Self-Reporting Transposons Enable Simultaneous Readout of Gene Expression and Transcription Factor Binding in Single Cells

Graphical Abstract

Highlights
- Novel genetic element can be mapped from mRNA and directed by DNA-binding proteins
- Combined gene expression and transcription factor binding data from single cells
- Mapped multiple transcription factors in several cell lines and the mouse cortex
- Discovered bromodomain-dependent cell-state transitions in leukemic cells

Authors
Arnav Moudgil, Michael N. Wilkinson, Xuhua Chen, ..., Samantha A. Morris, Joseph D. Dougherty, Robi D. Mitra

Correspondence
rmitra@wustl.edu

In Brief
Moudgil et al. present a single-cell method for simultaneously capturing gene expression and transcription factor binding site data from the same cells, first in cell lines and then in the mouse brain.

Moudgil et al., 2020, Cell 182, 1–17
August 20, 2020 © 2020 Elsevier Inc.
https://doi.org/10.1016/j.cell.2020.06.037
Resource

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 Alexander J. Cammack,4 Michael J. Vasek,1,5 Tomás Lagunas, Jr.,1,5 Zongtai Qi,1,2 Matthew A. Lalli,1 Chuner Guo,1,3,6 Samantha A. Morris,1,6,7 Joseph D. Dougherty,1,5 and Robi D. Mitra1,2,8,*

1Department of Genetics, Washington University School of Medicine in St. Louis, St. Louis, MO 63110, USA
2Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine in St. Louis, St. Louis, MO 63110, USA
3Medical Scientist Training Program, Washington University School of Medicine in St. Louis, St. Louis, MO 63110, USA
4Department of Neurology, Washington University School of Medicine in St. Louis, St. Louis, MO 63110, USA
5Department of Psychiatry, Washington University School of Medicine in St. Louis, St. Louis, MO 63110, USA
6Department of Developmental Biology, Washington University School of Medicine in St. Louis, St. Louis, MO 63110, USA
7Center of Regenerative Medicine, Washington University School of Medicine in St. Louis, St. Louis, MO 63110, USA
8Lead Contact
*Correspondence: rmitra@wustl.edu
https://doi.org/10.1016/j.cell.2020.06.037

SUMMARY

Cellular heterogeneity confounds in situ assays of transcription factor (TF) binding. Single-cell RNA sequencing (scRNA-seq) deconvolves cell types from gene expression, but no technology links cell identity to TF binding sites (TFBS) in those cell types. We present self-reporting transposons (SRTs) and use them in single-cell calling cards (scCC), a novel assay for simultaneously measuring gene expression and mapping TFBS in single cells. The genomic locations of SRTs are recovered from mRNA, and SRTs deposited by exogenous, TF-transposase fusions can be used to map TFBS. We then present scCC, which map SRTs from scRNA-seq libraries, simultaneously identifying cell types and TFBS in those same cells. We benchmark multiple TFs with this technique. Next, we use scCC to discover BRD4-mediated cell-state transitions in K562 cells. Finally, we map BRD4 binding sites in the mouse cortex at single-cell resolution, establishing a new method for studying TF biology in situ.

INTRODUCTION

Transcription factors (TFs) regulate the gene expression patterns that specify cell state (Gurdon, 2016; Hafler et al., 2012; Mizuguchi et al., 2001; Zhu et al., 2012). They are central to a number of critical developmental processes including the maintenance of pluripotency (Liu et al., 2008; Takahashi and Yamanaka, 2006), fate choice (Mizuguchi et al., 2001; Zhu et al., 2012), and embryogenesis (Fogarty et al., 2017). Perturbing TF activity can disrupt cellular development, homeostasis, or function, resulting in altered morphology (Gonen et al., 2018; Kvon et al., 2012), cellular transdifferentiation (Davis et al., 1987), or increased susceptibility to disease (Lee and Young, 2013). A better understanding of TF binding during development and homeostasis would provide insights into how cellular diversity arises and is maintained under normal and abnormal biological conditions.

Single-cell RNA sequencing (scRNA-seq) has emerged as the de facto approach for characterizing cellular diversity in complex tissues and organisms (Campbell et al., 2017; Cao et al., 2017; Fincher et al., 2018; Han et al., 2018; Karaiskos et al., 2017; Zeisel et al., 2015). Recently, multi-modal scRNA-seq technologies have emerged (Angermueller et al., 2016; Cao et al., 2018; Clark et al., 2018; Dey et al., 2015; Macaulay et al., 2015; Peterson et al., 2017; Stoeckius et al., 2017) linking transcriptional information to other genomic assays. These methods address the fact that, while scRNA-seq can describe the current state of a biological system, it alone cannot explain how that state arose. A notable lacuna in the single-cell repertoire is a method for jointly assaying transcriptome and TF binding. Such a method would lead to the genome-wide identification of TF binding sites across multiple cell types in complex tissues. Chromatin immunoprecipitation sequencing (ChiP-seq) is the most popular approach to studying TF binding (Johnson et al., 2007), and, while a number of antibody-based single-cell methods to detect DNA-protein contacts have been reported (Ai et al., 2019; Carter et al., 2019; Grosselin et al., 2019; Hainer et al., 2019; Harada et al.,
assay TF binding (Wang et al., 2007, 2011, 2012a). This system or non-specifically, such as chromatin remodelers. Used to study transcriptional regulators that bind DNA indirectly says do not directly measure TF occupancy nor can they be inference to identify potential DNA binding proteins. These as-

free regions that may be bound by TFs, though they rely on motif

in the mediator element MED1 and the bromodomain protein

H3K27ac as well as Pol II and transcriptional coactivators like

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 (Ding et al., 2005) fused to the TF SP1 and shown that the resulting pattern of insertions reflects SP1’s binding preferences (Wang et al., 2012a). However, this method was only feasible in bulk preparations of thousands of cells.

Here, we present single-cell calling cards (scCC), an extension of transposon calling cards that simultaneously profiles mRNA content and TF binding at single-cell resolution. The key component of our work is the self-reporting transposon (SRT), a novel element whose genomic location can be mapped from mRNA. We show that the RNA-based calling card method is more efficient than our standard DNA-based protocol and can be used to map TF binding sites with a directed transposase. We also 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-type-specific mapping of SRTs from scRNA-seq libraries. Thus, in one experiment, we can cluster cells by transcriptional identity and identify TF binding sites within those cell types. We highlight the range of this technology using a breadth of TFs in a variety of cell lines. We then use scCC to discover bromodomain-dependent cell-state dynamics in K562 cells. Finally, we identify cell-type-specific BRD4 binding sites in vivo in the postnatal mouse cortex. These results demonstrate that scCC could be a broadly applicable tool to study specific TF binding interactions across multiple cell types within heterogeneous systems.

RESULTS

SRTs Can Be Mapped from mRNA Instead of Genomic DNA

To combine scRNA-seq with calling cards, we first developed a transposon whose genomic location could be determined from mRNA. We created a piggyBac SRT by removing the polyadenylation signal (PAS) downstream of the reporter gene (Figure 1A; Methods S1) in the transposon. RNA polymerase II (Pol II) transcribes the SRT reporter and continues through the terminal repeat (TR) into the flanking genomic sequence. Thus, SRTs “self-report” their locations through the unique genomic sequence found in the 3’ untranslated regions (UTRs) of the reporter gene transcripts. While previously published gene- or enhancer-trap transposons (Cadiñanos and Bradley, 2007) could also encode local positional information in RNA, they are resolution-limited to the nearest gene or enhancer, respectively. In contrast, SRT-derived transcripts contain the transposon-genome junction, so insertions can be mapped 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 PASs or templated polyadenine tracts in genomic DNA downstream of the SRT insertion site (Figure 1B). We then use a modified tagmentation protocol to enrich for the transposon-genome junction (STAR Methods). We confirmed SRTs generate reproducible libraries, require a functional transposase, and can be recovered from virtually any chromatin state (Methods S1).

To compare how the new RNA-based approach fares against our standard DNA-based method (Wang et al., 2012a), we tested both protocols on the same population of cells. Our DNA-based library yielded 31,001 insertions, while the RNA-based protocol recovered 62,500 insertions (Table S1). Importantly, 80% of the insertions found by DNA calling cards were also recovered in the RNA-based library (25,060 insertions; Figure 1C). Thus, at the level of individual transpositions, RNA-based mapping is highly sensitive. Moreover, the RNA protocol recovered a further 37,440 insertions that were not found in the DNA-based library. We analyzed the distribution of insertions by both genetic annotation (Figure 1D) and chromatin state (Methods S1) and found no appreciable differences in either case between the DNA and RNA libraries. Finally, we also confirmed that SRTs could still be used to study TF binding using established TF-piggyBac constructs and quantified the rediectability of these fusions (Methods S1). Thus, RNA-based recovery of transposons appears to be unbiased with respect to our established, DNA-based protocol.

Clustering of Undirected piggyBac Insertions Identifies BRD4-Bound SEs

Previous studies have shown that undirected piggyBac preferentially inserts transposons near SEs (Yoshida et al., 2017), unique regulatory elements involved in regulating cell identity (Hnisz et al., 2013). SEs are enriched for the histone modification H3K27ac as well as Pol II and transcriptional coactivators like the mediator element MED1 and the bromodomain protein...
BRD4 (Hnisz et al., 2013; Lovén et al., 2013; Whyte et al., 2013). piggyBac has a strong biophysical affinity for BRD4, as these proteins can be co-immunoprecipitated (Gogol-Döring et al., 2016). Thus, we hypothesized that, given the millions of insertions recoverable by SRTs (Table S1), we would be able to identify BRD4-bound SEs simply from the localization of undirected piggyBac transpositions.

In HCT-116 cells, undirected piggyBac showed non-uniform densities of insertions at BRD4-bound loci (Figure 2A; for guidance on interpreting a calling card track, see STAR Methods). At statistically significant peaks of piggyBac calling cards, piggyBac showed high reproducibility of normalized insertions between biological replicates ($R^2 > 0.99$; Figure 2B). We calculated the mean BRD4 enrichment, as assayed by ChIP-seq (McCleland et al., 2016), over all piggyBac peaks, which showed significantly increased BRD4 signal compared to a permuted control set (Figure 2C; Kolmogorov-Smirnov [KS] test $p < 10^{-5}$). Maximum BRD4 ChIP-seq signal was observed at calling card peak centers and decreased symmetrically in both directions. Moreover, piggyBac peaks showed striking overlap with ChIP-seq profiles for several histone modifications (Sloan et al., 2016; ENCODE Project Consortium, 2012), in particular, an enrichment for H3K27 acetylation (Figure 2D).

Since bromodomains bind acetylated histones, this observation further supports the notion that undirected piggyBac insertions can be used to map BRD4 binding. Peaks were also enriched in H3K4me1, another canonical enhancer mark, and depleted for H3K9me3 and H3K27me3, modifications associated with heterochromatin (Lawrence et al., 2016). In all, piggyBac insertion density is highly correlated with BRD4

---

**Figure 1. Self-Reporting Transposons Are Mapped More Efficiently from RNA Compared to DNA**

(A) Schematics of a self-reporting piggyBac transposon with puromycin reporter gene (PB-SRT-Puro) and undirected (PBase) and SP1-directed (SP1-PBase) piggyBac transposases.

(B) Molecular workflow for mapping SRTs from bulk RNA libraries.

(C) Overlap of SRTs recovered by DNA- or RNA-based protocols in HCT-116 cells.

(D) Distribution of insertions with respect to genetic annotation between SRT libraries prepared from either DNA or RNA. TR, terminal repeat; Puro, puromycin; PAS, polyadenylation signal.
binding throughout the genome and that regions enriched for undirected \textit{piggyBac} insertions share features common to enhancers.

We next assessed whether undirected \textit{piggyBac} transpositions can accurately assay BRD4-bound SEs. We constructed receiver-operator characteristic curves based on our ability to detect SEs from \textit{piggyBac} transpositions. Across a range of sensitivities, calling card peaks are highly specific and have high positive predictive value (AUPRC = 0.92; Figure 2F). These trends also hold true for the hyperactive \textit{piggyBac} mutant (Methods S1).

Thus, undirected \textit{piggyBac} transpositions can accurately assay BRD4-bound SEs.

We also investigated how similar \textit{piggyBac} transposition is to that of Tn5, the transposase used in ATAC-seq (Buenrostro et al., 2013, 2015) to identify open chromatin. Since BRD4 and H3K27ac co-occur at accessible loci, it may be that undirected calling cards and ATAC-seq provide redundant information. If that were the case, we should be able to identify BRD4-bound SEs with high sensitivity from ATAC-seq data alone, much as we have shown for \textit{piggyBac}. We called SEs using publicly available ATAC-seq data from HCT-116 cells (Ponnaluri et al., 2017) in the same manner that we did for BRD4 ChIP-seq. We found...
almost no overlap between BRD4-bound SEs and these so-called SEs from ATAC-seq data (Figure S1A). Moreover, there are a small number (4.3%) of piggyBac peaks that are not found in accessible chromatin (Figure S1B), suggesting that there may be regulatory elements in closed chromatin that calling cards are better able to detect. Globally, over 20% of Trn5 insertions are directed to accessible sites, starkly higher than undirected piggyBac but comparable to TF-piggyBac fusions (Figure S1C). That piggyBac’s preference for targetting open chromatin can be markedly increased by a covalently linked TF highlights both piggyBac’s baseline insensitivity for accessible sites and the efficacy of TF redirection. Finally, we find that piggyBac peaks are an order of magnitude larger than ATAC-seq peaks and, as a result, capture more BRD4 binding (Figure S1D). We conclude that unfused piggyBac reflects BRD4’s binding preferences whereas Trn5 reports on all accessible chromatin; as a result, undirected calling cards are not equivalent to ATAC-seq.

scCC Enables Simultaneous Identification of Cell Type and Cell-Type-Specific BRD4 Binding Sites

We next sought to recover SRTs from scRNA-seq libraries, which would let us identify cell types from transcriptomic clustering and, using the same source material, simultaneously profile TF binding in those cell types. We adopted the 10x Chromium platform due to its high efficiency of cell and transcript capture as well as its ease of use (Zheng et al., 2017) but with a modified protocol (Methods S1). We split the first-strand synthesis product in two: one half is used to generate a scRNA-seq library, while the other half undergoes specific amplification for SRTs followed by circularization. The circularization step brings the cell barcode and unique molecular index (UMI), found at the 3’ ends of each transcript, next to the transposon-genome junction. In this way, SRTs can be mapped and assigned to single cells using high-throughput short read sequencing (Figure 3A). After sequencing, the cell barcodes shared between both libraries are used to connect individual insertions to specific cell types. We call this protocol scCC.

We first validated scCC by performing a species-mixing experiment with human HCT-116 cells and mouse N2a cells transfected with hyperactive piggyBac (HyPBase) and PB-SRT-Puro. The resulting scRNA-seq library showed strong species separation with an estimated multiplet rate of 3.2% (Figure S2A). We restricted our calling card analysis to those insertions whose cell barcodes were observed in the scRNA-seq library (Table S2). The distribution of insertions across these cells reflected a continuum from pure mouse to pure human (Figures S2B and S2C). Since intramolecular ligation in the circularization step or the subsequent PCR may introduce artifacts, such as the mis-assignment of a barcode from a mouse cell to an insertion site in a human cell, we required that a given insertion in a given cell must have at least two different UMIs associated with it. This filter greatly improved the number of pure mouse and human cells (Figure S2D), yielding clear species separation with an estimated multiplet rate of 7.9% (Figure 3B). Thus, scCC can accurately map SRT insertions in single cells.

We then asked whether scCC could discern cell-type-specific BRD4 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 distinguished these two cell types (Figures 3C and S3A). We 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 (Table S2). The distribution of insertions per cell varied by cell type (Figure S3D) and was not explained by differences in total RNA content (Figures S3B and S3C). Over 93% and 96% of HCT-116 and K562 cells, respectively, had at least one insertion event (Table S2). Using the scCC insertion data, 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 (Figure 3D). Both HCT-116 and K562 peaks showed statistically significant enrichment for BRD4 ChIP-seq signal over randomly permuted peaks (Figures S3E and S3F; KS test p < 10−9 in both instances). Furthermore, 57% of HCT-116 peaks and 81% of K562 peaks were specifically bound in their respective cell type. We estimated that with a p value cutoff of 10−9, our sensitivity for detecting BRD4-bound SRTs would be approximately 60% (Methods S1), while the actual sensitivity at this level was 67%. Finally, at statistically significant peaks, normalized insertion counts were highly concordant between biological replicates in both cell types (R2 = 0.91 and 0.94, respectively; Figures S3G and S3H).

In all, these experiments demonstrate that scCC can be used to identify and deconvolve cell-type-specific BRD4 binding sites.

scCC Identifies Binding Sites across a Spectrum of TFs and in a Variety of Cell Types

Our success mapping BRD4 SEs in single cells gave us confidence that we would also be able to map TF binding with scCC. We transfected HCT-116 and K562 cells with an SP1 fusion construct (SP1-HyPBase) and performed scCC (Table S2). As was observed in bulk (Methods S1), SP1-HyPBase-directed insertions recovered from single cells localized to SP1 binding sites in both HCT-116 and K562 cells (Figures 4A and 4E). In both cell lines, we observed significant enrichment of SP1 ChIP-seq signal at scCC peaks (Figures 4B and 4C and Figures 4F and 4G) and motif analysis identified the SP1 DNA binding motif (Figure 4D and Figure 4H) (p < 10−30 in each instance). SP1 is known to preferentially bind near transcription start sites (TSSs) and is also thought to play a role in demethylating CpG islands (Brandeis et al., 1994; Macleod et al., 1994; Philipsen and Suske, 1999). Accordingly, we observed significant enrichments for insertions near TSSs, CpG islands, and unmethylated CpG islands in particular (Figures S4A and S4B; G test of independence p < 10−9 in each instance).

We next performed scCC in HepG2 cells with the pioneer factor FOXA2 (Table S2), which has been shown to be required for normal liver development and drives core transcriptional networks in cancer cells (Fourrier et al., 2016; Lee et al., 2005). As with SP1, we observed a specific enrichment of insertions at FOXA2 binding sites (Figure 4I). Peaks called from scCC FOXA2 data were enriched in FOXA2 ChIP-seq signal (Figures 4J and 4K) and motif analysis was able to infer the core FOXA2 DNA binding motif (Figure 4L).

Last, we mapped the binding of BAP1 in the uveal melanoma cell line OCM-1A (Yen et al., 2018) using scCC (Table S2). Unlike SP1 and FOXA2, BAP1 does not bind DNA directly; instead, it is
drawn to chromatin in a complex (Carbone et al., 2013; Yu et al., 2010) where it acts as a histone deubiquitinase. Despite this indirect interaction, we were able to resolve sharp BAP1-directed peaks (Figure 4M). These peaks showed high concordance with bulk RNA calling card data that we also generated in this system (Figures 4N and 4O; Table S1). Sequence analysis elicited the motif of YY1 (Figure 4P), a DNA binding TF and known member of the BAP1 complex (Yu et al., 2010). BAP1 is known to preferentially bind promoters (Dey et al., 2012), and, as such, we observed a significant enrichment for BAP1-directed insertions near TSSs (Figure S4C; Fisher’s exact test p < 10^{-3}). This suggests that, in this model of uveal melanoma, promoter-bound BAP1 primarily acts as a repressor of gene expression.

Collectively, these results indicate that scCC can successfully map DNA-protein interactions for a range of TFs and in a variety of cell types. Furthermore, scCC showed high reproducibility in all four tested conditions (R^2 between 0.71 and 0.95; Figures S4E–S4H). Although TF-piggyBac fusions have been previously reported to decrease transposase activity (Wu et al., 2006), our findings were more equivocal: some fusions showed less activity per cell than undirected HyPBase, while others were more efficient (Figures S4I–S4L). Thus, there may be some variability in the number of recovered insertions depending on the TF and cell type of interest. Overall, however, the method is robust.

**scCC Reveal Bromodomain-Dependent Cell-State Dynamics in K562 Cells**

SEs and BRD4 are thought to mark genes important for specifying cell identity, and, while the strongest evidence for this comes from comparisons between organ systems or between...
Figure 4. scCC Works with a Variety of Transcription Factors and Cell Lines

(A–D) scCC with SP1-HyPBase in HCT-116 cells reveal SP1 binding sites. (A) Browser view of a peak from SP1 scCC. (B) Mean SP1 ChIP-seq signal at scCC SP1 peaks. (C) Heatmap of SP1 ChIP-seq signal across all scCC SP1 peaks. (D) Core SP1 motif elicited from SP1 scCC peaks.

(E–H) Same as (A)–(D) but in K562 cells.

(I–M) scCC with FOXA2-HyPBase in HepG2 cells reveal FOXA2 binding sites. (I) Browser view of a peak from FOXA2 scCC. (J) Mean FOXA2 ChIP-seq signal at scCC FOXA2 peaks. (K) Heatmap of FOXA2 ChIP-seq signal across all scCC FOXA2 peaks. (L) Core FOXA2 motif elicited from FOXA2 scCC peaks.

(N–P) scCC with BAP1-HyPBase in OCM-1A cells reveal BAP1 binding sites. (N) Browser view of a peak from BAP1 scCC. (O) Mean bulk BAP1 ChIP-seq signal at scCC BAP1 peaks. (P) YY1 motif elicited from BAP1 scCC peaks.

See also Figure S4. FC, fold change.

We first scored cells based on a principal-component analysis (PCA) of gene expression (Figures S5A and S5B), which revealed a gradient of identities along a stem-like-to-differentiated cell-state axis (Figure 5A). We then separated cells into low and high clusters (Figures S5C and S5D) and asked whether we could detect any differences in BRD4 binding between them in our scCC data. Indeed, we found multiple peaks that showed significant differential binding (Figure 5B). We corroborated these hits by comparing our peak calls to bulk BRD4 and H3K27ac ChIP-seq data, as well as to RNA pol II ChIA-PET data, which connects putative enhancers to actively transcribed genes (Fullwood et al., 2009). We highlight two genes that showed both differential binding and expression: VMP1, bound...
Figure 5. scCC Uncovers Bromodomain-Dependent Cell-State Dynamics in K562 Cells

(A) Gradient of cell states from scRNA-seq analysis of K562 cells.
(B) Differential BRD4 binding analysis of undirected HyPBase peaks in K562 cells.

Please cite this article in press as: Moudgil et al., Self-Reporting Transposons Enable Simultaneous Readout of Gene Expression and Transcription Factor Binding in Single Cells, Cell (2020), https://doi.org/10.1016/j.cell.2020.06.037
more in the CD24high stem-like cells, and PVT1, bound more in the differentiated, CD24low cells (Figures S5F and S5G). VMP1 overexpression is sufficient to induce autophagy (Ropolo et al., 2007), which is important for hematopoietic stem cell function (Folkerts et al., 2019; Ho et al., 2017) and may be one pathway recruited during these dynamic state transitions. PVT1 can act as both a tumor-suppressor and oncogene, in both instances acting on the MYC locus (Cho et al., 2018).

We next investigated whether the observed differences in BRD4 binding might be causally responsible for establishing these two cell states. Downregulating BRD4 has been shown to influence cell identity across a range of cell types (Di Micco et al., 2014; Kfouri et al., 2017; Najafova et al., 2017). Thus, we hypothesized that BRD4 inhibition would change the distribution of cells in the stem-like and differentiated states. Moreover, due to the asymmetric nature of significant hits (Figure 5B), there is a subset of peaks specific to the CD24high state that are not shared by the CD24low state, suggesting that there may be a gene regulatory network that is recruited as cells transit from the differentiated to stem-like state and lost as they return. Hence, not only should the distribution of CD24high/CD24low cells change upon BRD4 perturbation but the stem-like CD24high population should be more susceptible to such an intervention.

We tested this hypothesis by treating cells with JO1, a small-molecule bromodomain inhibitor often used to disrupt BRD4 binding and alter target gene expression (Delmore et al., 2011; Garcia-Carpizo et al., 2018; Lovén et al., 2013; Sdelci et al., 2019). JO1 treatment shifted the population from one containing equal proportions of CD24high/CD24low cells to one composed of almost exclusively CD24low cells (>95%,Figure 5C). This conversion took place rapidly over the first 2 days, plateaued by day 4, and remained stable 1 week after treatment. In contrast, the control cells remained evenly split between the two states at this time point (Figure 5D; two-way ANOVA p < 0.01). JO1 was not selectively cytotoxic to CD24high cells as there were no significant differences in the levels of annexin V, an early marker of apoptosis, between CD24high and CD24low cells, regardless of whether they had been exposed to JO1 or DMSO (Figure S6A; three-way ANOVA p = 0.84). Additionally, we examined whether CD24 is a direct target of BRD4, which would imply that the loss of CD24 staining was an unremarkable consequence of JO1 treatment. We did not find evidence of BRD4 binding sites, either by ChIP-seq or calling cards, or of elevated H3K27 acetylation in the vicinity of CD24 (Figure S6B). We also compared the relative changes in mRNA levels of MYC, a known BRD4 target (Knochel et al., 2014; Lovén et al., 2013; Rathert et al., 2015; Zuber et al., 2011), to that of CD24 during the first 24 h of JO1 exposure. While MYC levels fell within the first 3 h of exposure, transcript levels of CD24 decreased most precipitously between 3 and 9 h after JO1 induction (Figure S6C). This delayed response suggests a regulatory cascade rather than direct transcriptional control. Thus, JO1 treatment does not trivially downregulate a cell-surface marker but rather likely perturbs gene regulatory networks that include CD24.

While JO1 shows greatest affinity for BRD4, it does have some promiscuity toward other bromodomains, including the orthologs BRD2 and BRD3 (Filippakopoulos et al., 2010). To address whether off-target effects were responsible for the shift in cell states, we downregulated BRD4 expression with CRISPR interference (CRISPRi). We confirmed that our BRD4 guide RNA (gRNA) specifically reduced expression of BRD4 and not BRD2 nor BRD3 (Figure S6D; Welch’s t test p < 0.05). As with JO1, we observed a significant decrease in the proportion of CD24high cells with the BRD4 gRNA compared to the non-targeting (NT) gRNA (Figure 5E; Welch’s t-test p < 0.01), though not to the same levels as JO1. This suggests that, while BRD4 is necessary for the observed cell-state dynamics between CD24high and CD24low cells, it is likely that other bromodomains also play a role.

CD24high/CD24low cells have been previously shown to have different chemosensitivities, with the latter population showing more apoptosis when exposed to imatinib (Litzenburger et al., 2017). Therefore, we asked whether BRD4 inhibition increases imatinib sensitivity in K562 cells. If so, it would imply that the observed state shift functionally alters K562 cells as opposed to simply modulating a cell-surface marker. We first pretreated K562 cells with either DMSO or JO1 and then challenged each pretreatment group with either DMSO or imatinib and stained for apoptosis. In the DMSO pretreatment group, the percentage of CD24high cells rose to 54% on average, while for JO1-pretreated cells the mean was 17% (Figure 5F). When imatinib was added, a substantially greater fraction of JO1-pretreated cells underwent apoptosis relative to DMSO-pretreated cells (Figures 5F and 5G; two-way ANOVA p < 0.01). Thus, JO1 sensitizes K562 cells to imatinib. Furthermore, BRD4 CRISPRi partially phenocopied this sensitization (Figures S6E and S6F; Tukey’s honestly significant difference p = 0.68). This phenomenon is likely dosage dependent: in our experiments, CRISPRi reduced BRD4 mRNA levels by less than 50% (Figure S6D), whereas the JO1 concentration we used is expected to almost completely abolish BRD4 activity (Filippakopoulos et al., 2010). Thus, while a mild knockdown can reduce CD24 expression, greater inhibition may be necessary to induce imatinib sensitivity. Nevertheless, these results establish that BRD4 inhibition functionally and phenotypically shifts the underlying cell state of K562 cells.

Finally, we examined whether the JO1-induced K562 cell-state shift was a non-specific response to generic drug treatment.
We treated K562 cultures with a panel of cell-cycle inhibitors, another class of commonly used antineoplastic agents. We first confirmed that all drugs altered the proportions of cells in either G1 or G2/M phase (Figure S6G). Cultures remained under drug treatment until 5 days had elapsed, at which point we measured CD24 levels and stained for apoptosis (Figure S6H). JQ1 caused the greatest reduction in CD24high cells (one-way ANOVA p < 0.01) and induced significantly less apoptosis than lovastatin, its closest competitor (one-way ANOVA p < 0.01). Thus, JQ1’s effect on cell state appears to be mediated by a unique mechanism of action that is not readily replicated by cell-cycle perturbation.

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 a fluorescent reporter (Figure 6A) and tested this reagent in cell culture. When this element was transfected without transposase, merely 3.4% of cells registered as positive, likely due to the action of the self-cleaving ribozyme downstream of the transposon. However, when the construct was co-transfected with piggyBac, this figure rose as high as 48%, a 16-fold increase in signal (Figure 6B). Thus, this new construct, PB-SRT-tdTomato, allows us to collect cells carrying calling card insertions by fluorescence activated cell sorting (FACS).

Next, we delivered PB-SRT-tdTomato and HyPBase to the postnatal mouse cortex by performing adeno-associated viral (AAV) transduction (Cammack et al., 2020) in P0–2 pups and then generating scRNA-seq and scCC libraries between P14 and P28. Most transduced cells were neurons or astrocytes (Tables S2 and S3; Figure 6C; Methods S1), which is consistent with the known tropism of AAV9 (Cammack et al., 2020; Schuster et al., 2014). Therefore, we analyzed insertions in neurons (excluding neuroblasts and doublets) and astrocytes to determine whether scCC could recover biological differences between cell types in vivo. After calling peaks, we

Figure 6. scCC Deconvolves BRD4-Bound Loci in the Mouse Cortex

(A) Schematic of PB-SRT-tdTomato.

(B) Distribution of fluorescence intensity in K562 cells transfected with PB-SRT-tdTomato with and without piggyBac transposase.

(C) Neuron and astrocyte clusters from scRNA-seq analysis of mouse cortex libraries transduced with AAV-HyPBase and AAV-PB-SRT-tdTomato.

(D) Browser view of scCC HyPBase peaks in astrocytes and neurons alongside whole-cortex H3K27ac ChIP-seq.

(E) Expression specificity distributions of genes overlapping astrocyte or neuron peaks; horizontal lines indicate medians of the distributions. See also Figure S7.

TR, terminal repeat; Rz, ribozyme.
identified astrocyte-specific, neuron-specific, and shared BRD4 binding sites (Figure 6D). Since BRD4 ChIP-seq has not yet been reported for the mouse brain, we compared our peak calls to a recent cortical H3K27ac ChIP-seq dataset (Stroud et al., 2017), and, while this dataset is a mélange of all cell types in the brain, scCC peaks in both astrocytes and neurons showed statistically significant enrichment of H3K27ac signal (Figures S7A and S7C; KS test p < 10^{-9} in each case). Moreover, genes near astrocyte peaks were more likely to be specifically expressed in astrocytes and vice versa for genes near neuron peaks (Figure 4E; STAR Methods). Furthermore, Gene Ontology enrichment analysis (Mi et al., 2017) on the set of genes near astrocyte peaks included terms like “gliogenesis,” and “glial cell differentiation,” as well as copper metabolism (Figure S7B), a known function of astrocytes (Scheiber and Dringen, 2013), while the set of genes near neuronal peaks was enriched for terms related to synapse assembly, axonal guidance, and neuron development (Figure S7D). We conclude that scCC can accurately identify cell-type-specific BRD4 binding sites in vivo.

Last, we asked whether scCC in vivo could discriminate BRD4 binding between closely related cell types, much as we had shown in vitro with K562 cells. From our scRNA-seq data, we identified upper- and lower-layer cortical excitatory neurons (Figure 7B; Table S3; Methods S1) and compared HY-PBase scCC data between them to identify shared and specific BRD4-bound loci. As a positive control, we found a shared BRD4 binding site at the Pou3f2 (Brn-1) locus (Figure 7A), which was broadly expressed in both populations (Figure 7C) and has been used to label layers 2–4 of the postnatal cortex (Molyneaux et al., 2007; Pucilowska et al., 2012). Differential binding analysis showed specific BRD4 enrichment at Pou3f2 (Brn-1) in upper-layer neurons, which is more restricted to layers 2–4 than Pou3f3 (Fan et al., 2008; Molyneaux et al., 2007), while lower cortical neurons showed BRD4 binding at Bcl11b (Ctip2) and Foxp2, common markers of layer 5 and layer 6 neurons, respectively (Figure 7A; one-tailed Poisson p < 10^{-9} in each instance) (Molyneaux et al., 2007; Rasin et al., 2007). The expression patterns of these genes mirrored BRD4’s binding specificity, with Pou3f2’s expression mostly contained in the layer 2–4 cluster and the expression of Bcl11b and Foxp2 restricted to the layer 5–6 neuron population (Figure 7C). Thus, scCC can identify differentially bound loci between very similar cell types in vivo.

**DISCUSSION**

scCC enables simultaneous characterization of gene expression and TF binding in heterogeneous systems. The method is robust and flexible: we have demonstrated that it can map multiple kinds of DNA binding proteins—from sequence-specific TFs like SP1 and FOXA2, to indirect, chromatin-associated factors like BRD4 and BAP1—in a variety of in vitro systems and in vivo in the mouse cortex. Furthermore, our finding that cell-state transitions in K562 cells are mediated by bromodomain proteins including BRD4 demonstrates how scCC can lead to new hypotheses about transcriptional regulation in dynamic systems. Our approach fills a recognized void in the field (Shapiro et al., 2013; Shema et al., 2019) and is readily compatible with high-throughput droplet microfluidic platforms such as the 10x Chromium. We anticipate this technique will empower researchers to study TF binding in a variety of challenging ex vivo and in situ models.

The defining feature of scCC is the SRT. While here we have reported piggyBac and Sleeping Beauty SRTs (Methods S1), the self-reporting paradigm may be broadly generalizable. Expanding the palette of SRT systems could yield further insight into chromatin dynamics (Yoshida et al., 2017). Moreover, SRTs may enable multiplexed studies of TF binding, either through the simultaneous expression of many TFs, each tagged to a different transposase, or through the use of multiple barcode TF-piggyBac fusions expressed polyclonally in culture.
Since SRTs can be widely dispersed through the genome, full-length sequencing of self-reporting transcripts may find new PASs (Methods S1). Finally, SRTs could lead to new single-cell transposon-based assays. For example, just as CRISPR/Cas9 has been combined with scRNA-seq to assess the transcriptional effects of many single gene perturbations in parallel (Datlinger et al., 2017; Dixit et al., 2016), SRTs could enable massively multiplexed transposon mutagenesis screens to be read out by scRNA-seq.

One concern with calling cards is the potential for insertional mutagenesis of target genes leading to cell death and, consequently, false negatives. Previous work in diploid yeast found that calling cards are deposited into the promoters of essential and non-essential genes at comparable frequencies (Wang et al., 2011). Since mammalian genomes have much larger intergenic regions than yeast, human and mice genomes are likely also to tolerate calling card transpositions. Long-term follow-up of mice transduced intracranially with AAV calling cards showed no significant tissue pathology, behavioral deficits, developmental defects, or metabolic dysregulation (Cammack et al., 2020). This suggests calling cards imposes, at most, a small mutagenic burden, though more studies are needed to verify this.

Another potential drawback of calling cards is that exogenous expression of a TF at supraphysiological levels may lead to ectopic binding and, consequently, false positives. We note that over 90% of our peaks from scCC of SP1 in HCT-116 cells and FOXA2 in HepG2 cells were within 1,000 bp of a ChIP-seq peak from the respective TF. This suggests that calling card peaks reflect endogenous binding, though this behavior may vary by factor. Overexpression might also alter the transcriptome of transfected cells. Comparing gene expression levels between cells treated with TF-piggyBac and the undirected piggyBac control cells can determine whether there is transcriptional perturbation and to what extent. Tagging the endogenous TF locus with piggyBac ensures native expression levels and would alleviate both concerns.

The relatively few insertions recovered on a per-cell basis inflates the number of cells that must be analyzed. We recommend processing enough cells to obtain at least 15,000 insertions to analyze BRD4-bound SEs with undirected piggyBac, and at least 30,000 insertions for both constructs in TF-directed experiments. This should achieve moderate sensitivities (~50%; Methods S1) that can be increased by collecting more insertions. The scant data recovered on a per-cell level likely stem from limited transposase activity—up to 15–30 insertions per cell for PBase (Kettlun et al., 2011; Saridey et al., 2009; Wang et al., 2008; Wilson et al., 2007) and potentially up to 100 for HyPBase (Kalhor et al., 2018; Yusa et al., 2011)—and the low capture rate of mRNA transcripts in droplet scRNA-seq (Hwang et al., 2018). This sparsity precludes certain kinds of analyses, such as multimodal data integration. Moreover, piggyBac’s strict preference for TTAA tetramers also contributes to broader peaks with lower spatial resolution. While we overcame the latter constraint by focusing on peak centers and narrow peaks, peak width is inversely correlated with the number of insertions analyzed; as such, improving recovery of SRTs from single cells should be prioritized. Some of these gains may come organically as the transcript capture rates of scRNA-seq technologies improve.

Since the per-cell costs for scRNA-seq are falling exponentially (Svensson et al., 2018), combining scCC with sample multiplexing strategies like cell hashing (Stoeckius et al., 2018) or combinatorial barcoding (Rosenberg et al., 2018) may be an attractive approach to increase sensitivity.

Finally, calling card insertions, being integrated into the genome and preserved through mitosis, could serve as a molecular memory for recording TF binding events. The use of an inducible transposase (Qi et al., 2017) would enable the recording and identification of temporarily 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 (Rodriguez-Fraticelli et al., 2018; Sun et al., 2014). Multiplexing several SRTs across every cell in an organism could code lineage in a cumulative and combinatorially diverse fashion, generating high-resolution cellular phylogenies.

**STAR METHODS**

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **RESOURCE AVAILABILITY**
  - Lead Contact
  - Materials Availability
  - Data and Code Availability
- **EXPERIMENTAL MODEL AND SUBJECT DETAILS**
- **METHOD DETAILS**
  - DNA- versus RNA-based recovery
  - In vitro bulk calling card experiments
  - Isolation and RT of bulk RNA
  - Amplifying self-reporting transcripts from RNA
  - Generation of bulk RNA calling card libraries
  - Colony formation assay
  - In vitro single cell calling card experiments
  - Single cell RNA-seq library preparation
  - Single cell calling cards library preparation
  - Staining protocols for K562 cells
  - JQ1 treatment of K562 cells
  - BRD4 CRISPRi of K562 cells
  - Imatinib treatments of K562 cells
  - Cell cycle perturbation of K562 cells
  - SRT-tdTomato fluorescence validation
  - In vivo scCC experiments
- **QUANTIFICATION AND STATISTICAL ANALYSIS**
  - Interpreting calling card tracks
  - Sequencing and analysis: bulk DNA CC libraries
  - Sequencing and analysis: bulk RNA CC libraries
  - Sequencing and analysis: scRNA-seq libraries
  - Sequencing and analysis: scCC libraries
  - Peak calling on calling card data
  - TF binding analysis
SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cell.2020.06.037.

ACKNOWLEDGMENTS


## STAR★METHODS

### KEY RESOURCES TABLE

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Antibodies</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brilliant Violet 421 anti-human CD24 Antibody (clone ML5)</td>
<td>BioLegend</td>
<td>Cat# 311121; RRID: AB_10915556</td>
</tr>
<tr>
<td>Brilliant Violet 421 Mouse IgG2a, κ Isotype Ctrl Antibody (clone MOPC-173)</td>
<td>BioLegend</td>
<td>Cat# 400259; RRID: AB_10895919</td>
</tr>
<tr>
<td>APC anti-human CD24 Antibody (clone ML5)</td>
<td>BioLegend</td>
<td>Cat# 311117; RRID: AB_1877150</td>
</tr>
<tr>
<td>APC Rat IgG2a, κ Isotype Ctrl (clone RTK2758)</td>
<td>BioLegend</td>
<td>Cat# 400511; RRID: AB_2814702</td>
</tr>
<tr>
<td><strong>Bacterial and Virus Strains</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AAV9-PB-SRT-tdTomato</td>
<td>Joseph D. Dougherty (Cammack et al., 2020)</td>
<td>N/A</td>
</tr>
<tr>
<td>AAV9-HyPBase</td>
<td>Joseph D. Dougherty (Cammack et al., 2020)</td>
<td>N/A</td>
</tr>
<tr>
<td>Lenti-dCas9-KRAB</td>
<td>This study</td>
<td>N/A</td>
</tr>
<tr>
<td>Lenti-BRD4-CRISPRi</td>
<td>This study</td>
<td>N/A</td>
</tr>
<tr>
<td>Lenti-NT-CRISPRi</td>
<td>This study</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Chemicals, Peptides, and Recombinant Proteins</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DMEM</td>
<td>GIBCO</td>
<td>Cat# 11965-084</td>
</tr>
<tr>
<td>Antibiotic-Antimycotic (100X)</td>
<td>GIBCO</td>
<td>Cat# 15240-062</td>
</tr>
<tr>
<td>FBS</td>
<td>Peak Serum</td>
<td>Cat# PS-FB3</td>
</tr>
<tr>
<td>RPMI 1640 Medium</td>
<td>GIBCO</td>
<td>Cat# 11875-085</td>
</tr>
<tr>
<td>Lipofectamine 3000 Transfection Reagent</td>
<td>Invitrogen</td>
<td>Cat# L3000015</td>
</tr>
<tr>
<td>Trypsin-EDTA solution</td>
<td>Sigma-Aldrich</td>
<td>Cat# T4049</td>
</tr>
<tr>
<td>DPBS, no calcium, no magnesium</td>
<td>GIBCO</td>
<td>Cat# 14190-136</td>
</tr>
<tr>
<td>RNAProtect Cell Reagent</td>
<td>QIAGEN</td>
<td>Cat# 76526</td>
</tr>
<tr>
<td>2-Mercaptoethanol</td>
<td>GIBCO</td>
<td>Cat# 21985-023</td>
</tr>
<tr>
<td>RNase-Free DNase Set</td>
<td>QIAGEN</td>
<td>Cat# 79254</td>
</tr>
<tr>
<td>Maxima H Minus Reverse Transcriptase</td>
<td>Thermo Scientific</td>
<td>Cat# EP0752</td>
</tr>
<tr>
<td>Advantage® UltraPure PCR Deoxynucleotide Mix</td>
<td>Takara Bio</td>
<td>Cat# 639125</td>
</tr>
<tr>
<td>RNaseOUT Recombinant Ribonuclease Inhibitor</td>
<td>Invitrogen</td>
<td>Cat# 10777019</td>
</tr>
<tr>
<td>TransIT-LT1 Transfection Reagent</td>
<td>Mirus</td>
<td>Cat# MIR2304</td>
</tr>
<tr>
<td>RNase H</td>
<td>New England BioLabs</td>
<td>Cat# M0297S</td>
</tr>
<tr>
<td>HiFi HotStart ReadyMix (2X)</td>
<td>Kapa Biosoys</td>
<td>Cat# K2601</td>
</tr>
<tr>
<td>AMPure XP beads</td>
<td>Beckman Coulter</td>
<td>Cat# A63880</td>
</tr>
<tr>
<td>Puromycin dihydrochloride</td>
<td>Sigma-Aldrich</td>
<td>Cat# P8833</td>
</tr>
<tr>
<td>Crystal violet</td>
<td>Sigma-Aldrich</td>
<td>Cat# C0775</td>
</tr>
<tr>
<td>Methanol</td>
<td>Fisher Scientific</td>
<td>Cat# A452-4</td>
</tr>
<tr>
<td>Formaldehyde</td>
<td>Fisher Scientific</td>
<td>Cat# BP31-500</td>
</tr>
<tr>
<td>High Sensitivity D1000 Reagents</td>
<td>Agilent</td>
<td>Cat# 5067-5585</td>
</tr>
<tr>
<td>Ficoll PM400 (Dry Powder)</td>
<td>GE Healthcare</td>
<td>Cat# 17030010</td>
</tr>
<tr>
<td>NxGen® RNase Inhibitor</td>
<td>Lucigen</td>
<td>Cat# 30281-1</td>
</tr>
<tr>
<td>Dynabeads MyOne Silane</td>
<td>Life Technologies</td>
<td>Cat# 37002D</td>
</tr>
<tr>
<td>IDTE pH 8.0 (1X TE Solution)</td>
<td>IDT</td>
<td>Cat# 11-05-01-13</td>
</tr>
<tr>
<td>High Sensitivity D5000 Reagents</td>
<td>Agilent</td>
<td>Cat# 5067-5593</td>
</tr>
<tr>
<td>NEBuffer 2</td>
<td>New England BioLabs</td>
<td>Cat# B7002S</td>
</tr>
<tr>
<td>Buffer EB</td>
<td>QIAGEN</td>
<td>Cat# 19086</td>
</tr>
<tr>
<td>Hibernate-A Medium</td>
<td>GIBCO</td>
<td>Cat# A1247501</td>
</tr>
</tbody>
</table>

(Continued on next page)
<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-(+)-Trehalose dihydrate</td>
<td>Sigma-Aldrich</td>
<td>Cat# T9531</td>
</tr>
<tr>
<td>B-27 Supplement (50X), serum free</td>
<td>GIBCO</td>
<td>Cat# 17504044</td>
</tr>
<tr>
<td>0.5M EDTA, pH 8.0</td>
<td>Corning</td>
<td>Cat# 46-034-CI</td>
</tr>
<tr>
<td>Papain, Lyophilized</td>
<td>Worthington Biochemical</td>
<td>Cat# LS003118</td>
</tr>
<tr>
<td>Deoxyribonuclease I, Filtered</td>
<td>Worthington Biochemical</td>
<td>Cat# LS002060</td>
</tr>
<tr>
<td>Trypsin Inhibitor, Ovomucoid</td>
<td>Worthington Biochemical</td>
<td>Cat# LS003087</td>
</tr>
<tr>
<td>Bovine Serum Albumin</td>
<td>Sigma-Aldrich</td>
<td>Cat# A9418</td>
</tr>
<tr>
<td>OptiPrep Density Gradient Medium</td>
<td>Sigma-Aldrich</td>
<td>Cat# D1556</td>
</tr>
<tr>
<td>HBSS (10X)</td>
<td>GIBCO</td>
<td>Cat# 14185052</td>
</tr>
<tr>
<td>Magnesium chloride</td>
<td>Sigma-Aldrich</td>
<td>Cat# M4880</td>
</tr>
<tr>
<td>Magnesium sulfate</td>
<td>Sigma-Aldrich</td>
<td>Cat# M2643</td>
</tr>
<tr>
<td>Calcium chloride dihydrate</td>
<td>Sigma-Aldrich</td>
<td>Cat# C7902</td>
</tr>
<tr>
<td>D-(+)-Glucose</td>
<td>Sigma-Aldrich</td>
<td>Cat# G7021</td>
</tr>
<tr>
<td>Dimethyl sulfoxide (DMSO)</td>
<td>Sigma-Aldrich</td>
<td>Cat# D2650</td>
</tr>
<tr>
<td>Cell Staining Buffer</td>
<td>BioLegend</td>
<td>Cat# 420201</td>
</tr>
<tr>
<td>Annexin V Binding Buffer</td>
<td>BioLegend</td>
<td>Cat# 422201</td>
</tr>
<tr>
<td>SuperScript VILO cDNA Synthesis Kit</td>
<td>Invitrogen</td>
<td>Cat# 11754250</td>
</tr>
<tr>
<td>PowerUp SYBR Green Master Mix</td>
<td>Applied Biosystems</td>
<td>Cat# 25742</td>
</tr>
<tr>
<td>(+)-JQ1</td>
<td>Selleck Chemicals</td>
<td>Cat# S7110</td>
</tr>
<tr>
<td>Propidium iodide (PI)</td>
<td>Invitrogen</td>
<td>Cat# P3566</td>
</tr>
<tr>
<td>Hoechst 33342</td>
<td>Thermo Scientific</td>
<td>Cat# 62249</td>
</tr>
<tr>
<td>Blasticidin S HCl</td>
<td>GIBCO</td>
<td>Cat# A1113903</td>
</tr>
<tr>
<td>Lenti-X Concentrator</td>
<td>Takara Bio</td>
<td>Cat# 631232</td>
</tr>
<tr>
<td>Lipofectamine 2000 Transfection Reagent</td>
<td>Invitrogen</td>
<td>Cat# 11668030</td>
</tr>
<tr>
<td>Polybrene Infection / Transfection Reagent</td>
<td>Sigma-Aldrich</td>
<td>Cat# TR-1003</td>
</tr>
<tr>
<td>Esp3I</td>
<td>New England BioLabs</td>
<td>Cat# R0734S</td>
</tr>
<tr>
<td>T4 DNA Ligase</td>
<td>New England BioLabs</td>
<td>Cat# M0202S</td>
</tr>
<tr>
<td>IMDM</td>
<td>GIBCO</td>
<td>Cat# 12440046</td>
</tr>
<tr>
<td>Penicillin-Streptomycin (10,000 U/mL)</td>
<td>GIBCO</td>
<td>Cat# 15140122</td>
</tr>
<tr>
<td>Imatinib mesylate</td>
<td>Sigma-Aldrich</td>
<td>Cat# SML1027</td>
</tr>
<tr>
<td>Lovastatin</td>
<td>Sigma-Aldrich</td>
<td>Cat# M2147</td>
</tr>
<tr>
<td>Nocodazole</td>
<td>Sigma-Aldrich</td>
<td>Cat# M1404</td>
</tr>
<tr>
<td>CVT-313</td>
<td>Sigma-Aldrich</td>
<td>Cat# 238803</td>
</tr>
<tr>
<td>RO-3306</td>
<td>Sigma-Aldrich</td>
<td>Cat# SML0569</td>
</tr>
<tr>
<td>Annexin V-FITC</td>
<td>BioLegend</td>
<td>Cat# 640905</td>
</tr>
</tbody>
</table>

Critical Commercial Assays

- **Neon Transfection System 100 μL Kit** | Invitrogen | Cat# MPK10025 |
- **RNasey Plus Mini Kit** | QIAGEN | Cat# 74134 |
- **Qubit R1NA HS Assay Kit** | Invitrogen | Cat# Q32852 |
- **Qubit dsDNA HS Assay Kit** | Invitrogen | Cat# Q32851 |
- **Nextera XT DNA Library Preparation Kit** | Illumina | Cat# FC-131-1024 |
- **High Sensitivity D1000 ScreenTape** | Agilent | Cat# 5067-5584 |
- **Chromium Single Cell 3’ Library & Gel Bead Kit v2** | 10x Genomics | Cat# PN-120267 |
- **High Sensitivity D5000 ScreenTape** | Agilent | Cat# 5067-5592 |
- **Nextera Mate Pair Library Prep Kit** | Illumina | Cat# FC-132-1001 |

Deposited Data

- **K562 CpG islands** | Richard Myers | GEO: GSM1014203 |
- **HCT-116 SP1 ChIP-seq** | Richard Myers | ENCODE: ENCFF000PCT |
Continued

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCT-116 CTCF ChIP-seq</td>
<td>Richard Myers</td>
<td>ENCODE: ENCF0000ZC</td>
</tr>
<tr>
<td>HCT-116 ChIP-seq input control (SP1, CTCF)</td>
<td>Richard Myers</td>
<td>ENCODE: ENCF0000B0</td>
</tr>
<tr>
<td>HCT-116 BRD4 ChIP-seq</td>
<td>Ron Firestein</td>
<td>SRA: SRR2481799</td>
</tr>
<tr>
<td>HCT-116 ChIP-seq input control (BRD4)</td>
<td>Ron Firestein</td>
<td>SRA: SRR2481800</td>
</tr>
<tr>
<td>HCT-116 H3K27ac ChIP-seq</td>
<td>Bradley Bernstein</td>
<td>ENCODE: ENCF082JPN, ENCF0176BXC</td>
</tr>
<tr>
<td>HCT-116 H3K4me1 ChIP-seq</td>
<td>Bradley Bernstein</td>
<td>ENCODE: ENCF088BW, ENCF084MJI</td>
</tr>
<tr>
<td>HCT-116 H3K4me2 ChIP-seq</td>
<td>Bradley Bernstein</td>
<td>ENCODE: ENCF0936MM, ENCF097OOL</td>
</tr>
<tr>
<td>HCT-116 H3K4me3 ChIP-seq</td>
<td>Bradley Bernstein</td>
<td>ENCODE: ENCF183OZ, ENCF065FPR</td>
</tr>
<tr>
<td>HCT-116 H3K9me2 ChIP-seq</td>
<td>Bradley Bernstein</td>
<td>ENCODE: ENCF760OZN, ENCF065FPR</td>
</tr>
<tr>
<td>HCT-116 H3K9me3 ChIP-seq</td>
<td>Bradley Bernstein</td>
<td>ENCODE: ENCF65FPR</td>
</tr>
<tr>
<td>HCT-116 H3K27me3 ChIP-seq</td>
<td>Bradley Bernstein</td>
<td>ENCODE: ENCF281SBT, ENCF124GII</td>
</tr>
<tr>
<td>HCT-116 H3K36me3 ChIP-seq</td>
<td>Bradley Bernstein</td>
<td>ENCODE: ENCF850EAH, ENCF312RKB</td>
</tr>
<tr>
<td>HCT-116 H3K79me2 ChIP-seq</td>
<td>Bradley Bernstein</td>
<td>ENCODE: ENCF865KWP, ENCF947YPU</td>
</tr>
<tr>
<td>HCT-116 H4K20me1 ChIP-seq</td>
<td>Bradley Bernstein</td>
<td>ENCODE: ENCF070JUDY, ENCF334HHB</td>
</tr>
<tr>
<td>HCT-116 ChIP-seq input control (H3K27ac, H3K4me1, H3K4me2, H3K4me3, H3K9me2, H3K9me3, H3K27me3, H3K36me3, H3K79me2, H4K20me1)</td>
<td>Bradley Bernstein</td>
<td>ENCODE: ENCF048ZQO, ENCF827YX</td>
</tr>
<tr>
<td>K562 BRD4 ChIP-seq</td>
<td>Bradley Bernstein</td>
<td>ENCODE: ENCF413R4Q</td>
</tr>
<tr>
<td>K562 H3K27ac ChIP-seq</td>
<td>Bradley Bernstein</td>
<td>ENCODE: ENCF0008XH</td>
</tr>
<tr>
<td>K562 ChIP-seq input control (BRD4, H3K27ac)</td>
<td>Bradley Bernstein</td>
<td>ENCODE: ENCF000BWE</td>
</tr>
<tr>
<td>K562 SP1 ChIP-seq</td>
<td>Michael Snyder</td>
<td>ENCODE: ENCF0029PL, ENCF0029EC</td>
</tr>
<tr>
<td>K562 ChIP-seq input control (SP1)</td>
<td>Michael Snyder</td>
<td>ENCODE: ENCF0029EI, ENCF0029EA</td>
</tr>
<tr>
<td>GFP2 FOX2A ChIP-seq</td>
<td>Richard Myers</td>
<td>ENCODE: ENCF000000X</td>
</tr>
<tr>
<td>GFP2 ChIP-seq input control (FOX2A)</td>
<td>Richard Myers</td>
<td>ENCODE: ENCF000000PV</td>
</tr>
<tr>
<td>OCM-1A H3K27ac ChIP-seq</td>
<td>Michael Onken</td>
<td>ENCODE: ENCF000000KH</td>
</tr>
<tr>
<td>Mouse cortex H3K27ac ChIP-seq</td>
<td>Michael Greenberg</td>
<td>SRA: SRR6129714</td>
</tr>
<tr>
<td>Mouse cortex ChIP-seq input control (H3K27ac)</td>
<td>Michael Greenberg</td>
<td>SRA: SRR6129795</td>
</tr>
<tr>
<td>K562 RNA Pol II ChIA-PET</td>
<td>Yijun Ruan</td>
<td>ENCODE: ENCF000000KH</td>
</tr>
<tr>
<td>HCT-116 DNase-seq</td>
<td>John Stamatoyannopoulos</td>
<td>ENCODE: ENCF000000DK</td>
</tr>
<tr>
<td>HCT-116 ATAC-seq</td>
<td>Srinarsa Pradhan</td>
<td>SRA: SRR5453778</td>
</tr>
<tr>
<td>HCT-116 ATAC-seq control</td>
<td>Michael Guertin</td>
<td>GEO: GSE92674</td>
</tr>
<tr>
<td>HCT-116 CpG islands</td>
<td>Richard Myers</td>
<td>GEO: GSM1014209</td>
</tr>
<tr>
<td>Sequencing data and processed output</td>
<td>This study</td>
<td>GEO: GSE148448</td>
</tr>
</tbody>
</table>

Experimental Models: Cell Lines

<table>
<thead>
<tr>
<th>Cell Line</th>
<th>Source</th>
<th>Cat#</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neuro-2a (N2a)</td>
<td>ATCC</td>
<td>CCL-131</td>
</tr>
<tr>
<td>K562</td>
<td>ATCC</td>
<td>CCL-243</td>
</tr>
<tr>
<td>Hep G2</td>
<td>ATCC</td>
<td>CRL-11268</td>
</tr>
<tr>
<td>OCM-1A</td>
<td>Michael Onken</td>
<td>HB-8065</td>
</tr>
<tr>
<td>HCT 116</td>
<td>ATCC</td>
<td>CCL-247</td>
</tr>
<tr>
<td>293T/17 [HEK293T/17]</td>
<td>ATCC</td>
<td>CCL-11268</td>
</tr>
</tbody>
</table>

Experimental Models: Organisms/Strains

Mouse: C57BL/6J | Joseph D. Dougherty | N/A |
| Cat# CRL-11268 |

(Continued on next page)
### REAGENT or RESOURCE SOURCE IDENTIFIER

<table>
<thead>
<tr>
<th>Oligonucleotides</th>
<th>This study, see Table S4</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primers and oligonucleotides</td>
<td>This study</td>
<td>N/A</td>
</tr>
<tr>
<td>Recombinant DNA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pRM1024: PBase</td>
<td>This study</td>
<td>N/A</td>
</tr>
<tr>
<td>pRM1114: HyPBase</td>
<td>This study</td>
<td>N/A</td>
</tr>
<tr>
<td>pRM1023: SP1-PBase</td>
<td>This study</td>
<td>N/A</td>
</tr>
<tr>
<td>pRM1677: SP1-HyPBase</td>
<td>This study</td>
<td>N/A</td>
</tr>
<tr>
<td>pRM1882: FOXA2-HyPBase</td>
<td>This study</td>
<td>N/A</td>
</tr>
<tr>
<td>pRM1863: BAP1-HyPBase</td>
<td>This study</td>
<td>N/A</td>
</tr>
<tr>
<td>pRM1304: PB-SRT-Puro</td>
<td>This study</td>
<td>RRID: Addgene_154884</td>
</tr>
<tr>
<td>pRM1535: PB-SRT-tdTomato</td>
<td>This study</td>
<td>RRID: Addgene_154885</td>
</tr>
<tr>
<td>pCMV(CAT)T7-SB100</td>
<td>Zsuzsanna Izsák</td>
<td>RRID: Addgene_34879</td>
</tr>
<tr>
<td>pRM1665: SP1-SB100X</td>
<td>This study</td>
<td>RRID: Addgene_154887</td>
</tr>
<tr>
<td>pRM1668: SB-SRT-Puro</td>
<td>This study</td>
<td>RRID: Addgene_154888</td>
</tr>
<tr>
<td>pRM1217: AAV-HyPBase</td>
<td>Joseph D. Dougherty (Cammack et al., 2020)</td>
<td>N/A</td>
</tr>
<tr>
<td>pRM1648: AAV-PB-SRT-tdTomato</td>
<td>Joseph D. Dougherty (Cammack et al., 2020)</td>
<td>RRID: Addgene_154889</td>
</tr>
<tr>
<td>pUC19 Vector</td>
<td>New England BioLabs</td>
<td>Cat# N3041S</td>
</tr>
<tr>
<td>Lenti-dCas9-KRAB-blast</td>
<td>Gary Hon</td>
<td>RRID: Addgene_89567</td>
</tr>
<tr>
<td>sgOpti</td>
<td>Eric Lander &amp; David Sabatini</td>
<td>RRID: Addgene_85681</td>
</tr>
<tr>
<td>pMD2.G</td>
<td>Didier Trono</td>
<td>RRID: Addgene_12259</td>
</tr>
<tr>
<td>psPAX2</td>
<td>Didier Trono</td>
<td>RRID: Addgene_12260</td>
</tr>
<tr>
<td>pRM1889: BRD4 CRISPRi plasmid</td>
<td>This study</td>
<td>RRID: Addgene_154890</td>
</tr>
<tr>
<td>pRM1890: Non-targeting CRISPRi plasmid</td>
<td>Robi D. Mitra (Lalli et al., 2019)</td>
<td>RRID: Addgene_154891</td>
</tr>
</tbody>
</table>

### Software and Algorithms

- cutadapt 1.16 Martin, 2011 RRID: SCR_011841
- NovoAlign 3 Novocraft Technologies RRID: SCR_014818
- Cell Ranger 2.1.0 10x Genomics RRID: SCR_017344
- scanpy 1.3.7 Wolf et al., 2018 RRID: SCR_018139
- Drop-seq tools 1.11 Macosko et al., 2015 RRID: SCR_018142
- astroPy 3.2.1 Robitaille et al., 2013 RRID: SCR_018148
- WashU Human Epigenome Browser 46 Zhou et al., 2011 RRID: SCR_006208
- MEME-ChIP 4.11.2 Machanick and Bailey, 2011 RRID: SCR_001783
- Tomtom 5.1.0 Gupta et al., 2007 RRID: SCR_001783
- MACS 1.4.1 Zhang et al., 2008 RRID: SCR_013291
- BEDTools 2.27.1 Quinlan and Hall, 2010 RRID: SCR_006646
- NumPy 1.17.2 Oliphant, 2015 RRID: SCR_008633
- SciPy 1.4.1 Virtanen et al., 2020 RRID: SCR_008058
- statsmodels 0.10.1 Seabold and Perktold, 2010 RRID: SCR_016074
- matplotlib 3.0.3 Hunter, 2007 RRID: SCR_008624
- deeptools 3.0.1 Ramirez et al., 2016 RRID: SCR_016366
- ChromHMM 1.15 Ernst et al., 2011 RRID: SCR_018141
- liftOver Hinrichs et al., 2006 RRID: SCR_018160
- FlowCal 1.2.0 Castillo-Hair et al., 2016 RRID: SCR_018140
- PANTHER 14.0 Mi et al., 2017 RRID: SCR_004869

(Continued on next page)
RESOURCE AVAILABILITY

Lead Contact
Further information and requests for resources and reagents should be directed to, and will be fulfilled by, the Lead Contact, Robi D. Mitra (rmitra@wustl.edu).

Materials Availability
Plasmids generated in this study have been deposited to Addgene, where possible, and are available to the community. Plasmids encoding the piggyBac transposase are not available through Addgene due to licensing restrictions. These plasmids are available upon request to the Lead Contact.

Data and Code Availability
Data generated in this study have been submitted to the Gene Expression Ominbus (GEO) with accession number GSE148448. All code used to analyze the data is available online at https://github.com/arnavm/calling_cards.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

HCT-116, N2a, HEK293T, and HepG2 cells were cultured in Dulbecco’s Modified Eagle Medium (DMEM) supplemented with 10% fetal bovine serum (FBS) and 1% of either penicillin-streptomycin or antibiotic-antimycotic. K562 (unless otherwise indicated) and OCM-1A cells were grown under the same conditions as described above, replacing DMEM with RPMI 1640 Medium. Cells were grown at 37°C with 5% carbon dioxide (CO2). Media was replenished every 2 days. HepG2 cells were a gift from the Genome Engineering iPSC Center (GEiC) at Washington University in St. Louis School of Medicine. OCM-1A cells were a gift from Dr. Michael Onken. For the CD24high/CD24low cell state analyses, K562 cells were grown in IMDM containing 10% v/v FBS and 1% penicillin-streptomycin at 37°C.

Continued
concentration of 800,000 cells/ml. For experiments, cells were seeded at mid-log phase concentrations, around 400,000 cells/ml. At this point, ratio of CD24high/CD24low cells was approximately 1:1, as determined by flow cytometry.

All mouse experiments were done following procedures described in (Cammack et al., 2020). In brief, we cloned the PB-SRT-ttdTomato 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 $1.1 \times 10^{13}$ and $2.2 \times 10^{13}$ 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-ttdTomato only. After 2 to 4 weeks, we sacrificed mice and dissected the cortex (8 libraries) or hippocampus (1 library). The sex of mice was not taken into consideration. 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.

METHOD DETAILS

**DNA- versus 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 and 2.5 μg of the PB-SRT-Puro plasmid using Lipofectamine 3000 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 to a final concentration of 2 μg/ml and colonies were grown under selection for two weeks. We obtained approximately 2,300 colonies. All cells were pooled together and split into two populations. One half was subjected to DNA extraction, self-igation, and inverse PCR, as described previously (Wang et al., 2012a), with the following modification: digestion with MspI was not performed as the SRT construct contained an second MspI cut site near the terminal repeat. The other half of cells 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. Each replicate contained 2×10⁶ 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. 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 (the latter a gift from Dr. Zsuzsanna Izsövé; Mátés et al., 2009) plasmids. After transfection, each replicate was plated into a 10 cm dish. For the OCM-1A library, we transfected 1.25 μg of PB-SRT-Puro and 1.25 μg of either HyPBase or BAP1-HyPBase (the latter a gift from Dr. Michael Onken; Yen et al., 2018) using the TransIT-LT1 transfection reagent following manufacturer’s protocol for 6-well plates. Puromycin was added after 24 hours to a final concentration of 2 μg/ml. Cells were grown under selection for one week, by which time almost all negative control transfectants were dead. After 7 days, we dissociated each replicate with trypsin-EDTA and created single cell suspensions in phosphate-buffered saline (PBS). 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 −80°C in RNAProtect Cell Reagent.

**Isolation and RT of bulk RNA**

Total RNA was isolated from each replicate using the RNEasy Plus Mini Kit following manufacturer’s instructions. Briefly, cell pellets were resuspended in 600 μl of Buffer RLT Plus with 1% 2-mercaptoethanol. Cells were homogenized 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 DNasel was also performed. After washing, RNA was eluted in 40 μl RNase-free H2O. RNA was quantitated using the Qubit RNA HS Assay Kit.

We performed first strand synthesis on each replicate with Maxima H Minus Reverse Transcriptase. We mixed 2 μl of total RNA with 1 μl 10 mM dNTPs and 1 μl of 50 μM SMART_dT18VN primer (for a complete list of oligonucleotides, see Table S4), brought the total volume up to 14 μl, and incubated at 65°C for 5 minutes. After transferring to ice and letting rest for 1 minute, we added 4 μl 5X Maxima RT Buffer, 1 μl RNaseOUT, 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 50°C for 1 hour followed by heat inactivation at 85°C for 10 minutes. Finally, we digested with 1 μl RNaseH at 37°C for 30 minutes. cDNA was stored at −20°C.

**Amplifying self-reporting transcripts from 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, 0.5 μl 25 μM SMART_dT18VN 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_ttdTomato_F1 primer (in the case of tdTomato screening). The mixture was brought up to 25 μl with ddH2O. Thermocycling parameters were as follows: 95°C for 3 minutes; 20 cycles of: 98°C for 20 s–65°C for 30 s–72°C for 5 minutes; 72°C for 10 minutes; hold at 4°C forever. As a control, cDNA quality can be assessed with exon-spanning primers for β-actin [see Table S4 for examples of human primers (Raff et al., 1997)] under the same thermocycling settings.
PCR products were purified using AMPure XP beads. 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 s 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 ddH2O 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 using the dsDNA High Sensitivity Assay Kit.

Generation of bulk RNA calling card libraries
Calling card libraries from bulk RNA were generated using the Nextera XT DNA Library Preparation Kit. One nanogram of PCR product was resuspended in 5 µl ddH2O. 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 55 °C. 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: 95 °C for 30 s; 13 cycles of: 95 °C for 10 s–50 °C for 30 s; 72 °C for 5 minutes; hold at 4 °C 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 ddH2O and quantitated on an Agilent TapeStation 4200 System using the High Sensitivity D1000 ScreenTape.

Colony formation assay
For the piggyBac transfectants, we electroporated 500,000 HCT-116 cells with 750 ng of PB-SRT-Puro and 750 ng of either HyPBase or SP1-HyPBase plasmid using the aforementioned Neon settings. For the SRT-only conditions, cells received 750 ng of PB-SRT-Puro. We used the same design for the Sleeping Beauty transfectants, replacing the transposases with either SB100X or SP1-SB100X and using SB-SRT-Puro as the transposase. Each set of 500,000 cells were plated in a single well of a 6-well plate and allowed to recover overnight. We then added puromycin to a final concentration of 2 µg/ml. We cultured cells under selection for one week. Colonies were visualized using a solution comprising 1X PBS, 1% formaldehyde, 1% methanol, and 0.05% w/v crystal violet. After aspirating median, we covered cells with this solution, staining plates for 20 minutes washing under cold water and air drying.

In vitro single cell calling card experiments
All cell lines (HCT-116, K562, N2a, HepG2, and OCM-1A) were cultured as described above. HCT-116 cells were transfected using Neon electroporation with the aforementioned settings. K562 cells were electroporated with the following settings–pulse voltage: 1,450 V; pulse width: 10 ms; pulse number: 3. N2a cells were electroporated with the following settings–pulse voltage: 1,050 V; pulse width: 30 ms; pulse number: 2. HepG2 cells were electroporated with the following settings–pulse voltage: 1,200 V; pulse width: 50 ms; pulse number: 1. Each replicate for electroporation was comprised of 2x10⁶ cells. All cells were allowed to recover for 24 hours before undergoing puromycin selection. A negative control replicate, transfected only with PB-SRT-Puro, was treated identically in parallel. Replicates were harvested once the negative control cells had died. For the species mixing experiment, we transfected one replicate each of HCT-116 and N2a cells with 5 µg PB-SRT-Puro and 5 µg HyPBase. For the cell line mixing experiment, we transfected four replicates each of HCT-116 and K562 cells with 5 µg PB-SRT-Puro and 5 µg HyPBase. In all cases, cells were cultured independently and mixed immediately prior to generating single cell emulsions. For single cell calling cards analysis of SP1 binding in HCT-116 and K562 cells, we transfected four replicates each with 5 µg PB-SRT-Puro and 5 µg SP1-HyPBase. These libraries were not mixed. We used the demultiplexed data from the cell line mixing experiment with HyPBase as controls. For single cell calling cards analysis of FOXA2 binding in HepG2 cells, we transfected six replicates each with 5 µg PB-SRT-Puro; three of these replicates were co-transfected with 5 µg HyPBase, while the other three were co-transfected with 5 µg FOXA2-HyPBase. We used the mouse ortholog of FOXA2, which has 97% primary sequence identity with human FOXA2. For single cell calling cards analysis of BAP1 binding in OCM-1A cells, we lipofected (as described above) six replicates each with 1.25 µg PB-SRT-Puro; three of these replicates were co-transfected with 1.25 µg BAP1-HyPBase, while the other three were co-transfected with 1.25 µg BAP1-HyPBase.

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. 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. Gel-in-emulsion (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 −20°C 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, 10 μl of 10 mM dNTPs, 2.5 μl RNase Inhibitor 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 ddH₂O. Finally, we added 5 μl Maxima H Minus Reverse Transcriptase, mixed by flicking, and centrifuged briefly. This reaction was incubated at 25°C for 30 minutes followed by 50°C for 90 minutes and heat inactivated at 85°C for 5 minutes.

The solution was purified using DynaBeads MyOne Silane 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 ddH₂O. We then prepared a PCR primer cocktail comprising 5 μl of 100 μM Bio_Ilumina_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. One μl of this cocktail was added to the PCR mixture and placed in a thermocycler. Thermocycling settings were as follows: 98°C for 3 minutes; 20-22 cycles of 98°C for 20 s–67°C for 30 s–72°C for 5 minutes; 72°C for 10 minutes; 4°C 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.

Single cell calling card library preparation was performed using the Nextera Mate Pair Sample Prep Kit 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 ddH₂O, 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 30°C. 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 s; temperature: 6°C.

The library preparation was performed on manufacturer’s instructions until adaptor ligation. We designed custom adapters (Table S4) 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, then heating in a thermocycler at 95°C for 5 minutes, then holding at 70°C for 15 minutes, then ramping down at 1°C until it reached 25°C, holding at that temperature for 5 minutes, before keeping at 4°C forever. One microliter of this custom adapter mix was used in place of the manufacturer’s recommended DNA Adaptor 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 ddH₂O 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: 98°C for 30 s; 15 cycles of: 98°C for 10 s–60°C for 30 s–72°C for 2 minutes; 72°C for 5 minutes; 4°C 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 and quantitated on an Agilent TapeStation 4200 System using the High Sensitivity D1000 ScreenTape.

**Staining protocols for K562 cells**

CD24 surface protein was quantified using monoclonal human antibodies. Cells were spun down at 300g for 3 minutes and washed twice with 1 ml of Cell Staining Buffer. The cell pellet was then resuspended in 50 μl of Cell Staining Buffer containing 0.2 μg/ml of either CD24-APC or CD24-BV421. The tube was rotated at 4°C in the dark for 30 minutes. After, cells were washed twice (as before) and finally resuspended in 200 μl of Cell Staining Buffer. Cells were excited with 450/45 and 660/20 lasers (wavelength/filter bandwidth, both in nm). For concomitant analysis of DNA content, we used CD24-APC. Cells were incubated with 10 μg/ml Hoechst 33342 in 5 ml of growth medium for 30 minutes prior to the staining protocol. For simultaneous assessment of apoptosis, cells were stained with CD24-BV421. After the final wash, instead of resuspending in 200 μl of Cell Staining Buffer, cells were washed twice with Annexin V Staining Buffer. Cells were then incubated in 50 μl Annexin V Staining Buffer containing 0.2 μg/ml Annexin V-FITC and 100 μg/ml propidium iodide (PI). The reaction was incubated for 15 minutes at room temperature in the dark. Afterward, we added 150 μl of Annexin V Staining Buffer and proceeded to flow cytometry. All samples were measured on a Beckman-Coulter CytoFLEX S flow cytometer. Cells were excited with 450/45, 525/40, and 610/20 lasers. We collected 10,000 events per sample. The resulting data were processed with FlowJo Software for Mac Version 10.

**JQ1 treatment of K562 cells**

For the longitudinal treatment of K562 cells with JQ1, we seeded cells at log phase growth and treated them with growth medium containing DMSO (~0.4% final concentration) or 250 nM JQ1 (dissolved in DMSO). Medium was replaced every 48 hours without splitting. On days 1, 2, 3, 4, and 7, cells were split in half: one half was stained for CD24 and DNA content, while the other half was stained for CD24 and apoptosis (both described above). Experiments were performed with three biological replicates.
For qRT-PCR, we cultured K562 cells in either DMSO or 250 nM JQ1, in triplicate, and collected cells at 0, 3, 6, 9, 12, and 24 hours of treatment. Cells were pelleted, resuspended in 300 μl of RNA CellProtect, and stored at −80°C. When we were ready to extract RNA, we thawed prepared, prepared samples using QIAGEN RNEasy Plus Mini Kit, and quantitated with the Qubit RNA High Sensitivity kit. We reverse transcribed 500 ng of RNA with the SuperScript VILO CDNA Synthesis Kit in a 20 μl reaction, with the following thermocycling parameters: 25°C for 10 minutes; 42°C for 2 hours; 85°C for 5 minutes. We then performed PCR with 2 μl of the RT product as template, 1 μl each of forward and reverse primer (10 μM), 6 μl ddH2O, and 10 μl PowerUp SYBR Green Master Mix. We ran the PCR on an ABI QuantStudio 3 with the following settings: 2 minutes at 50°C, then 2 minutes at 95°C (hot start); 45 cycles of 95°C for 15 s followed by 60°C for 1 minute. We generated melt curves after each PCR and all samples yielded a single peak. Gene-specific primers were obtained from PrimerBank (Wang et al., 2012b). Data were normalized to the levels of β-actin.

**BRD4 CRISPRi of K562 cells**

For CRISPRi, we first made lentivirus expressing dCas9-KRAB (Fulco et al., 2016; Xie et al., 2017) from Addgene plasmid #89567, a gift from Gary Hon, packaged in HEK293T cells along with pMD2.G (Addgene plasmid #12259) and psPAX2 (Addgene plasmid #12260), both gifts from Didier Trono. We cloned a BRD4 guide RNA, selected from the Dolcetto collection (Sanson et al., 2018), into the sgOpti plasmid (Addgene plasmid #85681, a gift from Eric Lander & David Sabatini) using Golden Gate assembly with Esp3I. We used an in-house pipeline to design a non-targeting gRNA sequence, which was cloned into CROP-seq-opti (Lalli et al., 2019). Plasmids were transfected into HEK293T cells using Lipofectamine 2000. Media was collected after 24 and 48 hours, and subsequently concentrated using Lenti-X Concentrator. Viral titers were functionally assed on HEK293T cells using the appropriate antibiotic (blasticidin or puromycin).

Next, we generated a polyclonal pool of dCas9-KRAB-expressing K562 cells. We seeded each well of a 6-well plate with 200,000 cells each containing 2 mL of growth media supplemented with 4 μg/ml polybrene and 1,000,000 infectious lentiviral particles for an estimated multiplicity of infection (MOI) of 5. Plates were centrifuged at 2,000g for 30 minutes and returned to the incubator. After 48 hours, cells were split to mid-log phase concentration (~400,000 cells/ml) and selected on blastcidin (10 μg/ml) for 48 hours. We made frozen stocks from these cells.

For the knockdown experiments, cells were thawed and allowed to recover for 4 days. We confirmed that the proportions of CD24(high)/CD24(low) was approximately equal at this point. We then seeded 200,000 cells into each well of a 6-well plate. Three wells remained cells in each well were split between two new wells. One well continued to receive medium supplemented with DMSO, while the other was treated with medium containing imatinib mesylate at a concentration of 1 μM. Remaining cells in each well were split to mid-log phase concentration (~400,000 cells/ml) and selected on blastcidin (10 μg/ml) for 48 hours. We made frozen stocks from these cells.

For the knockdown experiments, cells were thawed and allowed to recover for 4 days. We confirmed that the proportions of CD24(high)/CD24(low) was approximately equal at this point. We then seeded 200,000 cells into each well of a 6-well plate. Three wells received the BRD4 gRNA lentivirus, while the other three received the non-targeting gRNA lentivirus, at MOI 2.5. We followed the same transduction protocol described above. After 48 hours of incubation, puromycin was added to the medium at a final concentration of 2 μg/ml. After a further 48 hours, cells were passaged 1:1 into 10 cm dishes containing 10 mL of growth medium. The surviving cells were expanded for a further 5 days before being stained for CD24 (nine days after gRNA transduction).

The BRD4 gRNA was validated by performing qRT-PCR on RNA samples from treated cells with primers for either BRD2, BRD3, or BRD4, as described above.

**Imatinib treatments of K562 cells**

Cells were challenged with imatinib either after JQ1 treatment or BRD4 CRISPRi. For the former, we plated 200,000 cells each well of a 6-well plate with 2 mL of growth medium. Half of the wells received DMSO while the other half received 250 nM JQ1. Cells were incubated for 5 days, with fresh media changes on days 1, 2, and 3. On day 5, a portion of each well was stained for CD24. The remaining cells in each well were split to mid-log phase concentration (~400,000 cells/ml) and selected on blastcidin (10 μg/ml) for 48 hours. We made frozen stocks from these cells.

We perturbed the cell cycle with lovastatin and nocodazole, two drugs classically used to synchronize cells in culture (Jackman and O’Connor, 2001), as well as the cyclin-dependent kinase inhibitors CVT-313 (Brooks et al., 1997) and RO-3306 (Vassilev et al., 2006). All drugs were dissolved in DMSO except nocodazole, which was dissolved in ethanol. We treated 200,000 cells per well in 6-well plates with either DMSO, ethanol (~0.4% final concentration), 250 nM JQ1, 12 μM lovastatin, 40 ng/μl nocodazole (in ethanol), 2 μM CVT-313, or 4.5 μM RO-3306. Media was refreshed every 48 hours. After 36 hours of treatment, we stained for CD24 levels and nuclear DNA content. We gated for live, single cells using the forward scatter (FSC) and side scatter channels (SSC). Univariate cell cycle analysis was performed with FlowJo. We confirmed that all drugs perturbed cell cycle by altering the proportions of cells in either G1 or G2/M phase (Figure S6G). CVT-313 caused a significant increase in G1 arrest cells (one-way ANOVA p < 0.05) and both nocodazole and RO-3306 caused significant G2 arrest (one-way ANOVA p < 0.01). Lovastatin has been reported to arrest cells in G1, in our hands it caused a significant decrease in G1 phase K562 cells (one-way ANOVA p < 0.01). Cultures remained under drug treatment until five days had elapsed, at which point we measured CD24 levels and stained for apoptosis (Figure S6H). As before, we set gates to exclude debris to quantitate annexin V and PI, and measured CD24 in live cells gated on FSC and SSC. The G2 inhibitors, in particular, had very few cells in the FSC/SSC gate (typically below 5%).
SRT-tdTomato fluorescence validation

To test the fluorescence properties of the SRT-tdTomato construct, we transfected K562 cells as previously described with either SRT-tdTomato fluorescence validation Resource figures were created with Python using matplotlib (Hunter, 2007). Statistical details for individual experiments have been provided as well as R 3.5.3 using the multcomp package (Hothorn et al., 2008). Flow cytometry figures were created with FlowJo. All other event whose genomic coordinate is along the x axis. The y axis is the number of reads supporting each insertion on a Calling card tracks depict recovered transposons as discrete data points. Each circle in the track is an independent transposition in FACS buffer: 1x HBSS, 2 mM MgCl₂, 2 mM MgSO₄, 1.25 mM CaCl₂, 1 mM D-glucose, 0.02% BSA, and 5% v/v trehalose. Cells cells were resuspended in FACS buffer at a concentration appropriate for 10x Chromium 3. Resequence values exceeded the gate. After sorting, cells were centrifuged for 3 minutes at 250 g. The supernatant was discarded and SRT-tdTomato only. We then collected cells from brains transfected with AAV9-PB-SRT-tdTomato and AAV9-HyPBase whose fluorescence activity was assayed on an Attune NxT Flow Cytometer with an excitation wavelength of 561 nm. Flow cytometry data were visualized using FlowCal (Castillo-Hair et al., 2016). 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 (Methods S1).

In vivo scCC experiments

We separately packaged the PB-SRT-tdTomato and HyPBase constructs in AAV9 viral particles (Cammack et al., 2020) 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 generated both scRNA-seq and scCC libraries.

Mouse cortical tissues were dissociated to single suspensions following a modification of previously published methods (Avey et al., 2018; Saxena et al., 2012). We incubated samples in a papain solution containing Hibernate-A with 5% v/v trehalose, 1x B-27 Supplement, 0.7 mM EDTA, 70 µM 2-mercaptoethanol, and 2.8 mg/ml papain. After incubation at 37°C, cells were treated with DNaseI, 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, 0.5% Bovine Serum Albumin (BSA), 33 µg/ml DNaseI (Worthington), 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 mixed with an equal volume of 1x Hank’s Balanced Salt Solution (HBSS) 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 800g at 12°C. 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 MgCl₂, 2 mM MgSO₄, 1.25 mM CaCl₂, 1 mM D-glucose, 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.

Quantification and Statistical Analysis

Statistical analyses were performed in Python 3.7.3 using Scipy (Virtanen et al., 2020) and statsmodels (Seabold and Perktold, 2010) as well as R 3.5.3 using the multcomp package (Hothorn et al., 2008). Flow cytometry figures were created with FlowJo. All other figures were created with Python using matplotlib (Hunter, 2007). Statistical details for individual experiments have been provided in the main text, figure legends, and Method Details. In general, we used 10-12 replicates for bulk RNA calling cards experiments; and at least three separate libraries for single cell calling cards experiments; and three biological replicates for the K562 cell state experiments.

Interpreting calling card tracks

Calling card tracks depict recovered transposons as discrete data points. Each circle in the track is an independent transposition event whose genomic coordinate is along the x axis. The y axis is the number of reads supporting each insertion on a log10 scale. The total, genome-wide library size is shown at left (n). To better compare transpositions across libraries with different numbers of insertions, we also plotted the normalized local insertion rate as a density track.

Sequencing and analysis: bulk DNA CC libraries

DNA calling card libraries were sequenced 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 3 base-pair barcode TAG followed by the end of the transposon terminal repeat, culminating with the piggyBac insertion site motif TTTA. Reads that had exact matches to these sequences were hard trimmed using cutadapt (Martin, 2011) with the following settings: -g "^TAGTTTAGCAGACTATTTTTAGGTTAA"—minimum-length 1–discard-untrimmed -e 0–no-indels. Reads passing this filter were then trimmed of vector sequence along read 2 using
cutadapt with the following settings: -g "\ATCACTTTAGCCGTTAC"—minimum-length 1—discard-untrimmed -e 0—no-indels. The remaining reads were aligned to the human genome (build hg38) with NovoAlign 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) using custom programs (available at https://github.com/amarn/calling_cards) and visualized on the WashU Epigenome Browser v46 (Zhou et al., 2011) (http://epigenomewidget.wustl.edu/legacy/).

**Sequencing and analysis: bulk RNA CC libraries**

Multiple calling card libraries were pooled together for sequencing on the Illumina HiSeq 2500 platform with 50% phiX. 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 with the following settings: -g "\NNNGGGTCA ATTTTAGCCGACTTCTTCTTAGGGTTAA"—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 "\NNNTAAGTGTATGAAAACTTCCGACTTCAACTGTA"—minimum-length 1—discard-untrimmed -e 0—no-indels. Reads passing this filter were then trimmed of any trailing Nextera adaptor sequence, again using cutadapt and the following settings: -a "CGTCTCCTTATACATCTCCGAGCCCACGAGACTNNNNNNNNNTCTGATGCGGTCTCCTGCTT"—minimum-length 1. The remaining reads were aligned to the human genome (build hg38) with NovoAlign 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) and visualized on the WashU Epigenome Browser v46 (Zhou et al., 2011) (http://epigenomewidget.wustl.edu/legacy/).

**Sequencing and analysis: scRNA-seq libraries**

scRNA-seq libraries were sequenced on either Illumina HiSeq 2500 or NovaSeq machines. Reads were analyzed using 10x Genomics’ Cell Ranger with the following settings:—expect-cells = 6000—chemistry = SC3Pv2—loci = 16—localmem = 30. The digital gene expression matrices from 10x were then further processed with scranpy (Wolf et al., 2018) for identification of highly variable genes, batch correction, dimensionality reduction, and Louvain clustering. Processed scRNA-seq datasets were stored as .loom files (http://loompy.org). We cross-referenced gene expression data with published datasets (Rosenberg et al., 2018; Rouillard et al., 2016; Saunders et al., 2018; Tasic et al., 2018; Zeisel et al., 2018) to assign cell types. The species mixing analysis was performed using Drop-seq tools (Macosko et al., 2015).

**Sequencing and analysis: 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 (Table S4). Read 1 sequenced the cell barcode and unique molecular index of each self-reporting transcript. Read 2 began with GGTAAA (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 adaptor sequence, again using cutadapt and with the following settings: -a "AGAGACTGCGACCAGTGACCATCCTACCAGATGNNNNNNNNNTCTCGATGCCGTCCTGCTT"—minimum-length 1. Reads passing these filters were aligned using 10x Genomics’ cellranger with the following settings:—expect-cells = 6000—nosecondary—chemistry = SC3Pv2—loci = 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). Finally, to minimize the presence of intermolecular artifacts, we required that each insertion must have been tagged by at least two different UMs. 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 individual 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). More details on these steps are also provided in the associated protocols. 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.

**Peak calling on calling card data**

We called peaks in calling card data using Bayesian blocks (Scargle et al., 2013), a noise-tolerant algorithm for segmenting discrete, one-dimensional data, using the astropy implementation (Robitaille et al., 2013; The Astropy Collaboration et al., 2018). 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 $B = \{b_1, b_2, \ldots, b_n\}$ 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 $y_j$ 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 $y_j$: $P(k \geq x_i| \lambda = y_j)$. We accepted all blocks that were significant beyond a particular $p$ value threshold.

For the analysis of TF-directed insertions, either in bulk or in single cells, we added a pseudocount of 1 to $y_j$, the number of insertions in block $b_i$ in the undirected experiment. We selected all blocks whose $p$ values were significant at a Benjamini-Hochberg false discovery rate of 5% (Benjamini and Hochberg, 1995). 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 $y_j$, 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 TTAAs in the genome by the total number of undirected insertions, then multiplying this value by the number of TTAAs in block $b_i$. Then, for each block we calculated the Poisson $p$ value $P(k \geq x_i| \lambda = x_j)$. 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 (Pott and Lieb, 2015; Whyte et al., 2013).

For the bulk PBase and HyPBase analysis, we used $p$ value cutoffs of $10^{-30}$ and $10^{-62}$, respectively. (We chose these stringent thresholds to better resolve super-enhancers, which is our primary focus here.) For both in vitro and in vivo single cell HyPBase analyses, we used a $p$ value cutoff of $10^{-9}$. To identify the differentially-bound loci between CD24high/CD24low K562 cells, as well as between upper and lower cortical layer neurons (i.e., Pou3f2/Brn-2, Bcl11b/Ctip2, and Foxp2), we used the same framework as described above for TF-directed analysis but did reciprocal enrichment analyses, where one dataset was used as the “experiment” track and the other as the “control” track, and vice-versa. This results in two one-sided hypothesis tests. When analyzing differential binding between upper and lower cortical layer neurons, we used a $p$ value cutoff of $10^{-9}$. For the CD24high/CD24low K562 analysis, we restricted our hypothesis testing to BRD4-bound peaks found in the cell line mixing experiment that had at least 20 insertions between both groups. For each peak, we normalized the number of insertions from each population by a library-specific scaling factor and calculated the fold change in binding as $\log_2$.

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 $y_j$, 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 TTAAs in the genome by the total number of undirected insertions, then multiplying this value by the number of TTAAs in block $b_i$. Then, for each block we calculated the Poisson $p$ value $P(k \geq x_i| \lambda = x_j)$. 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 (Pott and Lieb, 2015; Whyte et al., 2013).

For the bulk PBase and HyPBase analysis, we used $p$ value cutoffs of $10^{-30}$ and $10^{-62}$, respectively. (We chose these stringent thresholds to better resolve super-enhancers, which is our primary focus here.) For both in vitro and in vivo single cell HyPBase analyses, we used a $p$ value cutoff of $10^{-9}$. To identify the differentially-bound loci between CD24high/CD24low K562 cells, as well as between upper and lower cortical layer neurons (i.e., Pou3f2/Brn-2, Bcl11b/Ctip2, and Foxp2), we used the same framework as described above for TF-directed analysis but did reciprocal enrichment analyses, where one dataset was used as the “experiment” track and the other as the “control” track, and vice-versa. This results in two one-sided hypothesis tests. When analyzing differential binding between upper and lower cortical layer neurons, we used a $p$ value cutoff of $10^{-9}$. For the CD24high/CD24low K562 analysis, we restricted our hypothesis testing to BRD4-bound peaks found in the cell line mixing experiment that had at least 20 insertions between both groups. For each peak, we normalized the number of insertions from each population by a library-specific scaling factor and calculated the fold change in binding as $\log_2$.

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 and dividing by the length of the block in kilobases (insertions per kilobase per million map-}

Custom code to facilitate these analyses is available online (https://github.com/arnavm/calling_cards). Detailed instructions on how to analyze calling card data are provided in the linked protocols.

**TF binding analysis**

We compared our TF-directed calling card peaks to publicly available ChIP-seq datasets. See below for more details on aligning and analyzing ChIP-seq data. We collated a list of unique transcription start sites (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 and K562 cells and their methylation status were derived from previously-published Methyl-seq data (Brunner et al., 2009). We used the liftOver tool (Hinrichs et al., 2006) to convert coordinates from hg18 to hg38. We tested for enrichment in SP1-directed insertions at TSSs, CpG islands, and unmethylated CpG islands with the G test of independence. We used the same test when testing enrichment of BAP1-directed insertions at TSSs. For motif discovery, we restricted our analysis to peaks less than 5,000 bp in length. We then used MEME-ChIP (Machanick and Bailey, 2011) with a dinucleotide shuffled control and the following settings: -dna -nmeme 600 -seed 0 -ccut 250 -meme-mod zoops -meme-

**BRD4 sensitivity, specificity, and precision**

We used a published BRD4 ChIP-seq dataset (McCleland et al., 2016) to identify BRD4-bound super-enhancers in HCT-116 cells, following previously-described methods (Loven et al., 2013; Whyte et al., 2013). We first called peaks using MACS 1.4.1 (Zhang et al., 2008) at $p < 10^{-9}$ (using the parameters -p 1e−9--keep-dup = "auto" -f BAM -g hs -w -S-space = 50), then fed this into ROSE. 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 (Quinlan and Hall, 2010) 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 downsamped insertions at a range of p value cutoffs. Linear interpolation was performed using NumPy (Oliphant, 2015) and visualized using matplotlib (Hunter, 2007). 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.

**Analysis of external datasets**

For ChIP-seq, ATAC-seq, and DNase-seq data, 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 deepools (Ramirez et al., 2016). 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 as previously described (Ernst et al., 2011). For each chromatin state, we plotted the mean and standard deviation of the rate of normalized insertions (IPKM). We called peaks on SP1 ChIP-seq, DNase- and ATAC-seq data using MACS 2 with the following settings: -q 0.05–keep-dup = "auto." For the analysis of “super-enhancers” from ATAC-seq data, we used control data derived from ATAC-seq on deproteinized human genomic DNA (Martins et al., 2018) and followed the same steps for calling super-enhancers from BRD4 ChIP-seq data (above). If necessary, files were converted to hg38 using liftOver (Hinrichs et al., 2006).

**Cell state analyses of K562: scRNA-seq and scCC**

Cell state analysis was performed on batch-corrected K562 scRNA-seq data derived from the HyPBase cell line mixing experiment. Principal components analysis (PCA) of single cell gene expression (Figure S5A) revealed CD24 as one of the top genes in PC1, while PC2 was enriched in hemoglobin genes, particularly the fetal-specific markers HBE1 and HBZ. Furthermore, the expression of top PC1 and PC2 genes appear to be anticorrelated: cells that strongly expressed CD24 are not likely to express HBZ, and vice-versa (Figure S5B), suggesting mutually exclusive states. We then scored cells based on the expression of VIM, TMSB4X, HBG1, and HBG2, revealing a gradient of cell states along a stem-like-to-differentiated axis (Figure S5A). We then modeled the distribution of this state score as a 3 component Gaussian mixture model, drawing cutoffs where adjacent Gaussian distributions intersected (Figure S5C). These cutoffs were then used to label cells as either stem-seq (CD24high), differentiated (CD24low), or intermediate (Figure S5D). The expression levels of CD24 and HBZ, which were not used to score cells, showed high specificity for the stem-like and differentiated clusters (Figure S5E). Differentially bound peaks were called as described above.

**Analysis of K562 experiments**

We analyzed the JQ1 time course experiment using a two-way ANOVA with treatment and day as the independent variables and the percentage of CD24high cells as the dependent variable. For the analysis of annexin V levels in either JQ1- or DMSO-treated CD24high and CD24low cells, we used a three-way ANOVA with treatment, cell state, and day as independent variables. The imatinib experiments following either JQ1 or BRD4 CRISPRi pretreatment were analyzed using a two-way ANOVA with pretreatment (JQ1/ DMSO or NT/BRD4 gRNA) and treatment as the independent variables. Multiple hypothesis correction was performed using Tukey’s honestly significant difference. For the cell cycle inhibitor experiment, data were analyzed using a one-way ANOVA with Dunnett’s post hoc test using either DMSO or EtOH (for RO-3306) as controls.

**In vivo scCC 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 (Rosenberg et al., 2018; Saunders et al., 2018; Tasic et al., 2018; Zeisel et al., 2018). BRD4-bound peak calls were validated by comparing to a previously published cortical H3K27ac ChIP-seq dataset (Stroud et al., 2017). 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 (Gasperini et al., 2019). 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 (Zhang et al., 2014) 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, we calculated the specificity as...
AstrocyteFPKM(g) / [AstrocyteFPKM(g) + NeuronFPKM(g)]. 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. After accounting for differences in library size, we identified 383 genes near astrocyte peaks and 184 genes near neuron peaks, with 46 genes found in both datasets. We plotted the distributions of gene expression specificity for these gene sets. (Figure 4E). Gene Ontology analysis was performed on the same sets of genes using PANTHER (Mi et al., 2017) on the “GO biological process complete” database. Fisher’s exact test was used to compute p values, which were then subject to Bonferroni correction.

ADDITIONAL RESOURCES

We have created a number of protocols describing how to perform all aspects of bulk and single cell calling cards, from molecular biology and sequencing through data analysis and visualization. While these are listed in the Key Resources Table, we have also created a publicly accessible portal for easy access to all our workflows: https://www.protocols.io/groups/calling-cards/. Moving forward, this resource should contain the most up-to-date information.
Figure S1. BRD4 Calling Cards with Undirected piggyBac Are Not Equivalent to ATAC-Seq, Related to Figure 2

(A) Overlap of BRD4 super-enhancers, as inferred from BRD4 ChIP-seq, and “super-enhancers” inferred from open chromatin ATAC-seq peaks in HCT-116 cells.

(B) Browser view of a BRD4 calling card peak that is not detected by ATAC-seq nor DNase-seq. (C) Comparison of transposase predilections for accessible chromatin. (D) Comparison of peak sizes and BRD4 ChIP-seq enrichment as called by DNase-seq, ATAC-seq, and undirected piggyBac calling cards, respectively. Peaks are scaled to the median peak width (denoted by the start and end ticks) and are flanked by 3 kb in either direction. SE: super-enhancer; DHS: DNase hypersensitivity site; FC: fold change; kb: kilobase.
Figure S2. Filtering Single-Cell SRTs Reduce Intermolecular Artifacts, Related to Figure 3

(A) Barnyard plot from scRNA-seq of human HCT-116 and mouse N2a cells. (B) Barnyard plot from scCC of HCT-116 and N2a cells without filtering (estimated multiplet rate of 25.1%). (C) Distribution of cell barcode purity from unfiltered scCC data. The x axis is the proportion of transcripts mapping to the human or mouse genomes. (D) Distribution of species purity after filtering scCC data. UMI: unique molecular indexes.
Figure S3. Validation and Performance of Undirected In Vitro scCC, Related to Figure 3

(A) Single cell expression levels of three marker genes in a mixed scRNA-seq library of human HCT-116 and K562 cells. (B) Distributions of genes per cell by cell type. (C) Distributions of transcripts per cell by cell type. (D) Distributions of HyPBase insertions recovered per cell in HCT-116 and K562 cells. (E-F) Mean BRD4 ChIP-seq signal at HyPBase peaks in HCT-116 and K562 cells, respectively, compared to randomly permuted peaks (KS test p < 10^{-9} in each case). (G-H) Reproducibility of normalized insertions deposited by HyPBase and recovered by scCC at BRD4 binding sites in HCT-116 and K562 cells, respectively. KS: Kolmogorov-Smirnov.
Figure S4. Validation and Performance of TF-Directed In Vitro scCC, Related to Figure 4

(A-B) Enrichment of SP1-HyPBase-directed insertions to TSSs, CGIs, and unmethylated CGIs in single HCT-116 and K562 cells, respectively (G test of independence p < 10^{-9}). (C) Enrichment of BAP1-HyPBase-directed insertions TSSs in single OCM-1A cells (G test of independence p < 10^{-9}). (D) Percent of BAP1 targets that increase expression upon BAP1 KD stratified by binding site (Fisher’s exact test p < 10^{-9}). The dashed gray line represents the overall fraction of genes that increased expression upon KD. (E-H) Reproducibility of normalized insertions deposited by either HyPBase or TF-HyPBase fusions and recovered by scCC at TF binding sites, for the respective TF-cell line pair. (I-L) The distribution of recovered insertions per cell by construct (HyPBase versus TF-HyPBase) and cell type. TF: transcription factor; TSS: transcription start site; CGI: CpG island; KD: knockdown; IPM: insertions per million mapped insertions; n.s.: not significant.
Figure S5. Clustering of K562 Cells into Stem-like and Differentiated States, Related to Figure 5
(A) Principal component analysis of K562 scRNA-seq data. (B) Relative expression levels of highest-ranking genes in PC1 (top) and PC2 (bottom). (C) Gaussian mixture modeling of a cell-state score to define stem-like and differentiated K562 clusters. (D) Visualization of assigned cell clusters in the UMAP projection. (E) Specific expression of \textit{CD24} and \textit{HBZ} in the stem-like and differentiated clusters, respectively. (F) Genome browser view of scGC in the stem-like and differentiated clusters alongside bulk BRD4 and H3K27ac ChIP-seq as well as RNA Pol II ChIA-PET. (G) Expression of \textit{VMP1} and \textit{PVT1} in the stem-like and differentiated clusters. PC: principal component.
Figure S6. Validation of Bromodomain-Dependent K562 Cell States, Related to Figure 5

(A) Annexin V staining in CD24<sup>high</sup> (red) or CD24<sup>low</sup> (blue) K562 cells treated with DMSO (dashed line) or JQ1 (solid line) over a seven-day time course. (B) Genome browser view of the CD24 locus. (C) qRT-PCR for MYC and CD24 expression levels in bulk K562 cells treated with JQ1 relative to DMSO-treated controls. (D) Expression changes in BRD2, BRD3, and BRD4 in K562 cells transduced with dCas9-KRAB and BRD4 CRISPRi gRNA (Welch’s t test p < 0.05). (E) Annexin V and PI co-staining in cells subjected to either non-targeting (top) or BRD4 (bottom) CRISPRi followed by either DMSO (left) or imatinib (right) treatment. (F) Average percent of annexin V/PI double positive cells in either the non-targeted or BRD4 CRISPRi replicates, stratified by either DMSO or imatinib exposure (two-way ANOVA p < 0.01). (G) Percent of K562 cells in either G1 (left) or G2 (right) phase after 36 hours of drug treatment (one-way ANOVA with Dunnett’s test *p < 0.05, **p < 0.01). (H) Percent of K562 cells in the CD24<sup>high</sup> state (left) after 5 days, and the percent of annexin V/PI double positive cells (right) at the same time point (p < 0.01 in each instance, one-way ANOVA with Dunnett’s test). Bars/points represent means; error bars denote standard deviations. Experiments were performed in triplicate. DMSO: dimethyl sulfoxide; n.s.: not significant; FC: fold change; SSC: side scatter; CRISPRi: CRISPR interference; NT: non-targeting; gRNA: guide RNA; IMA: imatinib; PI: propidium iodide.
Figure S7. Validation of In Vivo BRD4 Binding in Astrocytes and Neurons, Related to Figure 6

(A) Mean H3K27ac ChIP-seq signal at HyPBase peaks in astrocytes compared to randomly permuted peaks (KS test $p < 10^{-9}$). (B) GO term enrichment analysis of genes near astrocytic BRD4 binding sites. (C) Mean H3K27ac ChIP-seq signal at HyPBase peaks in neurons compared to randomly permuted peaks (KS test $p < 10^{-9}$). (D) GO term enrichment analysis of genes near neuronal BRD4 binding sites. (B and D) The white line indicates the Bonferroni-adjusted p value threshold at $\alpha = 0.05$. GO: Gene Ontology; KS: Kolmogorov-Smirnov; FC: fold change.