Self-reporting transposons enable simultaneous readout of gene expression and transcription factor binding in single cells

Arnav Moudgil[1,2,3], Michael N. Wilkinson[1,2], Xuhua Chen[1,2], June He[1,2], Alex J. Cammack[4], Michael J. Vasek[1,6], Tomas Lagunas, Jr.[1,6], Zongtai Qi[1,2], Samantha A. Morris[1,5], Joseph D. Dougherty[1,6], Robi D. Mitra[1,2]

1. Department of Genetics, Washington University in St. Louis School of Medicine, St. Louis MO 63110, USA
2. Edison Family Center for Genome Sciences and Systems Biology, Washington University in St. Louis School of Medicine, St. Louis MO 63110, USA
3. Medical Scientist Training Program, Washington University in St. Louis School of Medicine, St. Louis MO 63110, USA
4. Department of Neurology, Washington University in St. Louis School of Medicine, St. Louis MO 63110, USA
5. Department of Developmental Biology, Washington University in St. Louis School of Medicine, St. Louis MO 63110, USA
6. Department of Psychiatry, Washington University in St. Louis School of Medicine, St. Louis MO 63110, USA

To whom correspondence should be addressed:

Rob Mitra, PhD
Washington University in St. Louis School of Medicine
Department of Genetics
660 S. Euclid Ave, Campus Box 8510
St. Louis, MO 63110
E-mail: rmitra@wustl.edu
Tel: +1-314 362-2751
Fax: +1-314-362-2157

Abstract

In situ assays of transcription factor (TF) binding are confounded by cellular heterogeneity and represent averaged profiles in complex tissues. Single cell RNA-seq (scRNA-seq) is capable of resolving different cell types based on gene expression profiles, but no technology exists to directly link specific cell types to the binding pattern of TFs in those cell types. Here, we present self-reporting transposons (SRTs) and their use in single cell calling cards (scCC), a novel assay for simultaneously capturing gene expression profiles and mapping TF binding sites in single cells. First, we show how the genomic locations of SRTs can be recovered from mRNA. Next, we demonstrate that SRTs deposited by the *piggyBac* transposase can be used to map the genome-wide localization of the TFs SP1, through a direct fusion of the two proteins, and BRD4, through its native affinity for *piggyBac*. We then present the scCC method, which maps SRTs from scRNA-seq libraries, thus enabling concomitant identification of cell types and TF binding sites in those same cells. As a proof-of-concept, we show recovery of cell type-specific BRD4 and SP1 binding sites from cultured cells. Finally, we map Brd4 binding sites in the mouse cortex at single cell resolution, thus establishing a new technique for studying TF biology *in situ.*

Introduction

Transcription factors (TFs) regulate gene expression during the most critical junctures in the specification of cell fate [1-4]. They are central to the maintenance of stem cell pluripotency [5,6] and required for normal organogenesis during development [7]. Overexpression of certain TFs can transdifferentiate one cell type into another [8], while abolishing TF binding sites can result in striking global phenotypes [9,10]. Furthermore, the pattern of TF binding is often dysregulated during disease states [11]. A better understanding of TF binding during tissue development and homeostasis would provide insights into how cellular diversity arises and is maintained under normal and abnormal biological conditions.

In the past few years, single cell RNA-seq (scRNA-seq) has emerged as the *de facto* approach for characterizing cellular diversity in complex tissues and organisms [12-17]. More recently, multi-modal scRNA-seq technologies have been developed [18-24] that combine transcriptional information with other genomic assays. These methods are motivated by the realization that while scRNA-seq can describe the current state of a biological system, it alone cannot explain how that state arose. Thus, for a given population of cells, one can now simultaneously measure transcriptome and genome [18,19], or methylome [20,21], or chromatin accessibility [21,22], or cell-surface markers [23,24]. These techniques enable greater insight into the regulatory elements driving individual transcriptional programs.

A notable lacuna in the single cell repertoire is a method for simultaneously assaying transcriptome and TF binding. Such a method would allow for the genome-wide identification of TF binding sites across multiple cell types in complex tissues. ChIP-seq is the most popular technique for studying TF binding [25], and while single cell ChIP-seq has been previously described [26], this technique has only been employed to map highly abundant proteins such as methylated histones. DamID can recover TF binding sites by identifying nearby exogenously methylated adenines [27,28], but in single cells it has only been used to study laminin-associated domains [29-31]. Importantly, both methods yield sparse data, either in small numbers of cells [31] or without simultaneously capturing mRNA [26]. Thus, each can only be used in heterogeneous samples if the cell type is known *a priori* and if sufficient numbers of cells are obtained by selection or sorting to overcome sparsity. In contrast, single cell assay for transposase-accessible chromatin (scATAC-seq) [32] can be used to identify nucleosome-free regions that may be bound by TFs across large numbers of mixed cells. However, it can only suggest potential DNA binding proteins by motif inference. Thus, it does not directly measure TF occupancy, and moreover it cannot be used to

study transcriptional regulators that bind DNA indirectly or non-specifically, such as chromatin remodelers.

Our lab has previously developed transposon calling cards as an alternative assay of TF binding [33-35]. This system relies on two components: a fusion between a TF and a transposase, and a transposon carrying a reporter gene. The fusion transposase deposits transposons near TF binding sites; these insertions are subsequently amplified from genomic DNA and subjected to high-throughput sequencing. Thus, the redirected transposase leaves "calling cards" at the genomic locations it has visited, which can be identified later in time. The result is a genome-wide assay of all binding sites for that particular TF. In mammalian cells, we have heterologously expressed the *piggyBac* transposase [36] fused to the TF SP1 and shown that the resulting pattern of insertions reflects SP1's DNA binding preferences [35]. However, the method was only feasible in bulk preparations.

Here we present single cell calling cards (scCC), an extension of transposon calling cards that simultaneously profiles mRNA abundance and TF binding at single cell resolution. The key component of our work is a novel construct called the self-reporting transposon (SRT). Using SRTs, the genomic locations of inserted transposons can be mapped from either mRNA or DNA, but the use of mRNA leads to both higher efficiency and compatibility with single-cell transcriptomics. We first establish that TF-directed SRTs, in bulk, retain the ability to accurately identify TF binding sites. Next, we demonstrate that the unfused *piggyBac* transposase, through its native affinity for the bromodomain TF BRD4, can be used to identify BRD4-bound super-enhancers (SEs). We then present the scCC method, which allows cell-specific mapping of SRTs from scRNA-seq libraries. This enables, in one experiment, concomitant assignment of cell types and identification of TF binding sites within those cells. As a proof-of-concept, we use scCC to map BRD4 and SP1 sites in mixtures of cultured human cells. We conclude by identifying cell type-specific Brd4 binding sites *in vivo* in the postnatal mouse cortex. These results demonstrate that scCC could be a broadly applicable tool for the study of specific TF binding interactions across all cell types within heterogeneous tissues.

Results

Self-reporting transposons can be mapped from mRNA instead of genomic DNA

In order to combine scRNA-seq with calling cards, we sought to develop a transposon whose genomic position could be determined from mRNA. We created a *piggyBac* self-reporting transposon (SRT) by removing the polyadenylation signal from our standard DNA-based calling card vector (Fig. 1A). This enables RNA polymerase II (Pol II) to transcribe the reporter gene contained in the transposon and continue through the terminal repeat (TR) into the flanking genomic sequence. Thus, SRTs "self-report" their locations through the unique genomic sequence found within the 3' untranslated regions (UTRs) of the reporter gene transcripts. Although previously published gene- or enhancer-trap transposons could, in principle, also capture local positional information via RNA, they are resolution-limited to the nearest gene or enhancer, respectively [37]. In contrast, the 3' UTRs of SRT-derived transcripts contain the transposon-genome junction in the mRNA sequence, so we can map insertions with base-pair precision.

SRTs are mapped following reverse transcription (RT) and PCR amplification of self-reporting transcripts. These transcripts contain stretches of adenines that are derived from either cryptic polyadenylation signals (PAS) or polyadenine tracts encoded in genomic DNA downstream of the SRT insertion point (Fig. 1B). A poly(T) RT primer hybridizes with these transcripts and introduces a universal priming site at one end of the transcripts. A pair of nested PCRs with an intermediate tagmentation [38] step enable recovery of the transposon-genome junction. After adapter trimming and alignment, the 5' coordinates of these reads identify the genomic locations of insertions in the library. Libraries generated without transposase produce very few genomically mapped reads but the protocol is highly efficient when transposase is added (Supp. Fig. 1A).

To compare transposon recovery between the new RNA-based protocol and our standard DNA-based inverse PCR protocol [35], we transfected HCT-116 cells with a plasmid carrying a *piggyBac* SRT (PB-SRT-Puro) and a plasmid encoding a fusion of the TF SP1 and *piggyBac* transposase (SP1-PBase; Fig. 1A). After two weeks of selection, we obtained approximately 2,300 puromycin-resistant clones. We split these cells in half: one half underwent inverse PCR while the other half were processed with our new RNA-based method. With inverse PCR, we obtained 31,001 insertions (mean coverage: 709 reads per insertion), while the RNA-based protocol recovered 62,500 insertions (mean coverage: 240 reads per insertion). About 80% of insertions recovered by DNA calling cards were also recovered in the RNA-based library (25,060 insertions; Fig. 1C), an overlap comparable to that between technical replicates of

the RNA workflow (Supp. Fig. 1B). However, the RNA protocol recovered a further 37,440 insertions that were not found in the DNA-based library. To determine if these extra insertions were genuine, we analyzed the distribution of insertions by genetic annotation (Fig. 1D) or chromatin state (Supp. Fig. 1C; Supp. Table 1). Transposons mapped from either the DNA or the RNA libraries showed comparable distribution into annotated domains of particular functional or chromatin states, indicating that RNA recovery of transposons appears to be unbiased with respect to our established, DNA-based protocol.

Since *piggyBac* is known to preferentially insert near active chromatin [39], we wondered whether SRT recovery was biased towards euchromatic regions. Prior studies have shown that the *Sleeping Beauty* transposase [40,41] has very little preference for chromatin state [39]. We created a self-reporting *Sleeping Beauty* transposon and compared its genome-wide distribution to that of SRTs deposited by wild-type *piggyBac* (Supp. Fig. 2A-B). Undirected *piggyBac* transposases appeared to modestly prefer transposing into promoter and enhancers, which is consistent with previous reports [39,42] (Supp. Table 1). By contrast, *Sleeping Beauty* showed largely uniform rates of insertions across all chromatin states, including repressed and inactive chromatin (Supp. Fig. 2B). 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.

A common artifact observed in DNA-based transposon recovery is a large fraction of reads mapping back 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 [35], these methods do not completely eliminate background and are not compatible with all experimental paradigms, viral transduction in particular. To reduce this artifact, we included a hammerhead ribozyme [43] in the SRT plasmid downstream of the 5' 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 SP1-PBase DNA library aligned to the plasmid, compared to fewer than 1% of reads from the RNA library (Supp. Fig. 1D). Thus, the addition of a self-cleaving ribozyme virtually eliminates recovery of un-excised transposons.

SP1 fused to piggyBac *directs SRT insertions to SP1 binding sites*

We next sought to confirm that RNA calling cards, in bulk, can still 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* [44] (SP1-HyPBase). As controls, we also transfected a similar number of replicates with undirected PBase or HyPBase, respectively. We obtained 411,287 insertions from SP1-PBase and 1,523,169 insertions from PBase. Similarly, we obtained 2,033,229 SP1-HyPBase insertions and 5,779,101 insertions from HyPBase.

Fig. 1E and Supp. Fig. 4A show the redirection of SRT calling cards by SP1-PBase and SP1-HyPBase, respectively, to three representative SP1-bound regions of the genome. Each circle in the insertions track is an independent transposition event whose genomic position is on the *x*-axis. The *y*-axis is the number of reads supporting each insertion on a log_{10} scale. To better compare transposition across libraries with different numbers of insertions, we plotted the normalized local insertion rate as a density track. All three of the loci depicted in Fig. 1E and Supp. Fig. 4A 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 = 0.84 and 0.96, respectively; Supp. Fig. 3A and Supp. Fig. 4B). Furthermore, calling card peaks were highly enriched for SP1 ChIP-seq signal at their centers, both on average (Supp. Fig. 3B and Supp. Fig. 4C) and in aggregate (Supp. Fig. 3C and Supp. Fig. 4D). SP1 is known to preferentially bind near TSSs [45,46] and is also thought to play a role in demethylating CpG islands [47-49]. We confirmed that the SP1-directed transposases preferentially inserted SRT calling cards near TSSs, CpG islands, and unmethylated CpGs at statistically significant frequencies ($p < 10^{-9}$ in each instance, *G* test of independence; Supp. Fig. 3D and Supp. Fig. 4E). Moreover, compared to undirected *piggyBac*, SP1-directed *piggyBac* showed a striking preference for depositing insertions into promoters (Supp. Fig. 2A-B). Lastly, regions targeted by SP1-PBase and SP1-HyPBase were enriched for the canonical SP1 DNA binding motif ($p < 10^{-70}$ in each instance; Supp. Fig. 3E and Supp. Fig. 4F). Taken together, these results indicate that SP1 can redirect *piggyBac* SRTs near SP1 binding sites.

Clustering of undirected piggyBac *insertions identifies BRD4-bound super-enhancers*

Previous studies have shown that the undirected *piggyBac* transposase preferentially inserts transposons near super-enhancers (SEs) [39], a unique regulatory element that is thought to play a critical role in regulating cell identity [50]. SEs are often enriched for the histone modification H3K27ac as well as RNA polymerase II and general transcription factors like the mediator element MED1 and the bromodomain

protein BRD4 [50-52]. Moreover, the *piggyBac* transposase has a strong biophysical affinity for BRD4, as these proteins can be co-immunoprecipitated [42]. We hypothesized that, given the millions of insertions we assayed from the undirected PBase and HyPBase controls in the SP1-directed experiments (Fig. 1E, Supp. Fig. 4A), we would be able to identify BRD4-bound SEs simply from the localization of undirected *piggyBac* transpositions.

Both undirected PBase and HyPBase showed non-uniform densities of insertions at loci bound by BRD4 (Fig. 2A, Supp. Fig. 7). At statistically significant peaks of *piggyBac* calling cards, PBase and HyPBase showed high reproducibility of normalized insertions between biological replicates (Fig. 2B, Supp. Fig. 5B). Next, we calculated the mean BRD4 enrichment, as assayed by ChIP-seq [53], across these peaks. *piggyBac* peaks showed significantly increased BRD4 signal compared to a genome-wide permutation of the peaks ($p < 10^{-9}$ in both instances, Kolmogorov-Smirnov test; Fig. 2C and Supp. Fig. 5C). Maximum BRD4 ChIP-seq signal was observed at calling card peak centers and decreased symmetrically in both directions. We also found that *piggyBac* peaks show striking ChIP-seq patterns for several histone modifications [54,55], in particular an enrichment for H3K27ac ChIP-seq signal (Fig. 2D, Supp. Fig. 5D). Since bromodomains bind acetylated histones, this observation further supports the hypothesis that undirected *piggyBac* insertions can be used to map BRD4 binding. These peaks were also enriched in H3K4me1, another canonical enhancer mark, and depleted for H3K9me3 and H3K27me3, modifications associated with repressed chromatin [56]. Taken together, these results demonstrate that *piggyBac* insertion density is highly correlated with BRD4 binding throughout the genome and that regions enriched for undirected *piggyBac* insertions share features common to enhancers.

We next assessed whether *piggyBac* peaks can be used to identify BRD4-bound SEs. We used BRD4 ChIP-seq data from HCT-116 cells [53] to create a reference list of BRD-bound SEs [51,57] (Fig. 2A, Supp. Fig. 5A). We then constructed receiver-operator characteristic curves based on our ability to detect SEs from PBase- and HyPBase-derived peaks (Fig. 2E and Supp. Fig. 5E). The high areas under the curve (0.98 in each instance) indicate that we can robustly identify BRD4-bound SEs from *piggyBac* transpositions. Across a range of sensitivities, calling card peaks are highly specific and have high positive predictive value (AUPRC = 0.92 in each instance; Fig. 2F and Supp. Fig. 5F). Thus, undirected *piggyBac* transpositions are an accurate assay of BRD4-bound SEs.

To better understand the relationship between SE sensitivity and the number of insertions recovered, we downsampled the data from the PBase and HyPBase experiments in half-log increments (Supp. Fig. 6A-B). These heatmaps show that sensitivity increases with the total number of insertions recovered. Since

we cannot predict how many, or few, insertions future experiments will yield, we also performed linear interpolation on the downsampled data. The resulting contour plots (Supp. Fig. 6C-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%.

Single cell calling cards enables simultaneous identification of cell type and cell type-specific TF binding sites

We next sought to recover SRTs from scRNA-seq libraries. This would enable us to identify cell types from transcriptomic clustering and, using the same source material, profile TF binding in those cell types. We adopted the 10x Chromium platform given its high efficiency of cell and transcript capture as well as its ease of use [58]. Like many microfluidic scRNA-seq approaches [59,60], the cell barcode and unique molecular index (UMI) are attached to the 3' ends of transcripts. This poses a molecular challenge for SRTs since the junction between the transposon and the genome may be many kilobases away, precluding the use of high-throughput short read sequencing. To overcome this barrier, we developed a circularization strategy to physically bring the cell barcode in apposition to the insertion site (Fig. 3A).

We used a modified version of the bulk SRT amplification protocol where we amplified with primers that bound to the universal priming sequence next to the cell barcode and the terminal sequence of the *piggyBac* TR. These primers were biotinylated and carried a 5' phosphate group. The PCR products of this amplification were diluted and allowed to self-ligate overnight. They were then sheared and captured with streptavidin-coated magnetic beads. The rest of library was prepared on-bead and involved end repair, A-tailing, and adapter ligation. A final PCR step added the required Illumina sequences for high-throughput sequencing. The standard Illumina read 1 primer sequenced the cell barcode and UMI, while a custom read 2 primer, annealing to the end of the *piggyBac* 5' TR, sequenced into the genome. Thus, we collected both the location of a *piggyBac* insertion as well as its cell of origin. We call this method single cell calling cards (scCC).

We validated the method by performing a species-mixing experiment using human HCT-116 cells and mouse N2a cells. Cells were mixed prior to droplet generation and the resulting emulsion was processed through first strand synthesis. At this point, half of the RT product was amplified according to the standard 10x protocol. The resulting scRNA-seq revealed strong species separation with an estimated multiplet rate of 3.2% (Supp. Fig. 8A). The remainder of the first strand synthesis was used for the scCC protocol. We restricted our calling card analysis to those insertions whose cell barcodes were observed in the scRNA-seq library. The distribution of insertions across these cells reflected a continuum from pure

mouse to pure human (Supp. Fig. 8B-C). Since intramolecular ligation and subsequent PCR may introduce unwanted artifacts, such as mis-assignment of a barcode from cell type A to an insertion site in cell type B, we required that a given insertion in a given cell must have at least two different UMIs associated with it. Imposing this filter improved the number of pure mouse and human cells (Supp. Fig. 8D), yielding clear species separation with an estimated multiplet rate of 7.8% (Fig. 3B). This establishes that our method can accurately map calling card insertions in single cells.

We then asked whether scCC could discern cell type-specific TF binding. We transfected two human cell lines, HCT-116 and K562, with HyPBase and PB-SRT-Puro and mixed them together. The resulting scRNA-seq libraries clearly identified the two major cell populations (Fig. 3C; Supp. Fig. 9A). We then prepared scCC libraries from these cells and used the cell barcodes from the HCT-116 and K562 clusters to assign insertions to the two different cell types. We obtained 44,214 insertions from 12,891 HCT-116 cells (mean 3.4 insertions per cell; mean 136 reads per insertion) and 132,994 insertions from 11,912 K562 cells (mean 11 insertions per cell; mean 103 reads per insertion). The distribution of insertions per cell varied by cell type (Supp. Fig. 9D) and does not appear to be correlated with differences in total RNA content (Supp. Fig. 9B-C). Over 93% and 97% of HCT-116 and K562 cells, respectively, had at least one insertion event. Using scCC insertion data alone, we called peaks and successfully identified BRD4-bound loci that were specific to HCT-116 cells, shared between HCT-116 and K562, and specific to K562 cells, respectively (Fig. 3D). Both HCT-116 and K562 peaks showed statistically significant enrichment for BRD4 ChIP-seq signal ($p < 10^{-9}$ in both instances, Kolmogorov-Smirnov test; Supp. Fig. 9E-F). From our earlier downsampling analysis, we estimated that with a p-value cutoff of 10^{-9}, our sensitivity for detecting BRD4-bound SEs would be approximately 60% (Supp. Fig. 6D). The actual sensitivity at this level of recovery was 64%, validating that downsampling analysis can reasonably estimate the performance of scCC. In all, these experiments demonstrate that scCC can be used to deconvolve cell type-specific TF binding.

Since these Brd4 binding sites were identified using undirected HyPBase, we also sought to confirm that TF-*piggyBac* fusions would still work with scCC. We transfected HCT-116 cells with SP1-HyPBase and then performed scRNA-seq. We made scCC libraries from these experiments and identified 92,406 insertions from 30,682 cells (mean 3 insertions per cell; mean 129 reads per insertion). Over 84% of cells had at least one insertion. While previous studies have reported decreased activity of TF-*piggyBac* fusions [61], we observed similar distributions of insertions recovered per cell between HyPBase and SP1-HyPBase (Supp. Fig. 9G). As was observed in bulk (Supp. Fig. 4A), SP1-HyPBase-directed insertions recovered from single cells localize to SP1 binding sites (Fig. 3E). Finally, we investigated the

reproducibility of the scCC method. Both single cell HyPBase and SP1-HyPBase showed high concordance between biological replicates at statistically significant peaks (Supp. Fig. 9H-I). Collectively, these experiments establish that scCC can be used to identify cell type-specific binding sites of both bromodomain and DNA-binding TFs.

Single cell calling cards deconvolves cell type-specific Brd4 binding sites in the mouse cortex

To establish broad utility for scCC, we sought to record TF binding *in vivo*. Since *in vivo* models preclude puromycin selection, we designed an SRT carrying the fluorescent reporter tdTomato (Fig. 4A) and tested this reagent in cell culture. When this construct was transfected without transposase, 3.4% of cells register as tdTomato-positive, likely due to the action of the self-cleaving ribozyme downstream of the transposon. However, when the construct was co-transfected with PBase or HyPBase, this figure rose to 33% and 48%, respectively, corresponding to 11- and 16-fold increases in signal (Fig. 4B). In addition, cells transfected with only the fluorescent SRT produced very few reads that mapped to the genome, while the overwhelming majority of reads from cells co-transfected with transposase mapped to genomic insertions (Supp. Fig. 1A). Thus, this new construct, PB-SRT-tdTomato, allows us to select cells carrying calling card insertions by fluorescence activated cell sorting (FACS).

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 [12,62-65]. We separately packaged the PB-SRT-tdTomato and HyPBase constructs in AAV9 viral particles [66] and delivered mixtures of both viruses to the developing mouse cortex via intracranial injections at P1. After 2-4 weeks, we dissected the cortex, dissociated it to a single cell suspension, performed FACS to isolate tdTomato-positive cells, and analyzed these cells by scRNA-seq and scCC using the 10x Chromium platform. We collected nine libraries in total comprising 35,950 cells and 113,859 insertions (Supp. Table 2). We clustered cells by their mRNA profiles and used established marker genes to classify different cell types (Supp. Fig. 10A-B) [63-65]. The two major cell populations recovered were neurons and astrocytes (Fig. 4C, Supp. Table 2), which is consistent with the known tropism of AAV9 [67]. We also identified a spectrum of differentiating oligodendrocytes and trace amounts of microglial, vascular, and ependymal cells. We then used the cell barcodes shared between the scRNA-seq and scCC libraries to assign insertions to specific cell types.

To determine whether scCC could recover biological differences between cell types *in vivo*, we analyzed HyPBase insertions in neurons and astrocytes, excluding neuroblasts and astrocyte-neuron doublets. We collected 90,299 insertions from 25,158 neurons and 17,102 insertions across 4,727 astrocytes. We then

called peaks on the insertions within each cluster and identified astrocyte-specific, neuron-specific, and shared Brd4 binding sites (Fig. 4D). Brd4 ChIP-seq has not been reported for the mouse brain, but as Brd4 is known to bind acetylated histones, we compared our peak calls to a recent cortical H3K27ac ChIP-seq dataset [68]. Although this dataset was agglomerated over all cell types in the brain, we nevertheless found that peaks in both astrocytes and neurons showed statistically significant enrichment of H3K27ac signal (Supp. Fig. 11A, C; Kolmogorov-Smirnov p-value < 10^{-9} in each case). Brd4 is also thought to mark cell type-specific genes, so we identified genes that overlapped or were near astrocyte or neuron peaks and evaluated the specificity of expression of these genes. We identified 399 genes near astrocyte peaks and 211 genes near neuron peaks. We used bulk RNA-seq data from purified populations of cells [69] to assign gene expression values for each gene and plotted the distribution of these values along a continuum from purely astrocytic expression to purely neuronal expression. Genes near astrocyte peaks were more likely to be specifically expressed in astrocytes, and vice-versa for genes near neuron peaks (Fig. 4E). Gene Ontology enrichment analysis [70] on the astrocyte gene list included "gliogenesis," and "glial cell differentiation," as well as copper metabolism (Supp. Fig. 11B), a known function of astrocytes [71]; while the neuronal gene list was enriched for terms related to synapse assembly and neuron development (Supp. Fig. 11D). Overall, we conclude that scCC can accurately identify cell type-specific Brd4 binding sites *in vivo.*

Finally, we wondered if scCC could discriminate Brd4 binding between closely related cell types. From our scRNA-seq data (Fig. 5B; Supp. Fig. 10A-B), we identified upper and lower layer cortical excitatory neurons and compared HyPBase scCC data between them to identify shared and specific Brd4-bound loci (Fig. 5A). From 9,083 upper cortical neurons we obtained 30,225 insertions, which was on par with the 32,434 insertions collected from 6,980 lower cortical neurons. As a positive control, we identified a shared Brd4 binding site at the *Pou3f3* (*Brn-1*) locus (Fig. 5A, $p < 10^{-9}$). *Pou3f3* was broadly expressed in both populations (Fig. 5C) and has been used to label layers 2-5 of the postnatal cortex [72,73]. We then identified differentially-bound regions in each cluster using insertions from the other cluster as a control. Upper cortical neurons showed specific Brd4 binding at *Pou3f2* (*Brn-2*), which is more restricted to layers 2-4 than *Pou3f3* [73,74], while lower cortical neurons showed Brd4 binding at *Bcl11b* (*Ctip2*) and *Foxp2*, common markers of layer 5 and layer 6 neurons, respectively (Fig. 5A; $p < 10^{-9}$ in each instance) [73,75]. The expression patterns of these genes mirrored Brd4's binding specificity, with *Pou3f2*'s expression mostly retained to the layer 2-4 cluster and the expression of *Bcl11b* and *Foxp2* restricted to the layer 5-6 neuron population (Fig. 5C). This demonstrates that scCC can identify differentially bound loci between very similar cell types.

Discussion

Mapping TF binding in heterogeneous tissues is a challenging problem because traditional methods combine signals from multiple cell types into a single, agglomerated profile. The difficulty is further compounded if individual cell types are difficult to identify, isolate, or are rare, precluding their study. Single cell RNA-seq is a promising paradigm for handling such heterogeneity. Until now, it has been impossible to directly study the actions of individual TFs and connect them to specific cell states. We have presented a new method, single cell calling cards (scCC), that enables simultaneous identification of cell types and TF binding sites from complex mixtures and tissues. This is an important addition to the single cell repertoire and fills a recognized void in the field [76,77]. We anticipate this technique will enable researchers to study the consequences of TF binding in a variety of *ex vivo* and *in situ* models.

A concern with any transposon-based technique is the potential for deleterious interruption of target genes leading to cell death and, consequently, false negatives. Previous experiments in diploid yeast found that calling cards are deposited into promoters of essential and non-essential genes at comparable frequencies [34]. Since mammalian genomes have much larger intergenic regions than yeast, human and mice genomes are likely also able to tolerate calling card transpositions. Indeed, that we were able to deposit SRTs in the developing mouse brain into enhancers and super-enhancers suggests a small mutagenic burden.

One of the limitations of this technique is the relatively few insertions recovered on a per-cell basis, inflating the number of cells that must be analyzed to achieve good sensitivity. Previous studies have reported up to 15-30 insertions per cell for PBase [78-81], and likely higher for HyPBase [44,82]. While we observed similar performance from bulk RNA, we recovered fewer insertions per cell than this, on average, in our single cell experiments. This is likely due to the low capture rate of mRNA transcripts, which is common to all scRNA-seq methods [83]. The inclusion of cis-regulatory features known to enhance mRNA maturation and stability, such as the woodchuck hepatitis virus post-transcriptional regulatory element (WPRE) may increase representation of SRTs in scRNA-seq libraries. Furthermore, as the transcript capture rates of scRNA-seq technologies improve, we expect the sensitivity of our method will increase. The sensitivity of scCC can also be improved by simply analyzing larger numbers of cells, such as with Cell Hashing [84] or combinatorial barcoding [62]. Since the per-cell costs for scRNA-seq are exponentially falling [85], we expect that scCC can be used to analyze TF binding in even very rare cell types in the near future.

Our scCC experiments employed the *piggyBac* transposase, but for some applications, other transposases may prove advantageous. *piggyBac* inserts almost exclusively into TTAA tetranucleotides. For TFs that bind GC-rich regions or have high GC-content motifs, *piggyBac* fusions may have a difficult time finding nearby insertion sites. *Sleeping Beauty*, which inserts into TA dinucleotides, or *Tol2*, which does not have a strict insertion site preference [39], could be used to overcome these limitations. However, the natural affinity of the *piggyBac* transposase for BRD4 makes it the ideal choice for the study of BRD4-bound SEs, which play important regulatory roles in development and disease [52]. It is unclear why *piggyBac* shows such an affinity. BRD4 has an intrinsically disordered region and cooperative interactions between BRD4 and coactivators like MED1 may mediate the formation intranuclear condensates [86] at SEs. One hypothesis is that *piggyBac* has a similarly disordered domain that allows it to preferentially enter condensates and enrich SEs with insertions.

The defining feature of the scCC method is the self-reporting transposon (SRT). While here we have reported the *piggyBac* and *Sleeping Beauty* SRTs, the self-reporting paradigm should be generalizable to any transposon lacking a polyadenylation signal (PAS) in at least one terminal repeat. Expanding the palette of SRTs will illuminate the genome-wide behaviors of transposases and may yield further insight into chromatin dynamics [39]. Simultaneous expression of many TFs, each tagged to a different transposase, may also enable multiplexed studies of TF binding in the same cells. 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 PASs prevent anti-sense transcription [87], which suggests that PASs may be more common in the genome than previously appreciated. Targeted 3'-end sequencing [88,89] of SRT libraries should help resolve this question, while long-read sequencing of self-reporting transcripts may identify non-canonical PASs. Finally, SRTs could lead to new single cell transposon-based assays. For example, just as CRISPR/Cas9 has been combined with scRNA-seq to read out the transcriptional effects of gene deletion [90,91], SRTs will allow transposon mutagenesis screens to be read out by scRNA-seq in a highly parallel fashion.

Finally, as calling card insertions are genomically integrated and preserved through mitosis, they could serve as records of molecular memory. The use of an inducible transposase [92] would enable the recording and identification of temporally-restricted TF binding sites. This would help uncover the stepwise order of events underlying the regulation of specific genes and inform cell fate decision making.

More generally, transposon insertions could serve as barcodes of developmental lineage. Single transposition events have been used to delineate relationships during hematopoiesis [93,94]. Multiplexing several SRTs across every cell in an organism could code lineage in a cumulative and combinatorially diverse fashion, generating high-resolution cellular phylogenies.

Acknowledgements

We would like to thank Jessica Hoisington-Lopez and MariaLynn Crosby from the DNA Sequencing Innovation Lab at The Edison Family Center for Genome Sciences and Systems Biology for their sequencing expertise. Additional sequencing services were performed by the Genome Technology Access Center in the Department of Genetics at Washington University School of Medicine. The Center is partially supported by NCI Cancer Center Support Grant P30 CA91842 to the Siteman Cancer Center and by ICTS/CTSA Grant UL1 TR000448 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), and NIH Roadmap for Medical Research. This work was also supported by the Hope Center Viral Vectors Core and a P30 Neuroscience Blueprint Interdisciplinary Center Core award to Washington University (P30 NS057105). Finally, we would like to thank Donald Conrad and Ben Humphreys for their advice and constructive feedback during this project.

This publication is solely the responsibility of the authors and does not necessarily represent the official view of NCRR or NIH.

Funding

This work was supported by NIH grants R21 HG009750 (R.D.M. and S.A.M.), U01 MH109133 (J.D.D. and R.D.M.), and RF1 MH117070 (J.D.D. and R.D.M.), as well as a grant from the Children's Discovery Institute (#MC-II-2016-533; R.D.M.). A.M. was supported by NIH grants T32 GM007200, T32 HG000045, and F30 HG009986. A.J.C was supported by NIH T32 GM008151, M.J.V. was supported by NIH F32 NS105363, and T.L. was supported by NIH T32 GM007067.

Author Contributions

A.M., M.N.W., Z.Q., and R.D.M. developed the self-reporting transposon (SRT) technology. A.M. and M.N.W. created and optimized the molecular workflow for recovering SRTs from bulk RNA-seq libraries, with contributions from Z.Q. A.M. developed and optimized the protocol for recovering SRTs from single cell RNA-seq libraries. S.A.M. provided guidance and assistance with single cell experiments. Z.Q. and T.L. cloned SRT constructs. A.M., M.N.W., J.D.D., and R.D.M. designed the experiments. A.M., X.C., and J.H. performed the *in vitro* experiments. M.N.W., A.J.C., and M.J.V performed the *in vivo* experiments. A.M., X.C., and J.H. generated sequencing libraries. A.M., J.D.D., and R.D.M. analyzed the data. A.M. and R.D.M. wrote the manuscript in consultation with all authors.

References

1. Zhu X, Zuo H, Maher BJ, Serwanski DR, LoTurco JJ, Lu QR, et al. Olig2-dependent developmental fate switch of NG2 cells. Development. Oxford University Press for The Company of Biologists Limited; 2012;139: 2299–2307. doi:10.1242/dev.078873
2. Mizuguchi R, Sugimori M, Takebayashi H, Kosako H, Nagao M, Yoshida S, et al. Combinatorial Roles of Olig2 and Neurogenin2 in the Coordinated Induction of Pan-Neuronal and Subtype-Specific Properties of Motoneurons. Neuron. Elsevier; 2001;31: 757–771. doi:10.1016/S0896-6273(01)00413-5
3. Hafler BP, Surzenko N, Beier KT, Punzo C, Trimarchi JM, Kong JH, et al. Transcription factor Olig2 defines subpopulations of retinal progenitor cells biased toward specific cell fates. PNAS. National Academy of Sciences; 2012;109: 7882–7887. doi:10.1073/pnas.1203138109
4. Gurdon JB. Cell Fate Determination by Transcription Factors. Essays on Developmental Biology, Part A. Elsevier; 2016. pp. 445–454. doi:10.1016/bs.ctdb.2015.10.005
5. Takahashi K, Yamanaka S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. Cell. 2006;126: 663–676. doi:10.1016/j.cell.2006.07.024
6. Liu X, Huang J, Chen T, Wang Y, Xin S, Li J, et al. Yamanaka factors critically regulate the developmental signaling network in mouse embryonic stem cells. Cell Res. 2008;18: 1177–1189. doi:10.1038/cr.2008.309
7. Fogarty NME, McCarthy A, Snijders KE, Powell BE, Kubikova N, Blakeley P, et al. Genome editing reveals a role for OCT4 in human embryogenesis. Nature. 2017;9: 346. doi:10.1038/nature24033
8. Davis RL, Weintraub H, Lassar AB. Expression of a single transfected cDNA converts fibroblasts to myoblasts. Cell. Cell Press; 1987;51: 987–1000. doi:10.1016/0092-8674(87)90585-X
9. Kvon EZ, Kamneva OK, Melo US, Barozzi I, Osterwalder M, Mannion BJ, et al. Progressive Loss of Function in a Limb Enhancer during Snake Evolution. Cell. 2016;167: 633–642.e11. doi:10.1016/j.cell.2016.09.028
10. Gonen N, Futtner CR, Wood S, Alexandra Garcia-Moreno S, Salamone IM, Samson SC, et al. Sex reversal following deletion of a single distal enhancer of Sox9. Science. American Association for the Advancement of Science; 2018;360: 1469–1471. doi:10.1126/science.aas9408
11. Lee TI, Young RA. Transcriptional Regulation and Its Misregulation in Disease. Cell. Elsevier; 2013;152: 1237–1251. doi:10.1016/j.cell.2013.02.014
12. Zeisel A, Munoz-Manchado AB, Codeluppi S, Lonnerberg P, La Manno G, Jureus A, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. American Association for the Advancement of Science; 2015;347: 1138–1142. doi:10.1126/science.aaa1934
13. Fincher CT, Wurtzel O, de Hoog T, Kravarik KM, Reddien PW. Cell type transcriptome atlas for the planarian Schmidtea mediterranea. Science. American Association for the Advancement of Science; 2018;20: eaaq1736–757. doi:10.1126/science.aaq1736
14. Han X, Wang R, Zhou Y, Fei L, Sun H, Lai S, et al. Mapping the Mouse Cell Atlas by Microwell-Seq. Cell. Elsevier; 2018;172: 1091–1107.e17. doi:10.1016/j.cell.2018.02.001
15. Karaiskos N, Wahle P, Alles J, Boltengagen A, Ayoub S, Kipar C, et al. The Drosophilaembryo at single-cell transcriptome resolution. Science. 2017;8: eaan3235. doi:10.1126/science.aan3235
16. Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. Science. American Association for the Advancement of Science; 2017;357: 661–667. doi:10.1126/science.aam8940
17. Campbell JN, Macosko EZ, Fenselau H, Pers TH, Lyubetskaya A, Tenen D, et al. A molecular census of arcuate hypothalamus and median eminence cell types. Nature Neuroscience. 2017;20: 484–496. doi:10.1038/nn.4495
18. Dey SS, Kester L, Spanjaard B, Bienko M, van Oudenaarden A. Integrated genome and transcriptome sequencing of the same cell. Nat Biotechnol. Nature Publishing Group; 2015;33: 285–289. doi:10.1038/nbt.3129
19. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nature Methods. Nature Research; 2015;12: 519–522. doi:10.1038/nmeth.3370
20. Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nature Methods. Nature Publishing Group; 2016;13: 229–232. doi:10.1038/nmeth.3728
21. Clark SJ, Argelaguet R, Kapourani C-A, Stubbs TM, Lee HJ, Alda-Catalinas C, et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. Nat Commun. Nature Publishing Group; 2018;9: 390. doi:10.1038/s41467-018-03149-4
22. Cao J, Cusanovich DA, Ramani V, Aghamirzaie D, Pliner HA, Hill AJ, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science. American Association for the Advancement of Science; 2018;33: eaau0730. doi:10.1126/science.aau0730
23. Peterson VM, Zhang KX, Kumar N, Wong J, Li L, Wilson DC, et al. Multiplexed quantification of proteins and transcripts in single cells. Nat Biotechnol. 2017;9: 2579. doi:10.1038/nbt.3973
24. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, et al. Simultaneous epitope and transcriptome measurement in single cells. Nature Methods. Nature Publishing Group; 2017;14: 865–868. doi:10.1038/nmeth.4380

25. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science. American Association for the Advancement of Science; 2007;316: 1497–1502. doi:10.1126/science.1141319
26. Rotem A, Ram O, Shoresh N, Sperling RA, Goren A, Weitz DA, et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. Nat Biotechnol. Nature Publishing Group; 2015;33: 1165–1172. doi:10.1038/nbt.3383
27. Greil F, Moorman C, van Steensel B. DamID: Mapping of In Vivo Protein–Genome Interactions Using Tethered DNA Adenine Methyltransferase. DNA Microarrays, Part A: Array Platforms and Wet-Bench Protocols. Elsevier; 2006. pp. 342–359. doi:10.1016/S0076-6879(06)10016-6
28. Vogel MJ, Peric-Hupkes D, van Steensel B. Detection of in vivo protein–DNA interactions using DamID in mammalian cells. Nature Protocols. 2007;2: 1467–1478. doi:10.1038/nprot.2007.148
29. Kind J, Pagie L, Ortabozkoyun H, Boyle S, de Vries SS, Janssen H, et al. Single-Cell Dynamics of Genome-Nuclear Lamina Interactions. Cell. Cell Press; 2013;153: 178–192. doi:10.1016/j.cell.2013.02.028
30. Kind J, Pagie L, de Vries SS, Nahidiazar L, Dey SS, Bienko M, et al. Genome-wide Maps of Nuclear Lamina Interactions in Single Human Cells. Cell. Cell Press; 2015;163: 134–147. doi:10.1016/j.cell.2015.08.040
31. Rooijers K, Markodimitraki C, Rang F, de Vries S, Chialastri A, de Luca K, et al. Simultaneous quantification of protein-DNA contacts and transcriptomes in single cells. bioRxiv. Cold Spring Harbor Laboratory; 2019;: 529388. doi:10.1101/529388
32. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. Nature Research; 2015;523: 486–490. doi:10.1038/nature14590
33. Wang H, Johnston M, Mitra RD. Calling cards for DNA-binding proteins. Genome Res. Cold Spring Harbor Lab; 2007;17: 1202–1209. doi:10.1101/gr.6510207
34. Wang H, Mayhew D, Chen X, Johnston M, Mitra RD. Calling Cards enable multiplexed identification of the genomic targets of DNA-binding proteins. Genome Res. Cold Spring Harbor Lab; 2011;21: 748–755. doi:10.1101/gr.114850.110
35. Wang H, Mayhew D, Chen X, Johnston M, Mitra RD. "Calling Cards" for DNA-Binding Proteins in Mammalian Cells. Genetics. Genetics; 2012;190: 941–949. doi:10.1534/genetics.111.137315
36. Ding S, Wu X, Li G, Han M, Zhuang Y, Xu T. Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. Cell. 2005;122: 473–483. doi:10.1016/j.cell.2005.07.013
37. Cadiñanos J, Bradley A. Generation of an inducible and optimized piggyBac transposon system. Nucl Acids Res. Oxford University Press; 2007;35: e87–e87. doi:10.1093/nar/gkm446
38. Picelli S, Björklund ÅK, Reinius B, Sagasser S, Winberg G, Sandberg R. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. Genome Res. Cold Spring Harbor Lab; 2014;24: 2033–2040. doi:10.1101/gr.177881.114
39. Yoshida J, Akagi K, Misawa R, Kokubu C, Takeda J, Horie K. Chromatin states shape insertion profiles of the piggyBac, Tol2 and Sleeping Beauty transposons and murine leukemia virus. Scientific Reports. Nature Publishing Group; 2017;7: 43613. doi:10.1038/srep43613
40. Ivics Z, Hackett PB, Plasterk RH, Izsvák Z. Molecular Reconstruction of Sleeping Beauty, a Tc1-like Transposon from Fish, and Its Transposition in Human Cells. Cell. Cell Press; 1997;91: 501–510. doi:10.1016/S0092-8674(00)80436-5
41. Mátés L, Chuah MKL, Belay E, Jerchow B, Manoj N, Acosta-Sanchez A, et al. Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. Nat Genet. Nature Publishing Group; 2009;41: 753–761. doi:10.1038/ng.343
42. Gogol-Döring A, Ammar I, Gupta S, Bunse M, Miskey C, Chen W, et al. Genome-wide Profiling Reveals Remarkable Parallels Between Insertion Site Selection Properties of the MLV Retrovirus and the piggyBac Transposon in Primary Human CD4+ T Cells. Molecular Therapy. 2016;24: 592–606. doi:10.1038/mt.2016.11
43. Yen L, Svendsen J, Lee J-S, Gray JT, Magnier M, Baba T, et al. Exogenous control of mammalian gene expression through modulation of RNA self-cleavage. Nature. Nature Publishing Group; 2004;431: 471–476. doi:10.1038/nature02844
44. Yusa K, Zhou L, Li MA, Bradley A, Craig NL. A hyperactive piggyBac transposase for mammalian applications. Proc Natl Acad Sci USA. National Acad Sciences; 2011;108: 1531–1536. doi:10.1073/pnas.1008322108
45. Xi H, Yu Y, Fu Y, Foley J, Halees A, Weng Z. Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1. Genome Res. Cold Spring Harbor Lab; 2007;17: 798–806. doi:10.1101/gr.5754707
46. Frith MC, Valen E, Krogh A, Hayashizaki Y, Carninci P, Sandelin A. A code for transcription initiation in mammalian genomes. Genome Res. Cold Spring Harbor Lab; 2008;18: 1–12. doi:10.1101/gr.6831208
47. Brandeis M, Frank D, Keshet I, Siegfried Z, Mendelsohn M, Nemes A, et al. Sp1 elements protect a CpG island from de novo methylation. Nature. Nature Publishing Group; 1994;371: 435–438. doi:10.1038/371435a0
48. Macleod D, Charlton J, Mullins J, Bird AP. Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. Genes Dev. 1994;8: 2282–2292.
49. Philipsen S, Suske G. A tale of three fingers: the family of mammalian Sp/XKLF transcription factors. Nucl Acids Res. Oxford University Press; 1999;27: 2991–3000. doi:10.1093/nar/27.15.2991
50. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, et al. Super-Enhancers in the Control of Cell Identity and Disease. Cell. 2013;155: 934–947. doi:10.1016/j.cell.2013.09.053

51. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. Cell. 2013;153: 307–319. doi:10.1016/j.cell.2013.03.035
52. Lovén J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, et al. Selective Inhibition of Tumor Oncogenes by Disruption of Super-Enhancers. Cell. Cell Press; 2013;153: 320–334. doi:10.1016/j.cell.2013.03.036
53. McCleland ML, Mesh K, Lorenzana E, Chopra VS, Segal E, Watanabe C, et al. CCAT1 is an enhancer-templated RNA that predicts BET sensitivity in colorectal cancer. J Clin Invest. American Society for Clinical Investigation; 2016;126: 639–652. doi:10.1172/JCI83265
54. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. Nature Publishing Group; 2012;489: 57–74. doi:10.1038/nature11247
55. Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC, et al. ENCODE data at the ENCODE portal. Nucl Acids Res. Oxford University Press; 2016;44: D726–D732. doi:10.1093/nar/gkv1160
56. Lawrence M, Daujat S, Schneider R. Lateral Thinking: How Histone Modifications Regulate Gene Expression. Molecular Cell. 2016;32: 42–56. doi:10.1016/j.tig.2015.10.007
57. Pott S, Lieb JD. What are super-enhancers? Nat Genet. 2014;47: 8–12. doi:10.1038/ng.3167
58. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. Nature Publishing Group; 2017;8: 14049. doi:10.1038/ncomms14049
59. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell. 2015;161: 1202–1214. doi:10.1016/j.cell.2015.05.002
60. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. Cell. 2015;161: 1187–1201. doi:10.1016/j.cell.2015.04.044
61. Wu SC-Y, Meir Y-JJ, Coates CJ, Handler AM, Pelczar P, Moisyadi S, et al. piggyBac is a flexible and highly active transposon as compared to sleeping beauty, Tol2, and Mos1 in mammalian cells. PNAS. National Acad Sciences; 2006;103: 15008–15013. doi:10.1073/pnas.0606979103
62. Rosenberg AB, Roco CM, Muscat RA, Kuchina A, Sample P, Yao Z, et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science. American Association for the Advancement of Science; 2018;360: 176–182. doi:10.1126/science.aam8999
63. Zeisel A, Hochgerner H, Lönnerberg P, Johnsson A, Memic F, van der Zwan J, et al. Molecular Architecture of the Mouse Nervous System. Cell. Elsevier; 2018;174: 999–1014.e22. doi:10.1016/j.cell.2018.06.021
64. Saunders A, Macosko EZ, Wysoker A, Goldman M, Krienen FM, de Rivera H, et al. Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. Cell. Cell Press; 2018;174: 1015–1030.e16. doi:10.1016/j.cell.2018.07.028
65. Tasic B, Yao Z, Graybuck LT, Smith KA, Nguyen TN, Bertagnolli D, et al. Shared and distinct transcriptomic cell types across neocortical areas. Nature. 5 ed. Nature Publishing Group; 2018;563: 72–78. doi:10.1038/s41586-018-0654-5
66. Cammack AJ, Moudgil A, Lagunas T, Vasek MJ, Shabsovich M, He J, et al. Transposon-mediated, cell type-specific transcription factor recording in the mouse brain. bioRxiv. Cold Spring Harbor Laboratory; 2019;: 538504. doi:10.1101/538504
67. Schuster DJ, Dykstra JA, Riedl MS, Kitto KF, Belur LR, McIvor RS, et al. Biodistribution of adeno-associated virus serotype 9 (AAV9) vector after intrathecal and intravenous delivery in mouse. Front Neuroanat. Frontiers; 2014;8: 42. doi:10.3389/fnana.2014.00042
68. Stroud H, Su SC, Hrvatin S, Greben AW, Renthal W, Boxer LD, et al. Early-Life Gene Expression in Neurons Modulates Lasting Epigenetic States. Cell. 2017;171: 1151–1164.e16. doi:10.1016/j.cell.2017.09.047
69. Zhang Y, Chen K, Sloan SA, Bennett ML, Scholze AR, O'Keeffe S, et al. An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex. Journal of Neuroscience. Society for Neuroscience; 2014;34: 11929–11947. doi:10.1523/JNEUROSCI.1860-14.2014
70. Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, et al. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucl Acids Res. Oxford University Press; 2017;45: D183–D189. doi:10.1093/nar/gkw1138
71. Scheiber IF, Dringen R. Astrocyte functions in the copper homeostasis of the brain. Neurochemistry International. Pergamon; 2013;62: 556–565. doi:10.1016/j.neuint.2012.08.017
72. Pucilowska J, Puzerey PA, Karlo JC, Galán RF, Landreth GE. Disrupted ERK Signaling during Cortical Development Leads to Abnormal Progenitor Proliferation, Neuronal and Network Excitability and Behavior, Modeling Human Neuro-Cardio-Facial-Cutaneous and Related Syndromes. Journal of Neuroscience. Society for Neuroscience; 2012;32: 8663–8677. doi:10.1523/JNEUROSCI.1107-12.2012
73. Molyneaux BJ, Arlotta P, Menezes JRL, Macklis JD. Neuronal subtype specification in the cerebral cortex. Nature Reviews Neuroscience. Nature Publishing Group; 2007;8: 427–437. doi:10.1038/nrn2151
74. Fan X, Kim H-J, Bouton D, Warner M, Gustafsson J-Å. Expression of liver X receptor β is essential for formation of superficial cortical layers and migration of later-born neurons. PNAS. National Academy of Sciences; 2008;105: 13445–13450. doi:10.1073/pnas.0806974105
75. Rašin M-R, Gazula V-R, Breunig JJ, Kwan KY, Johnson MB, Liu-Chen S, et al. Numb and Numbl are required for maintenance of cadherin-based adhesion and polarity of neural progenitors. Nature Neuroscience. Nature Publishing Group; 2007;10: 819–827. doi:10.1038/nn1924

76. Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nature Reviews Genetics. Nature Research; 2013;14: 618–630. doi:10.1038/nrg3542
77. Shema E, Bernstein BE, Buenrostro JD. Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution. Nat Genet. Nature Publishing Group; 2018;51: 19–25. doi:10.1038/s41588-018-0290-x
78. Kettlun C, Galvan DL, George AL, Kaja A, Wilson MH. Manipulating piggyBac transposon chromosomal integration site selection in human cells. Molecular therapy : the journal of the American Society of Gene Therapy. 2011;19: 1636–1644. doi:10.1038/mt.2011.129
79. Wang W, Lin C, Lu D, Ning Z, Cox T, Melvin D, et al. Chromosomal transposition of PiggyBac in mouse embryonic stem cells. Proc Natl Acad Sci USA. National Academy of Sciences; 2008;105: 9290–9295. doi:10.1073/pnas.0801017105
80. Saridey SK, Liu L, Doherty JE, Kaja A, Galvan DL, Fletcher BS, et al. PiggyBac transposon-based inducible gene expression in vivo after somatic cell gene transfer. Molecular therapy : the journal of the American Society of Gene Therapy. 2009;17: 2115–2120. doi:10.1038/mt.2009.234
81. Wilson MH, Coates CJ, George AL. PiggyBac transposon-mediated gene transfer in human cells. Molecular therapy : the journal of the American Society of Gene Therapy. 2007;15: 139–145. doi:10.1038/sj.mt.6300028
82. Kalhor R, Kalhor K, Mejia L, Leeper K, Graveline A, Mali P, et al. Developmental barcoding of whole mouse via homing CRISPR. Science. American Association for the Advancement of Science; 2018;113: eaat9804. doi:10.1126/science.aat9804
83. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med. Nature Publishing Group; 2018;50: 96. doi:10.1038/s12276-018-0071-8
84. Stoeckius M, Zheng S, Houck-Loomis B, Hao S, Yeung BZ, Mauck WM, et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol. BioMed Central; 2018;19: 224. doi:10.1186/s13059-018-1603-1
85. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nature Protocols. Nature Publishing Group; 2018;13: 599–604. doi:10.1038/nprot.2017.149
86. Coactivator condensation at super-enhancers links phase separation and gene control. Science. 2018;361: eaar3958. doi:10.1126/science.aar3958
87. Transcriptional Pause Sites Delineate Stable Nucleosome-Associated Premature Polyadenylation Suppressed by U1 snRNP. Molecular Cell. 2018;69: 648–663.e7. doi:10.1016/j.molcel.2018.01.006
88. Chen W, Jia Q, Song Y, Fu H, Wei G, Ni T. Alternative Polyadenylation: Methods, Findings, and Impacts. Elsevier; 2017;15: 287–300. doi:10.1016/j.gpb.2017.06.001
89. Zheng D, Liu X, Tian B. 3"READS+, a sensitive and accurate method for 3" end sequencing of polyadenylated RNA. RNA. Cold Spring Harbor Lab; 2016;22: 1631–1639. doi:10.1261/rna.057075.116
90. Datlinger P, Rendeiro AF, Schmidl C, Krausgruber T, Traxler P, Klughammer J, et al. Pooled CRISPR screening with single-cell transcriptome readout. Nature Methods. Nature Research; 2017;14: 297–301. doi:10.1038/nmeth.4177
91. Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, et al. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell. 2016;167: 1853–1866.e17. doi:10.1016/j.cell.2016.11.038
92. Qi Z, Wilkinson MN, Chen X, Sankararaman S, Mayhew D, Mitra RD. An optimized, broadly applicable piggyBac transposon induction system. Nucl Acids Res. Oxford University Press; 2017;: gkw1290. doi:10.1093/nar/gkw1290
93. Sun J, Ramos A, Chapman B, Johnnidis JB, Le L, Ho Y-J, et al. Clonal dynamics of native haematopoiesis. Nature. Nature Research; 2014;514: 322–327. doi:10.1038/nature13824
94. Rodriguez-Fraticelli AE, Wolock SL, Weinreb CS, Panero R, Patel SH, Jankovic M, et al. Clonal analysis of lineage fate in native haematopoiesis. Nature. Nature Publishing Group; 2018;124: 1929. doi:10.1038/nature25168
95. Raff T, van der Giet M, Endemann D, Wiederholt T, Paul M. Design and Testing of β-Actin Primers for RT-PCR that Do Not Co-amplify Processed Pseudogenes. BioTechniques. Future Science Ltd London, UK; 1997;23: 456–460. doi:10.2144/97233st02
96. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011;17: 10–12. doi:10.14806/ej.17.1.200
97. Wolf FA, Angerer P, Theis FJ. SCANPY : large-scale single-cell gene expression data analysis. Genome Biol. BioMed Central; 2018;19: 15. doi:10.1186/s13059-017-1382-0
98. Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database (Oxford). Oxford University Press; 2016;2016: baw100. doi:10.1093/database/baw100
99. Scargle JD, Norris JP, Jackson B, Chiang J. STUDIES IN ASTRONOMICAL TIME SERIES ANALYSIS. VI. BAYESIAN BLOCK REPRESENTATIONS. ApJ. IOP Publishing; 2013;764: 167. doi:10.1088/0004-637X/764/2/167
100. VanderPlas J, Connolly AJ, Ivezic Z, Gray A. Introduction to astroML: Machine learning for astrophysics. IEEE; pp. 47–54. doi:10.1109/CIDU.2012.6382200
101. Brunner AL, Johnson DS, Kim SW, Valouev A, Reddy TE, Neff NF, et al. Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. Genome Res. Cold Spring Harbor Lab; 2009;19: 1044–1056. doi:10.1101/gr.088773.108
102. Machanick P, Bailey TL. MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics. Oxford University Press; 2011;27: 1696–1697. doi:10.1093/bioinformatics/btr189

103. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol. BioMed Central; 2008;9: R137. doi:10.1186/gb-2008-9-9-r137
104. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. Oxford University Press; 2010;26: 841–842. doi:10.1093/bioinformatics/btq033
105. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucl Acids Res. Oxford University Press; 2016;44: W160–W165. doi:10.1093/nar/gkw257
106. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. Nature Publishing Group; 2011;473: 43–49.
107. Castillo-Hair SM, Sexton JT, Landry BP, Olson EJ, Igoshin OA, Tabor JJ. FlowCal: A User-Friendly, Open Source Software Tool for Automatically Converting Flow Cytometry Data from Arbitrary to Calibrated Units. ACS Synth Biol. American Chemical Society; 2016;5: 774–780. doi:10.1021/acssynbio.5b00284
108. Avey D, Sankararaman S, Yim AKY, Barve R, Milbrandt J, Mitra RD. Single-Cell RNA-Seq Uncovers a Robust Transcriptional Response to Morphine by Glia. Cell Reports. Cell Press; 2018;24: 3619–3629.e4. doi:10.1016/j.celrep.2018.08.080
109. Saxena A, Wagatsuma A, Noro Y, Kuji T, Asaka-Oba A, Watahiki A, et al. Trehalose-enhanced isolation of neuronal sub-types from adult mouse brain. BioTechniques. Future Science Ltd London, UK; 2012;52: 381–385. doi:10.2144/0000113878
110. Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, Zhang MD, et al. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. Cell. Cell Press; 2019. doi:10.1016/j.cell.2018.11.029

Methods

Cell culture

HCT-116, N2a, and HEK293T cells were cultured in Dulbecco's Modified Eagle Medium (DMEM; Gibco #11965-084) supplemented with 10% fetal bovine serum (FBS; Peak Serum #PS-FB3) and 1% antibiotic-antimycotic (Anti-Anti; Gibco #15240-062). K562 cells were grown under the same conditions as the HCT-116 and N2a except replacing DMEM with RPMI 1640 Medium (Gibco #11875-085). Cells were grown at 37ºC with 5% carbon dioxide (CO2). Puromycin (Sigma #P8899) was added 24 hours after transfection at a final concentration of 2 µg/ml. Media was replenished every 2 days.

DNA- vs RNA-based recovery

Approximately 500,000 HCT-116 cells were plated in a single well of a 6-well plate. Cells were transfected with 2.5 µg of the SP1-PBase plasmid (for a full list of plasmids, see Supp. Table 3) and 2.5 µg of the PB-SRT-Puro plasmid using Lipofectamine 3000 (Thermo Fisher #L3000015) following manufacturer's instructions. After 24 hours, cells were split and plated 1:10 in each of three 10 cm dishes. Puromycin was then added and colonies were allowed to grow out under selection for two weeks. We obtained approximately 2,300 colonies. All cells were pooled together and split into two populations: one was subjected to DNA extraction, self-ligation, and inverse PCR, as described previously [35]; while the other underwent RNA extraction and SRT library preparation (see below).

In vitro *bulk calling card experiments*

We cotransfected 10-12 replicates of HCT-116 cells with 5 µg of PB-SRT-Puro plasmid and 5 µg PBase plasmid via Neon electroporation (Thermo Fisher #MPK10025). Each replicate contained $2x10^6$ cells. As a negative control, we transfected one replicate of HCT-116 cells with 5 µg PB-SRT-Puro plasmid only. We used the following settings–pulse voltage: 1,530 V; pulse width: 20 ms; pulse number: 1. Each replicate was allowed to recover in a single well of a 6-well plate for 24 hours before being split 1:1 into a 10 cm dish and adding puromycin. Cells were grown under selection for one week, by which time almost all negative control transfectants were dead. We used the same experimental setup for experiments with PB-SRT-Puro and each of SP1-PBase, HyPBase, and SP1-HyPBase plasmids, as well as with SB-SRT-Puro and SB100X plasmids. Each replicate was cultured independently under aforementioned media conditions. After 7 days, we dissociated each replicate with trypsin-EDTA (Sigma #T4049) and created single cell suspensions in phosphate-buffered saline (PBS; Gibco #14190-136). Aliquots of each replicate were cryopreserved in cell culture media (see above) supplemented with 5% DMSO. The remaining cells were pelleted by centrifugation at 300g for 5 minutes. Cell pellets were either processed immediately or kept at -80ºC in RNAProtect Cell Reagent (QIAGEN # 76526).

Isolation of bulk RNA and reverse transcription

Total RNA was isolated from each replicate using the RNEasy Plus Mini Kit (QIAGEN #74134) following manufacturer's instructions. Briefly, cell pellets were resuspended in 600 µl of Buffer RLT Plus with 1% 2-mercaptoethanol (Gibco #21985-023). Cells were homogenzied by vortexing. DNA was removed by running lysate through gDNA Eliminator spin columns, while RNA was bound by passing the flow-through over RNEasy spin columns. An on-column treatment with DNase (QIAGEN #79254) was also performed. After washing, RNA was eluted in 40 µl RNase-free H_2O. RNA was quantitated using the Qubit RNA HS Assay Kit (Thermo Fisher #Q32852).

We performed first strand synthesis on each replicate with Maxima H Minus Reverse Transcriptase (Thermo Fisher #EP0752). We mixed 2 µg of total RNA with 1 µl 10 mM dNTPs (Clontech #639125) and

1 µl of 50 µM SMART_dT18VN primer (for a complete list of primer sequences, see Supp. Table 4), brought the total volume up to 14 µl, and incubated it at 65ºC for 5 minutes. After transferring to ice and letting rest for 1 minute, we added 4 µl 5X Maxima RT Buffer, 1 µl RNaseOUT (Thermo Fisher #10777019), and 1 µl of 1:1 Maxima H Minus Reverse Transcriptase diluted in 1x RT Buffer (100 U). The solution was mixed by pipetting and incubated at 50ºC for 1 hour followed by heat inactivation at 85ºC for 10 minutes. Finally, we digested with 1 µl RNaseH (New England BioLabs #M0297S) at 37ºC for 30 minutes. cDNA was stored at -20ºC.

Amplification of self-reporting transcripts from bulk RNA

The PCR conditions for amplifying self-reporting transcripts (i.e. transcripts derived from self-reporting transposons) involved mixing 1 µl cDNA template with 12.5 µl Kapa HiFi HotStart ReadyMix (Kapa Biosystems #KK2601), 0.5 µl 25 µM SMART primer, and either 1 µl of 25 µM SRT_PAC_F1 primer (in the case of puromycin selection) or 0.5 µl of 25 µM SRT_tdTomato_F1 primer (in the case of tdTomato screening). The mixture was brought up to 25 µl with ddH_2O. Thermocycling parameters were as follows: 95ºC for 3 minutes; 20 cycles of: 98ºC for 20 seconds–65ºC for 30 seconds–72ºC for 5 minutes; 72ºC for 10 minutes; hold at 4ºC forever. As a control, cDNA quality can be assessed with exon-spanning primers for β-actin (see Supp. Table 4 for examples of human primers [95]) under the same thermocycling settings.

PCR products were purified using AMPure XP beads (Beckman Coulter #A63880). 12 µl of resuspended beads were added to the 25 µl PCR product and mixed homogenously by pipetting. After a 5-minute incubation at room temperature, the solution was placed on a magnetic rack for 2 minutes. The supernatant was aspirated and discarded. The pellet was washed twice with 200 µl of 70% ethanol (incubated for 30 seconds each time), discarding the supernatant each time. The pellet was left to dry at room temperature for 2 minutes. To elute, we added 20 µl ddH_2O to the pellet, resuspended by pipetting, incubated at room temperature for 2 minutes, and placed on a magnetic rack for one minute. Once clear, the solution was transferred to a clean 1.5 ml tube. DNA concentration was measured on the Qubit 3.0 Fluorometer (Thermo Fisher #Q33216) using the dsDNA High Sensitivity Assay Kit (Thermo Fisher #Q32851).

Generation of bulk RNA calling card libraries

Calling card libraries from bulk RNA were generated using the Nextera XT DNA Library Preparation Kit (Illumina #FC-131-1024). One nanogram of PCR product was resuspended in 5 µl ddH_2O. To this mixture we added 10 µl Tagment DNA (TD) Buffer and 5 µl Amplicon Tagment Mix (ATM). After pipetting to mix, we incubated the solution in a thermocycler preheated to 55ºC. The tagmentation reaction was halted by adding 5 µl Neutralization Tagment (NT) Buffer and was kept at room temperature for 5 minutes. The final PCR was set up by adding 15 µl Nextera PCR Mix (NPM), 8 µl ddH2O, 1 µl of 10 µM transposon primer (e.g. OM-PB-NNN) and 1 µl Nextera N7 indexed primer. The transposon primer anneals to the end of the transposon terminal repeat–*piggyBac,* in the case of OM-PB primers, or *Sleeping Beauty,* in the case of OM-SB primers–and contains a 3 base pair barcode sequence. Every N7 primer contains a unique index sequence that is demultiplexed by the sequencer. Each replicate was assigned a unique combination of barcoded transposon primer and indexed N7 primer, enabling precise identification of each library's sequencing reads.

The final PCR was run under the following conditions: 95ºC for 30 seconds; 13 cycles of: 95ºC for 10 seconds–50ºC for 30 seconds–72ºC for 30 seconds; 72ºC for 5 minutes; hold at 4ºC forever. After PCR, the final library was purified using 30 µl (0.6x) AMPure XP beads, as described above. The library was eluted in 11 µl ddH_2O and quantitated on an Agilent TapeStation 4200 System using the High Sensitivity D1000 ScreenTape (Agilent #5067-5584 and #5067-5585).

Sequencing and analysis of bulk RNA calling card libraries

Multiple calling card libraries were pooled together for sequencing on the Illumina HiSeq 2500 platform. To increase the complexity of the library, PhiX was added at a final loading concentration of 50%. Reads were demultiplexed by the N7 index sequences added during the final PCR. Read 1 began with the 3 base pair barcode followed by the end of the transposon terminal repeat, culminating with the insertion site motif (TTAA in the case of *piggyBac*; TA in the case of *Sleeping Beauty*) before entering the genome. *piggyBac* reads were checked for exact matches to the barcode, transposon sequence, and insertion site at the beginning of reads before being hard trimmed using cutadapt [96] with the following settings: -g "^NNNTTTACGCAGACTATCTTTCTAGGGTTAA" --minimum-length 1 --discard-untrimmed -e 0 --no-indels, where NNN is replaced with the primer barcode. *Sleeping Beauty* libraries were trimmed with the following settings: -g "^NNNTAAGTGTATGTAAACTTCCGACTTCAACTGTA" --minimum-length 1 --discard-untrimmed -e 0 --no-indels. Reads passing this filter were then trimmed of any trailing Nextera adapter sequence, again using cutadapt and the following settings: -a "CTGTCTCTTATACACATCTCCGAGCCCACGAGACTNNNNNNNNNNNTCTCGTATGCCGTCTTCTGCTTG" --minimum-length 1. The remaining reads were aligned to the human genome (build hg38) with Novoalign 3 (Novocraft Technologies) and the following settings: -n 40 -o SAM -o SoftClip. Aligned reads were validated by confirming that they mapped adjacent to the insertion site motif. Successful reads were then converted to calling card format (.ccf; see http://wiki.wubrowse.org/Calling_card) and visualized on the WashU Epigenome Browser v46 (http://epigenomegateway.wustl.edu/legacy/).

In vitro *single cell calling card experiments*

N2a and K562 cells were cultured and transfected identically as HCT-116 cells, with the following exceptions: K562 cells were grown in RPMI 1640 Medium (Gibco #11875-085); for K562 cells, Neon electroporation settings were–pulse voltage: 1,450 V; pulse width: 10 ms; pulse number: 3; for N2a cells, Neon electroporation settings were–pulse voltage: 1,050 V; pulse width: 30 ms; pulse number: 2. For N2a cells, one replicate ($2x10^6$ cells) was transfected with 5 µg PB-SRT-Puro and 5 µg HyPBase, while another replicate was transfected with 5 µg PB-SRT-Puro only. For K562 cells, 4 replicates received both plasmids and one received the SRT alone. After 1 week of selection, N2a or K562 cells were mixed with transfected HCT-116 cells and then underwent single cell RNA-seq library preparation. For the species mixing experiment, cells were classified as either human or mouse if at least 80% of self-reporting transcripts in that cell mapped to the human or mouse genome, respectively, and as a multiplet. The estimated multiplet rate was calculated by doubling the observed percentage of human-mouse multiplet, to account for human-human and mouse-mouse doublets.

Single cell RNA-seq library preparation

Single cell RNA-seq libraries were prepared using 10x Genomics' Chromium Single Cell 3' Library and Gel Bead Kit (v2 chemistry; #120267). Each replicate was targeted for recovery of 6,000 cells. Library preparation followed a modified version of the manufacturer's protocol. We prepared the Single Cell Master Mix without RT Primer, replacing it with an equivalent volume of Low TE Buffer. GEM generation and GEM-RT incubation proceeded as instructed. At the end of Post GEM-RT cleanup, we added 36.5 µl Elution Solution I and transferred 36 µl of the eluted sample to a new tube (instead of 35.5 µl and 35 µl, respectively). The eluate was split into two 18 µl aliquots and kept at –20ºC until ready for further processing. One fraction was kept for single cell calling cards library preparation (see next section), while the other half was further processed into a single cell RNA-seq library.

We then added the RT Primer sequence to the products in the scRNA-seq aliquot. We created an RT master mix by adding 20 µl of Maxima 5X RT Buffer, 20 µl of 20% w/v Ficoll PM-400 (GE Healthcare

#17030010), 10 µl of 10 mM dNTPs (Clontech #639125), 2.5 µl RNase Inhibitor (Lucigen), and 2.5 µl of 100 µM 10x_TSO. To this solution we added 18 µl of the first RT product and 22 µl of ddH_2O. Finally, we added 5 µl Maxima H Minus Reverse Transcriptase, mixed by flicking, and centrifuged briefly. This reaction was incubated at 25ºC for 30 minutes followed by 50ºC for 90 minutes and heat inactivated at 85ºC for 5 minutes.

The solution was purified using DynaBeads MyOne Silane (Thermo Fisher #37002D) following 10x Genomics' instructions, beginning at "Post GEM-RT Cleanup – Silane DynaBeads" step D. The remainder of the single cell RNA-seq protocol, including purification, amplification, fragmentation, and final library amplification, followed manufacturer's instructions.

Single cell calling cards library preparation

To amplify self-reporting transcripts from single cell RNA-seq libraries, we took 9 µl of RT product (the other half was kept in reserve) and added it to 25 µl Kapa HiFi HotStart ReadyMix and 15 µl ddH_2O. We then prepared a PCR primer cocktail comprising 5 µl of 100 µM Bio_Illumina_Seq1_scCC_10X_3xPT primer, 5 µl of 100 µM Bio_Long_PB_LTR_3xPT, and 10 µl of 10 mM Tris-HCl, 0.1 mM EDTA buffer (IDT #11-05-01-13). One µl of this cocktail was added to the PCR mixture and placed in a thermocycler (Eppendorf MasterCycler Pro). Thermocycling settings were as follows: 98ºC for 3 minutes; 20-22 cycles of 98ºC for 20 seconds–67ºC for 30 seconds–72ºC for 5 minutes; 72ºC for 10 minutes; 4ºC forever. PCR purification was performed with 30 µl AMPure XP beads (0.6x ratio) as described previously. The resulting library was quantitated on an Agilent TapeStation 4200 System using the High Sensitivity D5000 ScreenTape (Agilent #5067-5592 and #5067-5593).

Single cell calling card library preparation was performed using the Nextera Mate Pair Sample Prep Kit (Illumina #FC-132-1001) with modifications to the manufacturer's protocol. The library was circularized by bringing 300 fmol (approximately 200 ng) of DNA up to a final volume of 268 µl with ddH_2O, then adding 30 µl Circularization Buffer 10x and 2 µl Circularization Ligase (final concentration: 1 nM). This reaction was incubated overnight (12-16 hours) at 30ºC. After removal of linear DNA (following manufacturer's instructions), we sheared the library on a Covaris E220 Focused-ultrasonicator with the following settings–peak power intensity: 200; duty factor: 20%; cycles per burst: 200; time: 40 seconds; temperature: 6ºC.

The library preparation proceeded per manufacturer's instructions until adapter ligation. We designed custom adapters (Supp. Table 4) so that the standard Illumina sequencing primers would not interfere with our library. Adapters were prepared by combining 4.5 µl of 100 µM scCC_P5_adapter, 4.5 µl of 100 µM scCC_P7_adapter, and 1 µl of NEBuffer 2 (New England BioLabs #B7002S), then heating in a thermocycler at 95ºC for 5 minutes, then holding at 70ºC for 15 minutes, then ramping down at 1% until it reached 25ºC, holding at that temperature for 5 minutes, before keeping at 4ºC forever. One microliter of this custom adapter mix was used in place of the manufacturer's recommended DNA Adapter Index. The ligation product was cleaned per manufacturer's instructions. For the final PCR, the master mix was created by combining 20 µl Enhanced PCR Mix with 28 µl of ddH_2O and 1 µl each of 25 µM scCC_P5_primer and 25 µM scCC_P7_primer. This was then added to the streptavidin bead-bound DNA and amplified under the following conditions: 98ºC for 30 seconds; 15 cycles of: 98ºC for 10 seconds–60ºC for 30 seconds–72ºC for 2 minutes; 72ºC for 5 minutes; 4ºC forever. All of the PCR supernatant was transferred to a new tube and purified with 35 µl (0.7x) AMPure XP beads following manufacturer's instructions. The final library was eluted in 25 µl Elution Buffer (QIAGEN #19086) and quantitated on an Agilent TapeStation 4200 System using the High Sensitivity D1000 ScreenTape.

Sequencing and analysis of scRNA-seq libraries

scRNA-seq libraries were sequenced on either Illumina HiSeq 2500 or NovaSeq S1 machines. Reads were analyzed using 10x Genomics' cellranger 2.1.0 with the following settings: --expect-cells=6000 --chemistry=SC3Pv2 --localcores=16 --localmem=30. The digital gene expression matrices from 10x were then further processed with scanpy 1.3.7 [97] for identification of highly variable genes, dimensionality reduction, and Louvain clustering. We cross-referenced gene expression data with published datasets [98] to assign cell types. The species-mixing analysis was performed using Drop-seq_tools 1.11 [59].

Sequencing and analysis of scCC libraries

scCC libraries were sequenced on Illumina NextSeq 500 machines (v2 Reagent Cartridges) with 50% PhiX. We used the standard Illumina primers for read 1 and index 2 (BP10 and BP14, respectively), and custom primers for read 2 and index 1 (Supp. Table 4). Read 1 sequenced the cell barcode and unique molecular index of each self-reporting transcript. Read 2 began with GGTTAA (end of the *piggyBac* terminal repeat and insertion site motif) before continuing into the genome. Reads containing this exact hexamer were trimmed using cutadapt with the following settings: -g "^GGTTAA" --minimum-length 1 --discard-untrimmed -e 0 --no-indels. Reads passing this filter were then trimmed of any trailing P7 adapter sequence, again using cutadapt and with the following settings: -a "AGAGACTGGCAAGTACACGTCGCACTCACCATGANNNNNNNNNATCTCGTATGCCGTCTTCTGCTTG" --minimum-length 1. Reads passing these filters were aligned using 10x Genomics' cellranger with the following settings: --expect-cells=6000 --nosecondary --chemistry=SC3Pv2 --localcores=16 --localmem=30. This workflow also managed barcode validation and collapsing of UMIs. Aligned reads were validated by verifying that they mapped adjacent to TTAA tetramers. Reads were then converted to calling card format (.ccf, see above). Finally, to minimize the presence of intermolecular artifacts, we required that each insertion must have been tagged by at least two different UMIs. We used the set of validated cell barcodes from each scRNA-seq library to demultiplex library-specific barcoded insertions from the scCC data. This approach requires no shared cell barcodes between scCC (and scRNA-seq) libraries. As a result, we excluded insertions from non-unique cell barcodes, which represented a very small number of total cells lost (< 1% per multiplexed library).

Peak calling

We called peaks in calling card data using Bayesian blocks [99], a noise-tolerant algorithm for segmenting discrete, one-dimensional data, using the astroML 0.3 implementation [100]. Bayesian blocks segments the genome into non-overlapping blocks where the density of calling card insertions is uniform. By comparing the segmentation against a background model, we were able to use Poisson statistics to assess whether a given block shows statistically significant enrichment for insertions. Let $B = \{b_1, b_2, \dots b_n\}$ represent the set of blocks found by performing Bayesian block segmentation on all insertions from a TF-directed experiment (e.g. SP1-PBase). For each block b_i, let x_i be the number of insertions in that block in the TF-directed experiment. Similarly, let y_i' be the number of insertions in that block in the undirected experiment (e.g. PBase) normalized to the total number of insertions found in the TF-directed experiment. Then, for each block we calculated the Poisson p-value of observing at least x_i insertions assuming a Poisson distribution with expectation y_i': $P(k \geq x_i | \lambda = y_i')$. We accepted all blocks that were significant beyond a particular p-value threshold.

For bulk analysis of SP1-PBase and SP1-HyPBase insertions, we added a pseudocount of 0.1 to all blocks and used p-value cutoffs of 10^{-6} and 10^{-22}, respectively. For single cell analysis of SP1-HyPBase insertions, we added a pseudocount of 1 to all blocks and used a p-value cutoff of 10^{-9}. All three of these values were beyond a Bonferroni-corrected α of 0.05. We polished peak calls by merging statistically-significant blocks that were within 250 bases of each other and by aligning block edges to coincide with TTAAs.

To identify BRD4 binding sites from undirected *piggyBac* insertions, we segmented those insertions using Bayesian blocks. For each block b_i, we let x_i denote the number of undirected insertions in that block. We also calculated x_i', the expected number of insertions in block b_i assuming *piggyBac* insertions were distributed uniformly across the genome. We did this by dividing the total number of mappable TTAAs in the genome by the total number of undirected insertions, then multiplying this value by the number of mappable TTAAs in block b_i. Then, for each block we calculated the Poisson *p*-value $P(k \geq x_i | \lambda = x_i')$. We accepted all blocks that were significant beyond a particular *p*-value threshold. Finally, we merged statistically-significant blocks that were within 12,500 bases of each other [51,57].

For the bulk PBase and HyPBase analysis, we used *p*-value cutoffs of 10^{-30} and 10^{-62}, respectively. For both *in vitro* and *in vivo* single cell HyPBase analyses, we used a *p*-value cutoff of 10^{-9}. To call differentially-bound loci between upper and lower cortical layer neurons, we used the same framework as described above for SP1 but did reciprocal enrichment analyses where the upper layer insertions were used as the "experiment" track and the lower layer insertions were used as the "control" track, and vice-versa. Here again we used a *p*-value cutoff of 10^{-9}.

Density tracks were generated by taking the Bayesian blocks segmentation of each calling card dataset and, for each block, calculating the normalized number of insertions (insertions per million mapped insertions, or IPM) and dividing by the length of the block in kilobases. This was plotted as a bedgraph file with smoothing applied in the WashU Epigenome Browser (25 pixel windows).

SP1 binding analysis in HCT-116 cells

We compared our SP1 peak calls to a publicly-available ChIP-seq dataset [54] as well as an input control file (Supp. Table 5). See below for more details on aligning and analyzing ChIP-seq data. We collated a list of unique TSSs by taking the 5'-most coordinates of RefSeq Curated genes in the hg38 build (UCSC Genome Browser). A list of CpG islands in HCT-116 cells and their methylation statuses were derived from previously-published Methyl-seq data [101]. We used the liftOver tool (UCSC) to convert coordinates from hg18 to hg38. We tested for enrichment in SP1-directed insertions at TSSs, CpG islands, and unmethylated CpGs with the *G* test of independence. For motif discovery we used MEME-ChIP 4.11.2 [102] with a dinucleotide shuffled control and the following settings: -dna -nmeme 600 -seed 0 -ccut 250 -meme-mod zoops -meme-minw 4 -meme-nmotifs 5.

BRD4 sensitivity, specificity, and precision analysis in HCT-116 cells

We used a published BRD4 ChIP-seq dataset [53] to identify BRD4-bound super-enhancers in HCT-116 cells, following previously-described methods [51,52]. We first called peaks using MACS 1.4.1 [103] at $p < 10^{-9}$, then fed this list into ROSE 0.1 (http://younglab.wi.mit.edu/super_enhancer_code.html). We discarded artifactual loci less than 2,000 bp in size, yielding a final list of 162 super-enhancers. To evaluate sensitivity, we used bedtools 2.27.1 [104] to ask what fraction of *piggyBac* peaks, at various *p*-value thresholds, overlapped the set of BRD4-bound super-enhancers. To measure specificity, we created a list of regions predicted to be insignificantly enriched ($p > 0.1$) for BRD4 ChIP-seq signal. We then sampled bases from this region such that the distribution of peak sizes was identical to that of the 162 super-enhancers. We sampled to 642x coverage, sufficient to cover each base with one peak, on average. We then asked what fraction of our *piggyBac* peaks overlapped these negative peaks and subtracted that value from 1 to obtain specificity. Finally, we calculated precision, or positive predictive value, by dividing the total number of detected super-enhancer peaks by the sum of the super-enhancer peaks and the false positive peaks.

Downsampling and replication analysis

When performing downsampling analyses on calling card insertions, we randomly sampled insertions without replacement and in proportion to the number of reads supporting each insertion. Peaks were called on the downsampled insertions at a range of *p*-value cutoffs. Linear interpolation was performed using numpy 1.15 and visualized using matplotlib 3.0. Replication was assessed by splitting calling card insertions into two, approximately equal, files based on their barcode sequences. Each new file was treated as a single biological experiment. For each peak called from the joint set of all insertions, we plotted the number of normalized insertions (IPM) in one replicate on the *x*-axis and the other replicate on *y*-axis.

ChIP-seq and chromatin state analyses

We aligned raw reads using Novoalign with the following settings for single-end datasets: -o SAM -o SoftClip, while paired-end datasets were mapped with the additional flag -i PE 200-500. To calculate and visualize the fold enrichment in ChIP-seq signal at calling card peaks, we used deeptools 3.0.1 [105]. We tested for significant mean enrichment in BRD4 ChIP-seq signal at *piggyBac* peaks over randomly shuffled control peaks with the Kolmogorov-Smirnov test. Chromatin state analysis was performed using ChromHMM 1.15 as previously described [106]. For each chromatin state, we plotted the mean and standard deviation of the rate of normalized insertions per kilobase (IPM/kb).

SRT-tdTomato fluorescence validation

To test the fluorescence properties of the SRT-tdTomato construct, we transfected K562 cells as previously described with either 1 µg of pUC19 plasmid (New England BioLabs #N3041S); 0.5 µg of PB-SRT-tdTomato plasmid and 0.5 µg pUC19; 0.5 µg of PB-SRT-tdTomato and 0.5 µg PBase plasmid; and 0.5 µg of PB-SRT-tdTomato and 0.5 µg HyPBase plasmid. Cells were allowed to expand for 8 days, after which fluorescence activity was assayed on an Attune NxT Flow Cytometer (Thermo Fisher) with an excitation wavelength of 561 nm. Flow cytometery data were visualized using FlowCal 1.2.0 [107]. We also performed bulk RNA calling cards on HEK293T cells transfected with SRT-tdTomato with or without HyPBase plasmid. While these cells were not sorted based on fluorescence activity, the SRT library from cells transfected with both SRT and transposase were more complex and contained many more insertions than the library from cells receiving SRT alone (Supp. Fig. 1A).

In vivo *single cell calling cards experiments*

All mouse experiments were done following procedures described in [66]. In brief, we cloned the PB-SRT-tdTomato and HyPBase constructs into AAV vectors. The Hope Center Viral Vectors Core at Washington University in St. Louis packaged each construct in AAV9 capsids. Titers for each virus ranged between $1.1x10^{13}$ and $2.2x10^{13}$ viral genomes/ml. We mixed equal volumes of each virus and performed intracranial cortical injections of the mixture into newborn wild-type C57BL/6J pups (P0-2). As a gating control, we injected one litter-matched animal with AAV9-PB-SRT-tdTomato only. After 2 to 4 weeks, we sacrificed mice and dissected the cortex (8 libraries) or hippocampus (1 library). All animal practices and procedures were approved by the Washington University in St. Louis Institutional Animal Care and Use Committee (IACUC) in accordance with National Institutes of Health (NIH) guidelines.

Tissues were dissociated to single suspensions following a modification of previously published methods [108,109]. We incubated samples in a papain solution containing Hibernate-A (Gibco #A1247501) with 5% v/v trehalose (Sigma-Aldrich #T9531), 1x B-27 Supplement (Gibco #17504044), 0.7 mM EDTA (Corning #36-034-Cl), 70 µM 2-mercaptoethanol (Gibco #21985023), and 2.8 mg/ml papain (Worthington Chemical Corporation #LS003118). After incubation at 37ºC, cells were treated with DNaseI (Worthington Chemical Corporation #NC9924263), triturated through increasingly narrow fire-

polished pipettes, and passed through a 40-micron filter prewetted with resuspension solution: Hibernate-A containing 5% v/v trehalose, 0.5% Ovomucoid Trypsin Inhibitor (Worthington Chemical Corporation #NC9931428), 0.5% Bovine Serum Albumin (BSA; Sigma-Aldrich #A9418), 33 μg/ml DNaseI, and 1x B-27 Supplement. The filter was washed with 6 ml of resuspension solution. The resulting suspension was centrifuged for 4 minutes at 250 g. The supernatant was discarded. The pellet was then resuspended in 2 ml of resuspension solution and resuspended by gentle pipetting.

We eliminated subcellular debris using gradient centrifugation. We first prepared a working solution of 30% w/v OptiPrep Density Gradient Medium (Sigma-Aldrich #D1556) mixed with an equal volume of 1x Hank's Balanced Salt Solution (HBSS; Gibco #14185052) with 0.5% BSA. We then prepared solutions of densities 1.057, 1.043, 1.036, and 1.029 g/ml using by combining the working solution with resuspension solution at ratios of 0.33:0.67, 0.23:0.77, 0.18:0.82, and 0.13:0.87, respectively. We layered 1 ml aliquots of each solution in a 15 ml conical tube beginning with the densest solution on the bottom. The cell suspension was added last to the tube and centrifuged for 20 minutes at 800 g at 12ºC. The top layer was then aspirated and purified cells were isolated from the remaining layers. These cells were then resuspended in FACS buffer: 1x HBSS, 2 mM $MgCl_2$ (Sigma-Aldrich #M4880), 2 mM $MgSO_4$ (Sigma-Aldrich #M2643), 1.25 mM $CaCl_2$ (Sigma-Aldrich #C7902), 1 mM D-glucose (Sigma-Aldrich #G7021), 0.02% BSA, and 5% v/v trehalose. Cells were centrifuged for 4 minutes at 250 g, the supernatant was discarded, and the pellet was resuspended in FACS buffer by gentle pipetting.

Cells were then sorted based on fluorescence activity. As a gating control, we analyzed cells from cortices injected with AAV9-PB-SRT-tdTomato only. We then collected cells from brains transfected with AAV9-PB-SRT-tdTomato and AAV9-HyPBase whose fluorescence values exceeded the gate. After sorting, cells were centrifuged for 3 minutes at 250 g. The supernatant was discarded and cells were resuspended in FACS buffer at a concentration appropriate for 10x Chromium 3' scRNA-seq library preparation.

In vivo *single cell calling cards analysis and validation*

Single cell RNA-seq and single cell calling card libraries were prepared, sequenced, and analyzed as described above. Cell types were assigned based on the expression of key marker genes and cross-referenced with recent cortical scRNA-seq datasets [62-65]. Brd4-bound peak calls were validated by comparing to a previously published cortical H3K27ac ChIP-seq dataset [68] (Supp. Table 5). Read alignment and statistical analysis were performed as described above.

The specificity of Brd4-bound gene expression in astrocytes and neurons was analyzed by first identifying all genes within 10,000 bases of astrocyte and neuronal Brd4 peaks. Although assigning an enhancer to its target gene is a difficult problem, using the nearest gene is common practice [110]. To control for sensitivity of gene detection, we downsampled the neuron insertions to the same number of astrocyte insertions, then called peaks and identified nearby genes in this subset. We used gene expression data from a bulk RNA-seq dataset [69] to compute the specificity of gene expression between astrocytes and neurons. We first discarded genes whose expression was not measured, and then set the value for genes with 0.1 FPKM to zero (to better distinguish non-expressed genes from lowly-expressed genes). Finally, for each gene g_i, we calculated the specificity as $\frac{Astrocyte_{FPKM}(g_i)}{Astrocyte_{FPKM}(g_i)+Neuron_{FPKM}(g_i)}$. Thus, a value of 0 denotes a gene purely expressed in neurons, a value of 0.5 for a gene equally expressed in both cell types, and a value of 1 for a gene purely expressed in astrocytes. We plotted distributions of gene expression specificity for the set of astrocyte-bound genes and the downsampled astrocyte-bound genes. Gene Ontology analysis was performed on the same sets of genes using PANTHER 14.0 [70] on the "GO biological process complete" database. Fisher's exact test was used to compute *p*-values, which were then subject to Bonferroni correction.

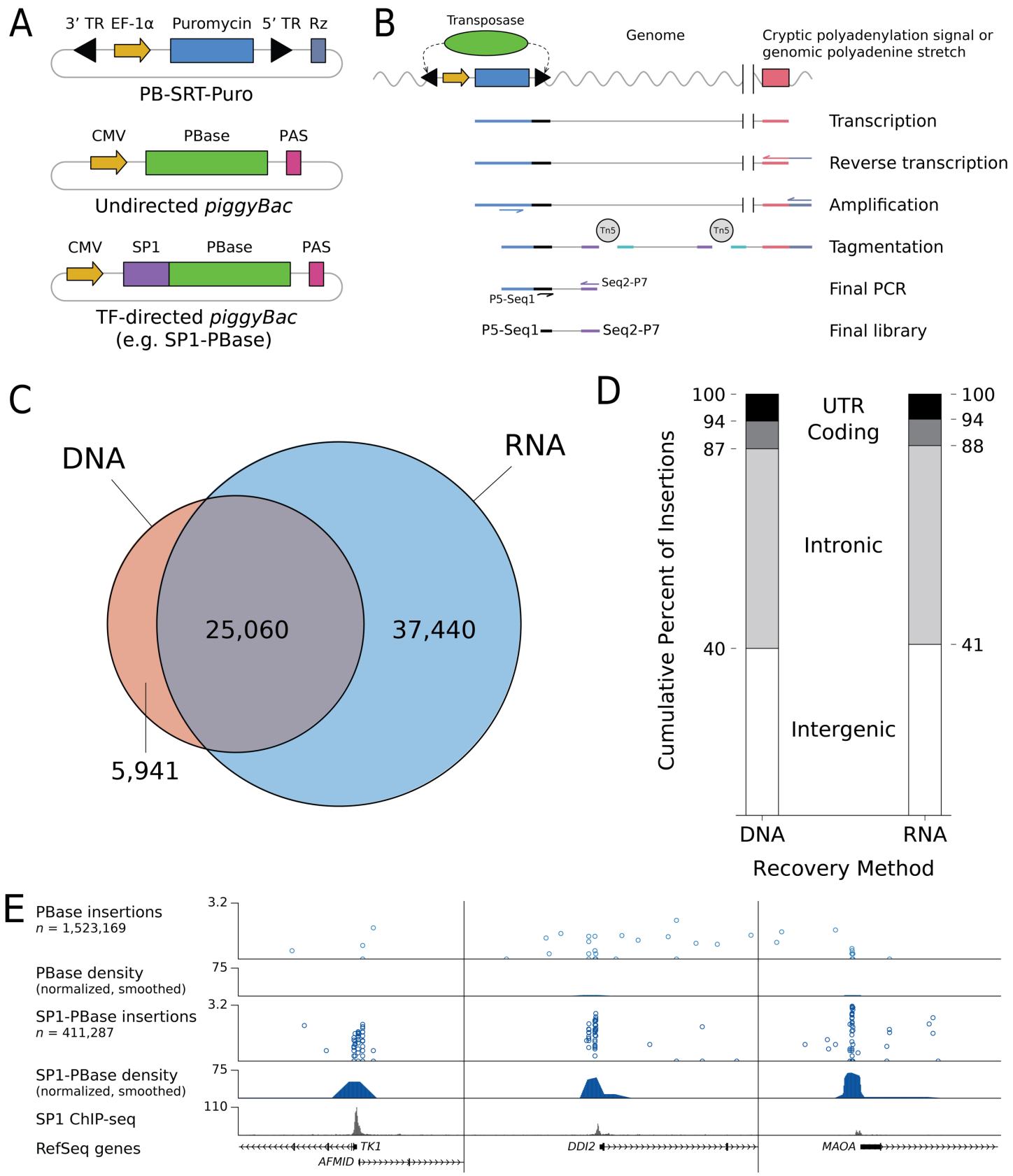

Figure 1: Self-reporting transposons (SRTs) are mapped more efficiently from RNA over DNA and, when directed SP1-PBase, identify SP1 binding sites. (A) Schematic of a self-reporting *piggyBac* transposon with puromycin marker (PB-SRT-Puro) and undirected (PBase) and SP1-directed (SP1-PBase) *piggyBac* transposases. SRTs are constructed by removing the PAS sequence between the end of the marker gene and the 5' TR. A self-cleaving ribozyme (Rz) on the delivery vector, downstream of the

SRT, prevents recovery of plasmid transposons. (B) SRTs are mapped by reverse transcribing RNA with a poly(T) primer followed by a series of nested PCRs and tagmentation. This final library is enriched for the junction between the transposon and the genome. (C) RNA-based recovery of SP1-directed SRTs in HCT-116 cells is more efficient than DNA-based recovery. The RNA protocol recovers 80% of the same insertions as the DNA protocol and recovers twice as many insertions overall. (D) The distribution of insertions with respect gene annotation is identical between transposons recovered by DNA and by RNA. (E) Insertions deposited by SP1-PBase show pronounced and specific clustering at SP1 ChIP-seq peaks over insertions left by undirected PBase. In the calling card track, each circle represents an independent insertion. Genomic position is on the *x*-axis and the number of reads supporting that insertion is on the *y*-axis on a log_{10}-transformed scale. The density tracks show the local rate of insertions in each experiment (IPM/kb), normalized for library size, and smoothed in the WashU Epigenome Browser. TR: terminal repeat; Puro: puromycin; PAS: polyadenylation signal; IPM: insertions per million mapped insertions; kb: kilobase.

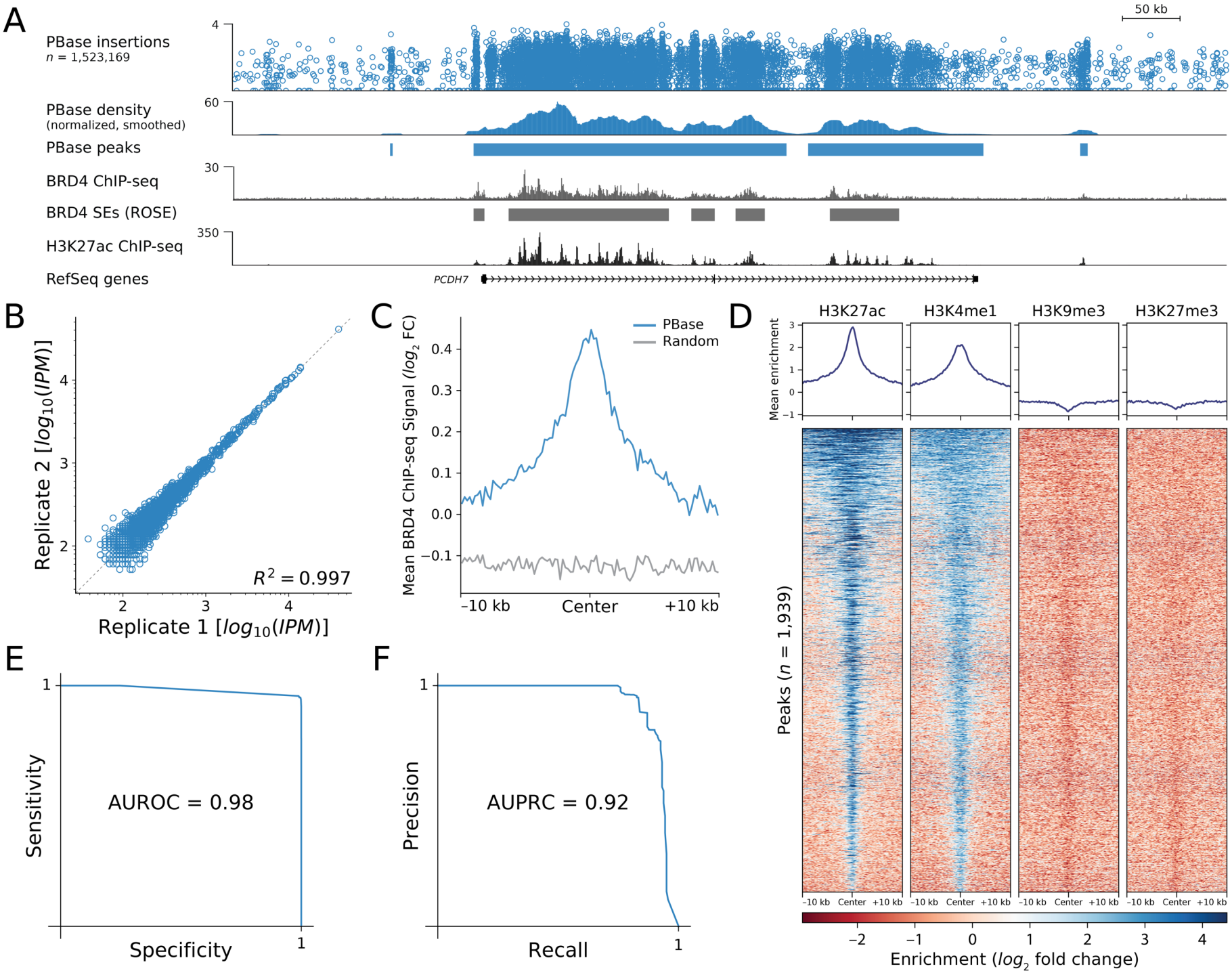
A
50 kb
PBase insertions
n = 1,523,169
PBase density
(normalized, smoothed)
PBase peaks
BRD4 ChIP-seq
BRD4 SEs (ROSE)
H3K27ac ChIP-seq
RefSeq genes
PCDH7
B
Replicate 2 [log10(IPM)]
Replicate 1 [log10(IPM)]
R2 = 0.997
C
PBase
Random
Mean BRD4 ChIP-seq Signal (log2 FC)
-10 kb
Center
+10 kb
D
H3K27ac
H3K4me1
H3K9me3
H3K27me3
Mean enrichment
Peaks (n = 1,939)
Enrichment (log2 fold change)
E
Sensitivity
Specificity
AUROC = 0.98
F
Precision
Recall
AUPRC = 0.92

Figure 2: Undirected *piggyBac* (PBase) insertions mark BRD4-bound super-enhancers. (A) Undirected PBase insertions are distributed non-randomly, with increased density overlapping BRD4-bound chromatin and H3K27 acetylated histones. Also shown are BRD4-bound SEs. (B) PBase peak calls are highly replicable, with biological replicates showing high concordance of normalized insertions at peaks. (C) PBase peaks show central enrichment for BRD4 ChIP-seq signal. These findings are statistically significant when compared to a genome-wide permutation of PBase peaks ($p < 10^{-9}$, KS test). (D) PBase peaks are centrally enriched for the histone modifications H3K27ac and H3K4me1, marks associated with enhancers. These same peaks show mild depletion for H3K9me and H3K27me, marks canonically associated with repressed chromatin. (E) Receiver-operator characteristic curve for SE detection using PBase peaks. (F) Precision-recall curve for SE detection using PBase peaks. SE: super-enhancer; IPM: insertions per million mapped insertions; AUROC: area under receiver-operator curve; AUPRC: area under precision-recall curve; KS: Kolmogorov-Smirnov; FC: fold change.

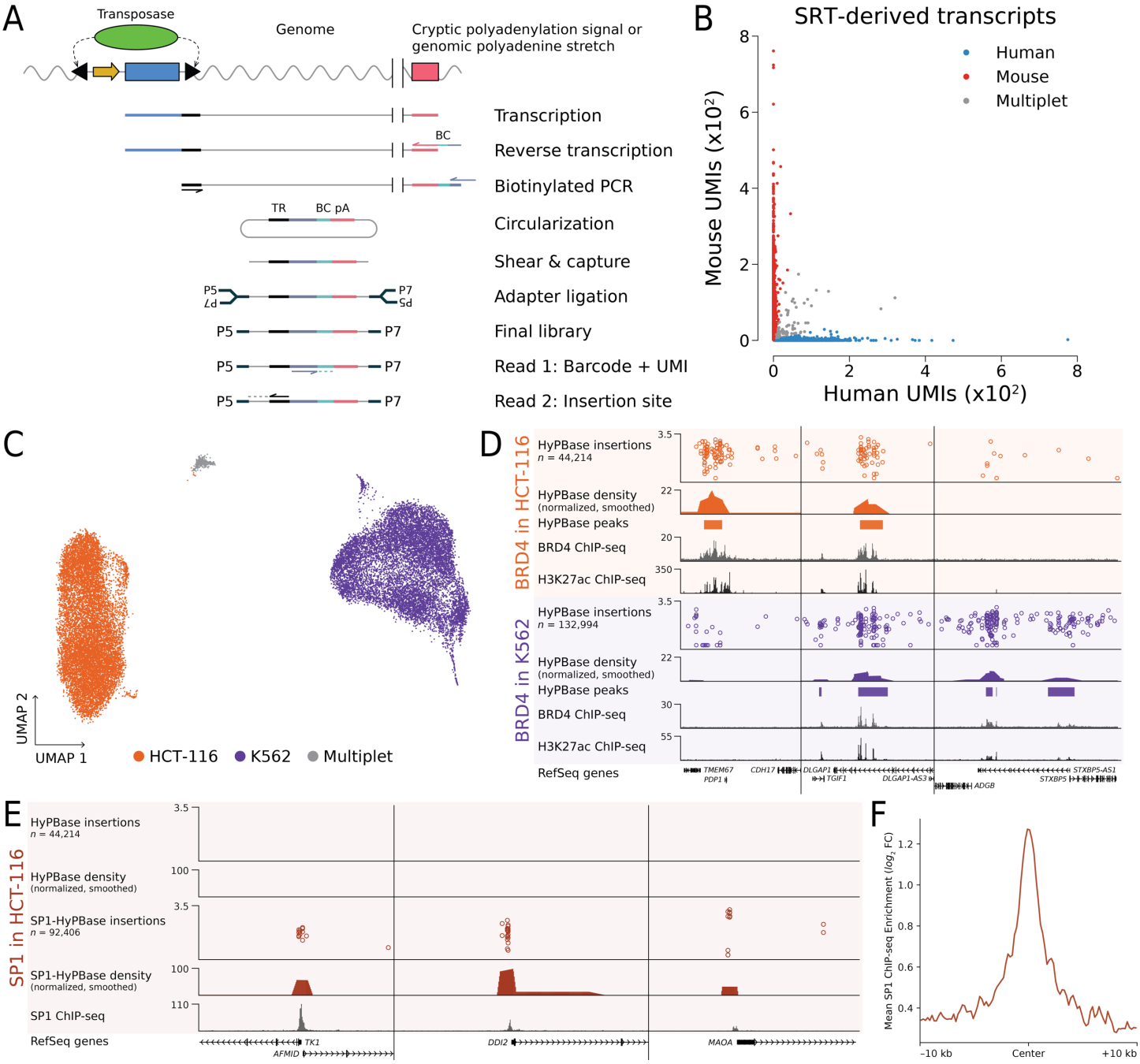

Figure 3: Single cell calling cards (scCC) maps BRD4 binding and SP1 in single cells. (A) Schematic of the scCC library preparation strategy from scRNA-seq libraries. Self-reporting transcripts are amplified using biotinylated primers and circularized, which brings the cell barcode and UMI in close proximity to the transposon-genome junction. Circularized molecules are sheared, captured with streptavidin, and Illumina adapters are ligated. Custom sequencing yields the cell barcode and UMI with read 1 and the genomic insertion site with read 2. (B) Barnyard plot of HCT-116 and N2a cells transfected with SRTs shows clean segregation of cell types. Most cells were assigned either human insertions or mouse insertions, with an estimated multiplet rate of 7.8%. (C) Human HCT-116 and K562 cells were transfected with PB-SRT-Puro and HyPBase and subsequently subjected to scRNA-seq. Two clear cell types emerge revealing each constituent cell population. (D) scCC deconvolves HyPBase insertions from HCT-116 and K562 cells, identifying shared and specific BRD4 binding sites. (E) scCC on HCT-116 cells transfected with SP1-HyPBase identifies SP1 binding sites. (F) SP1-HyPBase peaks from scCC data show strong central enrichment for SP1 ChIP-seq signal. TR: terminal repeat; BC: barcode; pA: poly(A) sequence; UMI: unique molecular index; Puro: puromycin; FC: fold change.

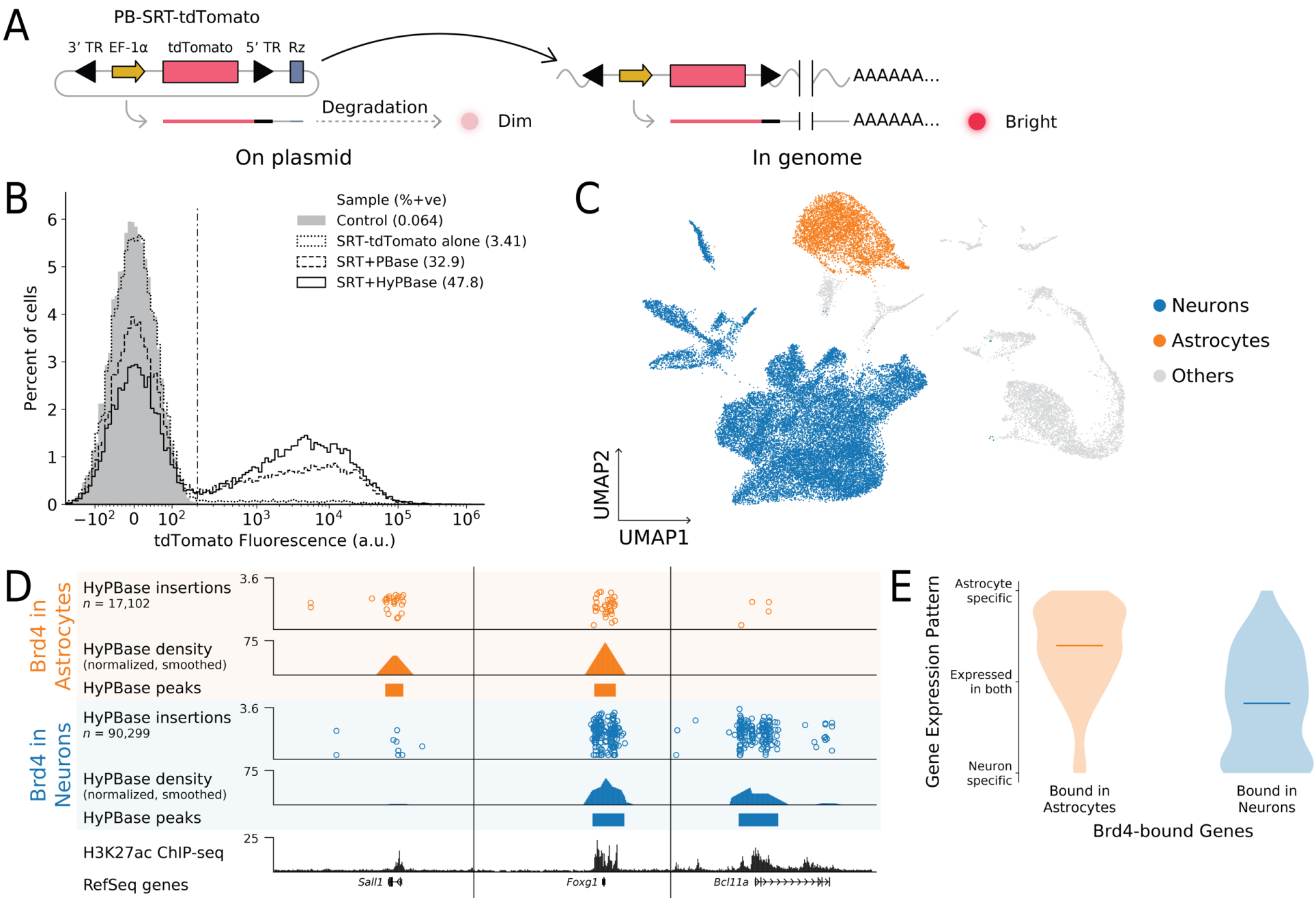

Figure 4: Single cell calling cards deconvolves Brd4-bound loci in the mouse cortex. (A) Schematic of PB-SRT-tdTomato, an SRT compatible with *in vivo* experiments. The pre-transposition tdTomato transcript (left) is degraded by the downstream ribozyme (Rz),

leading to low fluorescence intensity. After transposition into the genome, the self-reporting transcript is stabilized and results in a bright signal. (B) Validation of PB-SRT-tdTomato in K562 cells. Cells transfected with both SRT and transposase (either PBase or HyPBase) show bimodal fluorescence enabling sorting for cells with insertions. (C) scRNA-seq analysis of mouse cortex libraries transduced with PB-SRT-tdTomato and HyPBase reveals multiple cell types, including astrocytes (n = 4,727) and neurons (n = 25,158). (D) scCC analysis of HyPBase insertions in astrocytes and neurons identify shared and specific Brd4 binding sites. Whole cortex H3K27ac ChIP-seq shown for comparison. (E) Gene expression specificity of genes overlapping astrocyte or sensitivity-matched neuron peaks. Expression values were taken from bulk RNA-seq. Specificity for each gene was calculated by dividing the expression of the gene in astrocytes by the sum of the expression values in astrocytes and neurons. Horizontal lines indicate the medians of the distributions. TR: terminal repeat.

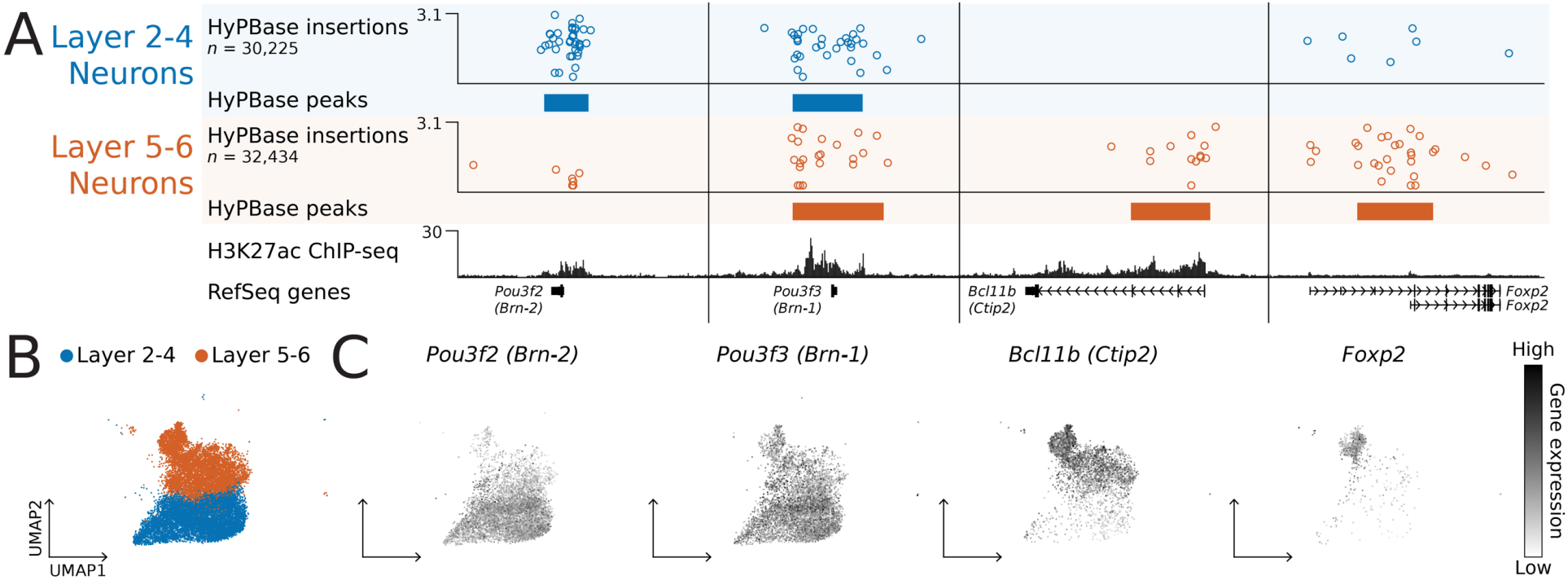

Figure 5: Single cell calling cards deconvolves Brd4 binding in cortical excitatory neurons and identifies known layer markers. (A) scCC analysis of HyPBase insertions in upper (layer 2-4) or lower (layer 5-6) cortical excitatory neurons identifies shared and specific Brd4 binding sites. Whole cortex H3K27ac ChIP-seq shown for comparison. (B) Layer 2-4 (n = 9,083) and layer 5-6 (n = 6,980) cortical excitatory neurons highlighted among the scRNA-seq clusters. (C) Gene expression patterns of the four genes from (A) mirrors the cell type-specificity of Brd4 binding.