illumina Mi-seq machine with other samples pooled in the same lane.

10X Experiment for scMPRA

We harvested both K562 and HEK293 cells 24 hours after transfection, then followed the cell preparation protocol of 10X genomics. We used the 10X V3.1 chromium kit for our single-cell RNA-seq protocol. All PCRs were performed on an Invitrogen PCR machine. We targeted 2000 cells per replicate for each experiment for the mixed cell experiment. We targeted 2500 cells per replicate for the K562 substrate experiment. We followed the 10X protocol (https://support.10xgenomics.com/single-cell-gene-expression/library-prep/doc/user-guide-chromium-single-cell-3-reagent-kits-user-guide-v31-chemistry) with 12 cycles of cDNA amplification. To amplify the Capture-Sequence captured reads. 0.25 µl of 100 uM SCARED P32 (5'-GTCAGATGTGTATAAGAGACAG-3') was added to the cDNA amplification mix. For step 2.2, we modify the clean-up protocol by saving both the beads and supernatents and for the supernatents, we use a final concentration of 1.2X beads to pull down the DNA fragments. We then take 25% of both the 0.6X and 1.2X pull down products for the next step of PCR. To construct the illumina sequencing library, we used a 3-step nested PCR strategy. Briefly, we first used Q5 (New England BioLabs M0515) polymerase to amplify the region containing the 2 barcodes with SCARED P17 and SCARED P18. We pooled 8 PCR reactions, each with 50 µl of total volume, with 10 cycles to reduce possible jackpotting. The annealing temperature is 61°C with an extension time of 10s. The product was then purified with the Monarch PCR&DNA Cleanup kit (New England BioLabs T1030L), and eluted with 20 µl of ddH2O. For the second PCR using the following 3 primers (SCARED P21: 5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGACGAGCTCTATAAGTAATCT-3', CAS PC2: 5'-CGAGATCTACACTCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-3', CAS PP2: 5'-ATCTACACTCTTTCCCTACACGACGCTCTTC-3'), we pulled 8 PCR reactions, each with 50 µl of total volume, with 10 cycles to reduce possible jackpotting , the annealing temperature is 61°C, and the extension time is 10s, a total of 10 cycles was done. The PCR product was cleaned up using the Monarch PCR&DNA Cleanup kit (New England BioLabs T1030L). For the

last PCR to add the P5 and P7 Illumina adapters (CAS P48: 5'-CAAGCAGAAGACGGCATACGAGATNNNNNNNN[index]GTGACTGGAGTTCAGAC-3', CAS PP4: 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACA-3', CAS PC4: 5'-AATGATACGGCGACCACCGAGATCTACACTCGTCG-3'), we pulled 8 PCR reactions, each with 50 µl of total volume, with 10 cycles to reduce possible jackpotting, a total of 10 cycles of PCR was done. The transcriptome is generated using the 10X Dual-Index Set TT expression kit (https://support.10xgenomics.com/single-cell-gene-expression/index/doc/technical-note-chromium-next-gem-single-cell-3-v31-dual-index-libraries).

The sequencing was done on an Illumina NextSeq machine. We used 40% of the barcode library, 40% of the balanced scRNA-seq transcriptome, and 20% Phi-X. Sequencing the constructed barcode library with transcriptome and Phi-X is crucial to reduce the sequencing error from the reporter constant sequence. On Read1, only 28 bps contains the 10X cell barcode and UMI wes amplified, to avoid sequencing the constant Poly(A) sequence; On Read2, 105 bps was sequenced. For the mixed experiment, we pool reads from a total of 2 runs of NextSeq High Throughput sequencing runs, and for the K562 cells, we pool 3 runs of NextSeq High Throughput runs.

scRNA-seq data processing

The single-cell RNAseq data were processed using Cellranger 6.0.1 (https://github.com/10XGenomics/cellranger) and Scanpy 1.8.1[49] (https://github.com/theislab/scanpy) following the standard pipeline. Briefly, different sequencing runs from the same biological replicate were pooled together and processed with CellRanger 6.1.1; the final output expression matrix was then imported into Scanpy for further normalization. We first removed cells with less than 1000 genes, and genes that were present in less than three cells. We then removed cells with high counts for mitochondrial genes. Next, we normalized the UMI counts to the total cell UMI counts. The normalized expression matrix was used for clustering and visualization with Scanpy. The clustering was done using the Leiden algorithm[50].

scMPRA data processing

The relevant script for processing single-cell MPRA reads can be found on a Github repository (https://github.com/szhao045/scMPRA). The final sequencing product for scMPRA with Read1 contains the cell and molecular information (cellBC and UMI), and Read2 contains the MPRA library information (cBC and rBC). First, we fuzzy-matched the constant sequences before and after both the promoter barcode and random barcode. In this step, we filtered out the reads without correct promoter barcode length, or random barcode length. To increase the speed, we wrote a stand-alone program (https://github.com/szhao045/scMPRA_parsingtools) written with Golang, and can be compiled to work on many operating systems. Second, we filtered out cell barcodes based on the cell barcode list from the CellRanger output barcode list, with error-correction with maximum hamming distance of 1. Third, to mitigate the effect of template-switching during the PCR steps, we plotted the rank read depth for each unique quad of 10X Cell Barcode, UMI, cBC, and rBC. We identified an elbow point with minimum depth of 1 (mixed

cell experiment) and 10 (K562 alone experiment), and kept any low-depth unique quad that contains the cBC-rBC pair at most hamming distance of 1 to a high depth pair. Lastly, we remove cells with less than 100 scMPRA-associated UMIs, since the scMPRA reads from those cells were poorly sampled.

Cell cycle analysis

Cell cycle analysis for the scRNA-seq experiment was done with Scanpy 1.8.1 with cell cycle genes[51]. The expression profile of each cell was projected onto a PCA plot based on the list of cell cycle genes using Scanpy.

Motif analysis

The core promoters were first clustered according to their expression levels in the different cell sub-state populations by hierarchical clustering. We categorized our data into up/down regulated clusters at the first branching point, aiming to preserve the large structure. We then identified core promoter motifs in each promoter according to the parameters in Zabidi et al[52]. using MAST v4.10.0[53] and plotted the proportion of promoters containing each motif in each promoter class.

Estimating intrinsic and extrinsic noise

Intrinsic and extrinsic noise were estimated using the statistical framework developed for the dual-reporter experiment[43]. We first extracted the pairwise expression level for cBCs that belong to the same promoter in every single cell. If more than two cBCs are found in the same cell, the pairwise expressions among them are recorded. We then removed promoters with less than 100 paired single-cell expression measurements (593 out of 676 promoters passed the filtering step). We then applied the statistical framework developed by Fu and Pachter[43]. The derivation is abbreviated and can be found in the original publication. Briefly, let C denote the expression for the first pBC in the cell and let Y denote the expression for the second pBC in the cell. Let η_{int} denote the intrinsic noise, and it can be calculated as:

$$\eta_{int} = \frac{1}{a}(\sum_{i=1}^{n} C_i Y_i - n\bar{C}\bar{Y}),$$

where

$$a = (n-1)(1+\frac{1}{n}) + \frac{1}{\rho^2}$$

$$\rho = \frac{Cov[C,Y]}{\sqrt{Var[C]}\sqrt{Var[Y]}}$$

where n denotes the number of cells.

Similarly, let η_{ext} denote the extrinsic noise, and it can be calculated as:

$$\eta_{ext} = \frac{1}{2a\bar{C}\bar{Y}}(\sum_{i=1}^{n}(C_i - Y_i)^2 - n(\bar{C} - \bar{Y})^2),$$

where

$$a = \frac{2n^3 - 7n + 6}{2(n^2 - n)} + \frac{2 - n}{n^2 - n}\frac{\rho}{1 - \rho} + \frac{1}{2(n^2 - n)}(\frac{\rho}{1 - \rho})^2$$

$$\rho = \frac{Cov[C, Y]}{\sqrt{Var[C]}\sqrt{Var[Y]}}$$

where n denotes the number of cells.

Statistical Analyses

All statistical analyses were done using Python 3.9.6, Numpy 1.12.1[54], Scipy 1.6.3 and R 4.0.2.

Data and Code Availability

Next-generation sequencing data that support the findings of the study are available in the Gene Expression Omnibus using accession code GSE188639.

The code that supports the findings of this study is available on Github Repository (https://github.com/szhao045/scMPRA).

Acknowledgements

We thank the members of the Cohen laboratory for their critical feedback on the manuscript. We thank Jess Hoistington-Lopez and MariaLynn Crosby for assistance with high-throughput sequencing. This work is supported by grants to B.A.C from the National Institutes of Health, R01 GM140711 and R01 GM092910.

Author Contributions

S.Z. and B.A.C. conceived and designed the project, S.Z. performed most of the experiments and analyses with significant technical contributions from C.K.Y.H. and D.M.G., S.Z. and B.A.C. wrote the manuscript with input and feedback from all authors.

Competing Interests

S.Z. and B.A.C. are inventors on a pending patent filed by Washington University in St. Louis which may encompass the methods, reagents, and data disclosed in this manuscript. B.A.C is on the scientific advisory board of Patch Biosciences.

References

1. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res.* **22**, 1748–1759 (2012).
2. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
3. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 9362–9367 (2009).
4. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
5. Vattikuti, S., Guo, J. & Chow, C. C. Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. *PLoS Genet.* **8**, e1002637 (2012).
6. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *Am. J. Hum. Genet.* **99**, 139–153 (2016).
7. Aygün, N. *et al.* Brain-trait-associated variants impact cell-type-specific gene regulation during neurogenesis. *Am. J. Hum. Genet.* **108**, 1647–1668 (2021).
8. Nott, A. *et al.* Brain cell type–specific enhancer–promoter interactome maps and disease-risk association. *Science* (2019).
9. Spielmann, M. & Mundlos, S. Looking beyond the genes: the role of non-coding variants in human disease. *Hum. Mol. Genet.* **25**, R157–R165 (2016).

10. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum. Mol. Genet.* **24**, R102–10 (2015).
11. Ong, C.-T. & Corces, V. G. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.* **12**, 283–293 (2011).
12. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
13. Kwasnieski, J. C., Mogno, I., Myers, C. A., Corbo, J. C. & Cohen, B. A. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 19498–19503 (2012).
14. Ireland, W. T. *et al.* Deciphering the regulatory genome of Escherichia coli, one hundred promoters at a time. *Elife* **9**, (2020).
15. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30**, 265–270 (2012).
16. Sharon, E. *et al.* Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* **30**, 521–530 (2012).
17. Kinney, J. B., Murugan, A., Callan, C. G., Jr & Cox, E. C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 9158–9163 (2010).
18. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271–277 (2012).
19. White, M. A. *et al.* A Simple Grammar Defines Activating and Repressing cis-Regulatory Elements in Photoreceptors. *Cell Rep.* **17**, 1247–1254 (2016).
20. Kwasnieski, J. C., Fiore, C., Chaudhari, H. G. & Cohen, B. A. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.* **24**, 1595–1602 (2014).
21. Chaudhari, H. G. & Cohen, B. A. Local sequence features that influence AP-1 cis-regulatory activity. *Genome Res.* **28**, 171–181 (2018).

22. Hughes, A. E. O., Myers, C. A. & Corbo, J. C. A massively parallel reporter assay reveals context-dependent activity of homeodomain binding sites in vivo. *Genome Res.* **28**, 1520–1531 (2018).
23. Tewhey, R. *et al.* Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165**, 1519–1529 (2016).
24. Hong, C. K. Y. & Cohen, B. A. Genomic environments scale the activities of diverse core promoters. *bioRxiv* 2021.03.08.434469 (2021) doi:10.1101/2021.03.08.434469.
25. Haberle, V. *et al.* Transcriptional cofactors display specificity for distinct types of core promoters. *Nature* **570**, 122–126 (2019).
26. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
27. Amabile, G. *et al.* Dissecting the role of aberrant DNA methylation in human leukaemia. *Nat. Commun.* **6**, 7091 (2015).
28. Juven-Gershon, T., Cheng, S. & Kadonaga, J. T. Rational design of a super core promoter that enhances gene expression. *Nat. Methods* **3**, 917–922 (2006).
29. Shao, H. *et al.* Core promoter binding by histone-like TAF complexes. *Mol. Cell. Biol.* **25**, 206–219 (2005).
30. Shaffer, S. M. *et al.* Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* **546**, 431–435 (2017).
31. Moudgil, A. *et al.* Self-Reporting Transposons Enable Simultaneous Readout of Gene Expression and Transcription Factor Binding in Single Cells. *Cell* **182**, 992–1008.e21 (2020).
32. Litzenburger, U. M. *et al.* Single-cell epigenomic variability reveals functional cancer heterogeneity. *Genome Biol.* **18**, 15 (2017).
33. Min, M. & Spencer, S. L. Spontaneously slow-cycling subpopulations of human cells originate from activation of stress-response pathways. *PLoS Biol.* **17**, e3000178 (2019).

624 34. Bonnet, D. & Dick, J. E. Human acute myeloid leukemia is organized as a hierarchy that
625 originates from a primitive hematopoietic cell. *Nat. Med.* **3**, 730–737 (1997).
626 35. Ishikawa, F. *et al.* Chemotherapy-resistant human AML stem cells home to and engraft
627 within the bone-marrow endosteal region. *Nat. Biotechnol.* **25**, 1315–1321 (2007).
628 36. Chang, H. H., Hemberg, M., Barahona, M., Ingber, D. E. & Huang, S. Transcriptome-wide
629 noise controls lineage choice in mammalian progenitor cells. *Nature* **453**, 544–547 (2008).
630 37. Emert, B. L. *et al.* Variability within rare cell states enables multiple paths toward drug
631 resistance. *Nat. Biotechnol.* (2021) doi:10.1038/s41587-021-00837-3.
632 38. Foreman, R. & Wollman, R. Mammalian gene expression variability is explained by
633 underlying cell state. *Mol. Syst. Biol.* **16**, e9146 (2020).
634 39. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a
635 single cell. *Science* **297**, 1183–1186 (2002).
636 40. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA synthesis
637 in mammalian cells. *PLoS Biol.* **4**, e309 (2006).
638 41. Hilfinger, A. & Paulsson, J. Separating intrinsic from extrinsic fluctuations in dynamic
639 biological systems. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 12167–12172 (2011).
640 42. Sherman, M. S., Lorenz, K., Lanier, M. H. & Cohen, B. A. Cell-to-cell variability in the
641 propensity to transcribe explains correlated fluctuations in gene expression. *Cell Syst* **1**,
642 315–325 (2015).
643 43. Fu, A. Q. & Pachter, L. Estimating intrinsic and extrinsic noise from single-cell gene
644 expression measurements. *Stat. Appl. Genet. Mol. Biol.* **15**, 447–471 (2016).
645 44. Chan, Y. K. *et al.* Engineering adeno-associated viral vectors to evade innate immune and
646 inflammatory responses. *Sci. Transl. Med.* **13**, (2021).
647 45. Byrne, L. C. *et al.* In vivo-directed evolution of adeno-associated virus in the primate retina.
648 *JCI Insight* **5**, (2020).
649 46. Wang, D., Tai, P. W. L. & Gao, G. Adeno-associated virus vector as a platform for gene

therapy delivery. *Nat. Rev. Drug Discov.* **18**, 358–378 (2019).

47. Bryant, D. H. *et al.* Deep diversification of an AAV capsid protein by machine learning. *Nat. Biotechnol.* **39**, 691–696 (2021).

48. Cohen, R. N., van der Aa, M. A. E. M., Macaraeg, N., Lee, A. P. & Szoka, F. C., Jr. Quantification of plasmid DNA copies in the nucleus after lipoplex and polyplex transfection. *J. Control. Release* **135**, 166–174 (2009).

49. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

50. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).

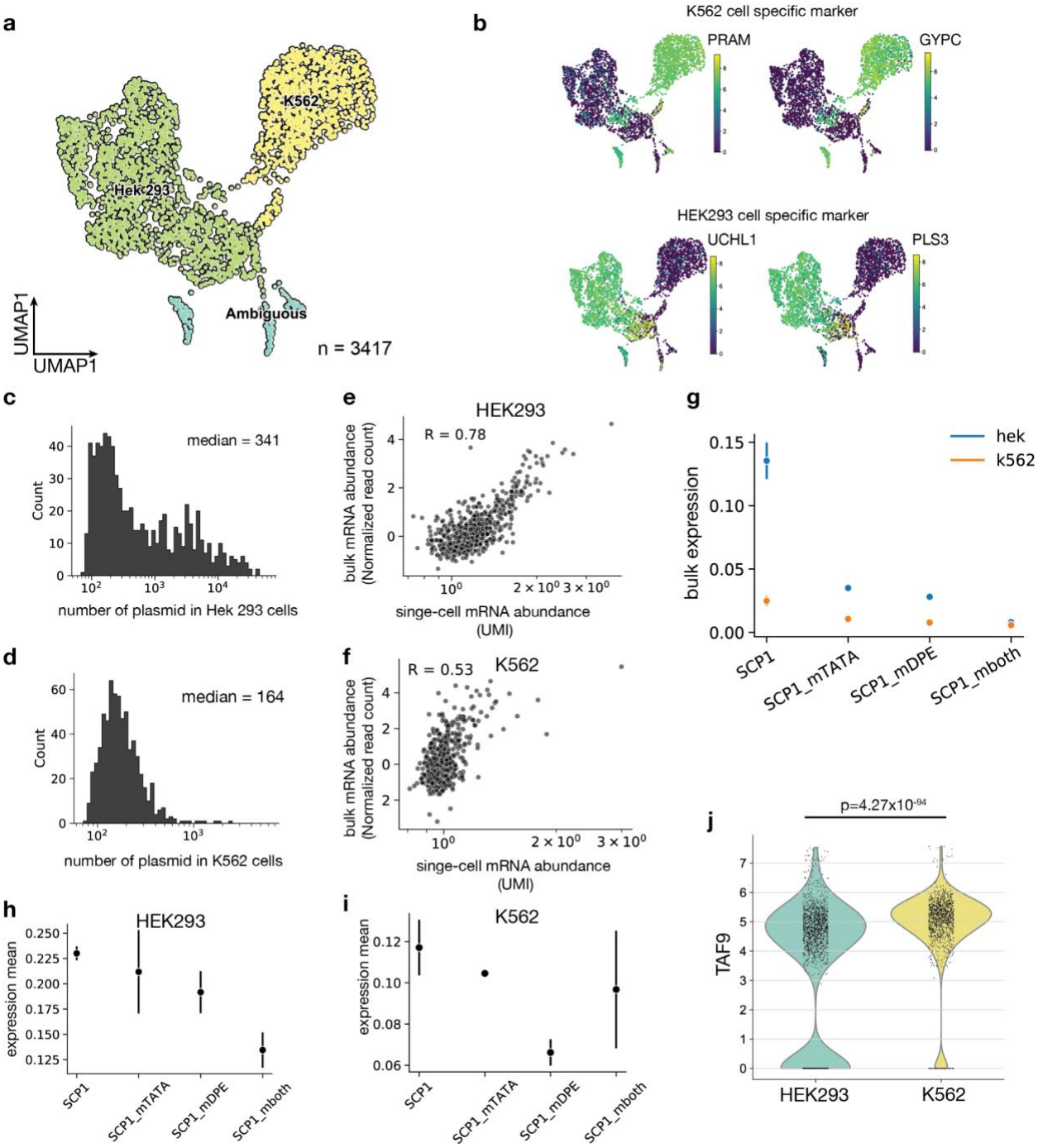
51. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).

52. Zabidi, M. A. *et al.* Enhancer–core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**, 556–559 (2014).

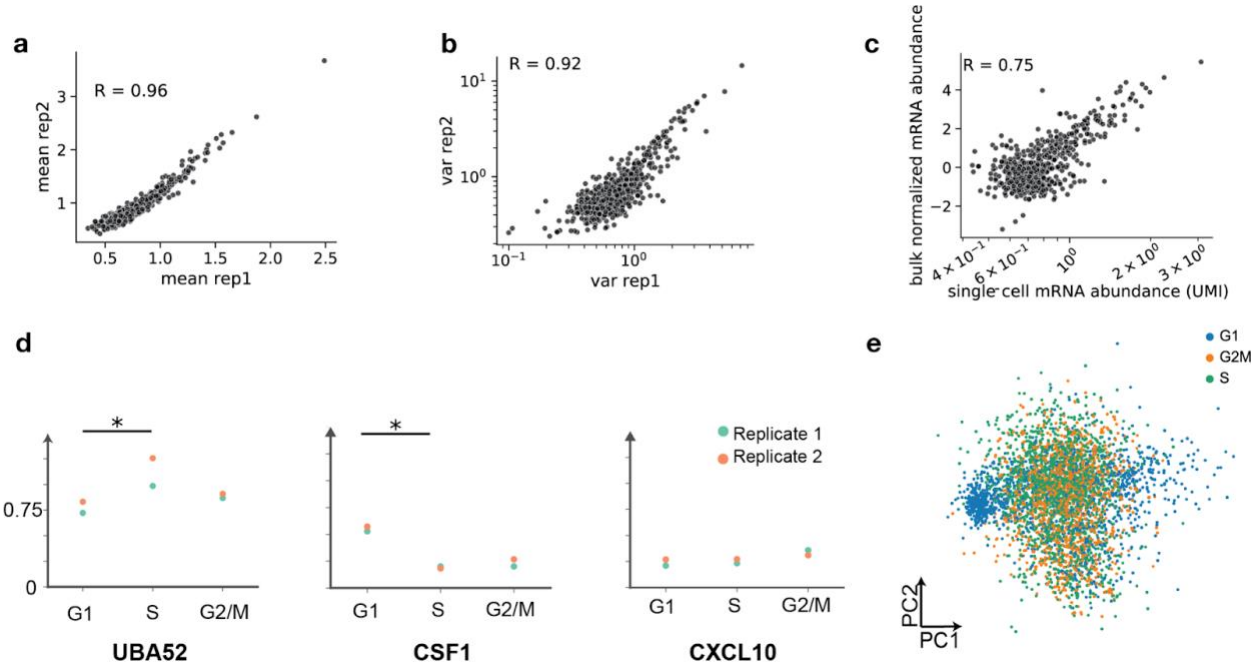
53. Bailey, T. L. & Gribskov, M. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **14**, 48–54 (1998).

54. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).

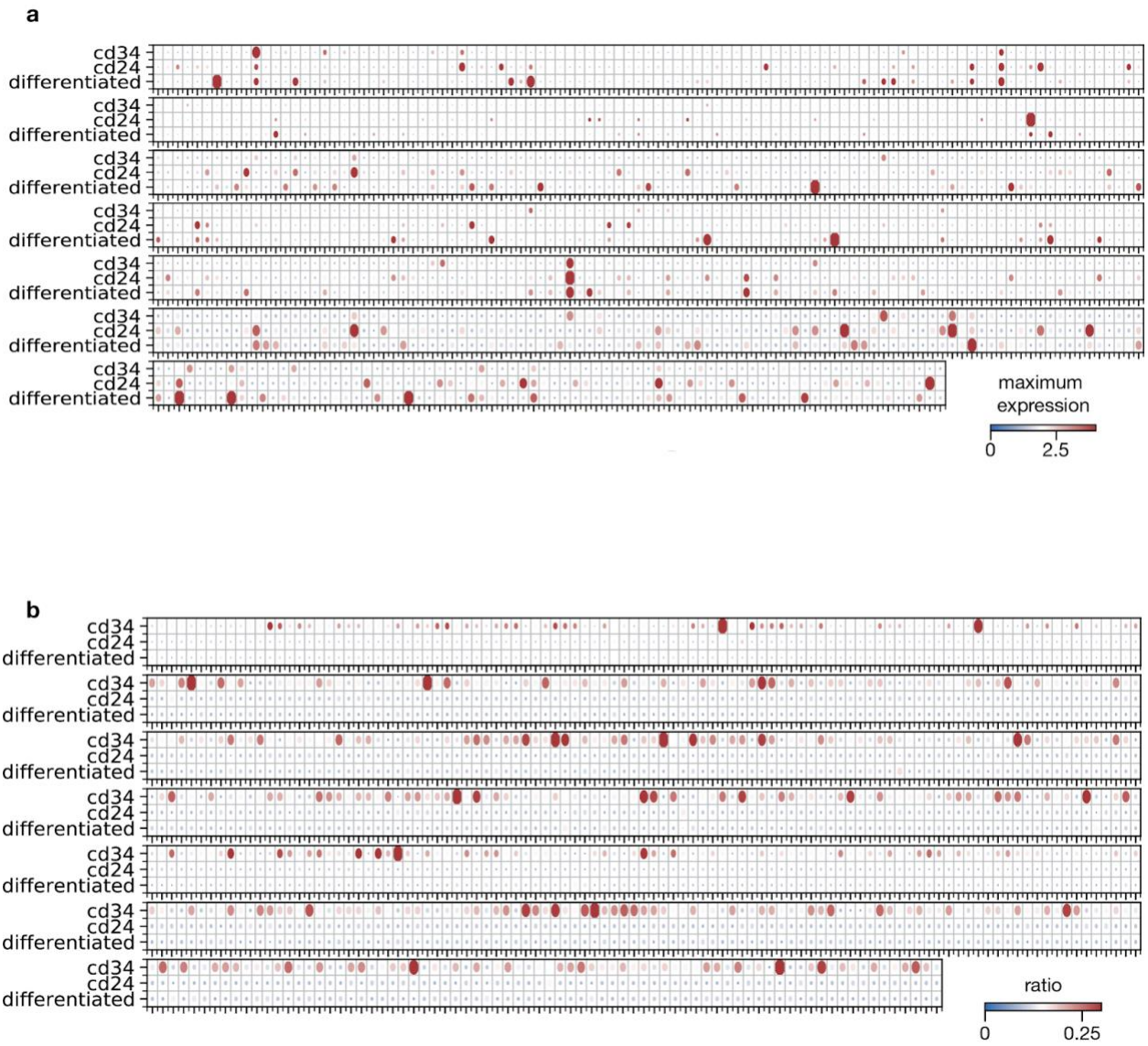
Supplementary Figures



Supplementary Figure 1. scMPRA measures cell-type specific CRS activity. (**a**) UMAP of the single-cell transcriptome from the mixed-cell experiment. 105 out of 3417 cells (3%) are labeled by both K562 and HEK293 cell genes. (**b**) UMAP of the mixed-cell experiment with cells marked by other representative markers for K562 and HEK293 cell expression. (**c-d**) Histogram of the number of plasmids transfected to K562 cells and HEK293 cells. (**e,f**) Scatterplot of bulk RNA-seq expression against expression mean from scMPRA (Pearson R for K562 cells: 0.53, Pearson R for HEK293 cells: 0.78). (**g**) Dot plot of the reporter activity of SCP1 and its mutants from bulk RNA-seq data (error bar: 1 s.d.). (**h**) Dot plot of the mean reporter activity of SCP1 and its mutants from scMPRA experiment for K562 cells. (**i**) Dot plot of the mean reporter activity of SCP1 and its mutants from scMPRA experiment for HEK293 cells.(**j**) Violin plot showing the expression distribution of TAF9 in K562 and HEK293 cells. (Wilcoxon rank sum test, $p = 4.27 \times 10^{-94}$).



Supplementary Figure 2 scMPRA measures CRS activity in K562 cell substates. (**a,b**) Reproducibility for expression mean and cell-to-cell variance (Pearson Correlation for mean: 0.96, for variance: 0.92). **(c)** Scatterplot of reproducibility of scMPRA mean expression with bulk MPRA measurement using UMI (Pearson Correlation: 0.75). **(d)** Different dynamics of expression. For UBA52, the promoter is most highly expressed in S phase; whereas for CSF1, the promoter is most highly expressed in G1 phase. For CXCL10, the promoter is expressed evenly through cell cycle (Stars indicate significance from Wilcoxon rank sum test, *: $p < 0.05$.) **(e)** Cells no longer cluster together based on cell cycle genes after normalization.



Supplementary Figure 3. CD34^{+}/CD38^{-} substate changes the probability of cells having higher expression, not the maximum expression level. (a) Dot plot showing the maximum single-cell expression for the core promoter library in CD34+/CD38-, CD24+, and Differentiated clusters. Color and size both indicate the maximum expression change. (**b**) Dot plot showing the percentage of cells in CD34+/CD38-, CD24+, and Differentiated clusters that are in the 90th percentile of expression level per promoter. Color and size both indicate the ratio change.