Methods S1. Self-reporting transposons measure single cell mRNA and transcription factor binding, related to STAR Methods

Arnav Moudgil, Michael N. Wilkinson, Xuhua Chen, June He, Alexander J. Cammack, Michael J. Vasek, Tomás Lagunas, Jr., Zongtai Qi, Matthew A. Lalli, Chuner Guo, Samantha A. Morris, Joseph D. Dougherty, and Robi D. Mitra

Notes

Figures and Tables

Molecular biology of self-reporting transposons (SRTs)

Self-reporting transposons (SRTs) are synthetic constructs that generate transcripts whose 3' untranslated regions (UTRs) contain the genomic sequence identifying the SRT's insertion site. These transcripts can be recovered using a poly(T) reverse transcription (RT) primer tailed with a universal priming site at one end of the transcripts. It is unclear whether self-reporting transcripts are truly polyadenylated like protein-coding mRNA (i.e. contain non-templated 3' adenines added by poly(A) polymerase after cleavage from RNA polymerase II) or contain templated stretches of adenine to which the RT primer can internally hybridize. Regardless, transcripts are amplified after first-strand synthesis. We then perform a pair of nested PCRs with an intermediate tagmentation step (Picelli et al., 2014) to recover the transposon-genome junction. After adapter trimming and alignment, the 5' coordinates of these reads specify the genomic locations of insertions in the library (Figure 1A).

Additional validation of SRTs

To confirm that our molecular protocol was specific for transposed SRTs (as opposed to the original plasmid copy), we generated libraries from cells transfected with either *piggyBac* SRTs alone or SRTs alongside *piggyBac* transposase. Libraries generated without transposase yielded very few reads mapping to the genome; however, protocol is highly efficient when transposase is added (Figure SM1A). Thus, a functional transposase is required to successfully map SRTs. This is likely due to the inclusion of a self-cleaving hammerhead ribozyme downstream of the SRT on the plasmid (below). We also found that technical replicates of the bulk SRT protocol showed high reproducibility, with over 80% concordance at the level on individual transpositions (Figure SM1B). To further compare the recovery of SRTs between the DNA- and RNA-based protocols, we generated chromatin state annotations based on epigenomic ChIP-seq data in HCT-116 cells (Methods, Figure SM2, and Table SM1). We observed that the distribution of insertions with respect to chromatin state was highly concordant between the DNA and RNA libraries (Figure SM1C).

Noise reduction

A common artifact observed in DNA-based transposon recovery is a large fraction of reads aligning to the donor transposon plasmid instead of the genome. Although this can be mitigated by long selection times or by digestion with the methyladenine-sensitive enzyme DpnI (Wang et al., 2012a), these methods do not completely eliminate background and are not compatible with all experimental paradigms (e.g. viral transduction). To reduce this artifact, we included a hammerhead ribozyme (Yen et al., 2004) in the SRT plasmid downstream of the 5' terminal repeat (TR). Before transposition, the ribozyme will cleave the nascent transcript originating from the marker gene, thus preventing RT. Transposition allows the SRT to escape the downstream ribozyme, leading to recovery of the self-reporting transcript. In our comparison of DNA- and RNA-based recovery, about 15% of reads from the DNA library aligned to the plasmid, compared to fewer than 1% of reads from the RNA library (Figure S1D). Thus, the addition of a self-cleaving ribozyme virtually eliminated recovery of un-excised transposons.

Uniform recovery of SRTs

Since SRT recovery relies on transcription, we wondered if SRTs deposited in euchromatic regions were recovered more efficiently than SRTs in less permissive chromatin states, which might lead to biases when mapping TF binding. As *piggyBac* is known to preferentially insert near active chromatin (Yoshida et al., 2017), this question cannot be easily answered using this transposon. Prior studies have shown that the *Sleeping Beauty* transposase (Ivics et al., 1997; Mátés et al., 2009) has very little preference for chromatin state (Yoshida et al., 2017). Therefore, we created a self-reporting *Sleeping Beauty* transposon and compared its genome-wide distribution to that of SRTs deposited by wild-type *piggyBac* (Table S1; Figure SM2A-B). Undirected *piggyBac* transposases appeared to modestly prefer transposing into promoter and enhancers, which is consistent with previous reports (Gogol-Döring et al., 2016; Yoshida et al., 2017) (Table SM1). By contrast, *Sleeping Beauty* showed largely uniform rates of insertions across all chromatin states, including repressed and inactive chromatin (Figure SM2B). These results affirm that while RNA-based recovery is more efficient, it still preserves the underlying genomic distributions of insertions. Furthermore, because SRTs can be recovered from virtually any chromatin state, RNA-based calling card recovery can be employed to analyze a variety of TFs with broad chromatin-binding preferences.

SRTs are compatible with TF-directed calling cards

Since the SRT is a new reagent, we sought to confirm that bulk RNA calling cards can, like DNA calling cards (Wang et al., 2012a), be used to identify TF binding sites. We transfected 10-12 replicates of HCT-116 cells with plasmids containing the PB-SRT-Puro donor transposon and SP1 fused to either *piggyBac* (SP1-PBase) or a hyperactive variant of *piggyBac* (Yusa et al., 2011) (SP1-HyPBase). As controls, we also transfected a similar number of replicates with undirected PBase or HyPBase, respectively. We obtained 410,588 insertions from SP1-PBase and 1,521,048 insertions from PBase; similarly, we obtained 2,029,931 SP1-HyPBase insertions and 5,771,207 insertions from HyPBase (Table S1).

Just as we had observed previously with DNA calling cards (Wang et al., 2012a), RNA calling cards were also redirected by SP1-PBase and SP1-HyPBase to SP1-bound regions of the genome (Figures SM3A and SM4A). All three of the loci shown in Figures SM3A and SM4A show a specific enrichment of calling card insertions in the SP1 fusion experiments that is not observed in the undirected control libraries. Next, we called peaks at all genomic regions enriched for SP1-directed transposition. The number of insertions observed at significant peaks for both SP1-PBase and SP1-HyPBase was highly reproducible between biological replicates (R^2 = 0.87 and 0.96, respectively; Figures SM3B and SM4B). Furthermore, calling card peaks were highly enriched for SP1 ChIP-seq signal at their centers, both on average (Figures SM3C and SM4C) and in aggregate (Figures SM3D and SM4D). SP1 is known to preferentially bind near transcription start sites (TSSs) and is also thought to play a role in demethylating CpG islands (Brandeis et al., 1994; Macleod et al., 1994; Philipsen and Suske, 1999). We confirmed that the SP1-directed transposases preferentially inserted SRT calling cards near TSSs, CpG islands, and unmethylated CpG islands at statistically significant frequencies ($p < 10^{-9}$ in each instance, *G* test of independence; Figures SM3E and SM4E). Moreover, compared to undirected *piggyBac*, SP1-directed *piggyBac* showed a striking preference for depositing insertions into promoters (Figure SM2A-B). Lastly, regions targeted by SP1-PBase and SP1-HyPBase were enriched for the core SP1 DNA binding motif ($p < 10^{-70}$ in each instance; Figures SM3F and SM4F). Taken together, these results indicate that the genome-wide binding of SP1 can be accurately mapped using *piggyBac* SRTs.

Hyperactive *piggyBac* insertions identify BRD4-bound super-enhancers

We analyzed hyperactive *piggyBac* transposase (HyPBase) data in identical fashion to the standard *piggyBac* transposase (PBase). Undirected HyPBase showed non-uniform densities of insertions at BRD4-bound loci (Figures SM5A and SM8). At statistically significant peaks, HyPBase showed high reproducibility of normalized insertions between biological replicates ($R^2 > 0.99$; Figure SM5B). We calculated the mean BRD4 enrichment, as assayed by ChIP-seq (McCleland et al., 2016), over all HyPBase peaks, which showed significantly increased BRD4 signal compared to a permuted control set ($p < 10^{-9}$, Kolmogorov-Smirnov test; Figure SM5C). Maximum BRD4 ChIP-seq signal was observed at calling card peak centers and decreased symmetrically in both directions. Furthermore, HyPBase peaks showed striking overlap with ChIP-seq profiles for several histone modifications (Sloan et al., 2016; The ENCODE Project Consortium, 2012), in particular an enrichment for the enhancer-associated acetylated H3K27 and monomethylated H4K4; and depletion for the heterochromatin marks H3K9me3 and H3K27me3 (Lawrence et al., 2016) (Figure SM5D). Finally, HyPBase-derived peaks showed high sensitivity, specificity, and precision for BRD4-bound super-enhancers (SEs; Figures SM5E-F).

Redirectability of *piggyBac*

piggyBac's baseline preference for BRD4 raises questions about how efficiently TF-*piggyBac* fusions can redirect insertions near TF binding sites. We further analyzed the bulk SP1-directed experiments and found that SP1-*piggyBac* increased insertion density at SP1-bound, BRD4-depleted regions by five- to seven-fold, on average (Figures SM6A and SM6C). We also saw a decrease in insertion density at non-SP1-bound BRD4 peaks on the order of 30 to 50 percent (Figures SM6B and SM6D). This suggests that, while the reduction of signal at BRD4-bound loci may be modest, the redirection to TF binding sites can be quite stark, explaining how TF binding sites can be accurately identified (Wang et al., 2012a). In contrast to *piggyBac*, *Sleeping Beauty* has a more uniform background distribution of insertions (Figure SM7), which suggests that the latter transposon system might be even more redirectable and allow us to perform TF-directed calling cards without the need for an undirected transposase control. Unfortunately, direct fusions of TFs to *Sleeping Beauty* almost completely abolish transposase activity (Wu et al., 2006). We confirmed this in a colony formation assay with SP1 fused to either *piggyBac* or *Sleeping Beauty.* The SP1-*Sleeping Beauty* fusion had virtually undetectable levels of transposition, whereas the SP1-*piggyBac* construct was still enzymatically functional (Figure SM8). Currently, *piggyBac* remains the practical choice for mammalian calling cards, but the prospect of a background-free calling card strategy should motivate future research.

Downsampling analysis

To project how transposon calling cards would scale to single cell experiments, where molecular techniques show broadly reduced sensitivity compared their bulk counterparts, we simulated assay performance under increasingly sparse conditions. We quantified the relationship between SE sensitivity and the number of insertions recovered in undirected calling cards experiments by downsampling the data from the PBase and HyPBase experiments in half-log increments and

calculating sensitivity (Figure SM9A-B). These heatmaps show that sensitivity increases with the total number of insertions recovered. Since we cannot predict how many, or few, insertions will be recovered in future experiments, we also performed linear interpolation on the downsampled data. The resulting contour plots (Figure SM9C-D) indicate the approximate sensitivity of BRD4-bound SE detection in HCT-116 cells. Our analysis suggests that even with as few as 10,000 insertions, we can still obtain sensitivities around 50%. Similarly, we investigated the reproducibility of SP1-directed peaks at a various downsampled numbers of insertions, using the peaks obtained from our bulk SP1-HyPBase experiment in HCT-116 cells as our reference set (Figure SM9E-F). We found that peak detection is directly proportional to the number of SP1-directed insertions recovered. At a lower limit of 10,000 insertions in both the experimental and control datasets, there was 40% overlap with peaks called from our bulk dataset. Together, these analyses provide a guide for how well calling cards will perform in the limit of insertion recovery.

Molecular biology of single cell calling cards

Single cell calling cards (scCC) uses a modified version of the bulk SRT amplification protocol since the cell barcode and transposon-genome junction may be too far to sequence with short-read technology. To overcome this, we amplify with primers that bind to the universal priming sequence next to the cell barcode and the terminal sequence of the *piggyBac* TR. These primers are biotinylated and carry a 5' phosphate group (Table S4). The PCR products of this amplification are diluted and allowed to circularize overnight. They are then sheared and captured with streptavidin-coated magnetic beads (Methods). The rest of the library is prepared on-bead and involves end repair, A-tailing, and adapter ligation (Methods). A final PCR step adds the required Illumina sequences for high-throughput sequencing (Methods). The standard Illumina read 1 primer sequences the cell barcode and UMI, while a custom read 2 primer, annealing to the end of the *piggyBac* 5' TR, sequences into the genome. Thus, we collect both the location of a *piggyBac* insertion as well as its cell of origin. Cells undergo droplet encapsulation and barcoding using the 10x Chromium protocol, with the omission of the template switch oligonucleotide (TSO; Methods) from the RT step. The resulting product is then split in two: one half is used to generate the scRNA-seq library (including adding back the TSO) while the other half undergoes the scCC protocol. After sequencing, the shared cell barcodes between both libraries are used to connect individual insertions to specific cell types.

Single cell RNA-seq of the mouse cortex

We chose the mouse cortex for our *in vivo* proof-of-concept because it is a heterogeneous tissue that has been the focus of several recent single cell studies (Rosenberg et al., 2018; Saunders et al., 2018; Tasic et al., 2018; Zeisel et al., 2015, 2018). We collected nine scRNA-seq libraries from P14-P28 mice, encompassing 35,950 cells and 111,382 insertions (Table S2). We clustered cells by their mRNA profiles and used established marker genes to classify different cell types (Figure SM10A-B) (Saunders et al., 2018; Tasic et al., 2018; Zeisel et al., 2018). Neurons and astrocytes were the two major cell populations we recovered, which is consistent with the known tropism of AAV9 (Cammack et al., 2020; Schuster et al., 2014), though we also identified a spectrum of differentiating oligodendrocytes and trace amounts of microglial, vascular, and ependymal cells (Table S3).

Discussion on SRT recovery and polyadenylation signals

Mapping SRTs using cellular RNA appears to be substantially more efficient than the DNA-based inverse PCR method, but the reasons for this are unclear. Some efficiency is likely gained by eliminating self-ligation, as well as having multiple mRNA copies of each insertion to buffer against PCR artifacts. It is also unknown what fraction of self-reporting transcripts are actually polyadenylated as opposed to merely containing A-rich genomic tracts. Non-genic polyadenylation signals (PASs) prevent anti-sense transcription (Chiu et al., 2018), which suggests that PASs may be more common in the genome than previously appreciated. Targeted 3'-end sequencing (Chen et al., 2017; Zheng et al., 2016) of SRT libraries should help resolve this question, while long-read sequencing of self-reporting transcripts may identify non-canonical PASs

Discussion on *piggyBac*'s affinity for BRD4

The natural affinity of *piggyBac* for BRD4 makes it ideal for studying BRD4-bound SEs, which play important regulatory roles in development and disease (Hnisz et al., 2013; Lovén et al., 2013; Whyte et al., 2013). It is unclear why *piggyBac* has this predilection. BRD4 has an intrinsically disordered region and cooperative interactions between BRD4 and coactivators like MED1 may mediate the formation of intranuclear condensates (Sabari et al., 2018) at SEs. One hypothesis is that *piggyBac* has a similarly disordered domain that allows it to preferentially enter condensates and enrich SEs with insertions. If such a domain exists, mutating it may make unfused *piggyBac* more uniform in its insertion profile, improving its utility for TF-directed calling cards.

Discussion on designing TF-*piggyBac* fusions

Although we exclusively used N-terminal fusions in this study, calling cards can also work with C-terminal fusions (Yen et al., 2018). For viral constructs where space is limited, we have also had success fusing a TF's binding domain to *piggyBac* (Cammack et al., 2020). In general, multiple fusion strategies should be tested to empirically determine the optimal construct, particularly if the binding domain lies near one of the termini. Finally, some TFs may not bind when fused to *piggyBac* and thus would not work with calling cards, though in our experience this is uncommon (less than 25% of the time or so).

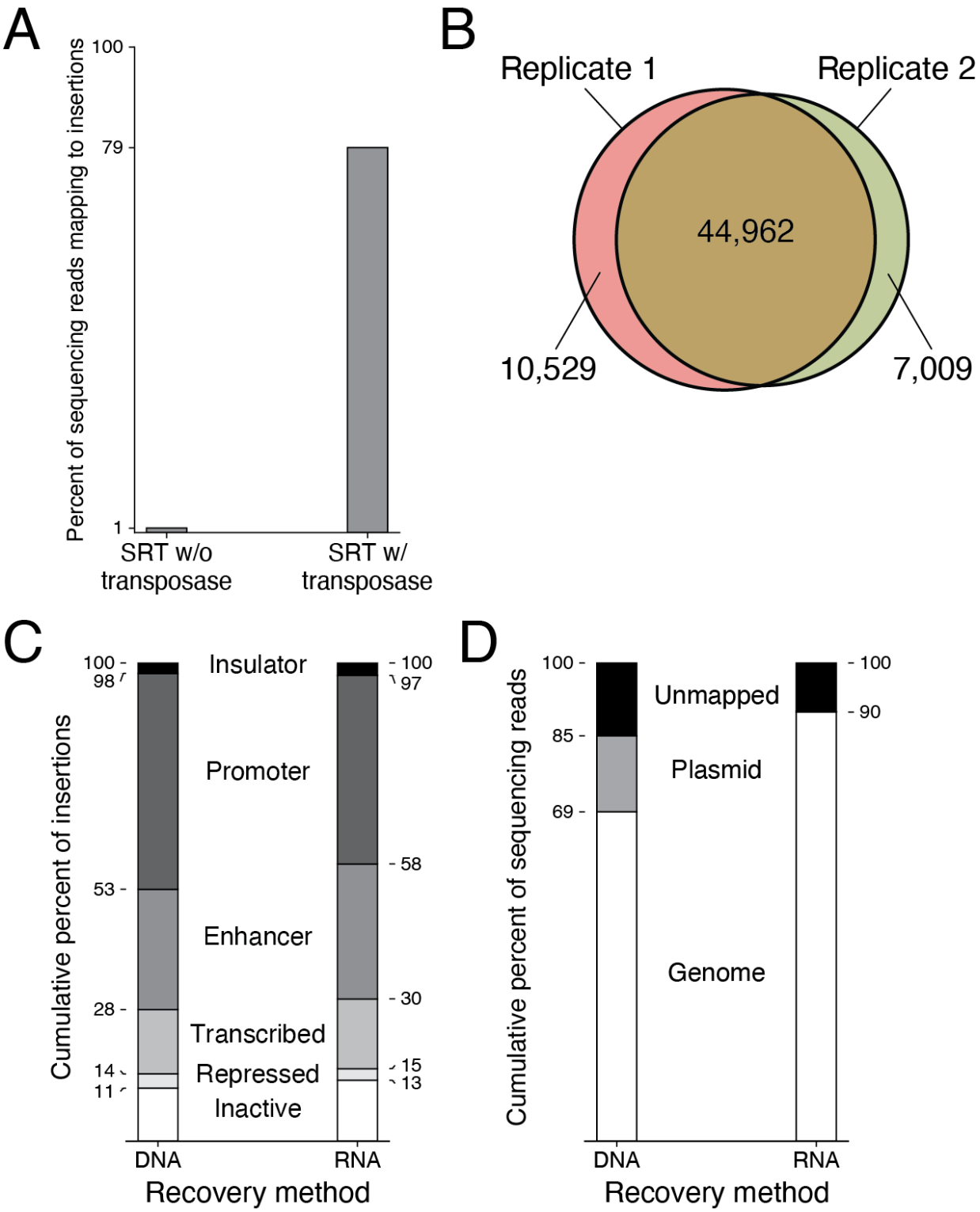

Figure SM1: Properties of self-reporting transposons (SRTs). (A) Efficiencies of bulk RNA calling card libraries prepared from HEK293T cells transfected with PB-SRT-tdTomato with or without HyPBase transposase. (B) Overlap of SRTs recovered by two technical replicates of bulk RNA calling cards in HCT-116 cells transfected with PB-SRT-Puro and SP1-PBase. (C) Distribution of insertions with respect to chromatin state between SRT libraries prepared from either DNA or RNA. (D) Breakdown of sequencing reads mapping to the genome or plasmid from SRT libraries prepared from either DNA or RNA.

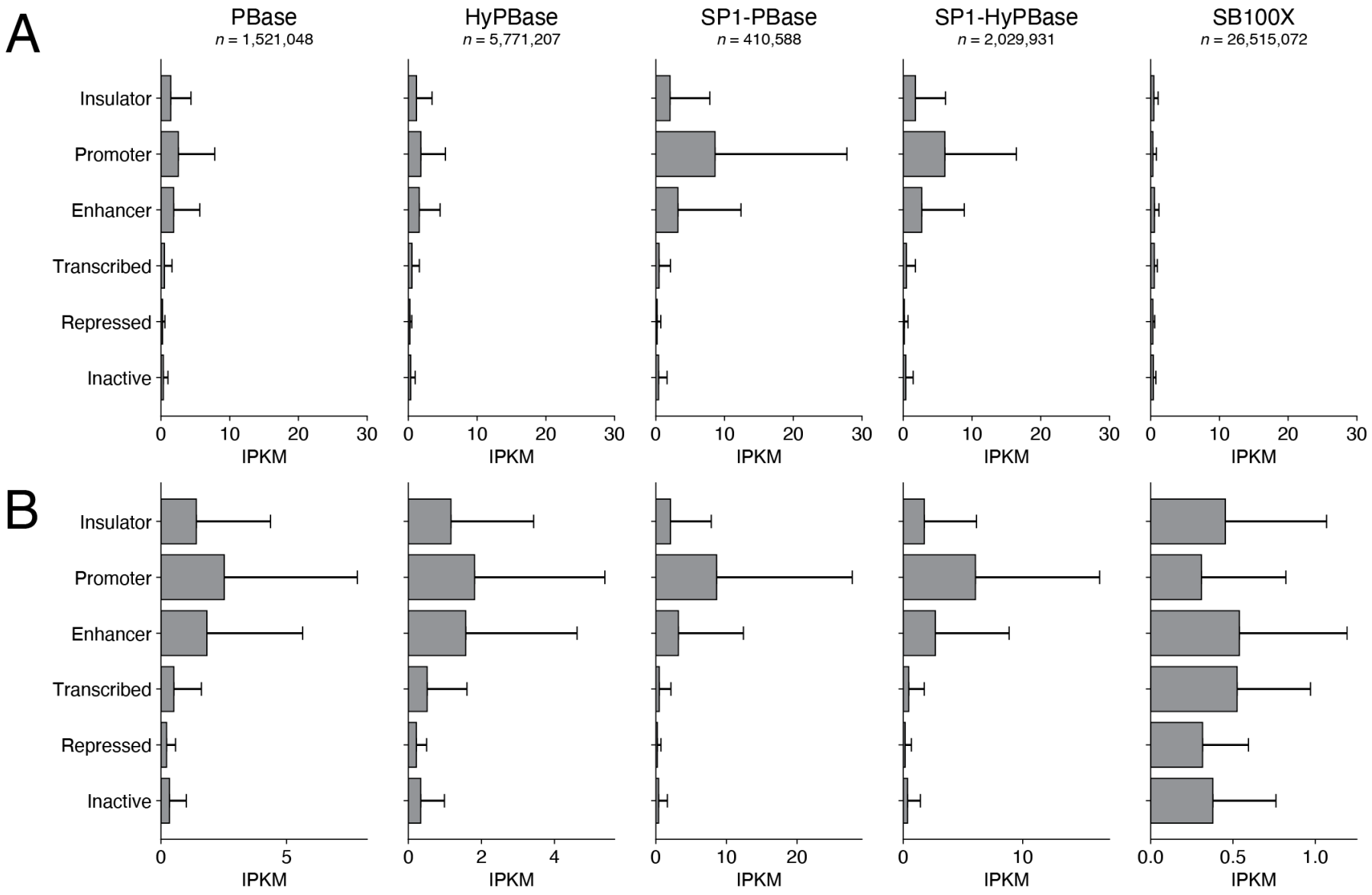

Figure SM2: *piggyBac*, SP1-*piggyBac* fusions, and *Sleeping Beauty* display different local transposition rates depending on chromatin state. (A) Chromatin state analysis on the local rates of transposition of undirected *piggyBac*, SP1-*piggyBac* fusions, and *Sleeping Beauty* transposases in HCT-116 cells. (B) Same data as (A) but with different *x*-axes for each graph. IPKM: insertions per kilobase per million mapped insertions.

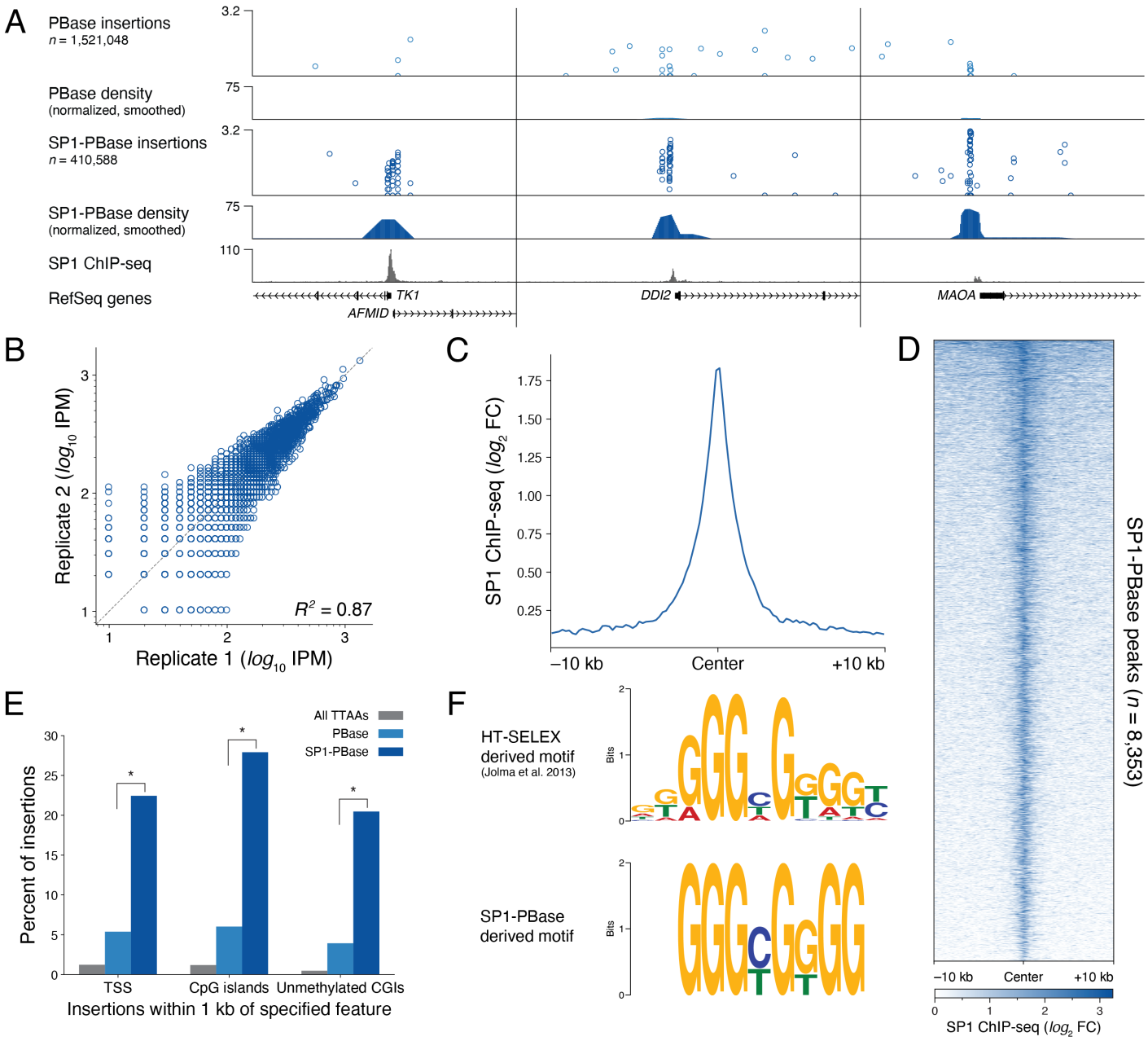

Figure SM3: SP1 fused to *piggyBac* (SP1-PBase) redirects SRTs to SP1 binding sites. (A) Browser view of bulk SP1-PBase calling cards in HCT-116 cells. (B) Reproducibility of normalized insertions at bulk SP1-PBase peaks. (C) Mean SP1 ChIP-seq signal at bulk SP1-PBase peaks. (D) Heatmap of SP1 ChIP-seq signal at bulk SP1-PBase peaks. (E) Enrichment of SP1-PBase-directed insertions to TSSs, CGIs, and unmethylated CGIs (*G* test of independence $p < 10^{-9}$). (F) SP1 core motif elicited from bulk SP1-PBase peaks. IPM: insertions per million mapped insertions; FC: fold change; TSS: transcription start sites; CGI: CpG island.

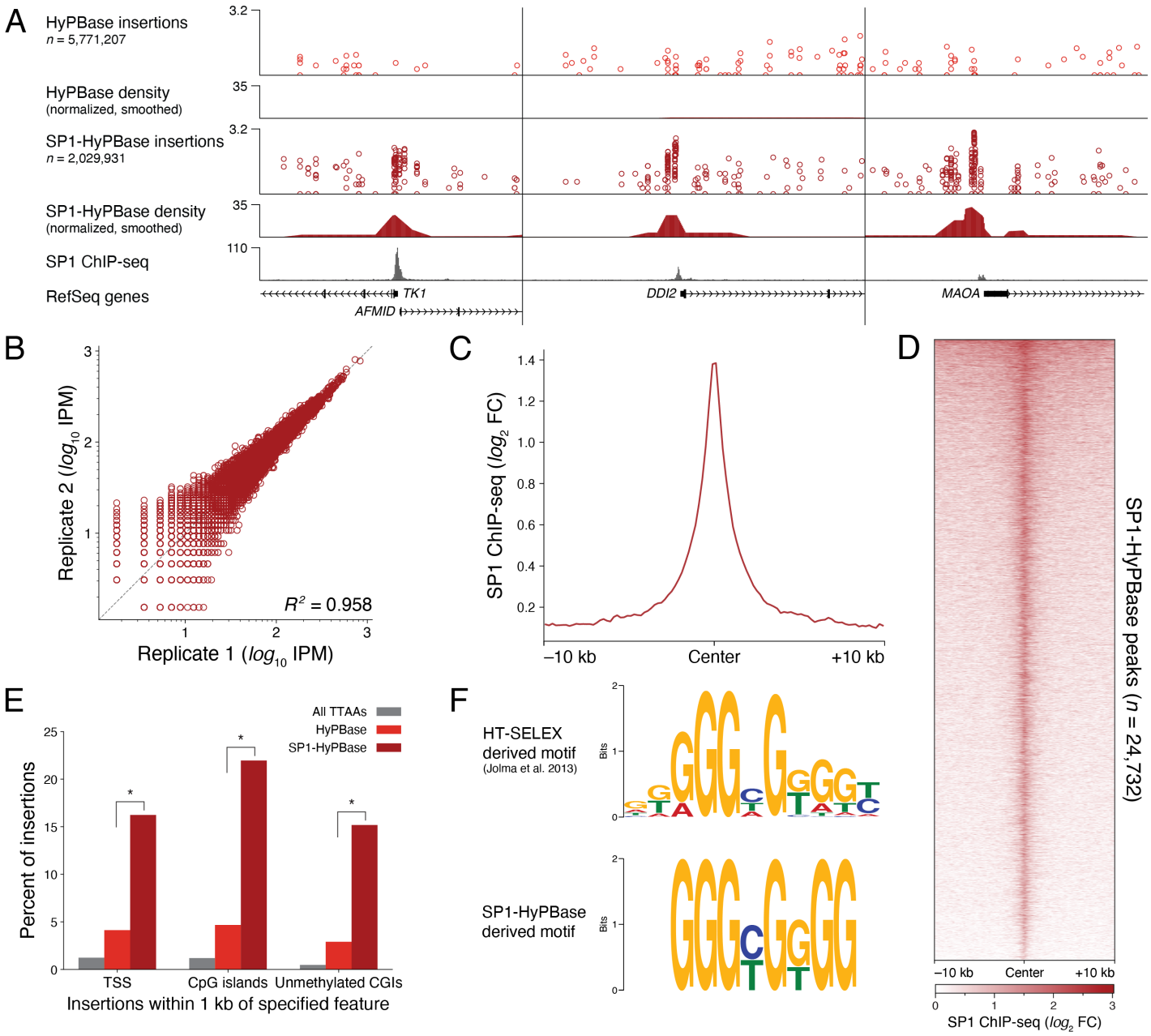

Figure SM4: SP1 fused to hyperactive *piggyBac* (SP1-HyPBase) also redirects SRTs to SP1 binding sites. (A) Browser view of bulk SP1-HyPBase calling cards in HCT-116 cells. (B) Reproducibility of normalized insertions at bulk SP1-HyPBase peaks. (C) Mean SP1 ChIP-seq signal at bulk SP1-HyPBase peaks. (D) Heatmap of SP1 ChIP-seq signal at bulk SP1-HyPBase peaks. (E) Enrichment of SP1-HyPBase-directed insertions to TSSs, CGIs, and unmethylated CGIs (*G* test of independence $p < 10^{-9}$). (F) SP1 core motif elicited from bulk SP1-HyPBase peaks. IPM: insertions per million mapped insertions; FC: fold change; TSS: transcription start sites; CGI: CpG island.

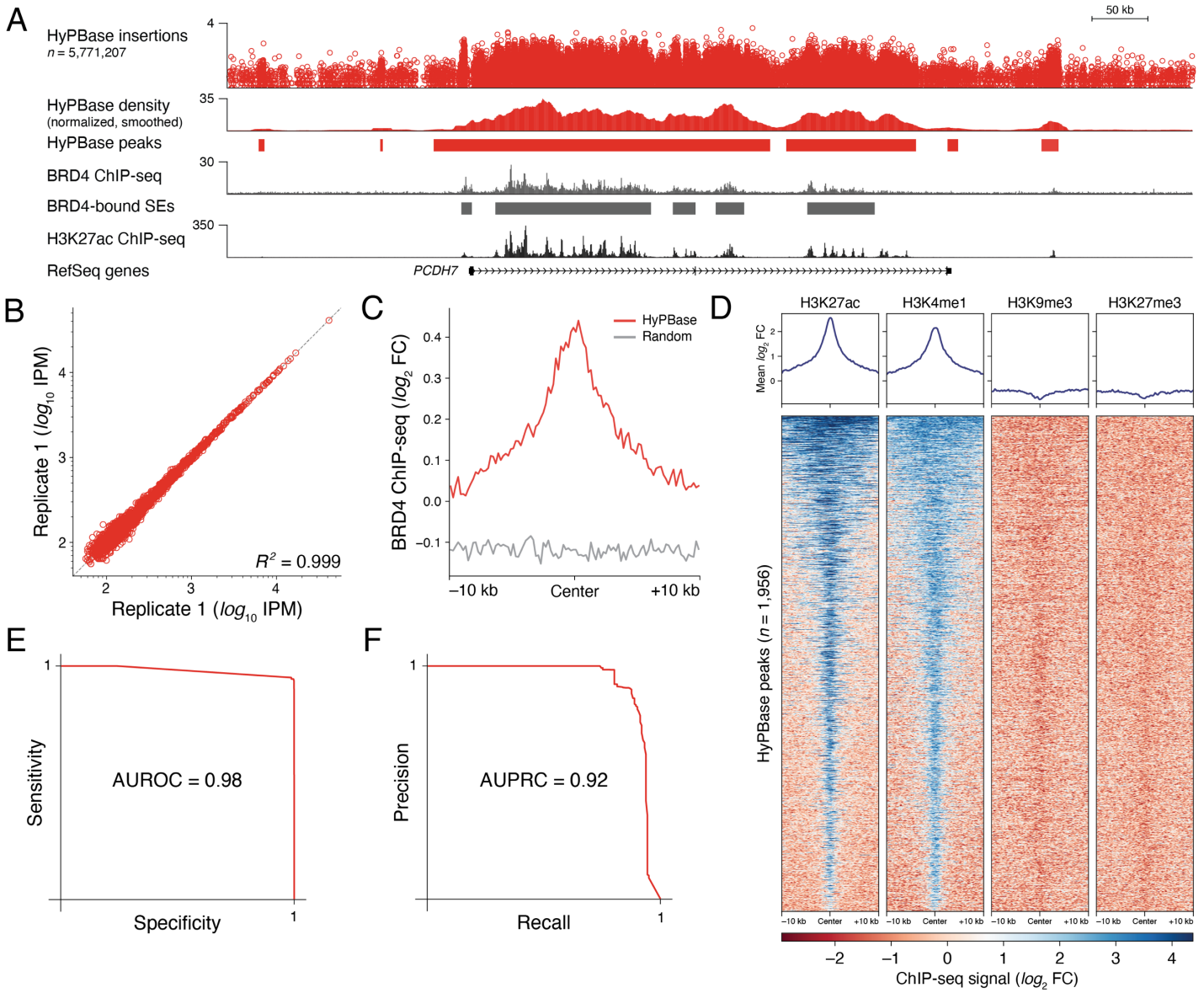

Figure SM5: Undirected hyperactive *piggyBac* (HyPBase) SRTs also mark BRD4-bound super-enhancers. (A) Browser view of undirected HyPBase insertions at a SE alongside BRD4 and H3K27ac ChIP-seq data in HCT-116 cells. (B) Reproducibility of normalized insertions at HyPBase peaks. (C) Mean BRD4 ChIP-seq signal at HyPBase peaks compared to permuted control set. (D) Heatmap of H3K27ac, H3K4me1, H3K9me3, and H3K27me3 ChIP-seq signal at HyPBase peaks. (E) Receiver-operator characteristic curve for SE detection using HyPBase peaks. (F) Precision-recall curve for SE detection using HyPBase peaks. SE: super-enhancer; IPM: insertions per million mapped insertions; AUROC: area under receiver-operator curve; AUPRC: area under precision-recall curve; FC: fold change.

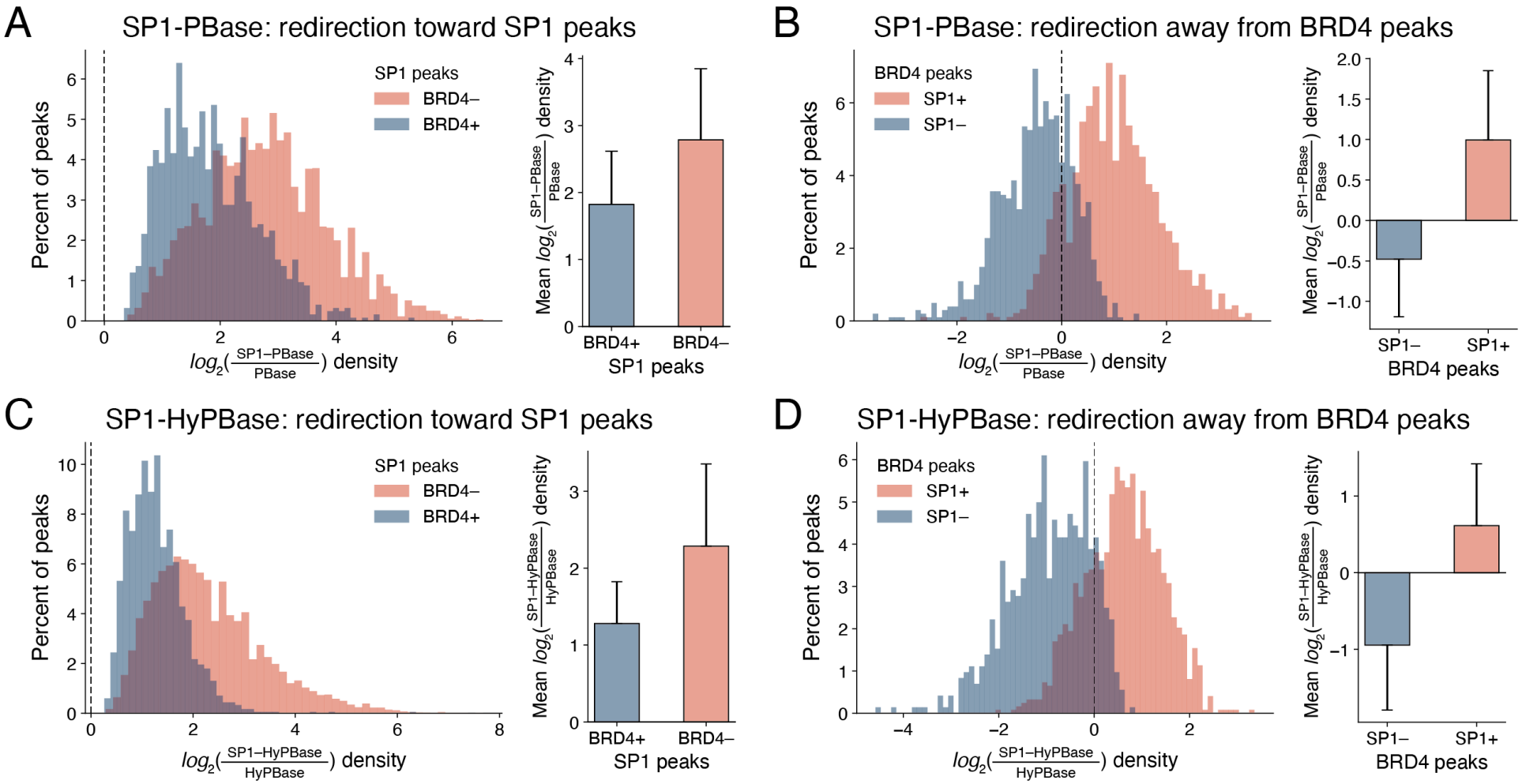

Figure SM6: Redirectability of SP1-*piggyBac* fusion constructs. (A) Left: distribution of insertion densities at SP1-PBase peaks that either overlap, or do not overlap, BRD4-directed PBase peaks (BRD4+ and BRD4–, respectively) in HCT-116 cells. Right: mean and SD of distributions. (B) Left: distribution of insertion densities at BRD4-directed, PBase peaks that either overlap, or do not overlap, SP1-Pbase peaks (SP1+ and SP1–, respectively). Right: mean and SD of distributions. (C-D) Similar analysis as (A-B) applied to the SP1-HyPBase and HyPBase datasets, respectively. SD: standard deviation.

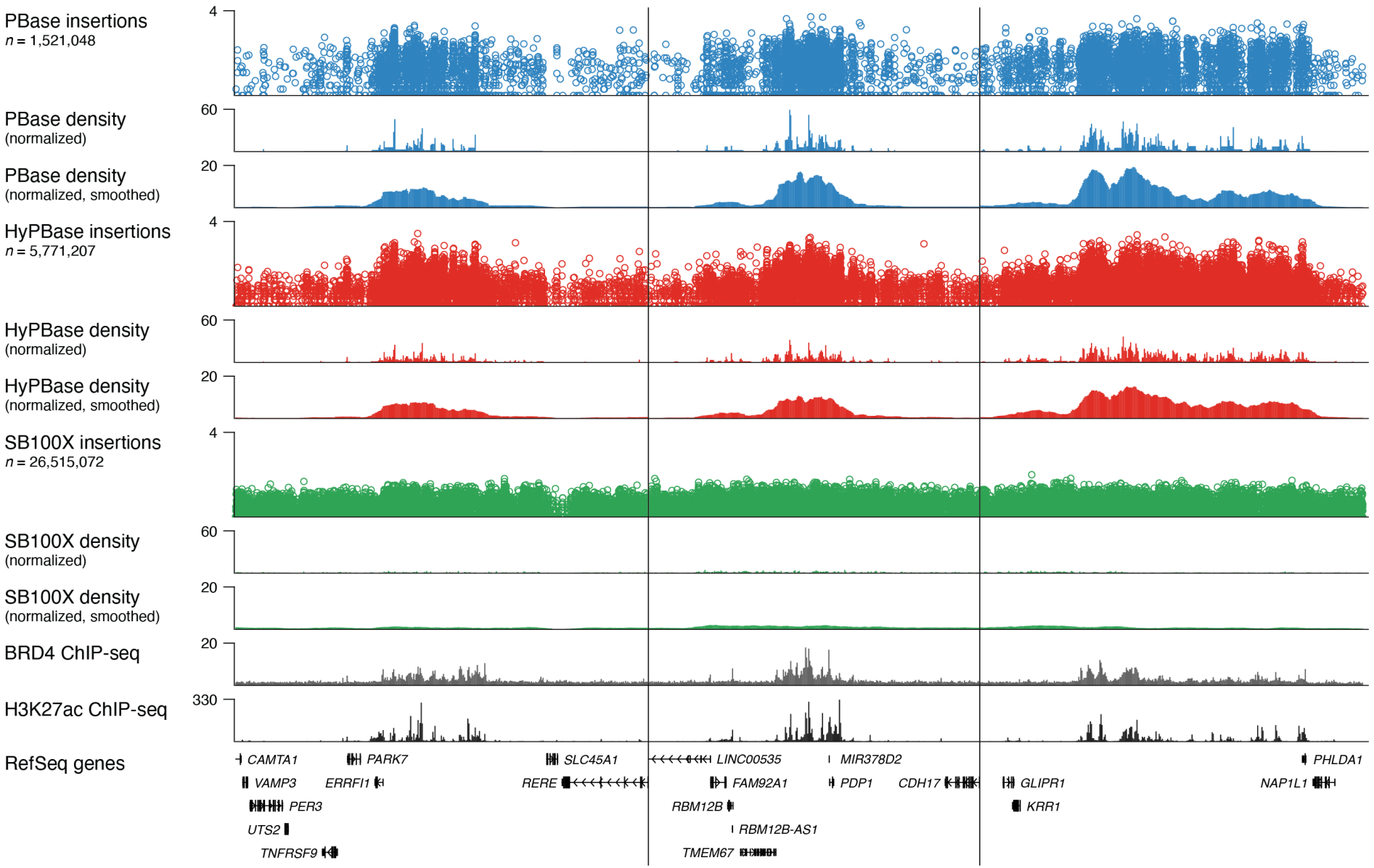

Figure SM7: Examples of BRD4-bound super-enhancers identified by bulk PBase and HyPBase calling cards in HCT-116 cells. Three different loci exhibiting non-uniform densitites of *piggyBac* insertions correlated with BRD4 and H3K27ac ChIP-seq data. *Sleeping Beauty* insertions at those same loci are more uniformly distributed. Density tracks are shown before and after smoothing.

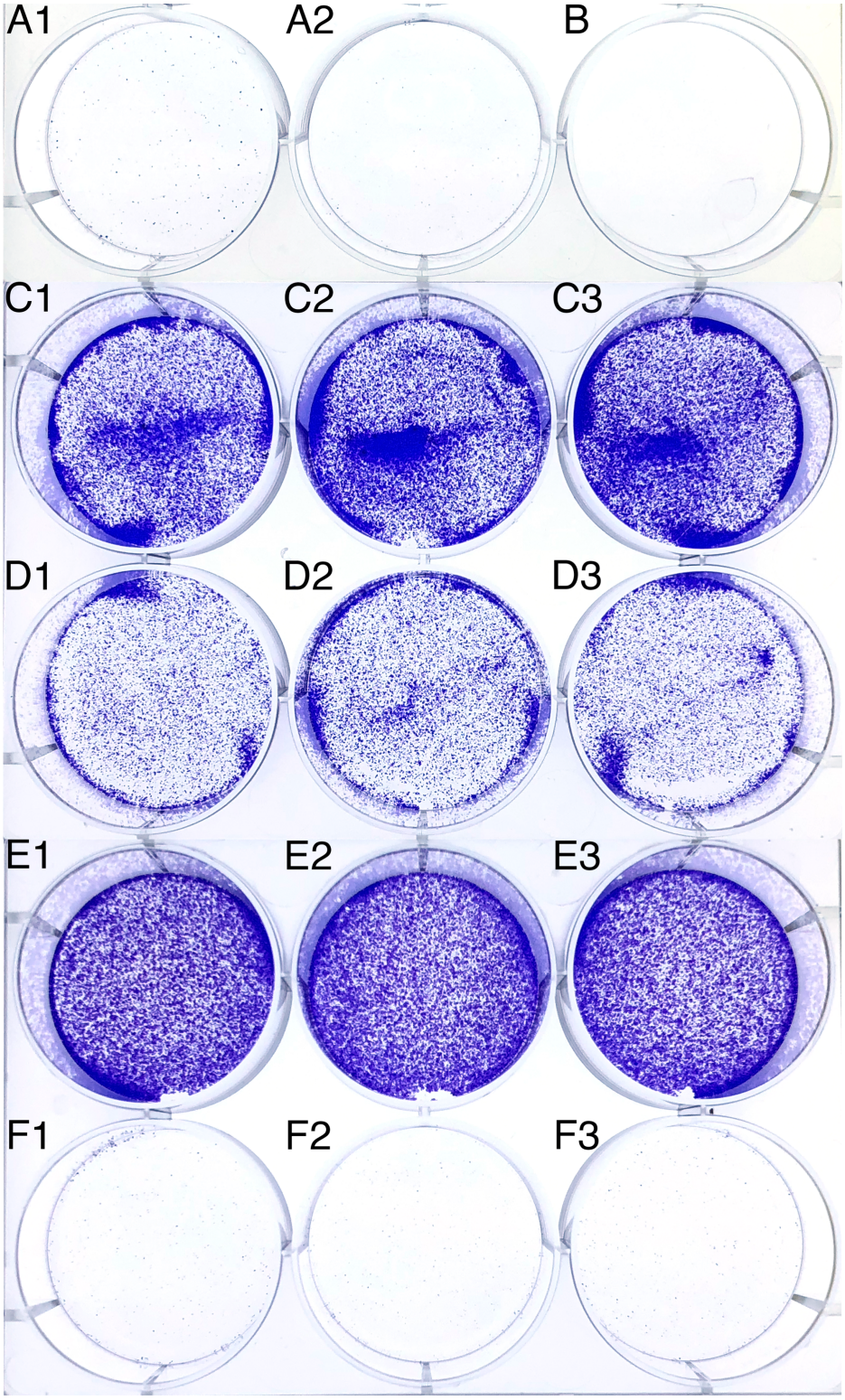

Figure SM8: *piggyBac* is more tolerant of transcription factor fusions than *Sleeping Beauty*. Colony formation assays of HCT-116 cells transfected with the specified construct(s), selected with puromycin, and stained with crystal violet. Numbers indicate biological replicates. (A) PB-SRT-Puro (B) Untransfected (no DNA). (C) PB-SRT-Puro and hyperactive *piggyBac* transposase (HyPBase). (D) PB-SRT-Puro and SP1 fused to hyperactive *piggyBac* (SP1-HyPBase). (E) SB-SRT-Puro and hyperactive *Sleeping Beauty* (SB100X). (F) SB-SRT-Puro and SP1 fused to hyperactive *Sleeping Beauty* (SP1-SB100X).

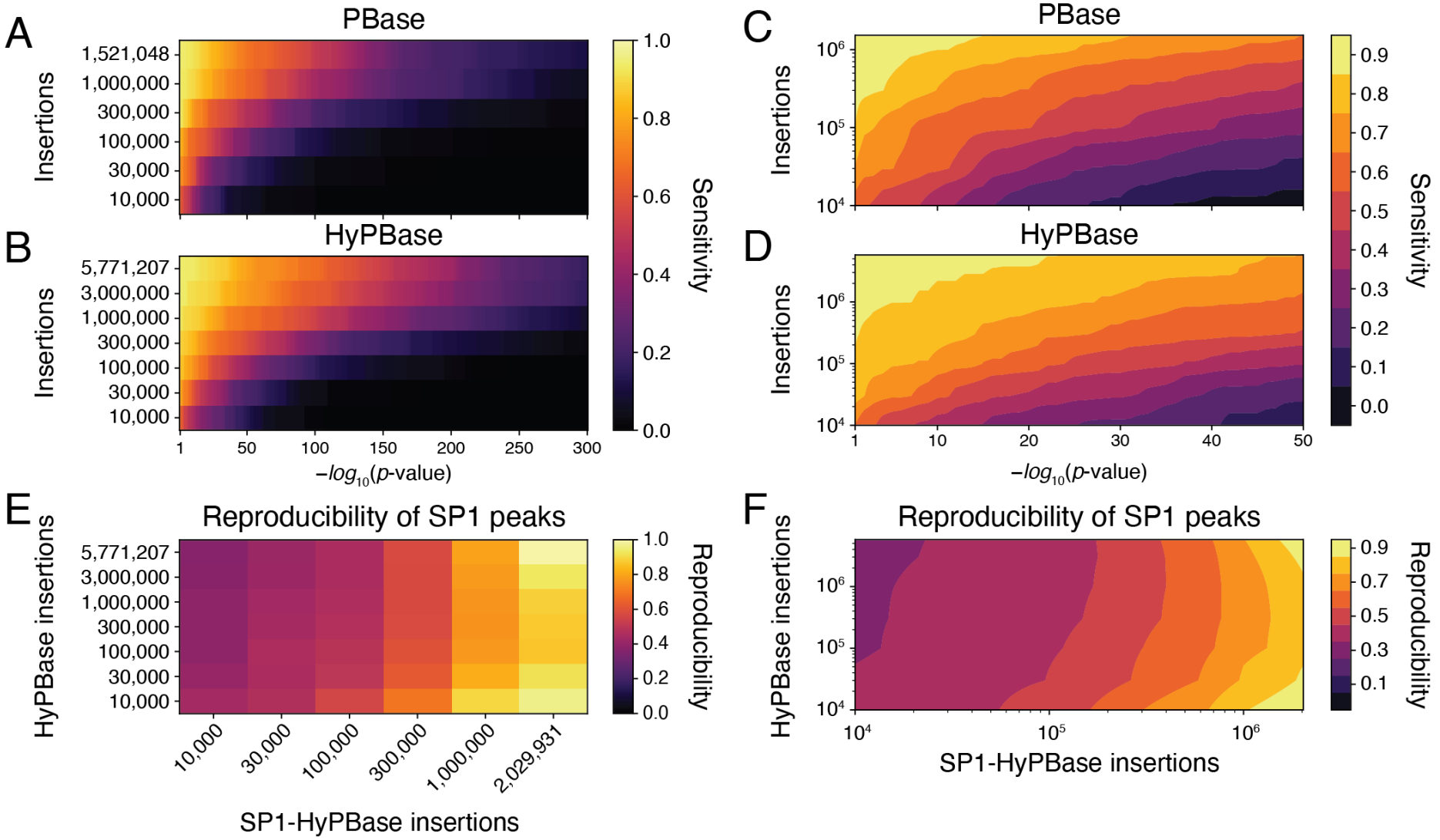

Figure SM9: Downsampling undirected and directed *piggyBac* insertions simulates assay performance. (A) Downsampling analysis of BRD4-bound SE detection by PBase insertions (in HCT-116 cells) at various *p*-value thresholds. (B) Downsampling analysis in (A) applied to HyPBase insertions. (C) Linear interpolation applied to (A) to predict SE sensitivity across a range of insertions. (D) Linear interpolation applied to (B). (E) Reproducibility of bulk SP1 calling card peaks at various numbers of HyPBase and SP1-HyPBase insertions, relative to the full dataset (top right corner). (F) Linear interpolation applied to (E) to predict peak reproducibility across a arange of experimental and control insertions.

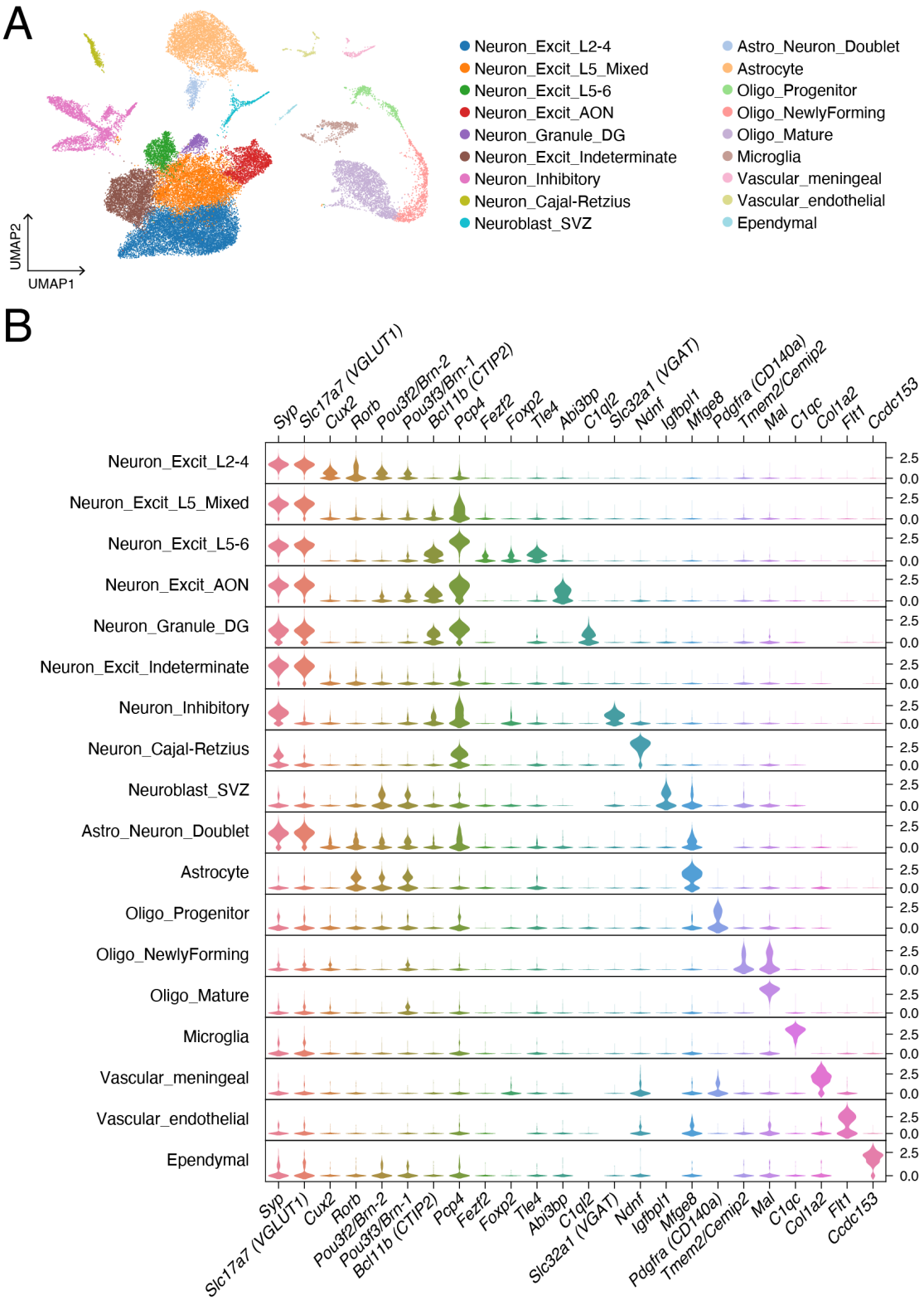

Figure SM10: Clustering of SRT-treated cortical cells and associated marker genes. (A) Two-dimensional UMAP embedding of nine mouse brain (P14-28; 8 cortical, 1 hippocampal) scRNA-seq libraries transduced with AAV9-PB-SRT-tdTomato and AAV9-HyPBase at P0-2. Louvain clustering analysis identified 18 populations. (B) Expression profiles of selected marker genes used to identify individual cell types. Excit: excitatory; oligo: oligodendrocytes; SVZ: sub-ventricular zone; AON: anterior olfactory nucleus; DG: dentate gyrus; L2-4: layer 2-4; L5-6: layer 5-6.

Table SM1: ChromHMM chromatin state annotations in HCT-116 cells

Chromatin states: Emission	CTCF	H3K9me2	H3K9me3	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	H3K79me2	Candidate state annotation
1	86	1	1	1	1	2	13	21	1	5	0	1	Insulator
2	23	1	1	2	0	1	38	97	87	20	29	3	Promoter
3	25	0	0	0	2	1	35	100	100	96	98	37	
4	14	0	1	0	2	1	89	98	36	95	41	4	Enhancer
5	3	2	2	1	1	2	59	57	1	9	1	3	
6	3	1	2	0	14	9	77	90	27	53	22	92	
7	1	1	4	0	13	10	11	1	0	5	0	85	Transcribed
8	0	1	3	0	14	5	4	0	0	2	0	12	
9	5	0	2	0	32	8	16	5	0	65	1	36	
13	0	3	6	1	1	3	4	0	0	0	0	2	
12	0	1	19	1	1	3	3	0	1	0	0	1	Repressed
15	0	2	3	30	0	5	3	0	0	0	0	1	
11	0	0	0	0	0	0	0	0	0	0	0	0	Inactive
14	0	1	1	1	0	1	2	0	0	0	0	0	
10	3	48	69	30	54	36	28	21	34	21	15	28	

Chromatin mark observation frequency (%)