nature biotechnology

Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions

Jason Ernst[1–5], Alexandre Melnikov[6], Xiaolan Zhang[6], Li Wang[6], Peter Rogov[6], Tarjei S Mikkelsen[6] & Manolis Kellis[6,7]

Massively parallel reporter assays (MPRAs) enable nucleotide-resolution dissection of transcriptional regulatory regions, such as enhancers, but only few regions at a time. Here we present a combined experimental and computational approach, Systematic high-resolution activation and repression profiling with reporter tiling using MPRA (Sharpr-MPRA), that allows high-resolution analysis of thousands of regions simultaneously. Sharpr-MPRA combines dense tiling of overlapping MPRA constructs with a probabilistic graphical model to recognize functional regulatory nucleotides, and to distinguish activating and repressive nucleotides, using their inferred contribution to reporter gene expression. We used Sharpr-MPRA to test 4.6 million nucleotides spanning 15,000 putative regulatory regions tiled at 5-nucleotide resolution in two human cell types. Our results recovered known cell-type-specific regulatory motifs and evolutionarily conserved nucleotides, and distinguished known activating and repressive motifs. Our results also showed that endogenous chromatin state and DNA accessibility are both predictive of regulatory function in reporter assays, identified retroviral elements with activating roles, and uncovered 'attenuator' motifs with repressive roles in active chromatin.

Epigenome maps predict thousands of putative regulatory regions through their *in vivo* epigenomic signatures and are widely used for studying gene regulation and disease[1–10]. However, such maps present only indirect evidence of regulatory function, have often limited resolution and do not distinguish activator from repressor elements[4,5,7]. DNA motif and sequence pattern analysis can complement epigenome maps, but also provides only indirect evidence and only identifies sequences that match enriched patterns[4,5,7,11,12].

Episomal reporter assays[3,6] and endogenous modulation[10,13] are two complementary approaches to characterize putative regulatory regions. Episomal reporters evaluate sequence function directly, independently of epigenetic effects, whereas endogenous perturbations capture endogenous context effects. Multiplexed endogenous or episomal assays have been used to dissect few regulatory regions at high resolution[14–22] or many at low resolution[20,22–28].

MPRAs[15,24] synthesize DNA sequences on programmable microarrays and integrate them in reporter gene plasmids that are then transfected into cell types of interest. Barcodes placed in reporter gene 3′ untranslated regions (UTRs) (to minimize their effect on pre-transcriptional control) provide a quantitative readout of gene expression. The limited number of array spots constrains the number of regions tested and the number of reporter constructs devoted to each region. Owing to the short length of synthesized fragments (~145 nucleotides), MPRAs require accurate knowledge of putative regulatory region position and boundaries, which are not generally known.

Here we overcome these limitations using dense tiling of MPRA constructs and computational analysis to infer activating and repressive nucleotides at high resolution across many regions. We termed the combined approach Sharpr-MPRA and the associated computational method SHARPR. We used Sharpr-MPRA to dissect over 15,000 putative regulatory regions from genome-wide epigenomic maps. We tiled each 295-base-pair (bp) region at 5-nucleotide offsets using overlapping 145-nucleotide constructs. We made 4.6 million nucleotide inferences, each in two cell types, and distinguished activating and repressive regulatory functions without the use of motifs or other sequence information. Inferred regulatory nucleotides were reproducible, high-resolution, cell-type-specific and supported by evolutionary conservation and regulatory motif evidence. Our strategy enabled gene-regulatory insights, including activating motifs lacking well-established regulators, 'dual-role' motifs with both activating and repressive roles, strongly activating repeat elements and 'attenuator' motifs that have repressive roles in active chromatin states.

RESULTS

Pilot design tiling 250 regions at 30-bp resolution

We first developed a low-resolution 'pilot' design, applied to 250 regions showing H3K27ac-marked enhancer chromatin states[3] (200 in liver carcinoma HepG2 cells and 50 in leukemia K562 cells). We tiled 385-nucleotide regions at 30-nucleotide offsets using 145-nucleotide constructs, and tested each unique sequence using 24 barcodes (**Fig. 1a** and **Supplementary Fig. 1**). We centered our tiling on H3K27ac

[1]Department of Biological Chemistry, University of California, Los Angeles, Los Angeles, California, USA. [2]Computer Science Department, University of California, Los Angeles, Los Angeles, California, USA. [3]Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research at University of California, Los Angeles, Los Angeles, California, USA. [4]Jonsson Comprehensive Cancer Center, University of California, Los Angeles, Los Angeles, California, USA. [5]Molecular Biology Institute, University of California, Los Angeles, Los Angeles, California, USA. [6]Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts, USA. [7]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. Correspondence should be addressed to J.E. (jason.ernst@ucla.edu) or M.K. (manoli@mit.edu).

Received 14 October 2015; accepted 16 August 2016; published online 3 October 2016; doi:10.1038/nbt.3678

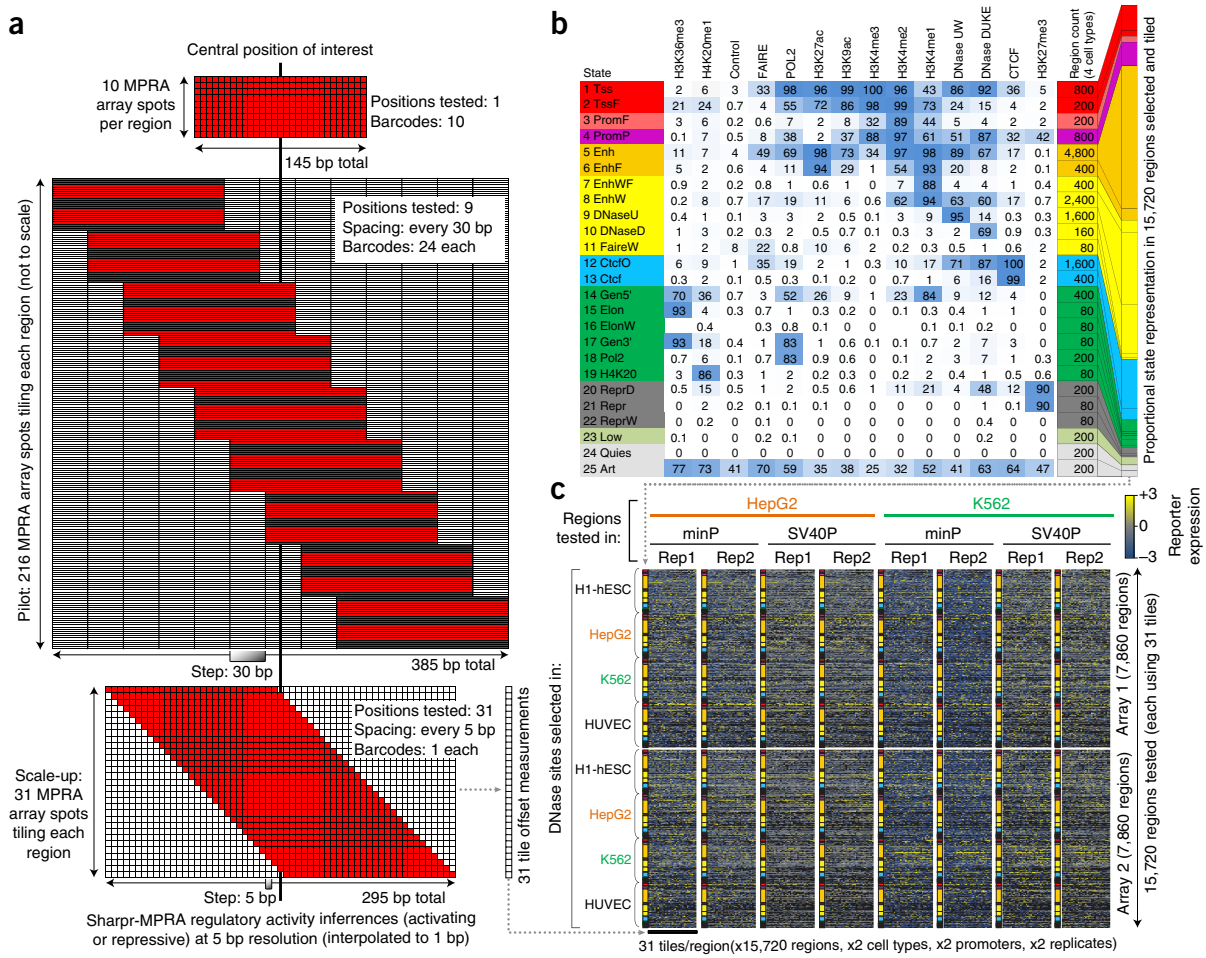

Figure 1 Experimental design. (**a**) Comparison of MPRA strategies for testing regulatory regions. Non-tiling approaches (top; for example, ref. 24) use multiple barcodes for the same tested sequence. Our pilot design (middle) tests each region using nine tile offsets, spaced at 30-bp increments, each tested using 24 barcodes (216 MPRA array spots per region). Our scaled-up design (bottom), tests each region using 31 tile offsets spaced at 5-bp increments, each tested using a single barcode per tile offset. The designs are to scale along the horizontal dimension. Only top and bottom are to scale in the vertical dimension. (**b**) The 25 chromatin states used in selecting regulatory regions for testing in the scale-up design[29,30] (**Supplementary Fig. 5**). Heatmap indicates the emission probabilities (scaled between 0 and 100) for each epigenomic feature (columns) in each chromatin state (rows). Tested regions were restricted to DNase sites in one of four cell types, with the number of regions selected based on a stratified random sampling as indicated. (**c**) Overview of experiments using the scale-up design (see **Supplementary Fig. 1** for the pilot design). We carried out 16 experiments, consisting of two sets of 7,860 regions (row groups) across 25 chromatin states (colors), with 31 tiles per region (individual columns), each tested in both HepG2 and K562 cells, each using both a minimal promoter (minP) and an SV40 promoter (SV40P), each in two replicates (Rep1 and Rep2). Heatmap shows MPRA reporter gene expression measurements (blue = low, yellow = high, black = missing) (**Supplementary Data 3**).

signal dips known to be indicative of nucleosome displacement owing to transcription factor (TF) binding, and thus likely to overlap regulatory nucleotides. For HepG2 cells, we selected 100 regions with strong dip scores, and 100 with wide-ranging dip scores (**Supplementary Fig. 1** and Online Methods). We profiled each region in both K562 and HepG2 cells, each in two replicates (**Supplementary Data 1** and **2**).

Among the nine tile positions, inner tiles (centered on H3K27ac dips) showed the highest level and frequency of activity (**Fig. 2a** and **Supplementary Fig. 2a–c**), and the highest variability across regions (**Supplementary Fig. 2d**), indicative of regulatory nucleotides. Regions with stronger H3K27ac dip scores showed higher activity and the cell-type specificity of epigenomic signals matched the cell-type specificity of reporter expression (**Fig. 2a** and **Supplementary Figs. 1** and **2**), suggesting that endogenous epigenomic information is indicative of reporter assay activity.

Biological replicates of the same tile showed reproducible median reporter activity (Pearson's correlation coefficient of 0.92 across replicates for HepG2 cells), but reporter activity for tiles offset by 30 bp sometimes differed substantially (Pearson's correlation = 0.57 in the same HepG2 cell experiment; **Fig. 2b**), with tiles separated by greater distances showing greater differences in reporter activity (**Supplementary Fig. 3a**). K562 cells showed similar results (Pearson's correlation of 0.66 vs. 0.34), with the lower correlation likely reflecting reduced transfection rates for K562 cells using our experimental protocols, as previously reported[24]. To gain insights into the sequences driving these differences, we focused on 637 pairs of neighboring positions in HepG2 cells and 142 pairs in K562 cells that showed significant differences in activity (false discovery rate 5%; **Supplementary Table 1** and **Supplementary Fig. 3b,c**), and searched for motif differences in sequence segments distinguishing consecutive tiles (**Fig. 2c,d**, **Supplementary Fig. 4a** and Online Methods). Segments with increased HepG2 cell activity were enriched for known liver-function motifs, including HNF4 and HNF1, whereas segments with increased K562 cell activity were enriched for known hematopoietic motifs, including GATA

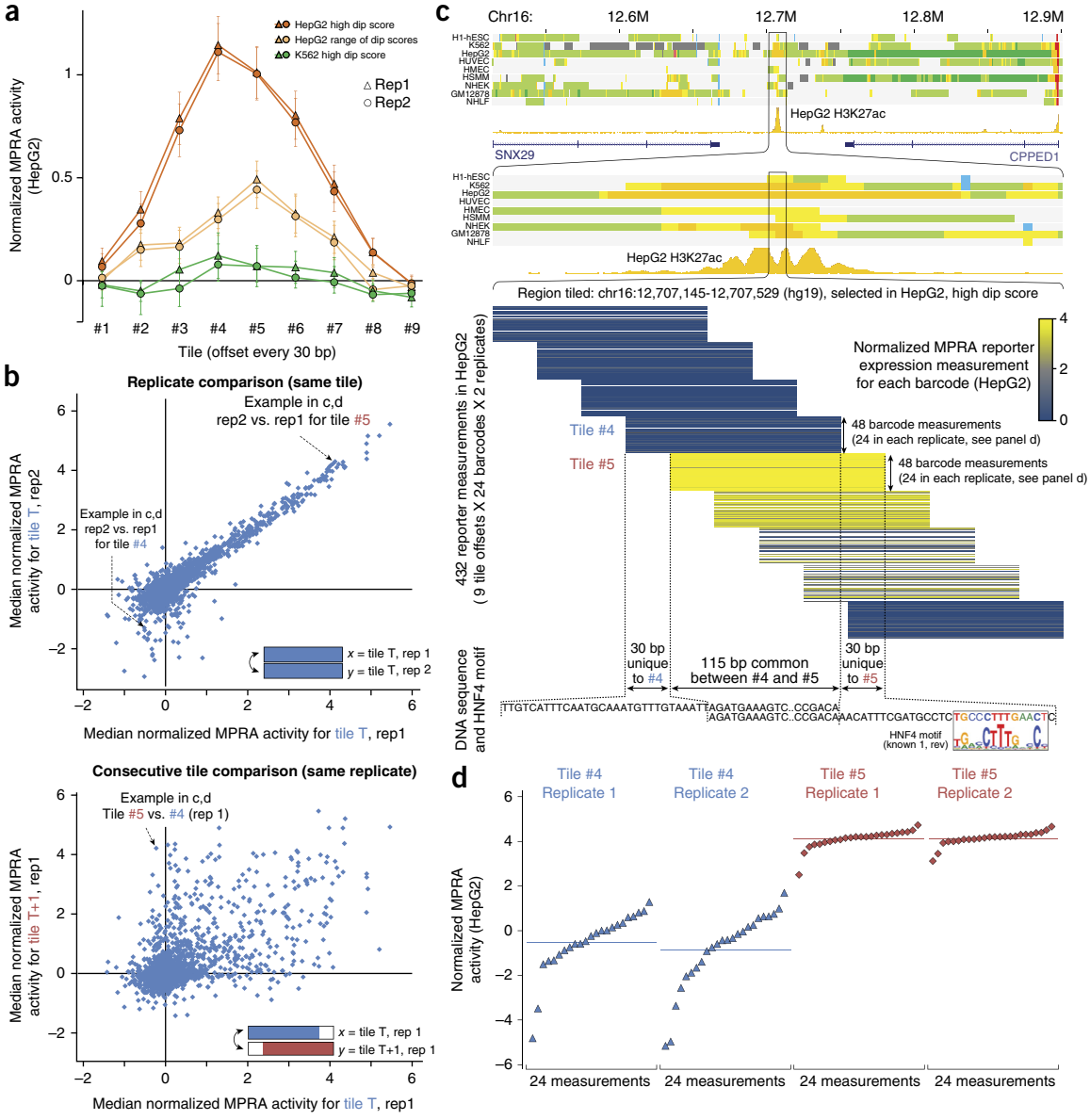

Figure 2 Tiling enhancer regions in pilot design revealed regulatory segments at 30-bp resolution. (**a**) Effect of tile offset and H3K27ac dip score on reporter expression. Average HepG2 cell reporter expression (*y* axis) at each of nine offsets (*x* axis) for three sets of regions: HepG2 cell candidate enhancers[3] with the highest H3K27ac dip scores (orange), candidate enhancers with a range of dip scores (light orange), and regions that were not predicted enhancers in HepG2 cells but were predicted enhancers with a high dip score in K562 (green) cells. Error bars, s.e.m. (n = 100, 100 and 50, respectively). (**b**) Consecutive tiles can differ in reporter expression. Comparison of median reporter activity between biological replicates in HepG2 cells (top; only first eight tile offsets are shown). Comparison of consecutive tiles T (*x* axis) and T + 1 (*y* axis) for the same biological replicate (rep1) (bottom). (**c**) Chromatin state annotations[3] in nine cell types and H3K27ac signal track in HepG2 cells over about 400 kb and 10 kb surrounding the tiled 385-bp region centered in the H3K27ac dip (top). Expanded view of tile reporter measurements (yellow blue color) across all nine tiles, 24 barcodes, and two replicates (middle). Tiles #4 and #5 share 115 bp in common (abbreviated), and have 30 bp unique to #4 or #5 (shown), indicating the potential presence of activating elements in the sequence unique to #5 and/or repressive elements in the sequence unique to #4 (bottom). Indeed, the 30-bp segment unique to #5 contains a candidate binding site for HNF4, a known activator of liver-related functions. (**d**) Expanded view of expression activity measurements for consecutive tiles #4 and #5 for all individual barcodes (points), sorted by their reporter expression levels. For replicate 1 of tile #4, 1 of 24 barcode measurements failed. The *y*-axis coordinates correspond to the ones shown in **b**. Horizontal lines indicate median normalized MPRA activity. See **Supplementary Figures 2–4** for additional results from the pilot design.

(**Supplementary Fig. 4b**). This confirmed that a tiling approach can reveal nucleotides important for cell-type-specific regulatory function.

Scale-up design tiling 15,720 regions at 5-bp resolution

We next scaled up our MPRA tiling design, increasing resolution, throughput, coverage and chromatin-state diversity. To achieve these goals, we made several modifications. (i) To achieve increased resolution, we positioned reporter sequences at 5-bp increments instead of 30 bp (**Fig. 1a**). (ii) To increase throughput, we used a single reporter construct per position instead of 24 (**Fig. 1a**), achieving robustness by exploiting the many reporter constructs overlapping most positions (15 on average; 25 for the central 105 nucleotides). We also tested smaller 295-bp regions instead of 385-bp regions, focusing on the most informative positions based on our pilot results (**Fig. 2a** and

Supplementary Fig. 2). (iii) To increase coverage, we used two 244K arrays for DNA synthesis (instead of a single 55K array), targeting 15,720 regions for tiling, each profiled in both HepG2 and K562 cells, using both a minimal TATA promoter (minP) and a strong SV40 promoter (SV40P), each in two replicates (**Fig. 1c**), resulting in a total of 3.9 million measurements (**Supplementary Data 3**). (iv) To enable analyses across diverse chromatin states of a 25-state ChromHMM model[29,30] (**Supplementary Fig. 5**), we centered regions on double-cut DNase I hypersensitive sites instead of H3K27ac dips. To ensure representation of all states, we used a tiered random sampling approach, which also favored enhancers and other DNase-hypersensitivity-enriched states (**Fig. 1b**). To include both active and inactive regulatory sites in the cell types profiled, we selected DNase sites from HepG2 and K562 cells, and two additional cell types, human umbilical vein endothelial cells (HUVECs) and human embryonic stem cells (H1-hESC) (**Fig. 1c**).

These design choices, and the SHARPR computational inference method we describe next, allowed us to infer regulatory activity at ~5-nucleotide resolution across 4.6 million nucleotides spanning over 15,000 regions, a sixfold increase in resolution and 60-fold increase in coverage compared to our pilot design.

Inference of activating and repressive nucleotides

We developed a computational method, SHARPR, that scores the relative activating or repressive potential of each 5-bp interval in tiled regions and interpolates these values to make predictions for individual nucleotides (**Fig. 3a**,**b**). The inclusion and exclusion of 5-bp nucleotide intervals between consecutive tiles is akin to perturbation experiments, allowing inferences at substantially higher resolution (5 bp) than with the original reporter constructs (145 bp) (**Fig. 3a**). We reasoned that activating intervals (for example, containing activator motifs) should increase the reporter expression for tiles overlapping them, whereas repressive intervals (for example, containing repressor motifs) should decrease reporter expression of overlapping tiles, as we showed in our pilot experiments (for example, **Fig. 2c**). Thus, modeling the relative activity of overlapping tiles should enable inference of activating and repressive nucleotide positions at high resolution (**Fig. 3b**).

We constructed a probabilistic graphical model (**Fig. 3a**) relating the unobserved regulatory activity of each 5-bp interval (hidden variables A_1–A_{59}) to the 145-bp reporter measurements (observed variables M_1–M_{31}). We modeled M_j using a normal distribution with mean the average of overlapping A_k and variance the empirical variance of all measurements in the experiment (Online Methods and **Supplementary Note 1**). We modeled A_k using a normal distribution with mean calculated as the average of all measurements in the experiment and variance σ_a^2 as a free parameter (smaller values resulting in more smoothed inferences). We used a low-variance prior and a high-variance prior, and combined those results (Online Methods and **Supplementary Note 2**). We inferred the 'most likely' values for the regulatory activity variables based on observed reporter measurements and their prior distributions, and standardized these using a z score to combine results from multiple replicates, promoters and variance settings (**Supplementary Fig. 6**). We carried out piecewise linear interpolation from the 5-bp activity estimates to infer the regulatory activity of each nucleotide in the tiled regions (Online Methods and **Supplementary Note 3**). We implemented the computational portion of Sharpr-MPRA in a software package for which we provide a public software release (http://www.biolchem.ucla.edu/labs/ernst/SHARPR/; **Supplementary Source Code**).

We used Sharpr-MPRA to make activating or repressive regulatory activity inferences for 4.6 million nucleotides, each in two cell types, each using two promoter types, each using two replicates. We inferred nucleotide activity for both minP and SV40P individually (combining two replicate experiments for each promoter), and for their combination (combinedP, using four experiments jointly), resulting in three activity tracks for each cell type (**Supplementary Data 4**, and **Supplementary Figs. 6b** and **7a**). We also assigned a minP, SV40P and combinedP score to each region (**Supplementary Fig. 7b**,**c**), using the signed (activating or repressive) score of the maximum absolute score position (MaxPos) (**Fig. 3b**). We provide visualizations showing all minP, SV40P and combinedP inferences for HepG2 and K562 cells in 31,440 figures at http://www.biolchem.ucla.edu/labs/ernst/SHARPR/, and for several selected subsets (**Supplementary Data 5–8**).

Reproducibility of Sharpr-MPRA regulatory activity scores

We evaluated the reproducibility of Sharpr-MPRA scores in multiple ways. We first evaluated the agreement between minP and SV40P inferences for each nucleotide position across regions. The central 101 positions showed on average 0.75 Pearson's correlation coefficient for HepG2 cells and 0.66 for K562 cells, which decreased toward outer positions (**Supplementary Fig. 8a**), attributable to fewer tiles overlapping outer positions and stronger regulatory activity closer to DNase site centers. Individual nucleotides maintained similar scores between promoter types (**Supplementary Fig. 9a**), with 83% of scores ≥ 1.5 for one promoter showing scores ≥ 1 for the other in HepG2 cells (71% for K562 cells), and 74% of scores < −0.5 for one promoter showing scores < 0 for the other in HepG2 cells (73% for K562 cells).

Individual replicates of each promoter type showed strong agreement for increasing absolute scores (**Supplementary Fig. 10a**) for both cell types and both promoter types (for example, Pearson's correlation of 0.7 on average for regions with |score| ≥ 2, and 0.9 for |score| ≥ 3 for minP). Between promoter types, regions with |score| ≥ 2 showed 0.8 correlation on average in HepG2 cells (0.7 in K562 cells), compared to < 0.1 expected by chance (**Supplementary Fig. 10b**). The MaxPos nucleotide position showed substantially greater concordance between replicates and between promoter types than expected by chance (**Supplementary Fig. 11a–c**), with the distance decreasing with increasing absolute score. Between minP replicates in HepG2 cells, MaxPos nucleotides were on average within 28 bp, 11 bp and 5 bp for |score| ≥ 2, 4 and 6, respectively, for HepG2 cells (26 bp, 13 bp and 7 bp, respectively, for K562 cells), compared to ~60 bp expected by chance when sampling MaxPos nucleotide positions.

We repeated these comparisons across different sets of barcodes using DNase sites that we independently selected in multiple cell types, and thus tested using multiple barcode sets (**Supplementary Data 5**). Forty-four regions overlapped completely and 212 overlapped partially, providing a resource for quantifying potential barcode effects. The position-specific correlation in Sharpr-MPRA scores between SV40P and minP showed a negligible change in HepG2 cells and a modest reduction in K562 cells when using the same vs. a different set of barcodes (0.72 vs. 0.71 and 0.69 vs. 0.60, respectively, for the central 101 nucleotide positions on average; **Supplementary Fig. 8b**), indicating a modest barcode effect relative to other sources of biological and technical variability. Individual nucleotide scores showed strong agreement across barcode sets for regions of both partial overlap and exact overlap (**Supplementary Fig. 9c**,**d**), with 90% of combinedP scores ≥ 1.5 in one showing combinedP score ≥ 1 in the other for HepG2 cells (80% in K562 cells) across all 256 multiply tiled regions. The 44 complete-overlap regions showed high score correlation

across barcode sets for regions with high absolute scores (0.96 for HepG2 cells and 0.77 for K562 cells for |score| ≥ 2; **Supplementary Fig. 10c**). The position of maximum absolute score was within 16 bp between barcode sets for HepG2 cells (21 bp for K562 cells) for |score| ≥ 2 (**Supplementary Fig. 11d**). A search for *k*-mer effects in the barcodes (Online Methods) revealed no influence on inferred

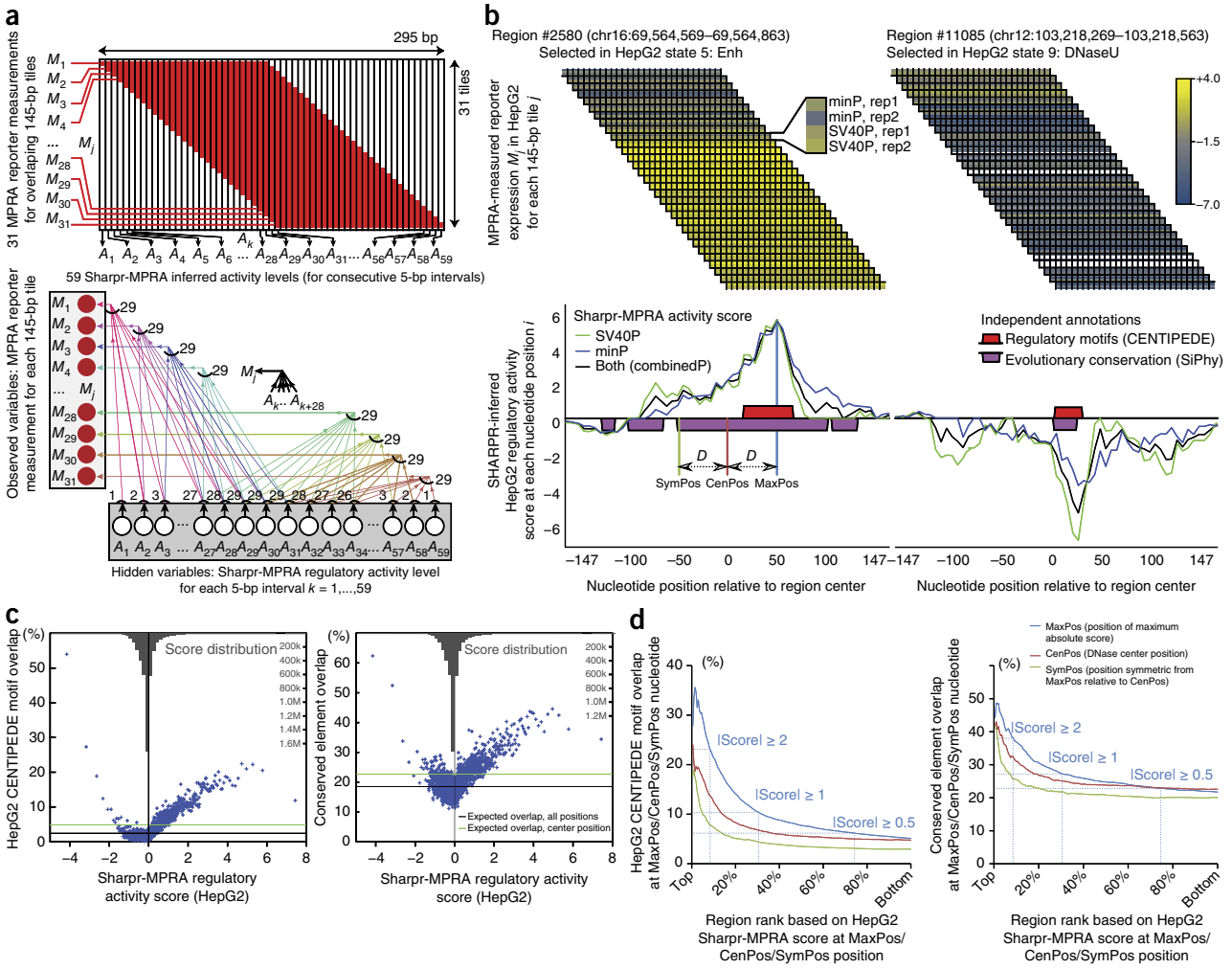

Figure 3 Scale-up design permits dissection of regulatory regions at high resolution. (**a**) Modeling scheme and probabilistic graphical model for the scale-up design. Variables M_1, ..., M_{31} represent the observed values of the reporter measurements for the 31 tiles (each 145 bp long), and variables A_1, ..., A_{59} represent the unobserved regulatory activity level of each 5-bp interval of the 295 bp covered, which is then normalized into the Sharpr-MPRA regulatory activity score. Probabilistic graphical model (bottom) used for high-resolution inference of activating and repressive intervals, with arrows $A_k \rightarrow M_j$ illustrating the dependencies between variables when tile M_j overlaps interval A_k, and the direction of information flow in the generative model. Conditional inference allows us to use the observed reporter measurements M_1, ..., M_{31} for the 31 tiles to infer the unobserved activity levels A_1, ..., A_{59} for the 59 intervals of length 5 bp each, which we interpolated to each nucleotide position i, under the modeling assumptions specified in Online Methods. (**b**) Observed reporter expression measurements for 145-bp segments (top) and inferred regulatory activity for 5-bp segments, interpolated to individual nucleotides (bottom) for two 295-bp regulatory regions in HepG2 cells. At each offset, the four rows correspond to four measurements of the same tile, using minP and SV40P, each in two replicates (top). Measurements for each tile are shown spanning all nucleotide positions the tile covers. White rows represent missing data for a promoter/replicate combination for a given 145-bp tile. Resulting inference of regulatory activity at each nucleotide i using all four measurements (black), only the two SV40P measurements (green), or only the two minP measurements (blue) (bottom). Predicted positions of highest activating (positive scores) or repressive (negative scores) activity capture CENTIPEDE[5] predicted binding sites (red boxes) and conserved elements identified by the SiPhy-PI method[33] (purple boxes), even though such information was not used in our inferences. These examples are shown (and boxed) in **Supplementary Data 6**. (**c**) Higher activating or repressive Sharpr-MPRA regulation activity score in HepG2 cells (x axis) resulted in higher overlap with transcription factor binding sites predicted by CENTIPEDE in HepG2 cells[5] (y axis, left), and higher overlap with conserved elements identified by SiPhy-PI[33] (y axis, right). Each point represents the average of 927 nucleotide positions in each of 5,000 quantiles. Horizontal black line shows the expected overlap averaged across all 295 nucleotide positions of each region, and the green line shows the expected overlap fraction at the center nucleotide position (a stringent control). Reversed gray barplot at the top of each panel shows the density (histogram) of the distribution of Sharpr-MPRA combinedP scores in HepG2 cells. (**d**) Sharpr-MPRA inferences capture regulatory nucleotides at high resolution. Cumulative overlap (y axis) with CENTIPEDE predicted transcription factor binding sites in HepG2 cells (left) and evolutionarily conserved elements (right) is higher for MaxPos, than for the stringent control of CenPos or for SymPos, indicating this is not a positional bias. Each set is ranked from highest (left) to lowest (right) absolute Sharpr-MPRA score in MaxPos/CenPos/SymPos nucleotides (x axis) in HepG2 cells (see **Supplementary Fig. 21** for K562 cells, and for individual promoter types). Dotted lines mark thresholds at absolute score ≥2, ≥1 and ≥ 0.5. MaxPos, CenPos and SymPos nucleotide positions are illustrated in the example of **b**.

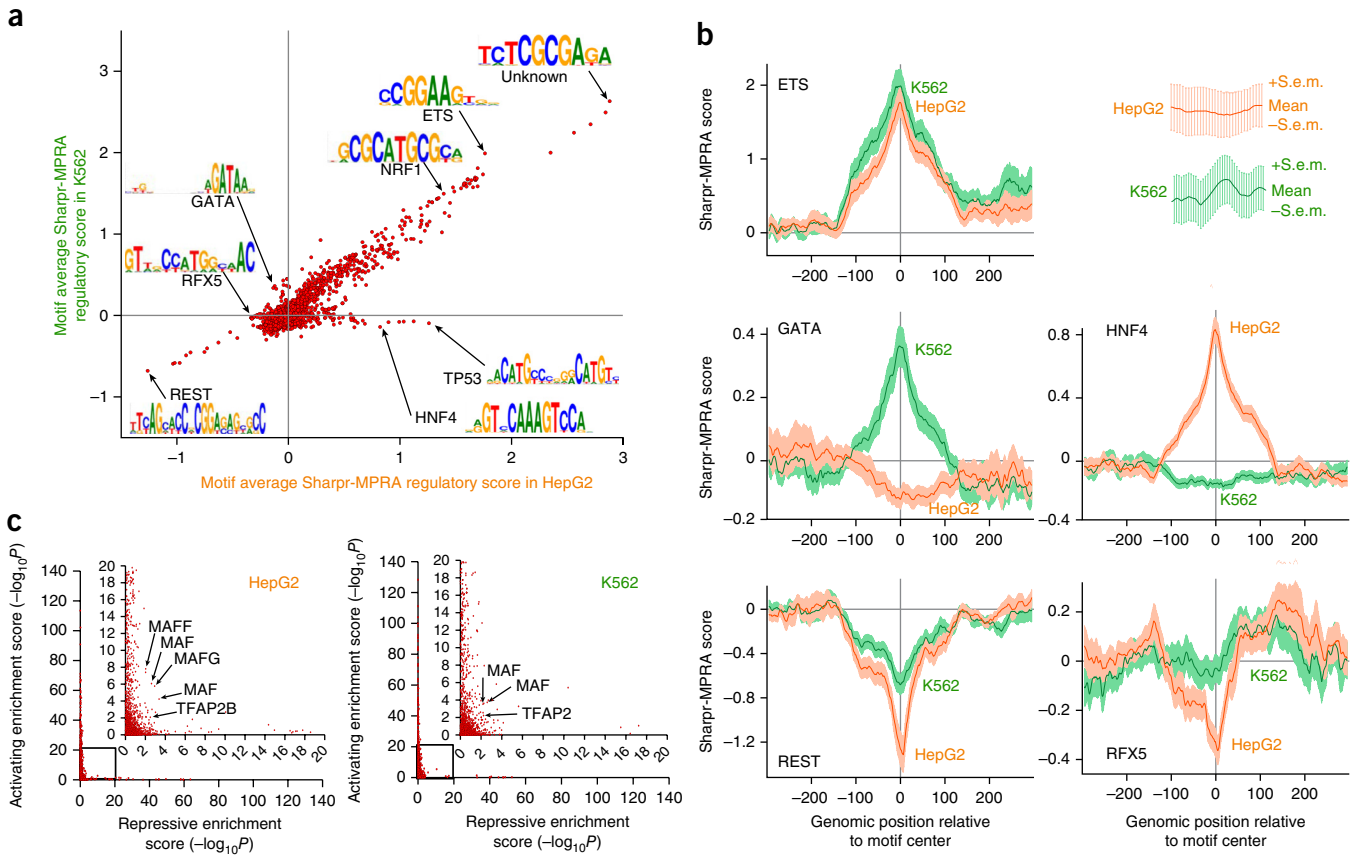

Figure 4 Comparison of Sharpr-MPRA with motif annotations. (**a**) Comparison of average Sharpr-MPRA score for regulatory motifs from a previously assembled compendium[11] (points) in HepG2 vs. K562 cells, averaged at the center position of all instances for each motif. Arrows highlight motif examples mentioned in the text (**Supplementary Table 2**). Only motifs with more than 10 instances are shown. (**b**) Aggregation plots of the regulation score (*y* axis) at increasing varying genomic positions relative to the motif center (*x* axis) for K562 and HepG2 cells for all motif instances, predicted independently of cell type in ref. 11, for ETS_known9, GATA_known14, REST_known2, HNF4_known18, and RFX5_known6 regulatory motifs. Error bar height is one s.e.m. (**c**) Activating enrichment score and repressive enrichment score for the regulatory motif compendium[11] (points) in HepG2 and K562 cells, based on the statistical significance ($-\log_{10} P$) for the enrichment of the center motif position for nucleotides with Sharpr-MPRA scores ≤ –1 (repressive) or ≥ 1 (activating), using a one-sided binomial test. Inset expands boxed region, and does not cover any points, as no motif was enriched beyond $-\log_{10}P = 20$ for both activating and repressing positions. Arrows highlight members of MAF and AP-2 motif families discussed in the text. Similar plots using top 5% activating and repressive nucleotides are shown in **Supplementary Figure 29**.

activity in HepG2 cells compared to random expectation, a small effect for shorter *k*-mers in K562 cells, and no noticeable effect for longer *k*-mers in either cell type (**Supplementary Fig. 12**).

Sharpr-MPRA recovers known motifs and conserved regions

To establish whether our inferred regulatory nucleotides were biologically relevant, we compared them to predictions of TF binding sites that we did not use to make our inferences, including DNase-based[4,5,7,31,32] and DNase-independent[11] predictions of TF-bound nucleotides, and both motif-based[5,11,32] and motif-independent[4,7,31] predictions of regulatory nucleotides. For example, CENTIPEDE[5] motif annotations showed strong agreement for both activating and repressive scores at the nucleotide level (**Fig. 3c** and **Supplementary Fig. 13**), at the region level (**Fig. 3b** and **Supplementary Data 6–8**) and for specific regulators (**Supplementary Fig. 14**), and CENTIPEDE nucleotides showed reproducible scores (**Supplementary Fig. 9b**,**e**,**f**). All six regulatory annotation sets tested showed better agreement with our inferences than with stringent controls (**Supplementary Figs. 15** and **16**).

We also compared our results to evolutionarily conserved elements across 29 mammals[33], and found enrichment for both activating and repressive nucleotides (**Fig. 3c**, and **Supplementary Figs. 6c** and **17**), also supporting that Sharpr-MPRA captures functional nucleotides in tiled regions.

K-mer-based DeltaSVM[12] predictions of nucleotides expected to have regulatory effects when mutated also agreed with our activating nucleotides (**Supplementary Fig. 18** and Online Methods). However, DeltaSVM predictions did not capture our repressive nucleotides, even though the latter agreed with both conserved nucleotides and CENTIPEDE motifs (**Fig. 3c**).

Sharpr-MPRA captures regulatory bases at high resolution

We next evaluated whether our inferences capture high-resolution information in tiled regions. We confirmed that CENTIPEDE motif and conserved element enrichments also held when focusing on MaxPos nucleotides (**Supplementary Fig. 19**), a substantial fraction of which disagreed with DNase site center locations (66% of MaxPos nucleotides were outside the 41 central nucleotides, and 44% were outside the 81 central nucleotides; **Supplementary Fig. 20**).

We compared the motif and conserved element enrichments of MaxPos nucleotides to those of DNase site center position (CenPos) nucleotides and their symmetric position (SymPos) nucleotides (equidistant from DNase site centers) (**Fig. 3b**). The CenPos comparison

was particularly stringent, given the higher activity expected at central nucleotides and the better power to detect activity with more overlapping constructs. Despite these biases favoring central positions, MaxPos nucleotides were substantially more enriched than CenPos nucleotides for both CENTIPEDE motifs and conserved elements, in both HepG2 and K562 cells, for all ranks of high activation or repression (**Fig. 3d** and **Supplementary Fig. 21**). By contrast, their SymPos nucleotides showed substantially lower enrichments than MaxPos nucleotides for all metrics, indicating that MaxPos enrichments do not stem from distance biases.

We evaluated the effect of tiling density on functional element recovery and replicate correlation, using increasingly spaced subsets of our reporter constructs. Higher density led to stronger CENTIPEDE motif and conserved element enrichments (**Supplementary Fig. 22**) and to higher correlation between replicates (**Supplementary Fig. 23**). Saturation was not reached at the 5-bp level used here, suggesting that smaller offsets might further increase discovery power, at the cost of more constructs per region and thus fewer tiled regions.

Cell-type-specific activating and repressing motifs

We used our Sharpr-MPRA results to study a compendium of 1,934 known and predicted regulatory motifs[11]. For each motif, we computed an average motif score (across all central motif positions), and separately an 'activating' and a 'repressive' enrichment score (based on its enrichment in nucleotides with scores ≥ 1 and ≤ −1, respectively) for each cell type (**Supplementary Table 2** and Online Methods). Motif scores were largely unchanged between experiments with minP and SV40P, and when using only the high-variance prior parameter (≥ 0.95 correlation; **Supplementary Fig. 24**).

Most motifs showed similar average scores between the two cell types (**Fig. 4a**). For example, motifs for the ETS and NRF1 regulators showed among the strongest activating average scores in both HepG2 and K562 cells, and known repressor REST motif showed the most repressive average score in both. We found the most activating average score in both cell types for variants of the TCTCGCGAGA palindrome, which was present in our compendium[11] based on its *de novo* discovery in chromatin immunoprecipitation-sequencing (ChIP-seq) experiments for diverse regulators (including NR3C1, BRCA1, ETS, CHD2, and ZBTB33; ref. 34). This motif lacks a well-established regulator *in vivo*[35], despite support for its importance from strong evolutionary conservation[36], high nearby gene expression[5], and other experimental and bioinformatics evidence[34,37].

A subset of motifs showed significant differences (using a paired *t*-test) in scores between HepG2 and K562 cells, (**Supplementary Fig. 25a**). Significantly different activating motifs included HNF4, RXRA, PPARA, HNF1A, HNF1B and FOXA in HepG2 cells, consistent with known liver-related roles, and GATA, SP1 and KLF in K562 cells, consistent with K562 cell roles[38] (**Supplementary Fig. 26a**). Significantly different repressive motifs included multiple RFX motifs in HepG2 cells (**Supplementary Fig. 26b**), consistent with previous evidence for one enhancer[39].

Cell-type-specificity was also reflected in the position-specific aggregated distribution of activity scores surrounding all instances of these motifs (**Fig. 4b**). Activating motifs (for example, ETS, GATA and HNF4), showed a positive peak surrounding their instances, and repressor motifs (for example, REST and RFX) showed a negative peak, in each case matching the expected cell types. These patterns were stronger when motifs occurred in more central positions (**Supplementary Fig. 27**), as expected given the higher reporter coverage and absolute activity of central positions.

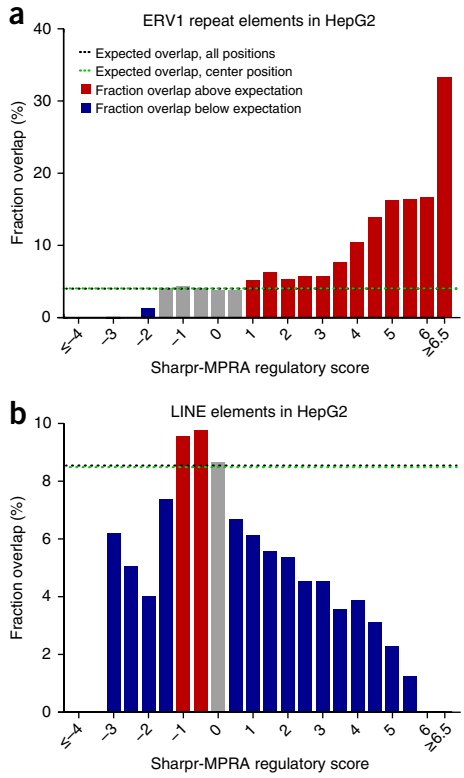

Figure 5 Regulatory activity of ERV1 and LINE repeats. (**a**,**b**) For nucleotides of varying Sharpr-MPRA regulatory activity score in HepG2 cells the fraction that overlaps with annotated repeat elements showed strong ERV1 repeat enrichment at the most activating nucleotides (**a**) and a depletion for LINE repeats at the most activating and most repressive nucleotides (**b**). Bins were formed by assigning each base to the nearest 0.5 value based on its regulatory score. Extreme bins contain extreme values as indicated. Horizontal lines denote expected overlap based on center position (CenPos, green), and all positions (black). Enrichments and depletions for K562 cells and for additional repeats are shown in **Supplementary Figure 30.**

The motif compendium resulted in more activating than repressive motif scores, both based on average scores (**Fig. 4a**), and based on activating vs. repressive enrichment scores (511 vs. 117 in HepG2 cells; 474 vs. 79 in K562 cells, respectively, at an uncorrected *P* value of 0.01; **Fig. 4c**, **Supplementary Fig. 25b** and **Supplementary Table 2**). The higher number of activating motifs also held for average scores of all 7-mer sequences (**Supplementary Fig. 28**), indicating it is not an ascertainment bias in the compendium used.

A small number of 'dual-role' motifs showed signatures of both activating and repressive function in the same cell type (14 in HepG2 and 15 in K562 cells, of which six were in common). These included several motifs for MAF proteins, consistent with previous reports of both activation and repression[40] and different motifs of the AP-2 family. Alternative cutoffs (top 5% activating and repressive nucleotides) also resulted in very few dual-role motifs (6 in HepG2 and 9 in K562 cells; **Supplementary Fig. 29**).

Enrichment of ERV1 repeats in activating nucleotides

The most strongly activating nucleotides in HepG2 cells showed substantial enrichment for long terminal repeat (LTR) elements (**Fig. 5** and **Supplementary Fig. 30**), consistent with previous reports of their ability to drive gene expression[41]. This helps explain why conserved

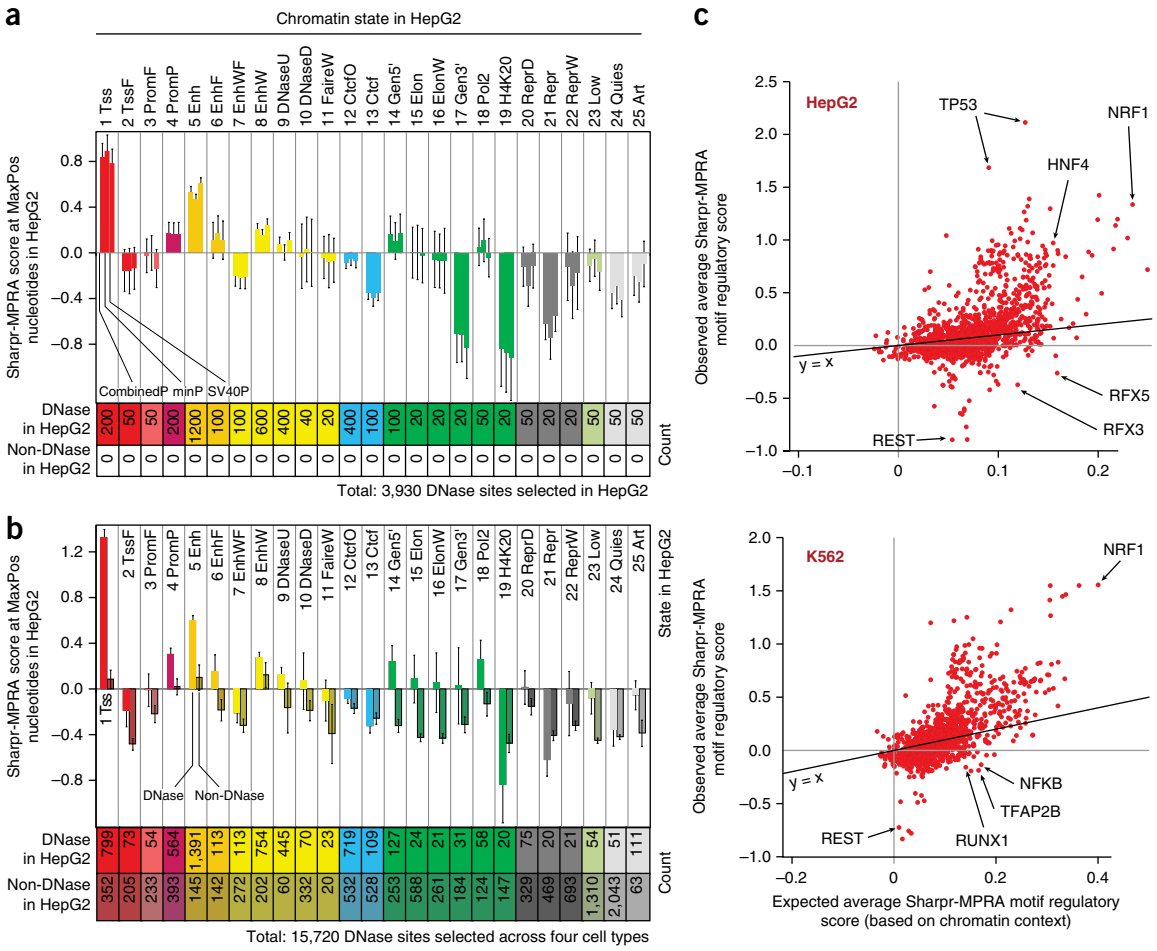

Figure 6 Endogenous chromatin state is predictive of reporter activity. (**a**,**b**) Average HepG2 cell Sharpr-MPRA regulatory score (*y* axis) and standard error (vertical error bars) for each chromatin state (columns) for all 3,930 DNase sites selected in HepG2 cells (**a**) and all 15,720 regions selected in all four cell types (**b**), evaluated at nucleotide positions of maximum absolute activity (MaxPos). In **a**, each group of consecutive bars shows the combinedP, minP and SV40P results. All 3,930 regions correspond to DNase sites in HepG2 cells, as they were selected in HepG2 cells. In **b**, the combinedP score is shown separately for regions corresponding to DNase (light shading) and non-DNase (darker shading) sites in HepG2 cells. Some DNase sites selected in other cell types were also DNase in HepG2, leading to an increased DNase count compared to data in **a**. All non-DNase sites in HepG2 cells were DNase sites in the cell type in which they were selected. The chromatin state of the center position is shown. K562 cells plots in **Supplementary Figure 31**. (**c**) For all motifs (ref. 11) (circles) in HepG2 cell-selected regions (top) and K562 cell-selected regions (bottom), relationship between their average combinedP Sharpr-MPRA score in the corresponding cell type (*y* axis) and their expected score based on the chromatin states in which the motif occurs (*x* axis), quantified as the median of randomized motif occurrences that preserve positional and chromatin state distributions (Online Methods). Only motifs with 20 or more evaluated instances in selected regions are shown. Randomization 95th percentile confidence intervals are shown in **Supplementary Figure 39**.

elements were less enriched in the most extreme activating scores (**Fig. 3c**), as no LTRs overlapped conserved elements with the most extreme activating scores.

Among LTRs, endogenous retroviral sequence 1 (ERV1) repeats showed the strongest enrichment in HepG2 activating nucleotides (**Fig. 5a**), overlapping 33% of the 820 nucleotides with the highest regulatory scores (≥6.5 bin), vs. only 4% expected on average (eight-fold enrichment). Regulatory roles have been previously hypothesized for ERV1 repeats, based on TF binding and RNA interference evidence[42,43], and our results indicate these repeats can function autonomously and lead to strong episomal expression.

By contrast, long interspersed nuclear elements (LINEs) were strongly depleted in both activating and repressive nucleotides in HepG2 and K562 cells (**Fig. 5b** and **Supplementary Fig. 30b**), indicating repeat-specific regulatory functions. Moreover, ERV1 and other LTR enrichments were weaker in K562-cell than HepG2-cell activating nucleotides (**Supplementary Fig. 30a**,**c**), indicating cell-type-specific repeat functions.

Epigenomic signatures are predictive of reporter activity

We analyzed the relationship between regulatory activity scores and endogenous chromatin state, enabled by inclusion of all chromatin states (defined in **Supplementary Fig. 5**), and both DNase and non-DNase sites in each cell type (by including DNase sites only active in other cell types).

Among DNase sites, endogenous chromatin state was predictive of regulatory function in reporter assays (quantified for each region by its MaxPos Sharpr-MPRA score; **Fig. 6a** and **Supplementary Fig. 31a**). Regions in active promoter or H3K27ac-marked enhancer chromatin states showed higher Sharpr-MPRA activating scores, regions in weak enhancer states showed intermediate activating scores, and regions in Polycomb-associated states showed repressive Sharpr-MPRA scores.

Conversely, among genomic locations in the same chromatin state, the DNA accessibility of the region in its endogenous context was predictive of Sharpr-MPRA reporter activity (**Fig. 6b** and **Supplementary Fig. 31b**), consistent with previous work in enhancer regions[24,28]. Together, these results indicate that the endogenous epigenomic signatures of DNA accessibility and chromatin state each capture unique information about regulatory function, and that sequence elements in these regions can show consistent activating or repressive regulatory functions outside their endogenous context.

The fraction of regions that showed activating and repressive MaxPos scores varied greatly between chromatin states and DNase classes (**Supplementary Fig. 32**). In HepG2 cells, activating regions with scores ≥ 1 included 36% of HepG2-selected DNase sites in an active promoter state (Tss) and 29% in an H3K27ac-marked enhancer state (Enh) (**Supplementary Fig. 32a**), compared to only 6% for non-DNase sites in the quiescent states (Quies) (**Supplementary Fig. 32d**) (41% and 32% vs. 5%, respectively, in K562 cells). Repressive regions with MaxPos scores ≤ −1 in HepG2 cells included 29% of HepG2-selected DNase sites in Polycomb repressed states ReprD and Repr (21% for K562 cells) (**Supplementary Fig. 32a**), compared to only 6% of all DNase sites in the active promoter state (10% in K562 cells) (**Supplementary Fig. 32c**). These comparisons allowed us to estimate false positive rates for both activating regions (for example, 6% for HepG2 cells and 5% for K562 cells) and for repressive regions (for example, 6% and 10%, respectively) relative to their respective backgrounds. These estimates are likely conservative, as all regions tested were in DNase sites in at least one cell type, and thus more likely to contain activating or repressive elements than random background nucleotides.

Beyond these thresholds, the full distribution of MaxPos scores across chromatin states and DNA accessibility (**Supplementary Fig. 33**), confirmed consistently higher activation scores for active promoter and H3K27ac-marked enhancer states across a broad range of ranked positions. For non-DNase sites, we found the strongest repressive scores for chromatin state 'DNaseD' (associated with single-cut DNase[44] and lack of double-cut DNase[8]), indicating that it contains repressive elements, consistent with a previously hypothesized repressive role[29] (**Supplementary Fig. 33c**).

The role of H3K27ac as a signature of active enhancer regions is well established[3,9,24,45] and is in agreement with our results here, but has been recently questioned in an isolated study[28] using a similar reporter assay (CRE-seq), which suggested that H3K27ac-marked regions show weaker reporter activity. That study[28] used a 7-state segmentation[30] that merged ChromHMM[46] and Segway[47] results, and tested smaller segments (130 bp) without a tiling approach and without anchoring on DNase sites, making the results dependent on positioning of the tested segments, and specifically whether DNase sites or their flanking elements were captured. Mapping their tested segments[28] on the 25-state ChromHMM annotations considered here, we found that the H3K27ac-marked enhancers selected in the study preferentially were outside DNase sites, and the non-H3K27ac enhancers selected preferentially were in DNase sites. Correcting for this bias by analyzing DNase and non-DNase sites separately, we found that H3K27ac enhancers had increased CRE-seq activity compared with non-H3K27ac enhancers (**Supplementary Fig. 34**), which is fully consistent with our results and the previous literature.

Similar to other studies[24,28,48,49], many predicted enhancer and promoter regions did not have reporter activity (**Supplementary Fig. 7b**). These regions showed distinct levels and patterns of endogenous TF binding and DNA accessibility, including less frequent endogenous TF binding in the tested regions, more frequent endogenous TF binding in the surrounding 2 kilobases, and proximity to other DNase sites (**Supplementary Figs. 35–38**). We interpret these findings to indicate that their endogenous activating signatures may arise at least in part from TF binding in nearby regions, consistent with their lower reporter gene expression when tested in isolation.

A subset of repressive motifs in active chromatin

For each motif, we analyzed the relationship between its observed average regulatory score and the average regulatory score that would be expected based on the chromatin states where the motif occurs, quantified as the median of randomized motif occurrences that preserve positional and chromatin-state distributions (Online Methods). Overall, the observed average score of a motif correlated with its expected average score (0.54 in HepG2 cells and 0.68 in K562 cells for motifs with ≥ 20 instances; **Fig. 6c**, **Supplementary Fig. 39**, **Supplementary Table 2**). For example, NRF1 showed both a high average regulation score and a high expected score in both HepG2 and K562 cells, indicating that it acts as an activator in active chromatin states.

Several motifs showed only moderate expected scores but very strong activating or repressive motif scores, suggesting they maintain their functions regardless of their genomic context. In HepG2 cells, for example, *TP53* showed only a moderate expected score, but the highest score among all evaluated motifs, consistent with its proposed role as a pioneer factor[50]. At the other end of the spectrum, REST showed only an intermediate expected score but had the most repressive motif score, indicating strong repressor functions irrespective of context.

A small number of motifs showed repressive motif scores, but among the highest expected scores, suggesting they play 'attenuator' repressive roles in activating chromatin contexts. RFX family motifs had among the most repressive motif scores in HepG2, but among the most activating expected scores (**Fig. 6c**). Consistent with 'attenuator' roles, they showed repressive (negative) activity in our positional activity analysis, but they were flanked by activating (positive) scores (**Fig. 4b**). Moreover, in our pilot analysis in HepG2 cells, RFX family motifs were discovered as enriched in segments inferred to be repressive, but were found in active enhancer regions (**Supplementary Fig. 4**). Indeed, a repressive role has been experimentally confirmed in an enhancer of the *CDX2* gene for a single RFX1 motif instance in HepG2 cells[39]. Our results indicate a broader repressive role in active regions, a discovery that stems directly from our ability to distinguish activating vs. repressive nucleotides using our tiling approach.

DISCUSSION

We presented Sharpr-MPRA, a combined experimental and computational approach for high-resolution mapping of activating and repressive nucleotides across thousands of genomic regions. We used dense tiling of MPRA constructs spanning 4.6 million nucleotides targeting 15,720 regions at a resolution typically not afforded without perturbation experiments, which are traditionally not applicable at this scale. Sharpr-MPRA distinguishes activating from repressive nucleotides, and directly assesses regulatory function in a reporter assay, thus complementing the endogenous epigenomic signatures surveyed by ENCODE[6], Roadmap Epigenomics[9] and related projects.

Nucleotides with stronger activating or repressive Sharpr-MPRA scores were enriched in evolutionarily conserved elements and predicted cell-type-specific TF binding sites. To our surprise, however, both enrichments were weaker for the most highly active nucleotides in HepG2 cells. Instead, the strongest reporter activity overlapped ERV1 endogenous retroviral repeat elements, which might speak to a potential role in regulatory turnover across even closely related species.

Endogenous epigenomic signatures were predictive of reporter gene expression, with chromatin state and DNA accessibility each providing relevant information. Segments with endogenous active promoter and H3K27ac-marked enhancer signatures drove the strongest average reporter gene activation, and segments showing endogenous Polycomb-associated signatures were among those with the strongest average reporter gene repression. These results indicate that even when tested outside their endogenous context, DNA sequence elements maintain the activating and repressive functions reflected in their endogenous epigenomic signatures.

Aggregation of activity scores at the motif level revealed cell-type-specific motifs, and distinguished activator and repressor motifs. Motif activity typically correlated with chromatin context, with activating motifs found in active chromatin states, and repressive motifs in repressive states. Notable exceptions included putative 'attenuator' motifs that showed repressive roles but were found in active chromatin states (for example, RFX motifs in HepG2 cells) and putative 'pioneer' motifs, which showed strong activity regardless of their chromatin-state context (for example, activator *TP53* and repressor *NRSF*), although directed experimentation and endogenous modulation will be needed to confirm these predictions. Most motifs showed activating-only or repressive-only signatures, but a small number of 'dual-role' motifs showed both activating and repressive signatures (for example, members of the MAF and AP-2 protein families). To our surprise, the sequence pattern with the strongest average activity in both cell types, TCTCGCGAGA, was not associated with a well-established regulator, highlighting our still incomplete understanding of regulatory elements, and the importance of unbiased dissection of regulatory regions.

Limitations include a need for longer sequences to show reporter activity in some regions, which might be overcome by improved DNA synthesis, and limited transfection efficiency in some cell types, which may require alternative delivery approaches (for example, viral transduction). Additionally, we assumed additive effects in our analysis, which may miss interactions between different nucleotide positions. Barcode effects and other factors may cause experimental noise, which could be overcome by higher density tiling or different experimental barcoding strategies[49]. We only tested elements in episomal assays, which provides direct information on regulatory activity, but does not capture potential effects of the endogenous chromatin context, and we only transfected unstimulated cells, although some sites may only function after specific stimulations.

We envision diverse uses for the results and methodology presented here. On one hand, the annotation of activating or repressive function for 4.6 million nucleotides can be useful to ask biological questions beyond the ones addressed here, and to train new computational models (for example, for predicting activating and repressive nucleotides outside the regions surveyed here, or predicting the effect of noncoding variation on regulatory function[48,49]). On the other hand, Sharpr-MPRA's combined experimental and computational strategy can be broadly useful for dissecting regulatory regions across individuals, species, cell types, conditions and disease states.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Gene Expression Omnibus: GSE71279.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank P. Kheradpour and J.-P. Vert for useful discussions related to this work. This work was supported by US National Institutes of Health (NIH) grants R01ES024995, U01HG007912 and U01MH105578 (J.E.), R01HG006785 (T.S.M.), R01GM113708, U01HG007610, R01HG004037, U54HG006991 and U41HG007000 (M.K.), an US National Science Foundation CAREER Award #1254200, and an Alfred P. Sloan Fellowship (J.E.).

AUTHOR CONTRIBUTIONS

J.E. and M.K. designed the sequences, developed the computational methods and analyzed the results. A.M., X.Z., L.W., P.R. and T.S.M. conducted the experimental work. T.S.M. oversaw the experimental work. J.E. and M.K. wrote the paper with substantial input from T.S.M.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Heintzman, N.D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
2. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* **28**, 817–825 (2010).
3. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
4. Boyle, A.P. *et al.* High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res.* **21**, 456–464 (2011).
5. Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* **21**, 447–455 (2011).
6. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
7. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90 (2012).
8. Thurman, R.E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
9. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
10. Claussnitzer, M. *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N. Engl. J. Med.* **373**, 895–907 (2015).
11. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.* **42**, 2976–2987 (2014).
12. Lee, D. *et al.* A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* **47**, 955–961 (2015).
13. Gröschel, S. *et al.* A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell* **157**, 369–381 (2014).
14. Patwardhan, R.P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27**, 1173–1175 (2009).
15. 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).
16. Patwardhan, R.P. *et al.* Massively parallel functional dissection of mammalian enhancers *in vivo*. *Nat. Biotechnol.* **30**, 265–270 (2012).
17. 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. USA* **109**, 19498–19503 (2012).
18. Vierstra, J. *et al.* Functional footprinting of regulatory DNA. *Nat. Methods* **12**, 927–930 (2015).
19. Canver, M.C. *et al.* BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527**, 192–197 (2015).
20. Shen, S.Q. *et al.* Massively parallel cis-regulatory analysis in the mammalian central nervous system. *Genome Res.* **26**, 238–255 (2016).
21. Rajagopal, N. *et al.* High-throughput mapping of regulatory DNA. *Nat. Biotechnol.* **34**, 167–174 (2016).
22. Korkmaz, G. *et al.* Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat. Biotechnol.* **34**, 192–198 (2016).
23. Arnold, C.D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
24. Kheradpour, P. *et al.* Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.* **23**, 800–811 (2013).
25. Gisselbrecht, S.S. *et al.* Highly parallel assays of tissue-specific enhancers in whole *Drosophila* embryos. *Nat. Methods* **10**, 774–780 (2013).
26. Dickel, D.E. *et al.* Function-based identification of mammalian enhancers using site-specific integration. *Nat. Methods* **11**, 566–571 (2014).
27. Murtha, M. *et al.* FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells. *Nat. Methods* **11**, 559–565 (2014).
28. 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).

29. Ernst, J. & Kellis, M. Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types. *Genome Res.* **23**, 1142–1154 (2013).
30. Hoffman, M.M. *et al*. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.* **41**, 827–841 (2013).
31. Piper, J. *et al*. Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res.* **41**, e201 (2013).
32. Sherwood, R.I. *et al*. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat. Biotechnol.* **32**, 171–178 (2014).
33. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
34. Raghav, S.K. *et al.* Integrative genomics identifies the corepressor SMRT as a gatekeeper of adipogenesis through the transcription factors C/EBPβ and KAISO. *Mol. Cell* **46**, 335–350 (2012).
35. Blattler, A. *et al.* ZBTB33 binds unmethylated regions of the genome associated with actively expressed genes. *Epigenetics Chromatin* **6**, 13 (2013).
36. Xie, X. *et al*. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434**, 338–345 (2005).
37. Mikula, M. *et al.* Comprehensive analysis of the palindromic motif TCTCGCGAGA: a regulatory element of the HNRNPK promoter. *DNA Res.* **17**, 245–260 (2010).
38. Hu, J.H., Navas, P., Cao, H., Stamatoyannopoulos, G. & Song, C.-Z. Systematic RNAi studies on the role of Sp/KLF factors in globin gene expression and erythroid differentiation. *J. Mol. Biol.* **366**, 1064–1073 (2007).
39. Watts, J.A. *et al*. Study of FoxA pioneer factor at silent genes reveals Rfx-repressed enhancer at Cdx2 and a potential indicator of esophageal adenocarcinoma development. *PLoS Genet.* **7**, e1002277 (2011).
40. Yang, Y. & Cvekl, A. Large Maf Transcription Factors: Cousins of AP-1 Proteins and Important Regulators of Cellular Differentiation. *Einstein J. Biol. Med.* **23**, 2–11 (2007).
41. Bannert, N. & Kurth, R. Retroelements and the human genome: new perspectives on an old relation. *Proc. Natl. Acad. Sci. USA* **101** (Suppl. 2), 14572–14579 (2004).
42. Wang, T. *et al.* Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc. Natl. Acad. Sci. USA* **104**, 18613–18618 (2007).
43. Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.* **42**, 631–634 (2010).
44. Song, L. *et al*. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res.* **21**, 1757–1767 (2011).
45. Creyghton, M.P. *et al*. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. USA* **107**, 21931–21936 (2010).
46. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
47. Hoffman, M.M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* **9**, 473–476 (2012).
48. Ulirsch, J.C. *et al.* Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell* **165**, 1530–1545 (2016).
49. Tewhey, R. *et al.* Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **165**, 1519–1529 (2016).
50. Sammons, M.A., Zhu, J., Drake, A.M. & Berger, S.L. TP53 engagement with the genome occurs in distinct local chromatin environments via pioneer factor activity. *Genome Res.* **25**, 179–188 (2015).

ONLINE METHODS

Pilot large-step massively parallel reporter assay design. As a pilot design, we selected 250 regulatory regions to test. Each selected regulatory region was tiled by nine sequence tiles of 145 bp in length placed in 30-bp offsets so that adjacent sequences would have 115 bp in common. Each tile was associated with 24 unique barcodes. In this design we used 216 barcodes per putative regulatory region.

Of the 250 regions, we selected 200 so that their center position came from a HepG2 H3K27ac-marked 'strong enhancer' chromatin states from the 15-state chromatin state model in ref. 3 (**Supplementary Fig. 1b**). To define the specific locations to test, we first defined a dip score based on the ENCODE hg18 HepG2 H3K27ac signal[3] on chromosomes 1–22 and chromosome X. We defined the dip score to be the sum of the signal from positions 200 bp away in both directions minus twice the signal at the dip center. We then ranked all positions at a 25-nucleotide resolution based on their dip score excluding from consideration positions that either (i) did not have the minimum HepG2 H3K27ac signal within 200 nucleotides, or (ii) were tied for the minimum H3K27ac signal with another position within 200 nucleotides and did not have a strictly greater dip score than the tied position. This requirement often enabled us to center our tiles in nucleosome-depleted regions. We excluded positions if they were within 2 kb of an annotated transcription start site based on the GENCODE v2b annotations. We also excluded from testing those sequences that contained a GGTACC, TCTAGA or GGCCNNNNNGGCC, as these were recognition sequences of the restriction enzymes. We selected 100 positions to be the highest ranked non-excluded positions. We selected an additional 100 positions to cover a range of dip scores among the non-excluded positions, grouping the remaining regions (after the top 100) into 100 dip score ranges, and selecting the region with the maximal dip score in each range. Formally, let m denote the dip score of the 101st ranked position. We defined an interval width w to be $(\ln(m) - \ln(10))/99$. For selections $i = 1, \ldots, 100$ made to cover a range of dip scores, we selected as the i^{th} range one to be the region with the greatest dip score, v, such that it was still the case that $\ln(v) \leq m - (i-1) \times w$. The center of the selected 25-bp position interval was used as the center of the center tile.

An additional 50 sequences were selected based on the same procedure to select the top 100 sequences for HepG2, but based on the K562 data, limited to the top 50, with the additional constraint that they were in low-activity states in HepG2 ('weak transcribed' or 'heterochromatin;low signal') (**Supplementary Fig. 1b**).

For the motif enrichment analysis, we converted the coordinates from this design to hg19 using the UCSC Genome Browser liftover tool.

Identification of significant adjacent changes in the pilot large-step data. Among all pairs of adjacent tiles we identified a set of pairs that had a significant difference in reporter expression level at a 5% false discovery rate. Our procedure for doing this was as follows. Let p_i denote the P value for the i^{th} adjacent pair (where $i = 1, \ldots, 2{,}000$ in our case) that there is a significant difference in the reporter expression between the first and second tiles of the pair. We let $p_{i,\mathrm{L}}$ denote the P value for the one sided test that the second tile has lower expression than the first and $p_{i,\mathrm{G}}$ the P value for a one-sided test that the second tile has greater expression than the first. We further let $p_{i,\mathrm{L},r}$ and $p_{i,\mathrm{G},r}$ for $r = 1, \ldots, R$, where $R = 2$ in our case, denote the P value for that in the i^{th} pair and the r^{th} replicate that the second tile had lower or greater expression than the first, respectively.

We computed the individual $p_{i,\mathrm{L},r}$ and $p_{i,\mathrm{G},r}$ P values based on a one-sided Mann-Whitney test on all the individual barcoded expression values for a tile, which was up to 24 in our case. To compute $p_{i,\mathrm{L},r}$, we first obtained a two-sided P value, v, for a Mann-Whitney test using Apache Commons Math 3.3 (https://commons.apache.org/). We then assigned the P value $v/2$ if the second tile had average lower or equal ranks than the first and the P value $1 - v/2$ otherwise; we made the opposite assignments for $p_{i,\mathrm{G},r}$. The $p_{i,\mathrm{L}}$ P value was computed based on Fisher's method combining the $p_{i,\mathrm{L},r}$ for $r = 1, \ldots, R$ P values, and likewise for the $p_{i,\mathrm{G}}$ P values. The P value p_i we defined to be $p_i = \min(1, 2 \times \min(p_{i,\mathrm{L}}, p_{i,\mathrm{G}}))$. We multiply by two here to correct for having tested both P values separately as one-sided tests. We obtained a false discovery rate for the set of P values p_i for $i = 1, \ldots, 2{,}000$ using a Benjamini-Hochberg procedure.

Scale-up Sharpr-MPRA assay design. We targeted 15,720 DNase sites, consisting of 3,930 tiled regions based on each of four cell types: HepG2, H1-hESC, K562 and HUVECs. The DNase sites were generated by the University of Washington ENCODE Group[8] and specifically we used the location of each peak call contained in the hg19 files: wgEncodeUwDnaseHepg2PkRep1.narrowPeak.gz, wgEncodeUwDnaseH1hescPkRep1.narrowPeak.gz, wgEncodeUwDnaseK562PkRep1.narrowPeak.gz, and wgEncodeUwDnaseHuvecPkRep1.narrowPeak.gz for the HepG2, H1-hESC, K562 and HUVEC cell types, respectively. The selection of the subset of 3,930 regulatory regions based on each cell type was conducted independently. To increase chromatin state diversity, we selected regions using a richer 25-state chromatin state model (ChromHMM[46] model from refs. 29,30), which was based on 14 input tracks, consisting of 8 histone modification marks, CTCF, POL2, DNase (single-cut, generated by Duke[44]; and double-cut, generated by University of Washington[8]), FAIRE[44] and input. The counts of each state were manually specified to ensure some coverage of each chromatin state, greater coverage in states more associated with DNase, and deeper coverage of enhancer chromatin states (**Fig. 1b**). The regions were then randomly selected given the counts for each state. As with the Pilot design, we excluded from testing those sequences that contained a GGTACC, TCTAGA or GGCCNNNNNGGCC as these were recognition sequences of the restriction enzymes, but we had no restriction in this design with respect to position relative to annotated genes. Selected regions were tiled with 31 sequence tiles each 145-bp long and placed at 5-bp offsets, centered on the DNase site. Each tile was associated with a single barcode. The tiled regions based on each cell type and each chromatin state were randomly and evenly divided between the two array designs, which led to each design having 7,860 tiled regions. If the same or overlapping regions were selected based on different cell types, we retained both tiled regions and considered them separately except in forming the browser tracks in which case we averaged all regulatory scores for a given nucleotide. In total, we targeted 15,720 regions, some of which overlapped, resulting in 15,455 unique non-overlapping regions.

Experimental procedure. The experimental methods for a massively parallel reporter assay are described in ref. 51. Oligonucleotide library synthesis was performed by Agilent Inc[52], and the cell culture, transfection and plasmid construction were done as in ref. 24. The MPRA vector backbone and promoter-reporter cassettes are available from Addgene (plasmids 49349, 49353 and 49354). The K562 and HepG2 cell lines were obtained directly from ATCC (CCL-243 and HB-8065). ATCC performed routine authentication using STR analysis. We did not perform separate authentication or mycoplasma testing.

For the pilot design, we used a single 55K-spot array for synthesis and designed sequences for 54,000 of them (2,250 unique sequences with 24 barcodes each). Transfection and barcode sequencing experiments were conducted in replicate in both K562 and HepG2 cells using the SV40 promoter (**Supplementary Fig. 1a**). The plasmid pools were amplified and sequenced in replicate and shared among the K562 and HepG2 experiments.

For the scale-up design, we used two 244K-spot arrays for synthesis and designed sequences for 243,660 of them of which 243,573 and 243,564 (99.96%) were included in the synthesis for the two arrays, leading to a total of 487,137 probed tiles. Transfection and barcode sequencing experiments were conducted for each array in both HepG2 and K562, each using both a minimal and SV40 promoter, each in two replicates (16 experiments total). For these experiments we amplified and sequenced four plasmid DNA pools, one for each combination of array design and promoter type, with RNA replicate experiments normalized to the same plasmid DNA pool.

Data normalization. The initial data processing for data from one experiment was as follows. We generated DNA and RNA counts based on the procedure described in ref. 24 and added a pseudocount of 1 to these counts for smoothing. We then divided all RNA values by the sum of all RNA values and divided the DNA values by the sum of all the DNA values. For the readout corresponding to a given barcode we computed the log base two ratio of the RNA to DNA counts. We treated as missing those barcodes that had less than 20 for the original DNA counts associated with them. For the pilot design, we used the median value among multiple barcodes of the same sequence tile. We then normalized these values by taking the difference with the average value for the expression of the tiles in the positions furthest from the H3K27ac dip centers (tiles #1 and #9) as an approximation for background in these experiment. For the scale-up design, additional normalization was conducted on the inferred activity values (see below).

doi:10.1038/nbt.3678

Sharpr-MPRA regulatory activity scores. To compute Sharpr-MPRA regulatory activity scores from tiled reporter data, we assumed each reporter sequence had length L, which was 145 bp in our case, and a consistent step size, s, between adjacent reporter sequences, which was 5 in our case. We assumed that L was divisible by s, and let $N = L/s$ denote the number of intervals of the step size overlapped by a reporter sequence, which was 29 in our case. We let J denote the number of overlapping reporter tiles covering a tiled region, which was 31 in our case. We let K denote the total number of non-overlapping intervals of step size s intervals that have coverage by at least one reporter sequence which is $N + (J - 1)$, or 59 in our case. We let $W = s \times K$ denote the total number of nucleotides covered by at least one reporter sequence, which was 295 in our case, and we index them using $i = 1,\ldots,295$. We let T denote the total number of tiled regions tested in a single design, which was 7,860 in our case. We let R denote the number of experiments for the design that we are combining, where $R = 2$ when we considered the SV40P and minP experiments separately and $R = 4$ when we combined all experiments for a design in the same cell type.

We let $A_{r,t,k}$ denote a random variable for the unobserved regulatory activity for the k^{th} s = 5-bp interval where $k = 1, \ldots, K = 59$ in the t^{th} tiled region in the r^{th} experiment being combined. We let $M_{r,t,j}$ denote a random variable for the normalized expression value for the reporter sequence at the j^{th} tile offset, where $j = 1, \ldots, J = 31$, of the t^{th} tiled region in the r^{th} experiment being combined (**Fig. 3a**). We let $m_{r,t,j}$ denote the corresponding observed value, and if there was no observed value it is set to null. Our objective is to infer the maximum a posteriori values of the $A_{r,t,k}$ variables conditioned on the observed values for the $M_{r,t,j}$ variables.

We assumed that each $A_{r,t,k}$ is normally distributed with mean μ_{a_r} and variance σ_a^2, that is

$$A_{r,t,k} \sim N(\mu_{a_r}, \sigma_a^2) \tag{1}$$

Let $\widehat{M_r} = \{m_{r,t,j} \mid m_{r,t,j} \neq null\}$ denote the multiset of all observed reporter values in the r^{th} experiment. μ_{a_r} was set to the empirical mean of the observed reporter values in the r^{th} experiment, that is,

$$\mu_{a_r} = \frac{1}{|\widehat{M_r}|} \sum_{m \in \widehat{M_r}} m \tag{2}$$

σ_a^2 is a free parameter. We performed the inference with σ_a^2 set both to 1 and 50 and combined the inferences using a procedure described below.

We assumed that each $M_{r,t,j}$ is normally distributed with mean $\mu_{m_{r,t,j}}$ and variance $\sigma_{m_r}^2$, that is

$$M_{r,t,j} \sim N(\mu_{m_{r,t,j}}, \sigma_{m_r}^2) \tag{3}$$

We assumed $\mu_{m_{r,t,j}}$ to be the mean of all the unobserved regulatory activity variables corresponding to intervals for which the j^{th} reporter tile overlaps, and we set $\sigma_{m_r}^2$ to the empirical variance of the observed reporter values in r^{th} experiment, that is,

$$\mu_{m_{r,t,j}} = \frac{1}{N} \sum_{l=0}^{N-1} A_{r,t,j+l}$$

$$\sigma_{m_r}^2 = \frac{1}{|\widehat{M_r}|} \sum_{m \in \widehat{M_r}} \left(m - \frac{1}{|\widehat{M_r}|} \sum_{m \in \widehat{M_r}} m \right)^2 \tag{4}$$

The vector $X_{r,t}$ comprised of the $A_{r,t,k}$ and $M_{r,t,j}$ variables in the r^{th} experiment for t^{th} tiled region, can be expressed as a multivariate normal distribution

$$X_{r,t} = \begin{bmatrix} A_{r,t} \\ M_{r,t} \end{bmatrix} \sim N(\mu_{x_{r,t}}, \Sigma_{x_{r,t}}) \tag{5}$$

where $A_{r,t} = [A_{r,t,1} \ldots A_{r,t,K}]^T$ and $M_{r,t} = [M_{r,t,1} \ldots M_{r,t,J}]^T$. If a $m_{r,t,j}$ is considered missing, that is, has a null value, the corresponding $M_{r,t,j}$ variable was omitted, but we did not re-index the remaining $M_{r,t,j}$ variables. We let

$$\mu_{x_{r,t}} = \begin{bmatrix} \mu_{A_{r,t}} \\ \mu_{M_{r,t}} \end{bmatrix}, \Sigma_{x_{r,t}} = \begin{bmatrix} \Sigma_{A_{r,t},A_{r,t}} & \Sigma_{A_{r,t},M_{r,t}} \\ \Sigma_{M_{r,t},A_{r,t}} & \Sigma_{M_{r,t},M_{r,t}} \end{bmatrix} \tag{6}$$

Both $\mu_{A_{r,t}}$ and $\mu_{M_{r,t}}$ are column vectors with all values equal to μ_{a_r}.

Conditioning on observing $M_{r,t} = m_{r,t}$ the maximum a posteriori values for values in $A_{r,t}$ was determined as the mean in a conditional multivariate normal distribution which is given by[53]

$$\mu_{A_{r,t}} + \Sigma_{A_{r,t},M_{r,t}} \Sigma^{-1}_{M_{r,t},M_{r,t}} (m_{r,t} - \mu_{M_{r,t}}) \tag{7}$$

For $\Sigma_{M_{r,t},M_{r,t}}$ we have:

$$\Sigma_{M_{r,t,u},M_{r,t,u}} = \sigma_{m_r}^2 + \frac{\sigma_a^2}{N}$$

$$\Sigma_{M_{r,t,u},M_{r,t,v}} = \frac{\sigma_a^2 (N - |u - v|)}{N^2} \text{ if } (0 < |u - v| < N)$$

$$\Sigma_{M_{r,t,u},M_{r,t,v}} = 0 \text{ if } (|u - v| \geq N) \tag{8}$$

For $\Sigma_{A_{r,t},M_{r,t}}$ we have:

$$\Sigma_{A_{r,t,k},M_{r,t,u}} = \frac{\sigma_a^2}{N} \text{ if } u \leq k < u + N$$

$$\Sigma_{A_{r,t,k},M_{r,t,u}} = 0 \text{ otherwise} \tag{9}$$

The above expressions are derived in **Supplementary Note 1**.

We denoted with $a_{r,t,k}$, the inferred value for the k^{th} interval in the t^{th} tiled region in the r^{th} experiment. We note that the modeling to infer these values can also be viewed as a specific instance of Bayesian linear regression.

We then standardized all inferred values within an experiment by subtracting the mean and dividing by the s.d. to define $z_{r,t,k}$, the standardized regulatory score for the k^{th} interval in the t^{th} tiled region in the r^{th} experiment (**Supplementary Note 2**). We also defined as our merged regulatory score $\hat{z}_{t,k}$ for the k^{th} interval in the t^{th} tiled region, which averages the standardized regulatory score from multiple experiments (**Supplementary Note 2**).

When conducting inference with two different values for σ_a^2, denoted $\sigma_{a_1}^2$ and $\sigma_{a_2}^2$, we denoted the merged regulatory scores for k^{th} interval in the t^{th} tiled region for these two parameter settings,

$$\hat{z}_{t,k}^{\sigma_{a_1}^2} \text{ and } \hat{z}_{t,k}^{\sigma_{a_2}^2},$$

respectively.

We combined them to obtain the value denoted $\hat{z}'_{t,k}$ with a procedure that takes the more conservative score when the signs of the values agree and 0 otherwise (**Supplementary Note 2**). Note that if $\sigma_{a_1}^2 = \sigma_{a_2}^2$, then

$$\hat{z}'_{t,k} = \hat{z}_{t,k}^{\sigma_{a_1}^2} = \hat{z}_{t,k}^{\sigma_{a_2}^2}.$$

We inferred activity values using both a low-variance prior ($\sigma_{a_1}^2 = 1$) and a high-variance prior ($\sigma_{a_2}^2 = 50$). We found this strategy for combining the results using two different variance prior parameters was more robust compared to using one specific parameter setting as it would reduce overfitting in cases when a single variance parameter was set to be too large, which sometimes led to unlikely high activating and repressive inferences in the same small region, while also reducing underfitting when a single variance parameter was set to be too small (**Supplementary Fig. 6**). We evaluated the enrichments of inferred highly activating and repressive nucleotides for conserved elements as a function of the variance prior parameters. We found them to be relatively robust across a substantial range of parameter settings, and in particular when using this strategy of using the more conservative inference from two substantially different settings for the variance prior (**Supplementary Fig. 6c**).

To obtain nucleotide regulatory activity scores, $b_{t,i}$, for each nucleotide $i = 0, \ldots, W - 1$ in the t^{th} tiled region we conducted piecewise linear interpolations between $\hat{z}'_{t,k}$ values (**Supplementary Note 3**).

The inference on the two designs was conducted separately, but they were combined for conducting the downstream analysis. The matrix operations were implemented using the library Apache Commons Math (https://commons.apache.org/proper/commons-math/) library v3.3. For the analysis on step sizes greater than 5 (**Supplementary Figs. 22** and **23**), we effectively applied the method here with a step size of 5, but treated entire positions of reporter tiles as missing.

Region and chromatin state scores. The region score for a tiled region t, denoted e_t was defined as $b_{t,i}$ (see above) where i is selected to maximize $|b_{t,i}|$ for $i = 0, \ldots, W-1$. The average region score for a chromatin state u, denoted s_u, based on matched data in a cell type, is the average value of e_t for all tiled regions t selected based on chromatin state u in the cell type.

Footprint, motif, transcription factor ChIP-seq, conservation and repeat data. The HepG2 and K562 CENTIPEDE elements were from ref. 5 obtained from http://centipede.uchicago.edu/. The footprints from ref. 7 were obtained from ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/supplementary/integration_data_jan2011/byDataType/footprints/jan2011/. The Wellington footprints in K562 were obtained from the supplementary data of ref. 31. The footprints of ref. 4 were the hg19 footprints obtained from http://fureylab.web.unc.edu/datasets/footprints/. The PIQ footprints[32] were the version 1 footprints obtained from http://piq.csail.mit.edu/ including both forward and reverse footprints. The motif instances were those provided by ref. 11. The ENCODE transcription factor binding peak call data sets used in the analyses of **Supplementary Figures 35** and **36** were the ENCODE2[6,54] uniform peak calls downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform which included 150 files in K562 cells and 77 for HepG2 cells. The conserved elements were the hg19 liftover of the SiPhy-PI[55] conserved elements from ref. 33. The repeats were based on RepeatMasker[56] obtained through the UCSC genome browser[57].

Motif analysis in the scale-up experiments data. The motif analysis comparing K562 and HepG2 cells (**Fig. 4**, **Supplementary Figure 25** and **Supplementary Table 2**) was computed based on averaging the $b_{t,i}$ values at the center of all instances for a motif. The P values were computed using a paired t-test over all instances tested using the Apache Commons Math library v3.3 implementation. The average motif score for the analysis in **Figure 6c** was based on just motif instances overlapping a tiled region selected based on HepG2 chromatin data when testing in HepG2 cells and likewise for K562 restricted to motifs with at least 20 such instances. The expected motif scores for these same set of instances were computed by permuting among tiled regions assigned to the same chromatin state and selected by the cell type of the measurements, which set of reporter values were assigned to which tiled regions. These permutations would preserve the same set of rows in a matrix where the rows correspond to reporter expression values and the columns the tile offsets. This was done for 1,000 permutations. For each motif the median average motif score across all permutations as well as the value of the 2.5% and 97.5% quantiles were recorded to form the expected motif values and 95% confidence intervals.

The P values for motif enrichment as an activator or repressor in **Figure 4c** and **Supplementary Figures 25**, **26** and **29** were computed based on one-sided binomial tests where the probability of success in the binomial distribution is the fraction of total nucleotides tested that had a regulatory score greater than or equal to the activation threshold for activators or less than or equal to the repression threshold for repressors. The number of trials is the number of instances of a motif with a center position overlapping a nucleotide tested. The number of successes is the number of instances of the motif with a center position having a regulatory score equal to or greater than the activation threshold for activators or less than or equal to the repression threshold for repressors. The P value threshold for defining activator and repressor motifs was an uncorrected P value of 0.01. In total, 1,934 motifs were tested. Motif instances that appeared on both strands at the same position were only counted once in the analysis.

Motif analysis in the pilot large-step experiments data. We defined four sets of sequences to analyze for motifs based on the pilot data. Two sets were defined based on adjacent pairs of tiles with significant differences at a 5% FDR in the HepG2 data, with one set corresponding to the sequences that on average had higher expression as determined based on the average ranks and the other set lower expression. The other two sets were based on the K562 data defined in the same way as for the HepG2 data. For each set of sequences we conducted motif analysis on the 30 bp that were unique to each sequence in the set compared to its corresponding adjacent tile plus ten additional base pairs into the common sequence. The motif enrichments with known motifs were computed using the program of ref. 11 modified so that the background set of motifs only included those overlapping sequences part of the array design. We ran *de novo* motif discovery using MEME[58] through the MEME suite with its default settings except requesting 10 motifs. The motifs were matched to a known motif using TOMTOM[59].

DeltaSVM comparison. For the comparison with important regulatory mutations predicted by the DeltaSVM approach in **Supplementary Figure 18a**, we identified a top 1% set of nucleotides associated with the maximum decrease in the sequence predicted to be regulatory when mutating the reference sequence. Specifically, we obtained the gkm-SVM 10-mer weights based on human ENCODE UW DHS from the website http://www.beerlab.org/deltasvm/ and used those in the files tup2_UwDnaseHepg2Aln_500_nc30_np_top10k_nsr1x1_gkm_1_10_6_3_weights.out and tup2_UwDnaseK562Aln_500_nc30_np_top10k_nsr1x1_gkm_1_10_6_3_weights.out for HepG2 and K562 cells, respectively. For each nucleotide tested, we computed the sum of the k-mer weights for the k-mers overlapping the nucleotide, which would be ten weights or fewer if the nucleotide was within the first or last nine nucleotides of the 295-bp region. We denoted this sum as s_{REF}. We also computed this sum for each of the three possible nucleotide substitutions to the reference sequence at the position denoted by s_{M1}, s_{M2} and s_{M3}. We ranked nucleotide position based on the extent to which they minimized $min(s_{M1}-s_{REF}, s_{M2}-s_{REF}, s_{M3}-s_{REF})$. The focus on the top 1% nucleotides was consistent with a percentage threshold used previously with DeltaSVM scores[12].

We also identified a top 1% set of nucleotides associated with the maximum increase in the sequence predicted to be regulatory when mutating the reference sequence (**Supplementary Fig. 18b**) using the same procedure as above except ranking nucleotides based on the extent to which they maximized the value of $max(s_{M1}-s_{REF}, s_{M2}-s_{REF}, s_{M3}-s_{REF})$.

Barcode k-mer analysis. For the analysis of barcode k-mer effect on inferred activity (**Supplementary Fig. 12**) we first define $e_{t,j}$ to be the barcode sequence for the t^{th} tiled region where $t = 1, \ldots, 15{,}720$, for the j^{th} reporter tile offset where $j = 1, \ldots, 31$, and $e_{t,j,p}$ the nucleotide at the p^{th} position in the barcode for $p = 1, \ldots, 10$.

For considering the occurrence of a k-mer sequence regardless of position within the barcode, we then defined for a k-mer sequence s the set $U_{s,j} = \{(t,p) \mid s = e_{t,j,p} \ldots e_{t,j,p+k-1}\}$, which gives all pairs of regions and barcode positions containing the k-mer sequence in the j^{th} reporter tile offset. We computed average inferred regulatory activity scores for all tuples (s, j, i) such that s is a sequence of length k found within at least one barcode sequence, $j = 1, \ldots, 31$ corresponds to one of the 31 reporter tile offsets, and $i = 1, \ldots, 295$ corresponds to one of the inferred activity positions. The average inferred regulatory activity average is then defined as

$$\frac{1}{|U_{s,j}|} \sum_{u \in U_{s,j}} b_{u.t,i} \tag{10}$$

where $u.t$ denotes the tiled region of u. We then ranked these averages to determine the cumulative distributions. To determine the expected distribution of these averages for this analysis we randomly reassigned each barcode sequence to reporter sequences, but still preserving the activity inferences based on the real assignments. We repeated the same analysis based on the randomized assignments. We did this for 400 randomizations of barcode assignments to reporter sequences and obtained 400 separate rankings of averages. For each rank in the ranking we determined the median value over the 400 randomizations and the 2.5 and 97.5$^{\text{th}}$ percentiles. This analysis was done separately for each value of $k = 1, \ldots, 6$.

Availability of data, software and Sharpr-MPRA scores. Raw data are available through GEO. Count data are available in **Supplementary Data 1** and **3**, from GEO, and http://www.biolchem.ucla.edu/labs/ernst/SHARPR/. The Sharpr-MPRA scores are available in text file, image and browser track formats from the website above and also in text format in **Supplementary Data 4**. The SHARPR software is also available from the website above, and the source code is maintained at http://github.com/jernst98/SHARPR.

doi:10.1038/nbt.3678

51. Melnikov, A., Zhang, X., Rogov, P., Wang, L. & Mikkelsen, T.S. Massively parallel reporter assays in cultured mammalian cells. *J. Vis. Exp.* **90**, 90, e51719 (2014).
52. LeProust, E.M. *et al.* Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res.* **38**, 2522–2540 (2010).
53. Bickel, P.J. & Doksum, K.A. *Mathematical Statistics: Basic Ideas and Selected Topics, Volume I, Second Edition*. (CRC Press, 2015).
54. Gerstein, M.B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91–100 (2012).
55. Garber, M. *et al.* Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, i54–i62 (2009).
56. Smit, A., Hubley, R. & Green, P. *RepeatMasker Open-3.0* (1996).
57. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
58. Bailey, T.L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol. ISMB Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28–36 (1994).
59. Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L. & Noble, W.S. Quantifying similarity between motifs. *Genome Biol.* **8**, R24 (2007).