PNAS
www.pnas.org

Supplementary Information for

Massively parallel discovery of human-specific substitutions that alter enhancer activity

Severin Uebbing, Jake Gockley, Steven K. Reilly, Acadia A. Kocher, Evan Geller, Neeru Gandotra, Curt Scharfe, Justin Cotney, James P. Noonan

James P. Noonan
Email: james.noonan@yale.edu

This PDF file includes:

Supplementary text
Figures S1 to S15
Tables S1 to S5
Legends for Datasets S1 to S4
SI References
Extended Methods

Other supplementary materials for this manuscript include the following:

Datasets S1 to S4

Supplementary Information Text

Section 1. Target Selection and Initial MPRA Library Design

We selected genomic regions with potential human-specific regulatory activity by taking all Human Accelerated Conserved Noncoding Sequences (1) and Human Accelerated Regions (version 2) (2) (collectively referred to as HARs) and all Human Gain Enhancers, that is, regions that show markedly higher ChIP-seq H3K27ac or H3K4me2 signal in human compared to mouse and rhesus macaque during corticogenesis (3) (Dataset S1). Within these regions in human genome GRCh37/hg19, we collected all human–chimpanzee substitutions, that were fixed for the chimpanzee state in a primate alignment (comprising chimpanzee panTro4, orangutan ponAbe2, rhesus macaque rheMac3 and common marmoset calJac3) and likely to be fixed or nearly fixed in human populations (not present as a SNP marked as "Common" in dbSNP (4) build 144, excluding indels). This resulted in a list of 32,776 human-specific Substitutions, or hSubs. We then designed 137 bp fragments centered on these hSubs. The fragment length was dictated by the maximal high-quality synthesis capacity at the time of 170 bp and the need to include primer sites for cloning. In case two hSubs were less than 137 bp apart, we designed additional fragments centered on the midpoint between those hSubs. This way, we designed 49,629 experimental fragments per species. We included 639 sequences per species from the same enhancer regions that did not contain any hSubs, for a total of 100,536 fragments. The final fragment library was synthesized by CustomArray.

For a more detailed description of the experimental procedure, see *Extended Methods* below. Briefly, after synthesis, the fragment library was first amplified in a low-cycle PCR to complement the single synthesized strand. In a second step, fragments were tagged with random, 16 bp barcode tags in an emulsion PCR. After transformation into *E. coli*, we selected a library with a complexity of about 80 barcode tags per fragment. The primer tails also contained restriction sites for cloning into the pMPRA1 vector and for cloning the *luc2* firefly luciferase ORF from pMPRAdonor2 in between fragment and barcode (Addgene plasmids #49349 and #49353 provided by T. Mikkelsen). We cloned the library into pMPRA1 and sequenced this "inert" library on an Illumina HiSeq 2500 (2x250 bp) to establish fragment–barcode connections (see *Section 2*). We then cloned the luciferase ORF in between the fragment and the barcode to obtain the "competent" library used for the MPRA experiment.

Section 2. Cell Culture, Transfection, Harvest, Sequencing and Data Preparation

We used H9-derived human neural stem cells (hNSCs; GIBCO #10419428). These cells show a gene expression profile consistent with an early neurodevelopmental state, expressing HES1, PAX6, NES and SOX2 according to RNA-seq (5). They are multipotent and capable of differentiation into neurons, oligodendrocytes and astrocytes. hNSCs were grown in DMEM/F12 (GIBCO #12660012) with GlutaMAX (GIBCO #35050061), StemPro Neural Supplement (GIBCO #A1050801), bFGF (GIBCO #PHG0021) and EGF (GIBCO #PHG0311) added in flasks coated with Matrigel (Corning) at 37°C and 5% CO_2. Cells were passaged using StemPro Accutase (GIBCO #A1110501) every 48 hours. We confirmed a uniform neural stem cell identity using immunofluorescent staining for SOX2 (Millipore #AB5603) and Nestin (Millipore #MAB5326) in parallel to transfection (Fig. S10).

To perform the MPRA experiment, the library was transfected into hNSCs using a Nucleofector 2b (Lonza #AAB-1001). We used the plasmid's luciferase activity to determine optimal experimental conditions (Fig. S11, Extended Methods). Under final experimental conditions, we transfected four replicates (two each at hNSC passages 12 and 18) of five times 4 million cells with 32 µg of library per replicate and incubated cells for 6 hours before harvest with Accutase.

We used 95% of the harvested cells for total RNA extraction and the rest for plasmid DNA (pDNA) extraction. Initially, genomic DNA was extracted using the Qiagen AllPrep Kit, after which pDNA was enriched using the Qiagen MinElute PCR Purification Kit. RNA was purified using the Qiagen RNeasy Mini Kit including two DNase digests to minimize pDNA carryover. RNA integrity numbers were obtained using a 2100 Bioanalzyer (Agilent) and were high (> 8) in each replicate. We synthesized cDNA with the Invitrogen SuperScript III reverse transcription kit using a mixture between a custom, library-specific primer and oligo-dT primers (Table S5). qRT-PCR was performed to estimate the optimal number of PCR cycles to amplify the library for sequencing, so as to stop amplification in the exponential phase of PCR to avoid over-amplification (Fig. S12).

pDNA and cDNA libraries were sequenced on an Illumina HiSeq 4000 (2x150 bp). Inert library sequencing reads (containing fragment and barcode) were trimmed of low-quality bases using FastX Trimmer (options -Q 33 -f 10) and assembled into read pair contigs using PEAR (-q 33). Outer adapter sequences were removed using Cutadapt (-e 0.16) (6). Finally, the inner adapter was removed with Cutadapt, leaving barcode tag and fragment sequence. Fragment sequences were mapped to the original library using Bowtie2 (--rg-id) (7) to generate a map connecting barcode sequences with fragment identity. Experimental pDNA and cDNA sequencing reads (containing the barcode) were trimmed using FastX Trimmer (-Q 33 -l 131 for read one, -Q 33 -l 126 for read two), oriented correctly (see scripts) and read pairs were assembled using PEAR (-q 33). Adapter sequences were removed using Cutadapt (-e 0.16) (6). Resulting barcode sequences were summed up and annotated using the barcode–fragment map. Once summarized, data was analyzed using R (8). All of our scripts are publicly available (see *Data and code availability* in the main text).

Section 3. Library Design and Experimental Procedures for the Second MPRA

Based on the results of the first MPRA, we designed a second MPRA library with two major components. First, we included all active fragments for replication (2,704 orthologous fragment pairs). Second, for all differentially active fragments, we designed artificial fragments that included all possible combinations of hSub states on both human and chimpanzee background sequences (14,963 fragments). In 972 loci (i.e., differentially active fragment pairs from the first MPRA), this library covered 1366 hSubs. After designing this library, we found 278 more “hSubs” that overlapped with “common” dbSNP (build 144) (4) SNPs, which we excluded post hoc. All numbers given in the article exclude those SNPs. This library also contained additional positive and negative controls. Positive controls included fragments from enhancers that were previously reported to actively drive transcription in hNSCs (enhancers regulating *NES*, *SOX2* and *TLX*) (9–11) and high confidence hNSC candidate enhancers overlapping phastCons elements (12). Negative controls were generated by permuting enhancer sequences to random, inactive regions in the genome and then permuting phastCons elements within those enhancers to generate ‘pseudo-phastCons elements’ in ‘pseudo-enhancers’ on which MPRA fragments were designed. The library contained a total of 24,543 unique fragments and was synthesized by Agilent.

The experiment was performed as before with the following changes. We used 17 bp barcode tags. We synthesized cDNA using a custom, library-specific primer only, with comparable results to the first MPRA. We also included an additional no-RT control to ensure pDNA carryover contamination was at acceptably low levels.

Libraries were sequenced paired-end for 150 cycles on an Illumina NovaSeq 6000. Mapping and summarization was done as in the first MPRA with the following changes. We mapped fragment sequences including their surrounding adapter sequences to force full-length mapping of highly similar sequences. We only accepted matches with >99% sequence identity and removed all barcode tags that were duplicated.

Section 4. Enrichment of Transcription Factor Binding Sites

We downloaded all vertebrate transcription factor binding site (TFBS) models for transcription factors (TFs) expressed in hNSCs from the JASPAR core database (version 2014) (13) and predicted binding sites in the human (GRCh37/hg19) and chimpanzee (panTro4) genomes using FIMO (option --max-stored-scores 10000000) (14). We then overlapped the union of human and chimpanzee TFBS predictions with measured, active, differentially active, human-biased and chimpanzee-biased fragments and tested for enrichment in each set of elements using the next largest set as a background (i.e., enrichment within differentially active fragments compared to all active fragments, *et cetera*) using 10,000 resampling replicates. We also overlapped predicted human and chimpanzee TFBSs with all 1,183 hSubs in active MPRA fragments and tested for an enrichment of TFBSs within the human or chimpanzee TFBS predictions relative to the union of both. The number of TFBSs overlapping hSubs within differentially active fragments was not sufficient for an informative enrichment test. However, since overall TFBS content did not differ significantly between active and differentially active fragments, comparing differences in TFBS representation between the human and chimpanzee sequences of active fragments may identify changes in TFBS recruitment that alter enhancer activity. *P* values were BH-corrected.

Additionally, we restricted our analysis to TFBSs present in accessible chromatin sites in hNSCs and cortical organoids (ATAC-seq from this study, ENCODE DNase-seq in H9-derived hNSCs and ATAC-seq in iPSC-derived, human and chimpanzee cortical organoids) (15,16). We could only detect enriched TFBSs in active fragments over all tested fragments but not in all other comparisons, likely because low power due to the smaller sample size.

Section 5. Comparison of our test results to other MPRA analysis methods

During the work on this project, several methods for the analysis of MPRA experiments were published. Here, we sought to analyze our dataset with those methods to compare our testing procedure to those methods. We could not find a method that appropriately models every aspect of our study design. We therefore had to approximate some aspects of the analysis to be able to obtain comparable results. We used two published methods (17,18) and a third approach that was used in a study of similar candidate enhancer sequences (19).

Study design

Our MPRA library contained about 50,000 fragment pairs of human and chimpanzee sequence origin. Each of the 100,00 single fragments were tagged with a random number of barcodes. In the final set of analyzed fragments, barcode number per fragment ranged from 12 (lower bound cutoff) to 1,320 with a mean of 69.3 barcodes per fragment. The entire library was tested in four experimental replicates.

Our study comprised a two stage analysis procedure. First we tested for fragment activity. Only if a fragment pair was active in at least one of the alleles, we proceeded to test for differential activity.

mpralm

mpralm (17) is a method for the detection of differential activity between the different alleles of an MPRA fragment. It is not designed to test for activity of fragments. The method uses the variance information from experimental replicates. Variance information from different barcodes, however, is not modeled but counts from different barcodes are summarized (either summed or averaged) prior to analysis. To compare this method to our results, we first tested our entire dataset for differential activity using mpralm and then filtered out those results that were not found active in our analysis (according to the permissive criteria, see Methods). Significance was declared using BH-corrected $P < 0.05$ (corrected first on the whole dataset and then separately on the reduced dataset).

Over the entire dataset, mpralm detected 5,268 significantly differentially active (DA) fragment pairs. We note that 27.7% DA fragment pairs had at least one replicate that disagreed in the direction of activity bias with other replicates.

When considering only those 3,219 fragment pairs that we tested for differential activity in our analysis, mpralm found 1,764 to be DA. Of these, 15.1% disagreed in the direction of bias between replicates. Consistent with the analyses presented in Myint *et al.* (17), our method was conservative compared to mpralm and our results are essentially a subset of the results of mpralm. Of our 673 significantly DA fragment pairs, mpralm also found 672 (99.9%).

MPRAnalyze

MPRAnalyze (18) is more flexible in its design and is capable to test fragments for activity and fragment pairs for differential activity. It considers both the variance from technical replicates and from barcodes and different fragments can have different numbers of barcodes. However, model fitting becomes computationally demanding with increasing number of barcodes. At the full barcode number in our dataset (>1,300), model fitting took more than 24 hours per fragment, after which we aborted the run. Assuming it took 24 hours per fragment, the time needed to finish the analysis for the entire dataset would exceed 100 core years on our cluster. We therefore subsampled fragments to 80 barcodes each and used this subset, which computed in about three days. We used this dataset to test for fragment activity in the human and chimpanzee alleles, and for differential activity between the two alleles. Similar to above, we first tested for differential activity over the entire dataset before we restricted the test to only those fragment pairs that were also tested in our analysis.

In the entire dataset, MPRAnalyze found 5,448 active fragments. Of those, 1,058 were also found active in our analysis. Only 6.8% of fragments were found to be active in both human and chimp, contrasting somewhat with the 27.8% of fragments that we found active in both human and chimpanzee in our original analysis.

Testing for differential activity over the entire dataset, MPRAnalyze detected 6,949 fragment pairs to be differentially active. Only 2,275 of those fragment pairs contained an active fragment according to the activity test of MPRAnalyze. Of the 6,949 DA fragment pairs, 64.4% disagreed in the direction of bias between replicates.

Among only the 3,219 fragment pairs that we tested for differential activity in our analysis, MPRAnalyze found 2,381 to be differentially active. Of these, 44.7% disagreed in the direction of bias between replicates. Of our 673 significantly DA fragment pairs, MPRAnalyze also found 451 (65.8%).

Ryu *et al.*

Ryu and colleagues (19) defined an MPRA fragment as significantly active if it was above the 75th percentile in all three of their technical replicates and in at least two of the four cell lines used (2 human and 2 chimpanzee cell lines). To apply this approach to our dataset, we defined a fragment as active if it was above the 75th percentile in all our four experimental replicates.

This way, we identified 7,684 fragments as significantly active. Of those, 3,155 were also found active in our original analysis. 1,967 fragments (34.4%) were found to be active in both human and chimpanzee. When summarizing fragments into HARs, all 51 of the HARs found active by Ryu *et al.* (19) and measured in our MPRA were found active using this approach applied to our data (meaning that each of those HARs contained at least one active fragment).

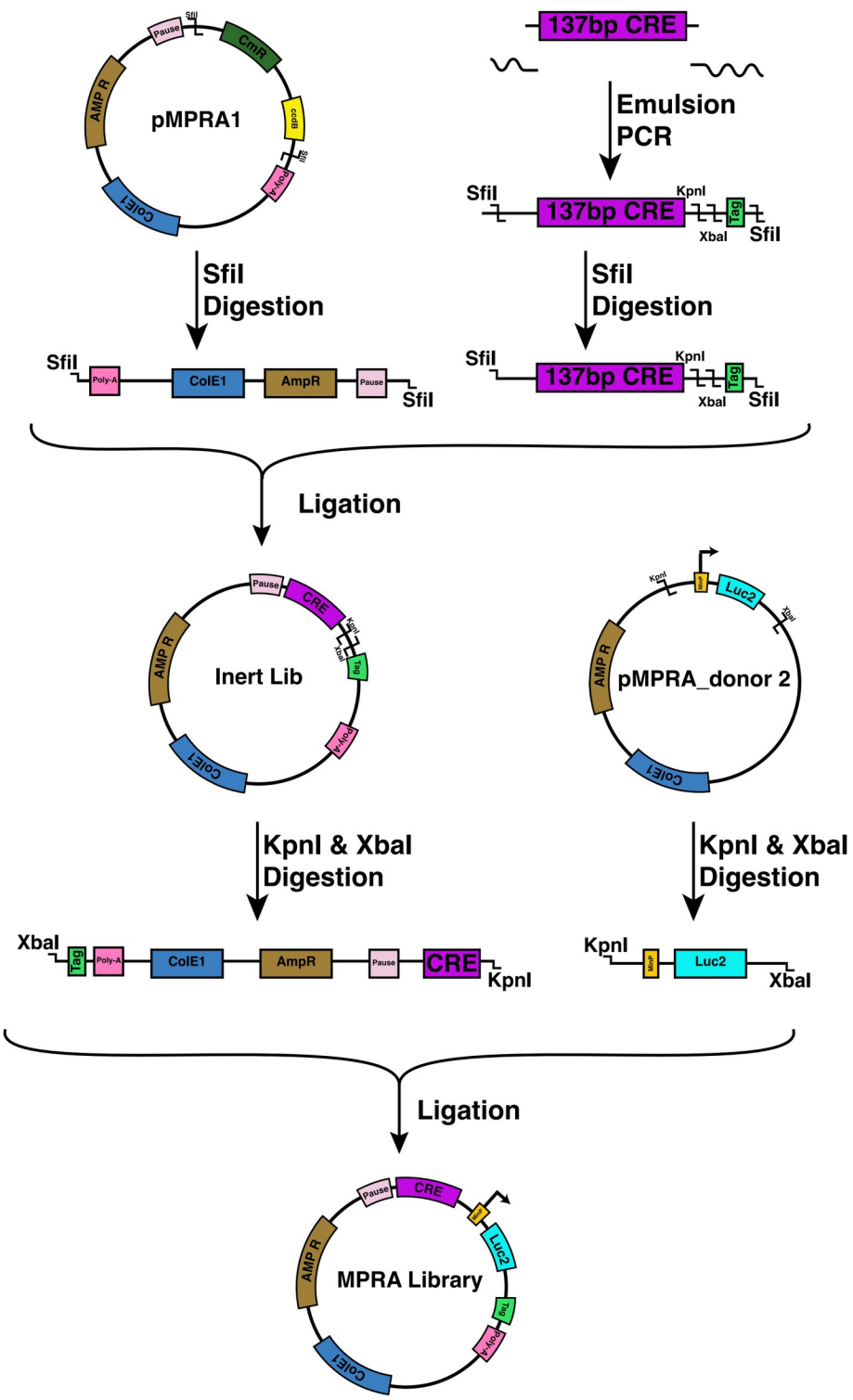

Figure S1. Cloning scheme to generate the MPRA library. CRE indicates the MPRA fragment (Candidate Regulatory Element). The location of the barcode tag is shown as a green box. See Supplemental Methods for details.

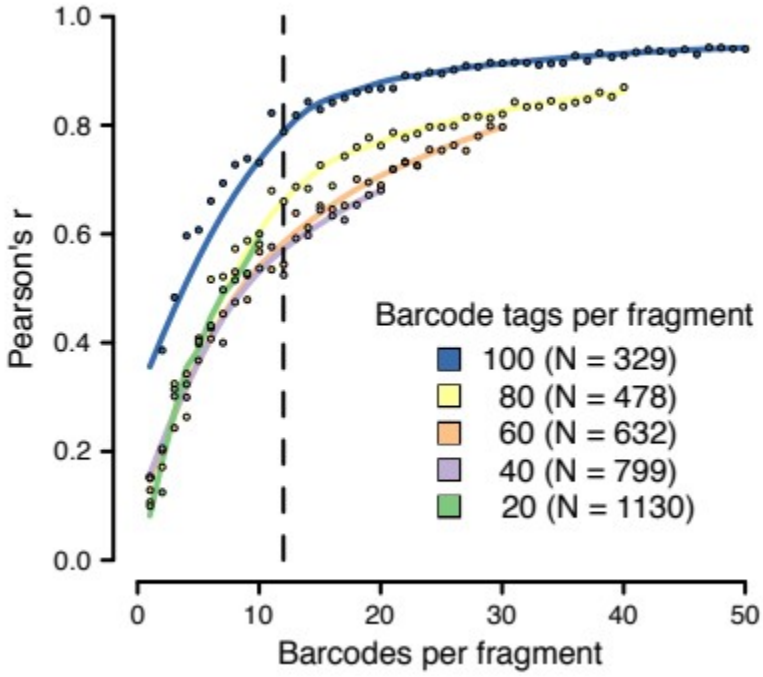

Figure S2. Correlations between replicates vs. barcode number. For all fragments having exactly N = 20, 40, 60, 80 or 100 barcode tags, two groups of barcodes per fragment were downsampled as N/2, (N/2) – 1, … , 2 barcode tags per group. Then the mean activity per fragment was calculated and Pearson's r was calculated between two sample replicates. By manual inspection of the inflection point at which the increase in correlation leveled off, we determined that fragments with twelve or more barcodes showed sufficiently high correlations to be included in downstream analyses.

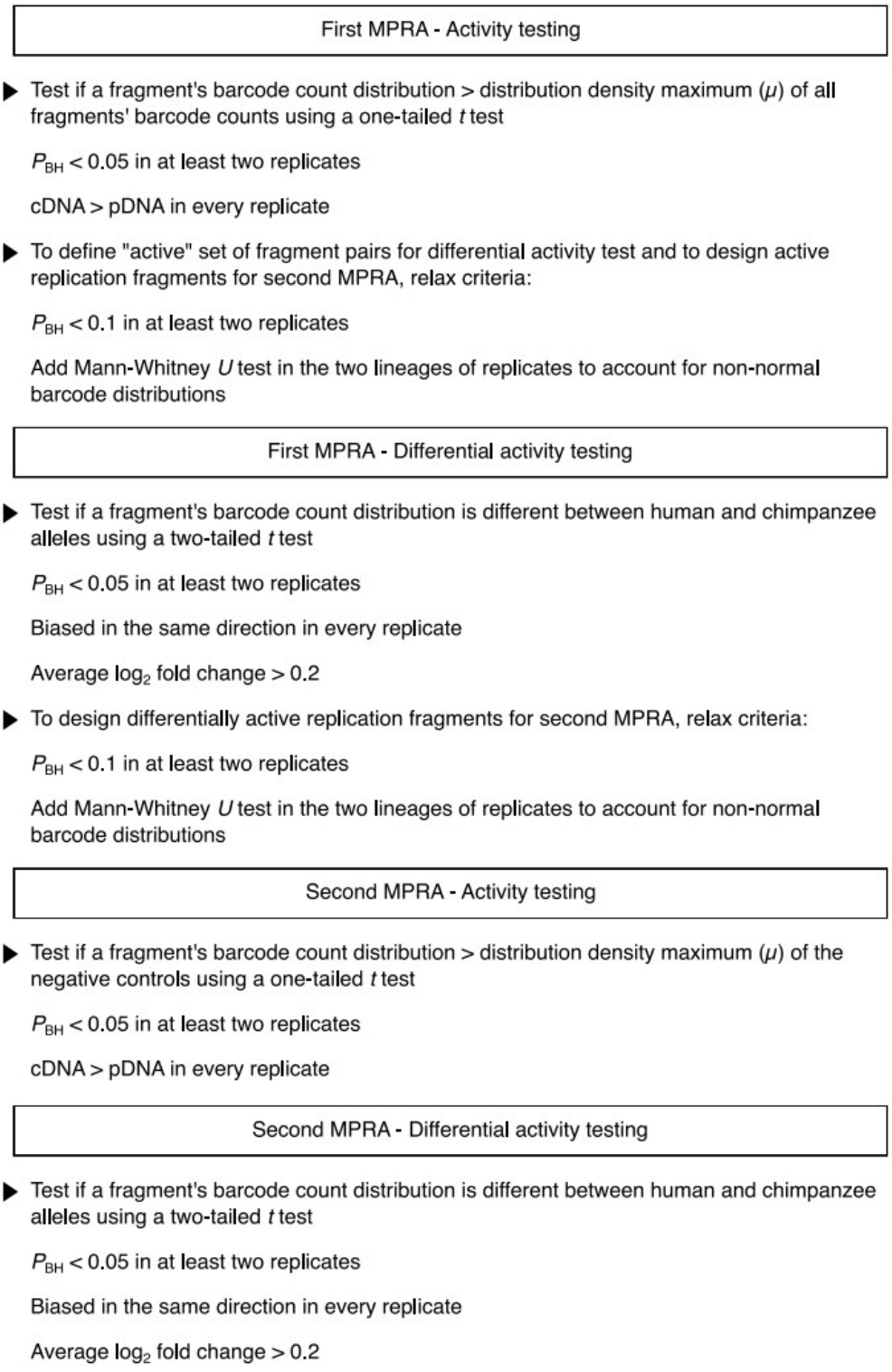

Figure S3. Overview over activity test procedures. To produce reported numbers, testing was done with the same criteria in both MPRAs. We relaxed stringency to design input fragments for the second MPRA and to generate the set of active fragment pairs for the differential activity test.

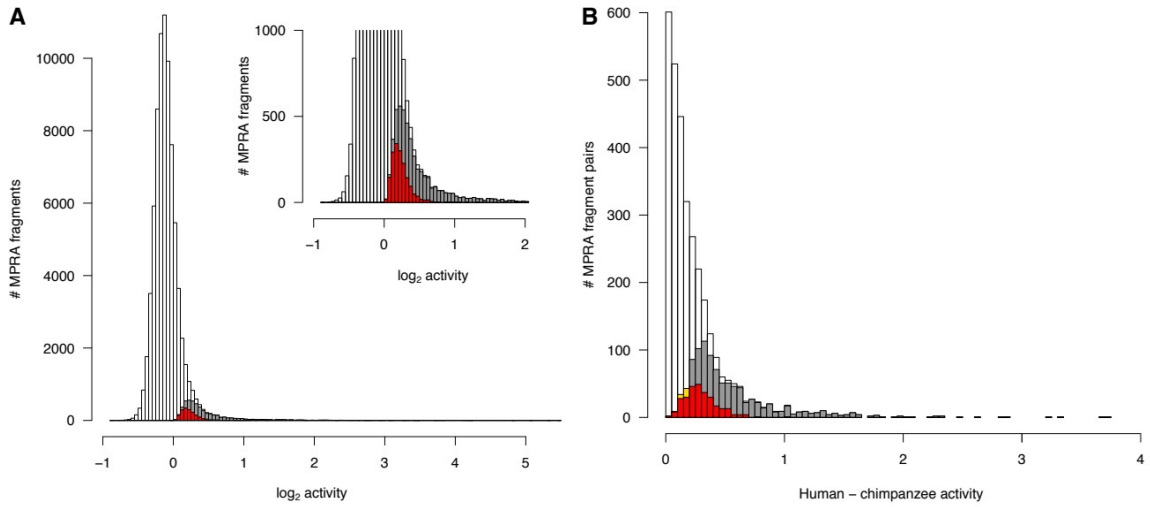

Figure S4. Fold change cutoff choice. (**A**) Histogram of $\log_2$ activity, with significantly active fragments indicated in gray. Red bars show fragments that were deemed active by *t* test but excluded, because they showed smaller cDNA than pDNA values in one or more replicates. The inset shows a magnification of the area of most active fragments. (**B**) Histogram of absolute $\log_2$ fold change between human and chimpanzee. Significantly differentially active fragment pairs are in gray. Red bars show fragment pairs that were deemed significantly differentially active in our *t* test but were excluded because replicates disagreed in their directionality. Yellow highlights 20 additional fragment pairs that were excluded because their average $\log_2$ fold change across replicates was < 0.2.

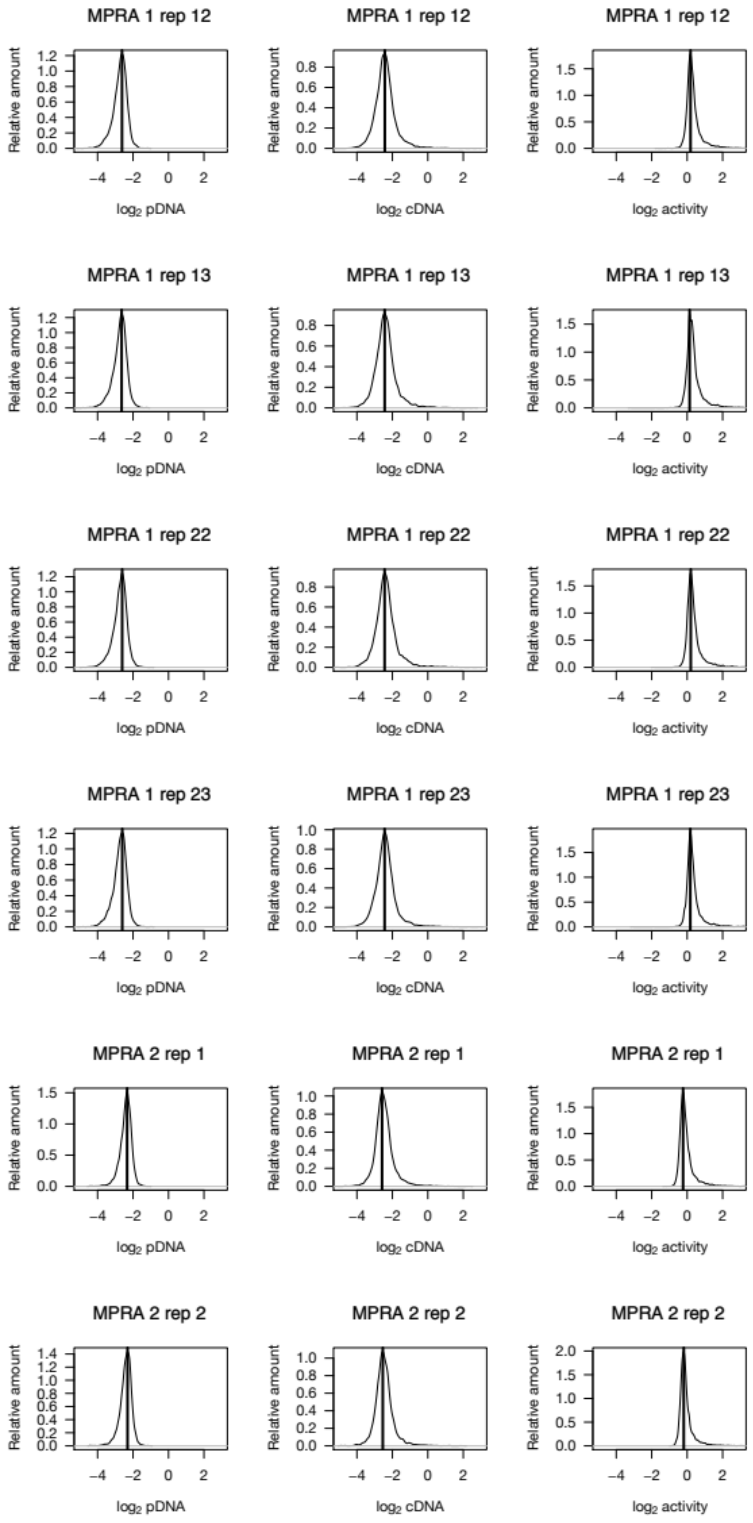

Figure S5. Density distributions for median fragment pDNA (left column), cDNA (middle) and activity (right) values for all four replicates of the first (row 1–4) and second MPRA (row 5–6). Note the following observations. 1) A lack of low activity values, preventing the detection of significantly low activity. 2) The markedly low activity in the second MPRA due to a combination

of high pDNA and low cDNA values. The high pDNA values are due to the fact that few unique fragments were transfected in the second MPRA compared to the first one, with the consequence that the sequenced pDNA reads (which were sequenced to similar depths in both MPRAs) are distributed over fewer unique fragments, leading to a higher average pDNA value per fragment in the second MPRA compared to the first. Similarly, because the transfected fragments were more active in the second MPRA than in the first, the cDNA reads (sequenced to similar depth between MPRAs) are being distributed over more fragments in the second MPRA than in the first, leading to a lower average cDNA count value in the second. Due to these differences in pDNA, cDNA and activity, absolute values are not comparable between different MPRA experiments.

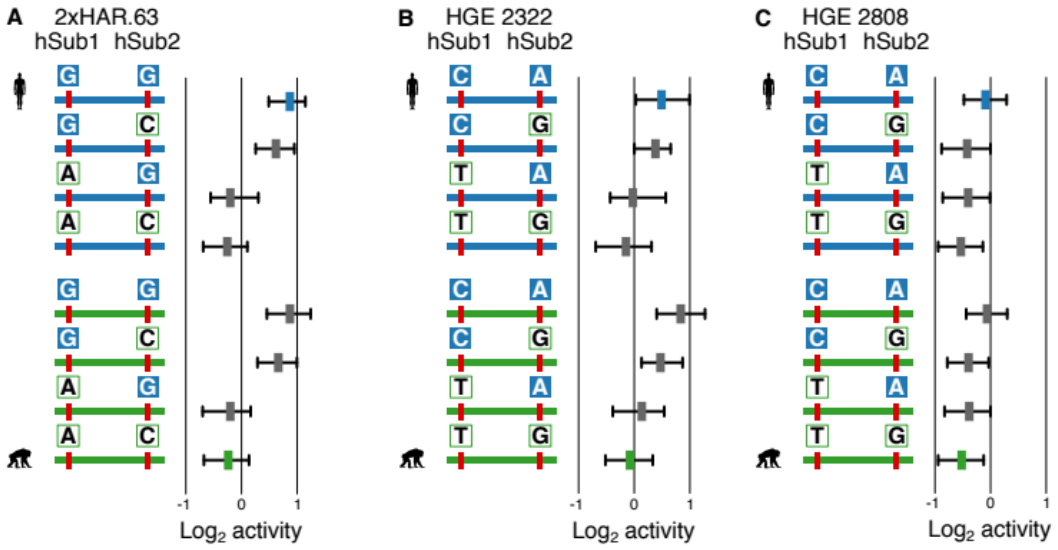

Figure S6. Three examples of combinatorial loci. Sketches (left) show the allele state. Filled blue boxes indicate the human hSub state, open green the chimpanzee state. Allele line color indicates background variation (blue, human; green, chimpanzee). Activity plots (right) show median and quartile ranges of fragment activity. (**A**) 2xHAR.63 contains the most human-biased locus with significant hSub effects in our dataset. The locus harbors four hSubs, two of which significantly affected regulatory activity in an additive manner. In addition, there was a slight but significant effect of background variation on activity. (**B**) HGE 2322 contains a locus that exemplifies visually additive hSub effects. This locus contains three hSubs, two of which showed significant effects on regulatory activity. Each of the two hSubs shown increases activity in the human allele state. Together in the same allele, these effects combine additively to produce the expected summed effect of both hSubs. There was no significant effect of the background variation. (**C**) HGE 2808 contains a locus that may serve as an example for an interactive effect. The locus contains three hSubs, two of which showed an effect on activity. Each of these two hSubs increases regulatory activity of the human allele slightly, but the combination of the two human states in the same allele increases activity more than expected by summing up their individual effects. The background variation present in this locus had no effect on activity.


Effect size [SD]
-6
-4
-2
0
2
4
-6
-4
-2
0
2
4
Effect size [SD]

Figure S7. Expanded version of Figure 4D showing all 401 fragments with significant hSub effects. Fragments are sorted along the Y axis according to their fragment effect size from most human-biased to most chimpanzee-biased. Additive (blue) and interactive (orange) hSub effects are shown as dots in the same column.

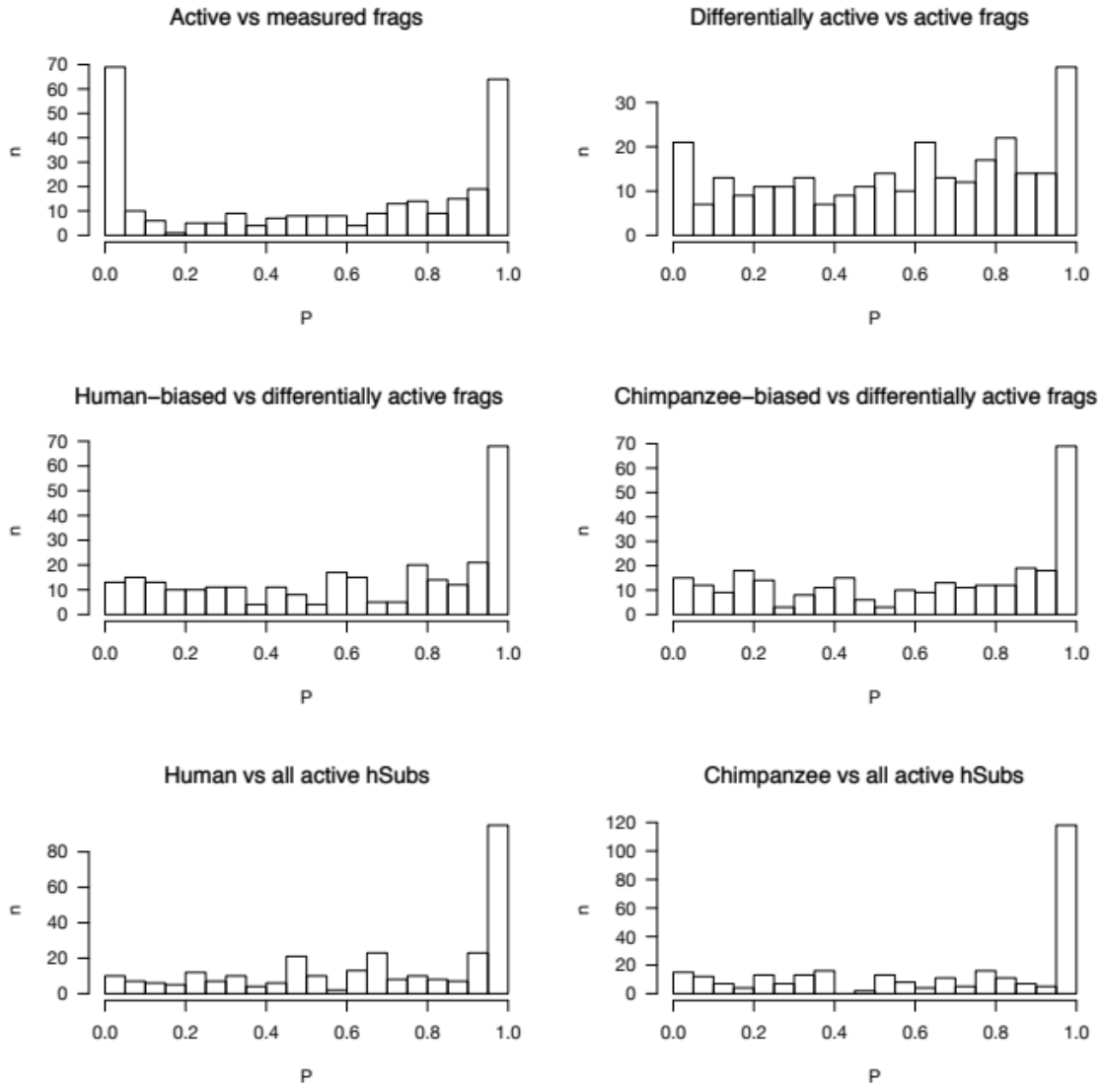

Figure S8. Uncorrected *P* value distributions for TFBS enrichment tests. See main text for details.

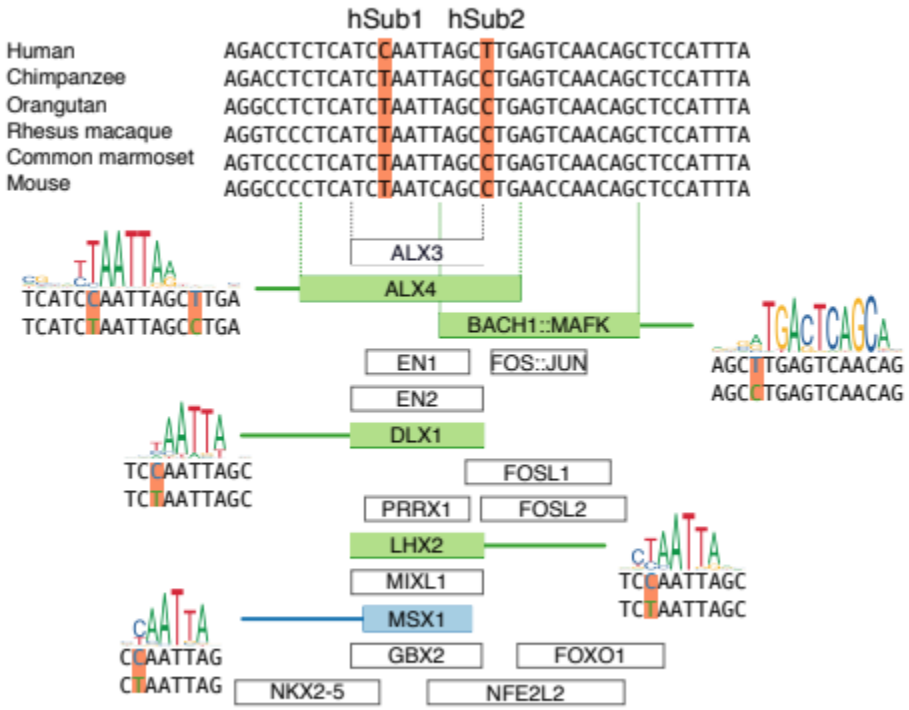

Figure S9. Extended Figure 6C showing all predicted TFBSs. Sites with stronger predicted binding in human than in chimpanzee are shown in blue, sites with stronger binding in chimpanzee are green, and sites predicted with similar binding strengths in both sequences are white.

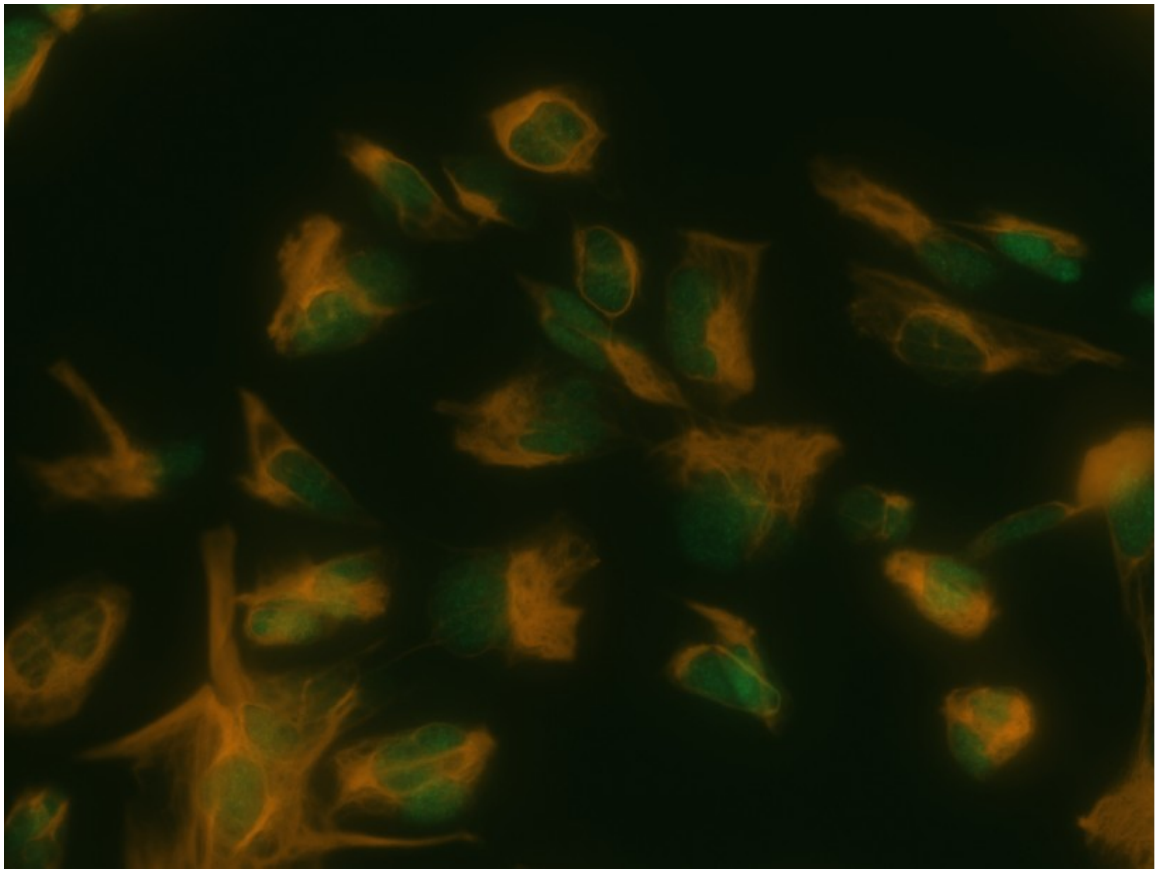

Figure S10. Nestin and SOX2 immunostaining confirms neural stem cell identity. Staining was performed in parallel to each MPRA transfection.

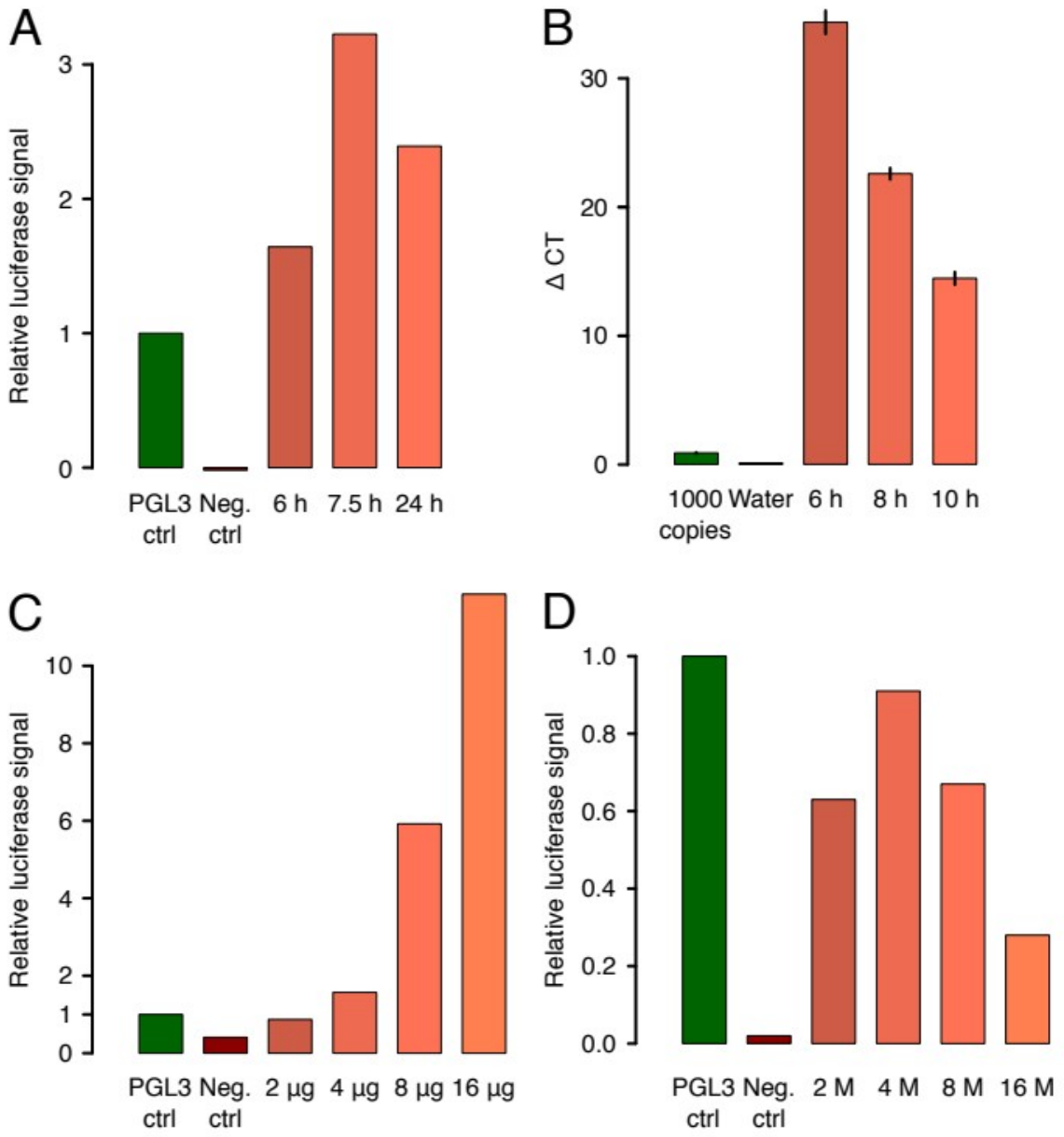

Figure S11. Optimization of transfection parameters based on the luciferase activity of the MPRA plasmid's luc2 reporter gene. (**A**) Luciferase signal after varying incubation times. We had previously established an incubation time of < 12 hours as most effective. (**B**) qRT-PCR of the transcribed luc2 gene to establish the optimal incubation time to harvest mRNA. This time is shorter than the time to maximal luciferase activity, potentially due to a translation time lag. (**C**) Transfecting varying amounts of plasmid per two million human neural stem cells. (**D**) Transfecting 16 µg of MPRA library with different numbers of cells to transfect. Note that the PGL3 control is extremely active in this experiment, while the negative control better reproduces previous results. Luciferase samples were standardized using 15 ng of co-transfected Renilla luciferase and luciferase signal is normalized relative to a vector containing the PGL3 promoter. In **C–D**, cells were grown for 24 h after transfection.

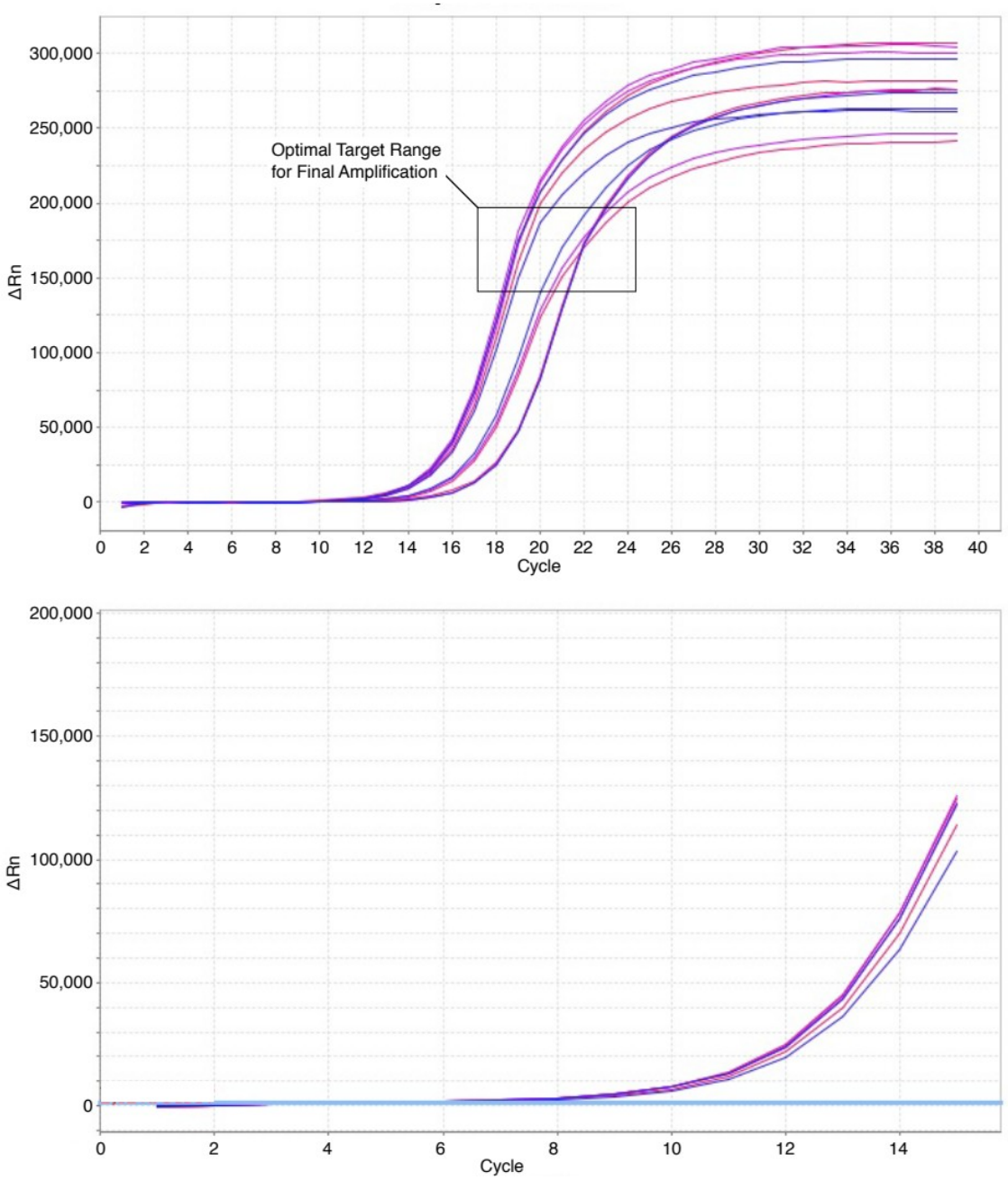

Figure S12. qRT-PCR to estimate optimal cycle number for experimental barcode amplification for sequencing. (Top) Amplification cycle number is first estimated in a full-length qRT-PCR run using a 10x dilution of pDNA and cDNA libraries. Given the target range, about 15 cycles should be sufficient for the undiluted library. (Bottom) qRT-PCR run of the undiluted library using 15 cycles shows amplification into the exponential phase. See Supplemental Methods for details.

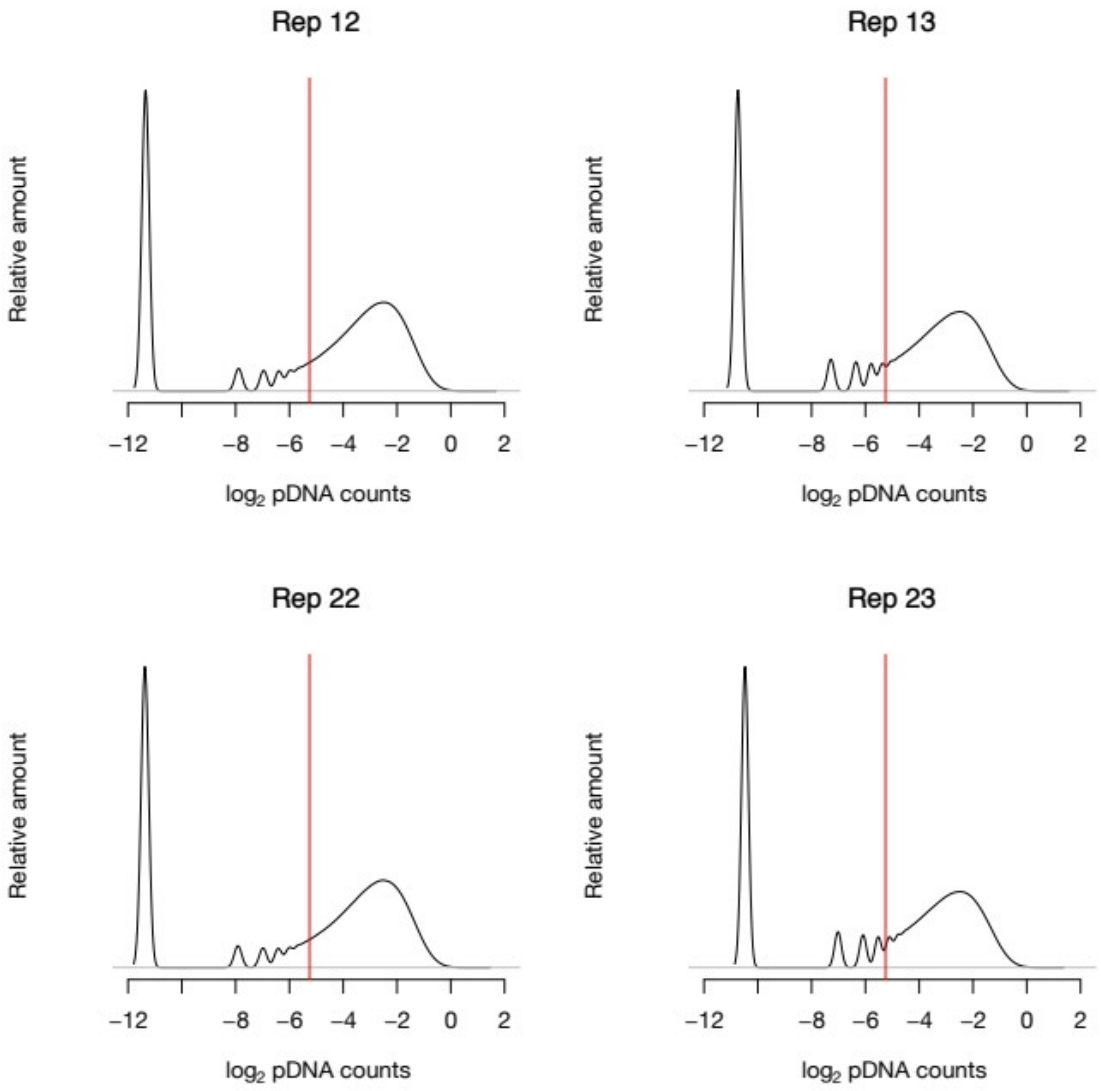

Figure S13. Density plots of $\log_2$ pDNA barcode counts showing Poisson error at the small range. We excluded all barcodes with a $\log_2$ count below -5.25, indicated by the red vertical line.

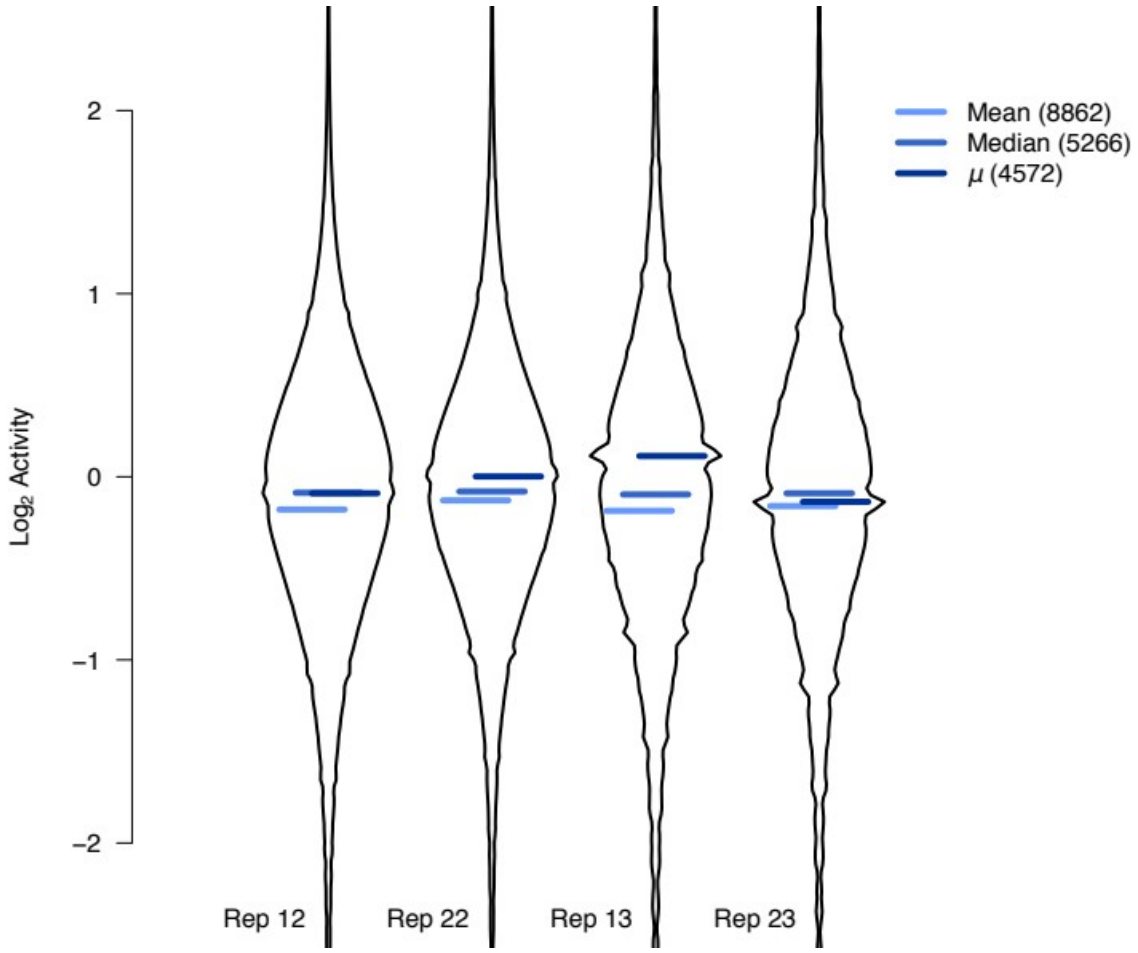

Figure S14. Comparison of three different summary statistics, Mean, Median and μ. Numbers in the legend refer to unique fragments identified as active by the *t* test shows that the results from our line of tests is dependent on the used summary statistic. (Note that the numbers reported in the main text are smaller due to additional constraints applied: Fragments had to have cDNA > pDNA in all replicates and an average activity >0.1.) We used μ as a measure of the distribution mean in our *t* test in the first MPRA because it was the most conservative and because it accounts best for the ragged activity distribution of replicate 13.

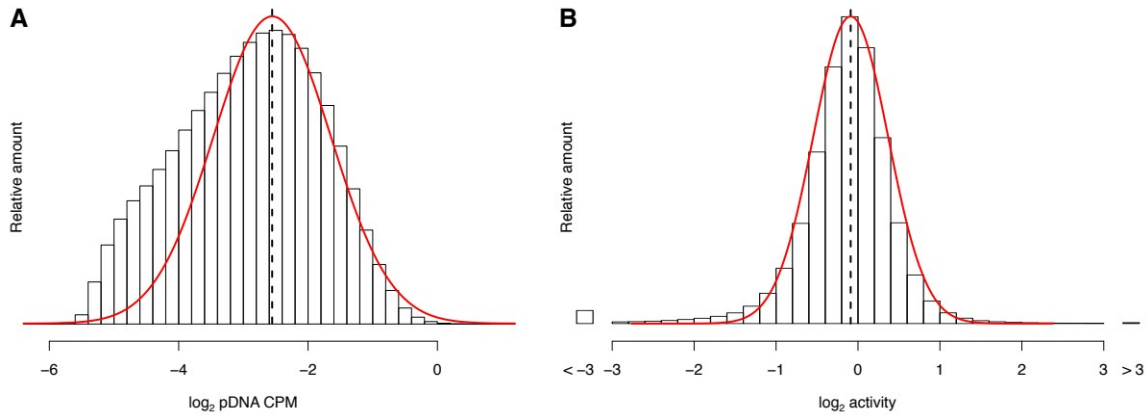

Figure S15. MPRA data is approximately log-normally distributed. (**A**) Distribution of the pDNA fraction from the first round of MPRA. Note that the lower end of the distribution is affected by the lower bound of detection (see Figure S13) and does not allow testing for fragments that show less representation in cDNA fraction vs. pDNA, which would indicated transcriptional repression. (**B**) Distribution of the activity from the first round of MPRA. The red line in both panels shows a normal density distribution simulated with mean and variance taken from actual data.

Table S1A. Statistics on the number of active and differentially active MPRA fragments in the two species and according to different test approaches.

Strict criteria[1]	**Inactive**	**Active**		**Total**
human	36,660	1,542		38,202
chimp	38,648	1,637		40,285
overlap[2]	32,722	1,392		34,114
union[3]	31,608	2,506		34,114
total	75,285	3,202		78,487
Permissive criteria[4]				
t test	73,941	4,546		78,487
U test	74,945	3,542		78,487
overlap	75,317	3,170		78,487
total	73,569	4,918		78,487
	Inactive pairs	**Active pairs**	**Differentially active**	**Total**
Paired differential activity test[5]	30,895	3,219	673	34,114

[1]Strict criteria are used for statistics
[2]The overlap of active fragments are those active in both orthologs. The number under "Inactive" indicates those that are measured in both species but active in no more than one.
[3]The union of active fragments are fragments active in either human, or chimpanzee, or both. The union of inactive fragments are those measured in both species but active in none.
[4]Active fragments identified using these criteria are the input for differential activity testing and the design of the second round of MPRA.
[5]The difference between these union number and that above is due to the fact that we used the permissive criteria to include more fragment pairs in the differential activity test. The differential activity test itself follows strict rules.

Table S1B. Test result replication between the two rounds of MPRA.

Active in round 1	3,203
Tested in round 2[6]	3,030
Active in round 2	2,111 (69.7%)
Differentially active in round 1	673
Tested in round 2	627
Active in round 2	559 (89.2%)
Differentially active in round 2	374 (66.9%)

[6]The set of fragments active in round 1 and measured again in round 2. All active fragments from round 1 were included in round 2 but not all had sufficient barcode coverage (>11).

Table S2. Overlap of inactive, active and differentially active MPRA fragments with ATAC-seq and H3K27ac ChIP-seq peaks measured in hNSCs. The second column shows the overlap of fragments with accessibility data from three datasets, ATAC-seq (this study), ATAC-seq for cortical organoids (16) and ENCODE DNase-seq for H9-derived NSCs (15).

	ATAC overlap	ATAC overlap 3 datasets	ChIP overlap
Inactive	2,121 (5.2%)	6,879 (16.7%)	10,308 (25.0%)
Active	332 (13.2%)	623 (24.8%)	782 (31.2%)
Differentially active	66 (9.8%)	144 (21.4%)	192 (28.5%)
All fragment pairs	2,453 (5.5%)	7,502 (17.2%)	11,090 (25.0%)

Table S3. hSub effect size distribution for additive and two-way interactive effects of human- and chimpanzee-biased hSubs. Because hSubs can be measured in more than one fragment both mean and maximum are given.

	max effect size		**average effect size**	
	Additive	**Interactive**	**Additive**	**Interactive**
Human-biased				
SD < 0.01	0	0	0	0
SD > 0.01	9	1	14	3
SD > 0.1	106	26	106	32
SD > 0.5	37	31	34	25
SD > 1.0	12	9	11	7
SD > 2.0	1	3	0	3
Max SD	2.26	3.36	1.79	3.36
Total	165	70	165	70
Chimpanzee-biased				
SD < 0.01	0	0	0	0
SD > 0.01	17	2	22	2
SD > 0.1	84	22	82	26
SD > 0.5	32	14	31	11
SD > 1.0	14	8	13	9
SD > 2.0	3	4	2	2
Max SD	2.90	4.01	2.76	4.01
Total	150	50	150	50
Grand total	315	120	315	120

Table S4. GO enrichment test results for genes interacting with differentially active MPRA enhancers. Enrichment test was performed using topGO (20) using standard options and 3C-type connectivity data was taken from (21–24). *P* values are uncorrected for multiple testing according to recommended use (20).

Table S4A Biological Process ontology

GO ID	Term	Annotated	Significant	Expected	*P* value
GO:2000737	negative regulation of stem cell differentiation	18	4	0.25	0.00010
GO:1903707	negative regulation of hemopoiesis	98	5	1.39	0.00015
GO:0034616	response to laminar fluid shear stress	14	3	0.2	0.00091
GO:0060216	definitive hemopoiesis	14	3	0.2	0.00091
GO:0045668	negative regulation of osteoblast differentiation	34	4	0.48	0.0013
GO:0060560	developmental growth involved in morphogenesis	195	7	2.76	0.0013
GO:0031069	hair follicle morphogenesis	16	3	0.23	0.0014
GO:0010812	negative regulation of cell-substrate adhesion	47	4	0.67	0.0019
GO:0071356	cellular response to tumor necrosis factor	188	6	2.66	0.0021
GO:0016477	cell migration	989	24	14.01	0.0023
GO:0070306	lens fiber cell differentiation	21	3	0.3	0.0031
GO:0003180	aortic valve morphogenesis	22	3	0.31	0.0035
GO:0048103	somatic stem cell division	20	3	0.28	0.0040
GO:0035023	regulation of Rho protein signal transduction	109	7	1.54	0.0040
GO:0035886	vascular smooth muscle cell differentiation	23	3	0.33	0.0040
GO:0070374	positive regulation of ERK1 and ERK2 cascade	110	6	1.56	0.0047
GO:0003156	regulation of animal organ formation	21	3	0.3	0.0052
GO:0003281	ventricular septum development	63	6	0.89	0.0055
GO:0060914	heart formation	19	3	0.27	0.0067
GO:0002067	glandular epithelial cell differentiation	28	3	0.4	0.0067
GO:0045598	regulation of fat cell differentiation	89	5	1.26	0.0082
GO:0071242	cellular response to ammonium ion	42	4	0.59	0.0082
GO:0001711	endodermal cell fate commitment	10	2	0.14	0.0083
GO:0002328	pro-B cell differentiation	10	2	0.14	0.0083
GO:0061314	Notch signaling involved in heart development	10	2	0.14	0.0083
GO:0048711	positive regulation of astrocyte differentiation	10	2	0.14	0.0083
GO:0048505	regulation of timing of cell differentiation	10	2	0.14	0.0083
GO:0003161	cardiac conduction system development	10	2	0.14	0.0083
GO:0003183	mitral valve morphogenesis	10	2	0.14	0.0083

GO:0014733	regulation of skeletal muscle adaptation	10	2	0.14	0.0083
GO:0030324	lung development	126	6	1.78	0.0084
GO:0045165	cell fate commitment	143	8	2.03	0.0091
GO:0043502	regulation of muscle adaptation	65	5	0.92	0.0098

Table S4B Cellular Component ontology

GO ID	Term	Annotated	Significant	Expected	*P* value
GO:0035102	PRC1 complex	15	4	0.21	0.000044
GO:0005796	Golgi lumen	49	4	0.68	0.0047
GO:0098797	plasma membrane protein complex	337	11	4.69	0.0086
GO:0005769	early endosome	288	8	4.01	0.0090

Table S4C Molecular Function ontology

GO ID	Term	Annotated	Significant	Expected	*P* value
GO:0032183	SUMO binding	16	3	0.22	0.0013
GO:0030145	manganese ion binding	54	4	0.74	0.0064

Table S5. List of oligonucleotide sequences used in this study.

Application	Sequence (5'-->3')
Initial low-cycle PCR - Fwd	ACTGGCCGCTTGACG
Initial low-cycle PCR - Rev	GCAGGAGCCGCAGTG
Emulsion PCR primers containing barcode tags and adapter sequences - Fwd	GCCAGAACATTTCTCTGGCCTAACTGGCC GCTTGACG
Emulsion PCR primers containing barcode tags and adapter sequences - Rev	CCGACTAGCTTGGCCGCCGAGGCCCGAC GCTCTTCCGATCTN[*16|17]TCTAGAGGTACCGCAGGAGCCGCAGTG
Library amplification for inert library sequencing - Fwd	NNNNNCAGGTGCCAGAACATTTCTCT
Library amplification for inert library sequencing - Rev	TTATCATGTCTGCTCGAAGCGG
Library amplification for competent library barcode sequencing - Fwd	NNNNNCAAGAAGGGCGGCAAGAT
Library amplification for competent library barcode sequencing - Rev	TTATCATGTCTGCTCGAAGCGG
Reverse transcription primer	CCGACTAGCTTGGCCGC
Transcript qPCR primers (targets Luciferase) - Fwd	AACACCCCAACATCTTCGAC
Transcript qPCR primers (targets Luciferase) - Rev	TCTCGGTCATGGTTTTACCG
Backbone qPCR primers - Fwd	ATTTGGTATCTGCGCTCTGC
Backbone qPCR primers - Rev	TTTGCCGGATCAAGAGCTAC
Library amplification for experimental barcode sequencing - Fwd	NNNNNCGAGGTGCCTAAAGGACTG
Library amplification for experimental barcode sequencing - Rev	CCGACGCTCTTCCGATCT

Dataset S1 (separate file). Excel table of all putative enhancers used in our study. Enhancer, enhancer name; Chr, Start, Stop, genomic coordinates in hg19; No_hSubs, number of hSubs within enhancer; NoFrags, number of fragments in enhancer; Active, YES if at least one fragment was active, otherwise NO; DiffActive, as before but for differential activity; sign_hSubs, number of actively regulating hSub within enhancer; DA_Kanton_etal, overlapping a differentially accessible peak in NPCs of human vs. chimpanzee organoids (16).

Dataset S2 (separate file). Excel table showing enrichment of *in silico*-predicted TFBSs in different classes of MPRA fragments (columns B-E) and over active hSubs (columns F-G). Column C shows results for enrichment restricted to TFBS overlapping chromatin accessibility according to one of three datasets, ATAC-seq (this study), ATAC-seq from cortical organoids (16) and ENCODE DNase-seq for H9-derived NSCs (15). Numbers are fold enrichment for significantly enriched motifs ($P_{BH} < 0.05$). The names of TFs involved in cell cycle and proliferation are in bold and those of TFs involved in differentiation in italic.

Dataset S3 (separate file). Excel table of looping events between genes expressed in hNSCs and differentially active enhancers according to chromosome conformation capture data. Column B indicates the number of hSubs with significant quantitative effects on MPRA activity. Note that all enhancers contain hSubs, but not all contain hSubs with with significant effects. Connectivity data is taken from (21–24).

Dataset S4 (separate file). Excel table of all 401 hSubs with significant effects on MPRA activity with rich information.

SI References

1. S. Prabhakar, J. P. Noonan, S. Pääbo, E. M. Rubin, Accelerated evolution of conserved noncoding sequences in humans. *Science* 314, 786–786 (2006).
2. K. Lindblad-Toh, *et al.*, A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478, 476–482 (2011).
3. S. K. Reilly, *et al.*, Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* 347, 1155–1159 (2015).
4. S. T. Sherry, *et al*., dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311 (2001).
5. J. Cotney, *et al.*, The autism-associated chromatin modifier CHD8 regulates other autism risk genes during human neurodevelopment. *Nat. Commun.* 6, 6404 (2015).
6. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17, 10–12 (2011).
7. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357 – 359 (2012).
8. R Core Team, R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria (2020).
9. M. M. Islam, *et al.*, Enhancer analysis unveils genetic interactions between *TLX* and *SOX2* in neural stem cells and in vivo reprogramming. *Stem Cell Rep.* 5, 805–815 (2015).
10. C. Lothian, U. Lendahl, An evolutionarily conserved region in the second intron of the human Nestin gene directs gene expression to CNS progenitor cells and to early neural crest cells. European *J. Neurosci.* 9, 452–462 (1997).
11. H. Y. Zhou, *et al.*, A *Sox2* distal enhancer cluster regulates embryonic stem cell differentiation potential. *Genes Dev.* 28, 2699–2711 (2014).
12. A. Siepel, *et al.*, Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15, 1034–1050 (2005).

13. A. Mathelier, X. Zhao, *et al*., JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 42, D142–D147 (2014).
14. C. E. Grant, T. L. Bailey, W. S. Noble, FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27, 1017–1018 (2011).
15. J. Vierstra, *et al.*, Global reference mapping of human transcription factor footprints. *Nature* 583, 729–736 (2020).
16. S. Kanton, M. J. Boyle, Z. He, *et al.*, Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature* 574, 418–422 (2019).
17. L. Myint, D. G. Avramopoulos, L. A. Goff, K. D. Hansen, Linear models enable powerful differential activity analysis in massively parallel reporter assays. *BMC Genomics* 20, 209 (2019).
18. T. Ashuach, D. S. Fischer, A. Kreimer, N. Ahituv, F. J. Theis, N. Yosef, MPRAnalyze: statistical framework for massively parallel reporter assays. *Genome Biol.* 20, 183 (2019).
19. H. Ryu, *et al.*, Massively parallel dissection of human accelerated regions in human and chimpanzee neural progenitors. *bioRxiv*, 256313 (2018).
20. A. Alexa, J. Rahnenführer, topGO: Enrichment Analysis for Gene Ontology. R package version 2.40.0 (2020).
21. H. Won, *et al.*, Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* 538, 523–527 (2016).
22. P. Rajarajan, T. Borrman, W. Liao, *et al.*, Neuron-specific signatures in the chromosomal connectome associated with schizophrenia risk. *Science* 362, eaat4311 (2018).
23. I. Jung, A. Schmitt, Y. Diao, *et al.*, A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nature Genet.* 51, 1442–1449 (2019).
24. M. Song, M.-P. Pebworth, X. Yang, *et al.*, Cell-type-specific 3D epigenomes in the developing human cortex. *Nature* 587, 644–649 (2020).