

Washington University in St. Louis

Washington University Open Scholarship

Arts & Sciences Electronic Theses and
Dissertations

Arts & Sciences

Winter 12-15-2021

Deconvolving Genomic Regulatory Heterogeneity with Self-Reporting Transposons

Arnav Moudgil

Washington University in St. Louis

Follow this and additional works at: https://openscholarship.wustl.edu/art_sci_etds



Part of the [Genetics Commons](#)

Recommended Citation

Moudgil, Arnav, "Deconvolving Genomic Regulatory Heterogeneity with Self-Reporting Transposons" (2021). *Arts & Sciences Electronic Theses and Dissertations*. 2617.
https://openscholarship.wustl.edu/art_sci_etds/2617

This Dissertation is brought to you for free and open access by the Arts & Sciences at Washington University Open Scholarship. It has been accepted for inclusion in Arts & Sciences Electronic Theses and Dissertations by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Computational and Systems Biology

Dissertation Examination Committee:

Robi D. Mitra, Chair
Donald F. Conrad
Joseph D. Dougherty
Benjamin D. Humphreys
Samantha A. Morris

Deconvolving Genomic Regulatory Heterogeneity with Self-Reporting Transposons
by
Arnav Moudgil

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2022
St. Louis, Missouri

© 2022, Arnav Moudgil

Table of Contents

List of Figures	viii
List of Tables	xi
Acknowledgments.....	xii
Abstract.....	xviii
Chapter 1: Introduction.....	1
1.1 Background, Significance, and Scope	1
1.2 References.....	14
Chapter 2: Self-reporting transposons enable simultaneous readout of gene expression and transcription factor binding in single cells.....	35
2.1 Abstract.....	35
2.2 Introduction.....	35
2.3 Results.....	39
2.3.1 SRTs can be mapped from mRNA instead of genomic DNA.....	39
2.3.2 SP1 fused to <i>piggyBac</i> directs SRT insertions to SP1 binding sites	45
2.3.3 Clustering of undirected <i>piggyBac</i> insertions identifies BRD4-bound super-enhancers	48
2.3.4 scCC simultaneously identifies cell type and cell type-specific BRD4 binding sites.....	58
2.3.5 scCC identifies binding sites across a spectrum of TFs and in a variety of cell types.....	63
2.3.6 scCC reveals bromodomain-dependent cell state dynamics in K562 cells	67
2.3.7 scCC deconvolves cell type-specific BRD4 binding sites in the mouse cortex	75
2.4 Discussion.....	82
2.5 Methods.....	87
2.5.1 Materials and data availability.....	89
2.5.2 Experimental model and subject details.....	94
2.5.3 DNA- vs RNA-based recovery.....	95
2.5.4 <i>In vitro</i> bulk calling card experiments.....	96
2.5.5 Isolation and RT of bulk RNA	97
2.5.6 Amplifying self-reporting transcripts from RNA.....	97
2.5.7 Generation of bulk RNA calling card libraries.....	98
2.5.8 <i>In vitro</i> single cell calling card experiments.....	99
2.5.9 Single cell RNA-seq library preparation	100
2.5.10 Single cell calling cards library preparation	101
2.5.11 Staining protocols for K562 cells.....	103
2.5.12 JQ1 treatment of K562 cells.....	104

2.5.13	BRD4 CRISPRi of K562 cells	104
2.5.14	Imatinib treatments of K562 cells	106
2.5.15	Cell cycle perturbation of K562 cells	106
2.5.16	SRT-tdTomato fluorescence validation	107
2.5.17	In vivo scCC experiments	107
2.5.18	Quantification and statistical analysis	109
2.5.19	Sequencing and analysis: bulk DNA CC libraries	109
2.5.20	Sequencing and analysis: bulk RNA CC libraries	110
2.5.21	Sequencing and analysis: scRNA-seq libraries	111
2.5.22	Sequencing and analysis: scCC libraries	111
2.5.23	Peak calling on calling card data	112
2.5.24	TF binding analysis	115
2.5.25	BRD4 sensitivity, specificity, and precision	115
2.5.26	Downsampling and replication analysis	116
2.5.27	Analysis of external datasets	116
2.5.28	Cell state analyses of K562: scRNA-seq and scCC	118
2.5.29	Analysis of K562 experiments	118
2.5.30	<i>In vivo</i> scCC analysis and validation	119
2.5.31	Additional resources	120
2.6	References	120
Chapter 3: The qBED track: a novel genome browser visualization for point processes		144
3.1	Abstract	144
3.1.1	Summary	144
3.1.2	Availability and Implementation	144
3.2	Introduction	144
3.3	Implementation	146
3.4	Applications	149
3.5	Conclusion	152
3.6	References	153
Chapter 4: Fast and optimal genome segmentation with Bayesian blocks		158
4.1	Abstract	158
4.2	Introduction	158
4.3	Results	161
4.3.1	Review of Bayesian blocks	161
4.3.2	Calling peaks using Bayesian blocks	166

4.3.2	Identifying CpG islands using Bayesian blocks.....	171
4.4	Discussion.....	174
4.5	Methods.....	177
4.6	Proofs.....	178
4.6.1	Properties of Poisson point processes.....	178
4.6.2	Derivation of the Bayesian blocks likelihood function.....	179
4.6.3	Adapting Bayesian blocks to PELT.....	181
4.7	References.....	182
Chapter 5: Self-reporting transposons reveal chromosomal compartmentalization.....		190
5.1	Introduction.....	190
5.2	Results.....	191
5.2.1	<i>Sleeping Beauty</i> SRTs are not uniformly distributed across the genome.....	191
5.2.2	Densities of <i>Sleeping Beauty</i> SRTs reveal chromosomal compartments.....	194
5.2.3	Inferring compartmentalization from sparse <i>Sleeping Beauty</i> data.....	199
5.3	Discussion.....	200
5.4	Methods.....	202
5.5	References.....	203
Chapter 6: Future directions.....		212
6.1	Future directions for <i>piggyBac</i> transposon calling cards.....	212
6.2	Future directions for <i>Sleeping Beauty</i> transposon calling cards.....	218
6.3	Industrial applications for SRTs.....	219
6.4	Tagmentation-free SRTs.....	220
6.5	Expanding the palette of self-reporting transposons.....	223
6.5.1	Mos1.....	223
6.5.2	Tol2.....	224
6.5.3	LINE-1.....	225
6.5.4	<i>Helraiser</i>	227
6.5.5	Final thoughts.....	229
6.6	The ecology of chromatin.....	229
6.7	References.....	230
Appendix 1: Mammalian Calling Cards Quick Start Guide.....		241
A1.1	Abstract.....	241
A1.2	Guidelines.....	241
A1.3	Materials.....	241
A1.4	Steps.....	243

A.1.4.1	Cloning and Sequence Validation of YFTF-HyPBase Fusions	243
A1.4.2	Functional Validation of YFTF-HyPBase Fusions	243
A1.4.3	Calling Card Library Preparation.....	245
A1.4.4	Sequencing, Analyzing, and Visualizing Calling Card Data	245
A1.4.5	Next Steps	246
Appendix 2:	Bulk Calling Cards Library Preparation	247
A2.1	Abstract	247
A2.2	Guidelines	247
A2.3	Materials	248
A2.4	Steps.....	250
A2.4.1	RNA Extraction with QIAGEN's RNEasy Plus Mini Kit.....	250
A2.4.2	cDNA Synthesis	252
A2.4.3	Amplification of Self-Reporting Transcripts	253
A2.4.4	Purification of PCR Products	255
A2.4.5	Generation of Bulk Calling Card Libraries.....	256
A2.4.6	Final Quantitation and Sequencing	258
Appendix 3:	Single Cell Calling Cards Library Preparation	260
A3.1	Abstract	260
A3.2	Guidelines	260
A3.3	Materials	261
A3.4	Steps.....	264
A3.4.1	Single Cell Barcoding and Reverse Transcription	264
A3.4.2	Single Cell RNA-seq Library Preparation and Sequencing.....	265
A3.4.3	Amplification of Self-Reporting Transcripts	266
A3.4.4	Purification of PCR Products	268
A3.4.5	Single Cell Calling Cards – Circularization.....	269
A3.4.6	Single Cell Calling Cards – Exonuclease and Setup.....	270
A3.4.7	Single Cell Calling Cards – Shearing and Capture	273
A3.4.8	Single Cell Calling Cards – End Repair, A-Tailing, and Adapter Ligation.....	274
A3.4.9	Single Cell Calling Cards – Final PCR and Purification	277
A3.4.10	Single Cell Calling Cards – Sequencing	280
Appendix 4:	Processing Bulk Calling Card Sequencing Data.....	282
A4.1	Abstract	282
A4.2	Guidelines	282
A4.3	Materials	283

A4.4	Steps.....	284
A4.4.1	Preamble.....	284
A4.4.2	Adapter Trimming.....	286
A4.4.3	Alignment.....	288
A4.4.4	Annotation.....	289
A4.4.5	Finishing Up.....	290
A4.4.6	Notes	291
Appendix 5:	Processing Single Cell Calling Card Sequencing Data.....	294
A5.1	Abstract.....	294
A5.2	Guidelines	294
A5.3	Materials	295
A5.4	Steps.....	296
A5.4.1	Preamble.....	296
A5.4.2	Adapter Trimming.....	300
A5.4.3	Alignment.....	302
A5.4.4	Annotation.....	304
A5.4.5	Demultiplexing.....	306
A5.4.6	Notes	311
Appendix 6:	Calling Peaks on <i>piggyBac</i> Calling Card Data	313
A6.1	Abstract.....	313
A6.2	Guidelines	313
A6.3	Materials	314
A6.4	Steps.....	315
A6.4.1	Preprocessing	315
A6.4.2	Bayesian Blocks.....	317
A6.4.3	Calling BRD4 Peaks	318
A6.4.4	Calling TF Peaks.....	323
A6.4.5	Final Thoughts	329
Appendix 7:	Visualizing Calling Card Data on the WashU Epigenome Browser	331
A7.1	Abstract.....	331
A7.2	Guidelines	331
A7.3	Materials	331
A7.4	Steps.....	331
A7.4.1	Introduction.....	331
A7.4.2	Uploading Data from an External Server.....	333

A7.4.3	Uploading Local Data Files	342
A7.4.4	Interacting with the Calling Card Track	344
Appendix 8:	Online Resources	347
A8.1	Links to Online Resources	347

List of Figures

Figure 1.1:	The landscape of multimodal scRNA-seq technologies	7
Figure 2.1:	Self-reporting transposons (SRTs) are mapped more efficiently from RNA compared to DNA.....	40
Figure 2.2:	Properties of self-reporting transposons (SRTs).....	42
Figure 2.3:	<i>piggyBac</i> , SP1- <i>piggyBac</i> fusions, and <i>Sleeping Beauty</i> display different local transposition rates depending on chromatin state	44
Figure 2.4:	SP1 fused to <i>piggyBac</i> (SP1-PBase) redirects SRTs to SP1 binding sites	47
Figure 2.5:	SP1 fused to hyperactive <i>piggyBac</i> (SP1-HyPBase) also redirects SRTs to SP1 binding sites.....	48
Figure 2.6:	Undirected <i>piggyBac</i> (PBase) SRTs mark BRD4-bound super-enhancers (SEs).....	50
Figure 2.7:	Undirected hyperactive <i>piggyBac</i> (HyPBase) SRTs also mark BRD4-bound super-enhancers (SEs).....	51
Figure 2.8:	Downsampling undirected and directed <i>piggyBac</i> insertions simulates assay performance.....	53
Figure 2.9:	Redirectability of SP1- <i>piggyBac</i> fusion constructs	54
Figure 2.10:	Examples of BRD4-bound super-enhancers identified by bulk PBase and HyPBase calling cards in HCT-116 cells.....	55
Figure 2.11:	<i>piggyBac</i> is more tolerant of transcription factor fusions than <i>Sleeping Beauty</i> ...	56
Figure 2.12:	BRD4 calling cards with undirected <i>piggyBac</i> is not equivalent to ATAC-seq....	58
Figure 2.13:	Single cell calling cards (scCC) maps BRD4 binding in single cells	59
Figure 2.14:	Filtering single cell SRTs reduces intermolecular artifacts	61
Figure 2.15:	Validation and performance of undirected <i>in vitro</i> single cell calling cards (scCC).....	63
Figure 2.16:	Single cell calling cards (scCC) works with a variety of transcription factors (TFs) and cell lines	64
Figure 2.17:	Validation and performance of TF-directed <i>in vitro</i> single cell calling cards (scCC).....	65
Figure 2.18:	Clustering of K562 cells into stem-like and differentiated states	68
Figure 2.19:	Single cell calling cards uncovers bromodomain-dependent cell state dynamics in K562 cells	69
Figure 2.20:	Validation of bromodomain-dependent K562 cell states.....	71
Figure 2.21:	Single cell calling cards (scCC) deconvolves BRD4-bound loci in the mouse cortex	76
Figure 2.22:	Clustering of SRT-treated cortical cells and associated marker genes	78
Figure 2.23:	Validation of <i>in vivo</i> BRD4 binding in astrocytes and neurons.....	80
Figure 2.24:	Single cell calling cards (scCC) deconvolves BRD4 binding in	

	cortical excitatory neurons and identifies known layer markers	82
Figure 3.1:	Overview of the qBED format and qBED tracks.....	148
Figure 3.2:	Application of the qBED specification to other genomic datasets	150
Figure 4.1:	Example calling cards tracks	160
Figure 4.2:	Overview of Bayesian blocks	162
Figure 4.3:	Effect of varying p_0 on segmentation.....	165
Figure 4.4:	Calling peaks on calling cards data with Bayesian blocks	166
Figure 4.5:	Additional benchmarking of peak calling.....	168
Figure 4.6:	TF-directed block sizes are a function of relative enrichment.....	170
Figure 4.7:	Detecting CpG islands with Bayesian blocks	171
Figure 4.8:	Length distributions of CpG islands	174
Figure 5.1:	SB insertion densities are correlated with chromosomal compartment.....	193
Figure 5.2:	Comparison of SB insertion densities by compartment.....	195
Figure 5.3:	Functional validation of SB compartments.....	197
Figure 5.4:	SB compartment analysis of a nucleolus organizing region (NOR).....	198
Figure 5.5:	Downsampling analysis of SB compartments	200
Figure 5.6:	Model of transposase activity as function of sub-nuclear localization	201
Figure 6.1:	Predicted disorder for the hyperactive <i>piggyBac</i> transposase.....	215
Figure 6.2:	Alignment of MMLV's BET interaction motif against <i>piggyBac</i>	218
Figure 6.3:	Alternative SRT library strategies without tagmentation	221
Figure 6.4:	Proposed workflow for mapping L1 SRTs	226
Figure 6.5:	Lineage tracing with self-reporting transposons.....	229
Figure A2.1:	Representative products of SRT amplification	255
Figure A2.2:	Representative TapeStation trace of bulk calling card libraries	259
Figure A3.1:	Representative TapeStation trace of SRT amplification from 10x 3' scRNA-seq library	269
Figure A3.2:	Representative TapeStation trace of scCC libraries.....	280
Figure A5.1:	Example of a 10x Chromium chip loaded for a single cell calling cards experiment.....	298
Figure A5.2:	Demultiplexing scCC libraries.....	307
Figure A6.1:	Example of a BRD4-directed calling cards peak.....	323
Figure A6.2:	Example of an SP1-directed calling cards peak.....	329
Figure A7.1:	WashU Epigenome Browser splash page	339
Figure A7.2:	Custom tracks pane	340
Figure A7.3:	Load custom data hub	340
Figure A7.4:	Successfully uploaded data hub	341
Figure A7.5:	Default view after uploading data hub.....	341
Figure A7.6:	Text track pane.....	342
Figure A7.7:	Uploading a calling cards text file	343
Figure A7.8:	Default view after uploading text file	343

Figure A7.9: Hovering over a calling cards insertion	344
Figure A7.10: Calling cards customization pane	345
Figure A7.11: Various customizations applied to calling card tracks.....	346

List of Tables

Table 2.1: Summary of bulk calling cards experiments	45
Table 2.2: Summary of single cell calling cards experiments	61
Table 2.3: Breakdown of cortical cell types and scCC HyPBase insertions per cluster	79
Table 2.4: Oligonucleotides referenced in this work.....	87
Table 2.5: Key Resources Table.....	89
Table 2.6: ChromHMM chromatin state annotations in HCT-116 cells	117

Acknowledgments

A PhD may have a clear start and end on a calendar, but it is one part of a larger trajectory of growth and learning. To properly acknowledge everyone who has helped me on this journey, I look both before and after in time.

My parents recognized, at an early age, my innate curiosity. I am told, for I have no recollection of these events, that I asked a lot of questions as a child about how things work and why the world is the way it is. I am ever grateful that they fostered this personality trait, encouraging my love for mathematics and science as well as supporting me in my pursuit of a PhD. Likewise, I am thankful for my siblings, Pranav and Divya, who inspire me to stay grounded and who are, perhaps, even more excited than I am about finishing this degree.

As an undergraduate, I had a superlative mentor in Jill Helms. She recognized my potential for a career in science years before I did. Whereas I went through periods of doubt, her convictions were steadfast. I can only hope to inspire the next generation of scientists to the same extent that she inspired me.

During these years, I was fortunate to have had the opportunity to work in two complementary laboratories. Marcus Feldman and Michael Palmer gave me my first shot at “real” research as a late blooming senior. During these years I first solidified my computational skills and pursued very theory-driven work. This experience has turned out to be more formative than I realized at the time, and I am grateful they took me under their wings.

Shortly after graduating, Peter Parham offered me a position in his laboratory. Those two years I spent as a research technician were invaluable for establishing my current research trajectory. Coming from a computational background, I picked up a solid foundation of wet lab

skills that would pay dividends during my PhD. It was this experience that illustrated the value of being equally adept at molecular and informatic methods. While there was a lot of hard work, there was also a lot of joy. For this, I can only credit my amazing colleagues. In particular, Libby Guethlein and Paul Norman were fantastic mentors and I was as eager to learn from them as (I hope) they were eager to teach me. Hugo Hilton, Amir Horowitz, Ana Goyos, Emily Wroblewski, and Jeroen Blokhuis were all postdocs at the time and created a welcoming, positive, and playful environment for a young scientist like me. To all of them I say, simply, thank you.

My support network during this time was a crew of close friends from college, collectively referred to as Kipling because that was the street on which they were renting a house. Douglas Stanford, Allison Dedrick, Adam Hepworth, Paul Schaffert, and JD Porter were Heather's and my closest friends during this time. At a time when we could have been obsessed with our careers, worried about whether we were "doing the right thing," they helped us celebrate the simple joy of good friendship. Thank you for being there for us.

I would also include Anish Mitra in this group, who bridges the pre- and post-MD/PhD arcs. Anish and I have been friends since college, and I owe him three key votes of thanks. He was the first person to encourage me to pursue MD/PhD training. What's more, he helped sell us on coming to St. Louis for this opportunity. And finally, he has been consistently supportive and encouraging throughout this journey. Even though he and his fiancée Rachel have moved away for residency, we still find time to play board games together over the Internet. Thank you for having been there for me all this time.

I have learned much from my cohort of fellow MD/PhD students, who find ways to continually surprise, inspire, and rejuvenate me. It has been an honor to have organized our

weekly journal clubs, which were cut short not by attrition but by a global pandemic. So, to Moises Arriaga, Willie Zhang, Gregg Fox, Olga Neyman, Jared Goodman, Kayla Berry, Jerry Fong, Lindsey Steinberg, Wilbur Song, Deng Pan, Matthew Matlock, Jennifer Stevens, Bernie Mulvey, Umber Dube, and Drew Sinha: thank you for the past seven years. Heather and I have grown especially close to Umber, his wife Amanda, Drew, and his wife Marilyn. Here's to more nights spent making pizza and playing Overcooked!

To the (as of this writing) 635 people following me on Twitter, as well as the countless people I have met at conferences, who have listened to me give a talk or stopped by my poster: thank you! A PhD is difficult, not necessarily due to physical labor, but because it taxes you mentally and emotionally. There have been difficult moments over the past few years and it has felt isolating at times. I am grateful for all your kind words, encouragement, and excitement. You inspire me to pay the warmth forward.

The chapters presented in this thesis give the illusion of a solo body of work. In reality, modern science is a team effort. While my coauthors are credited in print and online, I would like to thank them for here as well. Chapter 2 would not have been possible without help from Michael Wilkinson, Xuhua Chen, June He, Alexander Cammack, Michael Vasek, Tomás Lagunas, Jr., Zongtai Qi, Matthew Lalli, Chuner Guo, Samantha Morris, and Joseph Dougherty. Chapter 3 came together thanks to the concerted efforts Daofeng Li, Silas Hsu, Deepak Purushotham, and Ting Wang. Chapter 4 would not exist but for the generous help provided by Jeffrey Scargle and Rebecca Killick. Finally, I am indebted to Zoltán Ivics, Ivana Grabundzija, Fred Dyda, and Alex Holehouse for their contributions to Chapter 6. My thesis committee, which includes Sam and Joe as well as Don Conrad and Ben Humphreys, has been consistently

supportive. They all wrote me letters of reference for my NIH fellowship application and helped sharpen my writing. Thank you for guiding me along the path.

It takes a lab to raise a graduate student, and while the Mitra Lab has experienced turnover during my five years, one constant has been their collective push for rigor. One of my aims in graduate school was to build a solid foundation in scientific reasoning and methodology, and if I have succeeded it is thanks to my fellow lab members. Zongtai Qi taught me to be a better bench scientist. Justin Melendez taught me to think laterally about molecular biology. Christian Shively taught me to think deeply about my biological system. Jiayue Liu taught me to blend biology with computation. Xuhua Chen taught me to be consistent. Matthew Lalli taught me to carefully dissect biological arguments. Michael Wilkinson taught me to think big. Alex Cammack, an honorary lab member, taught me to relax. To all the rotation students I mentored, you have taught me to be a better teacher and leader. To all the new members who have joined recently, you have taught me to be optimistic for the future.

I am extremely fortunate and grateful to have had Rob Mitra as my mentor for graduate school. One reason why this partnership has been particularly successful is that Rob and I naturally think along similar lines. Thus, we have been on similar wavelengths for much of this journey. Beyond that, Rob is generous. For the vast majority of my PhD, we have had a standing, hour-long meeting every Thursday morning. What started as a way for him to make sure I was making progress gradually evolved into a more spontaneous conversation where I could take the lead. These meetings were effective precisely because Rob is open to talking about anything. In addition to the usual project-specific discussions, we have had freewheeling conversations ranging from the philosophical nature of science, the career challenges facing academics, fundamental problems in biology and mathematics, and practical advice for leadership. This last

point cannot be stressed enough: one of the main ways I have grown during my PhD is as a project manager, which Rob has enthusiastically supported. Finally, Rob is an eternal optimist and great motivator. I can come into his office prepared to wallow about something and I inevitably leave feeling better than when I walked in. This kind of buoyancy is key part of Rob's mentorship style. Rob, thank you for a transformative five years. I am excited to see what the next five have in store for both of us.

To close, I look to the new family that has nucleated around me in the past few years. My in-laws, Larry and Pita Benz, and their extended families, have been very positive during my PhD. I know they are incredibly proud and excited for me, and I am thrilled to have their support. Heather and I have been together for almost ten years now, married for almost three. When we first started dating, she had unwavering confidence in my decision to pursue MD/PhD training. To this day, her belief in me has not faltered. She reminds me to keep things in perspective, to take breaks as needed, and to have more fun outside. Neither of us expected to spend the past few months in self-imposed house arrest during a global pandemic while the world literally, and figuratively, burns to the ground. But I would not choose to spend them with anyone else.

During this time, we brought our daughter Ruby into the world. She is a shining light of unbridled potential and a reminder of what is most important in life. Ruby, you do not realize it yet, but your charming little smile is an instant cure for a stressful day. It is a privilege and joy to see you grow. Thank you.

Arnav Moudgil

Washington University in St. Louis

May 2022

Dedicated to Ruby, whose toothless smile sparks joy

ABSTRACT OF THE DISSERTATION

Deconvolving Genomic Regulatory Heterogeneity with Self-Reporting Transposons

by

Arnav Moudgil

Doctor of Philosophy in Biology and Biomedical Sciences

Computational and Systems Biology

Washington University in St. Louis, 2022

Professor Robi David Mitra, Chair

A cell's identity is a function of the genes expressed in that cell, which are in turn regulated by transcription factors. Over the last decade, single-cell RNA sequencing (RNA-seq) has emerged as a powerful class of techniques to characterize cellular diversity in heterogeneous tissues. These methods barcode transcripts by their cell-of-origin and assign them to specific genes. The resulting high-dimensional data are further processed to reveal clusters of cells sharing transcriptional states. Annotating these clusters, based on either known or discovered marker genes, offers a glimpse into the dynamic composition of an organ or biological process.

While single-cell RNA-seq excels at describing cell states, it alone does not inform us about the mechanisms maintaining a particular state. In recent years, multi-modal single cell technologies have flourished, combining single cell RNA-seq with at least one other genomic modality. As a result, joint assays now exist for assaying gene expression simultaneously with genotype, with methylation, with chromatin accessibility, or with lineage. Collectively, these methods aim to connect gene expression to regulatory processes in the genome, thereby gaining insight into the molecular foundations underpinning cellular identity.

Transcription factors are key protein regulators of gene expression. Master transcription factors organize gene regulatory networks to promote differentiation or homeostasis and are often used as markers of cell type. Unfortunately, no methods exist to measure single-cell RNA-seq and map transcription factor binding in those same cells. Such a technique would be uniquely poised to identify both the identity of a cell and candidate regulatory elements contributing to that identity. The Mitra Lab has developed transposon calling cards as an alternative assay to map transcription factor binding, using transcription factor-transposase fusions to mark binding sites with deposited transposon sequences. Here, I present a single cell extension of this technique using a novel construct, the self-reporting transposon, whose genomic location can be mapped from single-cell RNA-seq libraries. Thus, in one workflow, single cell calling cards identifies cell types in complex systems and deconvolves cell-type-specific regulatory elements bound by a transcription factor in those cell types.

The remainder of this dissertation is organized as follows. Chapter 1 reviews the biological and technological context for this work, with particular focus on single-cell RNA-seq techniques and methods to assay transcription factor binding sites. Chapter 2 presents the central advancement of this dissertation, the self-reporting transposon and its use in single cell calling cards to map cell-type-specific transcription factor binding sites in complex systems. Chapter 3 discusses the qBED track, a medium for visualizing calling cards data, and its accompanying data format for storing results. Chapter 4 examines the Bayesian blocks algorithm, a method adopted from the astrophysics community, and employs it to call peaks in calling cards data. Chapter 5 explores a new use for self-reporting transposons as surveyors of chromosomal compartmentalization. Chapter 6 concludes this dissertation, offering suggestions for future work and positing a broader role for self-reporting transposons in genomics.

Chapter 1: Introduction

1.1 Background, Significance, and Scope

Multicellular life is characterized by heterogenous populations of cells woven together to form tissues and complex organs. How such diversity arises is a fundamental and cross-disciplinary question in biology (Trapnell, 2015). Cataloging the diversity of cell types in an organism has inspired several cellular atlas projects (Han et al., 2018b, 2018b; Karaïskos et al., 2017; Regev et al., 2017) but these efforts are complicated by variability in cell state. For example, individual skeletal muscle cells may appear uniform under a microscope but can have diverse expression patterns of myosin isoforms (Biressi et al., 2007). Cell types traditionally thought to be relatively homogeneous, such as neutrophils (Silvestre-Roig et al., 2016) and tissue-specific stem cells (Goodell et al., 2015; Krieger and Simons, 2015), are proving to be much more functionally diverse than previously appreciated. Such heterogeneity is not simply restricted to normal development. Cellular diversity is also a hallmark of pathologies like cancer (Almendro et al., 2013). A tumor cell population can have widely varying morphologies, gene expression profiles, and chemotherapeutic resistance potentials (Brooks et al., 2015; Knoechel et al., 2014; Litzemberger et al., 2017; Michor and Polyak, 2010). This can present practical challenges for diagnostics, therapy, and prognosis. Characterizing cell type diversity, while a necessary first step, is not sufficient to mechanistically understand how cell identity is generated during development, maintained during homeostasis, and ultimately dysregulated during disease.

Single-cell RNA sequencing (scRNA-seq) is revealing cellular heterogeneity at unprecedented resolution. Investigators have used these methods to discover new subpopulations of cells in the immune compartment (Jaitin et al., 2014; Shalek et al., 2014), brain (Zeisel et al.,

2015), lungs (Plasschaert et al., 2018; Treutlein et al., 2014), intestine (Grün et al., 2015), and retina (Macosko et al., 2015); to profile rare circulating tumor cells in cancer patients (Cann et al., 2012; Miyamoto et al., 2015; Ramsköld et al., 2012); and uncover patterns of transcription during embryogenesis (Xue et al., 2013) as well as facilitate preimplantation diagnosis (Yan et al., 2013). In one particularly memorable example, researchers discovered a previously uncharacterized subset of planarian neoblast that, when transplanted into a carcass, completely regenerated an entire worm (Zeng et al., 2018). A major strength of scRNA-seq is that it allows researchers to sample tissues in an unbiased fashion and detect new cell types for which there is no known marker or culture method (Jaitin et al., 2014; Shapiro et al., 2013; Wen and Tang, 2016). This kind of discovery, in one sense a eukaryotic equivalent to bacterial metagenomics (Lasken, 2012), represents the power and promise of scRNA-seq.

Single-cell transcriptomic technologies are rapidly evolving. Following the first demonstration of single-cell RNA sequencing (Tang et al., 2009), researchers have developed STRT-seq (single-cell tagged reverse transcription) (Islam et al., 2011, 2012), Smart-seq/Smart-seq2 (Picelli et al., 2013; Ramsköld et al., 2012), CEL-seq/CEL-seq2 (cell expression by linear amplification) (Hashimshony et al., 2012, 2016), Quartz-seq (Sasagawa et al., 2013), MARS-seq (massively parallel single-cell RNA-sequencing) (Jaitin et al., 2014), MALBAC-RNA (multiple annealing and looping-based amplification cycles) (Chapman et al., 2015), Drop-seq (Macosko et al., 2015), inDrops (Klein et al., 2015), 10x Genomics (Zheng et al., 2017), sci-RNA-seq (Cao et al., 2017), and SPLiT-seq (Rosenberg et al., 2018), to name a few. (Many of these techniques are concisely reviewed in (Wen and Tang, 2016).)

Broadly speaking, scRNA-seq methods rely on isolating single cells into either individual wells of a culture plate or within microliter-sized droplets (Wen and Tang, 2016). The cells are

lysed, messenger RNA (mRNA) is reverse transcribed, and a complementary DNA (cDNA) library is created. The use of barcoded adapters and template-switching oligonucleotides uniquely labels molecules from each cell, which are then pooled and sequenced. The resulting reads are mapped to a reference genome to identify which genes are expressed and the barcode is used to link gene expression values to individual cells. A gene expression matrix can then be constructed with cell barcodes along one axis and gene expression values along the other. Clustering algorithms reduce the high-dimensional transcriptional profiles of individual cells into groups of cell types with shared transcriptomic profiles (Becht et al., 2019; Grün et al., 2015; Satija et al., 2015; Xu and Su, 2015; Zeisel et al., 2015). The pattern of gene expression within each cluster can then be used to assign cell type. Recently, researchers have been able to overlay directional information on top of clusters by comparing the relative ratios of intronic to exonic read counts (La Manno et al., 2018; Svensson and Pachter, 2018).

The assorted scRNA-seq methods have their strengths and weaknesses. STRT-seq only captures the 5' ends of transcripts while CEL-seq captures the 3' ends. Although this may be suitable for measuring gene expression, they cannot be used to detect alternative splicing or isoform switching. Smart-seq creates a shotgun library out of the entire transcript, which avoids biasing data towards either end, but is incompatible with the use of unique molecular indexes (UMIs) (Ziegenhain et al., 2017). Individual cells contain such little mRNA (Macaulay and Voet, 2014) that amplification bias, drop-out, and other PCR artifacts can skew the interpretability of the final dataset. UMIs enable digital counting of transcripts, which is a more noise-tolerant way of measuring transcript abundance than relative comparisons of mRNA content (Islam et al., 2013).

Methods relying on manual processing of individual cells in wells can sequence, at most, a few hundred cells per run (Wen and Tang, 2016). The Fluidigm C1 microfluidic platform has been another popular option for scRNA-seq (Achim et al., 2015; Bacher and Kendzierski, 2016; Xin et al., 2016) because of its ability to automatically sort cells into 96-well plate, perform reverse transcription, and generate sequencing libraries in a hands-off manner (Hebenstreit, 2012). Although it can be used to sequence several hundreds of cells in a single experiment (Wen and Tang, 2016), it is the most expensive platform for scRNA-seq due to its use of proprietary equipment and consumables (Ziegenhain et al., 2017). More self-reliant platforms, like Drop-seq (Macosko et al., 2015) and inDrops (Klein et al., 2015), provide microfluidic scRNA-seq at an affordable price point. 10x Genomics' Chromium platform has further increased adoption of scRNA-seq by offering easy-to-use kits and competitive cell recovery rates (Zheng et al., 2017). Recently, microwells (Han et al., 2018b) and combinatorial barcoding strategies (Cao et al., 2017; Datlinger et al., 2021; Rosenberg et al., 2018) have pushed scRNA-seq to ever larger library sizes, breaking the million-cell barrier (Cao et al., 2019).

Drop-seq provides a representative example of how microfluidic scRNA-seq techniques work. This method uses microparticle beads which are tagged polythymidine oligonucleotides. Each oligonucleotide on a bead has a constant sequence (SMART), a 12-bp barcode shared across each of probes (the cell barcode), and an 8-bp UMI which is unique to each oligonucleotide. The beads are sent through a microfluidic device in a lysis buffer where they intersect a parallel stream of single cells in suspension. Immediately afterwards, an oil stream splits the aqueous stream into distinct microfluidic droplets. These droplets now contain a single cell and as the cell lyses, polyadenylated mRNA molecules are captured by the polythymidine sequences on the bead. Library preparation occurs in bulk after droplets have been disrupted.

However, single-cell resolution is retained due to the incorporation of the cell barcode and UMI into the cDNA. Amplification of the cDNA followed by tagmentation results in a library of 3' transcript ends tagged with information connecting them to cell-of-origin (cell barcode) and transcript-of-origin (UMI).

This library is sequenced and the reads are mapped to a reference. To classify cell types from this data, a digital gene expression (DGE) matrix is generated. This table lists cell barcodes along one axis and genes along the other. The values in the matrix are integer counts of the number of UMIs observed for a given gene for a specified cell barcode. Each UMI can only have originated from a single transcript, but PCR jackpotting (Cha and Thilly, 1993) may lead to the preferential and disproportionate amplification of a few molecules (Lou et al., 2013). Restricting the values of the matrix to the UMI counts helps correct for these kinds of biases (Stegle et al., 2015).

A number of informatic strategies can be used to identify transcriptionally distinct groups of cells (Butler et al., 2018; Hsu and Culhane, 2020; Satija et al., 2015; Stuart et al., 2019; Wolf et al., 2018). Typically this process begins by identifying a subset of highly variable genes across the dataset, then performing a combination of principal components analysis (PCA) of these genes followed by either *t*-distributed stochastic neighbor embedding (t-SNE) (van der Maaten and Hinton, 2008) or uniform manifold approximation and projection (UMAP) (McInnes and Healy, 2018). PCA clusters cells according to vectors which are linear combination of genes; these vectors reflect genes that strongly covary with each other (Heimberg et al., 2016). This high-dimensional representation is then transformed into a two-dimensional projection (using t-SNE or UMAP) such that points that were close to each other in high-dimensional space remain close to each other in low-dimensional space. Clusters are identified within this representation

using either a density-based approach (Ester et al., 1996) or community detection algorithms (Blondel et al., 2008; Traag et al., 2019) Finally, the clusters are organized hierarchically and gene expression profiles within each cluster are used for cell type assignment. The original Drop-seq paper used this method to discover rare subpopulations of retinal amacrine cells that had not been previously described and proposed new marker genes for their identification (Macosko et al., 2015).

Limitations of Drop-seq, and droplet microfluidic methods more generally, include a low cell capture efficiency: the ratio of beads to cells is chosen to minimize the number of droplets containing two cells (Macosko et al., 2015). As a result, only about 5% of beads will be encapsulated with a cell, while the rest of the droplets will be empty (i.e. a bead without a cell) (Ziegenhain et al., 2017). Thus, relatively large number of cells are required as input into the microfluidic device, posing a challenge for rare cell type detection. Close packing of hydrogel beads containing barcoded reverse transcription primers improves cell recovery rates (Klein et al., 2015; Zheng et al., 2017). There are also limits to the amount of mRNA captured by the microparticles; estimates suggest that only 10-20% of all mRNA in a droplet is actually represented in the final library (Macosko et al., 2015), which could result in data biased towards abundantly transcribed genes. Despite this low-coverage libraries can still be used for cell type classification due to strong covariances within gene expression networks (Heimberg et al., 2016).

Multi-modal scRNA-seq techniques have emerged to mechanistically understand the factors governing a particular transcriptional state (Figure 1). We now have methods to simultaneously measure single cell transcripts and: genotype (Dey et al., 2015; Han et al., 2018a; Li et al., 2015; Macaulay et al., 2015); methylated cytosines (Angermueller et al., 2016; Hu et al., 2016); chromatin accessibility (Cao et al., 2018; Ma et al., 2020); protein levels

(Katzenelenbogen et al., 2020; Peterson et al., 2017; Stoeckius et al., 2017); genetic perturbation (Datlinger et al., 2017; Dixit et al., 2016; Jaitin et al., 2016); and lineage (Alemany et al., 2018; Bidy et al., 2018; Raj et al., 2018; Spanjaard et al., 2018; Wagner et al., 2018). However, one major class of studies that, until very recently, was absent from this suite of techniques were transcription factor binding assays.

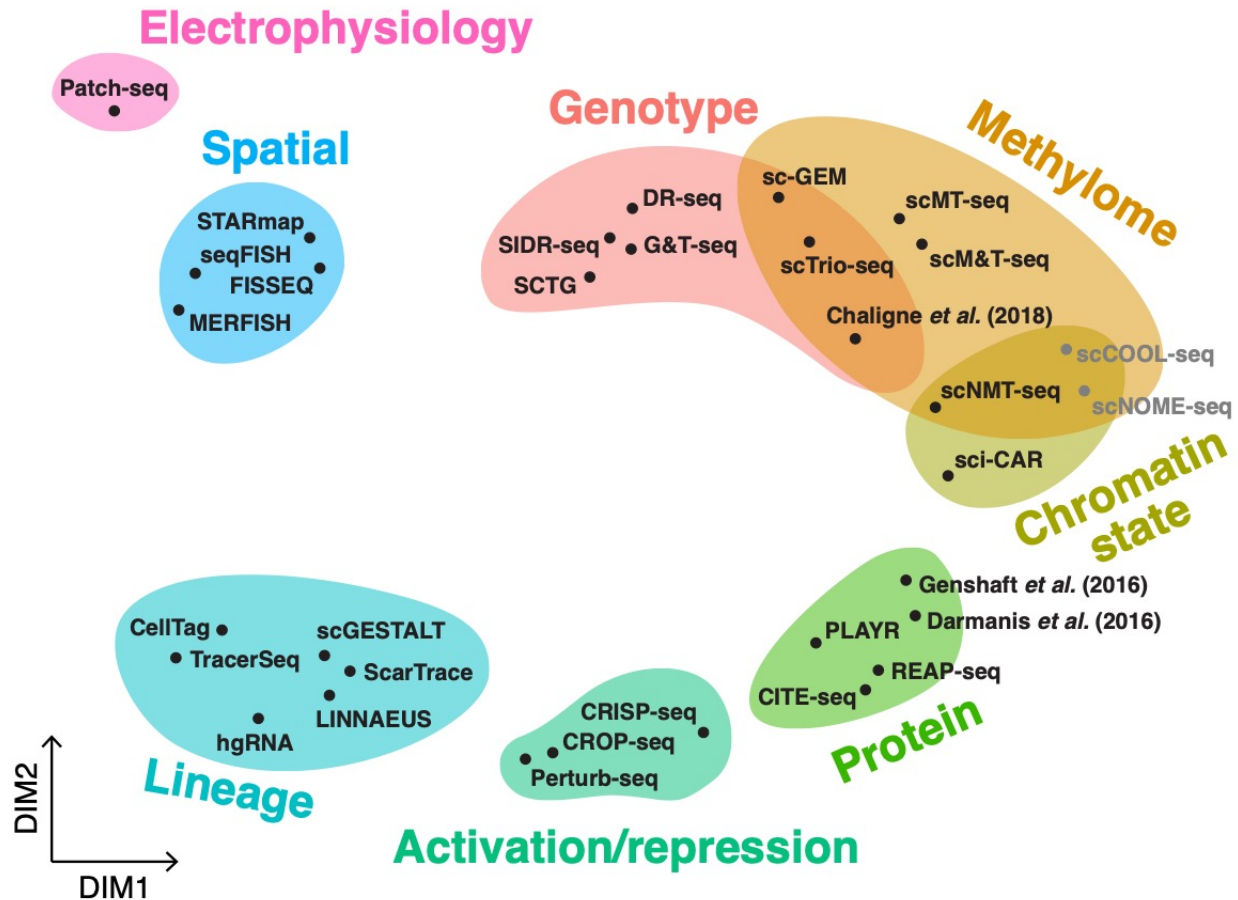


Figure 1.1: The landscape of multimodal scRNA-seq technologies. Methods in grey are theoretically compatible with scRNA-seq but have not demonstrated this in practice.

Transcription factors (TFs) encompass DNA-binding and chromatin-associated proteins that regulate gene expression. It is well-established that TFs are responsible for determining cell fate and maintaining cell identity (Baumgardt et al., 2007; Hevner, 2006; Holmberg and Perlmann, 2012; Iwafuchi-Doi and Zaret, 2016; Kuo and Grubbs, 1992; Orkin and Zon, 2002;

Roeder and Radtke, 2009; Wilson et al., 2003). TFs can directly bind promoters as well as more distant elements termed enhancers. A subset of enhancers known as super-enhancers are thought to regulate cell identity genes and are characterized by high densities of cell-type-specific TFs, permissive epigenetic marks, and transcriptional coactivators like BRD4 and MED1 (Di Micco et al., 2014; Hnisz et al., 2013; Niederriter et al., 2015; Whyte et al., 2013). Overexpression of TFs can force the transdifferentiation of one cell type into another (Holmberg and Perlmann, 2012; Liu et al., 2008; Megeney et al., 1996; Takahashi and Yamanaka, 2006; Tapscott, 2005), while removal of binding sites can profoundly alter morphology (Gonen et al., 2018; Kvon et al., 2016).

TF binding is commonly assayed using chromatin immunoprecipitation followed by sequencing (ChIP-seq) in which cellular proteins and DNA are chemically crosslinked with formaldehyde (Johnson et al., 2007; Park, 2009; Shlyueva et al., 2014). The DNA is then sonicated to an average of 200-600 bp fragments; optionally, fragmentation with an exonuclease (ChIP-exo), micrococcal nuclease (ChEC-seq), or the transposase Tn5 can be used to increase resolution (Furey, 2012; Schmid et al., 2004; Schmidl et al., 2015; Skene and Henikoff, 2017; Zentner et al., 2015). An antibody isolates fragments of genomic DNA bound by the TF of interest. The TF-DNA crosslinks are reversed and the freed DNA is used to construct a sequencing library whose reads should map back to regions bound by the TF. Another approach to identifying binding sites, DamID, relies on fusing TFs to *Escherichia coli* DNA adenine methyltransferase (*dam*) (Greil et al., 2006; Vogel et al., 2007). When the fusion protein binds DNA, the *dam* domain methylates adenines at nearby GATC sites. This modification does not occur naturally in mammalian cells and so is specific to the TF fusion. A methylation-specific

PCR amplifies bases bound by the TF and sequencing these reads should correspond to TF binding sites.

DNase-seq (Song and Crawford, 2010) and assay for transposase-accessible chromatin using sequencing (ATAC-seq) (Buenrostro et al., 2013) are two methods for identifying open chromatin. They work by either lightly digesting or tagmenting (Adey et al., 2010) the genome, which fragments unbound DNA. Sequencing these regions can lend insight into which genes may be actively transcribed; meanwhile, regions that were not found by sequencing may be silenced. By themselves, however, these methods cannot inform us as to which TFs may be regulating the activity of particular genes. Instead, they describe regions where certain binding motifs may be accessible or not (Stergachis et al., 2013), but this is correlative evidence at best, not causal. Different TFs can share very similar motifs, making it even more difficult to identify which protein was actually bound (Shively et al., 2019). In combination with TF-specific ChIP-seq, ATAC-seq data has been used to infer TF binding (Buenrostro et al., 2013), but this necessitates having good ChIP-seq data for a particular cell type.

No technologies exist to jointly assay gene expression and TF binding in single cells. Such a technique would be uniquely poised to make mechanistic connections between TF activity and cell identity. Several antibody-based single cell techniques have recently been developed to measure DNA-protein contacts (Ai et al., 2019; Carter et al., 2019; Grosselin et al., 2019; Hainer et al., 2019; Harada et al., 2019; Kaya-Okur et al., 2019; Rotem et al., 2015; Wang et al., 2019) but they are unable to also recover mRNA. As such, their use is restricted to pre-defined cell types and, in general, for highly abundant proteins like histones and master TFs. Single cell DamID has been reported (Kind et al., 2013, 2015) and has recently been combined with scRNA-seq (Rooijers et al., 2019) but has only been used for studying interactions between

the genome and the nuclear lamina. Thus, to measure TF binding in single cells, a new method is needed.

Our lab has previously developed transposon calling cards as a unique method of recording interactions between DNA-binding proteins and the genome (Wang et al., 2007, 2011, 2012). It relies on two components: a transposable donor element (the “calling card”), which becomes permanently integrated into the genome; and a helper, which is a TF fused to either a protein that directs a transposase—as in yeast; (Shively et al., 2019; Wang et al., 2007, 2011)—or a transposase itself—as in mammalian cells; (Cammack et al., 2020; Wang et al., 2012; Yen et al., 2018). The TF-transposase complex directs insertion of the donor element to genomic loci visited by the TF. After a period of time, genomic DNA is harvested, transposon insertions are recovered, sequenced, and analyzed.

For mammalian cells we use the *piggyBac* (PB) transposase (PBase), originally isolated from the moth *Trichoplusia ni* (Yusa, 2015), and its hyperactive version (HyPBase) (Yusa et al., 2011) as helpers. These proteins show robust activity in a variety of animal cell types (Di Matteo et al., 2014; Galvan et al., 2009; Wu et al., 2006; Yusa, 2013) and are amenable to N-terminal TF fusions (Wu et al., 2006). The donor element is a screenable or selectable marker gene (e.g. tdTomato, puromycin N-acetyltransferase) flanked by PB terminal repeat sequences. Between the marker gene and the downstream terminal repeat, we have added a bank of restriction sites, a barcode sequence (to uniquely identify a particular donor), and a next-generation sequencing site. In a typical calling cards experiment, a plasmid encoding the TF-PBase is transfected along with several barcoded donor plasmids. The TF-PBase directs insertion of the donor transposons near TF binding sites. The timing of transposase activity is critical for recording. One way to manage this is to simply transfect the helper plasmid when we want to record TF binding. We

can also fuse TF-PBase to the estrogen receptor ERT2 (Cadiñanos and Bradley, 2007) or a destabilization domain (Banaszynski et al., 2006) to create an inducible transposase (Higdon III, 2015; Mayhew, 2014; Qi et al., 2017). These constructs are inactive until treatment with tamoxifen, an estrogen analog, or the small molecule Shld1, respectively. When induced, the TF-PBase directs insertion of the donor into the genome; upon removal of these agents, TF-PBase activity stops (Qi et al., 2017).

At a later point, we harvest cells, isolate genomic DNA, and digest it. The fragmented libraries are then diluted and incubated overnight to promote self-circularization. Next, an inverse PCR is performed to specifically amplify transposon insertions that have successfully circularized. The products of this step are sent for high-throughput sequencing. One read will start within the terminal repeat, read through the PB insertion site motif (“TTAA”) and into genomic sequence. The other read will pick up the barcode labeling that donor sequence. The first read is trimmed and mapped back to the genome, which informs us to loci visited by the TF. To increase confidence that a given mapped read represents a true TF-genome interaction, we require multiple barcoded insertions at the same locus.

Integration of the donor sequence provides a permanent record of TF-genome interactions. These marks are stably passed on during mitosis, so interactions recorded in progenitor or stem cells are preserved in terminally differentiated cell types. The differentiated cells may also be more numerous than the progenitor cells, so in a sense the TF binding signal becomes naturally amplified. We have successfully used transposon calling cards to map TF binding in progenitor motor neurons in mouse embryoid bodies (Mayhew, 2014) and in zebrafish embryos (Higdon III, 2015).

One understandable criticism of transposon calling cards is the potential for mutagenesis caused by the TF-PBase fusion integrating donors into the genome. Off-target effects in a stem cell population may lead to aberrant development and thus may not represent normal transcriptional regulation. Moreover, gene trap constructs delivered with *piggyBac* has been used to discover cancer driver genes (Cadiñanos and Bradley, 2007). Our experience suggests that *in vivo* mutagenic effects are not a significant concern. Transfection of zebrafish embryos with calling cards plasmids revealed successful transposition in nearly every cell without a visible phenotypic effect (Higdon III, 2015). Typically, we expect 50-100 insertions per cell, a small number of relative to the size of vertebrate genomes. Our model organisms are also diploid, so in the event that one promoter or enhancer is disrupted, the other copy may still be haplosufficient. TF-PBase constructs tend to insert donors 100-300 bases away from the TF binding site (Wang et al., 2012), so the ability of the TF to bind and exert its function is likely preserved. Finally, in diploid yeast—which has much smaller intergenic regions than mammalian genomes—we recover essential genes at comparable frequencies to non-essential genes (Wang et al., 2011), suggesting at most a tolerable amount of mutagenesis.

Other limitations are related to sequence constraints. The *piggyBac* transposase almost always inserts into TTAA tetranucleotide motifs (Wang et al., 2012). In relatively GC-rich regions, there may be no suitable insertion site and so a potentially meaningful TF interaction may not be recorded. Furthermore, genomic sequence downstream of the insertion must contain one of the restriction sites found in the donor sequence. If no site is sufficiently close, the inverse PCR product will be too long to sequence on the Illumina platform (Wang et al., 2012). One potential way to overcome this is to use capture probes designed against the *piggyBac* terminal repeat, thereby isolating all the insertions in the genome (Mann et al., 2016).

Applying transposon calling cards to heterogeneous tissues is not straightforward. Bulk calling card data would have the same interpretability problem that faces ChIP-seq: we would observe an amalgam of binding sites without a clear link to any particular cell type. While scRNA-seq can successfully identify unique cell types, it is an RNA-based technique while transposon calling cards is a DNA-based assay. Dual-assay techniques that combine genome sequencing with scRNA-seq have recently been described (Macaulay et al., 2015) but amplifying genomic DNA from single cells is susceptible to allelic dropout and amplification bias (Gawad et al., 2016). Given the relatively small number (50-100) of insertions per cell we expect, relying on single-cell genome sequencing may result in noisy data that would not adequately reflect a TF's binding portfolio.

Our solution, as described over the remainder of this dissertation, is the “self-reporting” transposon. This is a novel calling card donor that can be isolated from RNA instead of DNA. This is one of our standard *piggyBac* donor plasmids with a puromycin resistance marker driven by the constitutively active promoter elongation factor 1-alpha (EF-1 α). Unlike prior donors, however, this construct lacks a polyadenylation signal. Thus, as RNA polymerase II transcribes the resistance gene, it continues through the downstream PB terminal repeat and into the flanking genomic sequence. Hence, it “self-reports” its genomic coordinates via mRNA.

Our new method, single cell calling cards, enables single cell recovery of SRTs. This protocol starts from scRNA-seq libraries, allowing us to characterize the transcriptomes our cell population and assign cell barcodes to distinct clusters. Transcripts from SRTs are captured and tagged with cell barcodes and UMIs alongside cellular mRNA. We can specifically amplify the SRT transcripts and create a library where both the junction of the genome-SRT junction and the cell barcode and UMI are on the same piece of DNA. High-throughput sequencing allows us to

map each insertion to a specific cell barcode. By aggregating insertion sites found across all barcodes within a cluster, we characterize the TF binding profiles of each cell type. Thus, with one single cell calling cards experiment we can assay TF binding across multiple cell types, free from the constraints of *in vitro* cultures.

Finally, while SRTs clearly enable single cell calling cards, the idea of SRTs may have broader utility than TF binding. Any genomic assay extracting location information could, in theory, be combined with scRNA-seq via SRTs. Deploying SRTs across the genome, particularly in an unbiased fashion, could be used to discover underappreciated regulatory mechanisms such as cryptic polyadenylation or splicing. More abstractly, the location of the SRT itself can be conceptualized as a positional barcode. Since transpositions are vertically transmitted, their distribution across a population of cells may reflect patterns of clonality and be used to infer lineage relationships. Finally, conditionally expressed SRTs may be useful as molecular sentinels of a kind, activating in response to upstream triggering events. Whether these constructs are widely adopted by the genomics community remains to be seen. In the interim, we might surprise and delight ourselves by thinking creatively and open-mindedly about SRTs and their application.

1.2 References

Achim, K., Pettit, J.-B., Saraiva, L.R., Gavriouchkina, D., Larsson, T., Arendt, D., and Marioni, J.C. (2015). High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nature Biotechnology* 33, 503–509.

Adey, A., Morrison, H.G., Asan, Xun, X., Kitzman, J.O., Turner, E.H., Stackhouse, B., MacKenzie, A.P., Caruccio, N.C., Zhang, X., et al. (2010). Rapid, low-input, low-bias

construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biology* 11, 1.

Ai, S., Xiong, H., Li, C.C., Luo, Y., Shi, Q., Liu, Y., Yu, X., Li, C., and He, A. (2019). Profiling chromatin states using single-cell *itChIP-seq*. *Nat Cell Biol* 21, 1164–1172.

Alemaný, A., Florescu, M., Baron, C.S., Peterson-Maduro, J., and van Oudenaarden, A. (2018). Whole-organism clone tracing using single-cell sequencing. *Nature* 556, 108–112.

Almendo, V., Marusyk, A., and Polyak, K. (2013). Cellular Heterogeneity and Molecular Evolution in Cancer. *Annual Review of Pathology: Mechanisms of Disease* 8, 277–302.

Angermueller, C., Clark, S.J., Lee, H.J., Macaulay, I.C., Teng, M.J., Hu, T.X., Krueger, F., Smallwood, S., Ponting, C.P., Voet, T., et al. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods* 13, 229–232.

Bacher, R., and Kendziorski, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biology* 17, 1.

Banaszynski, L.A., Chen, L., Maynard-Smith, L.A., Ooi, A.G.L., and Wandless, T.J. (2006). A Rapid, Reversible, and Tunable Method to Regulate Protein Function in Living Cells Using Synthetic Small Molecules. *Cell* 126, 995–1004.

Baumgardt, M., Miguel-Aliaga, I., Karlsson, D., Ekman, H., and Thor, S. (2007). Specification of Neuronal Identities by Feedforward Combinatorial Coding. *PLOS Biol* 5, e37.

- Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology* 37, 38–44.
- Biddy, B.A., Kong, W., Kamimoto, K., Guo, C., Waye, S.E., Sun, T., and Morris, S.A. (2018). Single-cell mapping of lineage and identity in direct reprogramming. *Nature* 564, 219–224.
- Biressi, S., Molinaro, M., and Cossu, G. (2007). Cellular heterogeneity during vertebrate skeletal muscle development. *Developmental Biology* 308, 281–293.
- Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.* 2008, P10008.
- Brooks, M.D., Burness, M.L., and Wicha, M.S. (2015). Therapeutic Implications of Cellular Heterogeneity and Plasticity in Breast Cancer. *Cell Stem Cell* 17, 260–271.
- Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* 10, 1213–1218.
- Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 36, 411–420.
- Cadiñanos, J., and Bradley, A. (2007). Generation of an inducible and optimized piggyBac transposon system. *Nucleic Acids Research* 35, e87–e87.

Cammack, A.J., Moudgil, A., Chen, J., Vasek, M.J., Shabsovich, M., McCullough, K., Yen, A., Lagunas, T., Maloney, S.E., He, J., et al. (2020). A viral toolkit for recording transcription factor–DNA interactions in live mouse tissues. *Proc Natl Acad Sci USA* *117*, 10003–10014.

Cann, G.M., Gulzar, Z.G., Cooper, S., Li, R., Luo, S., Tat, M., Stuart, S., Schroth, G., Srinivas, S., Ronaghi, M., et al. (2012). mRNA-Seq of Single Prostate Cancer Circulating Tumor Cells Reveals Recapitulation of Gene Expression and Pathways Found in Prostate Cancer. *PLoS ONE* *7*, e49144.

Cao, J., Packer, J.S., Ramani, V., Cusanovich, D.A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S.N., Steemers, F.J., et al. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* *357*, 661–667.

Cao, J., Cusanovich, D.A., Ramani, V., Aghamirzaie, D., Pliner, H.A., Hill, A.J., Daza, R.M., McFaline-Figueroa, J.L., Packer, J.S., Christiansen, L., et al. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* *33*, eaau0730.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature* *566*, 496–502.

Carter, B., Ku, W.L., Kang, J.Y., Hu, G., Perrie, J., Tang, Q., and Zhao, K. (2019). Mapping histone modifications in low cell number and single cells using antibody-guided chromatin tagmentation (ACT-seq). *Nat Commun* *10*, 3747.

Cha, R.S., and Thilly, W.G. (1993). Specificity, efficiency, and fidelity of PCR. *Genome Research* *3*, S18–S29.

Chapman, A.R., He, Z., Lu, S., Yong, J., Tan, L., Tang, F., and Xie, X.S. (2015). Single Cell Transcriptome Amplification with MALBAC. *PLoS ONE* *10*, e0120889.

Datlinger, P., Rendeiro, A.F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L.C., Kuchler, A., Alpar, D., and Bock, C. (2017). Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods* *14*, 297–301.

Datlinger, P., Rendeiro, A.F., Boenke, T., Senekowitsch, M., Krausgruber, T., Barreca, D., and Bock, C. (2021). Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. *Nat Methods* *18*, 635–642.

Dey, S.S., Kester, L., Spanjaard, B., Bienko, M., and van Oudenaarden, A. (2015). Integrated genome and transcriptome sequencing of the same cell. *Nature Biotechnology* *33*, 285–289.

Di Matteo, M., Samara-Kuko, E., Ward, N.J., Waddington, S.N., McVey, J.H., Chuah, M.K., and VandenDriessche, T. (2014). Hyperactive PiggyBac Transposons for Sustained and Robust Liver-targeted Gene Therapy. *Molecular Therapy* *22*, 1614–1624.

Di Micco, R., Fontanals-Cirera, B., Low, V., Ntziachristos, P., Yuen, S.K., Lovell, C.D., Dolgalev, I., Yonekubo, Y., Zhang, G., Rusinova, E., et al. (2014). Control of Embryonic Stem Cell Identity by BRD4-Dependent Transcriptional Elongation of Super-Enhancer-Associated Pluripotency Genes. *Cell Reports* *9*, 234–247.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* *167*, 1853-1866.e17.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, (AAAI Press), pp. 226–231.

Furey, T.S. (2012). ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nature Reviews Genetics* 13, 840–852.

Galvan, D.L., Nakazawa, Y., Kaja, A., Kettlun, C., Cooper, L.J.N., Rooney, C.M., and Wilson, M.H. (2009). Genome-wide Mapping of PiggyBac Transposon Integrations in Primary Human T Cells. *Journal of Immunotherapy* 32, 837–844.

Gawad, C., Koh, W., and Quake, S.R. (2016). Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics* 17, 175–188.

Gonen, N., Futtner, C.R., Wood, S., Alexandra Garcia-Moreno, S., Salamone, I.M., Samson, S.C., Sekido, R., Poulat, F., Maatouk, D.M., and Lovell-Badge, R. (2018). Sex reversal following deletion of a single distal enhancer of Sox9. *Science* 360, 1469–1471.

Goodell, M.A., Nguyen, H., and Shroyer, N. (2015). Somatic stem cell heterogeneity: diversity in the blood, skin and intestinal stem cell compartments. *Nature Reviews Molecular Cell Biology* 16, 299–309.

Greil, F., Moorman, C., and van Steensel, B. (2006). DamID: Mapping of In Vivo Protein–Genome Interactions Using Tethered DNA Adenine Methyltransferase. In *DNA Microarrays, Part A: Array Platforms and Wet-Bench Protocols*, (Elsevier), pp. 342–359.

Grosselin, K., Durand, A., Marsolier, J., Poitou, A., Marangoni, E., Nemati, F., Dahmani, A., Lameiras, S., Reyal, F., Frenoy, O., et al. (2019). High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nat Genet* 51, 1060–1066.

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., and van Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* 525, 251–255.

Hainer, S.J., Bošković, A., McCannell, K.N., Rando, O.J., and Fazzio, T.G. (2019). Profiling of Pluripotency Factors in Single Cells and Early Embryos. *Cell* 177, 1319-1329.e11.

Han, K.Y., Kim, K.-T., Joung, J.-G., Son, D.-S., Kim, Y.J., Jo, A., Jeon, H.-J., Moon, H.-S., Yoo, C.E., Chung, W., et al. (2018a). SIDR: simultaneous isolation and parallel sequencing of genomic DNA and total RNA from single cells. *Genome Res.* 28, 75–87.

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., et al. (2018b). Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* 172, 1091-1107.e17.

Harada, A., Maehara, K., Handa, T., Arimura, Y., Nogami, J., Hayashi-Takanaka, Y., Shirahige, K., Kurumizaka, H., Kimura, H., and Ohkawa, Y. (2019). A chromatin integration labelling method enables epigenomic profiling with lower input. *Nat Cell Biol* 21, 287–296.

Hashimshony, T., Wagner, F., Sher, N., and Yanai, I. (2012). CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Reports* 2, 666–673.

Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., Gennert, D., Li, S., Livak, K.J., Rozenblatt-Rosen, O., et al. (2016). CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biology* *17*, 1.

Hebenstreit, D. (2012). Methods, Challenges and Potentials of Single Cell RNA-seq. *Biology* *1*, 658–667.

Heimberg, G., Bhatnagar, R., El-Samad, H., and Thomson, M. (2016). Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. *Cell Systems* *2*, 239–250.

Hevner, R.F. (2006). From radial glia to pyramidal-projection neuron. *Molecular Neurobiology* *33*, 33–50.

Higdon III, C.W. (2015). Gene Expression and Enhancer Discovery in the Neural Crest. PhD Thesis.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-Enhancers in the Control of Cell Identity and Disease. *Cell* *155*, 934–947.

Holmberg, J., and Perlmann, T. (2012). Maintaining differentiated cellular identity. *Nature Reviews Genetics* *13*, 429–439.

Hsu, L.L., and Culhane, A.C. (2020). Impact of Data Preprocessing on Integrative Matrix Factorization of Single Cell Data. *Front. Oncol.* *10*, 973.

Hu, Y., Huang, K., An, Q., Du, G., Hu, G., Xue, J., Zhu, X., Wang, C.-Y., Xue, Z., and Fan, G. (2016). Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biol* *17*, 88.

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Research* *21*, 1160–1167.

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., and Linnarsson, S. (2012). Highly multiplexed and strand-specific single-cell RNA 5[prime] end sequencing. *Nature Protocols* *7*, 813–828.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2013). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods* *11*, 163–166.

Iwafuchi-Doi, M., and Zaret, K.S. (2016). Cell fate control by pioneer transcription factors. *Development* *143*, 1833–1837.

Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., et al. (2014). Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science* *343*, 776–779.

Jaitin, D.A., Weiner, A., Yofe, I., Lara-Astiaso, D., Keren-Shaul, H., David, E., Salame, T.M., Tanay, A., van Oudenaarden, A., and Amit, I. (2016). Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell* *167*, 1883-1896.e15.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science* 316, 1497–1502.

Karaiskos, N., Wahle, P., Alles, J., Boltengagen, A., Ayoub, S., Kipar, C., Kocks, C., Rajewsky, N., and Zinzen, R.P. (2017). The Drosophila embryo at single-cell transcriptome resolution. *Science* 358, 194–199.

Katzenelenbogen, Y., Sheban, F., Yalin, A., Yofe, I., Svetlichnyy, D., Jaitin, D.A., Bornstein, C., Moshe, A., Keren-Shaul, H., Cohen, M., et al. (2020). Coupled scRNA-Seq and Intracellular Protein Activity Reveal an Immunosuppressive Role of TREM2 in Cancer. *Cell* 182, 872-885.e19.

Kaya-Okur, H.S., Wu, S.J., Codomo, C.A., Pledger, E.S., Bryson, T.D., Henikoff, J.G., Ahmad, K., and Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat Commun* 10, 1930.

Kind, J., Pagie, L., Ortabozkoyun, H., Boyle, S., de Vries, S.S., Janssen, H., Amendola, M., Nolen, L.D., Bickmore, W.A., and van Steensel, B. (2013). Single-Cell Dynamics of Genome-Nuclear Lamina Interactions. *Cell* 153, 178–192.

Kind, J., Pagie, L., de Vries, S.S., Nahidiazar, L., Dey, S.S., Bienko, M., Zhan, Y., Lajoie, B., de Graaf, C.A., Amendola, M., et al. (2015). Genome-wide Maps of Nuclear Lamina Interactions in Single Human Cells. *Cell* 163, 134–147.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* 161, 1187–1201.

Knoechel, B., Roderick, J.E., Williamson, K.E., Zhu, J., Lohr, J.G., Cotton, M.J., Gillespie, S.M., Fernandez, D., Ku, M., Wang, H., et al. (2014). An epigenetic mechanism of resistance to targeted therapy in T cell acute lymphoblastic leukemia. *Nat Genet* 46, 364–370.

Krieger, T., and Simons, B.D. (2015). Dynamic stem cell heterogeneity. *Development* 142, 1396–1406.

Kuo, C.J., and Grubbs, G.R. (1992). Hepatocyte Differentiation. In *Development*, V.A. Russo, S. Brody, D. Cove, and S. Ottolenghi, eds. (Berlin), pp. 479–498.

Kvon, E.Z., Kamneva, O.K., Melo, U.S., Barozzi, I., Osterwalder, M., Mannion, B.J., Tissières, V., Pickle, C.S., Plajzer-Frick, I., Lee, E.A., et al. (2016). Progressive Loss of Function in a Limb Enhancer during Snake Evolution. *Cell* 167, 633-642.e11.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastrioti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. *Nature* 560, 494–498.

Lasken, R.S. (2012). Genomic sequencing of uncultured microorganisms from single cells. *Nature Reviews Microbiology* 10, 631–640.

Li, W., Calder, R.B., Mar, J.C., and Vijg, J. (2015). Single-cell transcriptogenomics reveals transcriptional exclusion of ENU-mutated alleles. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 772, 55–62.

Litzenburger, U.M., Buenrostro, J.D., Wu, B., Shen, Y., Sheffield, N.C., Kathiria, A., Greenleaf, W.J., and Chang, H.Y. (2017). Single-cell epigenomic variability reveals functional cancer heterogeneity. *Genome Biol* 18, 15.

Liu, X., Huang, J., Chen, T., Wang, Y., Xin, S., Li, J., Pei, G., and Kang, J. (2008). Yamanaka factors critically regulate the developmental signaling network in mouse embryonic stem cells. *Cell Research* 18, 1177–1189.

Lou, D.I., Hussmann, J.A., McBee, R.M., Acevedo, A., Andino, R., Press, W.H., and Sawyer, S.L. (2013). High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proceedings of the National Academy of Sciences of the United States of America* 110, 19872–19877.

Ma, S., Zhang, B., LaFave, L., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V.K., Law, T., Lareau, C., et al. (2020). Chromatin potential identified by shared single cell profiling of RNA and chromatin (Genomics).

van der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 2579–2605.

Macaulay, I.C., and Voet, T. (2014). Single Cell Genomics: Advances and Future Perspectives. *PLOS Genet* 10, e1004126.

Macaulay, I.C., Haerty, W., Kumar, P., Li, Y.I., Hu, T.X., Teng, M.J., Goolam, M., Saurat, N., Coupland, P., Shirley, L.M., et al. (2015). G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nature Methods* 12, 519–522.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* *161*, 1202–1214.

Mann, K.M., Newberg, J.Y., Black, M.A., Jones, D.J., Amaya-Manzanares, F., Guzman-Rojas, L., Kodama, T., Ward, J.M., Rust, A.G., van der Weyden, L., et al. (2016). Analyzing tumor heterogeneity and driver genes in single myeloid leukemia cells with SBCapSeq. *Nature Biotechnology*.

Mayhew, D. (2014). Recording Transcription Factor Binding During Development. PhD Thesis.

McInnes, L., and Healy, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv.Org*.

Megeney, L.A., Kablar, B., Garrett, K., Anderson, J.E., and Rudnicki, M.A. (1996). MyoD is required for myogenic stem cell function in adult skeletal muscle. *Genes & Development* *10*, 1173–1183.

Michor, F., and Polyak, K. (2010). The Origins and Implications of Intratumor Heterogeneity. *Cancer Prevention Research* *3*, 1361–1364.

Miyamoto, D.T., Zheng, Y., Wittner, B.S., Lee, R.J., Zhu, H., Broderick, K.T., Desai, R., Fox, D.B., Brannigan, B.W., Trautwein, J., et al. (2015). RNA-Seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance. *Science* *349*, 1351–1356.

Niederriter, A., Varshney, A., Parker, S., and Martin, D. (2015). Super Enhancers in Cancers, Complex Disease, and Developmental Disorders. *Genes* *6*, 1183–1200.

Orkin, S.H., and Zon, L.I. (2002). Hematopoiesis and stem cells: plasticity versus developmental heterogeneity. *Nature Immunology* 3, 323–328.

Park, P.J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics* 10, 669–680.

Peterson, V.M., Zhang, K.X., Kumar, N., Wong, J., Li, L., Wilson, D.C., Moore, R., McClanahan, T.K., Sadekova, S., and Klappenbach, J.A. (2017). Multiplexed quantification of proteins and transcripts in single cells. *Nature Biotechnology* 9, 2579.

Picelli, S., Björklund, A., Åsa K, Faridani, O.R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods* 10, 1096–1098.

Plasschaert, L.W., Žilionis, R., Choo-Wing, R., Savova, V., Knehr, J., Roma, G., Klein, A.M., and Jaffe, A.B. (2018). A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* 560, 377–381.

Qi, Z., Wilkinson, M.N., Chen, X., Sankararaman, S., Mayhew, D., and Mitra, R.D. (2017). An optimized, broadly applicable piggyBac transposon induction system. *Nucleic Acids Research* gkw1290.

Raj, B., Wagner, D.E., McKenna, A., Pandey, S., Klein, A.M., Shendure, J., Gagnon, J.A., and Schier, A.F. (2018). Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat Biotechnol* 36, 442–450.

Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O.R., Daniels, G.A., Khrebtukova, I., Loring, J.F., Laurent, L.C., et al. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology* 30, 777–782.

Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., et al. (2017). The Human Cell Atlas. *ELife* 6, e27041.

Roeder, I., and Radtke, F. (2009). Stem cell biology meets systems biology. *Development* 136, 3525–3530.

Rooijers, K., Markodimitraki, C.M., Rang, F.J., de Vries, S.S., Chialastri, A., de Luca, K.L., Mooijman, D., Dey, S.S., and Kind, J. (2019). Simultaneous quantification of protein–DNA contacts and transcriptomes in single cells. *Nat Biotechnol* 37, 766–772.

Rosenberg, A.B., Roco, C.M., Muscat, R.A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L.T., Peeler, D.J., Mukherjee, S., Chen, W., et al. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* 360, 176–182.

Rotem, A., Ram, O., Shores, N., Sperling, R.A., Goren, A., Weitz, D.A., and Bernstein, B.E. (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nature Biotechnology* 33, 1165–1172.

Sasagawa, Y., Nikaido, I., Hayashi, T., Danno, H., Uno, K.D., Imai, T., and Ueda, H.R. (2013). Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biology* 14, 1.

Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* 33, 495–502.

Schmid, M., Durussel, T., and Laemmli, U.K. (2004). ChIC and ChEC. *Molecular Cell* 16, 147–157.

Schmidl, C., Rendeiro, A.F., Sheffield, N.C., and Bock, C. (2015). CHIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nat Methods* 12, 963–965.

Shalek, A.K., Satija, R., Shuga, J., Trombetta, J.J., Gennert, D., Lu, D., Chen, P., Gertner, R.S., Gaublomme, J.T., Yosef, N., et al. (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* 510, 363–369.

Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics* 14, 618–630.

Shively, C.A., Liu, J., Chen, X., Loell, K., and Mitra, R.D. (2019). Homotypic cooperativity and collective binding are determinants of bHLH specificity and function. *Proc Natl Acad Sci USA* 116, 16143–16152.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics* 15, 272–286.

Silvestre-Roig, C., Hidalgo, A., and Soehnlein, O. (2016). Neutrophil heterogeneity: implications for homeostasis and pathogenesis. *Blood* 127, 2173–2181.

Skene, P.J., and Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *ELife* 6, e21856.

Song, L., and Crawford, G.E. (2010). DNase-seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells. *Cold Spring Harbor Protocols* 2010, pdb.prot5384-pdb.prot5384.

Spanjaard, B., Hu, B., Mitic, N., Olivares-Chauvet, P., Janjuha, S., Ninov, N., and Junker, J.P. (2018). Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. *Nature Biotechnology* 36, 469–473.

Stegle, O., Teichmann, S.A., and Marioni, J.C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics* 16, 133–145.

Stergachis, A.B., Neph, S., Reynolds, A., Humbert, R., Miller, B., Paige, S.L., Vernot, B., Cheng, J.B., Thurman, R.E., Sandstrom, R., et al. (2013). Developmental Fate and Cellular Maturity Encoded in Human Regulatory DNA Landscapes. *Cell* 154, 888–903.

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K., Swerdlow, H., Satija, R., and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods* 14, 865–868.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell* 177, 1888-1902.e21.

Svensson, V., and Pachter, L. (2018). RNA Velocity: Molecular Kinetics from Single-Cell RNA-Seq. *Molecular Cell* 72, 7–9.

Takahashi, K., and Yamanaka, S. (2006). Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* 126, 663–676.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* 6, 377–382.

Tapscott, S.J. (2005). The circuitry of a master switch: MyoD and the regulation of skeletal muscle gene transcription. *Development* 132, 2685–2695.

Traag, V.A., Waltman, L., and van Eck, N.J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* 9, 5233.

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. *Genome Research* 25, 1491–1498.

Treutlein, B., Brownfield, D.G., Wu, A.R., Neff, N.F., Mantalas, G.L., Espinoza, F.H., Desai, T.J., Krasnow, M.A., and Quake, S.R. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* 509, 371–375.

Vogel, M.J., Peric-Hupkes, D., and van Steensel, B. (2007). Detection of in vivo protein–DNA interactions using DamID in mammalian cells. *Nature Protocols* 2, 1467–1478.

Wagner, D.E., Weinreb, C., Collins, Z.M., Briggs, J.A., Megason, S.G., and Klein, A.M. (2018). Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* 360, 981–987.

Wang, H., Johnston, M., and Mitra, R.D. (2007). Calling cards for DNA-binding proteins. *Genome Research* *17*, 1202–1209.

Wang, H., Mayhew, D., Chen, X., Johnston, M., and Mitra, R.D. (2011). Calling Cards enable multiplexed identification of the genomic targets of DNA-binding proteins. *Genome Research* *21*, 748–755.

Wang, H., Mayhew, D., Chen, X., Johnston, M., and Mitra, R.D. (2012). “Calling Cards” for DNA-Binding Proteins in Mammalian Cells. *Genetics* *190*, 941–949.

Wang, Q., Xiong, H., Ai, S., Yu, X., Liu, Y., Zhang, J., and He, A. (2019). CoBATCH for High-Throughput Single-Cell Epigenomic Profiling. *Molecular Cell* *76*, 206-216.e7.

Wen, L., and Tang, F. (2016). Single-cell sequencing in stem cell biology. *Genome Biology* *17*, 1.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell* *153*, 307–319.

Wilson, M.E., Scheel, D., and German, M.S. (2003). Gene expression cascades in pancreatic development. *Mechanisms of Development* *120*, 65–80.

Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY : large-scale single-cell gene expression data analysis. *Genome Biology* *19*, 15.

Wu, S.C.-Y., Meir, Y.-J.J., Coates, C.J., Handler, A.M., Pelczar, P., Moisyadi, S., and Kaminski, J.M. (2006). piggyBac is a flexible and highly active transposon as compared to sleeping beauty,

Tol2, and Mos1 in mammalian cells. *Proceedings of the National Academy of Sciences* *103*, 15008–15013.

Xin, Y., Kim, J., Ni, M., Wei, Y., Okamoto, H., Lee, J., Adler, C., Cavino, K., Murphy, A.J., Yancopoulos, G.D., et al. (2016). Use of the Fluidigm C1 platform for RNA sequencing of single mouse pancreatic islet cells. *Proceedings of the National Academy of Sciences of the United States of America* *113*, 3293–3298.

Xu, C., and Su, Z. (2015). Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* *31*, 1974–1980.

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y.E., et al. (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* *500*, 593–597.

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., et al. (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nature Structural & Molecular Biology* *20*, 1131–1139.

Yen, M., Qi, Z., Chen, X., Cooper, J.A., Mitra, R.D., and Onken, M.D. (2018). Transposase mapping identifies the genomic targets of BAP1 in uveal melanoma. *BMC Med Genomics* *11*, 97.

Yusa, K. (2013). Seamless genome editing in human pluripotent stem cells using custom endonuclease-based gene targeting and the piggyBac transposon. *Nature Protocols* *8*, 2061–2078.

Yusa, K. (2015). piggyBac Transposon. *Microbiology Spectrum* 3.

Yusa, K., Zhou, L., Li, M.A., Bradley, A., and Craig, N.L. (2011). A hyperactive piggyBac transposase for mammalian applications. *Proceedings of the National Academy of Sciences of the United States of America* 108, 1531–1536.

Zeisel, A., Munoz-Manchado, A.B., Codeluppi, S., Lonnerberg, P., La Manno, G., Jureus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347, 1138–1142.

Zeng, A., Li, H., Guo, L., Gao, X., McKinney, S., Wang, Y., Yu, Z., Park, J., Semerad, C., Ross, E., et al. (2018). Prospectively Isolated Tetraspanin + Neoblasts Are Adult Pluripotent Stem Cells Underlying Planaria Regeneration. *Cell* 173, 1593-1608.e20.

Zentner, G.E., Kasinathan, S., Xin, B., Rohs, R., and Henikoff, S. (2015). ChEC-seq kinetics discriminates transcription factor binding sites by DNA sequence and shape in vivo. *Nature Communications* 6, 8733.

Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications* 8, 14049.

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., and Enard, W. (2017). Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell* 65, 631-643.e4.

Chapter 2: Self-reporting transposons enable simultaneous readout of gene expression and transcription factor binding in single cells

(A version of this chapter was published in *Cell* 182, pp. 992-1008)

2.1 Abstract

Cellular heterogeneity confounds *in situ* assays of transcription factor (TF) binding. Single cell RNA-seq (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 maps 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*.

2.2 Introduction

Transcription factors (TFs) regulate gene expression during the most critical junctures in the specification of cell fate (Gurdon, 2016; Hafler et al., 2012; Mizuguchi et al., 2001; Zhu et al., 2012). They are central to the maintenance of stem cell pluripotency (Liu et al., 2008; Takahashi and Yamanaka, 2006) and are required for normal organogenesis during development (Fogarty et al., 2017). Overexpression of certain TFs can transdifferentiate one cell type into another (Davis et al., 1987), while abolishing TF binding sites can result in striking global phenotypes (Gonen et

al., 2018; Kvon et al., 2016). Furthermore, the pattern of TF binding is often dysregulated in disease states (Lee and Young, 2013). A better understanding of TF binding during tissue development and homeostasis would provide insights into how cellular diversity arises and is maintained under normal and abnormal biological conditions.

In the past few years, single cell RNA-seq (scRNA-seq) has emerged as the *de facto* paradigm 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). More recently, multi-modal scRNA-seq technologies have been developed (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) that link transcriptional information to other genomic assays. These methods are motivated by the realization that while scRNA-seq can describe the current state of a biological system, it alone cannot explain how that state arose. Thus, for a given population of cells, one can now simultaneously measure transcriptome and genome (Dey et al., 2015; Macaulay et al., 2015), or methylome (Angermueller et al., 2016; Clark et al., 2018), or chromatin accessibility (Cao et al., 2018; Clark et al., 2018), or cell surface markers (Peterson et al., 2017; Stoeckius et al., 2017). These techniques enable greater insight into the regulatory processes driving individual transcriptional programs.

A notable lacuna in the single cell repertoire is a method for simultaneously assaying transcriptome and TF binding. Such a method would allow for the genome-wide identification of TF binding sites across multiple cell types in complex tissues. ChIP-seq is the most popular technique for studying TF binding (Johnson et al., 2007), relying on antibodies specific to the factor of interest to pull down bound DNA. While a number of antibody-based single cell epigenomic methods have been reported (Ai et al., 2019; Carter et al., 2019; Grosselin et al.,

2019; Hainer et al., 2019; Harada et al., 2019; Kaya-Okur et al., 2019; Rotem et al., 2015; Wang et al., 2019), these techniques have generally mapped highly abundant proteins, such as modified histones and CTCF. DamID can recover TF binding sites by identifying nearby exogenously methylated adenines (Greil et al., 2006; Vogel et al., 2007), but in single cells it has only been used to study lamina-associated domains (Kind et al., 2013, 2015; Rooijers et al., 2019). Moreover, while combined single cell DamID and transcriptome (scDamID&T) has been described (Rooijers et al., 2019), it is a plate-based assay which limits throughput. None of the other single cell epigenomic techniques simultaneously capture mRNA (Ai et al., 2019; Carter et al., 2019; Grosselin et al., 2019; Hainer et al., 2019; Harada et al., 2019; Kaya-Okur et al., 2019; Rotem et al., 2015; Wang et al., 2019), restricting their use to predetermined cell types. In contrast, single cell assays for transposase-accessible chromatin (Buenrostro et al., 2015; Cao et al., 2018) can be used to identify nucleosome-free regions that may be bound by TFs across large numbers of mixed cells. However, they can only suggest the identity of potential DNA binding proteins by motif inference. These assays do not directly measure TF occupancy; moreover, they cannot be used to study transcriptional regulators that bind DNA indirectly or non-specifically, such as chromatin remodelers.

Our lab has previously developed transposon calling cards as an alternative assay of TF binding (Wang et al., 2007, 2011, 2012a). This system relies on two exogenous components: a fusion between a TF and a transposase, and a transposon carrying a reporter gene. The fusion transposase deposits transposons near TF binding sites, which are subsequently amplified from genomic DNA and subjected to high-throughput sequencing. Thus, the redirected transposase leaves “calling cards” at the genomic locations it has visited, which can be identified later in time. The result is a genome-wide assay of all binding sites for that particular TF. In mammalian

cells, we have heterologously expressed the *piggyBac* transposase (Ding et al., 2005) fused to the TF SP1 and shown that the resulting pattern of insertions reflects SP1's DNA binding preferences (Wang et al., 2012a). However, this method was only feasible in bulk preparations of thousands to millions 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 a novel construct called the self-reporting transposon (SRT). The genomic coordinates of inserted SRTs can be mapped from either mRNA or DNA, but the use of mRNA leads to both higher efficiency and compatibility with single-cell transcriptomics. We first establish that TF-directed SRTs, in bulk samples, retain the ability to accurately identify TF binding sites. Next, we demonstrate that the unfused *piggyBac* transposase, through its native affinity for the bromodomain TF BRD4, can be used to identify BRD4-bound super-enhancers (SEs). We then present the scCC method, which allows cell-specific mapping of SRTs from scRNA-seq libraries. This enables, in one experiment, concomitant assignment of cell types and identification of TF binding sites within those cells. We highlight the range of this technology by mapping the binding of multiple TFs in a variety of cell lines. We then use scCC to discover bromodomain-dependent cell state dynamics in K562 cells. We conclude by identifying cell type specific BRD4 binding sites *in vivo* in the postnatal mouse cortex. These results demonstrate that scCC could be a broadly applicable tool for the study of specific TF binding interactions across multiple cell types within heterogeneous systems.

2.3 Results

2.3.1 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* self-reporting transposon (SRT) by removing the polyadenylation signal from our standard DNA-based calling card vector (Figure 2.1A). This enables RNA polymerase II (Pol II) to transcribe the reporter gene contained in the transposon and continue through the terminal repeat (TR) into the flanking genomic sequence. Thus, SRTs “self-report” their locations through the unique genomic sequence found within the 3' untranslated regions (UTRs) of the reporter gene transcripts. Although previously published gene- or enhancer-trap transposons (Cadiñanos and Bradley, 2007) could, in principle, also capture local positional information via RNA, they are resolution-limited to the nearest gene or enhancer, respectively. In contrast, the 3' UTRs of SRT-derived transcripts contain the transposon-genome junction in the mRNA sequence, so we can map insertions with base-pair precision.

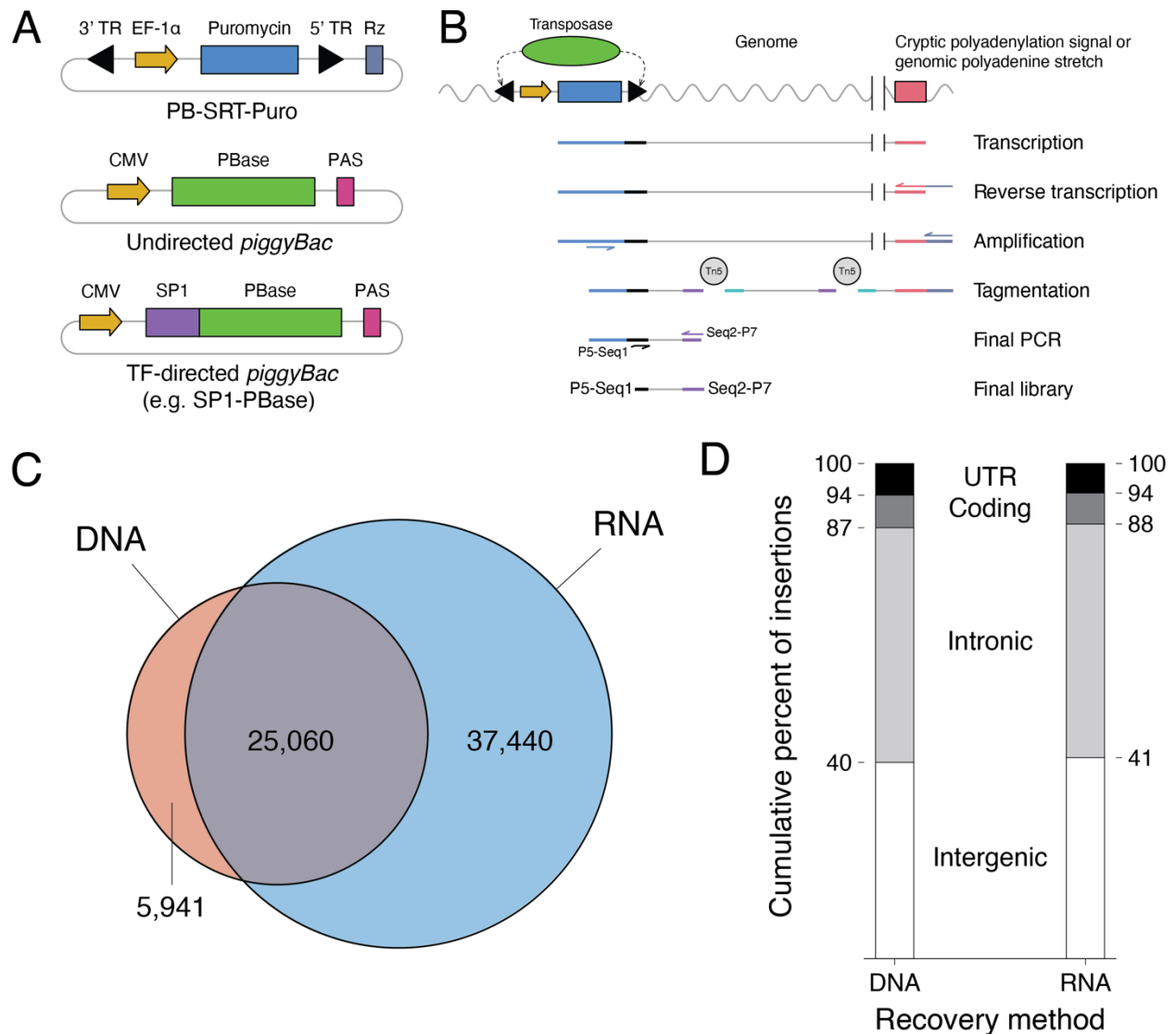


Figure 2.1: Self-reporting transposons (SRTs) are mapped more efficiently from RNA compared to DNA. (A) Schematics of a self-reporting *iggyBac* transposon with puromycin reporter gene (PB-SRT-Puro) and undirected (PBase) and SP1-directed (SP1-PBase) *iggyBac* transposases. (B) Molecular workflow for 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.

SRTs are mapped following reverse transcription (RT) and PCR amplification of self-reporting transcripts. These transcripts contain stretches of adenines that are derived from either cryptic polyadenylation signals (PAS) or polyadenine tracts encoded in genomic DNA downstream of the SRT insertion point (Figure 2.1B). A poly(T) RT primer hybridizes with these transcripts and introduces a universal priming site at one end of the transcripts. We then perform

a pair of nested PCRs with an intermediate tagmentation step (Picelli et al., 2014) to recover the transposon-genome junction. After adapter trimming and alignment, the 5' coordinates of these reads identify the genomic locations of insertions in the library. Libraries generated without transposase yield very few reads that map to the genome, but the protocol is highly efficient when transposase is added (Figure 2.2A).

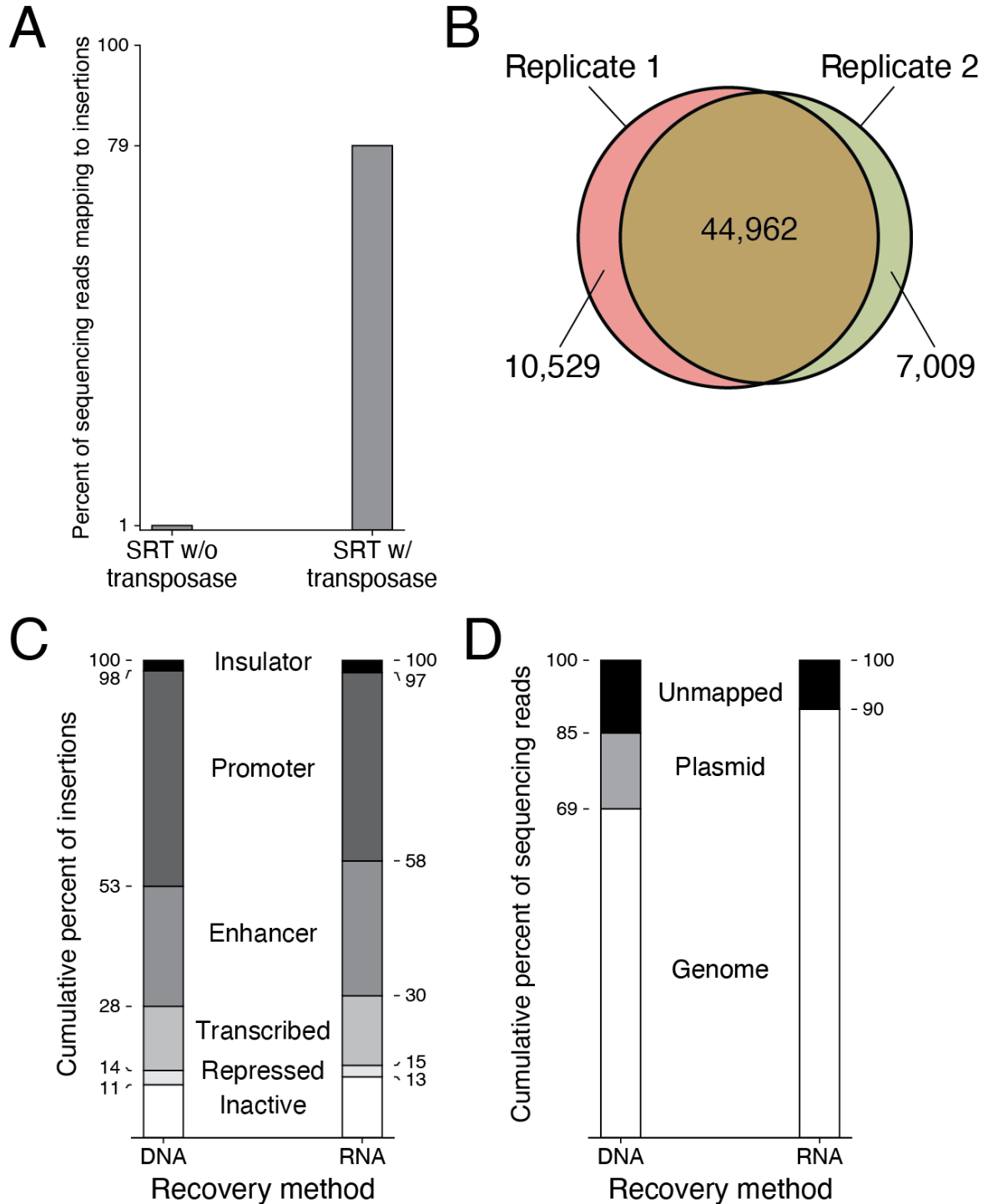


Figure 2.2: Properties of self-reporting transposons (SRTs). (A) Efficiencies of bulk RNA calling card libraries prepared from HEK293T cells transfected with PB-SRT-tdTomato with or without HyPBase transposase. (B) Overlap of SRTs recovered by two technical replicates of bulk RNA calling cards in HCT-116 cells transfected with PB-SRT-Puro and SP1-PBase. (C) Distribution of insertions with respect to chromatin state between SRT libraries prepared from either DNA or RNA. (D) Breakdown of sequencing reads mapping to the genome or plasmid from SRT libraries prepared from either DNA or RNA.

To compare transposon recovery between the new RNA-based protocol and our standard DNA-based inverse PCR protocol (Wang et al., 2012a), we transfected HCT-116 cells with a plasmid carrying a *piggyBac* SRT (PB-SRT-Puro) and a plasmid encoding a fusion of the TF SP1 and *piggyBac* transposase (SP1-PBase; Figure 2.1A). After two weeks of selection, we obtained approximately 2,300 puromycin-resistant clones. We split these cells into two populations: one half underwent inverse PCR while the other half were processed with our new RNA-based method. With inverse PCR, we obtained 31,001 insertions, while the RNA-based protocol recovered 62,500 insertions (mean coverage: 709 and 240 reads per insertion, respectively; Table 2.1). About 80% of the insertions recovered by DNA calling cards were also recovered in the RNA-based library (25,060 insertions; Figure 2.1C), an overlap comparable to that between technical replicates of the RNA workflow (Figure 2.2B). However, the RNA protocol recovered a further 37,440 insertions that were not found in the DNA-based library. To determine if these extra insertions were genuine, we analyzed the distribution of insertions by genetic annotation (Figure 2.1D) or chromatin state (Figure 2.2C). Transposons mapped from either the DNA or the RNA libraries showed comparable distributions with respect to genic annotation or chromatin states. This indicates that RNA-based recovery of transposons appears to be unbiased with respect to our established, DNA-based protocol.

Since SRT recovery relies on transcription, we wondered if SRTs deposited in euchromatic regions were recovered more efficiently than SRTs in less permissive chromatin states, which might lead to biases when mapping TF binding. Since *piggyBac* is known to preferentially insert near active chromatin (Yoshida et al., 2017), this question cannot be easily answered using this transposon. Prior studies have shown that the *Sleeping Beauty* transposase (Ivics et al., 1997; Mátés et al., 2009) has very little preference for chromatin state (Yoshida et

al., 2017). Therefore, we created a self-reporting *Sleeping Beauty* transposon and compared its genome-wide distribution to that of SRTs deposited by wild-type *piggyBac* (Table 2.1; Figure 2.3A-B). Undirected *piggyBac* transposases appeared to modestly prefer transposing into promoter and enhancers, which is consistent with previous reports (Gogol-Döring et al., 2016; Yoshida et al., 2017). By contrast, *Sleeping Beauty* showed largely uniform rates of insertions across all chromatin states, including repressed and inactive chromatin (Figure 2.3B). These results affirm that while RNA-based recovery is more efficient, it still preserves the underlying genomic distributions of insertions. Furthermore, because SRTs can be recovered from virtually any chromatin state, RNA-based calling card recovery can be employed to analyze a variety of TFs with broad chromatin-binding preferences.

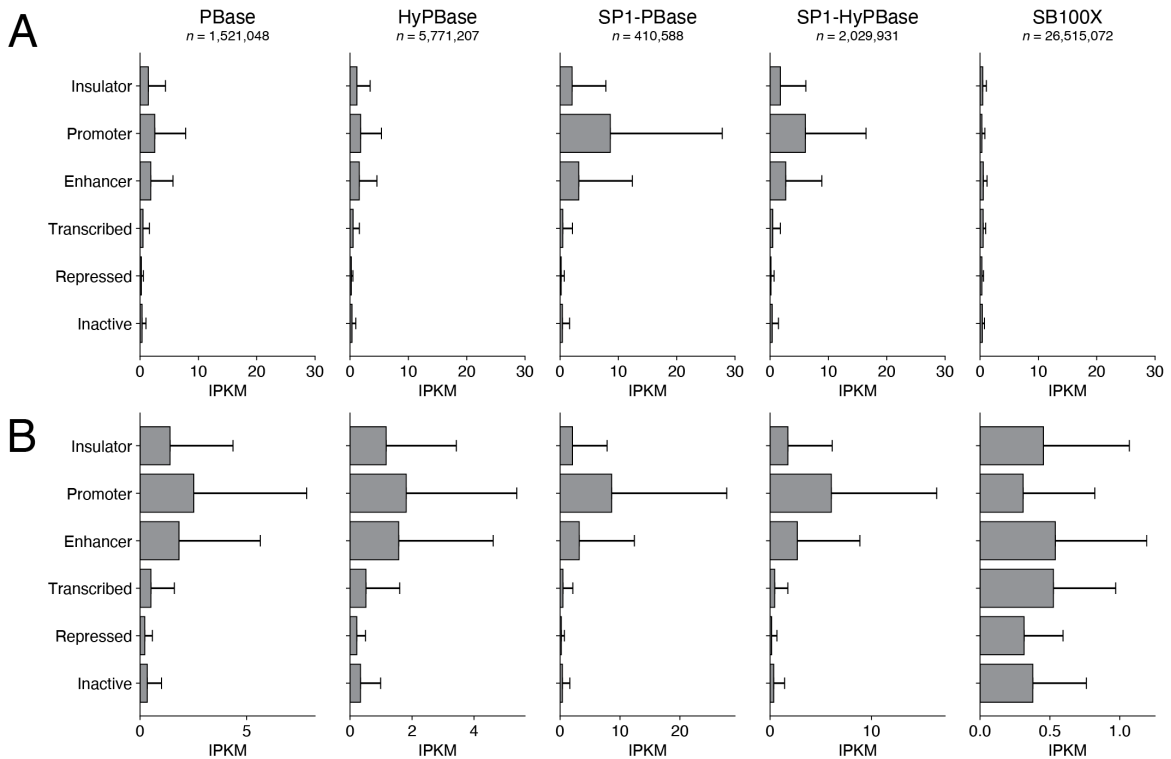


Figure 2.3: *piggyBac*, SP1-*piggyBac* fusions, and *Sleeping Beauty* display different local transposition rates depending on chromatin state. (A) Chromatin state analysis on the local rates of transposition of undirected *piggyBac*, SP1-*piggyBac* fusions, and *Sleeping Beauty* transposases in HCT-116 cells. (B) Same data as (A) but with different x-axes for each graph. IPKM: insertions per kilobase per million mapped insertions.

Table 2.1: Summary of bulk calling cards experiments

Sample	Construct	Modality	Replicates ^a	Insertions	Reads	Mean coverage
HCT-116	SP1-PBase	DNA ^b	1	31,001	21,975,948	708.9
HCT-116	SP1-PBase	RNA ^b	1	62,500	14,993,901	239.9
HCT-116	PBase	RNA	10	1,521,048	58,316,389	38.3
HCT-116	SP1-PBase	RNA	10	410,588	35,526,586	86.5
HCT-116	HyPBase	RNA	12	5,771,207	47,572,324	8.2
HCT-116	SP1-HyPBase	RNA	11	2,029,931	40,214,827	19.8
HCT-116	SB100X	RNA	12	26,515,072	67,650,985	2.6
OCM-1A	HyPBase	RNA	10	5,951,669	261,476,361	43.9
OCM-1A	BAP1-HyPBase	RNA	10	5,740,754	293,332,813	51.1

^a Biological replicates. ^b These experiments were used to assess DNA- vs. RNA-based recovery (Figure 2.1C).

A common artifact observed in DNA-based transposon recovery is a large fraction of reads aligning to the donor transposon plasmid instead of the genome. Although this can be mitigated by long selection times or by digestion with the methyladenine-sensitive enzyme DpnI (Wang et al., 2012a), these methods do not completely eliminate background and are not compatible with all experimental paradigms (e.g. viral transduction). To reduce this artifact, we included a hammerhead ribozyme (Yen et al., 2004) in the SRT plasmid downstream of the 5' TR. Before transposition, the ribozyme will cleave the nascent transcript originating from the marker gene, thus preventing RT. Transposition allows the SRT to escape the downstream ribozyme, leading to recovery of the self-reporting transcript. In our comparison of DNA- and RNA-based recovery, about 15% of reads from the SP1-PBase DNA library aligned to the plasmid, compared to fewer than 1% of reads from the RNA library (Figure 2.2D). Thus, the addition of a self-cleaving ribozyme virtually eliminated recovery of un-excised transposons.

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

Since the SRT is a new reagent, we sought to confirm that bulk RNA calling cards can, like DNA calling cards (Wang et al., 2012a), be used to identify TF binding sites. We transfected 10-12 replicates of HCT-116 cells with plasmids containing the PB-SRT-Puro donor transposon and SP1 fused to either *piggyBac* (SP1-PBase) or a hyperactive variant of *piggyBac* (Yusa et al., 2011) (SP1-HyPBase). As controls, we also transfected a similar number of replicates with

undirected PBase or HyPBase, respectively. We obtained 410,588 insertions from SP1-PBase and 1,521,048 insertions from PBase; similarly, we obtained 2,029,931 SP1-HyPBase insertions and 5,771,207 insertions from HyPBase (Table 2.1).

Just as we had observed previously with DNA calling cards (Wang et al., 2012a), RNA calling cards were also redirected by SP1-PBase and SP1-HyPBase to SP1-bound regions of the genome (Figure 2.4A and Figure 2.5A). Each circle in the insertions 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 \log_{10} scale. To better compare transposition across libraries with different numbers of insertions, we plotted the normalized local insertion rate as a density track. All three of the loci shown in Figure 2.4A and Figure 2.5A show a specific enrichment of calling card insertions in the SP1 fusion experiments that is not observed in the undirected control libraries. Next, we called peaks at all genomic regions enriched for SP1-directed transposition. The number of insertions observed at significant peaks for both SP1-PBase and SP1-HyPBase was highly reproducible between biological replicates ($R^2 = 0.87$ and 0.96 , respectively; Figure 2.4B and Figure 2.5B). Furthermore, calling card peaks were highly enriched for SP1 ChIP-seq signal at their centers, both on average (Figure 2.4C and Figure 2.5C) and in aggregate (Figure 2.4D and Figure 2.5D). SP1 is known to preferentially bind near TSSs and is also thought to play a role in demethylating CpG islands (Brandeis et al., 1994; Macleod et al., 1994; Philipson and Suske, 1999). We confirmed that the SP1-directed transposases preferentially inserted SRT calling cards near TSSs, CpG islands, and unmethylated CpG islands at statistically significant frequencies ($p < 10^{-9}$ in each instance, G test of independence; Figure 2.4E and Figure 2.5E). Moreover, compared to undirected *piggyBac*, SP1-directed *piggyBac* showed a striking preference for depositing insertions into promoters (Figure 2.3A-B). Lastly,

regions targeted by SP1-PBase and SP1-HyPBase were enriched for the core SP1 DNA binding motif ($p < 10^{-70}$ in each instance; Figure 2.4F and Figure 2.5F). Taken together, these results indicate that the genome-wide binding of SP1 can be accurately mapped using *piggyBac* SRTs.

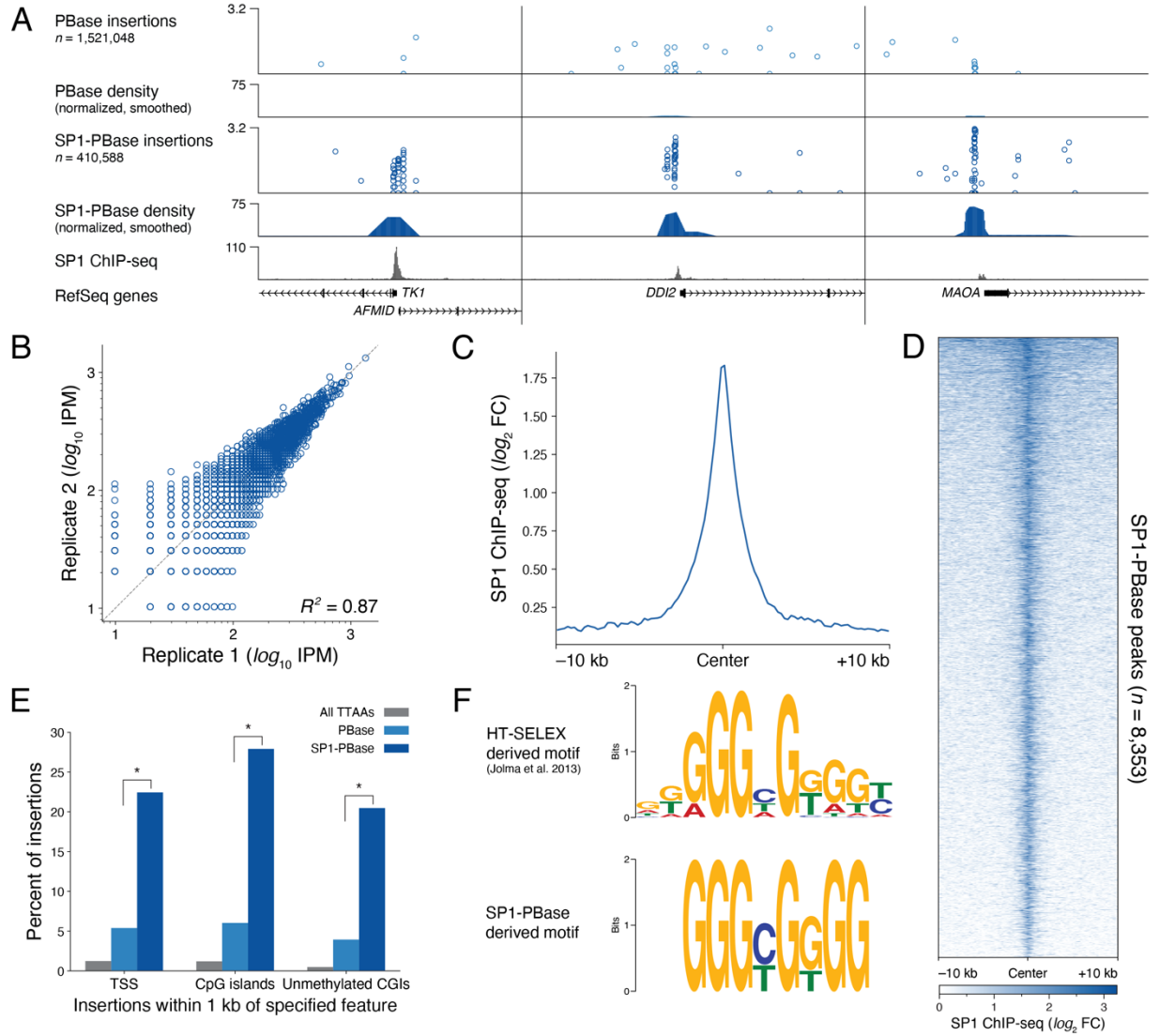


Figure 2.4: SP1 fused to *piggyBac* (SP1-PBase) redirects SRTs to SP1 binding sites. (A) Browser view of bulk SP1-PBase calling cards in HCT-116 cells. (B) Reproducibility of normalized insertions at bulk SP1-PBase peaks. (C) Mean SP1 ChIP-seq signal at bulk SP1-PBase peaks. (D) Heatmap of SP1 ChIP-seq signal at bulk SP1-PBase peaks. (E) Enrichment of SP1-PBase-directed insertions to TSSs, CGIs, and unmethylated CGIs (G test of independence $p < 10^{-9}$). (F) SP1 core motif elicited from bulk SP1-PBase peaks. IPM: insertions per million mapped insertions; FC: fold change; TSS: transcription start sites; CGI: CpG island.

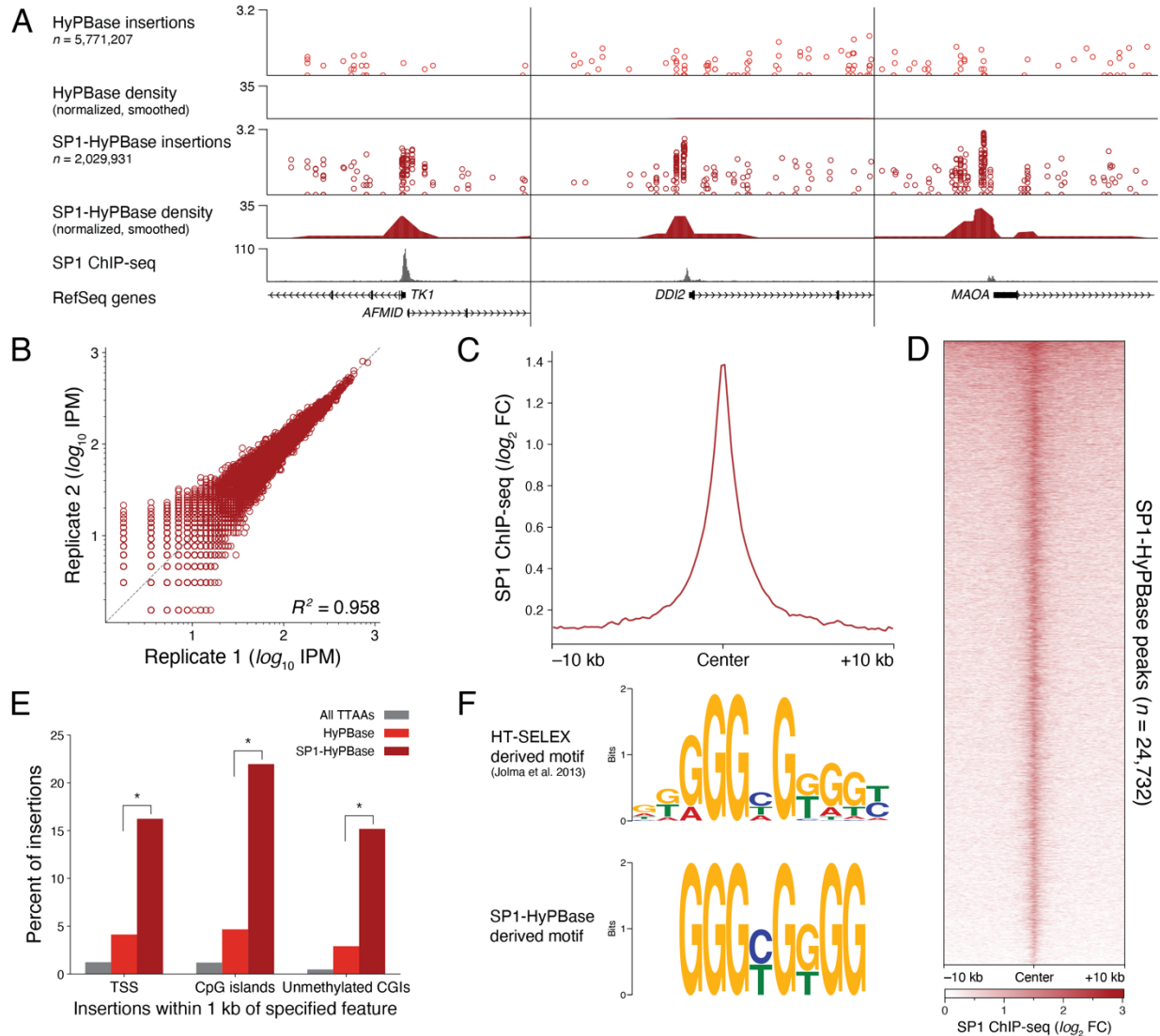


Figure 2.5: SP1 fused to hyperactive *piggyBac* (SP1-HyPBase) also redirects SRTs to SP1 binding sites. (A) Browser view of bulk SP1-HyPBase calling cards in HCT-116 cells. (B) Reproducibility of normalized insertions at bulk SP1-HyPBase peaks. (C) Mean SP1 ChIP-seq signal at bulk SP1-HyPBase peaks. (D) Heatmap of SP1 ChIP-seq signal at bulk SP1-HyPBase peaks. (E) Enrichment of SP1-HyPBase-directed insertions to TSSs, CGIs, and unmethylated CGIs (G test of independence $p < 10^{-9}$). (F) SP1 core motif elicited from bulk SP1-HyPBase peaks. IPM: insertions per million mapped insertions; FC: fold change; TSS: transcription start sites; CGI: CpG island.

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

Previous studies have shown that undirected PBase preferentially inserts transposons near super-enhancers (SEs) (Yoshida et al., 2017), a unique regulatory element that is thought to play a critical role in regulating cell identity (Hnisz et al., 2013). SEs are often enriched for the histone

modification H3K27ac as well as RNA polymerase 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). Moreover, PBase has a strong biophysical affinity for BRD4, as these proteins can be co-immunoprecipitated (Gogol-Döring et al., 2016). We hypothesized that, given the millions of insertions we assayed from the undirected PBase and HyPBase controls in the SP1-directed experiments (Figure 2.4A, Figure 2.5A), we would be able to identify BRD4-bound SEs simply from the localization of undirected *piggyBac* transpositions.

Both undirected PBase and HyPBase showed non-uniform densities of insertions at BRD4-bound loci (Figure 2.6A and Figure 2.7A). At statistically significant peaks of *piggyBac* calling cards, PBase and HyPBase showed high reproducibility of normalized insertions between biological replicates ($R^2 > 0.99$ in each instance; Figure 2.6B and Figure 2.7B). We calculated the mean BRD4 enrichment, as assayed by ChIP-seq (McClelland et al., 2016), over all *piggyBac* peaks, which showed significantly increased BRD4 signal compared to a permuted control set ($p < 10^{-9}$ in both instances, Kolmogorov-Smirnov test; Figure 2.6C and Figure 2.7C). Maximum BRD4 ChIP-seq signal was observed at calling card peak centers and decreased symmetrically in both directions. Furthermore, *piggyBac* peaks showed striking overlap with ChIP-seq profiles for several histone modifications (Sloan et al., 2016; The ENCODE Project Consortium, 2012), in particular an enrichment for H3K27 acetylation (Figure 2.6D, Figure 2.7D). Since bromodomains bind acetylated histones, this observation further supports the hypothesis 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). Taken together, these results demonstrate that *piggyBac* insertion density is highly correlated with BRD4 binding

throughout the genome and that regions enriched for undirected *piggyBac* insertions share features common to enhancers.

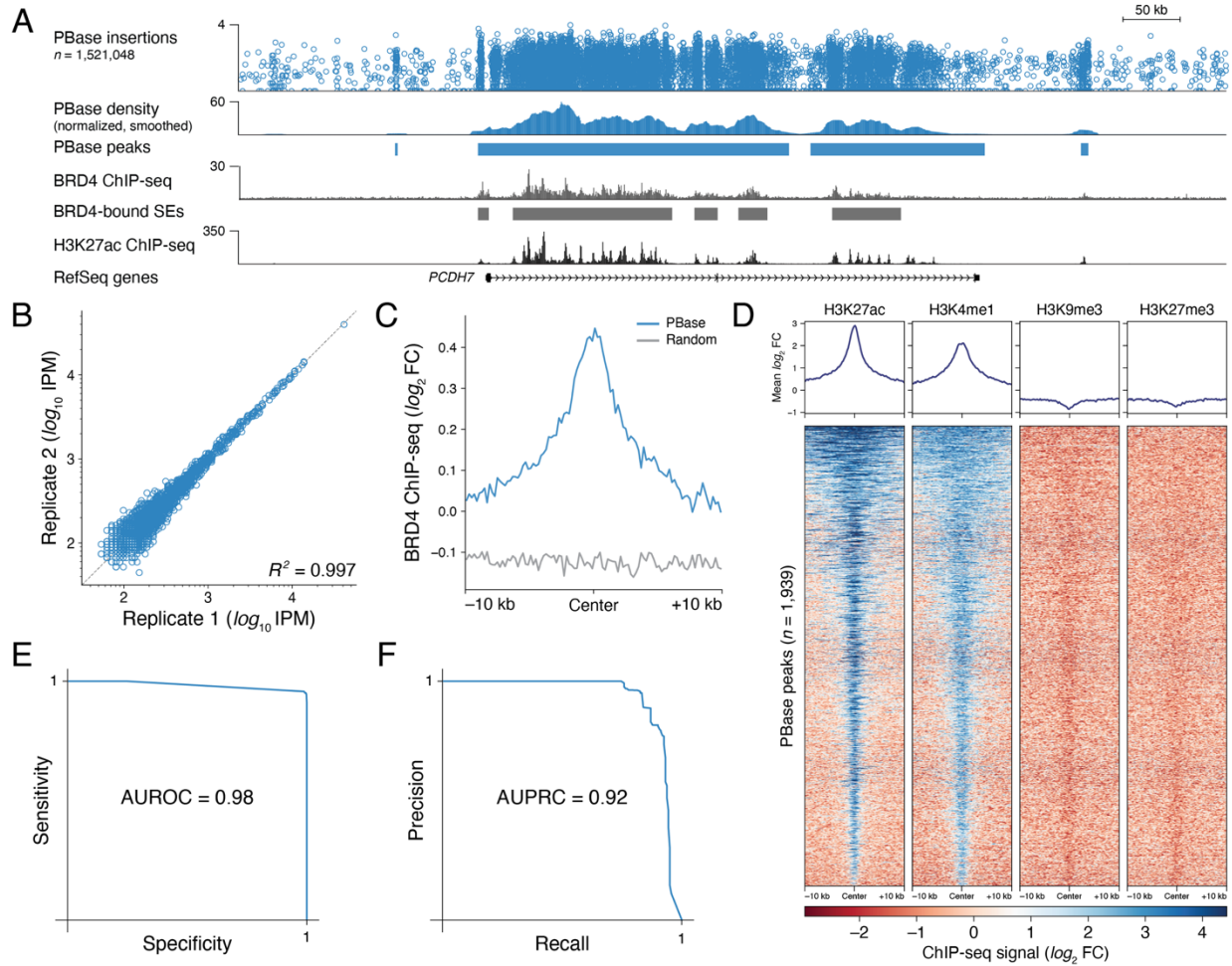


Figure 2.6: Undirected *piggyBac* (PBase) SRTs mark BRD4-bound super-enhancers (SEs). (A) Browser view of an undirected PBase insertions in HCT-116 cells at a SE alongside BRD4 and H3K27ac ChIP-seq data. (B) Reproducibility of normalized insertions at PBase peaks. (C) Mean BRD4 ChIP-seq signal at PBase peaks compared to a permuted control set. (D) Heatmap of H3K27ac, H3K4me1, H3K9me3, and H3K27me3 ChIP-seq signal at PBase peaks. (E) Receiver-operator characteristic curve for SE detection using PBase peaks. (F) Precision-recall curve for SE detection using PBase peaks. See also Figure S1. SE: super-enhancer; IPM: insertions per million mapped insertions; AUROC: area under receiver-operator curve; AUPRC: area under precision-recall curve; FC: fold change.

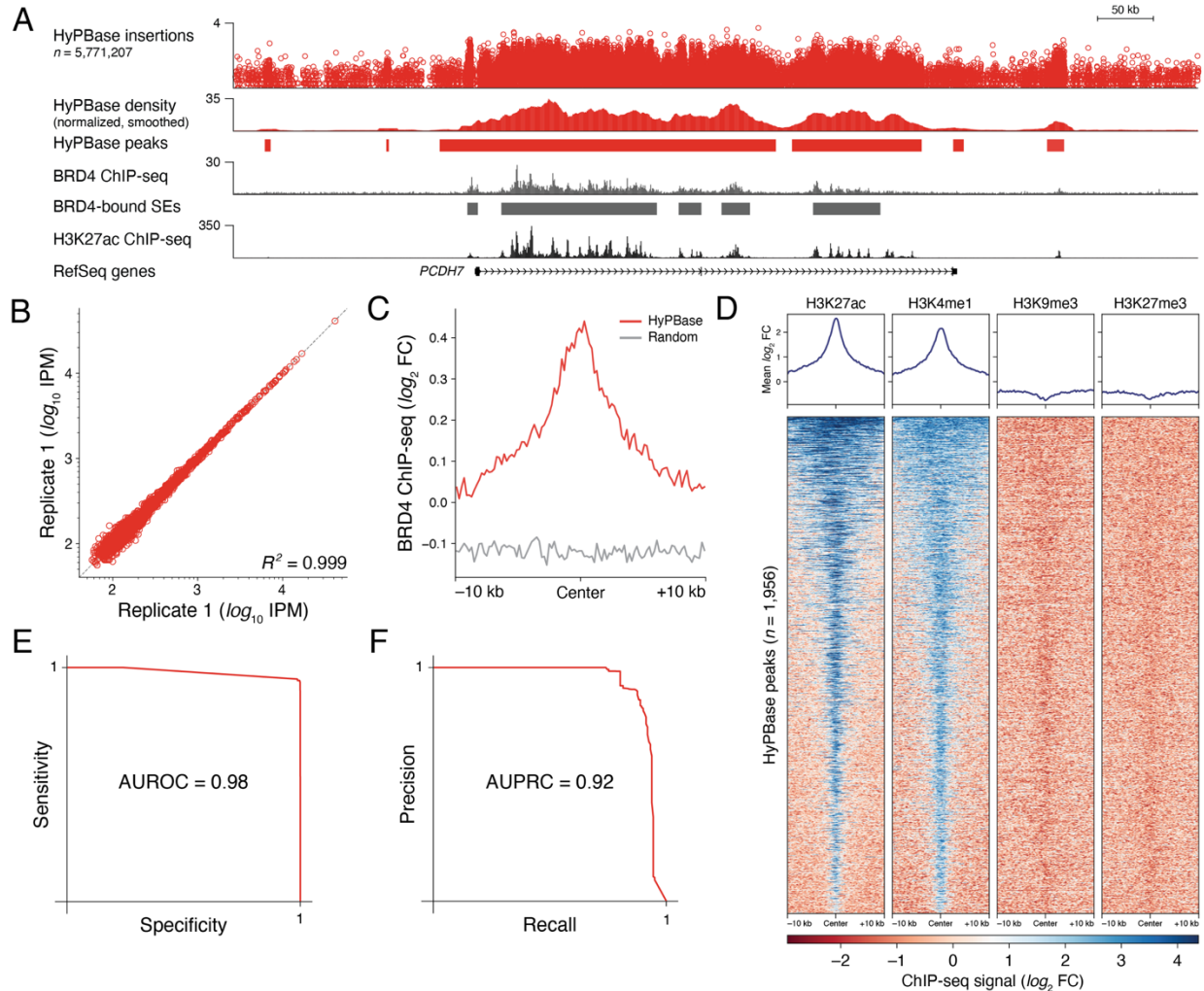


Figure 2.7: Undirected hyperactive *piggyBac* (HyPBase) SRTs also mark BRD4-bound super-enhancers (SEs). (A) Browser view of undirected HyPBase insertions at a SE alongside BRD4 and H3K27ac ChIP-seq data in HCT-116 cells. (B) Reproducibility of normalized insertions at HyPBase peaks. (C) Mean BRD4 ChIP-seq signal at HyPBase peaks compared to permuted control set. (D) Heatmap of H3K27ac, H3K4me1, H3K9me3, and H3K27me3 ChIP-seq signal at HyPBase peaks. (E) Receiver-operator characteristic curve for SE detection using HyPBase peaks. (F) Precision-recall curve for SE detection using HyPBase peaks. SE: super-enhancer; IPM: insertions per million mapped insertions; AUROC: area under receiver-operator curve; AUPRC: area under precision-recall curve; FC: fold change.

We next assessed whether undirected *piggyBac* peaks can be used to identify BRD4-bound SEs. We used BRD4 ChIP-seq data from HCT-116 cells (McClelland et al., 2016) to create a reference list of BRD-bound SEs (Pott and Lieb, 2014; Whyte et al., 2013) (Figure 2.6A, Figure 2.7A). We then constructed receiver-operator characteristic curves based on our ability to detect SEs from PBase- and HyPBase-derived peaks (Figure 2.6E and Figure 2.7E). The high

areas under the curves (0.98 in each instance) indicate that we can robustly identify BRD4-bound SEs from *piggyBac* transpositions. Across a range of sensitivities, calling card peaks are highly specific and have high positive predictive value (AUPRC = 0.92 in each instance; Figure 2.6F and Figure 2.7F). Thus, undirected *piggyBac* transpositions are an accurate assay of BRD4-bound SEs.

To project how calling cards would scale to single cell experiments, where molecular techniques show broadly reduced sensitivity compared their bulk counterparts, we simulated assay performance under increasingly sparse conditions. We quantified the relationship between SE sensitivity and the number of insertions recovered in undirected calling cards experiments by downsampling the data from the PBase and HyPBase experiments in half-log increments and calculating sensitivity (Figure 2.8A-B). These heatmaps show that sensitivity increases with the total number of insertions recovered. Since we cannot predict how many, or few, insertions will be recovered in future experiments, we also performed linear interpolation on the downsampled data. The resulting contour plots (Figure 2.8C-D) indicate the approximate sensitivity of BRD4-bound SE detection in HCT-116 cells. Our analysis suggests that even with as few as 10,000 insertions, we can still obtain sensitivities around 50%. Similarly, we investigated the reproducibility of SP1-directed peaks at a various downsampled numbers of insertions, using the peaks obtained from our bulk SP1-HyPBase experiment as our reference set (Figure 2.8E-F). We found that peak detection is directly proportional to the number of SP1-directed insertions recovered. At a lower limit of 10,000 insertions in both the experimental and control datasets, there was 40% overlap with peaks called from our bulk dataset. Together, these analyses provide a guide for how well calling cards will perform in the limit of insertion recovery.

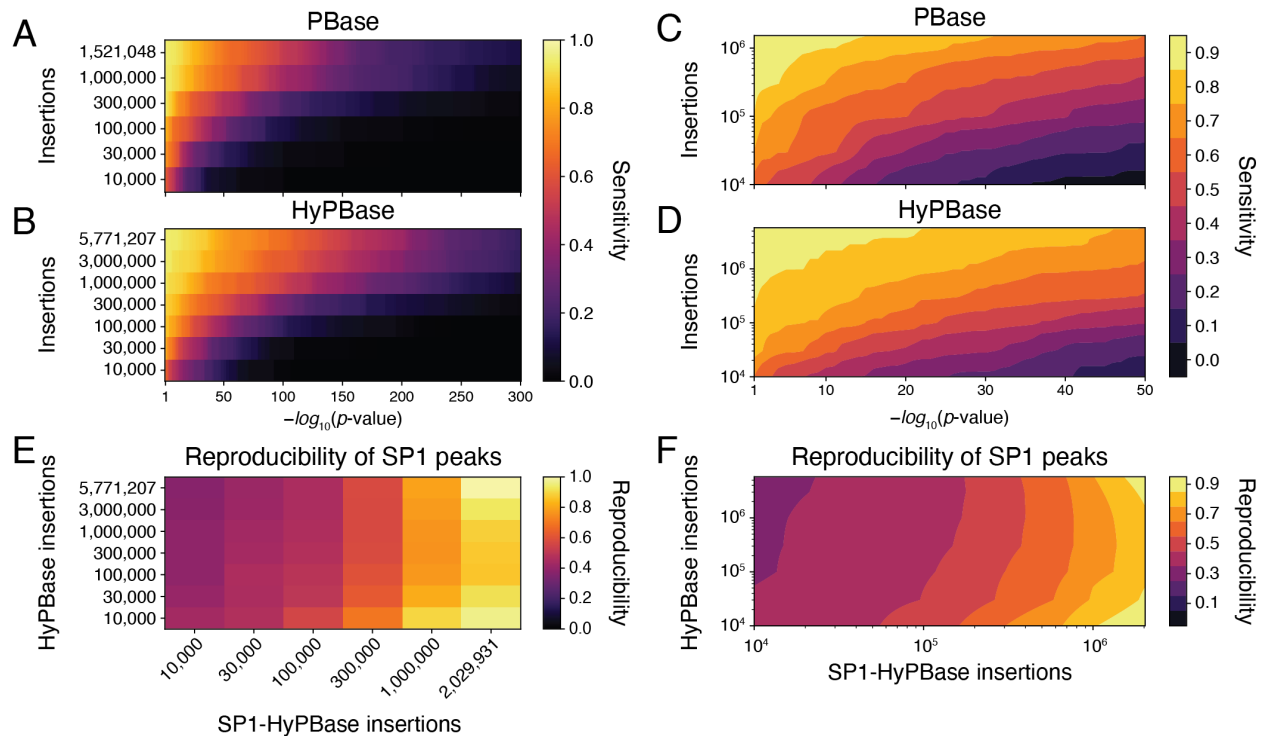


Figure 2.8: Downsampling undirected and directed *piggyBac* insertions simulates assay performance. (A) Downsampling analysis of BRD4-bound SE detection by PBBase insertions (in HCT-116 cells) at various p -value thresholds. (B) Downsampling analysis in (A) applied to HyPBase insertions. (C) Linear interpolation applied to (A) to predict SE sensitivity across a range of insertions. (D) Linear interpolation applied to (B). (E) Reproducibility of bulk SP1 calling card peaks at various numbers of HyPBase and SP1-HyPBase insertions, relative to the full dataset (top right corner). (F) Linear interpolation applied to (E) to predict peak reproducibility across a range of experimental and control insertions.

piggyBac's baseline preference for BRD4 raises questions about how efficiently TF-*piggyBac* fusions can redirect insertions near TF binding sites. We further analyzed the bulk SP1 directed experiments and found that SP1-*piggyBac* increased insertion density at SP1-bound, BRD4-depleted regions by five- to seven-fold, on average (Figure 2.9A, C). We also saw a decrease in insertion density at non-SP1-bound BRD4 peaks on the order of 30 to 50 percent (Figure 2.9B, D). This suggests that, while the reduction of signal at BRD4-bound loci may be modest, the redirection to TF binding sites can be quite stark, explaining how TF binding sites can be accurately identified (Wang et al., 2012a). In contrast to *piggyBac*, *Sleeping Beauty* has a more uniform background distribution of insertions (Figure 2.10), which suggests that the latter

transposon system might be even more redirectable and allow us to perform TF-directed calling cards without the need for an undirected transposase control. Unfortunately, direct fusions of TFs to *Sleeping Beauty* almost completely abolish transposase activity (Wu et al., 2006). We confirmed this in a colony formation assay with SP1 fused to either *piggyBac* or *Sleeping Beauty*. The SP1-*Sleeping Beauty* fusion had virtually undetectable levels of transposition, whereas the SP1-*piggyBac* construct was still enzymatically functional (Figure 2.11). Currently, *piggyBac* remains the practical choice for mammalian calling cards, while the prospect of a background-free calling card strategy should motivate future research.

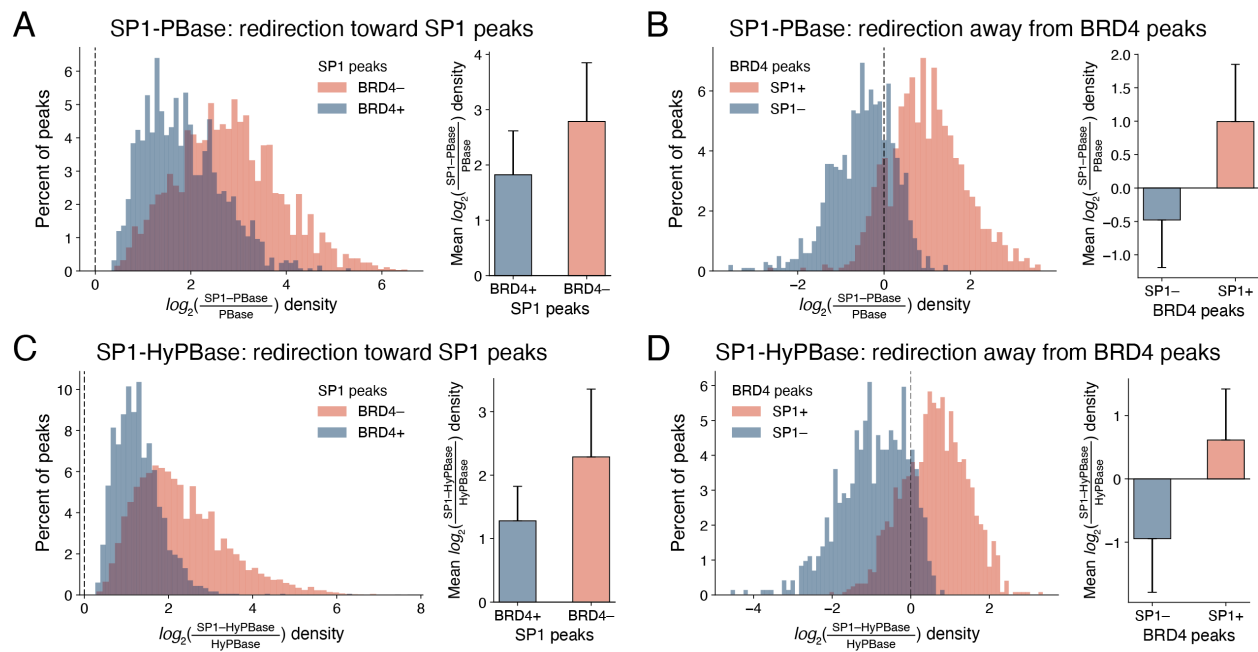


Figure 2.9: Redirectability of SP1-*piggyBac* fusion constructs. (A) Left: distribution of insertion densities at SP1-PBase peaks that either overlap, or do not overlap, BRD4-directed PBase peaks (BRD4+ and BRD4-, respectively) in HCT-116 cells. Right: mean and SD of distributions. (B) Left: distribution of insertion densities at BRD4-directed, PBase peaks that either overlap, or do not overlap, SP1-Pbase peaks (SP1+ and SP1-, respectively). Right: mean and SD of distributions. (C-D) Similar analysis as (A-B) applied to the SP1-HyPBase and HyPBase datasets, respectively. SD: standard deviation.

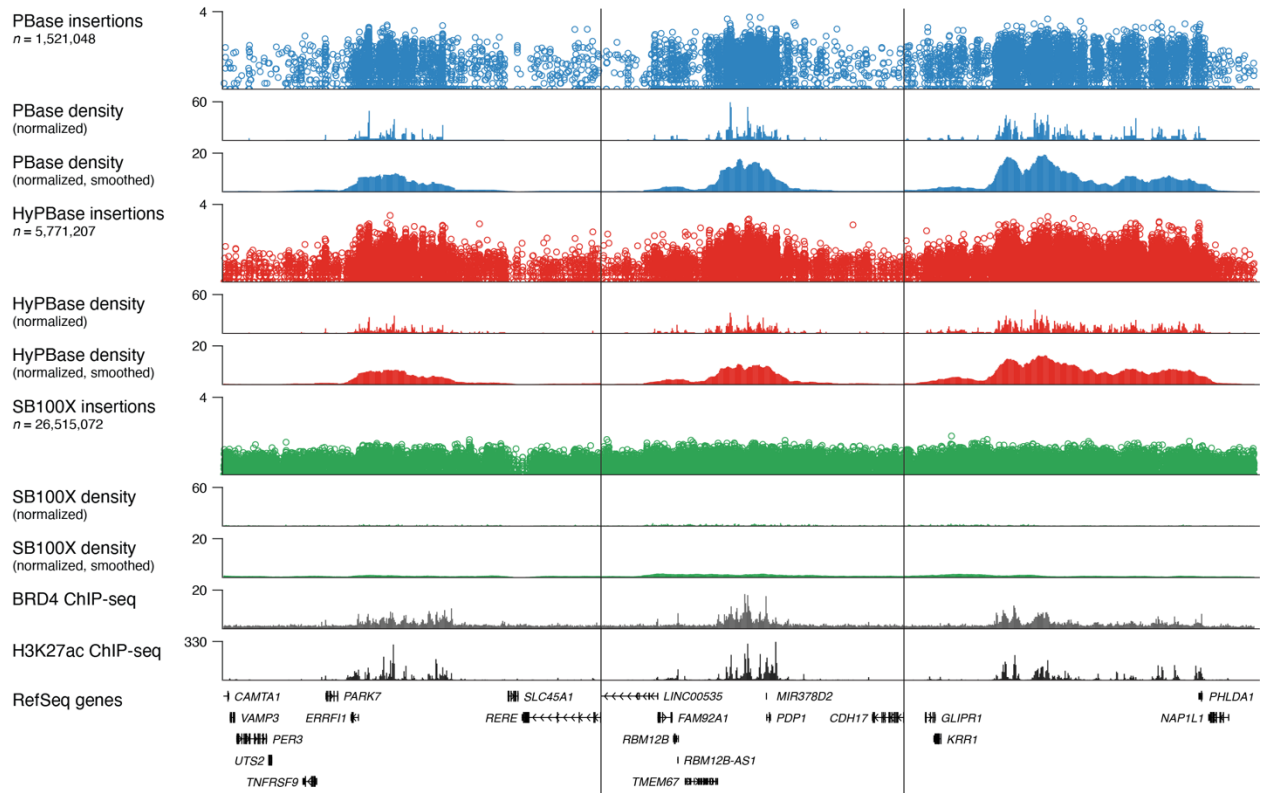


Figure 2.10: Examples of BRD4-bound super-enhancers identified by bulk PBase and HyPBase calling cards in HCT-116 cells. Three different loci exhibiting non-uniform densities of *piggyBac* insertions correlated with BRD4 and H3K27ac ChIP-seq data. *Sleeping Beauty* insertions at those same loci are more uniformly distributed. Density tracks are shown before and after smoothing.

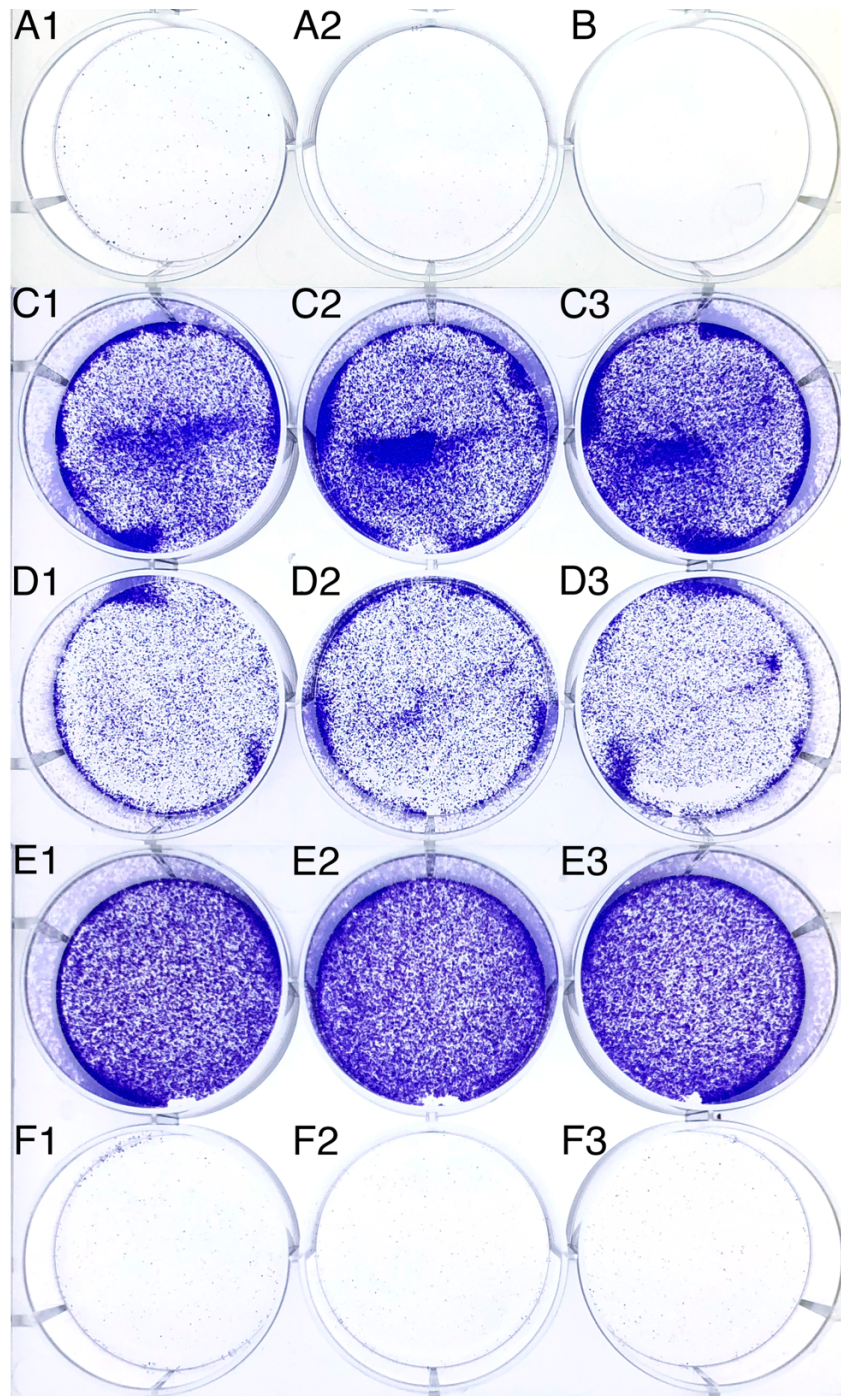


Figure 2.11: *piggyBac* is more tolerant of transcription factor fusions than *Sleeping Beauty*. Colony formation assays of HCT-116 cells transfected with the specified construct(s), selected with puromycin, and stained with crystal violet. Numbers indicate biological replicates. (A) PB-SRT-Puro (B) Untransfected (no DNA). (C) PB-SRT-Puro and hyperactive *piggyBac* transposase (HyPBase). (D) PB-SRT-Puro and SP1 fused to hyperactive *piggyBac* (SP1-HyPBase). (E) SB-SRT-Puro and hyperactive *Sleeping Beauty* (SB100X). (F) SB-SRT-Puro and SP1 fused to hyperactive *Sleeping Beauty* (SP1-SB100X).

Lastly, we investigated how similar undirected *piggyBac* transposition is to that of Tn5, the transposase at the heart of ATAC-seq (Buenrostro et al., 2013, 2015), which preferentially inserts into open chromatin. Since BRD4 and H3K27ac tend to accumulate at accessible chromatin, 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 SE's with high sensitivity from ATAC-seq data alone, much as we have shown for *piggyBac*. We took publicly available ATAC-seq data from HCT-116 cells (Ponnaluri et al., 2017) and called “super-enhancers” 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—of the 162 SEs in our gold standard set, only 1 was identified through our ATAC-seq analysis (Figure 2.12A). Moreover, there are a small number (4.3%) of *piggyBac* peaks that are not found in accessible chromatin (Figure 2.12B), suggesting that there are regulatory elements in closed chromatin that calling cards is better able to detect. Globally, we find that over 20% of Tn5 insertions are directed to accessible sites, as defined by DNase-seq, while undirected *piggyBac*, hyperactive *piggyBac*, and *Sleeping Beauty* show starkly decreased affinities for such loci (Figure 2.12C). Interestingly, fusing SP1 to *piggyBac* appears to rescue this behavior, highlighting the efficacy of transposase redirection to TF binding sites which tend to fall in accessible chromatin. Finally, we examined BRD4 ChIP-seq signal at DNase-seq peaks, ATAC-seq peaks, and undirected *piggyBac* peaks (Figure 2.12D). We find that *piggyBac* peaks are an order-of-magnitude larger than either DNase- or ATAC-seq peaks and capture more BRD4 binding than either of the other two assays. We conclude that, as expected, unfused *piggyBac* reflects BRD4's binding preferences, whereas Tn5 reports on all accessible chromatin; as a result, undirected calling cards is not equivalent to ATAC-seq.

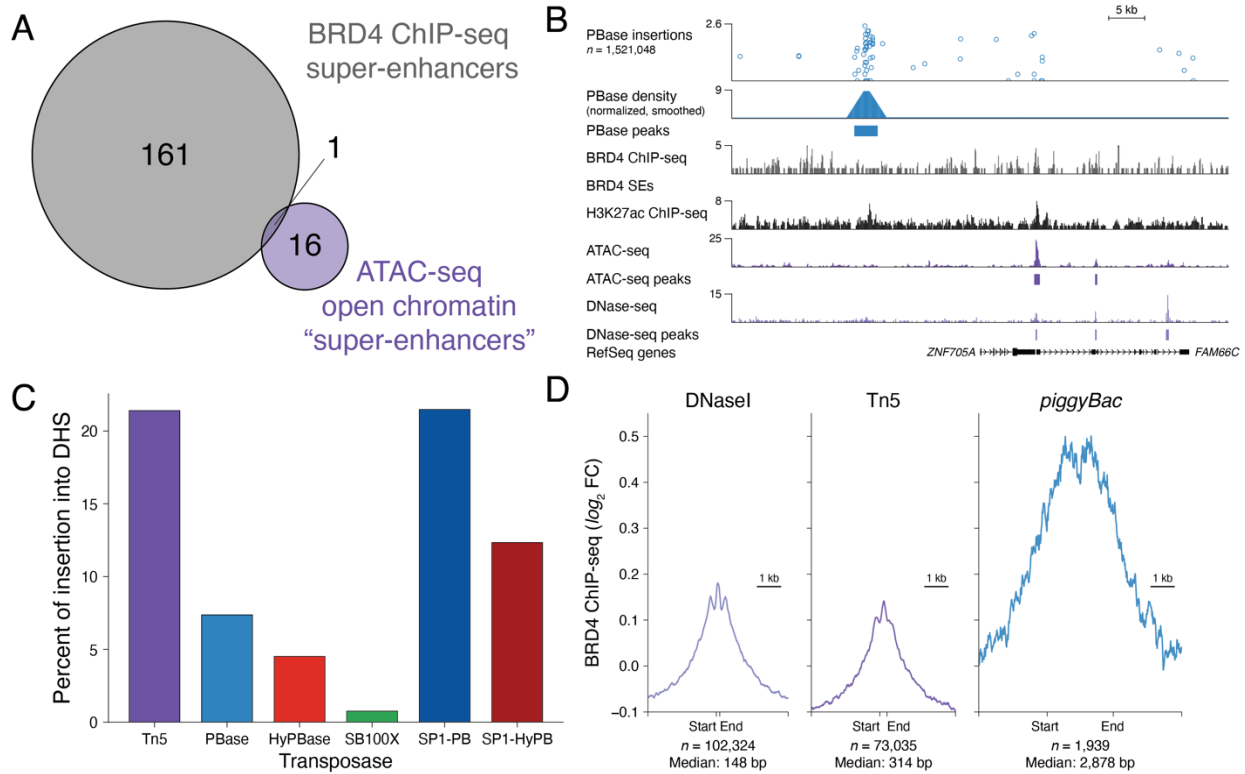


Figure 2.12: BRD4 calling cards with undirected *piggyBac* is not equivalent to ATAC-seq. (A) Overlap of BRD4 super-enhancers, as inferred from BRD4 ChIP-seq, and “super-enhancers” inferred from open chromatin ATAC-seq peaks in HCT-116 cells. (B) Browser view of a BRD4 calling card peak that is not detected 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: DNaseI hypersensitivity site; FC: fold change; kb: kilobase.

2.3.4 scCC simultaneously identifies cell type and cell type-specific BRD4 binding sites

We next sought to recover SRTs from scRNA-seq libraries. This would let us identify cell types from transcriptomic clustering and, using the same source material, profile TF binding in those cell types. We adopted the 10x Chromium platform due to its high efficiency of cell and transcript capture as well as its ease of use (Zheng et al., 2017). Like many microfluidic scRNA-seq approaches (Klein et al., 2015; Macosko et al., 2015), the cell barcode and unique molecular index (UMI) are attached to the 3' ends of transcripts. This poses a molecular challenge for SRTs since the junction between the transposon and the genome may be many kilobases away,

precluding the use of high-throughput short read sequencing. To overcome this barrier, we developed a circularization strategy to physically bring the cell barcode in apposition to the insertion site (Figure 2.13A).

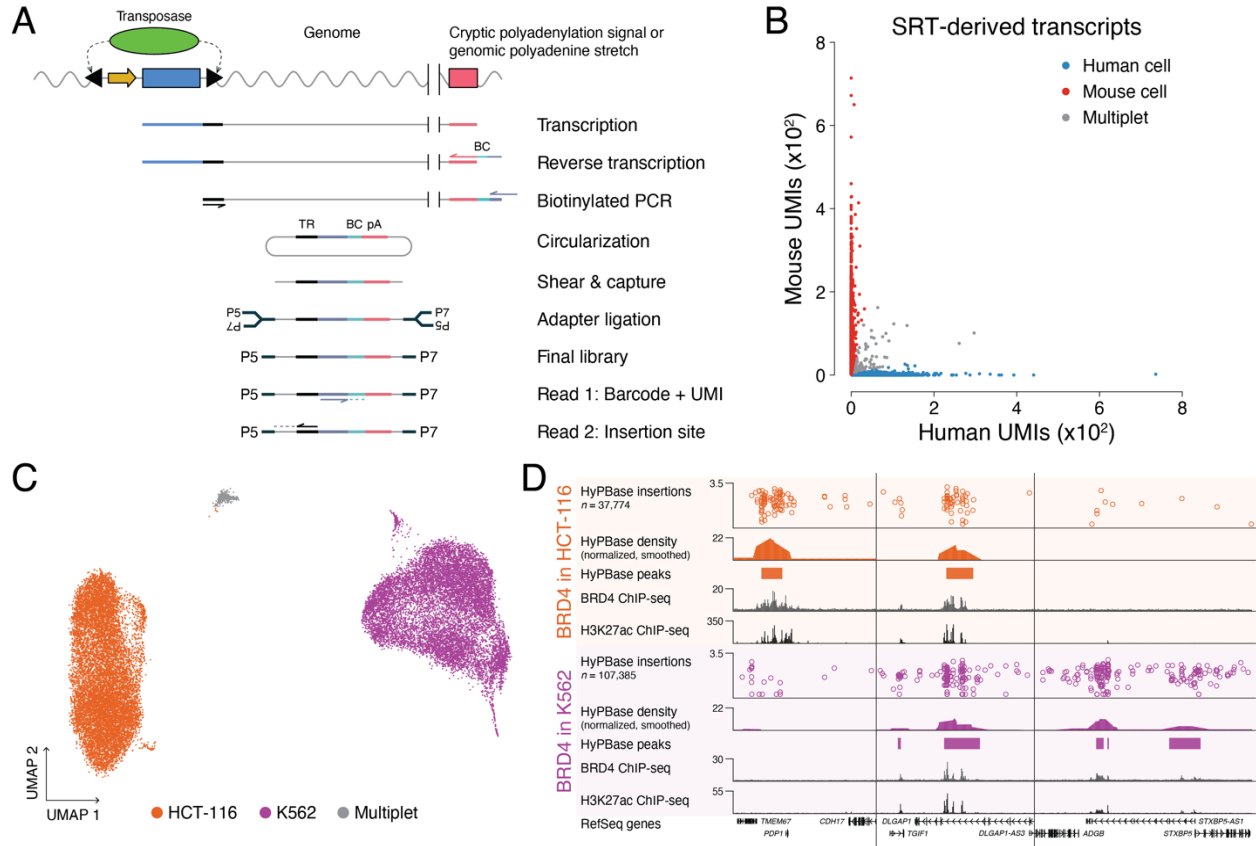


Figure 2.13: Single cell calling cards (scCC) maps BRD4 binding in single cells. (A) Schematic of the scCC library preparation strategy from scRNA-seq libraries. (B) Barnyard plot of scCC on a mixture of human HCT-116 and mouse N2a cells. (C) UMAP of scRNA-seq of a mixture of human HCT-116 and K562 cells. (D) Browser view of BRD4 peaks specific to HCT-116 and K562 cells deconvolved using scCC. See also Figures S2 and S3. TR: terminal repeat; BC: barcode; pA: poly(A) sequence; UMI: unique molecular index.

We used a modified version of the bulk SRT amplification protocol where we amplified with primers that bound to the universal priming sequence next to the cell barcode and the terminal sequence of the *piggyBac* TR. These primers were biotinylated and carried a 5' phosphate group. The PCR products of this amplification were diluted and allowed to self-ligate overnight. They were then sheared and captured with streptavidin-coated magnetic beads. The rest of the library was prepared on-bead and involved end repair, A-tailing, and adapter ligation.

A final PCR step added the required Illumina sequences for high-throughput sequencing. The standard Illumina read 1 primer sequenced the cell barcode and UMI, while a custom read 2 primer, annealing to the end of the *piggyBac* 5' TR, sequenced into the genome. Thus, we collected both the location of a *piggyBac* insertion as well as its cell of origin. We call this method single cell calling cards (scCC).

We validated the method by performing a species-mixing experiment with human HCT-116 cells and mouse N2a cells transfected with HyPBase and PB-SRT-Puro. Cells were cultured independently and mixed prior to droplet generation. The resulting emulsion was processed through first strand synthesis using the standard 10x Chromium 3' protocol. We then took half of the RT and finished preparing scRNA-seq libraries from it. The resulting analysis revealed strong species separation with an estimated multiplet rate of 3.2% (Figure 2.14A). The remainder of the first strand synthesis product was used for the scCC protocol. We restricted our calling card analysis to those insertions whose cell barcodes were observed in the scRNA-seq library (Table 2.2). The distribution of insertions across these cells reflected a continuum from pure mouse to pure human (Figure 2.14B-C). Since intramolecular ligation and subsequent PCR may introduce unwanted artifacts, such as the mis-assignment of a barcode from an N2a cell to an insertion site in an HCT-116 cell, we required that a given insertion in a given cell must have at least two different UMIs associated with it. Imposing this filter improved the number of pure mouse and human cells (Figure 2.14D), yielding clear species separation with an estimated multiplet rate of 7.9% (Figure 2.13B). This establishes that our method can accurately map calling card insertions in single cells.

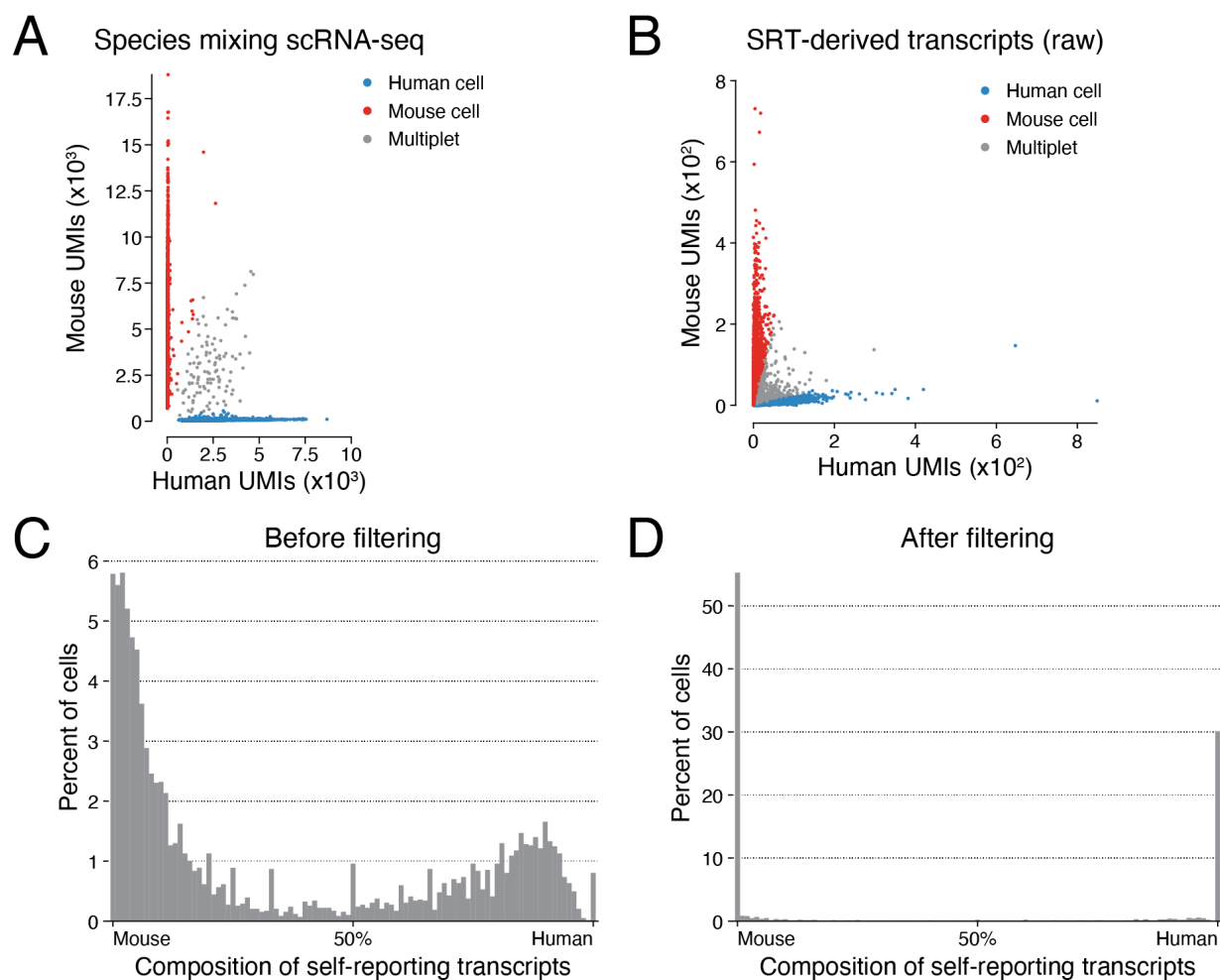


Figure 2.14: Filtering single cell SRTs reduces intermolecular artifacts. (A) Barnyard plot from scRNA-seq of human HCT-116 and mouse N2a cells. (B) Barnyard plot from scCC of HCT-116 and N2a cells without filtering (estimated multiplet rate of 25.1%). (C) Distribution of cell barcode purity from unfiltered 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.

Table 2.2: Summary of single cell calling cards experiments

Sample	Construct	Libraries	Cells	Insertions	Reads	Mean coverage	Mean IPC	Median IPC	% cells with ≥ 1 insertion
HCT-116 & N2a	HyPBase	1 ^a	6,068	33,223	1,710,525	51.5	5.4	4	91.8
HCT-116	HyPBase	4 ^b	12,891	37,774	4,768,230	126.2	3.0	2	93.4
K562	HyPBase	4 ^b	11,912	107,385	10,404,042	96.9	9.5	6	96.9
HCT-116	SP1-	4	30,411	77,210	9,874,157	127.9	2.6	2	83.8
K562	HyPBase	4	21,554	327,465	44,851,070	137.0	15.3	9	95.8
	SP1-								
HepG2	HyPBase	3	17,195	144,176	20,035,606	139.0	8.4	6	96.1
HepG2	FOXA2-	3	16,623	105,000	15,677,152	149.3	6.3	4	96.0
	HyPBase								
OCM-1A	HyPBase	3	23,978	150,707	19,794,848	131.3	6.3	4	96.2
OCM-1A	BAP1-	3	19,572	215,330	27,666,808	128.5	11.0	7	97.6

Mouse	HyPBase								
cortex	HyPBase	9 ^c	35,950	111,382	12,204,369	109.6	3.1	3	73.7

^a This library was from a species-mixing experiment (Figures 2.13B and 2.14). ^b These libraries were demultiplexed from a cell line-mixing experiment (Figures 2.13C-D and 2.15). ^c This experiment is further stratified by cell type in Table 2.3. IPC: insertions per cell.

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 identified the two major cell populations (Figure 2.13C; Figure 2.15A). 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. We obtained 37,774 insertions from 12,891 HCT-116 cells and 107,385 insertions from 11,912 K562 cells (Table 2.2). The distribution of insertions per cell varied by cell type (Figure 2.15D) and does not appear to be correlated with differences in total RNA content (Figure 2.15B-C). Over 93% and 96% of HCT-116 and K562 cells, respectively, had at least one insertion event (Table 2.2). Using scCC insertion data alone, we called peaks and successfully identified BRD4-bound loci that were specific to HCT-116 cells, shared between HCT-116 and K562, and specific to K562 cells, respectively (Figure 2.13D). Both HCT-116 and K562 peaks showed statistically significant enrichment for BRD4 ChIP-seq signal over randomly permuted peaks ($p < 10^{-9}$ in both instances, Kolmogorov-Smirnov test; Figure 2.15E-F). Furthermore, 57% of HCT-116 peaks and 81% of K562 peaks were specifically bound in their respective cell type. From our earlier downsampling analysis, we estimated that with a p -value cutoff of 10^{-9} , our sensitivity for detecting BRD4-bound SEs would be approximately 60% (Figure 2.8D). The actual sensitivity at this level of recovery was 67%, validating that downsampling analysis can reasonably estimate the performance of scCC. To conclude, we investigated the reproducibility of the scCC method. Single cell HyPBase insertions showed high concordance between

biological replicates at statistically significant peaks in both HCT-116 and K562 cells ($R^2 = 0.91$ and 0.94, respectively; Figure 2.15G-H). In all, these experiments demonstrate that scCC can be used to deconvolve cell type specific BRD4 binding.

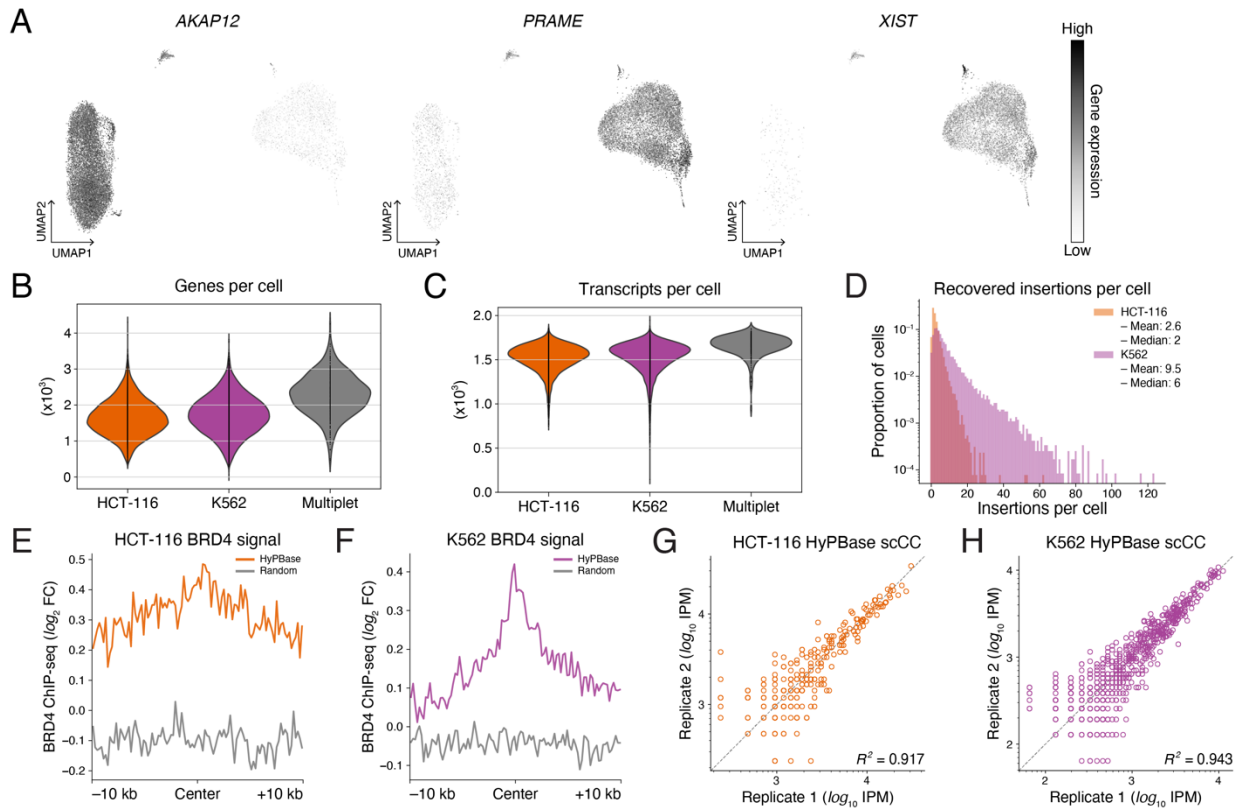


Figure 2.15: Validation and performance of undirected *in vitro* single cell calling cards (scCC). (A) Single cell expression levels of three marker genes in a mixed scRNA-seq library of human HCT-116 and K562 cells. (B) Distributions of genes per cell by cell type. (C) Distributions of transcripts per cell by cell type. (D) Distributions of HyPBases insertions recovered per cell in HCT-116 and K562 cells. (E-F) Mean BRD4 ChIP-seq signal at HyPBases 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 HyPBases and recovered by scCC at BRD4 binding sites in HCT-116 and K562 cells, respectively. KS: Kolmogorov-Smirnov.

2.3.5 scCC identifies binding sites across a spectrum of TFs and in a variety of cell types

Our success mapping BRD4 binding in single cells gave us confidence that we would also be able to map TF binding with scCC. We tested the SP1-HyPBases construct in both HCT-116 and K562 cells. We recovered 77,210 insertions from 30,682 HCT-116 cells and 327,465 insertions from 21,554 K562 cells (Table 2.2). We used the data generated when we mapped BRD4

binding in these cells as control datasets. As was observed in bulk (Figure 2.5A), SP1-HypBase-directed insertions recovered from single cells localized to SP1 binding sites in both HCT-116 and K562 cells (Figure 2.16A and 2.16E). In both cell lines, we observed significant enrichment of SP1 ChIP-seq signal at peaks (Figure 2.16B-C and Figure 2.16F-G) and motif analysis identified the SP1 DNA binding motif (Figure 2.16D and Figure 2.16H) ($p < 10^{-30}$ in each instance). Finally, as with bulk SP1 calling cards (Figure 2.16D and Figure 2.5E), scCC SP1 calling cards showed significant enrichments for insertions near TSSs, CpG islands, and unmethylated CpG islands (G test of independence $p < 10^{-9}$ in each instance; Figure 2.17A-B).

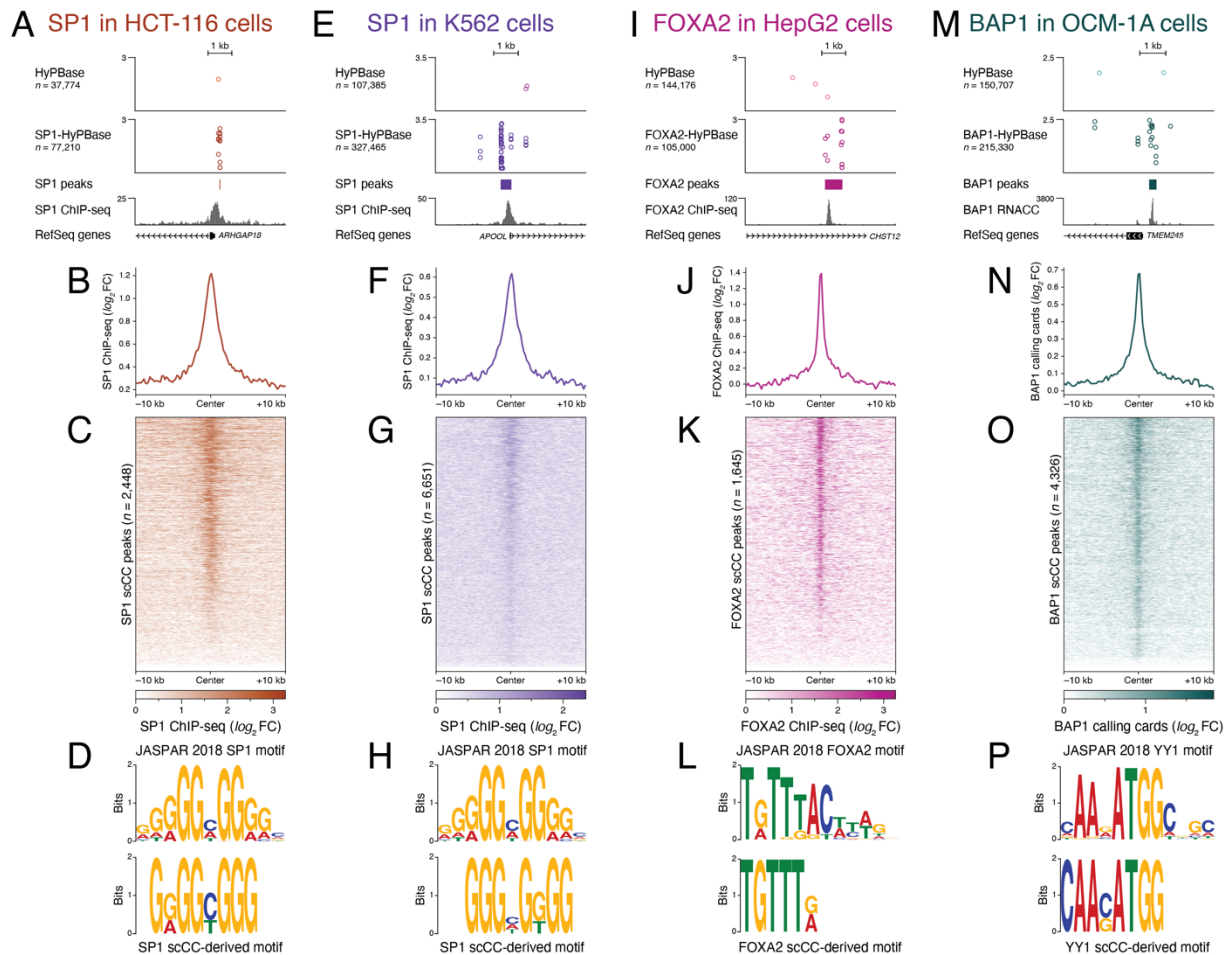


Figure 2.16: Single cell calling cards (scCC) works with a variety of transcription factors (TFs) and cell lines. (A-D) scCC with SP1-HypBase in HCT-116 cells reveals SP1 binding sites. (A) Browser view of a peak from SP1 scCC. (B) Mean SP1 ChIP-seq signal at scCC SP1 peaks. (C) Heatmap of SP1 ChIP-seq 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-L) scCC with FOXA2-HyPBase in HepG2 cells reveals 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. (M-P) scCC with BAP1-HyPB in OCM-1A cells reveals BAP1 binding sites. (M) Browser view of a peak from BAP1 scCC. (N) Mean bulk BAP1 calling cards signal at scCC BAP1 peaks. (O) Heatmap of bulk BAP1 calling cards signal across all scCC BAP1 peaks. (P) YY1 motif elicited from BAP1 scCC peaks. See also Figure S4. FC: fold change

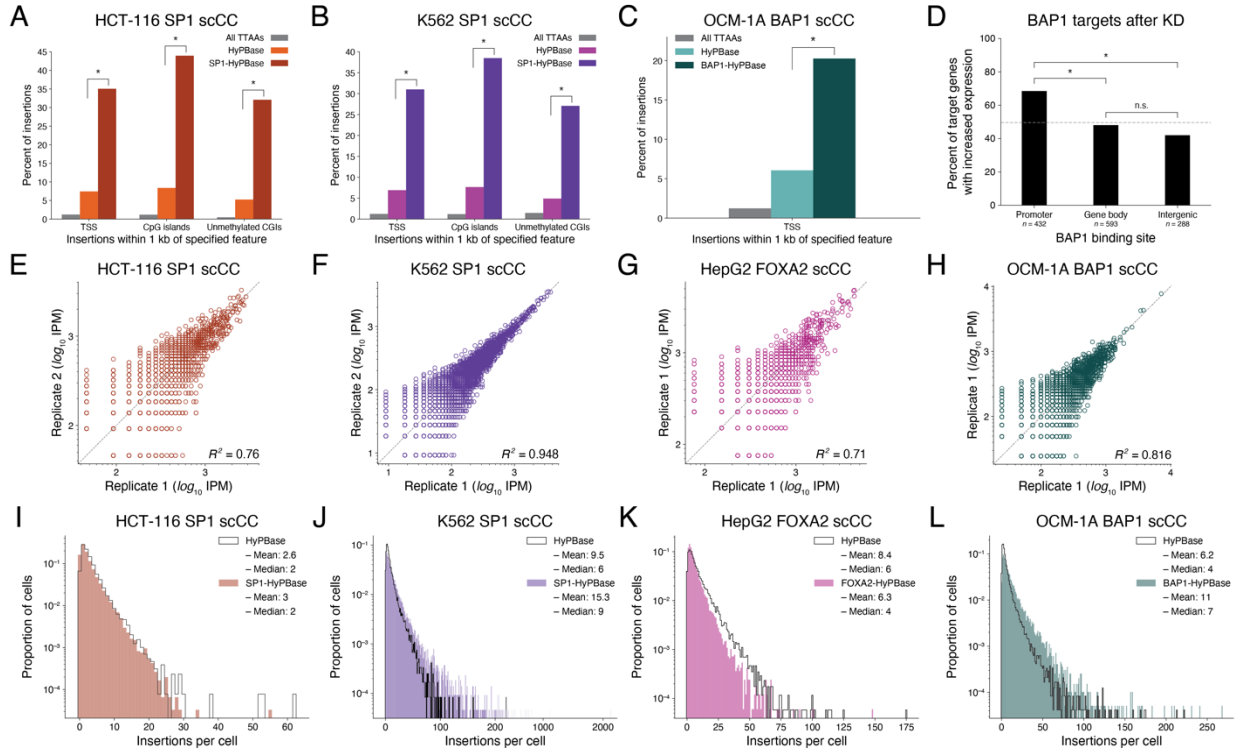


Figure 2.17: Validation and performance of TF-directed *in vitro* single cell calling cards (scCC). (A-B) Enrichment of SP1-HyPBase-directed insertions to TSSs, CGIs, and unmethylated CGIs in single HCT-116 and K562 cells, respectively (G test of independence $p < 10^{-9}$). (C) Enrichment of BAP1-HyPBase-directed 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 vs 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.

We next performed scCC in HepG2 cells with the pioneer factor FOXA2, which has been shown to be required for normal liver development and drives core transcriptional networks in cancer cells (Fournier et al., 2016; Lee et al., 2005). We mapped 144,176 undirected HyPBase insertions in 17,195 HepG2 cells, and from a further 16,623 cells we recovered 105,000 FOXA2-HyPBase insertions (Table 2.2). As with SP1, we observed a specific enrichment of insertions at

FOXA2 binding sites (Figure 2.16I). Peaks called from scCC FOXA2 data were enriched in FOXA2 ChIP-seq signal (Figure 2.16J-K), and motif analysis was able to infer the core FOXA2 DNA binding motif (Figure 2.16L).

Lastly, we mapped the binding of BAP1 in the uveal melanoma cell line OCM-1A (Yen et al., 2018) using scCC. Unlike SP1 and FOXA2, BAP1 does not bind DNA directly; instead, it is drawn to chromatin in a complex by cofactors (Carbone et al., 2013; Yu et al., 2010) where it acts as a histone deubiquitinase. We retrieved 150,707 undirected HyPBase insertions from 23,978 OCM-1A cells and 215,330 BAP1-HyPBase insertions from another 19,572 cells (Table 2.2). Despite the fact that this protein associates with chromatin indirectly, we were able to resolve sharp BAP1-directed peaks (Figure 2.16M). These peaks showed high concordance with bulk RNA calling card data that we also generated in this system (Figure 2.16N-O; Table 2.1). Sequence analysis elicited the motif of YY1 (Figure 2.16P), 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); accordingly, we observed a significant enrichment for BAP1-directed insertions near TSS (G test of independence $p < 10^{-9}$; Figure 2.17C). While BAP1 is a member of the Polycomb repressive complex, there are conflicting reports as to whether it activates or represses gene expression (Campagne et al., 2019; Matatall et al., 2013; Yu et al., 2010). We cross-referenced our single cell BAP1 peaks against published RNA-seq data obtained in unperturbed and BAP1 knockdown OCM-1A cells (Yen et al., 2018). Genes where BAP1 is bound at the promoter, as opposed to in the gene body or at a nearby intergenic locus, are significantly more likely to have increased expression upon BAP1 knockdown (Fisher's exact test $p < 10^{-9}$; Figure 2.17D). This suggests that, in OCM-1A cells, promoter-bound BAP1 primarily acts as a repressor of gene expression.

Collectively, these results indicate that single cell calling cards 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; Figure 2.17E-H). While TF-*piggyBac* fusions have been reported to decrease transposase activity (Wu et al., 2006), our findings were more equivocal: BAP1 and SP1 fusions in K562 cells showed greater activity than undirected HyPBase; FOXA2 showed decreased activity; and SP1 in HCT-116 cells showed roughly the same activity (Figure 2.17I-L). This suggests that there may be some variability in the number of recovered insertions depending on the TF and cell type of interest but, overall, the method is robust.

2.3.6 scCC reveals bromodomain-dependent cell state dynamics in K562 cells

K562 is a chronic myelogenous leukemia (CML) cell line first isolated in 1970 (Lozzio and Lozzio, 1975) and has been a workhorse of molecular biology ever since (Zhou et al., 2019). Recently, K562 cultures have been shown to be mixtures of a stem-like state characterized by high levels of the surface marker CD24, and a more differentiated, erythroleukemic state marked by low CD24 expression, with individual cells dynamically oscillating between these two extremes (Litzenburger et al., 2017). Since we profiled BRD4 binding in K562 cells with scCC, we wondered whether we could see evidence of these two states in the scRNA-seq data. Principal components analysis (PCA) of single cell gene expression (Figure 2.18A) 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 2.18B), suggesting mutually exclusive states. Scoring single cells on a subset of top PC genes revealed a gradient of cell states along a stem-like-to-differentiated

axis (Figure 2.19A). We then clustered cells on the basis of this state score to define stem-like and differentiated populations (Figure 2.18C-D), which faithfully recapitulate the expression differences detected by PCA (Figure 2.18E).

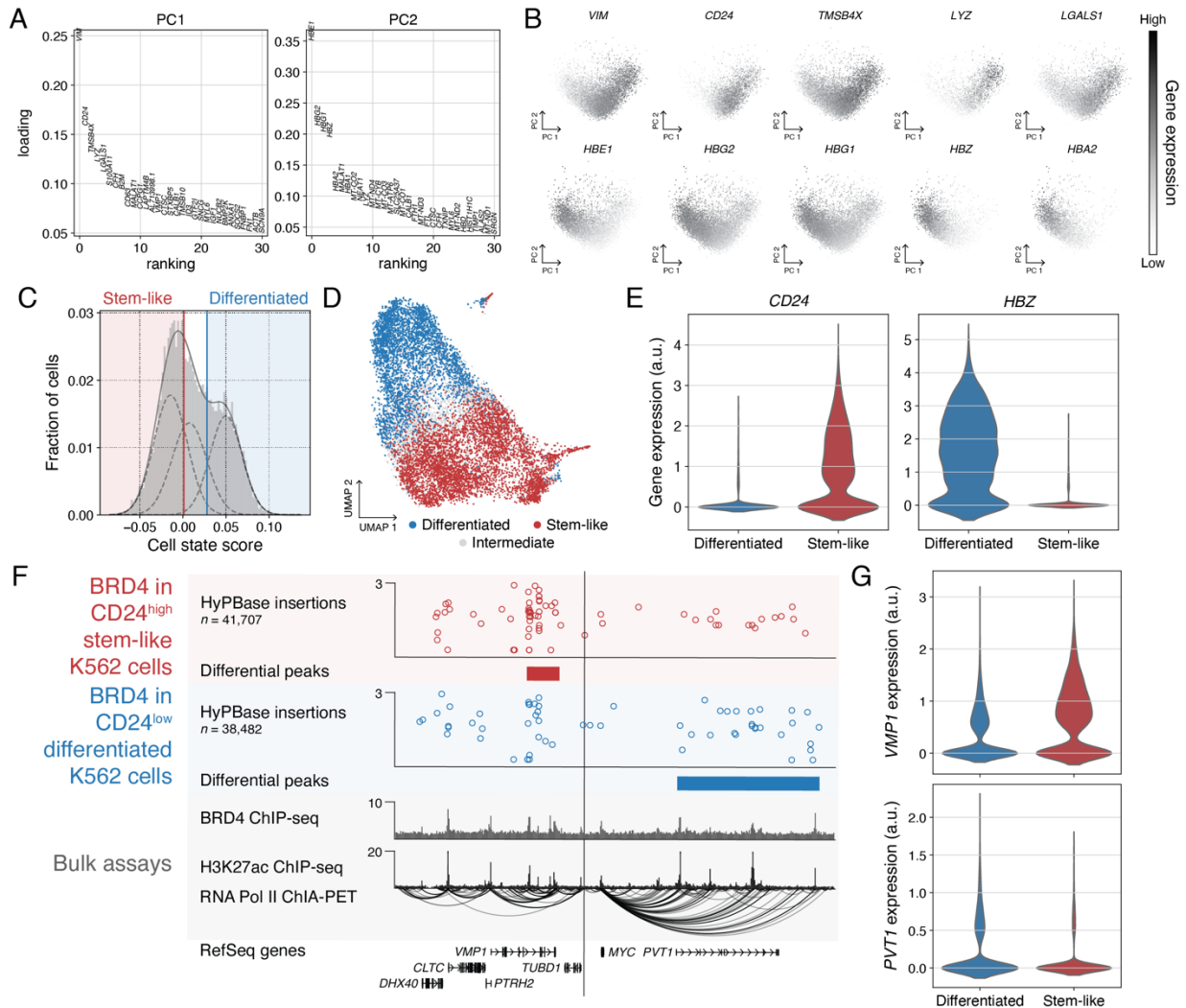


Figure 2.18: Clustering of K562 cells into stem-like and differentiated states. (A) Principal component analysis of K562 scRNA-seq data. (B) Relative expression levels of highest-ranking genes in PC1 (top) and PC2 (bottom). (C) Gaussian mixture modeling of a cell-state score to define stem-like and differentiated K562 clusters. (D) Visualization of assigned cell clusters in the UMAP projection. (E) Specific expression of *CD24* and *HBZ* in the stem-like and differentiated clusters, respectively. (F) Genome browser view of scCC 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 *VMP1* and *PVT1* in the stem-like and differentiated clusters. PC: principal component.

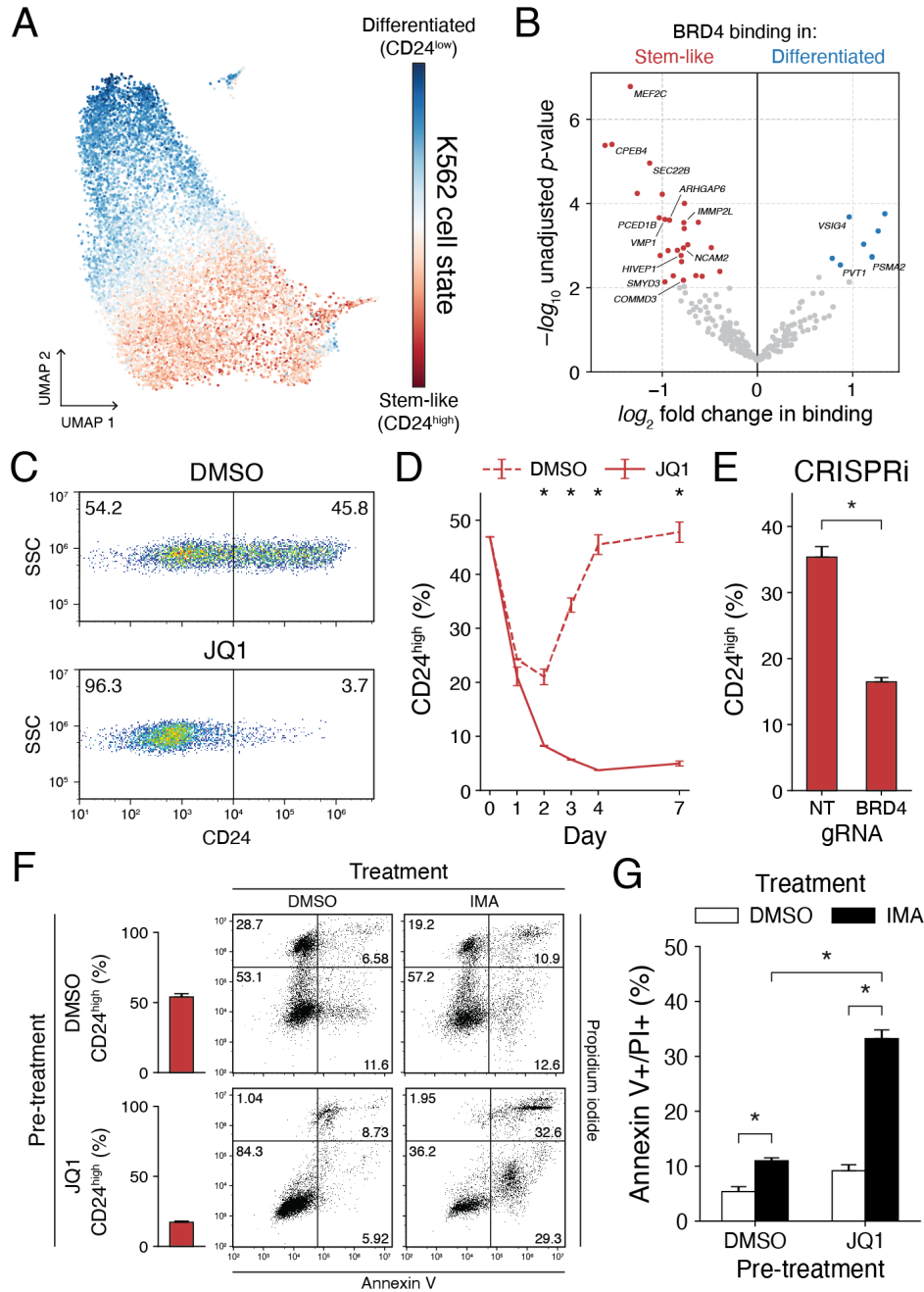


Figure 2.19: Single cell calling cards uncovers bromodomain-dependent cell state dynamics in K562 cells. (A) Gradient of cell states from scRNA-seq analysis of K562 cells. (B) Differential BRD4 binding analysis of undirected HyPBase peaks in K562 cells. (C) Representative distributions of CD24^{high} and CD24^{low} cells after either 96 hours of DMSO (top) or JQ1 (bottom) treatment. (D) Proportion of CD24^{high} cells over a seven-day time course of JQ1 treatment (three-way ANOVA $p < 0.01$). (E) Proportion of CD24^{high} cells after BRD4 CRISPRi (Welch's t-test $p < 0.01$). (F) Representative plots of annexin V and PI staining in K562 cells pretreated with either DMSO or JQ1 (250 nM) and subsequently treated for 48 hours with either DMSO or imatinib (1 μ M). (G) Quantification of (F) (two-way ANOVA $p < 0.01$). Bars represent means; error bars denote standard deviations. Experiments were performed in triplicate. See also Figures S5 and S6. DMSO: dimethyl sulfoxide; SSC: side scatter; CRISPRi: CRISPR interference; NT: non-targeting; gRNA: guide RNA; IMA: imatinib; PI: propidium iodide.

Super-enhancers 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 and sharply delineated disease states (Hnisz et al., 2013; Whyte et al., 2013), recent studies have shown that even closely related subpopulations of the same cell type can show subtle changes in BRD4 enrichment and enhancer utilization (Knoechel et al., 2014; Rathert et al., 2015). Therefore, we asked whether we could detect any differences in BRD4 binding between CD24^{high} and CD24^{low} cells. We first stratified scCC insertions by cell state, assigning 41,707 to the stem-like state and 38,482 to the differentiated cluster (Figure 2.20F). We then analyzed the peaks generated across all K562 cells and quantified differential binding between the two clusters. Indeed, we found multiple peaks that showed significant differential binding at a false-discovery rate threshold of 10% (Figure 2.19B). 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: *VMPI*, bound more in the CD24^{high} stem-like cells; and *PVTI*, bound more in the differentiated, CD24^{low} cells (Figure 2.18F-G). *VMPI* 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. *PVTI* can act as both a tumor-suppressor and oncogene, in both instances acting on the *MYC* locus (Cho et al., 2018).

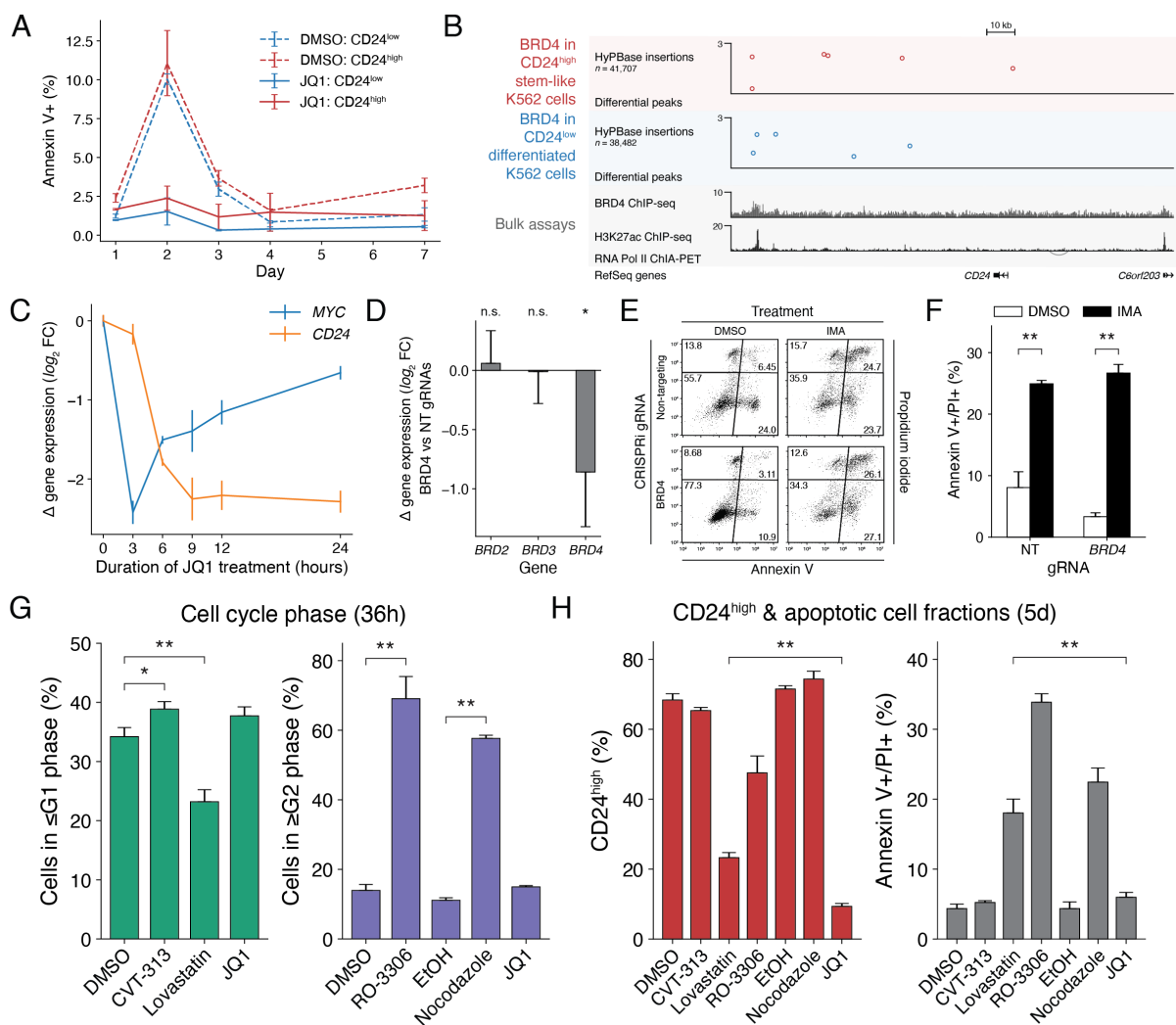


Figure 2.20: Validation of bromodomain-dependent K562 cell states. (A) Annexin V staining in CD24^{high} (red) or CD24^{low} (blue) K562 cells treated with DMSO (dashed line) or JQ1 (solid line) over a seven-day time course. (B) Genome browser view of the *CD24* locus. (C) qRT-PCR for *MYC* and *CD24* 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^{high} 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.

We next investigated whether the observed differences in BRD4 binding might be causally responsible for establishing these two cell states. Since modulation of this epigenetic

reader has been previously shown to influence cell identity across a range of tissues (Di Micco et al., 2014; Kfoury et al., 2017; Najafova et al., 2017), we hypothesized that perturbing BRD4 would change the distribution of cells in the stem-like and differentiated states. Moreover, due to the asymmetric nature of significant hits in Figure 2.19B, there is a subset of peaks specific to the CD24^{high} state that are not shared by the CD24^{low} 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. Thus, we predicted that not only should the distribution of CD24^{high}/CD24^{low} cells change upon BRD4 perturbation, but also that the stem-like CD24^{high} population should be more susceptible to such a perturbation.

To test this hypothesis, we treated cells with the small molecule bromodomain inhibitor JQ1, commonly 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). We observed that JQ1 exposure was sufficient to shift the population from one containing equal proportions of CD24^{high}/CD24^{low} cells to one comprised of almost exclusively CD24^{low} cells (> 95% CD24^{low} cells, Figure 2.19C). A time course analysis showed that this conversion takes place rapidly over the first two days, plateaus at day four, and remains stable one week after treatment; in contrast the control cells remain evenly split between the two states at this timepoint (Figure 2.19D; two-way ANOVA $p < 0.01$). We ruled out the possibility that JQ1 is selectively cytotoxic to CD24^{high} cells as there were no significant differences in levels of the early apoptotic marker annexin V between CD24^{high} and CD24^{low} cells, regardless of whether they had been exposed to JQ1 or DMSO (Figure 2.20A; three-way ANOVA $p = 0.84$). We also investigated whether *CD24* is a direct target of BRD4, which would imply that the loss of CD24^{high} cells does not reflect a true change in cell state but is, instead, a trivial transcriptional consequence of downregulating BRD4

by JQ1. To do so, we examined genomic signals at the *CD24* locus and did not find any prominent BRD4 binding sites, either by ChIP-seq or calling cards, or elevated levels of H3K27 acetylation in the vicinity of *CD24* (Figure 2.20B). We also compared the relative changes in mRNA levels of *MYC*, a known BRD4 target (Knoechel et al., 2014; Lovén et al., 2013; Rathert et al., 2015; Zuber et al., 2011), to that of *CD24* during the first 24 hours of JQ1 exposure. Whereas *MYC* levels fell within the first 3 hours of exposure, transcript levels of *CD24* decreased most precipitously somewhere between 3 and 9 hours after JQ1 induction (Figure 2.20C). This delayed response suggests that *CD24* is not a direct target of BRD4, but instead its expression changes as the result of downstream regulatory factors. These results argue that JQ1 treatment does not simply downregulate a cell surface marker, but rather perturbs transcriptional networks that ultimately include *CD24*.

While JQ1 shows greatest affinity for BRD4, it does have some promiscuity toward other bromodomains, such as those of the related bromodomain and extraterminal domain (BET) proteins BRD2 and BRD3 (Filippakopoulos et al., 2010). Thus, it was possible that the observed state shift may be arising through off-target effects and not through BRD4 itself. To address this, we specifically downregulated *BRD4* expression with CRISPRi using a dCas9-KRAB (Fulco et al., 2016; Xie et al., 2017) fusion directed to the *BRD4* locus. We confirmed, with qRT-PCR, that our *BRD4* guide RNA (gRNA) resulted in knockdown of *BRD4* and not *BRD2* nor *BRD3* (Figure 2.20D; Welch's *t*-test $p < 0.05$). As with JQ1, we observed a significant decrease in the proportion of CD24^{high} cells with the *BRD4* gRNA compared to the non-targeting (NT) gRNA (Figure 2.19E; Welch's *t*-test $p < 0.01$), though not to the same levels as JQ1. This result suggests that BRD4 is necessary for the observed cell state dynamics between CD24^{high} and CD24^{low} K562 cells, though it is likely that other bromodomains also play a role.

We next sought to obtain further evidence that bromodomain inhibition shifts K562 cell state by performing a direct functional assay. The CD24^{high}/CD24^{low} K562 cell states have been previously shown to have different chemosensitivities, with the latter population showing more apoptosis when exposed to imatinib (Litzenburger et al., 2017). We wondered whether bromodomain perturbation similarly increased imatinib sensitivity, or if its effect was restricted to modulating *CD24*. We tested this by first pre-treating K562 cells with either DMSO or JQ1 for five days. In the DMSO-treated group, the fraction of CD24^{high} cells rose to 54% on average, while the mean for JQ1-treated cells was 17% (Figure 2.19F). We then challenged each pretreatment group with either DMSO or imatinib and measured apoptosis by staining for annexin V and propidium iodide (PI). We observed a significant increase in annexin V/PI double positive cells in imatinib-treated cells over those pre-treated with DMSO (Figure 2.19F-G; two-way ANOVA $p < 0.01$), indicating that JQ1 sensitizes K562 cells to imatinib. We also found that BRD4 CRISPRi partially phenocopied this sensitization, though again not to the same effect size as JQ1 (Figure 2.20E-F; 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 2.20D), whereas JQ1, at this concentration, is expected to almost completely abolish BRD4 activity (Filippakopoulos et al., 2010). Thus, while a mild knockdown can reduce CD24 levels, a higher level of inhibition may be necessary to induce imatinib sensitivity. Nevertheless, these results establish that bromodomain inhibition functionally, in addition to phenotypically, shifts the underlying cell state of K562 cells.

Finally, we asked whether the JQ1-induced K562 cell state shift could be a non-specific response to generic drug treatment. To test this, we treated K562 cultures with cell cycle inhibitors, another class of commonly used antineoplastic agents. We used lovastatin and

nocodazole, two drugs classically used to synchronize cells in culture (Jackman and O'Connor, 1998), as well as the cyclin-dependent kinase inhibitors CVT-313 (Brooks et al., 1997) and RO-3306 (Vassilev et al., 2006). We first confirmed that all drugs perturbed cell cycle by altering the proportions of cells in either G1 or G2/M phase (Figure 2.20G). 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 $p < 0.01$). While 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 apoptotic activity (Figure 2.20H). JQ1 caused the greatest reduction in CD24^{high} cells (one-way ANOVA $p < 0.01$) and induced significantly less apoptosis than its closest competitor, lovastatin (one-way ANOVA $p < 0.01$). While all cell cycle inhibitors caused cell death, the mitotic inhibitors nocodazole and RO-3306 had very few surviving cells after five days of treatment. 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.

2.3.7 scCC deconvolves cell type-specific BRD4 binding sites in the mouse cortex

To establish broad utility for scCC, we sought to record TF binding *in vivo*. Since *in vivo* models preclude puromycin selection, we designed an SRT carrying the fluorescent reporter tdTomato (Figure 2.21A) and tested this reagent in cell culture. When this construct was transfected without transposase, merely 3.4% of cells registered as tdTomato-positive, likely due to the action of the self-cleaving ribozyme downstream of the transposon. However, when the construct was co-transfected with PBase or HyPBase, this figure rose to 33% and 48%, respectively, corresponding to 11- and 16-fold increases in signal (Figure 2.21B). In addition, cells transfected

with only the fluorescent SRT produced very few reads that mapped to the genome, while the overwhelming majority of reads from cells co-transfected with transposase mapped to genomic insertions (Figure 2.2A). Thus, this new construct, PB-SRT-tdTomato, allows us to collect cells carrying calling card insertions by fluorescence activated cell sorting (FACS).

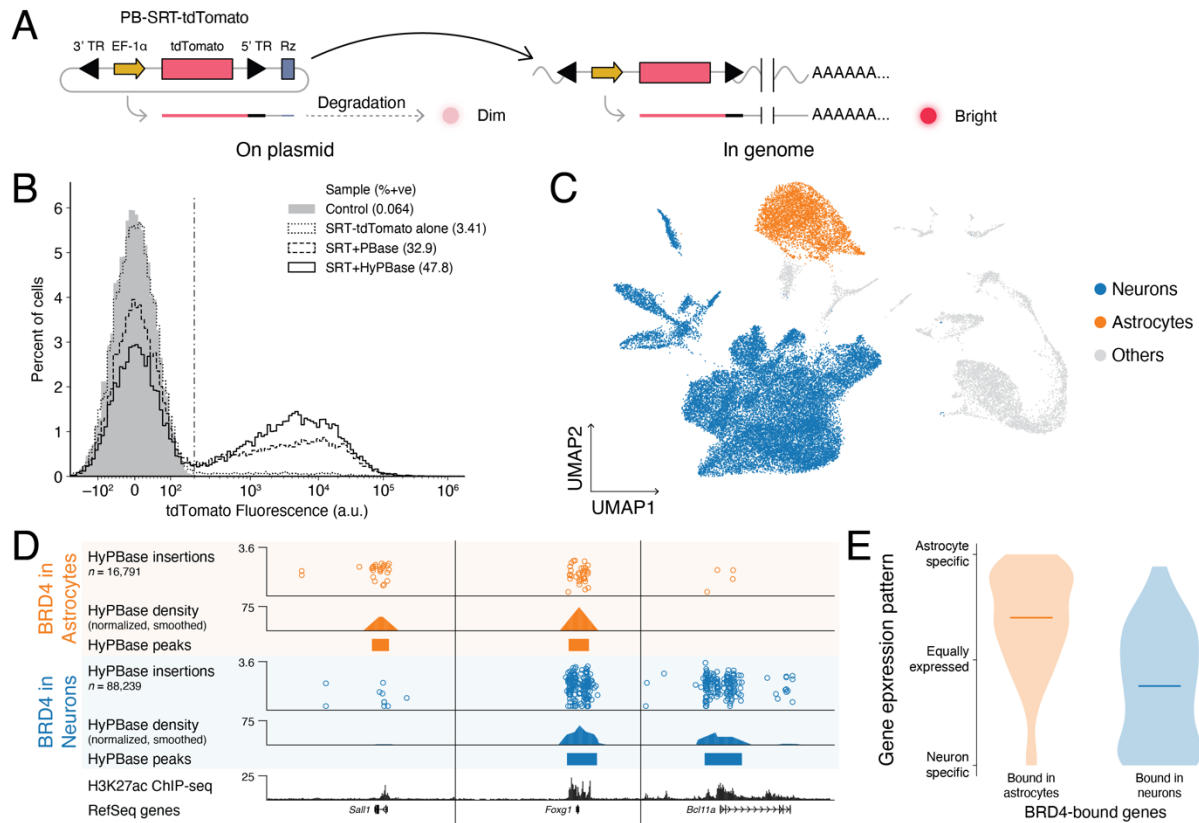


Figure 2.21: Single cell calling cards (scCC) deconvolves BRD4-bound loci in the mouse cortex. (A) Schematic of PB-SRT-tdTomato. (B) Distribution of fluorescence intensity in K562 cells transfected with PB-SRT-tdTomato with and without *piggyBac* transposase. (C) Neuron and astrocyte clusters from scRNA-seq analysis of mouse cortex libraries transduced with AAV-HyPBBase and AAV-PB-SRT-tdTomato. (D) Browser view of scCC HyPBBase 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.

We chose the mouse cortex for our *in vivo* proof-of-concept because it is a heterogeneous tissue that has been the focus of several recent single cell studies (Rosenberg et al., 2018; Saunders et al., 2018; Tasic et al., 2018; Zeisel et al., 2015, 2018). We separately packaged the PB-SRT-tdTomato and HyPBBase 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 analyzed these cells by scRNA-seq and scCC using the 10x Chromium platform. We collected nine libraries in total, encompassing 35,950 cells and 111,382 insertions (Table 2.2). We clustered cells by their mRNA profiles and used established marker genes to classify different cell types (Figure 2.22A-B) (Saunders et al., 2018; Tasic et al., 2018; Zeisel et al., 2018). Neurons and astrocytes were the two major cell populations we recovered (Figure 2.21C, Table 2.3), which is consistent with the known tropism of AAV9 (Cammack et al., 2020; Schuster et al., 2014). We also identified a spectrum of differentiating oligodendrocytes and trace amounts of microglial, vascular, and ependymal cells. We then used the cell barcodes shared between the scRNA-seq and scCC libraries to assign insertions to specific cell types.

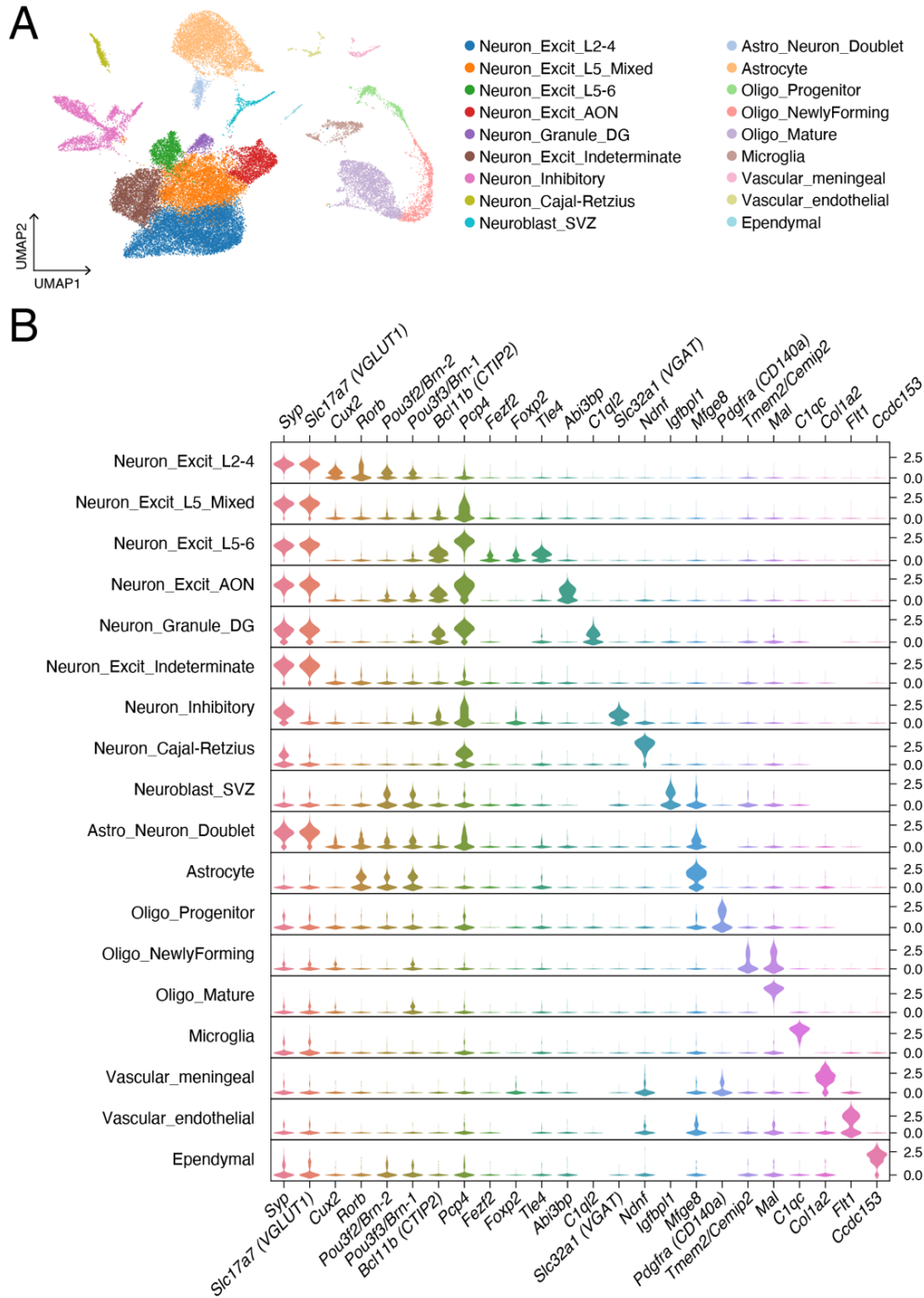


Figure 2.22: Clustering of SRT-treated cortical cells and associated marker genes. (A) Schematic of PB-SRT-tdTomato. (B) Distribution of fluorescence intensity in K562 cells transfected with PB-SRT-tdTomato with and without *piggyBac* transposase. (C) Neuron and astrocyte clusters from scRNA-seq analysis 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.

Table 2.3: Breakdown of cortical cell types and scCC HyPBase insertions per cluster

Cluster	Cells	Insertions	Mean IPC
Astrocyte	4,727	16,791	3.6
Astro_Neuron_Doublet	394	1,653	4.2
Ependymal	107	153	1.4
Microglia	569	238	0.4
Neuroblast_SVZ	369	1,084	2.9
Neuron_Cajal-Retzius	552	4,363	7.9
Neuron_Excit_AON	1,939	8,190	4.2
Neuron_Excit_Indeterminate	3,660	6,377	1.7
Neuron_Excit_L2-4	9,083	29,465	3.2
Neuron_Excit_L5	5,544	26,437	4.8
Neuron_Excit_L6	1,436	5,169	3.6
Neuron_Granule_DG	535	1,674	3.1
Neuron_Inhibitory	2,409	6,564	2.7
Oligo_Mature	2,740	1,729	0.6
Oligo_NewlyForming	959	674	0.7
Oligo_Progenitor	504	477	0.9
Vascular_endothelial	196	69	0.4
Vascular_meningeal	227	275	1.2

IPC: insertions per cell.

To determine whether scCC could recover biological differences between cell types *in vivo*, we analyzed HyPBase insertions in neurons and astrocytes, excluding neuroblasts and astrocyte-neuron doublets. We collected 88,239 insertions from 25,158 neurons and 16,791 insertions across 4,727 astrocytes (Table 2.3). We then called peaks on the insertions within each cluster and identified astrocyte-specific, neuron-specific, and shared BRD4 binding sites (Figure 2.21D). 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). Although this ChIP-seq dataset was agglomerated over all cell types in the brain, we nevertheless found that scCC peaks in both astrocytes and neurons showed statistically significant enrichment of H3K27ac signal (Figure 2.23A, C; Kolmogorov-Smirnov $p < 10^{-9}$ in each case). BRD4 is also thought to mark cell type-specific genes, so we identified genes that overlapped or that were near astrocyte or neuron peaks and evaluated the specificity of expression of these genes. 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 used bulk RNA-seq data from purified populations of cells (Zhang et al., 2014) to assign gene expression values for each gene and plotted the distribution of these values along a continuum from purely astrocytic expression to purely neuronal expression. Genes near astrocyte peaks were more likely to be specifically expressed in astrocytes, and vice-versa for genes near neuron peaks (Figure 2.16E). Gene Ontology enrichment analysis (Mi et al., 2017) on the astrocyte gene list included terms like “gliogenesis,” and “glial cell differentiation,” as well as copper metabolism (Figure 2.23B), a known function of astrocytes (Scheiber and Dringen, 2013); while the neuronal gene list was enriched for terms related to synapse assembly, axonal guidance, and neuron development (Figure 2.23D). Overall, we conclude that scCC can accurately identify cell type specific BRD4 binding sites *in vivo*.

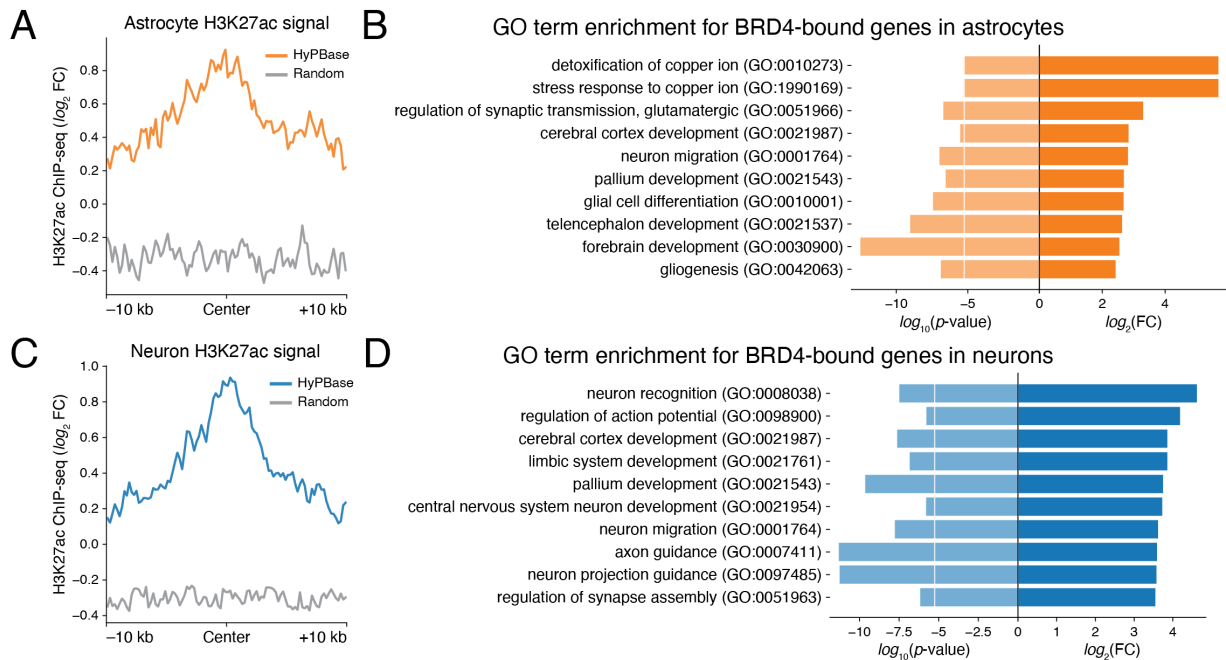


Figure 2.23: Validation of *in vivo* BRD4 binding in astrocytes and neurons. (A) Mean H3K27ac ChIP-seq signal at HyPBase peaks in astrocytes compared to randomly permuted peaks (KS test $p < 10^{-9}$). (B) GO term enrichment analysis of genes near astrocytic BRD4 binding sites. (C) Mean H3K27ac ChIP-seq 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.

Finally, we wondered if 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 (Figure 2.22A-B; Figure 2.24B), we identified upper and lower layer cortical excitatory neurons and compared HyPBase scCC data between them to identify shared and specific BRD4-bound loci (Figure 2.24A). From 9,083 upper cortical neurons we obtained 29,465 insertions, which was on par with the 31,606 insertions collected from 6,980 lower cortical neurons (Table 2.3). As a positive control, we identified a shared BRD4 binding site at the *Pou3f3* (*Brn-1*) locus (Figure 2.24A, $p < 10^{-9}$). *Pou3f3* was broadly expressed in both populations (Figure 2.24C) and has been used to label layers 2-5 of the postnatal cortex (Molyneaux et al., 2007; Pucilowska et al., 2012). We then identified differentially bound regions in each cluster using insertions from the other cluster as a control. Upper cortical neurons showed specific BRD4 binding at *Pou3f2* (*Brn-2*), which is more restricted to layers 2-4 than *Pou3f3* (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 2.24A; $p < 10^{-9}$ in each instance) (Molyneaux et al., 2007; Rašin et al., 2007). The expression patterns of these genes mirrored BRD4's binding specificity, with *Pou3f2*'s expression mostly retained to the layer 2-4 cluster and the expression of *Bcl11b* and *Foxp2* restricted to the layer 5-6 neuron population (Figure 2.24C). Thus, scCC can successfully identify differentially bound loci between very similar cell types *in vivo*.

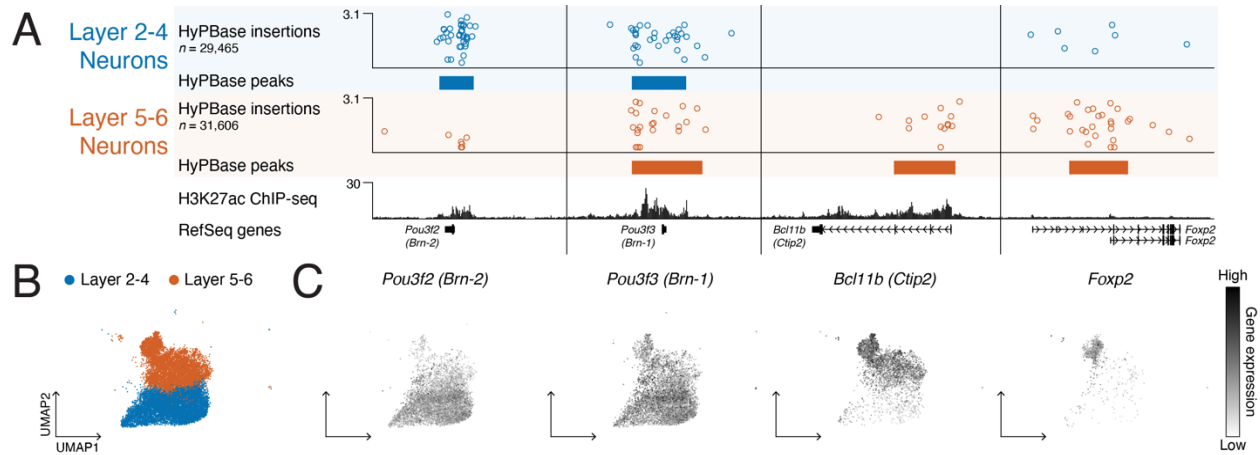


Figure 2.24: Single cell calling cards (scCC) deconvolves BRD4 binding in cortical excitatory neurons and identifies known layer markers. (A) Browser view of scCC HyPBase peaks in upper (layer 2-4) or lower (layer 5-6) cortical excitatory neurons alongside whole cortex H3K27ac ChIP-seq. (B) Layer 2-4 and layer 5-6 cortical excitatory neurons highlighted among the scRNA-seq clusters. (C) Single cell gene expression patterns of the four genes from (A).

2.4 Discussion

Mapping TF binding in heterogeneous tissues is a challenging problem because traditional methods combine signals from multiple cell types into a single, agglomerated profile. This difficulty is further compounded if individual cell types are difficult to identify, isolate, or are rare. Single cell RNA-seq is the dominant paradigm for handling such heterogeneity; here, we have combined it with transposon calling cards and developed a novel method to simultaneously discover cell types and TF binding sites within specific subpopulations. Moreover, we have shown that single cell calling cards (scCC) is robust and flexible: it can map multiple kinds of DNA binding proteins—from sequence-specific TFs like SP1 and FOXA2, to histone-associated factors like BRD4 and BAP1 that bind DNA indirectly—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, heterogeneous systems. scCC is an important addition to the single cell repertoire, fills a recognized void in the field (Shapiro et al.,

2013; Shema et al., 2018), and, unlike scDamID&T (Rooijers et al., 2019), is 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.

A concern with any transposon-based technique is the potential for deleterious interruption of target genes leading to cell death and, consequently, false negatives. Previous experiments in diploid yeast found that calling cards are deposited into promoters of essential and non-essential genes at comparable frequencies (Wang et al., 2011). Since mammalian genomes have much larger intergenic regions than yeast, human and mice genomes are likely also able 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.

The relatively few insertions recovered on a per-cell basis is one of the limitations of this technique, inflating the number of cells that must be analyzed to achieve good sensitivity. Based on our experiments, we recommend processing enough cells to obtain at least 15,000 insertions to analyze BRD4-bound super-enhancers with undirected *piggyBac*; and at least 30,000 insertions for both constructs in TF-directed experiments. These minimum recommendations should achieve moderate sensitivities (~50%) that can be increased by collecting additional insertions (Figure 2.8). The limited amount of data recovered on a per-cell level likely stems from a combination of 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 likely

higher 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, and leads to broader peaks with lower spatial resolution. We overcame the latter constraint by focusing on peak centers and, particularly for motif analysis, on relatively narrow peaks (e.g., less than 5,000 bp in length). Nevertheless, peak width is inversely correlated with the number of insertions analyzed; as such, improving the per-cell recovery of insertions should be prioritized. The inclusion of cis-regulatory features known to enhance mRNA maturation and stability, such as the woodchuck hepatitis virus post-transcriptional regulatory element (WPRE) may increase representation of SRTs in scRNA-seq libraries. Furthermore, as the transcript capture rates of scRNA-seq technologies improve, we expect the sensitivity of our method will increase. Sensitivity can also be improved by simply analyzing larger numbers of cells, such as with Cell Hashing (Stoeckius et al., 2018) or combinatorial barcoding (Rosenberg et al., 2018). Since the per-cell costs for scRNA-seq are falling exponentially (Svensson et al., 2018), we expect that scCC could be used to analyze TF binding in even very rare cell types in the near future.

Another potential limitation of this technique is the exogenous expression of a TF at supraphysiological levels. One possibility is that overexpression of the TF-*piggyBac* construct will lead to ectopic binding and, consequently, false positives. However, 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 activity, 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 undirected *piggyBac* can determine whether there is

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

Our scCC experiments employed the *piggyBac* transposase, but for some applications, other transposases may prove advantageous. *piggyBac* inserts almost exclusively into TTAA tetranucleotides. For TFs that bind GC-rich regions or have high GC-content motifs, *piggyBac* fusions may have a difficult time finding nearby insertion sites. *Sleeping Beauty*, which inserts into TA dinucleotides, or *Tol2*, which does not have a strict insertion site preference (Yoshida et al., 2017), could be used to overcome these limitations. This direction would also improve peak resolution by increasing the number of potential transposition sites. Neither of these transposases tolerate TF fusions, however, which would be the first obstacle to address (Meir et al., 2011; Wu et al., 2006). Meanwhile, the natural affinity of *piggyBac* for BRD4 makes it ideal for studying BRD4-bound SEs, which play important regulatory roles in development and disease (Hnisz et al., 2013; Lovén et al., 2013; Whyte et al., 2013). It is unclear why *piggyBac* has this predilection. BRD4 has an intrinsically disordered region and cooperative interactions between BRD4 and coactivators like MED1 may mediate the formation of intranuclear condensates

(Sabari et al., 2018) at SEs. One hypothesis is that *piggyBac* has a similarly disordered domain that allows it to preferentially enter condensates and enrich SEs with insertions. If such a domain exists, mutating it may make unfused *piggyBac* more uniform in its insertion profile, improving its utility for TF-directed calling cards.

The defining feature of the scCC method is the self-reporting transposon (SRT). While here we have reported the *piggyBac* and *Sleeping Beauty* SRTs, the self-reporting paradigm should be generalizable to any transposon lacking a polyadenylation signal (PAS) in at least one terminal repeat. Expanding the palette of SRTs will illuminate the genome-wide behaviors of transposases and may yield further insight into chromatin dynamics (Yoshida et al., 2017). Simultaneous expression of many TFs, each tagged to a different transposase, may also enable multiplexed studies of TF binding in the same cells. Mapping SRTs using cellular RNA appears to be substantially more efficient than the DNA-based inverse PCR method, but the reasons for this are unclear. Some efficiency is likely gained by eliminating self-ligation, as well as having multiple mRNA copies of each insertion to buffer against PCR artifacts. It is also unknown what fraction of self-reporting transcripts are actually polyadenylated as opposed to merely containing A-rich genomic tracts. Non-genic PASs prevent anti-sense transcription (Chiu et al., 2018), which suggests that PASs may be more common in the genome than previously appreciated. Targeted 3'-end sequencing (Chen et al., 2017; Zheng et al., 2016) of SRT libraries should help resolve this question, while long-read sequencing of self-reporting transcripts may identify non-canonical PASs. 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.

Finally, calling card insertions, being integrated into the genome and preserved through mitosis, could serve as a molecular memory for recording TF binding events. For example, the use of an inducible transposase (Qi et al., 2017) would enable the recording and identification of temporally-restricted TF binding sites. This would help uncover the stepwise order of events underlying the regulation of specific genes and inform cell fate decision making. More generally, transposon insertions could serve as barcodes of developmental lineage. Single transposition events have been used to delineate relationships during hematopoiesis (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.

2.5 Methods

Table 2.4: Oligonucleotides referenced in this work

Name	Sequence	Purification	Notes
SMART_dT18VN	AAGCAGTGGTATCAACGCAGAGTACGTTTTTTTTTT TTTTTTTTTTTTTTTTTTTTTVN	Standard desalt	RT primer for bulk RNA calling card recovery
SMART	AAGCAGTGGTATCAACGCAGAGT	Standard desalt	PCR primer for bulk RNA calling card amplification
SRT_PAC_F1	CAACCTCCCCTTCTACGAGC	Standard desalt	Puromycin marker in SRT
SRT_tdTomato_F1	TCCTGTACGGCATGGACGAG	Standard desalt	tdTomato marker in SRT
Raff_ACTB_F	CCTCGCCTTTGCCGATCCG	Standard desalt	Human ACTB primer (for RT control)
Raff_ACTB_R	GGATCTTCATGAGGTAGTCAGTCAGGTCC	Standard desalt	Human ACTB primer (for RT control)
OM-PB-ACG	AATGATACGGCGACCACCGAGATCTACACTCTTTCC CTACACGACGCTCTTCCGATCTACGTTTACGCAGAC TATCTTTCTAG	Standard desalt	For use with <i>piggyBac</i> SRTs
OM-PB-CTA	AATGATACGGCGACCACCGAGATCTACACTCTTTCC CTACACGACGCTCTTCCGATCTCTATTTACGCAGAC TATCTTTCTAG	Standard desalt	For use with <i>piggyBac</i> SRTs
OM-PB-GAT	AATGATACGGCGACCACCGAGATCTACACTCTTTCC CTACACGACGCTCTTCCGATCTGATTTTACGCAGAC TATCTTTCTAG	Standard desalt	For use with <i>piggyBac</i> SRTs
OM-PB-TGC	AATGATACGGCGACCACCGAGATCTACACTCTTTCC CTACACGACGCTCTTCCGATCTTGC TTTACGCAGAC TATCTTTCTAG	Standard desalt	For use with <i>piggyBac</i> SRTs
OM-PB-TAG	AATGATACGGCGACCACCGAGATCTACACTCTTTCC CTACACGACGCTCTTCCGATCTTAGTTTACGCAGAC TATCTTTCTAG	Standard desalt	For use with <i>piggyBac</i> SRTs
OM-PB-ATC	AATGATACGGCGACCACCGAGATCTACACTCTTTCC CTACACGACGCTCTTCCGATCTATCTTTACGCAGAC TATCTTTCTAG	Standard desalt	For use with <i>piggyBac</i> SRTs

OM-PB-CGT	AATGATACGGCGACCACCGAGATCTACACTCTTTCC CTACACGACGCTCTTCCGATCTCGT ^{TTT} TACGCAGAC TATCTTTCTAG	Standard desalt	For use with <i>piggyBac</i> SRTs
OM-PB-GCA	AATGATACGGCGACCACCGAGATCTACACTCTTTCC CTACACGACGCTCTTCCGATCTGCATTTACGCAGAC TATCTTTCTAG	Standard desalt	For use with <i>piggyBac</i> SRTs
OM-SB-ACG	AATGATACGGCGACCACCGAACACTCTTTCCCTACA CGACGCTCTTCCGATCTAC ^{TAAGT} GTATGTAAACT TCCGACTTCAA	Standard desalt	For use with <i>Sleeping Beauty</i> SRTs
OM-SB-CTA	AATGATACGGCGACCACCGAACACTCTTTCCCTACA CGACGCTCTTCCGATCTCTATAAGTGTATGTAAACT TCCGACTTCAA	Standard desalt	For use with <i>Sleeping Beauty</i> SRTs
OM-SB-GAT	AATGATACGGCGACCACCGAACACTCTTTCCCTACA CGACGCTCTTCCGATCTGAT ^{TAAGT} GTATGTAAACT TCCGACTTCAA	Standard desalt	For use with <i>Sleeping Beauty</i> SRTs
OM-SB-TGC	AATGATACGGCGACCACCGAACACTCTTTCCCTACA CGACGCTCTTCCGATCTT ^{GT} TAAGTGTATGTAAACT TCCGACTTCAA	Standard desalt	For use with <i>Sleeping Beauty</i> SRTs
OM-SB-TAG	AATGATACGGCGACCACCGAACACTCTTTCCCTACA CGACGCTCTTCCGATCTTAG ^{TAAGT} GTATGTAAACT TCCGACTTCAA	Standard desalt	For use with <i>Sleeping Beauty</i> SRTs
OM-SB-ATC	AATGATACGGCGACCACCGAACACTCTTTCCCTACA CGACGCTCTTCCGATCTATCTAAGTGTATGTAAACT TCCGACTTCAA	Standard desalt	For use with <i>Sleeping Beauty</i> SRTs
OM-SB-CGT	AATGATACGGCGACCACCGAACACTCTTTCCCTACA CGACGCTCTTCCGATCTCG ^T TAAGTGTATGTAAACT TCCGACTTCAA	Standard desalt	For use with <i>Sleeping Beauty</i> SRTs
OM-SB-GCA	AATGATACGGCGACCACCGAACACTCTTTCCCTACA CGACGCTCTTCCGATCTGC ^{ATAAGT} GTATGTAAACT TCCGACTTCAA	Standard desalt	For use with <i>Sleeping Beauty</i> SRTs
N7 indexed primer	CAAGCAGAAGACGGCATAACGAGAT[index]GTCTC GTGGGCTCGG	Standard desalt	Uniquely identifies each bulk RNA calling card library in conjunction with barcoded transposon primer
10x_TSO	AAGCAGTGGTATCAACGCAGAGTACATrGrGrG	Standard desalt	For continuing 10x scRNA-seq prep after splitting first RT product in half
Bio_Illumina_Seq1_scCC_10X_3xPT	/5Phos/ACACTCTTTCCC/iBiodT/ACACGACGC TCTTCCGA*T*C*T	HPLC	Single cell calling card primer for use with 10x Chromium 3' v2 kit
Bio_Long_PB_LTR_3xPT	/5Phos/GCGTCAATTTTACGCAGAC/iBiodT/AT CTTTC*T*A*G	HPLC	Single cell calling card primer for use with <i>piggyBac</i> SRTs
scCC_P5_adapter	AATGATACGGCGACCACCGAGATCTTCACTCATTCC ACACGACTCCTTGCCAGTCTC*T	Standard desalt	Adapter for scCC (needs to be pre-annealed with scCC P7 adapter)
scCC_P7_adapter	/5Phos/GAGACTGGCAAGTACACGTCGCACTCACC ATGA[index]ATCTCGTATGCCGCTTCTGCTTG	Standard desalt	Adapter for scCC (needs to be pre-annealed with scCC P5 adapter)
scCC_P5_primer	AATGATACGGCGACCACCGAGATC	Standard desalt	For final scCC library PCR
scCC_P7_primer	CAAGCAGAAGACGGCATAACGAGAT	Standard desalt	For final scCC library PCR
scCC_PB_CustomRead2	CGTGTAGGGAAGAGTGTGCGTCAATTTTACGCAGA CTATCTTTCTAG	PAGE	For custom sequencing of <i>piggyBac</i> scCC libraries; read 2 should begin with GGTTAA
scCC_CustomIndex1	GAGACTGGCAAGTACACGTCGCACTCACCATGA	PAGE	For custom sequencing of scCC libraries
ACTB PrimerBank F	CATGTACGTTGCTATCCAGGC	Standard desalt	For qRT-PCR
ACTB PrimerBank R	CTCCTTAATGTCACGCACGAT	Standard desalt	For qRT-PCR
CD24 PrimerBank F	CTCTACCCACGCAGATTTATTC	Standard desalt	For qRT-PCR
CD24 PrimerBank R	AGAGTGAGACCACGAAGAGAC	Standard desalt	For qRT-PCR
MYC PrimerBank F	GTC AAGAGGCGAACACACAAC	Standard desalt	For qRT-PCR
MYC PrimerBank R	TTGGACGGACAGGATGTATGC	Standard desalt	For qRT-PCR
BRD2 PrimerBank F	AATGGCACAACGCTGGAAAA	Standard desalt	For qRT-PCR
BRD2 PrimerBank R	CACTGGTAACACTGCCTTG	Standard desalt	For qRT-PCR
BRD3 PrimerBank F	TGCAAGCGAATGTATGCAGGA	Standard desalt	For qRT-PCR
BRD3 PrimerBank R	CATCTGGGCCACTTTTGTAGAA	Standard desalt	For qRT-PCR
BRD4 PrimerBank F	GAGTACCCACAGAAGAAACC	Standard desalt	For qRT-PCR
BRD4 PrimerBank R	GAGTCGATGCTTGAGTTGTGTT	Standard desalt	For qRT-PCR
BRD4 CRISPRi gRNA	GCGGCTGCCGGCGGTGCCCG	N/A	For knockdown of BRD4 with CRISPRi

NT CRISPRi gRNA	GGAGGCGAGGTAAGACGCGG	N/A	Control non-targeting gRNA for CRISPRi
-----------------	----------------------	-----	--

2.5.1 Materials and data 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 generated in this study have been submitted to the Gene Expression Omnibus (GEO) and are available at accession GSE148448. All code used to analyze the data is online at https://github.com/arnavm/calling_cards. A complete listing of reagents, datasets, and software used in this study is given in Table 2.5.

Table 2.5: Key Resources Table

Reagent or Resource	Source	Identifier
Antibodies		
Brilliant Violet 421™ anti-human CD24 Antibody (clone ML5)	BioLegend	Cat#311121; RRID:AB_10915556
Brilliant Violet 421™ Mouse IgG2a, κ Isotype Ctrl Antibody (clone MOPC-173)	BioLegend	Cat#400259; RRID:AB_10895919
APC anti-human CD24 Antibody (clone ML5)	BioLegend	Cat#311117; RRID:AB_1877150
APC Rat IgG2a, κ Isotype Ctrl (clone RTK2758)	BioLegend	Cat#400511; RRID:AB_2814702
Bacterial and Virus Strains		
AAV9-PB-SRT-tdTomato	Joseph D. Dougherty (Cammack et al., 2020)	N/A
AAV9-HyPBBase	Joseph D. Dougherty (Cammack et al., 2020)	N/A
Lenti-dCas9-KRAB	This study	N/A
Lenti-BRD4-CRISPRi	This study	N/A
Lenti-NT-CRISPRi	This study	N/A
Chemicals, Peptides, and Recombinant Proteins		
DMEM	Gibco	Cat#11965-084
Antibiotic-Antimycotic (100X)	Gibco	Cat#15240-062
FBS	Peak Serum	Cat#PS-FB3
RPMI 1640 Medium	Gibco	Cat#11875-085
Lipofectamine™ 3000 Transfection Reagent	Invitrogen	Cat#L3000015
Trypsin-EDTA solution	Sigma-Aldrich	Cat#T4049
DPBS, no calcium, no magnesium	Gibco	Cat#14190-136
RNAprotect Cell Reagent	QIAGEN	Cat#76526
2-Mercaptoethanol	Gibco	Cat#21985-023

RNase-Free DNase Set	QIAGEN	Cat#79254
Maxima H Minus Reverse Transcriptase	Thermo Scientific	Cat#EP0752
Advantage® UltraPure PCR Deoxynucleotide Mix	Takara Bio	Cat#639125
RNaseOUT™ Recombinant Ribonuclease Inhibitor	Invitrogen	Cat#10777019
TransIT®-LT1 Transfection Reagent	Mirus	Cat#MIR2304
RNase H	New England BioLabs	Cat#M0297S
HiFi HotStart ReadyMix (2X)	Kapa Biosystems	Cat#KK2601
AMPure XP beads	Beckman Coulter	Cat#A63880
Puromycin dihydrochloride	Sigma-Aldrich	Cat#P8833
Crystal violet	Sigma-Aldrich	Cat#C0775
Methanol	Fisher Scientific	Cat#A452-4
Formaldehyde	Fisher Scientific	Cat#BP531-500
High Sensitivity D1000 Reagents	Agilent	Cat#5067-5585
Ficoll PM400 (Dry Powder)	GE Healthcare	Cat#17030010
NxGen® RNase Inhibitor	Lucigen	Cat#30281-1
Dynabeads™ MyOne™ Silane	Life Technologies	Cat#37002D
IDTE pH 8.0 (1X TE Solution)	IDT	Cat#11-05-01-13
High Sensitivity D5000 Reagents	Agilent	Cat#5067-5593
NEBuffer™ 2	New England BioLabs	Cat#B7002S
Buffer EB	QIAGEN	Cat#19086
Hibernate™-A Medium	Gibco	Cat#A1247501
D-(+)-Trehalose dihydrate	Sigma-Aldrich	Cat#T9531
B-27™ Supplement (50X), serum free	Gibco	Cat#17504044
0.5M EDTA, pH 8.0	Corning	Cat#46-034-CI
Papain, Lyophilized	Worthington Biochemical	Cat#LS003118
Deoxyribonuclease I, Filtered	Worthington Biochemical	Cat#LS002060
Trypsin Inhibitor, Ovomuroid	Worthington Biochemical	Cat#LS003087
Bovine Serum Albumin	Sigma-Aldrich	Cat#A9418
OptiPrep™ Density Gradient Medium	Sigma-Aldrich	Cat#D1556
HBSS (10X)	Gibco	Cat#14185052
Magnesium chloride	Sigma-Aldrich	Cat#M4880
Magnesium sulfate	Sigma-Aldrich	Cat#M2643
Calcium chloride dihydrate	Sigma-Aldrich	Cat#C7902
D-(+)-Glucose	Sigma-Aldrich	Cat#G7021
Dimethyl sulfoxide (DMSO)	Sigma-Aldrich	Cat#D2650
Cell Staining Buffer	BioLegend	Cat#420201
Annexin V Binding Buffer	BioLegend	Cat#422201
SuperScript™ VILO™ cDNA Synthesis Kit	Invitrogen	Cat#11754250
PowerUp™ SYBR™ Green Master Mix	Applied Biosystems	Cat#25742
(+)-JQ1	Selleck Chemicals	Cat#S7110
Propidium iodide (PI)	Invitrogen	Cat#P3566
Hoechst 33342	Thermo Scientific	Cat#62249

Annexin V-FITC	BioLegend	Cat#640905
Blasticidin S HCl	Gibco	Cat#A1113903
Lenti-X™ Concentrator	Takara Bio	Cat#631232
Lipofectamine™ 2000 Transfection Reagent	Invitrogen	Cat#11668030
Polybrene Infection / Transfection Reagent	Sigma-Aldrich	Cat#TR-1003
Esp31	New England BioLabs	Cat#R0734S
T4 DNA Ligase	New England BioLabs	Cat#M0202S
IMDM	Gibco	Cat#12440046
Penicillin-Streptomycin (10,000 U/mL)	Gibco	Cat#15140122
Imatinib mesylate	Sigma-Aldrich	Cat#SML1027
Lovastatin	Sigma-Aldrich	Cat#M2147
Nocodazole	Sigma-Aldrich	Cat#M1404
CVT-313	Sigma-Aldrich	Cat#238803
RO-3306	Sigma-Aldrich	Cat#SML0569
Critical Commercial Assays		
Neon™ Transfection System 100 µL Kit	Invitrogen	Cat#MPK10025
RNeasy Plus Mini Kit	QIAGEN	Cat#74134
Qubit™ RNA 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
HCT-116 CTCF ChIP-seq	Richard Myers	ENCODE:ENCFF000OZC
HCT-116 ChIP-seq input control (SP1, CTCF)	Richard Myers	ENCODE:ENCFF000PBO
HCT-116 BRD4 ChIP-seq	Ron Firestein	SRA:SRR2481799
HCT-116 ChIP-seq input control (BRD4)	Ron Firestein	SRA:SRR2481800
HCT-116 H3K27ac ChIP-seq	Bradley Bernstein	ENCODE:ENCFF082JPN; ENCODE:ENCFF176BXC
HCT-116 H3K4me1 ChIP-seq	Bradley Bernstein	ENCODE:ENCFF088BWP; ENCODE:ENCFF804MJI
HCT-116 H3K4me2 ChIP-seq	Bradley Bernstein	ENCODE:ENCFF936MMN; ENCODE:ENCFF937OOL
HCT-116 H3K4me3 ChIP-seq	Bradley Bernstein	ENCODE:ENCFF183OZI; ENCODE:ENCFF659FPR
HCT-116 H3K9me2 ChIP-seq	Bradley Bernstein	ENCODE:ENCFF760OZN; ENCODE:ENCFF565FDP
HCT-116 H3K9me3 ChIP-seq	Bradley Bernstein	ENCODE:ENCFF578MDZ; ENCODE:ENCFF033XOG
HCT-116 H3K27me3 ChIP-seq	Bradley Bernstein	ENCODE:ENCFF281SBT; ENCODE:ENCFF124GII

HCT-116 H3K36me3 ChIP-seq	Bradley Bernstein	ENCODE:ENCFF850EAH; ENCODE:ENCFF312RKB
HCT-116 H3K79me2 ChIP-seq	Bradley Bernstein	ENCODE:ENCFF865KPW; ENCODE:ENCFF947YPU
HCT-116 H4K20me1 ChIP-seq	Bradley Bernstein	ENCODE:ENCFF070JDY; ENCODE:ENCFF334HHB
HCT-116 ChIP-seq input control (H3K27ac, H3K4me1, H3K4me2, H3K4me3, H3K9me2, H3K9me3, H3K27me3, H3K36me3, H3K79me2, H4K20me1)	Bradley Bernstein	ENCODE:ENCFF048ZOQ; ENCODE:ENCFF827YXC
HCT-116 H3K9ac ChIP-seq	Bradley Bernstein	ENCODE:ENCFF408RRT
HCT-116 ChIP-seq input control (H3K9ac)	Bradley Bernstein	ENCODE:ENCFF413RQG
K562 BRD4 ChIP-seq	Bradley Bernstein	ENCODE:ENCFF335PHG
K562 H3K27ac ChIP-seq	Bradley Bernstein	ENCODE:ENCFF000BXH
K562 ChIP-seq input control (BRD4, H3K27ac)	Bradley Bernstein	ENCODE:ENCFF000BWK
K562 SP1 ChIP-seq	Michael Snyder	ENCODE:ENCFF002DPL; ENCODE:ENCFF002EGC
K562 ChIP-seq input control (SP1)	Michael Snyder	ENCODE:ENCFF002EGI; ENCODE:ENCFF002EGA
HepG2 FOXA2 ChIP-seq	Richard Myers	ENCODE:ENCFF000PIX
HepG2 ChIP-seq input control (FOXA2)	Richard Myers	ENCODE:ENCFF000POV
OCM-1A HyPBase DNA calling cards	Michael Onken	DOI:10.1186/s12920-018-0424-0
OCM-1A BAP1-HyPBase DNA calling cards	Michael Onken	DOI:10.1186/s12920-018-0424-0
OCM-1A RNA-seq (BAP1 and control shRNA)	Michael Onken	GEO:GSE110193
Mouse cortex H3K27ac ChIP-seq	Michael Greenberg	SRA:SRR6129714
Mouse cortex ChIP-seq input control (H3K27ac)	Michael Greenberg	SRA:SRR6129695
K562 RNA Pol II ChIA-PET	Yijun Ruan	ENCODE:ENCFF000KYH
HCT-116 DNase-seq	John Stamatoyannopoulos	ENCODE:ENCFF001DCK
HCT-116 ATAC-seq	Sriharsa Pradhan	SRA:SRR5453778
HCT-116 ATAC-seq control	Michael Guertin	GEO:GSE92674
HCT-116 CpG islands	Richard Myers	GEO:GSM1014209
Sequencing data and processed output	This study	GEO:GSE148448
Experimental Models: Cell Lines		
Neuro-2a (N2a)	ATCC	Cat#CCL-131
K-562	ATCC	Cat#CCL-243
Hep G2	ATCC	Cat#HB-8065
OCM-1A	Michael Onken (Yen et al., 2018)	N/A
HCT 116	ATCC	Cat#CCL-247
293T/17 [HEK 293T/17]	ATCC	Cat#CRL-11268
Experimental Models: Organisms/Strains		
Mouse: C57BL/6J	Joseph D. Dougherty (Cammack et al., 2020)	N/A
Oligonucleotides		
Primers and oligonucleotides	This study, see Table	N/A

	2.4	
Recombinant DNA		
pRM1024: PBase	This study	N/A
pRM1114: HyPBase	This study	N/A
pRM1023: SP1-PBase	This study	N/A
pRM1677: SP1-HyPBase	This study	N/A
pRM1882: FOXA2-HyPBase	This study	N/A
pRM1863: BAP1-HyPBase	This study	N/A
pRM1304: PB-SRT-Puro	This study	RRID:Addgene_154884
pRM1535: PB-SRT-tdTomato	This study	RRID:Addgene_154885
pCMV(CAT)T7-SB100	Zsuzsanna Izsvák	RRID:Addgene_34879
pRM1665: SP1-SB100X	This study	RRID:Addgene_154887
pRM1668: SB-SRT-Puro	This study	RRID:Addgene_154888
pRM1217: AAV-HyPBase	Joseph D. Dougherty (Cammack et al., 2020)	N/A
pRM1648: AAV-PB-SRT-tdTomato	Joseph D. Dougherty (Cammack et al., 2020)	RRID:Addgene_154889
pUC19 Vector	New England BioLabs	Cat#N3041S
Lenti-dCas9-KRAB-blast	Gary Hon	RRID:Addgene_89567
sgOpti	Eric Lander & David Sabatini	RRID:Addgene_85681
pMD2.G	Didier Trono	RRID:Addgene_12259
psPAX2	Didier Trono	RRID:Addgene_12260
pRM1889: BRD4 CRISPRi plasmid	This study	RRID:Addgene_154890
pRM1890: Non-targeting CRISPRi plasmid	Robi D. Mitra (Lalli et al., 2019)	RRID:Addgene_154891
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	Ramírez 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
ROSE 0.1	Whyte et al., 2013 & Lovén et al., 2013	RRID:SCR_017390
FlowJo™ Software for Mac Version 10	Becton, Dickson and Company	RRID:SCR_008520
Multcomp 1.4-12	Hothorn et al., 2008	RRID:SCR_018255
Custom calling card code	This study	https://github.com/arnavm/calling_cards
Other		
Qubit® 3.0 Fluorometer	Thermo Fisher	Cat#Q33216
4200 TapeStation System	Agilent	Cat#G2991AA
E220 Focused-ultrasonicator	Covaris	N/A
MasterCycler Pro PCR System	Eppendorf	Cat#950030010
Attune NxT Flow Cytometer	Thermo Fisher	N/A
CytoFLEX S	Beckman-Coulter	Cat#B75442
QuantStudio™	Applied Biosystems	Cat#A28567
Protocol: Mammalian Calling Cards Quick Start Guide	This study	DOI:10.17504/protocols.io.xurfnv6
Protocol: Bulk Calling Cards Library Preparation	This study	DOI:10.17504/protocols.io.xwhfpb6
Protocol: Single Cell Calling Cards Library Preparation	This study	DOI:10.17504/protocols.io.xwifpce
Protocol: Processing Bulk Calling Card Sequencing Data	This study	DOI:10.17504/protocols.io.xwjfpcn
Protocol: Processing Single Cell Calling Card Sequencing Data	This study	DOI:10.17504/protocols.io.4phgvj6
Protocol: Calling Peaks on <i>piggyBac</i> Calling Card Data	This study	DOI:10.17504/protocols.io.bb9xir7n
Protocol: Visualizing Calling Card Data on the WashU Epigenome Browser	This study	DOI:10.17504/protocols.io.bca8ishw

2.5.2 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 CD24^{high}/CD24^{low} cell state analyses, K562 cells were grown in IMDM containing 10% v/v FBS and 1% penicillin-streptomycin at 37 °C with 5% CO₂. Frozen aliquots were thawed and passaged every 48 hours until they reached a maximum 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 CD24^{high}/CD24^{low} 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-tdTomato and HyPBase constructs into AAV vectors. The Hope Center Viral Vectors Core at Washington University in St. Louis packaged each construct in AAV9 capsids. Titers for each virus ranged between 1.1×10^{13} and 2.2×10^{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-tdTomato only. After 2 to 4 weeks, we sacrificed mice and dissected the cortex (8 libraries) or hippocampus (1 library). All animal practices and procedures were approved by the Washington University in St. Louis Institutional Animal Care and Use Committee (IACUC) in accordance with National Institutes of Health (NIH) guidelines.

2.5.3 DNA- vs RNA-based recovery

Approximately 500,000 HCT-116 cells were plated in a single well of a 6-well plate. Cells were transfected with 2.5 µg of the SP1-PBase plasmid 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-ligation, 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).

2.5.4 *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 PBBase plasmid via Neon electroporation. Each replicate contained 2×10^6 cells. As a negative control, we transfected one replicate of HCT-116 cells with 5 µg PB-SRT-Puro plasmid only. We used the following settings—pulse voltage: 1,530 V; pulse width: 20 ms; pulse number: 1. We used the same experimental setup for experiments with PB-SRT-Puro and each of SP1-PBase, HyPBase, and SP1-HyPBase plasmids, as well as with SB-SRT-Puro and SB100X plasmids. 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 (a gift from Dr. Michael Onken) 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.

2.5.5 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 DNaseI 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 µg of total RNA with 1 µl 10 mM dNTPs and 1 µl of 50 µM SMART_dT18VN primer (Table 2.4), brought the total volume up to 14 µl, and incubated it 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.

2.5.6 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 primer, and either 1 µl of 25 µM SRT_PAC_F1 primer (in the case of puromycin selection) or 0.5 µl of 25 µM SRT_tdTomato_F1 primer (in the case of

tdTomato screening). The mixture was brought up to 25 μ l with ddH₂O. Thermocycling parameters were as follows: 95°C for 3 minutes; 20 cycles of: 98°C for 20 seconds–65°C for 30 seconds–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 (Table 2.4; (Raff et al., 1997)) under the same thermocycling settings.

PCR products were purified using AMPure XP beads. 12 μ l of resuspended beads were added to the 25 μ l PCR product and mixed homogeneously 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 μ l of 70% ethanol (incubated for 30 seconds each time), discarding the supernatant each time. The pellet was left to dry at room temperature for 2 minutes. To elute, we added 20 μ l 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.

2.5.7 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 μ l ddH₂O. To this mixture we added 10 μ l Tagment DNA (TD) Buffer and 5 μ l 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 μ l Neutralization Tagment (NT) Buffer and was kept at room temperature for 5 minutes. The final PCR was set up by adding 15 μ l Nextera PCR Mix (NPM), 8 μ l ddH₂O, 1 μ l of 10 μ M transposon primer (e.g. OM-PB-NNN) and 1 μ l 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 seconds; 13 cycles of: 95°C for 10 seconds—50°C for 30 seconds—72°C for 30 seconds; 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.

2.5.8 *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 2×10^6 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 HyPBBase. For the cell line mixing

experiment, we transfected four replicates each of HCT-116 and K562 cells with 5 μ g PB-SRT-Puro and 5 μ g HyPBase. 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 μ g PB-SRT-Puro and 5 μ g 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 μ g PB-SRT-Puro; three of these replicates were co-transfected with 5 μ g HyPBase, while the other three were co-transfected with 5 μ g 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 μ g PB-SRT-Puro; three of these replicates were co-transfected with 1.25 μ g HyPBase, while the other three were co-transfected with 1.25 μ g BAP1-HyPBase.

2.5.9 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 μ l Elution Solution I and transferred 36 μ l of the eluted sample to a new tube (instead of 35.5 μ l and 35 μ l, respectively). The eluate was split into two 18 μ l 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 μ l of Maxima 5X RT Buffer, 20 μ l of 20% w/v Ficoll PM-400, 10 μ l of 10 mM dNTPs, 2.5 μ l RNase Inhibitor and 2.5 μ l of 100 μ M 10x_TSO. To this solution we added 18 μ l of the first RT product and 22 μ l of ddH₂O. Finally, we added 5 μ l 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.

2.5.10 Single cell calling cards library preparation

To amplify self-reporting transcripts from single cell RNA-seq libraries, we took 9 μ l of RT product (the other half was kept in reserve) and added it to 25 μ l Kapa HiFi HotStart ReadyMix and 15 μ l ddH₂O. We then prepared a PCR primer cocktail comprising 5 μ l of 100 μ M Bio_Illumina_Seq1_scCC_10X_3xPT primer, 5 μ l of 100 μ M Bio_Long_PB_LTR_3xPT, and 10 μ l of 10 mM Tris-HCl, 0.1 mM EDTA buffer. One μ l 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 seconds–67°C for 30 seconds–72°C for 5 minutes; 72°C for 10 minutes; 4°C forever. PCR purification was performed with 30 μ l 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 μ l with ddH₂O, then adding 30 μ l Circularization Buffer 10x and 2 μ l 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 seconds; temperature: 6°C.

The library preparation was performed per manufacturer's instructions until adapter ligation. We designed custom adapters (Table 2.4) so that the standard Illumina sequencing primers would not interfere with our library. Adapters were prepared by combining 4.5 μ l of 100 μ M scCC_P5_adapter, 4.5 μ l of 100 μ M scCC_P7_adapter, and 1 μ l 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% 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 Adapter Index. The ligation product was cleaned per manufacturer's instructions. For the final PCR, the master mix was created by combining 20 μ l Enhanced PCR Mix with 28 μ l of ddH₂O and 1 μ l each of 25 μ M scCC_P5_primer and 25 μ M scCC_P7_primer. This was then added to the streptavidin bead-bound DNA and amplified under the following conditions: 98°C for 30 seconds; 15 cycles of: 98°C for 10 seconds—60°C for 30 seconds—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 μ l (0.7x) AMPure XP beads following manufacturer's instructions. The final library was eluted in 25 μ l Elution Buffer and quantitated on an Agilent TapeStation 4200 System using the High Sensitivity D1000 ScreenTape.

2.5.11 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 μ l of Cell Staining Buffer containing 0.2 μ g 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 μ l 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 μ g/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 μ l of Cell Staining Buffer, cells were washed twice with Annexin V Staining Buffer. Cells were then incubated in 50 μ l Annexin V Staining Buffer containing 0.2 μ g Annexin V-FITC and 100 μ g/ml propidium iodide (PI). The reaction was incubated for 15 minutes at room temperature in the dark. Afterwards, we added 150 μ l 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.

2.5.12 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 μ l of RNA CellProtect, and stored at -80 °C. When we were ready to extract RNA, we thawed cells, 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 μ l 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 μ l of the RT product as template, 1 μ l each of forward and reverse primer (10 μ M), 6 μ l ddH₂O, and 10 μ l 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 seconds 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.

2.5.13 BRD4 CRISPRi of K562 cells

For CRISPRi, we first made lentivirus expressing dCas9-KRAB from Addgene plasmid #89567, a gift from Gary Hon, packaged in HEK 293T 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 HEK 293T cells using Lipofectamine 2000. Media was collected after 24 and 48 hours, and subsequently concentrated using Lenti-X™ Concentrator. Viral titers were functionally assayed on HEK 293T 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 blasticidin (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 an MOI of 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 allowed to expand for a further 5 days before being stained for CD24 (nine days after gRNA transduction.)

2.5.14 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 levels. The remaining 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 μ M. After 48 hours, every well was stained for CD24 levels and apoptotic activity, as previously described. Cells undergoing BRD4 or non-targeted CRISPRi were split in two and treated with either DMSO or imatinib (1 μ M) as described and in triplicate. The resulting data were processed with FlowJo™. We set gates such that we could exclude debris but that we would capture both live and dying cells. This gate was used to calculate levels of annexin V and PI.

2.5.15 Cell cycle perturbation of K562 cells

The cell cycle inhibitors lovastatin, nocodazole, CVT-313 and RO-3306 were purchased from Sigma-Aldrich. 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 μ M lovastatin, 40 ng/ μ l nocodazole (in ethanol), 2 μ M CVT-313, or 4.5 μ M 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™. After 5 days of treatment, we stained for CD24 levels and apoptotic activity. 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%).

2.5.16 SRT-tdTomato fluorescence validation

To test the fluorescence properties of the SRT-tdTomato construct, we transfected K562 cells as previously described with either 1 µg of pUC19 plasmid; 0.5 µg of PB-SRT-tdTomato plasmid and 0.5 µg pUC19; 0.5 µg of PB-SRT-tdTomato and 0.5 µg pBase plasmid; and 0.5 µg of PB-SRT-tdTomato and 0.5 µg HyPBase plasmid. Cells were allowed to expand for 8 days, after which fluorescence activity was assayed on an Attune NxT Flow Cytometer 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 (Figure 2.2A).

2.5.17 In vivo scCC experiments

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 2 narrow fire-polished pipettes, and passed through a 40-micron filter prewetted with resuspension solution: Hibernate-A containing 5% v/v trehalose, 0.5% Ovomuroid 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.

2.5.18 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; at least three separate libraries for single cell calling cards experiments; and three biological replicates for the K562 cell state experiments.

2.5.19 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 TTAA. Reads that had exact matches to these sequences were hard trimmed using cutadapt (Martin, 2011) with the following settings: `-g "^TAGTTTACGCAGACTATCTTTCTAGGGTTAA" --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 "^ATCACTTAAGCCGGTAC" --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/arnavm/calling_cards) and visualized on the WashU Epigenome Browser v46 (Zhou et al., 2011) (<http://epigenomegateway.wustl.edu/legacy/>).

2.5.20 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

```
"^NNNGCGTCAATTTTACGCAGACTATCTTTCTAGGGTTAA" --minimum-length 1 --discard-untrimmed -e 0 --no-indels, where NNN is replaced with the primer barcode. Sleeping Beauty libraries were trimmed with the following settings: -g
```

```
"^NNNTAAGTGTATGTAAACTTCCGACTTCAACTGTA" --minimum-length 1 --discard-untrimmed -e 0 --no-indels. Reads passing this filter were then trimmed of any trailing Nextera adapter sequence, again using cutadapt and the following settings: -a
```

```
"CTGTCTCTTATACACATCTCCGAGCCCACGAGACTNNNNNNNNNTCTCGTATGCCGTCTTCTGCTTG" --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://epigenomegateway.wustl.edu/legacy/>).

2.5.21 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 --localcores=16 --localmem=30`. The digital gene expression matrices from 10x were then further processed with scanpy (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)

2.5.22 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 2.4). Read 1 sequenced the cell barcode and unique molecular index of each self-reporting transcript. Read 2 began with GGTTAA (end of the *piggyBac* terminal repeat and insertion site motif) before continuing into the genome. Reads containing this exact hexamer were trimmed using cutadapt with the following settings: `-g "^GGTTAA" --minimum-length 1 --discard-untrimmed -e 0 --no-indels`.

Reads passing this filter were then trimmed of any trailing P7 adapter sequence, again using cutadapt and with the following settings: `-a`

`"AGAGACTGGCAAGTACACGTCGCACTCACCATGANNNNNNNNNATCTCGTATGCCGTCTTCTGCTTG" --minimum-length 1`. Reads passing these filters were aligned using 10x

Genomics' cellranger with the following settings: `--expect-cells=6000 --nosecondary --chemistry=SC3Pv2 --localcores=16 --localmem=30`. This workflow also managed barcode

validation and collapsing of UMIs. Aligned reads were validated by verifying that they mapped adjacent to TTAA tetramers. Reads were then converted to calling card format (.ccf). Finally, to minimize the presence of intermolecular artifacts, we required that each insertion must have been tagged by at least two different UMIs. We used the set of validated cell barcodes from each scRNA-seq library to demultiplex library-specific barcoded insertions from the scCC data. This approach requires no shared cell barcodes between 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.

2.5.23 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 $B = \{b_1, b_2, \dots, b_n\}$ represent the set of blocks found by performing Bayesian block segmentation on all insertions from a TF-directed experiment (e.g. SP1-PBase). For each block b_i , let x_i be the number of insertions in that block in the TF-directed experiment. Similarly, let y_i be the number

of insertions in that block in the undirected experiment (e.g. PBase) normalized to the total number of insertions found in the TF-directed experiment. Then, for each block we calculated the Poisson p -value of observing at least x_i insertions assuming a Poisson distribution with expectation y'_i : $P(k \geq x_i | \lambda = y'_i)$. We accepted all blocks that were significant beyond a particular p -value threshold.

For the analysis of TF-directed insertions, either in bulk or in single cells, we added a pseudocount of 1 to y'_i , 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 TTAAAs.

To identify BRD4 binding sites from undirected *piggyBac* insertions, we segmented those insertions using Bayesian blocks. For each block b_i , we let x_i denote the number of undirected insertions in that block. We also calculated x'_i , the expected number of insertions in block b_i assuming *piggyBac* insertions were distributed uniformly across the genome. We did this by dividing the total number of TTAAAs in the genome by the total number of undirected insertions, then multiplying this value by the number of TTAAAs in block b_i . Then, for each block we calculated the Poisson p -value $P(k \geq x_i | \lambda = x'_i)$. We accepted all blocks that were significant beyond a particular p -value threshold. Finally, we merged statistically-significant blocks that were within 12,500 bases of each other (Pott and Lieb, 2014; 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 CD24^{high}/CD24^{low} 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 CD24^{high}/CD24^{low} 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 $\log_2 \frac{\text{Normalized CD24 high insertions}}{\text{Normalized CD24 low insertions}}$. We then took the smaller of the two p -values and adjusted for multiple hypotheses at a Benjamini-Hochberg false discovery rate of 10%. This was plotted against the fold change values to generate the volcano plot.

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 mapped insertions, or IPKM). This was plotted as a bedgraph file with smoothing applied in the WashU Epigenome Browser (25 pixel windows).

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 Appendices 4-6.

2.5.24 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 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 statuses 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 kb 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-minw 4 -meme-nmotifs 10`. Motifs were aligned on the web version of Tomtom (Gupta et al., 2007) querying the “Vertebrates (In vivo and in silico)” database. We cross-referenced BAP1 scCC binding sites with publicly available BAP1 shRNA data (Yen et al., 2018), focusing on genes that showed a significant change in gene expression (adjusted p -value < 0.05).

2.5.25 BRD4 sensitivity, specificity, and precision

We used a published BRD4 ChIP-seq dataset (McClelland et al., 2016) to identify BRD4-bound super-enhancers in HCT-116 cells, following previously-described methods (Lovén 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.

2.5.26 Downsampling and replication analysis

When performing downsampling analyses on calling card insertions, we randomly sampled insertions without replacement and in proportion to the number of reads supporting each insertion. Peaks were called on the downsampled insertions at a range of p -value cutoffs. Linear interpolation was performed using NumPy (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.

2.5.27 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 deeptools (Ramírez 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 (Table 2.6) 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).

Table 2.6: ChromHMM chromatin state annotations in HCT-116 cells

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

11	0	0	0	0	0	0	0	0	0	0	0	0	Inactive
14	0	1	1	1	0	1	2	0	0	0	0	0	
10	3	48	69	30	54	36	28	21	34	21	15	28	

Chromatin mark observation frequency (%)

2.5.28 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 mixing experiment. We observed that the top genes in PC1 (*VIM*, *CD24*, *TMSB4X*, *LYZ*, and *LGALS1*) and PC2 (*HBE1*, *HBG2*, *HBG1*, *HBZ*, and *HBA2*) were anticorrelated with each other (Figure 2.18A-B), implying the existence of two mutually exclusive states. We scored cells based on the expression of *VIM*, *TMSB4X*, *HBG1*, and *HBG2*. We then modeled the distribution of this state score as a 3 component Gaussian mixture model, drawing cutoffs where adjacent Gaussian distributions intersected (Figure 2.18C). These cutoffs were then used to label cells as either stem-like ($CD24^{\text{high}}$), differentiated ($CD24^{\text{low}}$), or intermediate (Figure 2.18D). 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 2.18E). Differentially bound peaks were called as described above.

2.5.29 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 $CD24^{\text{low}}$ cells as the dependent variable. For the analysis of annexin V levels in either JQ1- or DMSO-treated $CD24^{\text{high}}$ and $CD24^{\text{low}}$ 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.

2.5.30 *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_i , we calculated the specificity as

$$\frac{\text{Astrocyte}_{FPKM}(g_i)}{\text{Astrocyte}_{FPKM}(g_i) + \text{Neuron}_{FPKM}(g_i)}$$

Thus, a value of 0 denotes a gene purely expressed in neurons,

a value of 0.5 for a gene equally expressed in both cell types, and a value of 1 for a gene purely expressed in astrocytes. We plotted distributions of gene expression specificity for the set of

astrocyte-bound genes and the downsampled astrocyte-bound genes. Gene Ontology analysis was performed on the same sets of genes using PANTHER (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.

2.5.31 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 (Table 2.5) and reproduced in the appendices, 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 folder should contain the most up-to-date information.

2.6 References

Ai, S., Xiong, H., Li, C.C., Luo, Y., Shi, Q., Liu, Y., Yu, X., Li, C., and He, A. (2019). Profiling chromatin states using single-cell *itChIP-seq*. *Nat. Cell Biol.* *21*, 1164–1172.

Angermueller, C., Clark, S.J., Lee, H.J., Macaulay, I.C., Teng, M.J., Hu, T.X., Krueger, F., Smallwood, S., Ponting, C.P., Voet, T., et al. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* *13*, 229–232.

Avey, D., Sankararaman, S., Yim, A.K.Y., Barve, R., Milbrandt, J., and Mitra, R.D. (2018). Single-Cell RNA-Seq Uncovers a Robust Transcriptional Response to Morphine by Glia. *Cell Rep.* *24*, 3619–3629.e4.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* 57, 289–300.

Brandeis, M., Frank, D., Keshet, I., Siegfried, Z., Mendelsohn, M., Nemes, A., Temper, V., Razin, A., and Cedar, H. (1994). Sp1 elements protect a CpG island from de novo methylation. *Nature* 371, 435–438.

Brooks, E.E., Gray, N.S., Joly, A., Kerwar, S.S., Lum, R., Mackman, R.L., Norman, T.C., Rosete, J., Rowe, M., Schow, S.R., et al. (1997). CVT-313, a Specific and Potent Inhibitor of CDK2 That Prevents Neointimal Proliferation. *J. Biol. Chem.* 272, 29207–29211.

Brunner, A.L., Johnson, D.S., Kim, S.W., Valouev, A., Reddy, T.E., Neff, N.F., Anton, E., Medina, C., Nguyen, L., Chiao, E., et al. (2009). Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. *Genome Res.* 19, 1044–1056.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* 10, 1213–1218.

Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 523, 486–490.

Cadiñanos, J., and Bradley, A. (2007). Generation of an inducible and optimized piggyBac transposon system. *Nucleic Acids Res.* 35, e87–e87.

Cammack, A.J., Moudgil, A., Chen, J., Vasek, M.J., Shabsovich, M., McCullough, K., Yen, A., Lagunas, T., Maloney, S.E., He, J., et al. (2020). A viral toolkit for recording transcription factor–DNA interactions in live mouse tissues. *Proc. Natl. Acad. Sci.* 201918241.

Campagne, A., Lee, M.-K., Zielinski, D., Michaud, A., Le Corre, S., Dingli, F., Chen, H., Shahidian, L.Z., Vassilev, I., Servant, N., et al. (2019). BAP1 complex promotes transcription by opposing PRC1-mediated H2A ubiquitylation. *Nat. Commun.* 10, 348.

Campbell, J.N., Macosko, E.Z., Fenselau, H., Pers, T.H., Lyubetskaya, A., Tenen, D., Goldman, M., Verstegen, A.M.J., Resch, J.M., McCarroll, S.A., et al. (2017). A molecular census of arcuate hypothalamus and median eminence cell types. *Nat. Neurosci.* 20, 484–496.

Cao, J., Packer, J.S., Ramani, V., Cusanovich, D.A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S.N., Steemers, F.J., et al. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* 357, 661–667.

Cao, J., Cusanovich, D.A., Ramani, V., Aghamirzaie, D., Pliner, H.A., Hill, A.J., Daza, R.M., McFaline-Figueroa, J.L., Packer, J.S., Christiansen, L., et al. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* 33, eaau0730.

Carbone, M., Yang, H., Pass, H.I., Krausz, T., Testa, J.R., and Gaudino, G. (2013). BAP1 and cancer. *Nat. Rev. Cancer* 13, 153–159.

Carter, B., Ku, W.L., Kang, J.Y., Hu, G., Perrie, J., Tang, Q., and Zhao, K. (2019). Mapping histone modifications in low cell number and single cells using antibody-guided chromatin tagmentation (ACT-seq). *Nat. Commun.* 10, 3747.

Castillo-Hair, S.M., Sexton, J.T., Landry, B.P., Olson, E.J., Igoshin, O.A., and Tabor, J.J. (2016). FlowCal: A User-Friendly, Open Source Software Tool for Automatically Converting Flow Cytometry Data from Arbitrary to Calibrated Units. *ACS Synth. Biol.* *5*, 774–780.

Chen, W., Jia, Q., Song, Y., Fu, H., Wei, G., and Ni, T. (2017). Alternative Polyadenylation: Methods, Findings, and Impacts. *Genomics Proteomics Bioinformatics* *15*, 287–300.

Chiu, A.C., Suzuki, H.I., Wu, X., Mahat, D.B., Kriz, A.J., and Sharp, P.A. (2018). Transcriptional Pause Sites Delineate Stable Nucleosome-Associated Premature Polyadenylation Suppressed by U1 snRNP. *Mol. Cell* *69*, 648-663.e7.

Cho, S.W., Xu, J., Sun, R., Mumbach, M.R., Carter, A.C., Chen, Y.G., Yost, K.E., Kim, J., He, J., Nevins, S.A., et al. (2018). Promoter of lncRNA Gene PVT1 Is a Tumor-Suppressor DNA Boundary Element. *Cell* *173*, 1398-1412.e22.

Clark, S.J., Argelaguet, R., Kapourani, C.-A., Stubbs, T.M., Lee, H.J., Alda-Catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., Marioni, J.C., et al. (2018). scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* *9*, 390.

Datlinger, P., Rendeiro, A.F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L.C., Kuchler, A., Alpar, D., and Bock, C. (2017). Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* *14*, 297–301.

Davis, R.L., Weintraub, H., and Lassar, A.B. (1987). Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell* *51*, 987–1000.

Delmore, J.E., Issa, G.C., Lemieux, M.E., Rahl, P.B., Shi, J., Jacobs, H.M., Kastiris, E., Gilpatrick, T., Paranal, R.M., Qi, J., et al. (2011). BET Bromodomain Inhibition as a Therapeutic Strategy to Target c-Myc. *Cell* 146, 904–917.

Dey, A., Seshasayee, D., Noubade, R., French, D.M., Liu, J., Chaurushiya, M.S., Kirkpatrick, D.S., Pham, V.C., Lill, J.R., Bakalarski, C.E., et al. (2012). Loss of the Tumor Suppressor BAP1 Causes Myeloid Transformation. *Science* 337, 1541–1546.

Dey, S.S., Kester, L., Spanjaard, B., Bienko, M., and van Oudenaarden, A. (2015). Integrated genome and transcriptome sequencing of the same cell. *Nat. Biotechnol.* 33, 285–289.

Di Micco, R., Fontanals-Cirera, B., Low, V., Ntziachristos, P., Yuen, S.K., Lovell, C.D., Dolgalev, I., Yonekubo, Y., Zhang, G., Rusinova, E., et al. (2014). Control of Embryonic Stem Cell Identity by BRD4-Dependent Transcriptional Elongation of Super-Enhancer-Associated Pluripotency Genes. *Cell Rep.* 9, 234–247.

Ding, S., Wu, X., Li, G., Han, M., Zhuang, Y., and Xu, T. (2005). Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. *Cell* 122, 473–483.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* 167, 1853–1866.e17.

Ernst, J., Kheradpour, P., Mikkelson, T.S., Shores, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473, 43–49.

Fan, X., Kim, H.-J., Bouton, D., Warner, M., and Gustafsson, J.-Å. (2008). Expression of liver X receptor β is essential for formation of superficial cortical layers and migration of later-born neurons. *Proc. Natl. Acad. Sci.* *105*, 13445–13450.

Filippakopoulos, P., Qi, J., Picaud, S., Shen, Y., Smith, W.B., Fedorov, O., Morse, E.M., Keates, T., Hickman, T.T., Felletar, I., et al. (2010). Selective inhibition of BET bromodomains. *Nature* *468*, 1067–1073.

Fincher, C.T., Wurtzel, O., de Hoog, T., Kravarik, K.M., and Reddien, P.W. (2018). Cell type transcriptome atlas for the planarian *Schmidtea mediterranea*. *Science* *20*, eaaq1736–757.

Fogarty, N.M.E., McCarthy, A., Snijders, K.E., Powell, B.E., Kubikova, N., Blakeley, P., Lea, R., Elder, K., Wamaitha, S.E., Kim, D., et al. (2017). Genome editing reveals a role for OCT4 in human embryogenesis. *Nature* *9*, 346.

Folkerts, H., Wierenga, A.T., van den Heuvel, F.A., Woldhuis, R.R., Kluit, D.S., Jaques, J., Schuringa, J.J., and Vellenga, E. (2019). Elevated VMP1 expression in acute myeloid leukemia amplifies autophagy and is protective against venetoclax-induced apoptosis. *Cell Death Dis.* *10*, 421.

Fournier, M., Bourriquen, G., Lamaze, F.C., Côté, M.C., Fournier, É., Joly-Beauparlant, C., Caron, V., Gobeil, S., Droit, A., and Bilodeau, S. (2016). FOXA and master transcription factors recruit Mediator and Cohesin to the core transcriptional regulatory circuitry of cancer cells. *Sci. Rep.* *6*, 34962.

Fulco, C.P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S.R., Perez, E.M., Kane, M., Cleary, B., Lander, E.S., and Engreitz, J.M. (2016). Systematic mapping of functional enhancer–promoter connections with CRISPR interference. *Science* 354, 769–773.

Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., et al. (2009). An oestrogen-receptor- α -bound human chromatin interactome. *Nature* 462, 58–64.

Garcia-Carpizo, V., Ruiz-Llorente, S., Sarmentero, J., Graña-Castro, O., Pisano, D.G., and Barrero, M.J. (2018). CREBBP/EP300 bromodomains are critical to sustain the GATA1/MYC regulatory axis in proliferation. *Epigenetics Chromatin* 11, 30.

Gasperini, M., Hill, A.J., McFaline-Figueroa, J.L., Martin, B., Kim, S., Zhang, M.D., Jackson, D., Leith, A., Schreiber, J., Noble, W.S., et al. (2019). A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell*.

Gogol-Döring, A., Ammar, I., Gupta, S., Bunse, M., Miskey, C., Chen, W., Uckert, W., Schulz, T.F., Izsvák, Z., and Ivics, Z. (2016). Genome-wide Profiling Reveals Remarkable Parallels Between Insertion Site Selection Properties of the MLV Retrovirus and the piggyBac Transposon in Primary Human CD4⁺ T Cells. *Mol. Ther.* 24, 592–606.

Gonen, N., Futtner, C.R., Wood, S., Alexandra Garcia-Moreno, S., Salamone, I.M., Samson, S.C., Sekido, R., Poulat, F., Maatouk, D.M., and Lovell-Badge, R. (2018). Sex reversal following deletion of a single distal enhancer of Sox9. *Science* 360, 1469–1471.

Greil, F., Moorman, C., and van Steensel, B. (2006). DamID: Mapping of In Vivo Protein–Genome Interactions Using Tethered DNA Adenine Methyltransferase. In *DNA Microarrays, Part A: Array Platforms and Wet-Bench Protocols*, (Elsevier), pp. 342–359.

Grosselin, K., Durand, A., Marsolier, J., Poitou, A., Marangoni, E., Nemati, F., Dahmani, A., Lameiras, S., Reyat, F., Frenoy, O., et al. (2019). High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nat. Genet.* *51*, 1060–1066.

Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W. (2007). Quantifying similarity between motifs. *Genome Biol.* *8*, R24.

Gurdon, J.B. (2016). Cell Fate Determination by Transcription Factors. In *Essays on Developmental Biology, Part A*, (Elsevier), pp. 445–454.

Hafler, B.P., Surzenko, N., Beier, K.T., Punzo, C., Trimarchi, J.M., Kong, J.H., and Cepko, C.L. (2012). Transcription factor Olig2 defines subpopulations of retinal progenitor cells biased toward specific cell fates. *Proc. Natl. Acad. Sci.* *109*, 7882–7887.

Hainer, S.J., Bošković, A., McCannell, K.N., Rando, O.J., and Fazzio, T.G. (2019). Profiling of Pluripotency Factors in Single Cells and Early Embryos. *Cell* *177*, 1319–1329.e11.

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., et al. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* *172*, 1091–1107.e17.

Harada, A., Maehara, K., Handa, T., Arimura, Y., Nogami, J., Hayashi-Takanaka, Y., Shirahige, K., Kurumizaka, H., Kimura, H., and Ohkawa, Y. (2019). A chromatin integration labelling method enables epigenomic profiling with lower input. *Nat. Cell Biol.* *21*, 287–296.

Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., et al. (2006). The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* *34*, D590–D598.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-Enhancers in the Control of Cell Identity and Disease. *Cell* *155*, 934–947.

Ho, T.T., Warr, M.R., Adelman, E.R., Lansinger, O.M., Flach, J., Verovskaya, E.V., Figueroa, M.E., and Passegué, E. (2017). Autophagy maintains the metabolism and function of young and old stem cells. *Nature* *543*, 205–210.

Hothorn, T., Bretz, F., and Westfall, P. (2008). Simultaneous Inference in General Parametric Models. *Biom. J.* *50*, 346–363.

Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* *9*, 90–95.

Hwang, B., Lee, J.H., and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* *50*, 96.

Ivics, Z., Hackett, P.B., Plasterk, R.H., and Izsvák, Z. (1997). Molecular Reconstruction of Sleeping Beauty, a Tc1-like Transposon from Fish, and Its Transposition in Human Cells. *Cell* *91*, 501–510.

Jackman, J., and O’Connor, P.M. (1998). Methods for Synchronizing Cells at Specific Stages of the Cell Cycle. *Curr. Protoc. Cell Biol.* *00*, 8.3.1-8.3.20.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science* 316, 1497–1502.

Kalhor, R., Kalhor, K., Mejia, L., Leeper, K., Graveline, A., Mali, P., and Church, G.M. (2018). Developmental barcoding of whole mouse via homing CRISPR. *Science* 361, eaat9804.

Karaiskos, N., Wahle, P., Alles, J., Boltengagen, A., Ayoub, S., Kipar, C., Kocks, C., Rajewsky, N., and Zinzen, R.P. (2017). The *Drosophila* embryo at single-cell transcriptome resolution. *Science* 358, 194–199.

Kaya-Okur, H.S., Wu, S.J., Codomo, C.A., Pledger, E.S., Bryson, T.D., Henikoff, J.G., Ahmad, K., and Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* 10, 1930.

Kettlun, C., Galvan, D.L., George, A.L., Kaja, A., and Wilson, M.H. (2011). Manipulating piggyBac transposon chromosomal integration site selection in human cells. *Mol. Ther. J. Am. Soc. Gene Ther.* 19, 1636–1644.

Kfoury, N., Qi, Z., Wilkinson, M., Broestl, L., Berrett, K., Moudgil, A., Sankararaman, S., Chen, X., Gertz, J., Mitra, R., et al. (2017). Brd4-bound enhancers drive critical sex differences in glioblastoma (bioRxiv).

Kind, J., Pagie, L., Ortazokoyun, H., Boyle, S., de Vries, S.S., Janssen, H., Amendola, M., Nolen, L.D., Bickmore, W.A., and van Steensel, B. (2013). Single-Cell Dynamics of Genome-Nuclear Lamina Interactions. *Cell* 153, 178–192.

Kind, J., Pagie, L., de Vries, S.S., Nahidiazar, L., Dey, S.S., Bienko, M., Zhan, Y., Lajoie, B., de Graaf, C.A., Amendola, M., et al. (2015). Genome-wide Maps of Nuclear Lamina Interactions in Single Human Cells. *Cell* *163*, 134–147.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* *161*, 1187–1201.

Knoechel, B., Roderick, J.E., Williamson, K.E., Zhu, J., Lohr, J.G., Cotton, M.J., Gillespie, S.M., Fernandez, D., Ku, M., Wang, H., et al. (2014). An epigenetic mechanism of resistance to targeted therapy in T cell acute lymphoblastic leukemia. *Nat. Genet.* *46*, 364–370.

Kvon, E.Z., Kamneva, O.K., Melo, U.S., Barozzi, I., Osterwalder, M., Mannion, B.J., Tissières, V., Pickle, C.S., Plajzer-Frick, I., Lee, E.A., et al. (2016). Progressive Loss of Function in a Limb Enhancer during Snake Evolution. *Cell* *167*, 633–642.e11.

Lalli, M.A., Avey, D., Dougherty, J.D., Milbrandt, J., and Mitra, R.D. (2019). High-throughput single-cell functional elucidation of neurodevelopmental disease-associated genes reveals convergent mechanisms altering neuronal differentiation (bioRxiv).

Lawrence, M., Daujat, S., and Schneider, R. (2016). Lateral Thinking: How Histone Modifications Regulate Gene Expression. *Mol. Cell* *32*, 42–56.

Lee, T.I., and Young, R.A. (2013). Transcriptional Regulation and Its Misregulation in Disease. *Cell* *152*, 1237–1251.

Lee, C.S., Friedman, J.R., Fulmer, J.T., and Kaestner, K.H. (2005). The initiation of liver development is dependent on Foxa transcription factors. *Nature* 435, 944–947.

Litzenburger, U.M., Buenrostro, J.D., Wu, B., Shen, Y., Sheffield, N.C., Kathiria, A., Greenleaf, W.J., and Chang, H.Y. (2017). Single-cell epigenomic variability reveals functional cancer heterogeneity. *Genome Biol.* 18, 15.

Liu, X., Huang, J., Chen, T., Wang, Y., Xin, S., Li, J., Pei, G., and Kang, J. (2008). Yamanaka factors critically regulate the developmental signaling network in mouse embryonic stem cells. *Cell Res.* 18, 1177–1189.

Lovén, J., Hoke, H.A., Lin, C.Y., Lau, A., Orlando, D.A., Vakoc, C.R., Bradner, J.E., Lee, T.I., and Young, R.A. (2013). Selective Inhibition of Tumor Oncogenes by Disruption of Super-Enhancers. *Cell* 153, 320–334.

Lozzio, C., and Lozzio, B. (1975). Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome. *Blood* 45, 321–334.

Macaulay, I.C., Haerty, W., Kumar, P., Li, Y.I., Hu, T.X., Teng, M.J., Goolam, M., Saurat, N., Coupland, P., Shirley, L.M., et al. (2015). G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* 12, 519–522.

Machanick, P., and Bailey, T.L. (2011). MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* 27, 1696–1697.

Macleod, D., Charlton, J., Mullins, J., and Bird, A.P. (1994). Sp1 sites in the mouse *aprt* gene promoter are required to prevent methylation of the CpG island. *Genes Dev.* 8, 2282–2292.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* *161*, 1202–1214.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal* *17*, 10–12.

Martins, A.L., Walavalkar, N.M., Anderson, W.D., Zang, C., and Guertin, M.J. (2018). Universal correction of enzymatic sequence bias reveals molecular signatures of protein/DNA interactions. *Nucleic Acids Res.* *46*, e9–e9.

Matatall, K.A., Agapova, O.A., Onken, M.D., Worley, L.A., Bowcock, A.M., and Harbour, J.W. (2013). BAP1 deficiency causes loss of melanocytic cell identity in uveal melanoma. *BMC Cancer* *13*, 371.

Mátés, L., Chuah, M.K.L., Belay, E., Jerchow, B., Manoj, N., Acosta-Sanchez, A., Grzela, D.P., Schmitt, A., Becker, K., Matrai, J., et al. (2009). Molecular evolution of a novel hyperactive *Sleeping Beauty* transposase enables robust stable gene transfer in vertebrates. *Nat. Genet.* *41*, 753–761.

McClelland, M.L., Mesh, K., Lorenzana, E., Chopra, V.S., Segal, E., Watanabe, C., Haley, B., Mayba, O., Yaylaoglu, M., Gnad, F., et al. (2016). CCAT1 is an enhancer-templated RNA that predicts BET sensitivity in colorectal cancer. *J. Clin. Invest.* *126*, 639–652.

Meir, Y.-J.J., Weirauch, M.T., Yang, H.-S., Chung, P.-C., Yu, R.K., and Wu, S.C.-Y. (2011). Genome-wide target profiling of piggyBac and Tol2 in HEK 293: pros and cons for gene discovery and gene therapy. *BMC Biotechnol.* *11*, 28.

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., and Thomas, P.D. (2017). PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.* *45*, D183–D189.

Mizuguchi, R., Sugimori, M., Takebayashi, H., Kosako, H., Nagao, M., Yoshida, S., Nabeshima, Y., Shimamura, K., and Nakafuku, M. (2001). Combinatorial Roles of Olig2 and Neurogenin2 in the Coordinated Induction of Pan-Neuronal and Subtype-Specific Properties of Motoneurons. *Neuron* *31*, 757–771.

Molyneaux, B.J., Arlotta, P., Menezes, J.R.L., and Macklis, J.D. (2007). Neuronal subtype specification in the cerebral cortex. *Nat. Rev. Neurosci.* *8*, 427–437.

Najafova, Z., Tirado-Magallanes, R., Subramaniam, M., Hossan, T., Schmidt, G., Nagarajan, S., Baumgart, S.J., Mishra, V.K., Bedi, U., Hesse, E., et al. (2017). BRD4 localization to lineage-specific enhancers is associated with a distinct transcription factor repertoire. *Nucleic Acids Res.* *45*, 127–141.

Oliphant, T.E. (2015). *Guide to NumPy* (Austin, Tex.: Continuum Press).

Peterson, V.M., Zhang, K.X., Kumar, N., Wong, J., Li, L., Wilson, D.C., Moore, R., McClanahan, T.K., Sadekova, S., and Klappenbach, J.A. (2017). Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.* *9*, 2579.

Philipsen, S., and Suske, G. (1999). A tale of three fingers: the family of mammalian Sp/XKLF transcription factors. *Nucleic Acids Res.* *27*, 2991–3000.

Picelli, S., Björklund, A., Asa K., Reinius, B., Sagasser, S., Winberg, G., and Sandberg, R. (2014). Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* *24*, 2033–2040.

Ponnaluri, V.K.C., Zhang, G., Estève, P.-O., Spracklin, G., Sian, S., Xu, S., Benoukraf, T., and Pradhan, S. (2017). NicE-seq: high resolution open chromatin profiling. *Genome Biol.* *18*, 122.

Pott, S., and Lieb, J.D. (2014). What are super-enhancers? *Nat. Genet.* *47*, 8–12.

Pucilowska, J., Puzerey, P.A., Karlo, J.C., Galán, R.F., and Landreth, G.E. (2012). Disrupted ERK Signaling during Cortical Development Leads to Abnormal Progenitor Proliferation, Neuronal and Network Excitability and Behavior, Modeling Human Neuro-Cardio-Facial-Cutaneous and Related Syndromes. *J. Neurosci.* *32*, 8663–8677.

Qi, Z., Wilkinson, M.N., Chen, X., Sankararaman, S., Mayhew, D., and Mitra, R.D. (2017). An optimized, broadly applicable piggyBac transposon induction system. *Nucleic Acids Res.* gkw1290.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* *26*, 841–842.

Raff, T., van der Giet, M., Endemann, D., Wiederholt, T., and Paul, M. (1997). Design and Testing of β -Actin Primers for RT-PCR that Do Not Co-amplify Processed Pseudogenes. *BioTechniques* *23*, 456–460.

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* *44*, W160–W165.

Rašin, M.-R., Gazula, V.-R., Breunig, J.J., Kwan, K.Y., Johnson, M.B., Liu-Chen, S., Li, H.-S., Jan, L.Y., Jan, Y.-N., Rakic, P., et al. (2007). Numb and Numbl are required for maintenance of cadherin-based adhesion and polarity of neural progenitors. *Nat. Neurosci.* *10*, 819–827.

Rathert, P., Roth, M., Neumann, T., Muerdter, F., Roe, J.-S., Muhar, M., Deswal, S., Cerny-Reiterer, S., Peter, B., Jude, J., et al. (2015). Transcriptional plasticity promotes primary and acquired resistance to BET inhibition. *Nature* *525*, 543–547.

Robitaille, T.P., Tollerud, E.J., Greenfield, P., Droettboom, M., Bray, E., Aldcroft, T., Davis, M., Ginsburg, A., Price-Whelan, A.M., Kerzendorf, W.E., et al. (2013). Astropy: A community Python package for astronomy. *Astron. Astrophys.* *558*, A33.

Rodriguez-Fraticelli, A.E., Wolock, S.L., Weinreb, C.S., Panero, R., Patel, S.H., Jankovic, M., Sun, J., Calogero, R.A., Klein, A.M., and Camargo, F.D. (2018). Clonal analysis of lineage fate in native haematopoiesis. *Nature* *124*, 1929.

Rooijers, K., Markodimitraki, C.M., Rang, F.J., de Vries, S.S., Chialastri, A., de Luca, K.L., Mooijman, D., Dey, S.S., and Kind, J. (2019). Simultaneous quantification of protein–DNA contacts and transcriptomes in single cells. *Nat. Biotechnol.* *37*, 766–772.

Ropolo, A., Grasso, D., Pardo, R., Sacchetti, M.L., Archange, C., Re, A.L., Seux, M., Nowak, J., Gonzalez, C.D., Iovanna, J.L., et al. (2007). The Pancreatitis-induced Vacuole Membrane Protein 1 Triggers Autophagy in Mammalian Cells. *J. Biol. Chem.* *282*, 37124–37133.

Rosenberg, A.B., Roco, C.M., Muscat, R.A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L.T., Peeler, D.J., Mukherjee, S., Chen, W., et al. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* 360, 176–182.

Rotem, A., Ram, O., Shoresh, N., Sperling, R.A., Goren, A., Weitz, D.A., and Bernstein, B.E. (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* 33, 1165–1172.

Rouillard, A.D., Gundersen, G.W., Fernandez, N.F., Wang, Z., Monteiro, C.D., McDermott, M.G., and Ma'ayan, A. (2016). The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* 2016, baw100.

Sabari, B.R., Dall'Agnese, A., Boija, A., Klein, I.A., Coffey, E.L., Shrinivas, K., Abraham, B.J., Hannett, N.M., Zamudio, A.V., Manteiga, J.C., et al. (2018). Coactivator condensation at super-enhancers links phase separation and gene control. *Science* 361, eaar3958.

Sanson, K.R., Hanna, R.E., Hegde, M., Donovan, K.F., Strand, C., Sullender, M.E., Vaimberg, E.W., Goodale, A., Root, D.E., Piccioni, F., et al. (2018). Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. *Nat. Commun.* 9, 5416.

Saridey, S.K., Liu, L., Doherty, J.E., Kaja, A., Galvan, D.L., Fletcher, B.S., and Wilson, M.H. (2009). PiggyBac transposon-based inducible gene expression in vivo after somatic cell gene transfer. *Mol. Ther. J. Am. Soc. Gene Ther.* 17, 2115–2120.

Saunders, A., Macosko, E.Z., Wysoker, A., Goldman, M., Krienen, F.M., de Rivera, H., Bien, E., Baum, M., Bortolin, L., Wang, S., et al. (2018). Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell* 174, 1015–1030.e16.

Saxena, A., Wagatsuma, A., Noro, Y., Kuji, T., Asaka-Oba, A., Watahiki, A., Gurnot, C., Fagiolini, M., Hensch, T.K., and Carninci, P. (2012). Trehalose-enhanced isolation of neuronal sub-types from adult mouse brain. *BioTechniques* 52, 381–385.

Scargle, J.D., Norris, J.P., Jackson, B., and Chiang, J. (2013). STUDIES IN ASTRONOMICAL TIME SERIES ANALYSIS. VI. BAYESIAN BLOCK REPRESENTATIONS. *Astrophys. J.* 764, 167.

Scheiber, I.F., and Dringen, R. (2013). Astrocyte functions in the copper homeostasis of the brain. *Neurochem. Int.* 62, 556–565.

Schuster, D.J., Dykstra, J.A., Riedl, M.S., Kitto, K.F., Belur, L.R., McIvor, R.S., Elde, R.P., Fairbanks, C.A., and Vulchanova, L. (2014). Biodistribution of adeno-associated virus serotype 9 (AAV9) vector after intrathecal and intravenous delivery in mouse. *Front. Neuroanat.* 8, 42.

Sdelci, S., Rendeiro, A.F., Rathert, P., You, W., Lin, J.-M.G., Ringler, A., Hofstätter, G., Moll, H.P., Gürtl, B., Farlik, M., et al. (2019). MTHFD1 interaction with BRD4 links folate metabolism to transcriptional regulation. *Nat. Genet.* 51, 990–998.

Seabold, S., and Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In 9th Python in Science Conference, p.

Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* 14, 618–630.

Shema, E., Bernstein, B.E., and Buenrostro, J.D. (2018). Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution. *Nat. Genet.* 51, 19–25.

Sloan, C.A., Chan, E.T., Davidson, J.M., Malladi, V.S., Strattan, J.S., Hitz, B.C., Gabdank, I., Narayanan, A.K., Ho, M., Lee, B.T., et al. (2016). ENCODE data at the ENCODE portal. *Nucleic Acids Res.* *44*, D726–D732.

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K., Swerdlow, H., Satija, R., and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* *14*, 865–868.

Stoeckius, M., Zheng, S., Houck-Loomis, B., Hao, S., Yeung, B.Z., Mauck, W.M., Smibert, P., and Satija, R. (2018). Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* *19*, 224.

Stroud, H., Su, S.C., Hrvatin, S., Greben, A.W., Renthal, W., Boxer, L.D., Nagy, M.A., Hochbaum, D.R., Kinde, B., Gabel, H.W., et al. (2017). Early-Life Gene Expression in Neurons Modulates Lasting Epigenetic States. *Cell* *171*, 1151–1164.e16.

Sun, J., Ramos, A., Chapman, B., Johnnidis, J.B., Le, L., Ho, Y.-J., Klein, A., Hofmann, O., and Camargo, F.D. (2014). Clonal dynamics of native haematopoiesis. *Nature* *514*, 322–327.

Svensson, V., Vento-Tormo, R., and Teichmann, S.A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* *13*, 599–604.

Takahashi, K., and Yamanaka, S. (2006). Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* *126*, 663–676.

Tasic, B., Yao, Z., Graybuck, L.T., Smith, K.A., Nguyen, T.N., Bertagnolli, D., Goldy, J., Garren, E., Economo, M.N., Viswanathan, S., et al. (2018). Shared and distinct transcriptomic cell types across neocortical areas. *Nature* *563*, 72–78.

The Astropy Collaboration, Price-Whelan, A.M., Sipőcz, B.M., Günther, H.M., Lim, P.L., Crawford, S.M., Conseil, S., Shupe, D.L., Craig, M.W., Dencheva, N., et al. (2018). The Astropy Project: Building an inclusive, open-science project and status of the v2.0 core package. *Astron. J.* *156*, 123.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* *489*, 57–74.

Vassilev, L.T., Tovar, C., Chen, S., Knezevic, D., Zhao, X., Sun, H., Heimbrook, D.C., and Chen, L. (2006). Selective small-molecule inhibitor reveals critical mitotic functions of human CDK1. *Proc. Natl. Acad. Sci.* *103*, 10660–10665.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* *17*, 261–272.

Vogel, M.J., Peric-Hupkes, D., and van Steensel, B. (2007). Detection of in vivo protein–DNA interactions using DamID in mammalian cells. *Nat. Protoc.* *2*, 1467–1478.

Wang, H., Johnston, M., and Mitra, R.D. (2007). Calling cards for DNA-binding proteins. *Genome Res.* *17*, 1202–1209.

Wang, H., Mayhew, D., Chen, X., Johnston, M., and Mitra, R.D. (2011). Calling Cards enable multiplexed identification of the genomic targets of DNA-binding proteins. *Genome Res.* *21*, 748–755.

Wang, H., Mayhew, D., Chen, X., Johnston, M., and Mitra, R.D. (2012a). “Calling Cards” for DNA-Binding Proteins in Mammalian Cells. *Genetics* *190*, 941–949.

Wang, Q., Xiong, H., Ai, S., Yu, X., Liu, Y., Zhang, J., and He, A. (2019). CoBATCH for High-Throughput Single-Cell Epigenomic Profiling. *Mol. Cell* *76*, 206-216.e7.

Wang, W., Lin, C., Lu, D., Ning, Z., Cox, T., Melvin, D., Wang, X., Bradley, A., and Liu, P. (2008). Chromosomal transposition of PiggyBac in mouse embryonic stem cells. *Proc. Natl. Acad. Sci. U. S. A.* *105*, 9290–9295.

Wang, X., Spandidos, A., Wang, H., and Seed, B. (2012b). PrimerBank: a PCR primer database for quantitative gene expression analysis, 2012 update. *Nucleic Acids Res.* *40*, D1144–D1149.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell* *153*, 307–319.

Wilson, M.H., Coates, C.J., and George, A.L. (2007). PiggyBac transposon-mediated gene transfer in human cells. *Mol. Ther. J. Am. Soc. Gene Ther.* *15*, 139–145.

Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY : large-scale single-cell gene expression data analysis. *Genome Biol.* *19*, 15.

Wu, S.C.-Y., Meir, Y.-J.J., Coates, C.J., Handler, A.M., Pelczar, P., Moisyadi, S., and Kaminski, J.M. (2006). piggyBac is a flexible and highly active transposon as compared to sleeping beauty, Tol2, and Mos1 in mammalian cells. *Proc. Natl. Acad. Sci.* *103*, 15008–15013.

Xie, S., Duan, J., Li, B., Zhou, P., and Hon, G.C. (2017). Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Mol. Cell* *66*, 285-299.e5.

Yen, L., Svendsen, J., Lee, J.-S., Gray, J.T., Magnier, M., Baba, T., D'Amato, R.J., and Mulligan, R.C. (2004). Exogenous control of mammalian gene expression through modulation of RNA self-cleavage. *Nature* *431*, 471–476.

Yen, M., Qi, Z., Chen, X., Cooper, J.A., Mitra, R.D., and Onken, M.D. (2018). Transposase mapping identifies the genomic targets of BAP1 in uveal melanoma. *BMC Med. Genomics* *11*, 97.

Yoshida, J., Akagi, K., Misawa, R., Kokubu, C., Takeda, J., and Horie, K. (2017). Chromatin states shape insertion profiles of the piggyBac, Tol2 and Sleeping Beauty transposons and murine leukemia virus. *Sci. Rep.* *7*, 43613.

Yu, H., Mashtalir, N., Daou, S., Hammond-Martel, I., Ross, J., Sui, G., Hart, G.W., Rauscher, F.J., Drobetsky, E., Milot, E., et al. (2010). The Ubiquitin Carboxyl Hydrolase BAP1 Forms a Ternary Complex with YY1 and HCF-1 and Is a Critical Regulator of Gene Expression. *Mol. Cell. Biol.* *30*, 5071–5085.

Yusa, K., Zhou, L., Li, M.A., Bradley, A., and Craig, N.L. (2011). A hyperactive piggyBac transposase for mammalian applications. *Proc. Natl. Acad. Sci. U. S. A.* *108*, 1531–1536.

Zeisel, A., Munoz-Manchado, A.B., Codeluppi, S., Lonnerberg, P., La Manno, G., Jureus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347, 1138–1142.

Zeisel, A., Hochgerner, H., Lönnerberg, P., Johnsson, A., Memic, F., van der Zwan, J., Häring, M., Braun, E., Borm, L.E., La Manno, G., et al. (2018). Molecular Architecture of the Mouse Nervous System. *Cell* 174, 999–1014.e22.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* 9, R137.

Zhang, Y., Chen, K., Sloan, S.A., Bennett, M.L., Scholze, A.R., O’Keeffe, S., Phatnani, H.P., Guarnieri, P., Caneda, C., Ruderisch, N., et al. (2014). An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex. *J. Neurosci.* 34, 11929–11947.

Zheng, D., Liu, X., and Tian, B. (2016). 3’READS+, a sensitive and accurate method for 3’ end sequencing of polyadenylated RNA. *RNA* 22, 1631–1639.

Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8, 14049.

Zhou, B., Ho, S.S., Greer, S.U., Zhu, X., Bell, J.M., Arthur, J.G., Spies, N., Zhang, X., Byeon, S., Pattni, R., et al. (2019). Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562. *Genome Res.* 29, 472–484.

Zhou, X., Maricque, B., Xie, M., Li, D., Sundaram, V., Martin, E.A., Koebbe, B.C., Nielsen, C., Hirst, M., Farnham, P., et al. (2011). The Human Epigenome Browser at Washington University. *Nat. Methods* 8, 989–990.

Zhu, X., Zuo, H., Maher, B.J., Serwanski, D.R., LoTurco, J.J., Lu, Q.R., and Nishiyama, A. (2012). Olig2-dependent developmental fate switch of NG2 cells. *Development* 139, 2299–2307.

Zuber, J., Shi, J., Wang, E., Rappaport, A.R., Herrmann, H., Sison, E.A., Magoon, D., Qi, J., Blatt, K., Wunderlich, M., et al. (2011). RNAi screen identifies Brd4 as a therapeutic target in acute myeloid leukaemia. *Nature* 478, 524–528.

Chapter 3: The qBED track: a novel genome browser visualization for point processes

(A version of this chapter was published in *Bioinformatics* 37(8), pp. 1168-1170)

3.1 Abstract

3.1.1 Summary

Transposon calling cards is a genomic assay for identifying transcription factor binding sites in both bulk and single cell experiments. Here we describe the qBED format, an open, text-based standard for encoding and analyzing calling card data. In parallel, we introduce the qBED track on the WashU Epigenome Browser, a novel visualization that enables researchers to inspect calling card data in their genomic context. Finally, through examples, we demonstrate that qBED files can be used to visualize non-calling card datasets, such as CADD scores and GWAS/eQTL hits, and may have broad utility to the genomics community.

3.1.2 Availability and Implementation

The qBED track is available on the WashU Epigenome Browser

(<http://epigenomegateway.wustl.edu/browser>), beginning with version 46. Source code for the

WashU Epigenome Browser with qBED support is available on GitHub

(<http://github.com/arnavm/eg-react> and <http://github.com/lidaof/eg-react>). A complete definition

of the qBED format is available as part of the WashU Epigenome Browser documentation

(<https://eg.readthedocs.io/en/latest/tracks.html#qbed-track>). We have also released a tutorial on

how to upload qBED data to the browser ([dx.doi.org/10.17504/protocols.io.bca8ishw](https://doi.org/10.17504/protocols.io.bca8ishw)).

3.2 Introduction

Advances in genomic technologies often lead to new data formats and new platforms to visualize those data. The Human Genome Project originated the popular Browser Extensible Data (BED)

standard for describing genomic intervals (Kent et al., 2002). Routine next-generation sequencing projects, such as whole-genome sequencing and RNA-seq, use the SAM format to store and visualize data (Li et al., 2009). Epigenetic modifications detected by bisulfite sequencing can be visualized using methylC tracks (Zhou et al., 2014). The bedGraph and wiggle (.wig/.bigWig) formats have emerged as flexible standards for encoding pseudo-continuous integer- and real-valued signals across the genome, such as from normalized ChIP- or ATAC-seq assays (Huy Hoang and Sung, 2014; Kent et al., 2010; Rosenbloom et al., 2010). Finally, the .hic and .cool formats (Abdennur and Mirny, 2019; Durand et al., 2016) encapsulate intra-chromosomal contact frequencies and have contributed to our understanding of chromatin organization.

Over the past several years, we have introduced, and developed, transposon calling cards to identify genome-wide transcription factor (TF) binding sites (TFBS) (Wang et al., 2007, 2008, 2011, 2012). This approach uses a TF of interest fused to a transposase. The fusion construct deposits transposons into the genome near TFBS, which can be recovered from either DNA or RNA libraries. Significantly enriched clusters of transposons indicate putative TFBS. Instead of plotting read coverage, as would be done in more traditional TF studies like ChIP-seq, we plot each insertion as a discrete point along the (genomic) x -axis and the number of reads supporting that particular insertion on the y -axis. The result resembles a scatterplot in which an increased density of insertions is typically observed near TFBS.

Historically, raw insertion data were visualized using GNASHY, an in-house file format and genome browser custom built for calling card data. While useful, the GNASHY browser suffered from two major limitations: first, it was restricted to visualizing one track—and therefore, one sample or experiment—at a time; and second, it did not support conventional genomic

formats like bedGraph or bigWig. Thus, any comparative analysis of calling card data with, say, ChIP-seq or ATAC-seq relied on manually aligning images from different browsers (Wang et al., 2012).

Calling card technology is currently undergoing a renaissance. We have recently used calling cards to study TF binding in bulk populations of cells *in vivo* (Cammack et al., 2020), and we have also combined calling cards with single cell RNA-seq to simultaneously profile cell identity in complex organs and heterogenous disease states (Moudgil et al., 2020). Calling cards has also been used to dissect TF binding in both steady state and dynamic contexts (Mayhew and Mitra, 2014; Shively et al., 2019). As the scope and application of the calling card technique grows, we anticipate greater interest and increasingly complex visualization demands. Here, to better support existing and future users, we describe the qBED format, a new text-based genomic data format for storing calling card data. We also describe the qBED track, an interface for visualizing calling card data on the WashU Epigenome Browser. Finally, we present examples of non-calling card genomic data visualized using the qBED standard to demonstrate the format's flexibility.

3.3 Implementation

We christened our format *q*BED because it stores multidimensional, *quantitative* information about *quantized* events, such as calling card transpositions. Formally, qBED follows a BED3+3 standard (Figure 3.1A). For calling card data, the first three columns denote the chromosome, start, and end coordinates of the transposon insertion. The width of the interval depends on the transposase used: mammalian calling cards, which employs the *piggyBac* transposase, uses a four base-pair width for the insertion coordinate as *piggyBac* overwhelmingly inserts into TTTA tetramers (Wang et al., 2012); whereas yeast calling cards uses single base-pair intervals as these

assays use the motif-agnostic Ty5 retrotransposon (Wang et al., 2007). qBED files inherit the BED format's 0-based, half-open intervals and are compatible with programs like bedtools (Quinlan and Hall, 2010) and bedops (Neph et al., 2012) for intersection analysis. The fourth column encodes a numerical value—in this case, the number of reads supporting each insertion—and is the last column required in qBED files. The fifth and sixth columns are optional, but recommended, as the former denotes the strand (+/-, or . if unspecified) that was targeted, while the latter encodes an annotation string. For calling card experiments, this is where sample-specific barcodes are registered (Figure 3.1A). Like BED files, qBED files can be compressed and indexed with bgzip and tabix, respectively (Li et al., 2009).

To visualize qBED files, we have created the qBED track and implemented it in the WashU Epigenome Browser (Li et al., 2019; Zhou et al., 2011), a leading portal for analyzing epigenomic data such as ChIP-seq, ATAC-seq, and Hi-C. (Prior to version 51.0.3, the qBED track was known as the calling card track). qBED tracks display circular markers for genomic features in two dimensions: genomic position along the x -axis and numerical value along the y -axis (Figure 3.1Bi). For calling card experiments, these represent transposon insertions and read counts, respectively. When the insertion coordinate spans more than one base, the marker is drawn at the midpoint of the interval. Moreover, as multiple insertions may occur at the same insertion site (e.g. from different replicates), multiple markers can co-occur at the same x -coordinate and stratify across the y -axis. qBED tracks support interactive exploration of data. As a cursor approaches a data point, a rollover pane appears (Figure 3.1Bii), displaying the read count, strand, and annotation (columns 4, 5 and 6, respectively). Near the top of the rollover pane is the track name and an approximate (to the nearest pixel) genomic location.

A

chr	start	end	value	strand	annotation
chr4	27464929	27464933	68	+	TGC
chr4	55548472	55548476	120	-	TAG
chr4	69653208	69653212	129	+	CTA
chr4	74667807	74667811	610	+	TGC
chr4	99980613	99980617	106	-	GAT



Figure 3.1: Overview of the qBED format and qBED tracks. (A) Example of a qBED file encoding transposon calling card data. The first three columns are inherited from the BED standard and encode the location of the insertion site. The fourth column stores the number of reads observed for each entry, while the fifth denotes strand. The sixth and final column is an annotation recording the sample-specific barcode for each insertion in the library. (B) Screenshot of qBED tracks depicting calling card data in the WashU Epigenome Browser. (i) qBED features appear on two-dimensional tracks, with genomic position along the x -axis and a numerical value on the y -axis (here, log-transformed read counts). (ii) An informational panel appears upon rollover of a calling card insertion, revealing read count, strand, barcode, and approximate location. (iii) Right-clicking on a qBED track pulls up a configuration panel. Tracks can be customized with respect to color, size, y -axis limits and transformations, marker size (iii-iv), opacity (v), and sample size (vi). (vii) Orthogonal datasets like transcription factor and histone ChIP-seq data can be directly displayed alongside calling card data.

Right-clicking on a qBED track leads to a customization panel (Figure 3.1Biii).

Individual tracks can be shaded in any RGB color (Figure 3.1Biii-vi), to better delineate different

samples. The size of the calling card marker can be made larger (Figure 3.1Biii) or smaller (Figure 3.1Biv), depending on user preference. The opacity of the track can also be adjusted (Figure 3.1Bv), which may help reveal structure in regions of pronounced insertion density. For very large datasets, a random subsample of the data can be displayed (Figure 3.1Bvi). This prevents overplotting of markers and can reduce the browser's memory consumption. Finally, and most importantly, the WashU Epigenome Browser enables calling card data to be natively visualized alongside other genomic datasets, such as ChIP-seq from the same cell type (Figure 3.1Bvii).

3.4 Applications

qBED files present genomic data as a discrete point process as opposed to a pseudo-continuous function of sequencing coverage. In addition to analyzing calling card experiments, this format may also be useful for existing genomic data types. Here we present two such examples.

Combined Annotation Dependent Depletion (CADD) scores integrate multiple streams of information to predict the deleteriousness of single nucleotide polymorphisms (SNPs) and indels (Kircher et al., 2014; Rentzsch et al., 2019). These are typically displayed as vertical lines depicting the maximum score observed for each base (Figure 3.2A). This approach, while useful as a summary statistic, does not allow for interactive exploration of individual mutations. We converted CADD scores for indels from variant call format (VCF) to a qBED file, using the numeric column to store the CADD score and the annotation column to store the mutation. When viewed on the WashU Epigenome Browser, individual polymorphisms can be inspected. A view of the homeobox gene *CRX* reveals a cluster of strongly deleterious indels in the terminal exon (Figure 3.2A). The qBED display emphasizes the density of variants along both the genomic (x)

and CADD (y) axes, offering an unvarnished look at the complete spectrum of deleteriousness in a dataset.

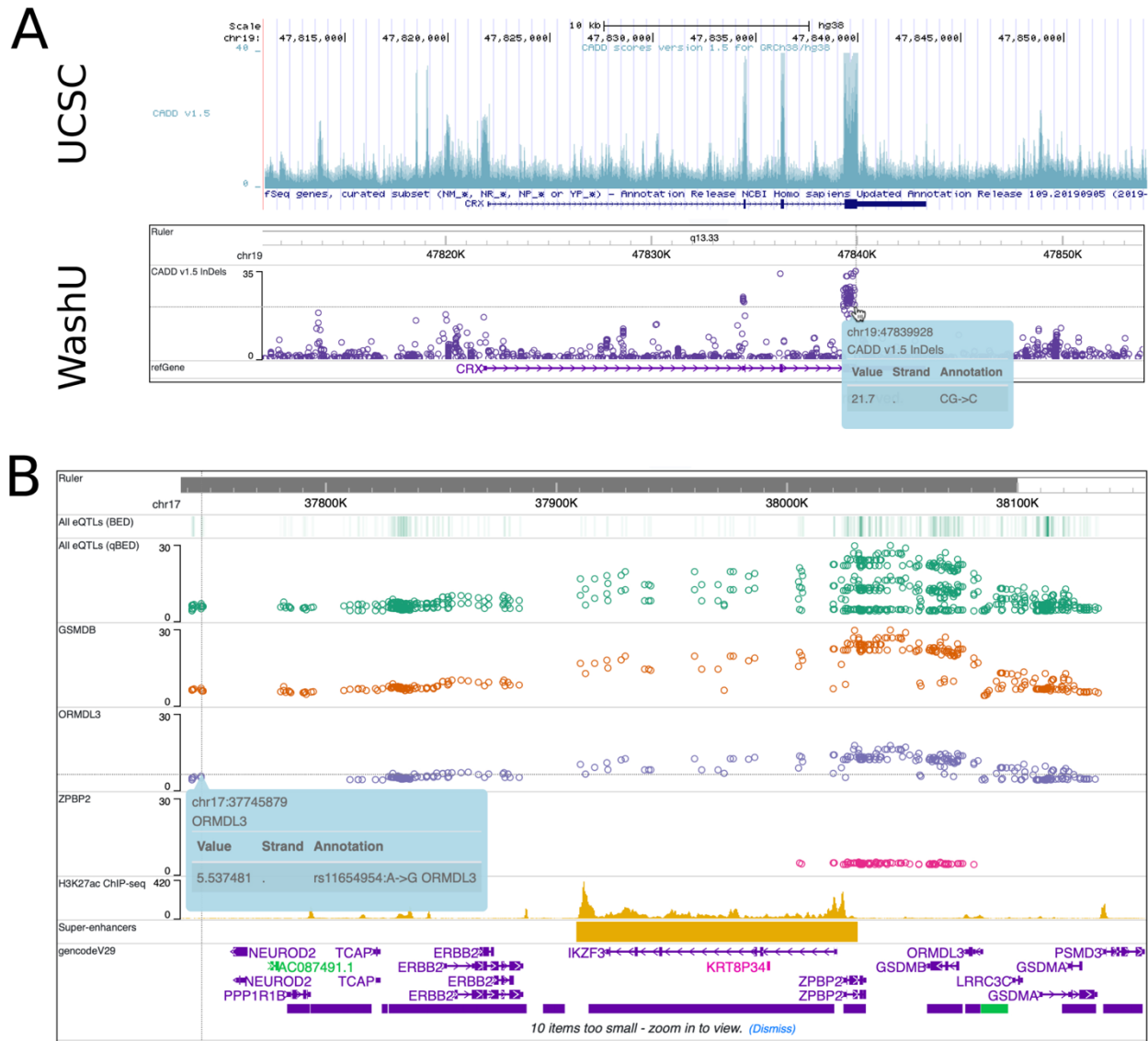


Figure 3.2: Application of the qBED specification to other genomic datasets. (A) Top: CADD scores for the gene *CRX*, as visualized on the UCSC Genome Browser. Bottom: CADD scores visualized on the WashU Epigenome Browser after conversion to qBED. Genomic position is along the x -axis and Phred-style CADD scores are along the y -axis. The mouseover pane reveals more information on an individual variant. (B) eQTLs for *CD20*⁺ B cells visualized as calling card tracks. The top track shows all significant eQTLs in view plotted as a BED (density) track, followed by a qBED representation of the same data. The y -axis represents the negative base-ten logarithm of the p -value. The next three tracks show significant eQTLs for the genes *GSDMB*, *ORMDL3*, and *ZPBP2*, respectively. Finally, we show H3K27ac ChIP-seq (coverage on the y -axis) and a super-enhancer for this cell type. A mouseover pane can reveal further details stored in the qBED file, including Reference SNP ID and mutation.

A second application of qBED files is in genome wide association studies (GWAS) and expression quantitative trait locus (eQTL) mapping, which aim to identify SNPs that are significantly correlated with either phenotypes or gene expression, respectively. Most significant SNPs fall in noncoding regions and their functional significance can be unclear (Gloss and Dinger, 2018; Tak and Farnham, 2015). One way to prioritize variants is by considering their regulatory and epigenetic context (Gloss and Dinger, 2018; Tak and Farnham, 2015); however, a quantitative view of SNPs is not supported by most genome browsers. Investigators either manually align separate views of SNPs with views of epigenetic profiles, or encode SNPs as BED tracks, which shows position but sacrifices the quantitative measure—usually the negative base-ten logarithm of the p -value—of the association (Farh et al., 2015).

We reasoned that the qBED track could display both the density and the quantitative value of SNPs in association studies. We used a publicly available eQTL dataset from CD20+ B cells (Schmiedel et al., 2018) and converted it to qBED format, storing the negative base-ten logarithm of the p -value of the eQTL association in the numeric column; and storing the reference SNP, mutation, and linked gene in the annotation field. We simultaneously plotted H3K27ac ChIP-seq data (Davis et al., 2018; The ENCODE Project Consortium, 2012) and a track of super-enhancers for the same cell type (Figure 3.2B). Such data would either have to be manually aligned with another browser shot or plotted as a BED track (shown) that only emphasizes the local density of variants. The qBED visualization shows both the density of variants and the significance of each variant, alongside epigenetic context, all in a single pane. We can also separate eQTLs by target gene and assign them to individual tracks, revealing how genes in close proximity to each other can have different eQTL effect sizes from the same genomic sequence. In particular, eQTLs associated with *GSDMB* and *ORMDL3* expression span

a large swath of flanking DNA, including overlapping an adjacent super-enhancer, while eQTLs associated with *ZFPBP2* expression are constrained to a much narrower segment. This example demonstrates how the qBED track can bridge the fields of association studies and epigenomics, simplifying certain kinds of analyses for researchers.

The qBED track is best positioned for exploring dense, quantitative data driven by high-resolution point processes. As such, it can be a useful complement to the popular lollipop track (Jay and Brouwer, 2016; Lee et al., 2019), either as a way to publish figures of raw data without the clutter of lollipop stems; or as a way to inspect data before choosing individual points for further annotation and emphasis. Moreover, by first specifying the x- and y- (columns 1-3 and 4, respectively) values, the qBED format prioritizes the two-dimensional relationship of the data. This may also have some advantages for data compression as the strand and annotation columns are not required and can be added if additional specificity is required. In contrast, interval-based formats like BED and ENCODE's tagAlign would require six fields to store the same data. Finally, where tagAlign stores the actual sequence of each read, qBED defers to sequence in the reference genome at the x coordinate. The annotation field can be used to record departures from the reference, similar to the way Variant Call Format (VCF) files encodes SNPs, but remains broadly flexible for the end user.

3.5 Conclusion

The qBED specification and the accompanying qBED track offer researchers the ability to visualize genomic point processes—such as transposon insertions, polymorphism deleteriousness, or phenotypic associations—by adding a numerical y-axis for stratifying features on the genomic x-axis. We envision investigators using this format not only for analyzing calling card experiments, but any data involving relatively small, quantitatively separable genomic features.

While we feel the six-column format presented here is complete enough for existing analyses, we leave open the possibility for future enhancements. In particular, extra columns could be added to encode secondary and/or tertiary information for each entry. These could be visualized, pending browser support, with either a numerical color scale, in the case of quantitative data, or different marker shapes, for categorical data.

3.6 References

Abdennur, N., and Mirny, L.A. (2019). Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* btz540.

Cammack, A.J., Moudgil, A., Chen, J., Vasek, M.J., Shabsovich, M., McCullough, K., Yen, A., Lagunas, T., Maloney, S.E., He, J., et al. (2020). A viral toolkit for recording transcription factor–DNA interactions in live mouse tissues. *Proc Natl Acad Sci USA* 201918241.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Research* 46, D794–D801.

Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems* 3, 95–98.

Farh, K.K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shores, N., Whitton, H., Ryan, R.J.H., Shishkin, A.A., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 518, 337–343.

Gloss, B.S., and Dinger, M.E. (2018). Realizing the significance of noncoding functionality in clinical genomics. *Exp Mol Med* 50, 97.

Huy Hoang, D., and Sung, W.-K. (2014). CWig: compressed representation of Wiggle/BedGraph format. *Bioinformatics* 30, 2543–2550.

Jay, J.J., and Brouwer, C. (2016). Lollipops in the Clinic: Information Dense Mutation Plots for Precision Medicine. *PLoS ONE* 11, e0160519.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The Human Genome Browser at UCSC. *Genome Research* 12, 996–1006.

Kent, W.J., Zweig, A.S., Barber, G., Hinrichs, A.S., and Karolchik, D. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* 26, 2204–2207.

Kircher, M., Witten, D.M., Jain, P., O’Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 46, 310–315.

Lee, C.M., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Gonzalez, J.N., Hinrichs, A.S., Lee, B.T., Nassar, L.R., Powell, C.C., et al. (2019). UCSC Genome Browser enters 20th year. *Nucleic Acids Research* gkz1012.

Li, D., Hsu, S., Purushotham, D., Sears, R.L., and Wang, T. (2019). WashU Epigenome Browser update 2019. *Nucleic Acids Research* 47, W158–W165.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079.

Mayhew, D., and Mitra, R.D. (2014). Transcription factor regulation and chromosome dynamics during pseudohyphal growth. *Molecular Biology of the Cell* 25, 2669–2676.

Moudgil, A., Wilkinson, M.N., Chen, X., He, J., Cammack, A.J., Vasek, M.J., Lagunas, T., Qi, Z., Lalli, M.A., Guo, C., et al. (2020). Self-Reporting Transposons Enable Simultaneous Readout of Gene Expression and Transcription Factor Binding in Single Cells. *Cell* S009286742030814X.

Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., Thomas, S., et al. (2012). BEDOPS: high-performance genomic feature operations. *Bioinformatics* 28, 1919–1920.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842.

Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research* 47, D886–D894.

Rosenbloom, K.R., Dreszer, T.R., Pheasant, M., Barber, G.P., Meyer, L.R., Pohl, A., Raney, B.J., Wang, T., Hinrichs, A.S., Zweig, A.S., et al. (2010). ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic Acids Research* 38, D620–D625.

Schmiedel, B.J., Singh, D., Madrigal, A., Valdovino-Gonzalez, A.G., White, B.M., Zapardiel-Gonzalo, J., Ha, B., Altay, G., Greenbaum, J.A., McVicker, G., et al. (2018). Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell* *175*, 1701-1715.e16.

Shively, C.A., Liu, J., Chen, X., Loell, K., and Mitra, R.D. (2019). Homotypic cooperativity and collective binding are determinants of bHLH specificity and function. *Proc Natl Acad Sci USA* *116*, 16143–16152.

Tak, Y.G., and Farnham, P.J. (2015). Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics & Chromatin* *8*, 57.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* *489*, 57–74.

Wang, H., Johnston, M., and Mitra, R.D. (2007). Calling cards for DNA-binding proteins. *Genome Research* *17*, 1202–1209.

Wang, H., Heinz, M.E., Crosby, S.D., Johnston, M., and Mitra, R.D. (2008). “Calling Cards” method for high-throughput identification of targets of yeast DNA-binding proteins. *Nature Protocols* *3*, 1569–1577.

Wang, H., Mayhew, D., Chen, X., Johnston, M., and Mitra, R.D. (2011). Calling Cards enable multiplexed identification of the genomic targets of DNA-binding proteins. *Genome Research* *21*, 748–755.

Wang, H., Mayhew, D., Chen, X., Johnston, M., and Mitra, R.D. (2012). “Calling Cards” for DNA-Binding Proteins in Mammalian Cells. *Genetics* 190, 941–949.

Zhou, X., Maricque, B., Xie, M., Li, D., Sundaram, V., Martin, E.A., Koebbe, B.C., Nielsen, C., Hirst, M., Farnham, P., et al. (2011). The Human Epigenome Browser at Washington University. *Nat Methods* 8, 989–990.

Zhou, X., Li, D., Lowdon, R.F., Costello, J.F., and Wang, T. (2014). methylC Track: visual integration of single-base resolution DNA methylation data on the WashU EpiGenome Browser. *Bioinformatics* 30, 2206–2207.

Chapter 4: Fast and optimal genome segmentation with Bayesian blocks

4.1 Abstract

Genome segmentation and peak calling are common tasks to identify regions enriched in a signal of interest. Typically, the signal is a dependent function of genomic position; examples include epigenetic marks assayed by ChIP-seq and accessible loci by ATAC-seq. Less common is the determination of peaks based on the local density of genomic features. Here, we present Bayesian blocks, an algorithm for optimally segmenting intervals based on the underlying density of data. While this algorithm was initially developed for astrophysics, we have adapted it for use in genomics. We first explain the mathematical foundations of the algorithm, including a linear runtime optimization. We next demonstrate Bayesian blocks on two genomic datasets: first, to call peaks in transposon calling cards data and identify transcription factor binding sites; and second, to identify CpG islands based on genomic sequence. We conclude that Bayesian blocks may be generally useful in genomics and have released an accompanying Python package (blockify) implementing support for genomic data formats.

4.2 Introduction

Genomic analysis frequently involves segmenting the genome into regions, often termed peaks, enriched for signals of interest. Peaks can identify fine-grained features, such as protein-bound DNA detected using chromatin immunoprecipitation and sequencing (ChIP-seq), or accessible loci detected by the assay for transposase-accessible chromatin (ATAC-seq) (Zhang et al., 2008). On a broader scale, segmentation can be used to identify open reading frames (Cleynen et al., 2014) and transcriptionally active regulatory elements (Wang et al., 2019) from RNA sequencing

(RNA-seq) data as well as larger domains of organized chromatin (Hansen et al., 2010; Keough et al., 2020) and structural variants such as duplications and deletions (Fan and Mackey, 2017).

Peak callers often look for a statistical enrichment of coverage, or read depth, relative to a control dataset or against flanking sequence (Zhang et al., 2008). Transcription factor (TF) ChIP-seq, for example, uses an antibody to pull down DNA crosslinked to a TF on interest (Johnson et al., 2007). Peaks are then called by comparing the relative amount of reads in the experimental track compared to the input control, which is typically the same sample but without antibody enrichment. Coverage is a pragmatic metric for assays that generate random genomic fragments, though read depth can be skewed by technical artifacts such as skewed representation and chimeras (Cha and Thilly, 1993; Kanagawa, 2003).

Over the past few years, we have developed transposon calling cards as an alternative assay to map TF binding sites (Wang et al., 2007, 2012). This technique uses a TF of interest fused to a transposase. As the TF visits its binding sites, the transposase deposits transposons nearby. Loci that show a high density of transpositions in the TF-transposase condition, but not in the undirected transposase dataset, represent sites likely bound by the TF. In contrast to ChIP-seq, we do not visualize our data as function of coverage. Rather, each transposition is drawn as a distinct marker (Figure 4.1). While insertions are stratified across two dimensions for visualization, we only consider the genomic coordinates for downstream analysis using count-based statistics.

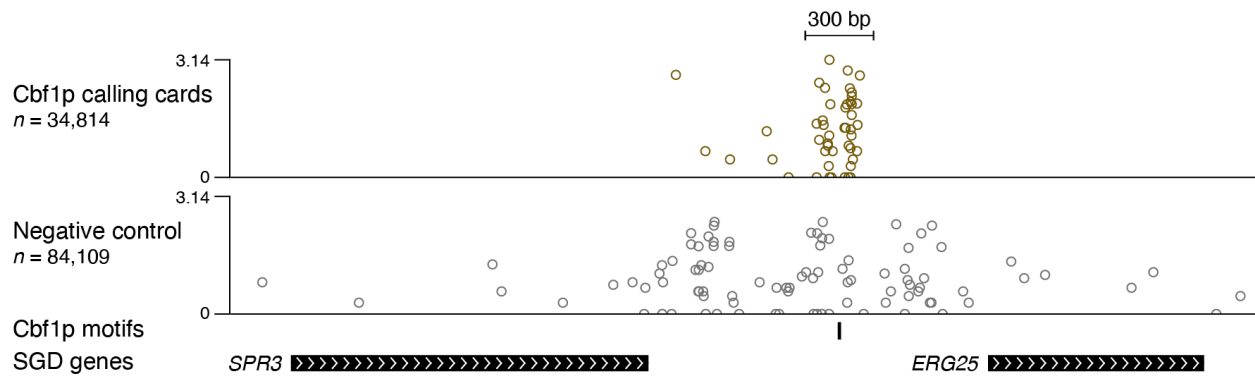


Figure 4.1: Example calling cards tracks. Two representative tracks from calling cards experiment. The top track are insertions from a Cbf1p-directed yeast calling cards experiment and the bottom track are insertions from the negative control. Insertions are drawn as circular markers with genomic coordinate on the x-axis and log₁₀-transformed read count on the y-axis. n denotes the total, genome-wide library size.

While calling cards can generate highly specific binding profiles, there are challenges with calling peaks on calling cards data. Mammalian calling cards uses the *piggyBac* transposase, which almost exclusively inserts into TTAA tetramers (Wang et al., 2012). As a result, the insertion profile can be sparse. Yeast calling cards can be deposited without such constraint, resulting in very smooth peaks. However, our analyses only looked for differences over entire promoters (Shively et al., 2019). One drawback of this is that we are not able to distinguish between sharp, TF-directed insertions against a diffuse background of undirected transpositions, leading to false negatives (Figure 4.1). Finally, while mammalian and yeast calling cards are philosophically identical and generate similar kinds of data, their analytical pipelines are divorced. We therefore set out to develop a new strategy for calling peaks that would be flexible enough to work with both yeast and mammalian datasets, but robust enough to handle sparsity.

We turned to Bayesian blocks, an algorithm developed by astrophysicists for counting photons (Scargle, 1998; Scargle et al., 2013). Among its strengths is the ability to generate a mathematically optimal segmentation of one-dimensional data by globally maximizing a piecewise Poisson likelihood function. It can also tolerate gaps, repeated values, and noise,

which makes it attractive for analyzing calling cards experiments. Calling card transpositions also share certain features with photons that make them appropriate for this approach, namely that they are discrete, independent events and can be aptly described by Poisson-based counting processes. We realized, as we began exploring Bayesian blocks, we that this algorithm may be generally useful in genomics. Indeed, Bayesian blocks has already made a few appearances in the genomics literature (Cang and Nie, 2020; Chan et al., 2017; Ish-Horowicz and Reid, 2017; Stumpf et al., 2017). However, we have not yet seen an introduction or implementation specifically tailored to the genomics community.

The remainder of the manuscript is structured as follows. We start by reviewing the mathematical foundations of the algorithm and implement an optimization strategy that greatly improves the runtime complexity. Next, we demonstrate how Bayesian blocks, in conjunction with control and experimental datasets, can be used as a peak caller for calling cards to identify TF binding sites. Finally, we turn Bayesian blocks onto a classic problem, identifying CpG islands from genomic sequence. We find that Bayesian blocks can be competitive with hidden Markov models (HMMs), the dominant paradigm for segmenting genomic data. We conclude that Bayesian blocks may have broad utility in genomics as a general-purpose approach to segmentation.

4.3 Results

4.3.1 Review of Bayesian blocks

Bayesian blocks seeks to segment a one-dimensional dataset into a series of contiguous blocks where the rate of events per block is piecewise-constant (Scargle et al., 2013). In doing so, it optimizes the global likelihood of the partition assuming a regularized Poisson likelihood function.

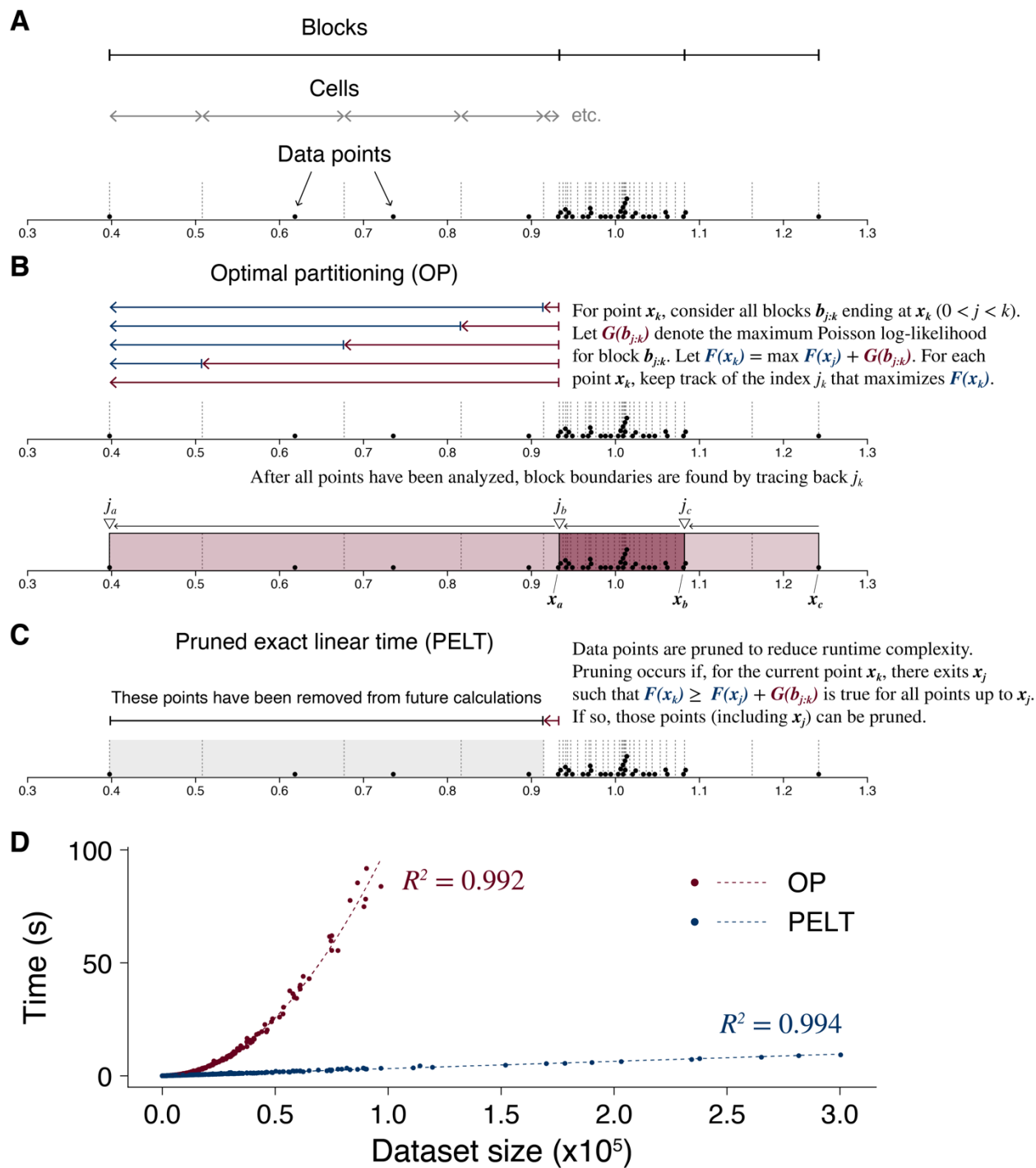


Figure 4.2: Overview of Bayesian blocks. (A) Sample dataset and terminology. Nearby data points are drawn arbitrarily stacked on top of each other. (B) Schematic of the Optimal Partitioning (OP) algorithm. (C) Explanation of the pruned exact linear time (PELT) optimization scheme. (D) Runtime comparison between OP and PELT on real-world calling cards datasets.

Formally, Bayesian blocks considers a set of ordered data points $X = \{x_1 \dots x_n\}$ along a one-dimensional interval (Figure 4.2A). The interval is divided into cells such that each unique

x_i belongs to a cell c_i , which is bounded by edges e_i and e_{i+1} . The cells form a Voronoi tessellation of X : $e_i = \frac{x_{i-1} + x_i}{2}$, with $e_1 = x_1$ and $e_{n+1} = x_n$. Blocks are composed of consecutive cells; we denote a block $b_{j:k} = \cup_{i=j}^k c_i$, where $j \leq k$. For a given block $b_{j:k}$, let $N_{j:k} = \sum_{i=j}^k 1$ represent the number of data points in the block, and let $T_{j:k} = e_{k+1} - e_j$ be the length of the block. The edges of blocks demarcate sites where the local density of events changes. Thus, Bayesian blocks fits into the class of change point detection algorithms. While change point detection is commonly performed on a time- or position-dependent signal, Bayesian blocks is distinct in that it operates on the underlying density of data along the interval itself.

Bayesian blocks proceeds using a dynamic programming strategy called Optimal Partitioning (OP) (Jackson et al., 2005) (Figure 4.2B). For a point x_k , the algorithm determines the optimal location of the last change point containing x_k . We denote the Poisson maximum log-likelihood function for block $b_{j:k}$ as $G(b_{j:k})$. (The objective function must be additive across blocks, which is why the likelihood is log-transformed. A complete derivation is provided in the **Proofs** section). We define the global log-likelihood of a partition containing x_k as $F(x_k) = \max F(x_j) + G(b_{j:k})$. Thus, for each data point, Bayesian blocks works backwards, considering each previous data point as possible block members and storing the index j_k that maximizes $F(x_k)$, i.e. $\text{argmax } F(x_k) = j_k$ is the location of the optimal last change point containing x_k . After all points have been considered, the algorithm performs a traceback: the first change point to be emitted corresponds to the index maximizing the log-likelihood function of the last data point (j_c for x_c in Figure 4.2B), the next change point corresponds to the index

maximizing the log-likelihood function for the data point immediately preceding j_c (j_b for x_b), and proceeding so forth until we reach x_1 .

This scheme offers two key benefits. First, by iterating through the interval in this manner, the algorithm performs $O(n^2)$ calculations, which is a significant improvement over the $O(2^n)$ total possible combinations of change points. Second, by maximizing the log-likelihood at every iteration, the algorithm is guaranteed to find the globally optimal log-likelihood partition. The proof for this relies on mathematical induction and is described in (Jackson et al., 2005).

While Bayesian blocks' quadratic runtime is tolerable for many applications, genomes readily span millions to billions of base pairs and our capacity to generate sequencing data grows exponentially. Even within our own lab, calling cards datasets have grown two orders of magnitude in the last five years. To keep pace with technological advances, we have implemented an optimization scheme called Pruned Exact Linear Time (PELT) (Killick et al., 2012) into Bayesian blocks (Figure 4.2C). The intuition here is that we can reduce how far back we search if, at some point, the log-likelihood function monotonically decreases. To take advantage of this, we have to show that there exists a constant K such that

$$G(b_{i:j}) + G(b_{j+1:k}) + K \geq G(b_{i:k}), \text{ for } i, j < k. \text{ (We show that that the objective function}$$

satisfies this criterion in the **Proofs** section). Next, if there exists x_j ($j < k$) such that

$$F(x_k) \geq F(x_j) + G(b_{j:k})$$

is true for all points up to and including x_j , then the last optimal change point for all points x_{k+1} and beyond will not be found beyond x_j . Put differently, the last optimal change point is bounded by x_{j+1} for the remaining data points. As such, we can prune all points up to and including x_j . This optimization achieves linear runtime in practice on real-world genomic datasets (Figure 4.2D).

Finally, to prevent overfitting, Bayesian blocks incorporates a regularization term. This penalty takes the form of a prior on the total number of change points. We use the original formulation of $\text{ncp_prior} = 4 - \log(73.53 * p_0 * N^{-0.478})$, where N is the total number of points in the dataset and p_0 represents a false positive rate in the interval $[0, 1]$. Tuning the value of p_0 alters the sensitivity of block detection. Modest values (e.g. $p_0 = 0.05$) can yield generally good segmentations but may fail to segment subtle differences in densities (Figure 4.3A). Conversely, a more aggressive value may excel at detecting fine structure but may also be more likely to spuriously partition noise (Figure 4.3B). Regardless, we note that both outputs look remarkably similar and partition the data in a manner that accords with our intuition. This example demonstrates the core strengths of Bayesian blocks: a minimal number of hyperparameters; robustness to hyperparameter choice; and block widths that adapt to local variability in signal-to-noise ratio.

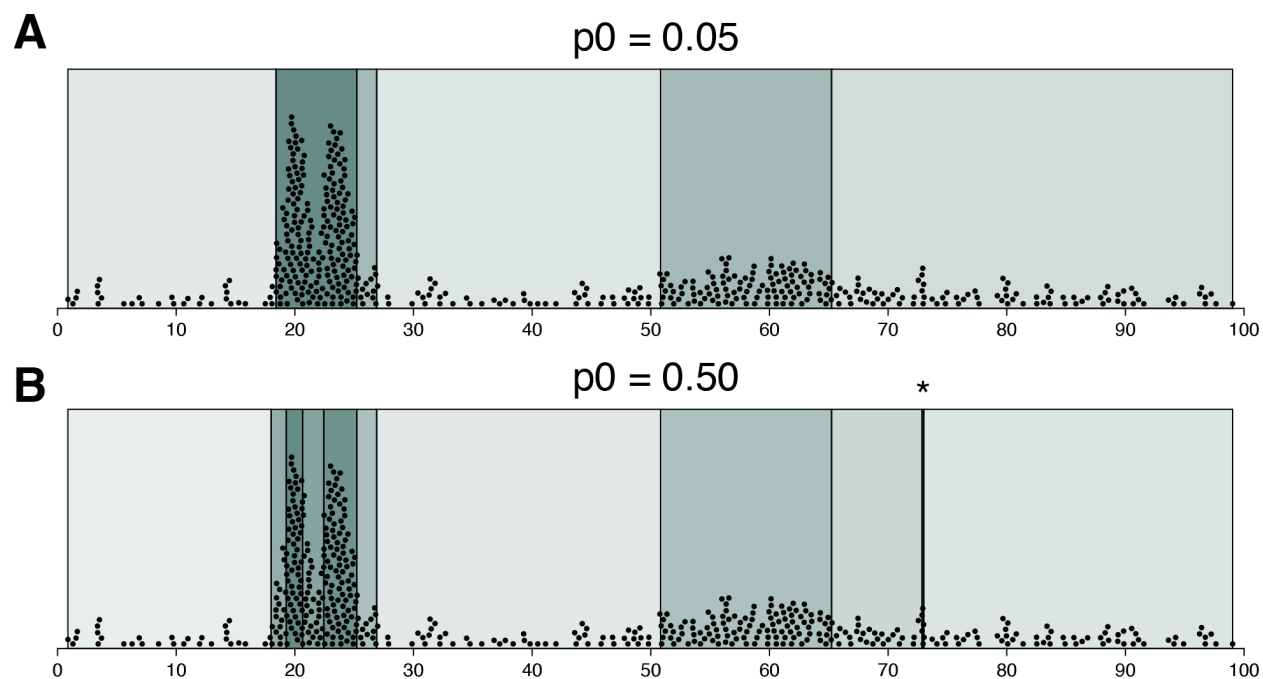


Figure 4.3: Effect of varying p_0 on segmentation. (A) Segmentation resulting from a modest choice of p_0 on multimodal data. (B) Same as (A) but with a more aggressive p_0 . The asterisk marks a potential false positive block.

4.3.2 Calling peaks using Bayesian blocks

We next use Bayesian blocks to call peaks in calling cards data. As mentioned earlier, calling cards are similar to photons in that they are discrete, independent events and can be appropriately described by Poisson-based counting processes. Furthermore, as calling cards data can be sparse in certain contexts, Bayesian blocks' ability to tolerate noise makes it a practical choice as the basis for a peak caller. We have previously used Bayesian blocks to call peaks in mammalian calling cards (Cammack et al., 2020; Moudgil et al., 2020a). Here, we demonstrate Bayesian blocks on publicly available yeast calling cards data and provide a thorough benchmark of the algorithm's performance.

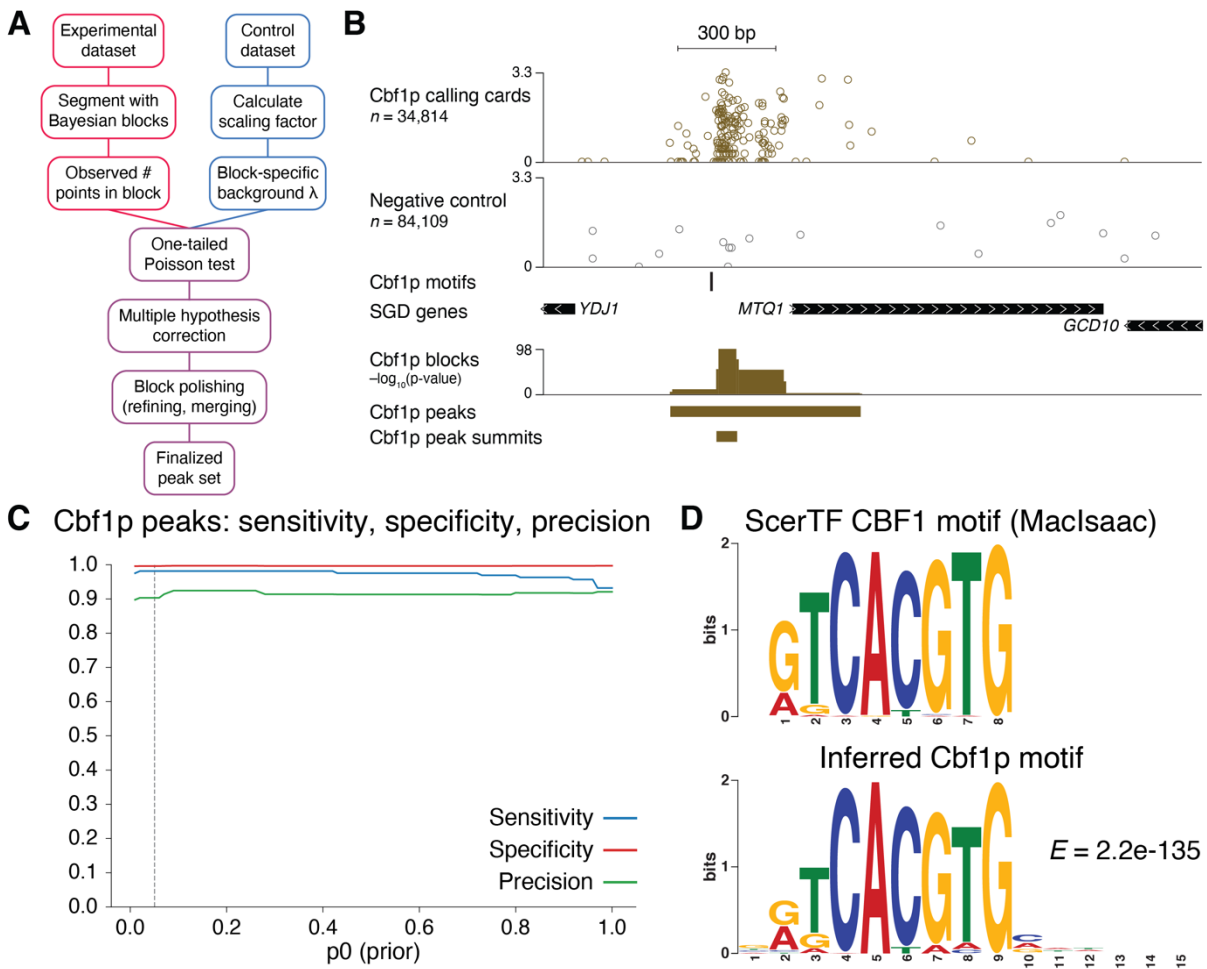


Figure 4.4: Calling peaks on calling cards data with Bayesian blocks. (A) Overview of the peak calling workflow. (B) Example of Cbf1p calling cards peak. Raw insertions in the Cbf1p-directed and negative control

datasets are drawn in the top-most tracks. Output of the Bayesian blocks segmentation is shown toward the bottom, scaled by unadjusted p-value on the y-axis. Peaks were formed by merging significant adjacent blocks. Peak summits denote blocks where the peak achieves maximum significance. (C) Sensitivity, specificity, and precision of Cbf1p peak calls as a function of varying p_0 . Dashed line represents the default value of 0.05. (D) Motif analysis of peaks identified from Cbf1p calling cards elicits the core Cbf1p motif.

Calling cards experiments rely on comparing data from a TF-directed transposase against a control dataset from the undirected transposase (Figure 4.4A). Since the signal we are most interested in are TF-directed densities, we begin by segmenting the TF-directed data using Bayesian blocks. As yeast and mammalian chromosomes are linear, each chromosome is partitioned independently. The resulting blocks are candidate peaks. To account for variable numbers of insertions between the control and experimental datasets, we scale the number of insertions in the control track to equalize library sizes. This scalar is then applied per block (b_i) to the control dataset (subject to a small pseudocount), which then establishes a block-specific mean (λ_i) parameterizing a block-specific Poisson process.

We then consider the data from the TF-directed experiment. For each block b_i , we perform a one-tailed Poisson hypothesis test on observing z_i events or greater in the block, where z_i is the number of insertions within b_i in from the TF-directed experiment, i.e. $P(X \geq z_i | \text{Pois}(\lambda_i))$. In other words, for each block, the null hypothesis is that insertions are distributed according to a Poisson process inferred from the control experiment. We then perform multiple hypothesis correction on these p-values, accepting those beyond a specified threshold. These candidates are then polished, such as by merging significant blocks within a small distance window, before generating a final set of peaks.

We analyzed the data from the basic-helix-loop-helix (bHLH) TF Cbf1p in this manner (Figure 4.4B), finding that our peak calling strategy accurately characterizes Cbf1p binding. We observed that the global likelihood of the segmentation was relatively invariant as a function of

p_0 (Figure 4.5A), which highlights Bayesian blocks' ability to robustly find an optimal segmentation. Moreover, while the number of significant blocks increased slowly as p_0 increased, the number of significant peaks remained constant (Figure 4.5B). Here, we polished peaks by only merging significant adjacent blocks. This suggests that much of the increased partitioning from raising p_0 is likely occurring under Cbf1p peaks, where we would expect to see signal, rather than at unbound loci with sparser insertion densities.

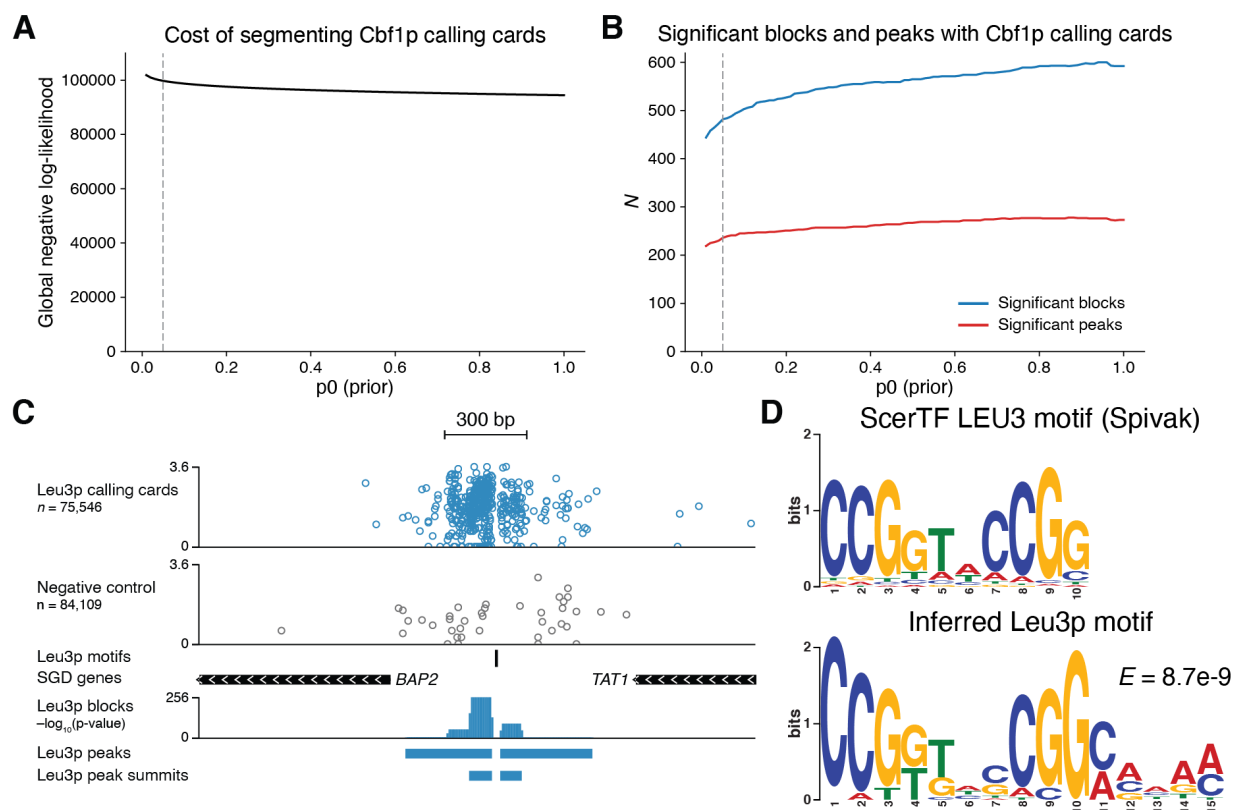


Figure 4.5: Additional benchmarking of peak calling. (A) Global negative log-likelihood from segmenting Cbf1p calling cards as a function of p_0 . (B) Total numbers of significant blocks and significant peaks in Cbf1p calling cards as a function of p_0 . Here, as in (A), dashed line represents the default value of 0.05. (C) Raw insertion data, block representation, peaks, and peak summits from Leu3p calling cards data. (D) Motif analysis of Leu3p peaks elicits the core Leu3p motif.

Finally, we measured the accuracy of our peaks by comparing them to a set of gold-standard set of promoters predicted to be either bound or not bound by Cbf1p (Shively et al., 2019). We called peaks using a default p_0 of 0.05 and calculated their sensitivity, specificity,

and precision as 98.1%, 99.6%, and 90.3%, respectively. Varying p_0 had little effect on sensitivity and precision and only affected specificity at the highest values. A motif analysis of our peaks returned a near-perfect match to the Cbflp motif, confirming the biological validity of our peak calls. To generalize across TF families, we also analyzed data generated from the zinc finger TF Leu3p. As with Cbflp, we identified stark peaks enriched for Leu3p-directed insertions (Figure 4.5C). Our sensitivity, specificity, and precision were 85.8%, 99.7%, and 89%, respectively (Shively et al., 2019), and we were able to elicit the Leu3p motif from our peak calls (Figure 4.5D). We conclude that Bayesian blocks can accurately identify TF binding sites from transposon calling cards data.

One limitation of calling peaks using Bayesian blocks is that it may generate relatively broad peaks (Figure 4.4B, 4.5C). By design, calling cards experiments use a TF-transposase fusion. When bound at its motif, the TF may sterically hinder the transposase, blocking it such that the transposase targets the flanking genomic sequence. This can result in doublets flanking the TF motif (Figure 4.5D). Similarly, high local concentrations of TF-transposase fusion proteins, such as through cooperative binding (Shively et al., 2019), can cause “spillover” of insertions into adjacent DNA (Figure 4.4B). Peaks are built from the underlying structure of blocks. In general, block sizes scale inversely with the relative enrichment of TF-directed insertions over background (Figure 4.6A), which reflects focal redirection to TF binding sites. This effect is also observed in mammalian calling cards on a coarser scale (Figure 4.6B-C), though there are important biological and technical considerations: the human genome is 250 times larger than that of yeast; the transposase targets a specific motif which may be less frequent in certain sequence contexts (Wang et al., 2012); and recovery of insertions can be sparse, particularly in single cell assays (Moudgil et al., 2020a). While we have successfully used

Bayesian blocks to map well-resolved TF binding sites in mammalian systems (Cammack et al., 2020; Moudgil et al., 2020a), there are opportunities for informatic and molecular improvements to further sharpen our analyses.

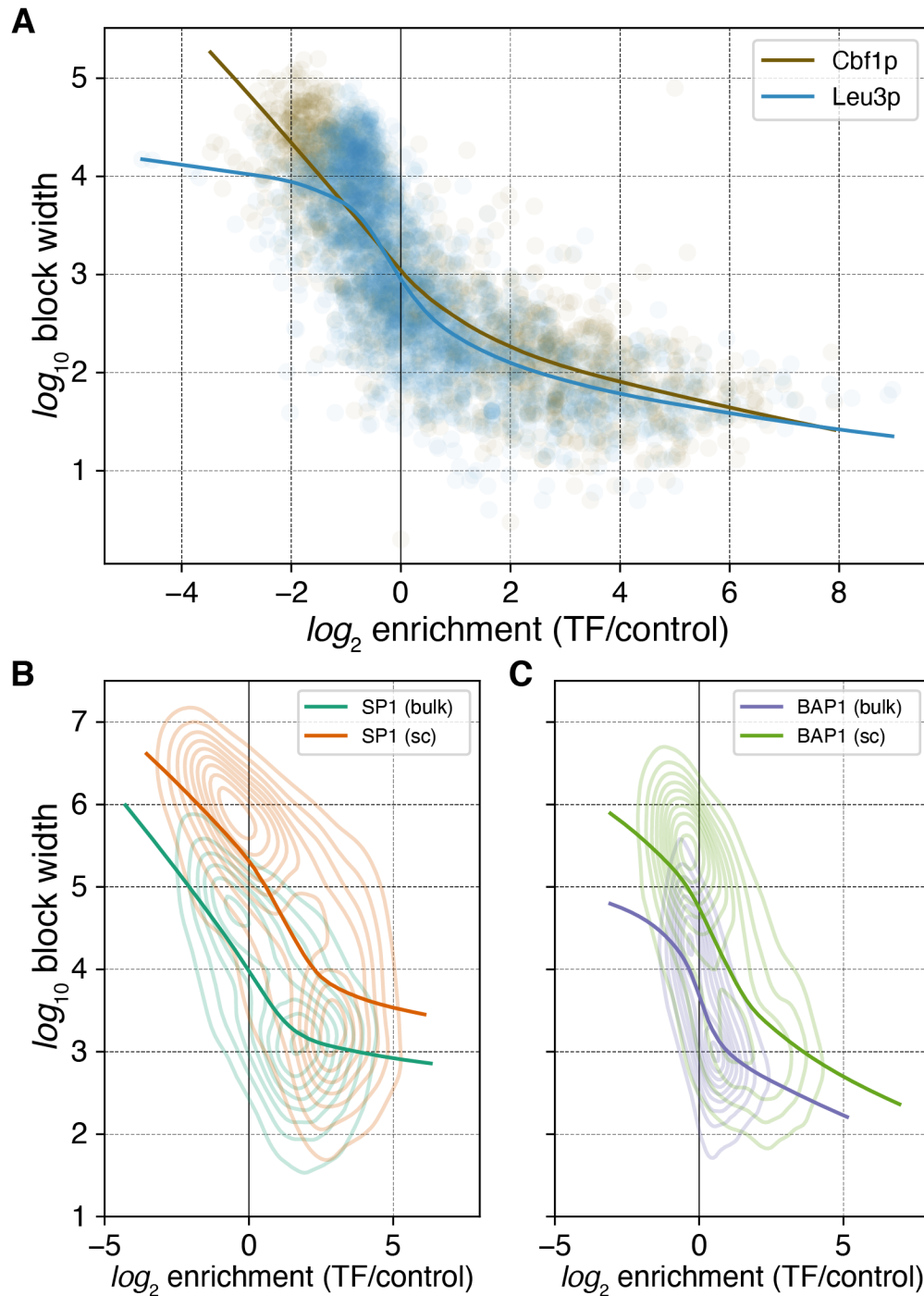


Figure 4.6: TF-directed block sizes are a function of relative enrichment. (A) Block widths in yeast calling cards experiments plotted as a function of \log_2 fold change in normalized insertions between the TF and control datasets.

(B-C) Same as in (A) but for human cell line calling cards experiments with SP1 (B) and BAP1 (C). Data from matched bulk and single cell (sc) experiments are shown. Contours are drawn to emphasize the density of data points. Lines are LOESS-smoothed curves.

4.3.2 Identifying CpG islands using Bayesian blocks

Finally, we consider an application of Bayesian blocks beyond transposon calling cards. CpG

islands represent clusters of CG dinucleotides in the genome. These base pairs decay over evolutionary time due to spontaneous cytidine deamination into thymine. Their occurrence, particularly in clusters, suggests functional selection, and CpG islands are often observed at gene promoters (Tahir et al., 2019). Identifying CpG islands from genomic sequence is a classic segmentation problem in genomics, with methods ranging from sliding window analyses (Gardiner-Garden and Frommer, 1987; Ponger and Mouchiroud, 2002; Takai and Jones, 2002) to more complex Markov models (Byung-Jun Yoon, 2004; Kakumani et al., 2012; Wu et al., 2010). We noted that the task of identifying CpG islands is a density detection problem similar to calling peaks in calling cards data. Therefore, we investigated whether Bayesian blocks could be used to find CpG islands.

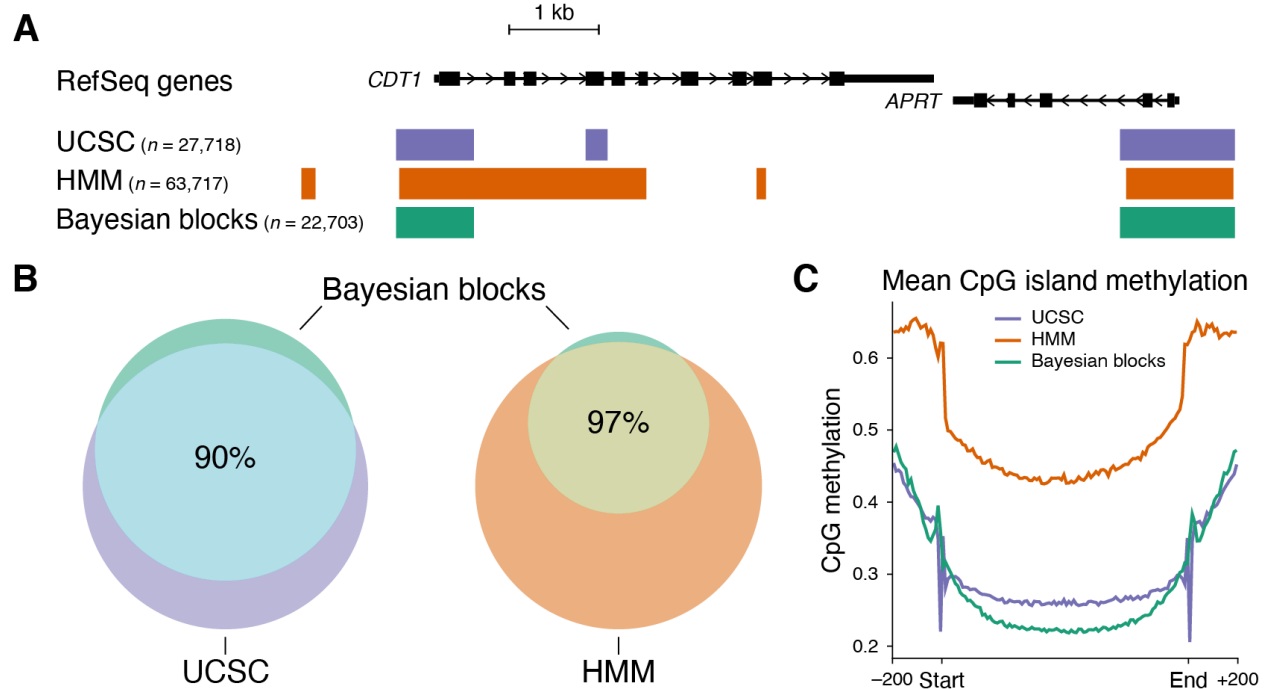


Figure 4.7: Detecting CpG islands with Bayesian blocks. (A) Representative CpG islands as derived from UCSC (Gardiner-Garden and Frommer, 1987), a hidden Markov model (HMM; Wu et al. 2010), and Bayesian blocks. n represents the total number of called islands. (B) Overlap between the Bayesian blocks annotations and the UCSC and HMM datasets, respectively. The number indicates the percent of CpG islands called by Bayesian blocks that overlap the respective comparison set. (C) Mean methylation status of CpG islands in H1 human embryonic stem cells by dataset. kb: kilobase.

We first segmented the genome based on the locations of CpG dinucleotides, a feat that was greatly aided by the PELT optimization. A block was considered a CpG island if it met the criteria laid out by (Gardiner-Garden and Frommer, 1987), whose own annotations form the default set of CpG islands in the UCSC Genome Browser and thus have considerable visibility to researchers. Since many repetitive elements are also rich in CG dinucleotides (Tahir et al., 2019; Wu et al., 2010), we performed our analysis on a repeat-masked genome. Our Bayesian blocks-based CpG islands showed strong overlap with the UCSC set (Figure 4.7A), with over 90% of our calls overlapping a UCSC CpG island (Figure 4.7B). This suggests that Bayesian blocks can be used sensitively and specifically detect CpG islands.

We were curious as to how Bayesian blocks compares to hidden Markov models (HMMs), which is a well-established method for genome segmentation (Chou and Danko, 2019; Durbin et al., 1998; Ernst et al., 2011; Ha et al., 2012; Keough et al., 2020; Malekpour et al., 2017; Munch and Krogh, 2006). We cross-referenced our CpG islands against a published set of HMM-based CpG islands (Wu et al., 2010) (Figure 4.7A-B). Once again, we observed high specificity of overlap, with over 97% of our CpG islands overlapping an HMM-based CpG island. However, there was a large fraction of HMM CpG islands that did not overlap the Bayesian blocks set, which may indicate that the latter is susceptible to false negatives. To investigate further, we looked at the average methylation status of these called CpG islands (Stevens et al., 2013). While CpGs are generally methylated, CpG islands show decreased methylation relative to surrounding sequence (Greenberg and Bourc'his, 2019). All three

datasets followed this pattern, though the UCSC and Bayesian blocks sets had markedly less methylation on average than the HMM set (Figure 4.7C). We conclude that the Bayesian blocks approach to calling CpG islands is not likely prone to false negatives.

To close, we looked at the size distribution of these called CpG islands. HMMs use transition probabilities to mark changes in state. As such, the probability of staying in the same state decreases exponentially and so HMMs should be more likely to generate short segmentations. We found that the HMM-based CpG islands had a greater enrichment for smaller segments over Bayesian blocks (Figure 4.8). CpG islands from Bayesian blocks also appear to avoid the short segment bias seen in the UCSC dataset. The tendency toward longer features is one key difference between HMMs and Bayesian blocks. For some applications, this may be a desirable feature.

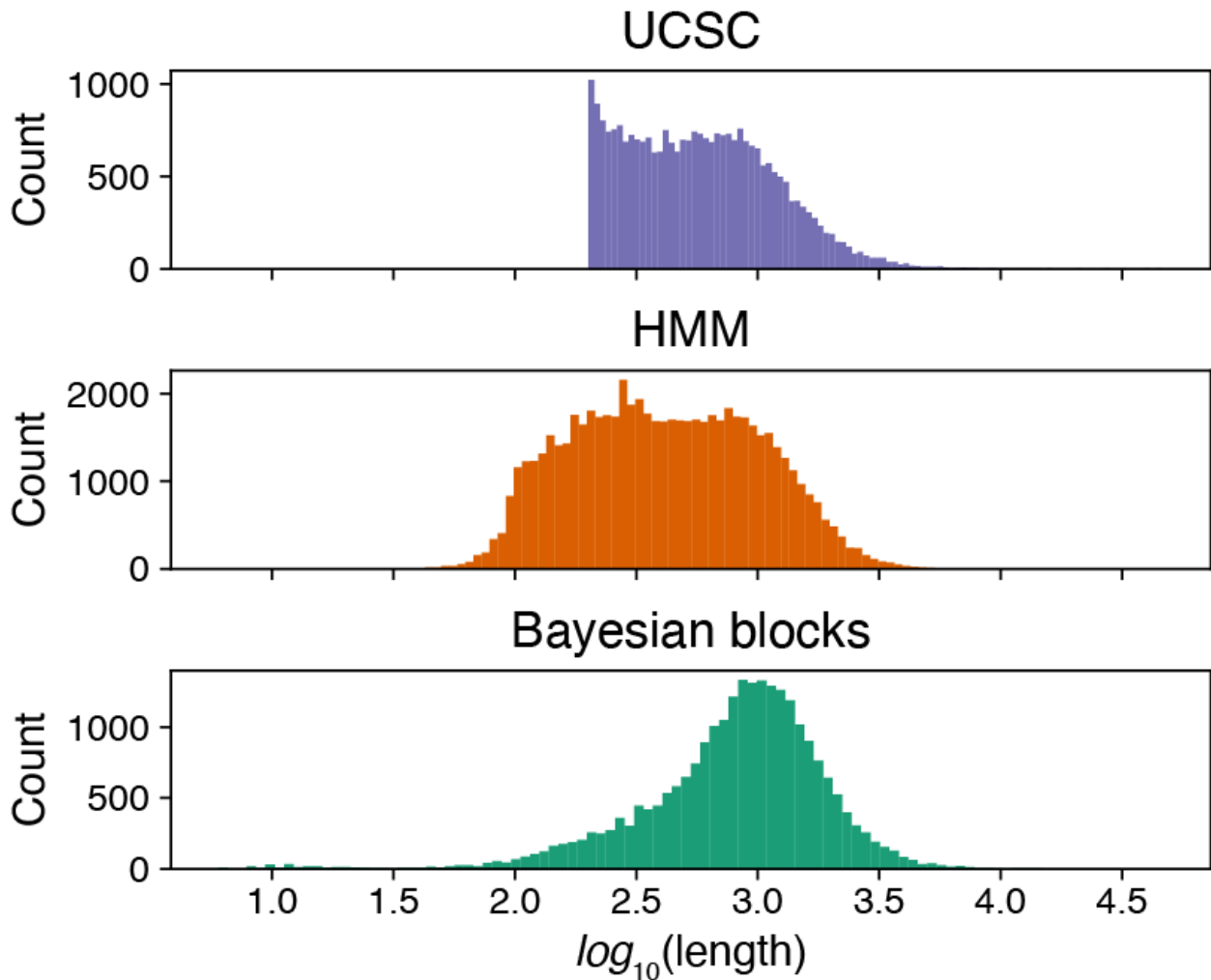


Figure 4.8: Length distributions of CpG islands. For this analysis, block lengths were log-transformed. In addition, we removed the 200 bp minimum length requirement for the Bayesian blocks dataset to better visualize the entire distribution.

4.4 Discussion

Segmentation is a common task in biology. PELT, for example, has been used to segment two-dimensional signals, such as coverage across the genome (Burke et al., 2018; Weinberg et al., 2019), neuronal firing over time (Gouwens et al., 2019; Jewell et al., 2019), and fluorescence intensity over spatial trajectories (Desai et al., 2019). Its adoption was facilitated by the R package `changept` (Killick and Eckley, 2014). However, there was no analogous software for segmenting one-dimensional data, such as the location of exogenous transposons or CG dinucleotides.

We fill that gap by introducing Bayesian blocks as a density-based approach to partitioning genomic features. In addition to this manuscript, we have released a Python package, blockify (<https://github.com/arnavm/blockify>), for analyzing genomic datasets with Bayesian blocks. Blockify directly operates on BED files, a popular standard for encoding genomic features; incorporates the PELT-based optimization for fast segmentation; comes with a peak caller to facilitate discriminative analyses; and is globally available via the Python Package Index. We have also created an online notebook illustrating Bayesian blocks, OP, and PELT within an interactive framework (<https://observablehq.com/d/d2cafaa7d8c1e018>). In particular, this allows users to see how changing the underlying distribution of data and hyperparameters affects segmentation.

We demonstrated how Bayesian blocks can be used to call peaks in transposon calling cards, comparing information from a TF-directed experiment against a control dataset to identify binding sites. That Bayesian blocks can find peaks with high accuracy and precision, tolerates noise, and is robust to hyperparameters makes it an attractive algorithm for this purpose. However, peak widths can be broad, for a variety of reasons outlined above. One strategy that could enrich for smaller peaks is to set a relatively lax size filter and accept all blocks with a minimum enrichment threshold. For example, one might draw quadrants on a plot like Figure 4.6B and take all blocks in the lower right corner. This would reduce the number of multiple comparisons when assessing statistical significance, which in turn may identify loci with weaker TF-directed signal. Adjusting the prior on the number of change points may also lead to sharper peaks. While we adopted a global prior per chromosome, a more sophisticated approach could be to dynamically vary the prior based on local genomic features. Information about chromatin state, derived from ChIP-seq and ATAC-seq, could be incorporated into a context-specific prior

that favors more aggressive segmentation in promoters and enhancers—where TF binding sites are more likely to reside—and decreases sensitivity in heterochromatin. This would balance the desire to pinpoint binding sites while minimizing the risk of false positives.

Our investigation of CpG islands suggests that Bayesian blocks can be competitive with HMMs, which is a standard technique for genome segmentation. However, we do not see Bayesian blocks replacing HMMs. HMMs can support multiple states, making them arbitrarily customizable, and they can encode additional layers of information, such as sequence composition. Bayesian blocks is responsive to data density and is better suited for relatively simple classifications, such as peak and non-peak regions. Additionally, the speed of the algorithm makes it ideal for exploratory analyses that can inform more complex models.

There are a number of future directions to develop the core algorithm. While here we use a piecewise-constant density function as a first-order approximation for TF binding, this simplification may not necessarily reflect biological reality. Bayesian blocks, at least with the OP algorithm, can also accommodate piecewise linear and exponential density functions. In our current implementation, blockify assumes linear chromosomes. However, Bayesian blocks can also operate on circular intervals, and this could be applied to analyzing metagenomic and mitochondrial genomes. Bayesian blocks can also be extended to two- or even multi-dimensional datasets (Scargle, 2002), which could prove useful in clustering single cell RNA-seq data. Expanding blockify to use non-Poisson likelihoods could also be useful, as long as block-additivity is preserved. Many RNA-seq datasets, for example, use the negative binomial distribution to model count distributions. A negative binomial likelihood function could be apt for discovering genomic “dark matter” like cryptic open reading frames (Wang et al., 2021). Finally, as ATAC-seq is another transposase-based assay, Bayesian blocks may be able to infer

the optimal width of accessible loci, increasing resolution and possibly better revealing dynamic changes.

4.5 Methods

All segmentation and peak calling were performed with blockify 0.1.0, which is available at <https://github.com/arnavm/blockify>. The Bayesian blocks algorithm was originally forked from astropy 3.2.1 (Robitaille et al., 2013; The Astropy Collaboration et al., 2018) and expanded with custom code. We have also implanted Bayesian blocks in JavaScript at <https://observablehq.com/d/d2cafaa7d8c1e018>. The OP and PELT timing analyses were performed on a MacBook Pro with a 2.9 GHz Quad-Core Intel Core i7 processor with 16 GB of RAM. General data analysis was performed with numpy 1.17.2 (Oliphant, 2015), matplotlib 3.0.3 (Hunter, 2007), statsmodels 0.11.1 (Seabold and Perktold, 2010), and Python 3.6.

Yeast Cbf1p, Leu3p, and control calling cards datasets were obtained from (Shively et al., 2019). Bulk mouse neuron and astrocyte *piggyBac* calling cards datasets were obtained from (Cammack et al., 2020). We also downloaded all bulk and *in vitro* calling cards datasets from (Moudgil et al., 2020a), as well as *in vivo* astrocyte and neuron data. Unless otherwise specified, segmentation was performed with $p_0 = 0.05$. For the yeast dataset, peaks were called by merging adjacent significant blocks (after Bonferroni correction with an adjusted p-value cutoff of 0.05) with a pseudocount of 1: `-d 0 -a 0.05 --correction "bonferroni" -c 1`. For the sensitivity, specificity, and precision analysis of Cbf1p and Leu3p peaks, we used data from (Shively et al., 2019). Specifically, we used p-value cutoffs of $< 1e-5$ to call true positive promoters and $p > 0.1$ to call true negative promoters. Sensitivity was defined as $TP / (TP + FN)$, specificity was $TN / (TN + FP)$, and precision as $TP / (TP + FP)$. Motif locations were found by scanning the *sacCer3* genome with PWMScan (Ambrosini et al., 2018) using the recommended Cbf1p and Leu3p

motifs and PWM scores from ScerTF (Spivak and Stormo, 2012). Motif analysis was performed using meme-chip (Machanick and Bailey, 2011) with the following settings: “-dna -nmeme 600 -seed 0 -ccut 250 -meme-mod zoops -meme-minw 4 -meme-nmotifs 5.” Calling card datasets were visualized as qBED tracks on the WashU Epigenome Browser (Li et al., 2019; Moudgil et al., 2021).

For the analysis of CpG islands, we first generated a list of all CG dinucleotides in a repeat-masked hg19 assembly using kmer.cc (<https://gist.github.com/arnavm/039e76a34a386a4f29b82682bc8e6c72>). We then segmented this using blockify ($p_0 = 0.05$). Finally, a block was considered a CpG island if it met the following criteria: length greater than 200 bp; minimum GC content of 50%; and a ratio of observed to expected CG dinucleotides of at least 0.6 (Gardiner-Garden and Frommer, 1987). When comparing length distributions (Figure 4.8), we dropped the first restriction from the Bayesian blocks set. A reference set of CpG islands in hg19 was downloaded from the UCSC Genome Browser. An orthogonal set of HMM-derived CpG islands in hg19 was obtained from (Wu et al., 2010). CpG methylation data in H1 human embryonic stem cells were derived from (Stevens et al., 2013). Mean methylation levels were calculated using deeptools 3.0.1 (Ramírez et al., 2016).

4.6 Proofs

4.6.1 Properties of Poisson point processes

Let $P(k | \lambda)$ be a Poisson random variable with value k parameterized by an expected value of λ . Thus,

$$P(k | \lambda) = \frac{e^{-\lambda} \lambda^k}{k!}$$

From this, we see that $P(0) = e^{-\lambda}$ and $P(1) = \lambda e^{-\lambda}$. If $N(t)$ is a Poisson process defined on the interval t and with rate parameter λ , the probability of k events within the interval is

$$P(N(t) = k) = \frac{e^{-\lambda t} (\lambda t)^k}{k!}$$

We can use this equation to calculate the instantaneous probability of an event over an infinitesimal interval dt :

$$P(N(dt) = 1) = e^{-\lambda dt} \lambda dt$$

As $dt \rightarrow 0$, the exponential term goes to 1, leaving $P(N(dt) = 1) = \lambda dt$. This can also be seen from the Taylor series expansion of $e^x = \sum_{i=0}^{\infty} x^i / i!$

$$= e^{-\lambda dt} \lambda dt$$

$$= \left[\frac{(-\lambda dt)^0}{0!} + \frac{(-\lambda dt)^1}{1!} + \frac{(-\lambda dt)^2}{2!} + \dots \right] \lambda dt$$

$$= \left[1 - \lambda dt + \frac{(-\lambda dt)^2}{2!} + \dots \right] \lambda dt$$

$$= \lambda dt - (\lambda dt)^2 + \frac{(-\lambda dt)^3}{2!} + \dots$$

As $dt \rightarrow 0$, the nonleading terms rapidly go to zero, resulting in $P(N(dt) = 1) = \lambda dt$.

4.6.2 Derivation of the Bayesian blocks likelihood function

Bayesian blocks finds the segmentation that maximizes the global Poisson likelihood.

The likelihood of a Poisson process parameterized by a mean of λ is given by

$$L(\text{Pois}(\lambda)) = P(x_1 \dots x_n | \text{Pois}(\lambda))$$

Let t_i denote the intervals between data points within a block, where $i \in \{0 \dots n\}$. Then, the likelihood is the product of the instantaneous probabilities at x_i and zero events in the intervals t_i :

$$L(Pois(\lambda)) = e^{-\lambda t_0} dx_1 e^{-\lambda t_1} dx_2 \dots e^{-\lambda t_n}$$

$$L(Pois(\lambda)) = e^{-\lambda \sum_0^n t_i} \lambda^n \prod_{i=1}^n dx_i$$

We can generalize simplify this by noting that the sum of the intervals t_i is the length of the block and the rest of the terms depend only on the number of data points within the block (N) . Thus,

$$L = e^{-\lambda T} \lambda^N (dx)^N$$

Bayesian blocks requires that the objective function be additive, which we achieve by taking the logarithm of the likelihood.

$$\log L = -\lambda T + N \log \lambda + N \log dx$$

We now need to find the value of λ that maximizes the likelihood function. We obtain this value by taking the derivative of the likelihood function and solving for zero.

$$\frac{dL}{d\lambda} = [-T e^{-\lambda T} \lambda^N + N e^{-\lambda T} \lambda^{N-1}] (dx)^N$$

$$\frac{dL}{d\lambda} = [e^{-\lambda T} \lambda^{N-1} (-T\lambda + N)] (dx)^N$$

The only nontrivial root of this equation is at $\hat{\lambda} = N/T$. Substituting this into the log-likelihood equation:

$$\log L = -T \frac{N}{T} + N \log \frac{N}{T}$$

$$\log L = -N + N(\log N - \log T)$$

$$\log L + N = N(\log N - \log T)$$

This matches Equation 19 from (Scargle et al., 2013). To simplify calculations, we note that $\log L \propto N(\log N - \log T)$ and so use the right-hand side of this relationship as the Bayesian blocks likelihood function:

$$G(b_{i:j}) = N_{i:j} (\log N_{i:j} - \log T_{i:j})$$

4.6.3 Adapting Bayesian blocks to PELT

PELT enables us to dynamically prune data points, improving the runtime from quadratic to

linear. To take advantage of this, we have to show that there exists a constant K such

that $G(b_{i:j}) + G(b_{j+1:k}) + K \geq G(b_{i:k})$, where $i < j < k$. The proof of this is as follows:

$$G(b_{i:k}) = N_{i:k} (\log N_{i:k} - \log T_{i:k})$$

$$G(b_{i:k}) = N_{i:k} \log N_{i:k} - N_{i:k} \log T_{i:k}$$

$$G(b_{i:k}) = (N_{i:j} + N_{j+1:k}) \log(N_{i:j} + N_{j+1:k}) - (N_{i:j} + N_{j+1:k}) \log(T_{i:j} + T_{j+1:k})$$

$$G(b_{i:k}) < (N_{i:j} + N_{j+1:k})(\log N_{i:j} + \log N_{j+1:k}) - (N_{i:j} + N_{j+1:k}) \log(T_{i:j} + T_{j+1:k})$$

$$\begin{aligned} G(b_{i:k}) < & (N_{i:j} \log N_{i:j} + N_{j+1:k} \log N_{j+1:k}) \\ & + (N_{i:j} \log N_{j+1:k} + N_{j+1:k} \log N_{i:j}) - N_{i:j} \log(T_{i:j} + T_{j+1:k}) \\ & - N_{j+1:k} \log(T_{i:j} + T_{j+1:k}) \end{aligned}$$

$$\begin{aligned} G(b_{i:k}) < & (N_{i:j} \log N_{i:j} + N_{j+1:k} \log N_{j+1:k}) \\ & + (N_{i:j} \log N_{j+1:k} + N_{j+1:k} \log N_{i:j}) - N_{i:j} \log T_{i:j} - N_{j+1:k} \log T_{j+1:k} \end{aligned}$$

$$\begin{aligned} G(b_{i:k}) < & (N_{i:j} \log N_{i:j} + N_{j+1:k} \log N_{j+1:k} - N_{i:j} \log T_{i:j} - N_{j+1:k} \log T_{j+1:k}) \\ & + (N_{i:j} \log N_{j+1:k} + N_{j+1:k} \log N_{i:j}) \end{aligned}$$

$$G(b_{i:k}) < N_{i:j} \log N_{i:j} + N_{j+1:k} \log N_{j+1:k} - N_{i:j} \log T_{i:j} - N_{j+1:k} \log T_{j+1:k}$$

$$G(b_{i:k}) < N_{i:j} \log N_{i:j} - N_{i:j} \log T_{i:j} + N_{j+1:k} \log N_{j+1:k} - N_{j+1:k} \log T_{j+1:k}$$

$$G(b_{i:k}) < N_{i:j}(\log N_{i:j} - \log T_{i:j}) + N_{j+1:k}(\log N_{j+1:k} - \log T_{j+1:k})$$

$$G(b_{i:k}) < G(b_{i:j}) + G(b_{j+1:k})$$

Thus, K exists and equals 0. Note that this inequality holds as long as $N_{i:j}, N_{j+1:k} \geq 2$ and so must be taken into consideration when implemented.

4.7 References

- Ambrosini, G., Groux, R., and Bucher, P. (2018). PWMScan: a fast tool for scanning entire genomes with a position-specific weight matrix. *Bioinformatics* 34, 2483–2484.
- Burke, J.E., Longhurst, A.D., Merkurjev, D., Sales-Lee, J., Rao, B., Moresco, J.J., Yates, J.R., Li, J.J., and Madhani, H.D. (2018). Spliceosome Profiling Visualizes Operations of a Dynamic RNP at Nucleotide Resolution. *Cell* 173, 1014-1030.e17.
- Byung-Jun Yoon, P.P.V. (2004). Identification of CPG islands using a bank of IIR lowpass filters. In 3rd IEEE Signal Processing Education Workshop. 2004 IEEE 11th Digital Signal Processing Workshop, 2004., (Taos Ski Valley, NM, USA: IEEE), pp. 315–319.
- Cammack, A.J., Moudgil, A., Chen, J., Vasek, M.J., Shabsovich, M., McCullough, K., Yen, A., Lagunas, T., Maloney, S.E., He, J., et al. (2020). A viral toolkit for recording transcription factor–DNA interactions in live mouse tissues. *Proc Natl Acad Sci USA* 117, 10003–10014.
- Cang, Z., and Nie, Q. (2020). Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nat Commun* 11, 2084.

- Cha, R.S., and Thilly, W.G. (1993). Specificity, efficiency, and fidelity of PCR. *Genome Research* 3, S18–S29.
- Chan, T.E., Stumpf, M.P.H., and Babbie, A.C. (2017). Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. *Cell Systems* 5, 251-267.e3.
- Chou, S.-P., and Danko, C.G. (2019). AlleleHMM: a data-driven method to identify allele specific differences in distributed functional genomic marks. *Nucleic Acids Research* 47, e64–e64.
- Cleynen, A., Koskas, M., Lebarbier, E., Rigaiil, G., and Robin, S. (2014). Segmentor3IsBack: an R package for the fast and exact segmentation of Seq-data. *Algorithms Mol Biol* 9, 6.
- Desai, V.P., Frank, F., Lee, A., Righini, M., Lancaster, L., Noller, H.F., Tinoco, I., and Bustamante, C. (2019). Co-temporal Force and Fluorescence Measurements Reveal a Ribosomal Gear Shift Mechanism of Translation Regulation by Structured mRNAs. *Molecular Cell* 75, 1007-1019.e5.
- Durbin, R., Eddy, Sean, Krogh, Anders, and Mitchison, Graeme (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids* (Cambridge, UK : New York: Cambridge University Press).
- Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shores, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473, 43–49.

- Fan, Z., and Mackey, L. (2017). Empirical Bayesian analysis of simultaneous changepoints in multiple data sequences. *Ann. Appl. Stat.* *11*, 2200–2221.
- Gardiner-Garden, M., and Frommer, M. (1987). CpG Islands in vertebrate genomes. *Journal of Molecular Biology* *196*, 261–282.
- Gouwens, N.W., Sorensen, S.A., Berg, J., Lee, C., Jarsky, T., Ting, J., Sunkin, S.M., Feng, D., Anastassiou, C.A., Barkan, E., et al. (2019). Classification of electrophysiological and morphological neuron types in the mouse visual cortex. *Nat Neurosci* *22*, 1182–1195.
- Greenberg, M.V.C., and Bourc’his, D. (2019). The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol* *20*, 590–607.
- Ha, G., Roth, A., Lai, D., Bashashati, A., Ding, J., Goya, R., Giuliany, R., Rosner, J., Oloumi, A., Shumansky, K., et al. (2012). Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Research* *22*, 1995–2007.
- Hansen, R.S., Thomas, S., Sandstrom, R., Canfield, T.K., Thurman, R.E., Weaver, M., Dorschner, M.O., Gartler, S.M., and Stamatoyannopoulos, J.A. (2010). Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences* *107*, 139–144.
- Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* *9*, 90–95.
- Ish-Horowicz, J., and Reid, J. (2017). Mutual information estimation for transcriptional regulatory network inference (bioRxiv).

Jackson, B., Scargle, J.D., Barnes, D., Arabhi, S., Alt, A., Gioumouisis, P., Gwin, E., San, P., Tan, L., and Tsai, T.T. (2005). An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters* *12*, 105–108.

Jewell, S.W., Hocking, T.D., Fearnhead, P., and Witten, D.M. (2019). Fast nonconvex deconvolution of calcium imaging data. *Biostatistics*.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science* *316*, 1497–1502.

Kakumani, R., Ahmad, O., and Devabhaktuni, V. (2012). Identification of CpG islands in DNA sequences using statistically optimal null filters. *J Bioinform Sys Biology* *2012*, 12.

Kanagawa, T. (2003). Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of Bioscience and Bioengineering* *96*, 317–323.

Keough, K.C., Shah, P.P., Wickramasinghe, N.M., Dundes, C.E., Chen, A., Salomon, R.E.A., Whalen, S., Loh, K.M., Dubois, N., Pollard, K.S., et al. (2020). An atlas of lamina-associated chromatin across thirteen human cell types reveals cell-type-specific and multiple subtypes of peripheral heterochromatin (bioRxiv).

Killick, R., and Eckley, I.A. (2014). changepoint: An R Package for Changepoint Analysis. *J. Stat. Soft.* *58*.

Killick, R., Fearnhead, P., and Eckley, I.A. (2012). Optimal Detection of Changepoints With a Linear Computational Cost. *Journal of the American Statistical Association* *107*, 1590–1598.

Li, D., Hsu, S., Purushotham, D., Sears, R.L., and Wang, T. (2019). WashU Epigenome Browser update 2019. *Nucleic Acids Research* 47, W158–W165.

Machanick, P., and Bailey, T.L. (2011). MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* 27, 1696–1697.

Malekpour, S.A., Pezeshk, H., and Sadeghi, M. (2017). PSE-HMM: genome-wide CNV detection from NGS data using an HMM with Position-Specific Emission probabilities. *BMC Bioinformatics* 18, 30.

Moudgil, A., Wilkinson, M.N., Chen, X., He, J., Cammack, A.J., Vasek, M.J., Lagunas, T., Qi, Z., Lalli, M.A., Guo, C., et al. (2020). Self-Reporting Transposons Enable Simultaneous Readout of Gene Expression and Transcription Factor Binding in Single Cells. *Cell* 182, 992-1008.e21.

Moudgil, A., Li, D., Hsu, S., Purushotham, D., Wang, T., and Mitra, R.D. (2021). The qBED track: a novel genome browser visualization for point processes. *Bioinformatics* 37, 1168–1170.

Munch, K., and Krogh, A. (2006). Automatic generation of gene finders for eukaryotic species. *BMC Bioinformatics* 7, 263.

Oliphant, T.E. (2015). *Guide to NumPy* (Austin, Tex.: Continuum Press).

Ponger, L., and Mouchiroud, D. (2002). CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics* 18, 631–633.

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* 44, W160–W165.

Robitaille, T.P., Tollerud, E.J., Greenfield, P., Droettboom, M., Bray, E., Aldcroft, T., Davis, M., Ginsburg, A., Price-Whelan, A.M., Kerzendorf, W.E., et al. (2013). Astropy: A community Python package for astronomy. *Astronomy & Astrophysics* 558, A33.

Scargle, J.D. (1998). Studies in Astronomical Time Series Analysis. V. Bayesian Blocks, a New Method to Analyze Structure in Photon Counting Data. *The Astrophysical Journal* 504, 405–418.

Scargle, J.D. (2002). Bayesian blocks in two or more dimensions: Image segmentation and cluster analysis. In *AIP Conference Proceedings*, (Baltimore, Maryland (USA): AIP), pp. 163–173.

Scargle, J.D., Norris, J.P., Jackson, B., and Chiang, J. (2013). STUDIES IN ASTRONOMICAL TIME SERIES ANALYSIS. VI. BAYESIAN BLOCK REPRESENTATIONS. *The Astrophysical Journal* 764, 167.

Seabold, S., and Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, p.

Shively, C.A., Liu, J., Chen, X., Loell, K., and Mitra, R.D. (2019). Homotypic cooperativity and collective binding are determinants of bHLH specificity and function. *Proc Natl Acad Sci USA* 116, 16143–16152.

Spivak, A.T., and Stormo, G.D. (2012). ScerTF: a comprehensive database of benchmarked position weight matrices for *Saccharomyces* species. *Nucleic Acids Research* 40, D162–D168.

Stevens, M., Cheng, J.B., Li, D., Xie, M., Hong, C., Maire, C.L., Ligon, K.L., Hirst, M., Marra, M.A., Costello, J.F., et al. (2013). Estimating absolute methylation levels at single-CpG

resolution from methylation enrichment and restriction enzyme sequencing methods. *Genome Research* 23, 1541–1553.

Stumpf, P.S., Smith, R.C.G., Lenz, M., Schuppert, A., Müller, F.-J., Babbie, A., Chan, T.E., Stumpf, M.P.H., Please, C.P., Howison, S.D., et al. (2017). Stem Cell Differentiation as a Non-Markov Stochastic Process. *Cell Systems* 5, 268-282.e7.

Tahir, R.A., Zheng, D., Nazir, A., and Qing, H. (2019). A review of computational algorithms for CpG islands detection. *J Biosci* 44, 143.

Takai, D., and Jones, P.A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *PNAS* 99, 3740–3745.

The Astropy Collaboration, Price-Whelan, A.M., Sipőcz, B.M., Günther, H.M., Lim, P.L., Crawford, S.M., Conseil, S., Shupe, D.L., Craig, M.W., Dencheva, N., et al. (2018). The Astropy Project: Building an inclusive, open-science project and status of the v2.0 core package. *AJ* 156, 123.

Wang, H., Johnston, M., and Mitra, R.D. (2007). Calling cards for DNA-binding proteins. *Genome Research* 17, 1202–1209.

Wang, H., Mayhew, D., Chen, X., Johnston, M., and Mitra, R.D. (2012). “Calling Cards” for DNA-Binding Proteins in Mammalian Cells. *Genetics* 190, 941–949.

Wang, M.F.Z., Mantri, M., Chou, S.-P., Scuderi, G.J., McKellar, D.W., Butcher, J.T., Danko, C.G., and De Vlaminck, I. (2021). Uncovering transcriptional dark matter via gene annotation independent single-cell RNA sequencing analysis. *Nat Commun* 12, 2158.

Wang, Z., Chu, T., Choate, L.A., and Danko, C.G. (2019). Identification of regulatory elements from nascent transcription using dREG. *Genome Res.* 29, 293–303.

Weinberg, D.N., Papillon-Cavanagh, S., Chen, H., Yue, Y., Chen, X., Rajagopalan, K.N., Horth, C., McGuire, J.T., Xu, X., Nikbakht, H., et al. (2019). The histone mark H3K36me2 recruits DNMT3A and shapes the intergenic DNA methylation landscape. *Nature* 573, 281–286.

Wu, H., Caffo, B., Jaffee, H.A., Irizarry, R.A., and Feinberg, A.P. (2010). Redefining CpG islands using hidden Markov models. *Biostatistics* 11, 499–514.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* 9, R137.

Chapter 5: Self-reporting transposons reveal chromosomal compartmentalization

5.1 Introduction

The three-dimensional conformation of chromatin can influence gene expression (Fullwood et al., 2009; Le Dily et al., 2014). To better understand this process, a number of sequencing-based techniques have been developed measuring DNA-DNA contacts and reconstructing spatial relationships within the genome (Dekker, 2002; Dostie et al., 2006; Simonis et al., 2006; Zhao et al., 2006). Perhaps most well-known is Hi-C (Lieberman-Aiden et al., 2009), a high-throughput method to sequence proximal contacts in an unbiased manner. In addition to inspiring several derivative techniques (Hsieh et al., 2015; Krietenstein et al., 2020; Quinodoz et al., 2018), Hi-C has revealed higher-order chromosomal structures such as compartments (Lieberman-Aiden et al., 2009) and topologically-associated domains (Dixon et al., 2012).

Hi-C, like all bulk assays, creates an averaged representation of genome conformation and is therefore most appropriate when studying a pure population of cells. Newer techniques have emerged that barcode and capture genomic interactions in individual cells (Arrastia et al., 2022; Nagano et al., 2013; Ramani et al., 2017). While these methods are revealing organizational heterogeneity within nuclei, they only recover conformation information and are still restricted to predefined cell types. Single cell RNA sequencing (scRNA-seq) can be used to discover cell types (Cao et al., 2017; Klein et al., 2015; Macosko et al., 2015; Plasschaert et al., 2018; Zheng et al., 2017) and has fueled multi-omic technologies that link gene expression to other genomic processes (Angermueller et al., 2016; Cao et al., 2018; Rooijers et al., 2019; Stoeckius et al., 2017). The ability to directly connect transcriptome to genome conformation

would greatly advance our understanding of regulatory mechanics in individual cell types, especially if this could be deployed *in vivo*. Unfortunately, no such method currently exists.

Here we offer one approach to bridging this gap. We have previously developed single cell calling cards to map transcription factor binding sites in a cell-type-specific manner (Moudgil et al., 2020). This technique relies on self-reporting transposons (SRTs), exogenous transposable elements that transcribe their genomic coordinates as mRNA. Thus, transpositions can be mapped from scRNA-seq libraries, which also contain cellular transcripts to identify cell type. We have shown that this method can be used both *in vitro*, with plasmid transfection, as well as *in vivo* via viral transduction.

In our prior work, we focused primarily on the *piggyBac* transposase. Now, we analyze the distribution of *Sleeping Beauty* (SB) SRTs in HCT-116 cells. We find that SB transposition is non-uniform across the genome, but this only becomes apparent at the megabase scale. Variation in SB activity follows the same pattern as chromosomal compartmentalization, with greater transposition seen in the transcriptionally active compartment A. This correlation is strong enough that we can call compartments from SB transposition itself. We provide several benchmarks to confirm the precision of our assignments. Finally, we quantify how accurately we can identify compartments in the limit of sparse data. These analyses lay the foundation for a potentially powerful, albeit unconventional, method for jointly measuring gene expression and genome topology at the single cell level.

5.2 Results

5.2.1 *Sleeping Beauty* SRTs are not uniformly distributed across the genome
Sleeping Beauty (SB) has previously been shown to have little chromatin preference (Yoshida et al., 2017), a finding we confirmed with a much larger dataset generated from self-reporting

transposons (SRTs) (Moudgil et al., 2020). However, at that time our analysis was focused on enhancers and super-enhancers, which are found at a scale of 10^3 to 10^5 bases (Whyte et al., 2013). Genomic DNA is hierarchically organized into organizational layers such as nucleosome clutches and chromatin fibers which are dynamically modulated by transcription and during differentiation (Ricci et al., 2015). Moreover, the nucleus is itself a highly heterogeneous organelle containing numerous phase-separated subdivisions (Hnisz et al., 2017; Klein et al., 2020; Padrón et al., 2019; Sabari et al., 2018) . Thus, while SB may have little transpositional variation at small length scales, it was unclear if this uniformity was preserved more globally.

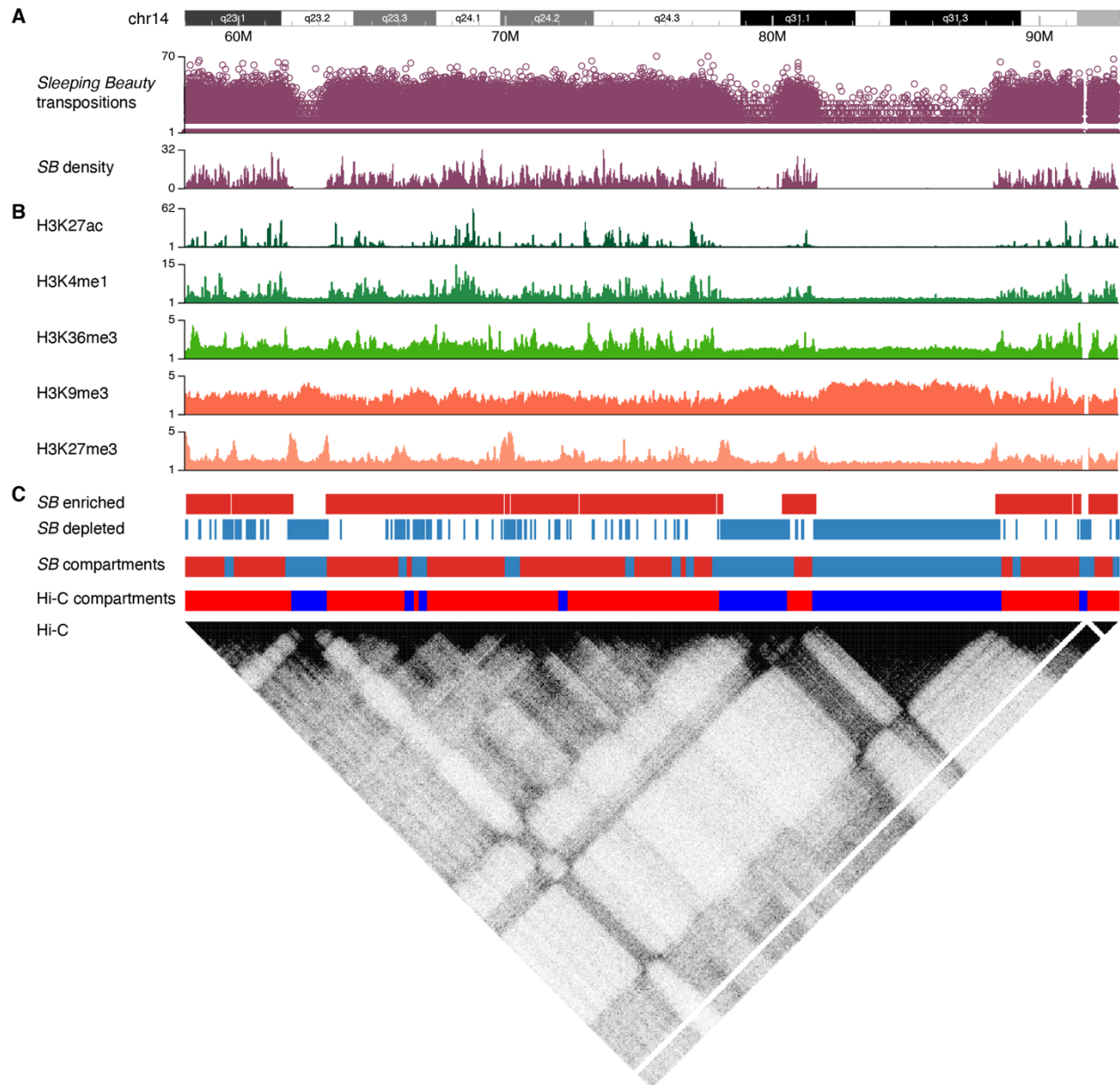


Figure 5.1: SB insertion densities are correlated with chromosomal compartment. (A) Raw insertions and insertion density profiles for SB SRTs in HCT-116 cells. (B) ChIP-seq for post-translationally modified histones. (C) Blocks of genomic DNA statistically enriched and depleted for SB insertions can be used to reconstruct chromosomal compartments. Hi-C compartmentalization and contact matrix are also shown.

We re-analyzed the SRT dataset and found that, on large length scales (10^6 to 10^7 bases), SB transposition is indeed non-uniform (Figure 5.1A). Insertions appear to fall into alternating regions of high and low, with graded transitions between them visible in the insertion track. While the density of insertions within a region seems constant, the variegation between adjacent

regions is only apparent when juxtaposed on such a large scale. Finally, it is worth noting that these high- and low-SB regions form relatively long and contiguous segments, with few interruptions.

To better understand factors contributing to this non-uniform transposition, we cross-referenced our insertion data against published modified histone chromatin immunoprecipitation sequencing (ChIP-seq) data. We observed that regions with high SB insertion frequency tended to also have higher levels of the epigenetic marks H3K27ac, H3K4me1, and H3K36me3. These modifications are frequently associated with actively transcribed regions of the genome (Lawrence et al., 2016). Conversely, regions with comparatively lower transposition rates were enriched for the repressive marks H3K9me3 and H3K27me3, with the latter most prominently seen flanking the former. We conclude that, in a broad sense, the local density of SB transposition is positively correlated with whether genomic DNA is likely to be transcribed.

5.2.2 Densities of *Sleeping Beauty* SRTs reveal chromosomal compartments

One level of nuclear organization binarizes chromosomal sequences into one of two compartments, “A” and “B”. These states are typically inferred from high-throughput assays, such as Hi-C, and reflect intrachromosomal contact frequencies (Lieberman-Aiden et al., 2009). Namely, “A” compartment sequences are more likely to be in close spatial proximity to one another (similarly for “B” compartment sequences) and are less likely to be near “B” compartment sequences. Moreover, compartment A tends to be centrally located in the nucleus and transcriptionally active, while the compartment B is found along the nuclear periphery and is less likely to be transcribed (Therizols et al., 2014).

Since SB insertions show density changes on the megabase scale, and these changes correlate with transcriptionally informative epigenetic marks, we hypothesized that the variation

seen in SB insertion density reflects the underlying chromosomal compartmentalization. To test this, we analyzed published Hi-C data (Rao et al., 2017) and identified chromosomal compartments (Durand et al., 2016). We then measured the distribution of insertion densities in A and B compartments. We observed that the density of insertions is indeed higher in A compartments over B compartments (Figure 5.2A). Although the magnitude of this effect is modest, the differences are not likely by chance (Mann-Whitney U p -value $< 10^{-9}$). Thus, we conclude that SB is biased by chromosomal compartmentalization.

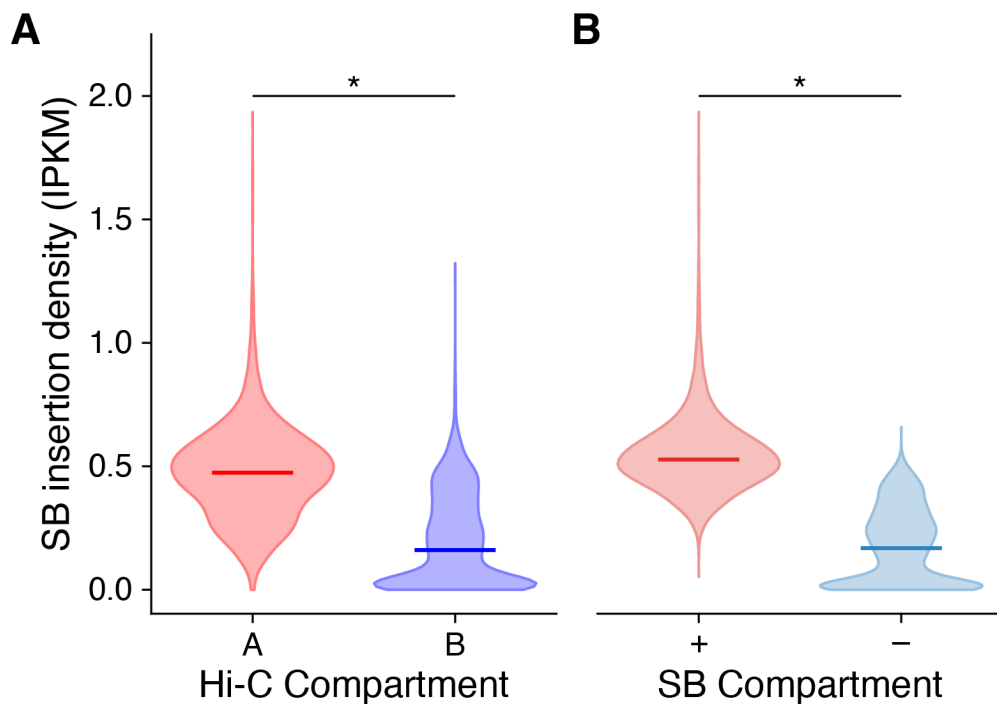


Figure 5.2: Comparison of SB insertion densities by compartment. (A) Distribution of insertions per kilobase per million mapped insertions (IPKM) in A and B compartments as inferred by Hi-C. (B) Distribution of IPKM in SB+ and SB- compartments. Horizontal lines represent medians of the distributions. * Mann-Whitney U p -value $< 10^{-9}$.

Since we observed such a striking connection between compartment state and insertion density, we wondered whether it was possible to identify chromosomal compartments from SB insertions alone. We first used a density-based approach (Chapter 4) to segment the genome into regions of extreme enrichment or depletion for SB insertions (Figure 5.1C). We then smoothed

these data using the same fixed-width bins as the Hi-C chromosomal compartments. To differentiate our analysis from the Hi-C-based compartments, we annotate our states as either SB⁺ or SB⁻, to reflect high- and low-density regions, respectively.

Visual inspection of SB compartments reveals close concordance with Hi-C compartments (Figure 5.1C). The Rand index, a metric used to evaluate the similarity between two segmentations (Truong et al., 2020), was 0.757 between SB and Hi-C compartments, considerably better than chance (0.5). In addition, the distributions of insertion densities within SB compartments is both qualitatively similar to those within Hi-C compartments and significantly different between SB⁺ and SB⁻ states (Figure 5.2B; Mann-Whitney U p -value < 10⁻⁹). To functionally validate our findings, we compared the genome-wide enrichments in ChIP-seq signals between SB⁺ and SB⁻ compartments. The latter showed greater average intensities for the repressive marks H3K9me3 and H3K27me3 (Figure 5.3A-B) than the former. We observed the opposite trend for activating marks, with SB⁺ being enriched for H3K27ac, H3K4me1, and H3K36me3 (Figure 5.3C-E). Thus, SB compartments appear to demarcate transcriptionally active and inactive domains.

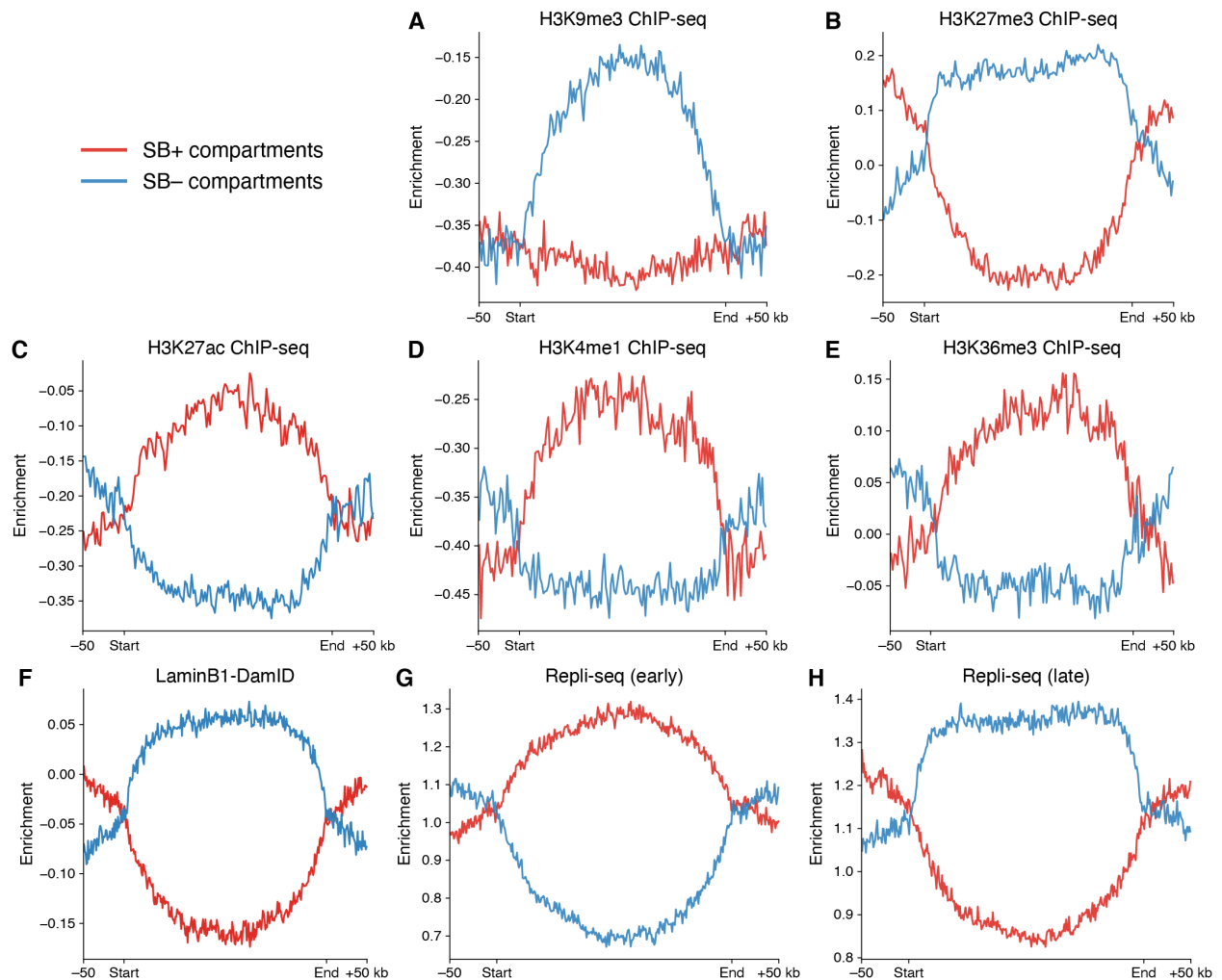


Figure 5.3: Functional validation of SB compartments. Differential enrichment in SB+ and SB– compartments for (A) H3K9me3 ChIP-seq, (B) H3K27me3 ChIP-seq, (C) H3K27ac ChIP-seq, (D) H3K4me1 ChIP-seq, (E) H3K36me3 ChIP-seq, (F) LaminB1-DamID, (G) early replicating DNA, and (H) late replicating DNA. The units for the y-axis in A-F is \log_2 fold-change; in G-H, it is normalized coverage.

The B compartment is often associated with the nuclear lamina, which serves to anchor and organize chromatin (van Schaik et al., 2019). We quantitated the enrichment of LaminB1, a major lamina component, in SB+ and SB– compartments. LaminB1 signal was greater in SB– compartments than SB+ (Figure 5.3F), suggesting that former is likely to be interacting with the nuclear lamina. Lamina-associated domains are among the last regions of the genome to replicate (Hansen et al., 2010). Cross-referencing our annotations with replication timing data revealed that SB+ compartments are more likely to replicate early and SB- compartments more likely to

replicate late (Figure 5.3G-H). This is additional evidence that SB transposition rates can discriminate chromosomal compartments. Furthermore, it supports the notion that SB densities reflect genome organization.

In addition to the nuclear lamina, the nucleolus is also an organizational focus for the nucleus. In particular, nucleolus organizer regions (NORs) on the short arms of the acrocentric autosomes help form the nucleolus (van Sluis et al., 2019). We investigated if SB can access NORs and whether insertion densities would resemble SB+ or SB- compartments. NORs are rich in ribosomal DNA (rDNA) repeats and therefore difficult to assemble. As such, only one NOR is present in the human reference genome. Our analysis revealed few insertions, classifying it as part of an SB- compartment (Figure 5.4). This suggests that SB does not readily penetrate the nucleolus, though we consider this finding preliminary. The repetitive nature of NORs makes it difficult to align reads to them, which would also result in reduced insertion densities. As genome assemblies and aligners continue to improve (Miga et al., 2020), it may be possible to confirm this finding with across multiple NORs.

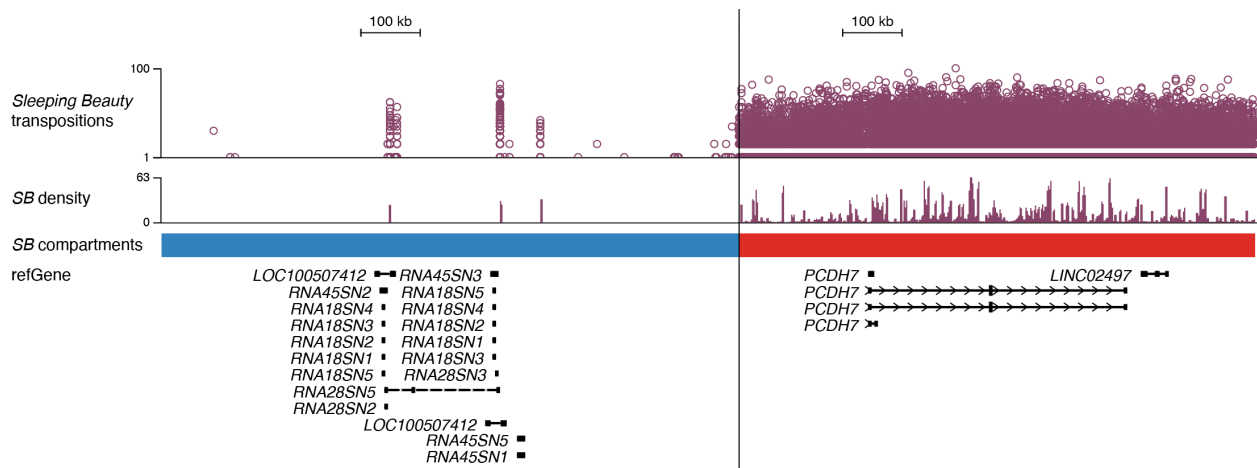


Figure 5.4: SB compartment analysis of a nucleolus organizing region (NOR). Left, NOR on chromosome 21. Right, euchromatic locus in SB+ compartment for comparison.

5.2.3 Inferring compartmentalization from sparse *Sleeping Beauty* data

We have recently shown that SRTs can be recovered from single cell RNA-seq (scRNA-seq)

libraries and can be stratified by cell type (Moudgil et al., 2020). This raises the possibility that

SB SRTs could also be recovered from scRNA-seq libraries, enabling simultaneous

identification of cell identity and cell-type-specific chromosomal compartmentalization.

However, single cell genomic libraries tend to be much sparser than their bulk counterparts. To

predict how accurately single cell SB libraries could identify compartments, we downsampled

our bulk SB dataset and measured the Rand index against a range of Hi-C compartment

resolutions (Figure 5.5A). We were able to call compartments with greater accuracy than

chance—corresponding to a Rand index above 0.5—with as few as 30,000 insertions, though in

practice we expect this to be an extreme lower bound. SB is a highly active transposase and we

should be able to routinely collect at least 100,000 insertions per single cell library, yielding a

minimum Rand index of 0.68. As single cell technologies improve, or by increasing the number

of cells analyzed, we may be able to push the Rand index closer to 0.75 at the coarsest scales.

We also quantitated, separately, our sensitivity to A and B compartments (Figure 5.5B-C). In the

range of a few hundred thousand insertions, SB insertions are about equally sensitive to A and B

compartments (69-77% and 66-69%, respectively). We conclude that in single cells SB insertion

profiles can be a reasonably accurate way to measure cell type-specific genome organization.

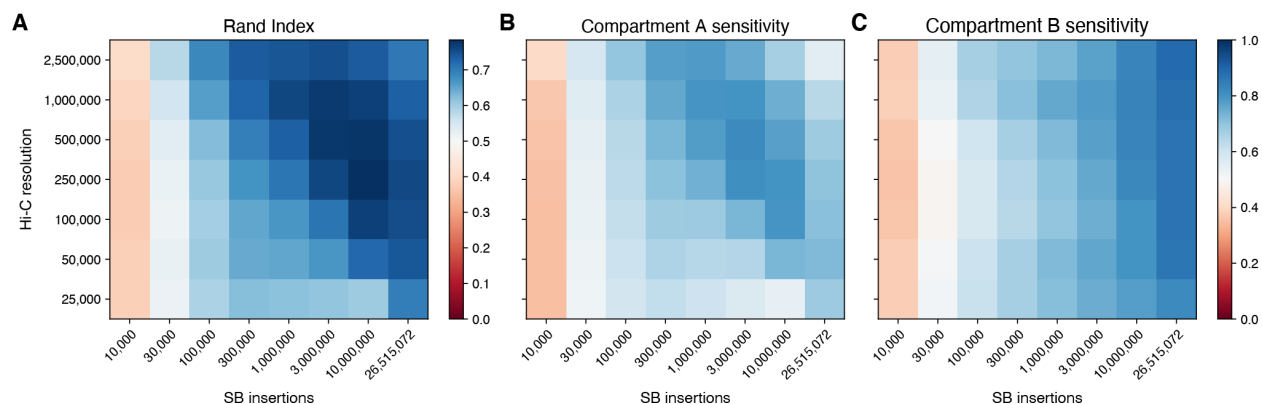


Figure 5.5: Downsampling analysis of SB compartments. (A) Rand index of similarity between Hi-C and SB compartments at a range of resolutions and library sizes. (B) Sensitivity of downsampled SB insertions to compartment A. (C) Sensitivity of downsampled SB insertions to compartment B.

5.3 Discussion

While *Sleeping Beauty* (SB) has previously been reported to have little chromatin preference (Yoshida et al., 2017), a more recent analysis claims that it has a modest bias towards transcriptionally active regions (Sultana et al., 2019). Here, we confirm the latter trend, finding that SB insertion profiles tend to align with chromosomal compartments. Regions with increased SB density tend to overlap with compartment A, which itself is positively correlated with gene expression. Conversely, we observed that regions depleted for SB insertions coincided with compartment B. What is remarkable about these findings is that we were able to infer two-dimensional spatial correlations from one-dimensional transposition data. Interestingly, other groups have also attempted to reconstruct compartments from epigenetic data but from more established techniques like ChIP-seq and the assay for transposase-accessible chromatin (ATAC-seq) (Fortin and Hansen, 2015; Zhu et al., 2016).

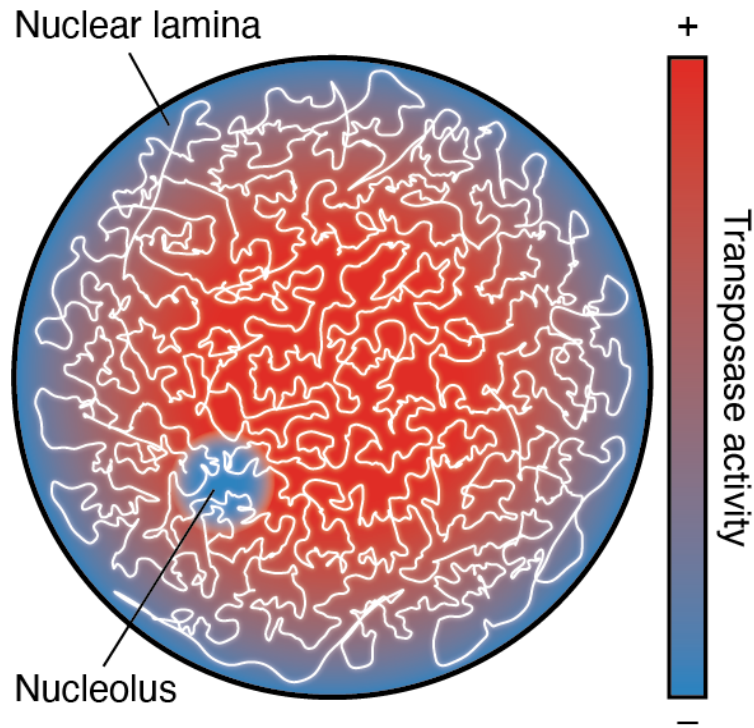


Figure 5.6: Model of transposase activity as function of sub-nuclear localization. White line represents genomic DNA.

Compartment A has been previously found to be more centrally located in the nucleus, enabling it to, among other things, replicate earlier in the cell cycle (Therizols et al., 2014). Compartment B, in contrast, tends to accumulate around the periphery, associating with the nuclear lamina (Ramani et al., 2016). Our data further suggests a spatial model of transposase activity, with transposition efficiency decreasing radially away from the nucleus's center (Figure 5.6). Indeed, this has already been suggested for LINE-1 retrotransposons (Sultana et al., 2019). We reanalyzed published *piggyBac* transposition data in light of our findings. Although *piggyBac* has particular localization preferences at the sub-compartmental level (Moudgil et al., 2020; Yoshida et al., 2017), globally it deposits transposons in a similar pattern to SB (data not shown). Thus, this framework may be generally applicable to many transposases. Quantitatively developing this model may enhance our understanding of native transposases and could better inform analyses of exogenous assays, such as transposon calling cards (Wang et al., 2012).

SRTs can be recovered from scRNA-seq libraries, which enables joint identification of cell type and cell-specific transpositions. Aggregated *piggyBac* transposition profiles have led to new insights into cell-type-specific regulatory processes (Moudgil et al., 2020). In a similar fashion, single cell recovery of SB SRTs could illuminate chromosomal compartmentalization resolved by cell type. Compartments can be dynamic, switching over the course of cellular differentiation (Criscione et al., 2016; Dixon et al., 2015). While our downsampling analysis suggests that single cell analyses of SB compartments can be reasonably accurate, these metrics were measured in a single cell line at steady state. Future work should explore whether SB transposition, particularly from single cells, is sensitive to changes in compartmentalization. The most straightforward approach would be *in vitro* perturbation with a small molecule (Kantidze et al., 2019) or in a cell culture model of differentiation (Reimer et al., 2021).

5.4 Methods

Sleeping Beauty SRT data in HCT-116 cells were obtained from (Moudgil et al., 2020). Raw insertion data were first segmented using blockify 0.2.1 (Chapter 4) to identify contiguous blocks with piecewise-constant density. We next determined whether a block was either enriched or depleted for insertions by performing one-tailed Poisson hypothesis tests. Specifically, we ran “blockify call,” using a Bonferroni-adjusted p -value threshold of 0.05, a pseudocount of 0, and a merge window of 0, and setting the “measure” parameter to either “enrichment” or “depletion.” The null hypothesis we used was that insertions should be uniformly distributed with respect to all TA dinucleotides (SB’s insertion motif) across the genome. Finally, to determine SB+ and SB- compartments, we partitioned the genome into non-overlapping, fixed width windows; intersected each window with the sets of enriched and depleted blocks; and calculated an average score per window. The score was calculated by assigning to each base a value: 1 if it overlapped

an enriched block, -1 if it overlapped a depleted block, and 0 otherwise. The window score was then the sum of all scored bases divided by the length of the window. From this, SB+ windows were those with positive scores and SB- windows were those with negative scores. For the sparsity analysis, downsampling was performed using blockify using a random seed of 0 .

Hi-C data were obtained from GEO datasets SRR6107782-SRR610781 and were processed using juicer (Durand et al., 2016) with the following settings: “-q general -l general -s MboI -Q 10080 -L 10080.” Data were aligned to the hg38 human genome assembly and visualized on the UCSC Genome Browser. ChIP-seq data and processing steps were taken from (Moudgil et al., 2020). Processed DamID and LMNB1-DamID data were downloaded from the 4D Nucleome Data Portal (the 4D Nucleome Network et al., 2017), accession numbers 4DNFI9XJQPIZ and 4DNFIMFJN73W, respectively (van Schaik et al., 2020). Aligned early and late Repli-seq datasets were also downloaded from the 4D Nucleome Data Portal, accession numbers 4DNFIBQM2JE2 and 4DNFIA5J1HN7, respectively. Epigenomic profile plots were generated using deeptools 3.0.0 (Ramírez et al., 2016). General data analysis and visualization was performed using numpy 1.16.2 (Oliphant, 2015), scipy 1.2.1 (Virtanen et al., 2020), matplotlib 3.0.3 (Hunter, 2007), statsmodels 0.8.0 (Seabold and Perktold, 2010), and Python 3.6.5.

5.5 References

Angermueller, C., Clark, S.J., Lee, H.J., Macaulay, I.C., Teng, M.J., Hu, T.X., Krueger, F., Smallwood, S., Ponting, C.P., Voet, T., et al. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods* 13, 229–232.

Arrastia, M.V., Jachowicz, J.W., Ollikainen, N., Curtis, M.S., Lai, C., Quinodoz, S.A., Selck, D.A., Ismagilov, R.F., and Guttman, M. (2022). Single-cell measurement of higher-order 3D genome organization with scSPRITE. *Nat Biotechnol* 40, 64–73.

Cao, J., Packer, J.S., Ramani, V., Cusanovich, D.A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S.N., Steemers, F.J., et al. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* 357, 661–667.

Cao, J., Cusanovich, D.A., Ramani, V., Aghamirzaie, D., Pliner, H.A., Hill, A.J., Daza, R.M., McFaline-Figueroa, J.L., Packer, J.S., Christiansen, L., et al. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* 33, eaau0730.

Ciscione, S.W., De Cecco, M., Siranosian, B., Zhang, Y., Kreiling, J.A., Sedivy, J.M., and Neretti, N. (2016). Reorganization of chromosome architecture in replicative cellular senescence. *Sci. Adv.* 2, e1500882.

Dekker, J. (2002). Capturing Chromosome Conformation. *Science* 295, 1306–1311.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 376–380.

Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature* 518, 331–336.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., et al. (2006). Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Research* 16, 1299–1309.

Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems* 3, 95–98.

Fortin, J.-P., and Hansen, K.D. (2015). Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biol* 16, 180.

Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., et al. (2009). An oestrogen-receptor- α -bound human chromatin interactome. *Nature* 462, 58–64.

Hansen, R.S., Thomas, S., Sandstrom, R., Canfield, T.K., Thurman, R.E., Weaver, M., Dorschner, M.O., Gartler, S.M., and Stamatoyannopoulos, J.A. (2010). Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences* 107, 139–144.

Hnisz, D., Shrinivas, K., Young, R.A., Chakraborty, A.K., and Sharp, P.A. (2017). A Phase Separation Model for Transcriptional Control. *Cell* 169, 13–23.

Hsieh, T.-H.S., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., and Rando, O.J. (2015). Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell* 162, 108–119.

Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* 9, 90–95.

Kantidze, O.L., Luzhin, A.V., Nizovtseva, E.V., Safina, A., Valieva, M.E., Golov, A.K., Velichko, A.K., Lyubitelev, A.V., Feofanov, A.V., Gurova, K.V., et al. (2019). The anti-cancer drugs curaxins target spatial genome organization. *Nat Commun* 10, 1441.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* 161, 1187–1201.

Klein, I.A., Boija, A., Afeyan, L.K., Hawken, S.W., Fan, M., Dall’Agnese, A., Oksuz, O., Henninger, J.E., Shrinivas, K., Sabari, B.R., et al. (2020). Partitioning of cancer therapeutics in nuclear condensates. *Science* 368, 1386–1392.

Krietenstein, N., Abraham, S., Venev, S.V., Abdennur, N., Gibcus, J., Hsieh, T.-H.S., Parsi, K.M., Yang, L., Maehr, R., Mirny, L.A., et al. (2020). Ultrastructural Details of Mammalian Chromosome Architecture. *Molecular Cell* 78, 554-565.e7.

Lawrence, M., Daujat, S., and Schneider, R. (2016). Lateral Thinking: How Histone Modifications Regulate Gene Expression. *Molecular Cell* 32, 42–56.

Le Dily, F., Baù, D., Pohl, A., Vicent, G.P., Serra, F., Soronellas, D., Castellano, G., Wright, R.H.G., Ballare, C., Filion, G., et al. (2014). Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes Dev.* 28, 2151–2162.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragozy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* 326, 289–293.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 161, 1202–1214.

Miga, K.H., Koren, S., Rhie, A., Vollger, M.R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G.A., et al. (2020). Telomere-to-telomere assembly of a complete human X chromosome. *Nature*.

Moudgil, A., Wilkinson, M.N., Chen, X., He, J., Cammack, A.J., Vasek, M.J., Lagunas, T., Qi, Z., Lalli, M.A., Guo, C., et al. (2020). Self-Reporting Transposons Enable Simultaneous Readout of Gene Expression and Transcription Factor Binding in Single Cells. *Cell* 182, 992-1008.e21.

Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* 502, 59–64.

Oliphant, T.E. (2015). *Guide to NumPy* (Austin, Tex.: Continuum Press).

Padrón, A., Iwasaki, S., and Ingolia, N.T. (2019). Proximity RNA Labeling by APEX-Seq Reveals the Organization of Translation Initiation Complexes and Repressive RNA Granules. *Molecular Cell* 75, 875-887.e5.

Plasschaert, L.W., Žilionis, R., Choo-Wing, R., Savova, V., Knehr, J., Roma, G., Klein, A.M., and Jaffe, A.B. (2018). A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* *560*, 377–381.

Quinodoz, S.A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J.M., Detmar, E., Lai, M.M., Shishkin, A.A., Bhat, P., Takei, Y., et al. (2018). Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell* *174*, 744-757.e24.

Ramani, V., Shendure, J., and Duan, Z. (2016). Understanding Spatial Genome Organization: Methods and Insights. *Genomics, Proteomics & Bioinformatics* *14*, 7–20.

Ramani, V., Deng, X., Qiu, R., Gunderson, K.L., Steemers, F.J., Disteche, C.M., Noble, W.S., Duan, Z., and Shendure, J. (2017). Massively multiplex single-cell Hi-C. *Nat Methods* *14*, 263–266.

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* *44*, W160–W165.

Rao, S.S.P., Huang, S.-C., Glenn St Hilaire, B., Engreitz, J.M., Perez, E.M., Kieffer-Kwon, K.-R., Sanborn, A.L., Johnstone, S.E., Bascom, G.D., Bochkov, I.D., et al. (2017). Cohesin Loss Eliminates All Loop Domains. *Cell* *171*, 305-320.e24.

Reimer, K.A., Mimoso, C.A., Adelman, K., and Neugebauer, K.M. (2021). Co-transcriptional splicing regulates 3' end cleavage during mammalian erythropoiesis. *Molecular Cell* *81*, 998-1012.e7.

Ricci, M.A., Manzo, C., García-Parajo, M.F., Lakadamyali, M., and Cosma, M.P. (2015). Chromatin Fibers Are Formed by Heterogeneous Groups of Nucleosomes In Vivo. *Cell* *160*, 1145–1158.

Rooijers, K., Markodimitraki, C.M., Rang, F.J., de Vries, S.S., Chialastri, A., de Luca, K.L., Mooijman, D., Dey, S.S., and Kind, J. (2019). Simultaneous quantification of protein–DNA contacts and transcriptomes in single cells. *Nat Biotechnol* *37*, 766–772.

Sabari, B.R., Dall’Agnese, A., Bojja, A., Klein, I.A., Coffey, E.L., Shrinivas, K., Abraham, B.J., Hannett, N.M., Zamudio, A.V., Manteiga, J.C., et al. (2018). Coactivator condensation at super-enhancers links phase separation and gene control. *Science* *361*, eaar3958.

van Schaik, T., Vos, M., Peric-Hupkes, D., and van Steensel, B. (2019). Cell cycle dynamics of lamina associated DNA (bioRxiv).

van Schaik, T., Vos, M., Peric-Hupkes, D., HN Celie, P., and van Steensel, B. (2020). Cell cycle dynamics of lamina-associated DNA. *EMBO Rep* *21*.

Seabold, S., and Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In 9th Python in Science Conference, p.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). *Nat Genet* *38*, 1348–1354.

van Sluis, M., Gailín, M.Ó., McCarter, J.G.W., Mangan, H., Grob, A., and McStay, B. (2019). Human NORs, comprising rDNA arrays and functionally conserved distal elements, are located within dynamic chromosomal regions. *Genes Dev.* *33*, 1688–1701.

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K., Swerdlow, H., Satija, R., and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods* *14*, 865–868.

Sultana, T., van Essen, D., Siol, O., Bailly-Bechet, M., Philippe, C., Zine El Aabidine, A., Pioger, L., Nigumann, P., Sacconi, S., Andrau, J.-C., et al. (2019). The Landscape of L1 Retrotransposons in the Human Genome Is Shaped by Pre-insertion Sequence Biases and Post-insertion Selection. *Molecular Cell* *74*, 555-570.e7.

the 4D Nucleome Network, Dekker, J., Belmont, A.S., Guttman, M., Leshyk, V.O., Lis, J.T., Lomvardas, S., Mirny, L.A., O’Shea, C.C., Park, P.J., et al. (2017). The 4D nucleome project. *Nature* *549*, 219–226.

Therizols, P., Illingworth, R.S., Courilleau, C., Boyle, S., Wood, A.J., and Bickmore, W.A. (2014). Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. *Science* *346*, 1238–1242.

Truong, C., Oudre, L., and Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing* *167*, 107299.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* *17*, 261–272.

Wang, H., Mayhew, D., Chen, X., Johnston, M., and Mitra, R.D. (2012). “Calling Cards” for DNA-Binding Proteins in Mammalian Cells. *Genetics* 190, 941–949.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell* 153, 307–319.

Yoshida, J., Akagi, K., Misawa, R., Kokubu, C., Takeda, J., and Horie, K. (2017). Chromatin states shape insertion profiles of the piggyBac, Tol2 and Sleeping Beauty transposons and murine leukemia virus. *Scientific Reports* 7, 43613.

Zhao, Z., Tavoosidana, G., Sjölander, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Singh Sandhu, K., Singh, U., et al. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet* 38, 1341–1347.

Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications* 8, 14049.

Zhu, Y., Chen, Z., Zhang, K., Wang, M., Medovoy, D., Whitaker, J.W., Ding, B., Li, N., Zheng, L., and Wang, W. (2016). Constructing 3D interaction maps from 1D epigenomes. *Nat Commun* 7, 10812.

Chapter 6: Future directions

The self-reporting transposon (SRT) is fascinating in that it broadcasts a traditional, DNA-based signal—genomic location—into mRNA, allowing it to be detected, mapped, and read out alongside bulk and single RNA-seq data. There are several future directions and potential applications for this technology. While some of these have been mentioned in previous chapters, they are collected here, expanded in scope and number.

6.1 Future directions for *piggyBac* transposon calling cards

In its current form, SRTs rely on the constitutive expression of a strong promoter (EF-1a) to generate self-reporting transcripts. This is useful for maximizing sensitivity across a range of cell types (Cammack et al., 2020) but could perturb local gene expression. Strong promoters can function as local enhancers (Medina-Rivera et al., 2018). The number of calling cards deposited per cell is relatively small (50-100) and while we do not expect widespread dysregulation of gene expression, we cannot eliminate the possibility of calling cards altering transcription of native genes. One way to mitigate this concern would be to use an inducible promoter (Qin et al., 2010), pulsing with the inducer just before collecting cells to minimize perturbation.

Silencing of promoters is another concern with sustained transcription of transgenes. Long-term expression can lead to methylation of promoters, decreasing expression over time. We did not suffer from appreciable silencing in either our short-course *in vitro* experiments (Moudgil et al., 2020) nor in long-term experiments with intracranial calling cards (Cammack et al., 2020). However, we were not specifically looking for silencing nor can we guarantee that it will not happen in the future. Ubiquitous chromatin-opening elements (UCOEs) are cis-

regulatory sequences that are resistant to DNA methylation and have been used to counteract silencing of lentiviral transgenes (Nair et al., 2011; Saunders et al., 2015; Zhang et al., 2007, 2010). Including a UCOE in the SRT could help maintain expression over longer time courses.

Finally, it may be worth considering non-Pol II promoters to drive SRTs. The simplest solution is to include a viral T7 RNA polymerase (RNAP) promoter just before the terminal repeat. Libraries could then be prepared from genomic DNA following an *in vitro* transcription reaction. This approach, already in use to amplify single cell genomes (Chen et al., 2017a), would evade the issues of ectopic enhancement and promoter silencing. The inclusion of the T7 RNAP promoter in Ty5 retrotransposons could also help bring SRTs to yeast calling cards (Mayhew and Mitra, 2016). There may also be uses for SRTs containing RNA polymerase III promoters, such as the popular U6 sequence for synthesizing Cas9 guide RNAs (Cong et al., 2013; Mali et al., 2013). However, this would only work with transposons that do not have polythymidine tracts in their terminal repeats.

One of the drawbacks of bulk RNA calling cards is the necessity for multiple biological replicates to build up statistical power. In DNA calling cards, this was accomplished using a polyclonal pool of donor plasmids containing different internal barcodes. This feature was dropped in our initial development of the SRT because of the prohibitive distance between the barcode and the transposon-genome junction. Recent work by Matthew Lalli demonstrates that the *piggyBac* terminal repeat can be mutated to generate barcoded vectors (Lalli et al., 2021). Moreover, structural analysis of the *piggyBac* transpososome revealed partial duplication of the terminal repeat can increase transposition efficiency (Chen et al., 2020). While these plasmids will immediately benefit bulk calling cards protocols, they may eventually be useful in single cell calling cards to further enhance our ability to recover SRTs.

Individual transcription factors often bind in complexes. For example, we showed that BAP1 calling cards can recover the motif of YY1, a known binding partner (Moudgil et al., 2020). Multiplexing calling cards to record the binding of several TFs would provide further mechanistic insight into the binding and activity of TF collectives. One way to achieve this is to use multiple transposase families, with each transposase fused to a different TF. Since transposons are specifically transposed by their cognate transposase, the localization of a given transposon would unequivocally reflect a single TF's binding specificity. This strategy may be difficult in practice because not all transposases tolerate fusion (see below). Alternatively, multiple TF-*piggyBac* fusions could be deployed at once *in vitro* and deconvolved with single cell calling cards. This approach requires each fusion to be barcoded by a TF-specific barcode sequence in the 3' untranslated region (UTR) downstream of the transposase (assuming the TF is fused at the N-terminus). Viral transduction of these fusions at a low multiplicity of infection (MOI) would ensure that a cell surviving selection received a single TF-*piggyBac* fusion. Single cell RNA-seq analysis of the *piggyBac* transcripts would be able to detect the TF barcode and identify which construct was delivered to each cell. SRT insertions could then be stratified by cells sharing the same TF. This would reveal TF co-binding patterns at a pseudobulk level. Similar approaches are already used in single cell CRISPR perturbation experiments (Datlinger et al., 2017; Dixit et al., 2016).

We and others have characterized *piggyBac*'s inordinate affinity for bromodomain proteins, particularly the coactivator BRD4 (Gogol-Döring et al., 2016; Moudgil et al., 2020; Yoshida et al., 2017). Western blot analysis of bromodomain truncations showed that it is the C-terminal domains of the bromodomain and extra-terminal motif (BET) proteins BRD2, BRD3, and BRD4 that directly interact with *piggyBac* (Gogol-Döring et al., 2016). Intriguingly, this is

the same part of BRD4 that is intrinsically disordered and is thought to play a role in cooperatively forming intranuclear transcriptional condensates (Sabari et al., 2018). Even more provocative is that *piggyBac* itself has disordered regions at its N- and C-termini (Figure 1). This suggests a potential target to wean *piggyBac* of its bromodomain affinity. If successful, these new transposases would have greater redirectability by TFs and resolve sharper peaks. They may also be useful for gene therapy, a field that has been frustrated by the inability of transposases to deliver transgenes with targeted site specificity (Tipanee et al., 2017; Vargas et al., 2016).

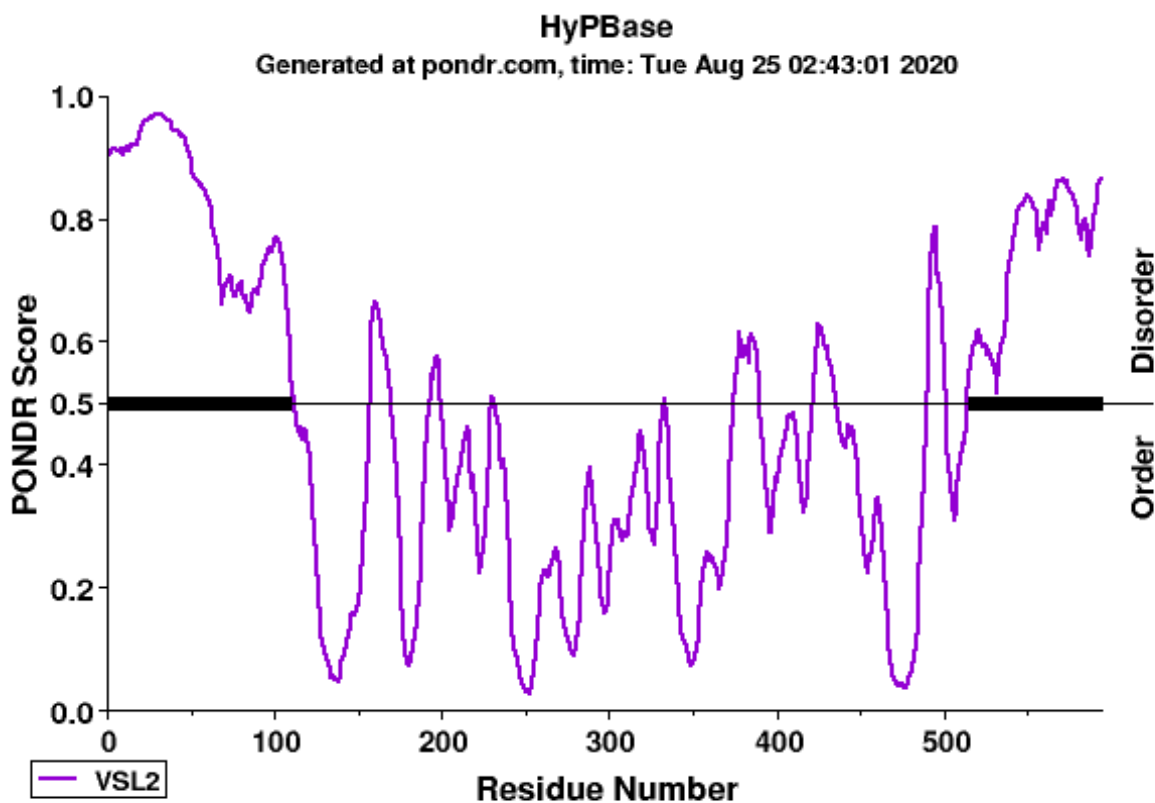


Figure 6.1: Predicted disorder for the hyperactive *piggyBac* transposase. The core catalytic domain lies between residues 130-522 (Morellet et al., 2018).

The C-terminal domain of *piggyBac*, while predicted to be disordered (Figure 6.1), is necessary for binding the terminal repeat sequences (Morellet et al., 2018). In contrast, little is known about the N-terminus and its function. In addition to the first 110 residues being

disordered, the first 80 or so resemble an acidic activation domain (Alex Holehouse, personal communication). Based on these observations, there are number of possible experiments to try. First, generating a set of N-terminal *piggyBac* truncations would elucidate which residues, if any, are necessary for transposition. Similar work has already been done but was complicated by the inclusion of additional peptide tags (Lalli, personal communication). Arginines are thought to favor condensation and the N-terminus of *piggyBac* is rich in them (Holehouse, personal communication). To see if they contribute to *piggyBac*'s bromodomain affinity, they could be mutated to lysines. Hydrophobicity is another sequence property that contributes to disorder (Holehouse, personal communication), so to test its contribution to *piggyBac*'s behavior, the hydrophobic amino acids can be mutated to either serine or glycine would abolish. Collectively, these experiments should provide a reasonable starting point for understanding the role of *piggyBac*'s N-terminus.

It is curious that N-terminus of *piggyBac* should resemble an acidic activation domain such as that of the yeast TF Gcn4p (Staller et al., 2018). From where did this domain originate? Was it part of an ancestral *piggyBac* transposase or was it co-opted from gene shuffling over evolutionary time? Molecular evolutionary analyses of various *piggyBac*-like and *piggyBac*-derived genes showed poor conservation of the N-terminal domain (Bouallègue et al., 2017; Sarkar et al., 2003). Intrinsically disordered regions, however, are challenging for multiple sequence alignment algorithms (Vriend et al., 2016). A more sophisticated approach, taking into account the disordered behavior, may prove more fruitful. Then there is the question of function: what fitness advantage does such a domain confer to *piggyBac*? It is not unreasonable that transposing into highly transcribed regions, especially in the germline, maximizes the chances of propagating to the next generation. Interestingly, two of the point mutations used to generate

hyperactive *piggyBac* (I30V and S103P; (Yusa et al., 2011)) fall within the N-terminal disordered region. These residues may have been under balancing selection in wild type *piggyBac* to minimize mutagenic burden. Finally, we can ask whether this domain has any ability to transactivate expression. Why the transposase should benefit from such activity is not clear, but neither is the possibility something we can rule out. Perhaps *piggyBac* has evolved a transcriptionally inactive mimic to other activation domains, leading it to active loci without necessarily perturbing gene expression itself. Complementation assays involving the exchange of *piggyBac*'s and Gcn4p's acidic domain, coupled to either growth assays or a reporter gene (Liu et al., 2020; Shively et al., 2019), would be one approach to answering this question.

Murine leukemia virus (MLV) also has an insertion profile strongly biased by bromodomain proteins (Gogol-Döring et al., 2016; Yoshida et al., 2017). MLV and other retroviruses have been extensively scrutinized because of their promise to deliver gene therapies and the notable adverse reactions they have caused in clinical trials (Bushman, 2007; Hacein-Bey-Abina, 2003; Hacein-Bey-Abina et al., 2003). Recently, a comparative analysis across gammaretroviruses revealed a conserved motif at the integrase C-terminus that mediated the association with bromodomains (El Ashkar et al., 2014). Deletion of these residues, or a single point mutation (W390A), largely abrogated the bromodomain preference (El Ashkar et al., 2014). Furthermore, fusing other chromatin-binding domains to this mutant MLV could redirect integrations (El Ashkar et al., 2017). These achievements are viral parallels to our goal of redirecting *piggyBac*. Out of curiosity, we aligned the conserved BET interaction motif of Moloney MLV (MMLV) against the first 110 residues of *piggyBac* and hyperactive *piggyBac* (Figure 2). We found that the end of *piggyBac*'s N-terminus does contain residues resembling MMLV's BET interaction motif. Most intriguing is the preserved tryptophan residue (denoted by

the asterisk) that is critical for MMLV's association with bromodomains. Therefore, targeted mutagenesis of these residues may offer an alternative path to disassociating *piggyBac* from BETs and, consequently, enhancing redirection by TFs.

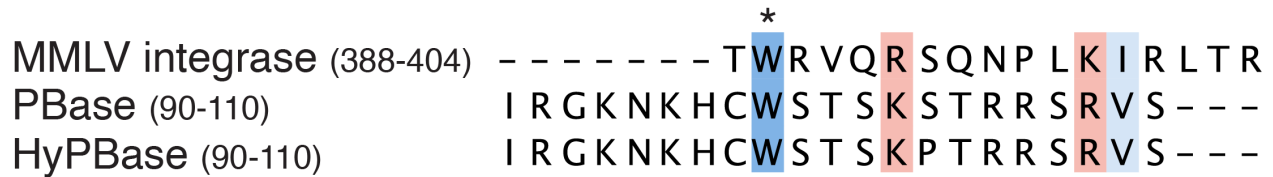


Figure 6.2: Alignment of MMLV's BET interaction motif against *piggyBac*. Multiple sequence alignment of Moloney murine leukemia virus's (MMLV) BET interaction motif against a portion of the N-termini of *piggyBac* (PBase) and hyperactive *piggyBac* (HyPBBase) transposases. Conserved residues are colored using the ClustalX palette (blue: hydrophobic; red: positively charged). The asterisk denotes the critical tryptophan necessary for MMLV's BET interaction.

6.2 Future directions for *Sleeping Beauty* transposon calling cards

We used *Sleeping Beauty*'s broad and considerably (though not completely) uniform insertion profile to identify patterns of chromosomal compartmentalization. There is certainly room for continued computational and biological work to, respectively, improve compartment assignment and study dynamic changes in genome organization. There is also the opportunity to simultaneously multiplex both *piggyBac* and *Sleeping Beauty* SRTs in the same cells, leading to triple measurements characterizing cell identity, enhancer usage and/or TF binding sites, and compartmentalization.

Furthermore, *Sleeping Beauty*'s ability to deposit transposons so widely across the genome may have additional applications. One long-standing question regarding SRTs is whether there exist cryptic polyadenylation signals in intergenic regions that terminate RNA polymerase II transcription. Distributing *Sleeping Beauty* SRTs widely, amplifying self-reporting transcripts, and then sequencing full-length molecules (Gonzalez-Garay, 2016; Gupta et al.,

2018) would determine if, and to what extent, non-templated adenines are being added post-transcriptionally. Genomic sequence near the start of such tails could then be analyzed for cryptic polyadenylation signals. This approach could also be used to screen for cryptic splicing signals. Finally, full-length reads from single cell *Sleeping Beauty* libraries may yield sufficient coverage to identify structural variants, offering another option to connect genotype to transcriptome.

6.3 Industrial applications for SRTs

While we developed SRTs in an academic setting, it is a useful exercise to consider whether SRTs have any applications in industry. One can view SRTs as a set of positional, heritable barcodes. Collectively, they should uniquely identify individual clones. One biological use for this is in lineage tracing, which is discussed in detail further below. A related application may be as a way of confirming the provenance of biological samples.

In recent years, synthetic yeast strains have been created capable of synthesizing pharmaceuticals and other commercially important chemicals at scale (König et al., 2015; Liu et al., 2019; Walker and Pretorius, 2018). As the central reagent here is a living, replicating organism, there are concerns about how to protect it as intellectual property (König et al., 2015). With the cost of sequencing ever decreasing accompanied by efforts to expand the scope of DNA synthesis (Boeke et al., 2016), there may come a time when copying and printing DNA on demand is trivially easy, enabling the theft and counterfeiting of genomic commodities. A small number of SRTs distributed across a genome may offer some protection as these elements could serve as a randomly generated checksum. The probability that two yeast strains independently tagged with SRTs sharing identical insertion coordinates is virtually zero. This would make it easier for manufacturers to prove the authenticity of their creations. Moreover, copying and

synthesizing SRTs may be prohibitively expensive since they naturally contain repetitive sequences. Clinical samples could similarly be tagged with SRTs to generate sample-specific ensemble barcodes. Instead of sequencing genomic variants, which can be used to uniquely identify individuals (Erlich et al., 2018; Harmanci and Gerstein, 2018), the locations of SRTs would authenticate patient-derived DNA. Without context, these data are simply random numbers. Thus, SRTs may add a layer of additional security around potentially sensitive information.

6.4 Tagmentation-free SRTs

One of the bottlenecks of preparing bulk RNA calling cards libraries is the tagmentation of self-reporting transcripts. Commercial Tn5 enzyme is expensive and while there are protocols for in-house purification (Picelli et al., 2014), it is a non-trivial process. Generating cost efficient, high-quality libraries that precisely map the transposon-genome junction would make calling cards even more approachable to new users and allow us to increase the scale of our experiments. Here are three proposals for tagmentation-free alternatives for mapping SRTs.

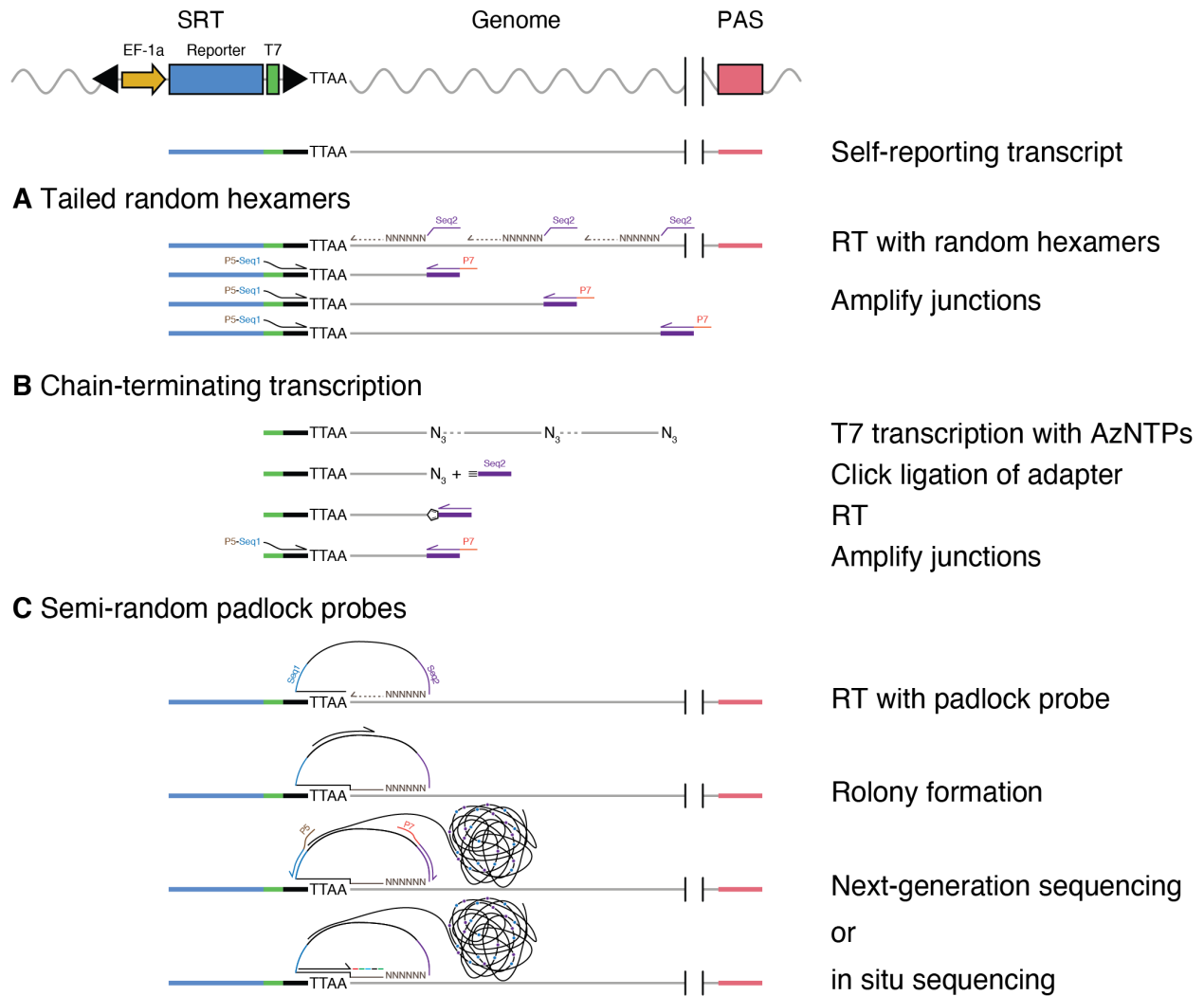


Figure 6.3: Alternative SRT library strategies without tagmentation. (A) Reverse transcription with tailed random hexamers. (B) Random chain termination of *in vitro* transcription. (C) Next-generation or *in situ* sequencing with semi-random padlock probes.

The first strategy involves reverse transcription (RT) with random hexamers instead of a polythymidine primer (Figure 6.3A). The hexamers would be tailed with a constant sequence that forms a universal priming site (Mäki and Tirola, 2018). The other primer would bind near the end of the terminal repeat to minimize the risk of PCR template switching (Kanagawa, 2003). Since libraries are random, a smear of product sizes should be seen on a gel. Shorter products can be enriched by increasing the concentration of hexamers (2006). Advantages of this approach are cost, ease, and detection of the entire transcriptome if bulk RNA-seq is

simultaneously needed. The latter is also a major drawback as SRT-specific amplification would be occurring against a background of all RNA. This could necessitate more amplification cycles, increasing the risk of PCR artifacts.

Another idea is to transcribe *in vitro* from an internal T7 RNAP promoter but use chain termination to create a random library (Figure 6.3B). This approach is inspired by bio-orthogonal click chemistry approaches for generating RNA-seq libraries (Routh et al., 2015, 2017). A small amount of rNTPs containing a 3' azide group (AzNTPs) would be mixed into the reaction. The elongating T7 polymerase would incorporate one of these at random and be unable to extend further. A second oligonucleotide, encoding a universal priming sequence and capped with a 5' alkyne moiety, would then be introduced. The click reaction would selectively and efficiently ligate the azide group to the alkyne, resulting in a triazole linkage. RT proceeds from the universal priming site upstream to the SRT followed by a junction-specific PCR. Advantages of this approach include the basal amplification of SRTs by T7 and the robustness of the click reaction. One downside is that the concentration of AzNTPs would have to be carefully titrated to get libraries that are long enough to capture the junction but short enough to sequence. Polymerase read-through of the triazole linkage is not particularly efficient (Routh et al., 2017), but this could be addressed by designing the RT primer to extend past the linkage with a small number of degenerate bases, acting as a splint oligonucleotide (Datlinger et al., 2019).

A third approach, aimed at integrating single cell calling cards with *in situ* sequencing techniques (Alon et al., 2020; Fürth et al., 2019; Lee et al., 2015), is to generate rolling circle colonies (rolonies) that linearly amplify the transposon-genome junction (Figure 6.3C). *In situ* sequencing typically uses highly specific padlock probes to achieve this (Chen et al., 2017b), which rely on two arms that hybridize to known sequences with or without a gap between them.

These ends are then ligated together and amplified with phi29. The problem for SRTs is that only one end, the transposon terminal repeat, has a fixed sequence, while the other is unknown. The solution we propose is to create a semi-random padlock probe: one end anneals near the end of the SRT while the other ends in a random hexamer. RT would start at the hexamer, proceed toward the 5' end of the probe which is phosphorylated and ligate it closed into a loop of single stranded DNA (ssDNA). To ensure that the hexamer binds at some minimum distance from the end of the transposon, dSpacer (abasic) residues could be added to the 5' end, then removed with Endonuclease VIII or APE1. The rolon can be used directly for *in situ* sequencing, enabling simultaneous readout of the transcriptome, SRTs, and spatial position of single cells; or, alternatively, the probe could be used to generate next-generation sequencing libraries from universal priming sites. To our knowledge, such a padlock design has not been reported, making it a riskier way to proceed.

6.5 Expanding the palette of self-reporting transposons

In this work, we developed self-reporting *piggyBac* and *Sleeping Beauty* transposons. That we were able to readily create these modified transposons suggests that the self-reporting paradigm may generalize to several transposases systems, provided that the terminal sequences do not contain a polyadenylation signal. We discuss several specific options below.

6.5.1 Mos1

Mos1 is a widely-used transposase for invertebrate transgenesis, particularly in *Caenorhabditis elegans* (Frøkjær-Jensen et al., 2008). Like *Sleeping Beauty*, it transposes into TA dinucleotides. Unlike *Sleeping Beauty*, however, Mos1 tolerates peptide fusions and can be redirected (Maragathavally et al., 2006). Rational design of Mos1 mutants identified a variant that is up to 800 times more active than wild-type Mos1 (Germon et al., 2009). Sequence changes to the

canonical Mos1 transposon can also lead to more efficient transposition (Jaillet et al., 2012): replacing the 5' inverted terminal repeat (ITR) with another copy of the 3' ITR increased transposition rates 20-fold. However, Mos1 shows markedly reduced activity in mammalian cells (Germon et al., 2009), though there are conflicting reports (Trubitsyna et al., 2017). One culprit could be post-translational modifications. Specifically, phosphorylation of S170 alters subcellular localization and drastically decreases transposition rate (Bouchet et al., 2014). The S170A mutant showed some activity but the pseudophosphorylated S170D mutant did not. It is not clear how the S170A mutant would behave in mammalian cells. Milder substitutions like S170V or S170C that better preserve the size of the residue and prevent phosphorylation could also be effective, especially in conjunction with the rationally designed hyperactive version (Germon et al., 2009).

6.5.2 Tol2

Tol2 is a popular transposase for zebrafish transgenesis (Ni et al., 2008). The transposase itself would likely not tolerate fusions as even small peptide tags abolish activity (Meir et al., 2011). However, undirected Tol2 appears to preferentially insert into regions enriched for the epigenetic marks H3K4me3 and H3K27me3 (Yoshida et al., 2017). These loci are thought to mark bivalent domains termed “poised” enhancers that are being primed for future activation (Bernstein et al., 2006). Single cell calling cards with Tol2 SRTs could identify poised loci during development or other dynamic processes. Such an analysis would likely have to be performed simultaneously with *piggyBac* single cell calling cards as it also has an affinity for H3K4me3 but not for H3K27me3 (Yoshida et al., 2017). Therefore, Tol2 peaks that do not overlap *piggyBac* peaks would most likely represent bivalent loci. Although we started developing Tol2 SRTs, molecular refinements to the protocol are necessary (Nicolas Ledru, personal communication).

6.5.3 LINE-1

Until now, we have focused on cut-and-paste transposases. However, copy-paste transposases can be uniquely powerful for certain applications. The best-known example is likely the LINE-1 (L1) retrotransposon (Garcia-Perez et al., 2015; Han and Boeke, 2005; Paço et al., 2014; Singer et al., 2010), which has been engineered for increased activity (An et al., 2006; Han and Boeke, 2004). The copy-paste activity of LINE-1 allows it to monotonically increase in number over time. This can be valuable for TF binding studies, particularly those at steady-state equilibrium, as it means donor elements are no longer limiting in number. This property may also make it easier to work in difficult-to-transfect systems: in theory, as long as few copies are introduced, they can be propagated indefinitely.

One of the challenges facing self-reporting L1 transposons is that, upon reintegration into the genome, the transposon truncates toward the 5' end, often in the middle of the ORF2 gene (An et al., 2006; Sultana et al., 2019). This breakpoint would indicate the junction between the transposon and the genome (SRTs could not point in the opposite direction due to the presence of a polyadenylation signal). Unfortunately, the lack of a constant sequence for priming the SRT PCR makes it difficult to enrich for insertions. One potential solution is inspired by our single cell calling cards protocol for *piggyBac* (Figure 6.4).

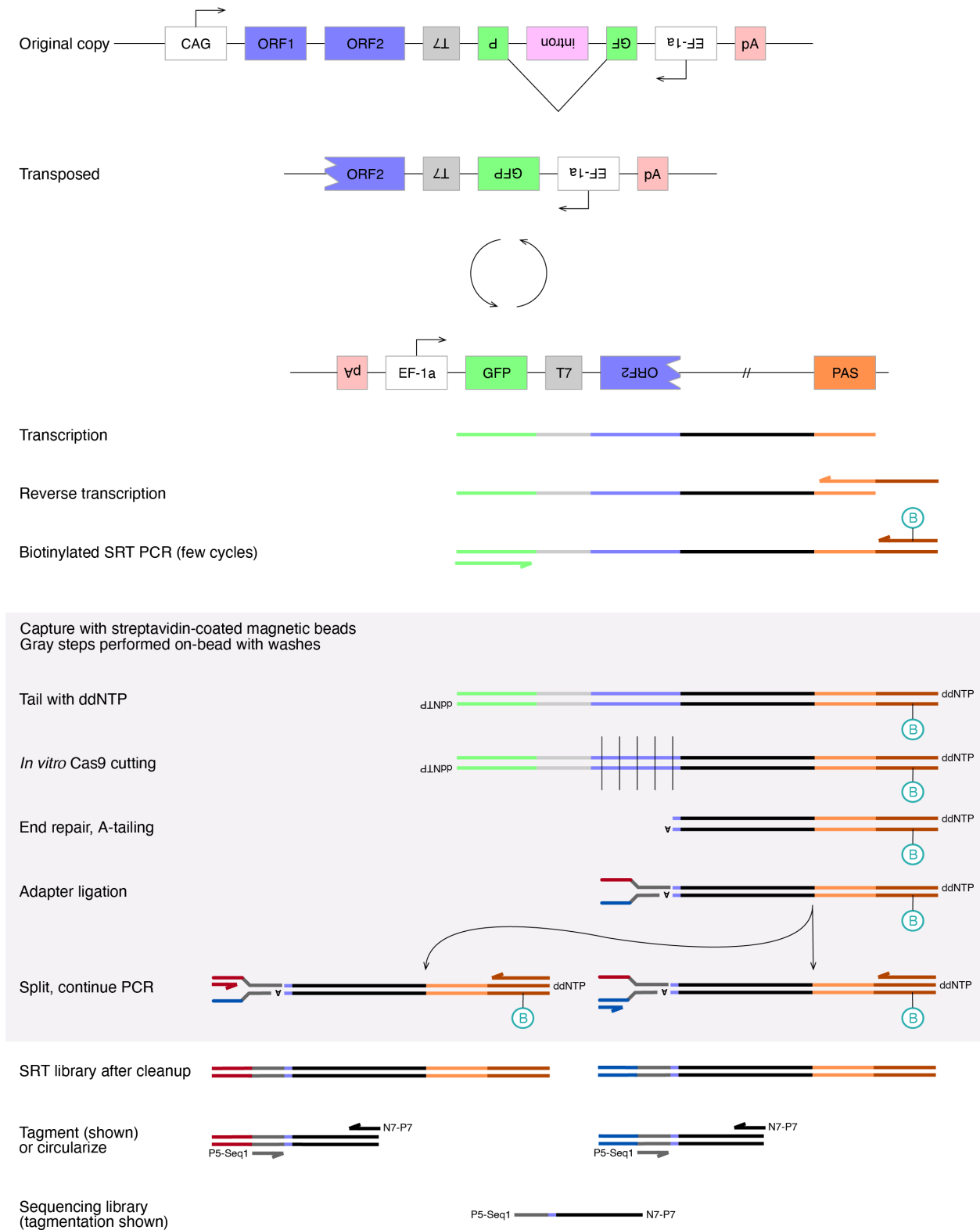


Figure 6.4: Proposed workflow for mapping L1 SRTs. pA: polyadenylation signal; PAS: polyadenine stretch.

The parental L1 SRT contains a reporter gene split by an artificial intron in the opposite orientation. Transcription of L1 removes the intron in derived copies; upon integration, the reporter gene can be properly transcribed and translated. The reporter's transcripts would contain the truncated ORF2 gene as part of its 3' UTR before encountering either a cryptic polyadenylation signal or a genomic polyadenine stretch (PAS). RT would proceed as per our bulk RNA calling card protocol (Appendix 2) followed by a few cycles of SRT amplification. One difference is that we would use a biotinylated primer at one end. The resulting products would then be captured on streptavidin-coated beads where they would be tailed with a dideoxynucleotide triphosphate (ddNTP). Since all sequence upstream of the breakpoint is unnecessary, we could direct Cas9 *in vitro* to cut these with a set of tiled gRNAs (Gu et al., 2016). This would eliminate most ORF2 sequence but preserve a small portion next to the breakpoint for quality control. We then would ligate a splinkerette adapter containing a next-generation sequencing primer onto the molecule. Due to the ddNTP, the adapter is added at the 5' end, ensuring we only sequence from ORF2 into the genome. A series of nested PCRs finalizes the library. Although involved, putting this protocol into practice would maximize the information yield of L1 SRT libraries.

6.5.4 *Helraiser*

Helraiser is a new transposase and the only known functional member of the *Helitron* family (Grabundzija et al., 2016, 2018). *Helitrons* are thought to replicate in a copy-paste fashion but, unlike retrotransposons, they do not proceed through an RNA intermediate (Feschotte and Wessler, 2001; Kapitonov and Jurka, 2001, 2007; Pritham and Thomas, 2015). Instead, the transposase nicks one strand of transposon DNA at one of the terminal sequences and use a rolling circle strategy to peel off a single strand of transposon DNA (Grabundzija et al., 2018).

This circular ssDNA intermediate can either transpose into genomic DNA, inserting into AT dinucleotides, or can be used to generate a complementary strand, creating a double-stranded transposon circle (Grabundzija et al., 2018; Kosek et al., 2021). Repeated cycles of peeling and pasting into the genome increase copy number over time.

One application for which *Helraiser* may be uniquely suited is lineage tracing. Since SRTs are vertically transmitted, their genomic coordinates can be seen as a heritable barcode of clonal origin. The cumulative set of self-reporting *Helitrons* in a cell should, in theory, completely describe its lineage. With cut-and-paste transposases, the risk is that SRTs may move around during the recording window, degrading clonal information (Figure 6.5A). We simulated lineage reconstruction with copy-paste SRTs, cut-and-paste SRTs, and binary CRISPR-based recorders (McKenna et al., 2016). We found that copy-paste SRTs were the most accurate, especially as the per-generation mutation rate increased (Figure 6.5B). Cut-and-paste SRTs, assuming a fixed population without re-transfection, came second and were slightly better than CRISPR recorders. Thus, copy-paste SRTs can be an accurate way to infer lineages and may be worth pursuing, particularly for whole organism cellular phylogenies. One advantage that *Helraiser* may have over LINE-1 is that, since the former was derived from the bat genome (Grabundzija et al., 2016), it may face fewer endogenous genome defense mechanisms in human or mouse cells (Ariumi, 2016).

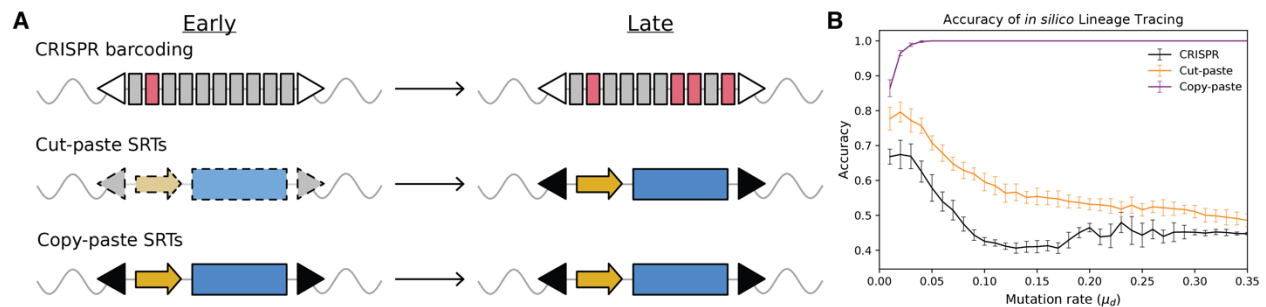


Figure 6.5: Lineage tracing with self-reporting transposons. (A) Schematic of binary CRISPR recorders, cut-and-paste SRTs, and copy-paste SRTs during lineage recording. (B) Accuracy of computational simulations modeling mutation of CRISPR recorders, cut-and-paste SRTs, and copy-paste SRTs. X-axis is the per-generation mutation rate. Accuracy, as measured by the Robinson-Foulds distance (Salvador-Martínez et al., 2019), is on the y-axis.

Although we have created a self-reporting *Helitron*, there is some concern about transposition rate. For lineage tracing, ideally there would be exactly one new insertion per cell cycle. However, in our hands, transposition is much slower. Our calling cards experiments in cell culture typically last only a few days, while other groups characterized *Helraiser* activity over a month-long time course (Grabundzija et al., 2016, 2018). Perhaps *Helitron* transposition starts slow but ramps up over time. Another open question is what components regenerate the *Helitron* strand complementary to the single stranded circular intermediate. If that is a rate-limiting step, finding ways to accelerate its synthesis may lead to faster transposition and, consequently, better resolved lineages.

6.5.5 Final thoughts

There are a number of other transposases active in vertebrate systems, such as *Minos*, *Passport*, and *Himar1* (Ni et al., 2008). The latter has recently been shown to tolerate fusions and can be strongly redirected in bacteria but additional validation is needed in mammalian cells. (Chen and Wang, 2019). Tn7 is a bacterial transposase with exquisite site specificity with reduced, but incompletely characterized, activity in human cells (Bainton, 1993; Kuduvalli, 2005). Tn5, the bacterial transposase used in the assay for transposase-accessible chromatin (ATAC-seq) (Buenrostro et al., 2013) has very low activity in live mammalian cells but synthetic dimers can

increase transposition efficiency (Blundell-Hunter et al., 2018). Finally, it is possible that the set of transposases we know of represents a fraction of the diversity present in nature. A shotgun metagenomics approach (Quince et al., 2017) to discover new enzymes may uncover even more redirectable, versatile, or otherwise intriguing transposases to explore.

6.6 The ecology of chromatin

The “ecology of the genome” was coined over twenty years ago, and later expounded upon, to describe how endogenous transposable elements live in, and interact with, the genome in ways akin to how plants and animals engage with ecosystems (Brookfield, 2005; Kidwell and Lisch, 1997). I argue there also exists an ecology of chromatin, encompassing the multifaceted and heterogeneous interactions between proteins and DNA. Euchromatin and heterochromatin, for example, are populated by different sets of proteins, are rooted by different post-translational epigenetic marks, and support different regulatory activities. Here, we employed transposases, tethered and free-range, to traverse this landscape. This work suggests the need to reevaluate what “accessibility” means for a transposase. Individual enzymes clearly show varying affinities for genomic regions at scales larger than traditionally “accessible” loci. To add, both *piggyBac* and *Sleeping Beauty* demonstrate how careful analyses of undirected transpositions can reveal distinct aspects of genomic regulation. As researchers embrace these tools, and as the tools become increasingly refined, it will be exciting to see what genomic wonders transposases, these navigators of the nucleus, discover.

6.7 References

Alon, S., Goodwin, D.R., Sinha, A., Wassie, A.T., Chen, F., Daugharthy, E.R., Bando, Y., Kajita, A., Xue, A.G., Marrett, K., et al. (2021). Expansion sequencing: Spatially precise in situ transcriptomics in intact biological systems. *Science* 371, eaax2656.

An, W., Han, J.S., Wheelan, S.J., Davis, E.S., Coombes, C.E., Ye, P., Triplett, C., and Boeke, J.D. (2006). Active retrotransposition by a synthetic L1 element in mice. *Proceedings of the National Academy of Sciences* 103, 18662–18667.

Ariumi, Y. (2016). Guardian of the Human Genome: Host Defense Mechanisms against LINE-1 Retrotransposition. *Frontiers in Chemistry* 4, 761.

Bainton, R. (1993). Tn7 transposition: Target DNA recognition is mediated by multiple Tn7-encoded proteins in a purified in vitro system. *Cell* 72, 931–943.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell* 125, 315–326.

Blundell-Hunter, G., Tellier, M., and Chalmers, R. (2018). Transposase subunit architecture and its relationship to genome size and the rate of transposition in prokaryotes and eukaryotes. *Nucleic Acids Research* 46, 9637–9646.

Boeke, J.D., Church, G., Hessel, A., Kelley, N.J., Arkin, A., Cai, Y., Carlson, R., Chakravarti, A., Cornish, V.W., Holt, L., et al. (2016). The Genome Project-Write. *Science* 353, 126–127.

Bouallègue, M., Rouault, J.-D., Hua-Van, A., Makni, M., and Capy, P. (2017). Molecular evolution of piggyBac superfamily: from selfishness to domestication. *Genome Biol Evol* evw292.

Bouchet, N., Jaillet, J., Gabant, G., Brillet, B., Briseno-Roa, L., Cadene, M., and Auge-Gouillou, C. (2014). cAMP protein kinase phosphorylates the Mos1 transposase and regulates its activity:

evidences from mass spectrometry and biochemical analyses. *Nucleic Acids Research* 42, 1117–1128.

Brookfield, J.F.Y. (2005). The ecology of the genome — mobile DNA elements and their hosts. *Nat Rev Genet* 6, 128–136.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* 10, 1213–1218.

Bushman, F.D. (2007). Retroviral integration and human gene therapy. *J. Clin. Invest.* 117, 2083–2086.

Cammack, A.J., Moudgil, A., Chen, J., Vasek, M.J., Shabsovich, M., McCullough, K., Yen, A., Lagunas, T., Maloney, S.E., He, J., et al. (2020). A viral toolkit for recording transcription factor–DNA interactions in live mouse tissues. *Proc Natl Acad Sci USA* 117, 10003–10014.

Chen, S.P., and Wang, H.H. (2019). An Engineered Cas-Transposon System for Programmable and Site-Directed DNA Transpositions. *The CRISPR Journal* 2, 376–394.

Chen, C., Xing, D., Tan, L., Li, H., Zhou, G., Huang, L., and Xie, X.S. (2017). Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). *Science* 356, 189–194.

Chen, Q., Luo, W., Veach, R.A., Hickman, A.B., Wilson, M.H., and Dyda, F. (2020). Structural basis of seamless excision and specific targeting by piggyBac transposase. *Nat Commun* 11, 3446.

Chen, X., Sun, Y.-C., Church, G.M., Lee, J.H., and Zador, A.M. (2018). Efficient in situ barcode sequencing using padlock probe-based BaristaSeq. *Nucleic Acids Research* 46, e22–e22.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., et al. (2013). Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* 339, 819–823.

Datlinger, P., Rendeiro, A.F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L.C., Kuchler, A., Alpar, D., and Bock, C. (2017). Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods* 14, 297–301.

Datlinger, P., Rendeiro, A.F., Boenke, T., Senekowitsch, M., Krausgruber, T., Barreca, D., and Bock, C. (2021). Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. *Nat Methods* 18, 635–642.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* 167, 1853-1866.e17.

El Ashkar, S., De Rijck, J., Demeulemeester, J., Vets, S., Madlala, P., Cermakova, K., Debyser, Z., and Gijsbers, R. (2014). BET-independent MLV-based Vectors Target Away From Promoters and Regulatory Elements. *Molecular Therapy - Nucleic Acids* 3, e179.

El Ashkar, S., Van Looveren, D., Schenk, F., Vranckx, L.S., Demeulemeester, J., De Rijck, J., Debyser, Z., Modlich, U., and Gijsbers, R. (2017). Engineering Next-Generation BET-Independent MLV Vectors for Safer Gene Therapy. *Molecular Therapy - Nucleic Acids* 7, 231–245.

Erlich, Y., Shor, T., Pe'er, I., and Carmi, S. (2018). Identity inference of genomic data using long-range familial searches. *Science* 362, 690–694.

Feschotte, C., and Wessler, S.R. (2001). Treasures in the attic: rolling circle transposons discovered in eukaryotic genomes. *Proceedings of the National Academy of Sciences* 98, 8923–8924.

Frøkjær-Jensen, C., Wayne Davis, M., Hopkins, C.E., Newman, B.J., Thummel, J.M., Olesen, S.-P., Grunnet, M., and Jorgensen, E.M. (2008). Single-copy insertion of transgenes in *Caenorhabditis elegans*. *Nat Genet* 40, 1375–1383.

Fürth, D., Hatini, V., and Lee, J.H. (2019). In Situ Transcriptome Accessibility Sequencing (INSTA-seq) (bioRxiv).

Garcia-Perez, J.L., Doucet, A.J., Kopera, H.C., Richardson, S.R., Moldovan, J.B., and Moran, J.V. (2015). The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. *Microbiology Spectrum* 3.

Germon, S., Bouchet, N., Casteret, S., Carpentier, G., Adet, J., Bigot, Y., and Augé-Gouillou, C. (2009). Mariner Mos1 transposase optimization by rational mutagenesis. *Genetica* 137, 265–276.

Gogol-Döring, A., Ammar, I., Gupta, S., Bunse, M., Miskey, C., Chen, W., Uckert, W., Schulz, T.F., Izsvák, Z., and Ivics, Z. (2016). Genome-wide Profiling Reveals Remarkable Parallels Between Insertion Site Selection Properties of the MLV Retrovirus and the piggyBac Transposon in Primary Human CD4⁺ T Cells. *Molecular Therapy* 24, 592–606.

Gonzalez-Garay, M.L. (2016). Introduction to Isoform Sequencing Using Pacific Biosciences Technology (Iso-Seq). In *Transcriptomics and Gene Regulation*, J. Wu, ed. (Dordrecht: Springer Netherlands), pp. 141–160.

Grabundzija, I., Messing, S.A., Thomas, J., Cosby, R.L., Bilic, I., Miskey, C., Gogol-Döring, A., Kapitonov, V., Diem, T., Dalda, A., et al. (2016). A Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. *Nature Communications* 7, 10716.

Grabundzija, I., Hickman, A.B., and Dyda, F. (2018). Helraiser intermediates provide insight into the mechanism of eukaryotic replicative transposition. *Nature Communications* 9, 1278.

Gu, W., Crawford, E.D., O'Donovan, B.D., Wilson, M.R., Chow, E.D., Retallack, H., and DeRisi, J.L. (2016). Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. *Genome Biology* 17, 2408.

Gupta, I., Collier, P.G., Haase, B., Mahfouz, A., Joglekar, A., Floyd, T., Koopmans, F., Barres, B., Smit, A.B., Sloan, S.A., et al. (2018). Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat Biotechnol* 36, 1197–1202.

Hacein-Bey-Abina, S. (2003). LMO2-Associated Clonal T Cell Proliferation in Two Patients after Gene Therapy for SCID-X1. *Science* 302, 415–419.

Hacein-Bey-Abina, S., von Kalle, C., Schmidt, M., Le Deist, F., Wulffraat, N., McIntyre, E., Radford, I., Villeval, J.-L., Fraser, C.C., Cavazzana-Calvo, M., et al. (2003). A Serious Adverse

Event after Successful Gene Therapy for X-Linked Severe Combined Immunodeficiency. *N Engl J Med* 348, 255–256.

Han, J.S., and Boeke, J.D. (2004). A highly active synthetic mammalian retrotransposon. *Nature* 429, 314–318.

Han, J.S., and Boeke, J.D. (2005). LINE-1 retrotransposons: Modulators of quantity and quality of mammalian gene expression? *BioEssays* 27, 775–784.

Harmanci, A., and Gerstein, M. (2018). Analysis of sensitive information leakage in functional genomics signal profiles through genomic deletions. *Nat Commun* 9, 2453.

Jaillet, J., Genty, M., Cambefort, J., Rouault, J.-D., and Augé-Gouillou, C. (2012). Regulation of Mariner Transposition: The Peculiar Case of Mos1. *PLoS ONE* 7, e43365.

Kanagawa, T. (2003). Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of Bioscience and Bioengineering* 96, 317–323.

Kapitonov, V.V., and Jurka, J. (2001). Rolling-circle transposons in eukaryotes. *Proceedings of the National Academy of Sciences* 98, 8714–8719.

Kapitonov, V.V., and Jurka, J. (2007). Helitrons on a roll: eukaryotic rolling-circle transposons. *Molecular Cell* 23, 521–529.

Kidwell, M.G., and Lisch, D. (1997). Transposable elements as sources of variation in animals and plants. *Proceedings of the National Academy of Sciences* 94, 7704–7711.

König, H., Dorado-Morales, P., and Porcar, M. (2015). Responsibility and intellectual property in synthetic biology: A proposal for using Responsible Research and Innovation as a basic framework for intellectual property decisions in synthetic biology. *EMBO Rep* 16, 1055–1059.

Kosek, D., Grabundzija, I., Lei, H., Bilic, I., Wang, H., Jin, Y., Peaslee, G.F., Hickman, A.B., and Dyda, F. (2021). The large bat Helitron DNA transposase forms a compact monomeric assembly that buries and protects its covalently bound 5'-transposon end. *Molecular Cell* 81, 4271-4286.e4.

Kuduvalli, P.N. (2005). Site-specific Tn7 transposition into the human genome. *Nucleic Acids Research* 33, 857–863.

Lalli, M., Yen, A., Thopte, U., Dong, F., Moudgil, A., Chen, X., Milbrandt, J., Dougherty, J.D., and Mitra, R.D. (2021). Measuring Transcription Factor Binding and Gene Expression using Barcoded Self-Reporting Transposon Calling Cards and Transcriptomes (Genomics).

Lee, J.H., Daugharthy, E.R., Scheiman, J., Kalhor, R., Ferrante, T.C., Terry, R., Turczyk, B.M., Yang, J.L., Lee, H.S., Aach, J., et al. (2015). Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat Protoc* 10, 442–458.

Liu, J., Shively, C.A., and Mitra, R.D. (2020). Quantitative analysis of transcription factor binding and expression using calling cards reporter arrays. *Nucleic Acids Research* 48, e50–e50.

Liu, Z., Zhang, Y., and Nielsen, J. (2019). Synthetic Biology of Yeast. *Biochemistry* 58, 1511–1520.

- Mäki, A., and Tirola, M. (2018). Directional high-throughput sequencing of RNAs without gene-specific primers. *BioTechniques* 65, 219–223.
- Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. (2013). RNA-guided human genome engineering via Cas9. *Science* 339, 823–826.
- Maragathavally, K.J., Kaminski, J.M., Coates, C.J., Maragathavally, K.J., Kaminski, J.M., and Coates, C.J. (2006). Chimeric Mos1 and piggyBac transposases result in site-directed integration. *FASEB j.* 20, 1880–1882.
- Mayhew, D., and Mitra, R.D. (2016). Transposon Calling Cards. *Cold Spring Harbor Protocols* 2016, pdb.top077776.
- McKenna, A., Findlay, G.M., Gagnon, J.A., Horwitz, M.S., Schier, A.F., and Shendure, J. (2016). Whole organism lineage tracing by combinatorial and cumulative genome editing. *Science* 353, aaf7907–aaf7907.
- Medina-Rivera, A., Santiago-Algarra, D., Puthier, D., and Spicuglia, S. (2018). Widespread Enhancer Activity from Core Promoters. *Trends in Biochemical Sciences* 43, 452–468.
- Meir, Y.-J.J., Weirauch, M.T., Yang, H.-S., Chung, P.-C., Yu, R.K., and Wu, S.C.-Y. (2011). Genome-wide target profiling of piggyBac and Tol2 in HEK 293: pros and cons for gene discovery and gene therapy. *BMC Biotechnol* 11, 28.
- Morellet, N., Li, X., Wieninger, S.A., Taylor, J.L., Bischerour, J., Moriau, S., Lescop, E., Bardiaux, B., Mathy, N., Assrir, N., et al. (2018). Sequence-specific DNA binding activity of the

cross-brace zinc finger motif of the piggyBac transposase. *Nucleic Acids Research* 46, 2660–2677.

Moudgil, A., Wilkinson, M.N., Chen, X., He, J., Cammack, A.J., Vasek, M.J., Lagunas, T., Qi, Z., Lalli, M.A., Guo, C., et al. (2020). Self-Reporting Transposons Enable Simultaneous Readout of Gene Expression and Transcription Factor Binding in Single Cells. *Cell* 182, 992-1008.e21.

Nair, A.R., Jinger, X., and Hermiston, T.W. (2011). Effect of different UCOE-promoter combinations in creation of engineered cell lines for the production of Factor VIII. *BMC Research Notes* 4, 178.

Ni, J., Clark, K.J., Fahrenkrug, S.C., and Ekker, S.C. (2008). Transposon tools hopping in vertebrates. *Briefings in Functional Genomics and Proteomics* 7, 444–453.

Paço, A., Adegá, F., and Chaves, R. (2014). LINE-1 retrotransposons: from ‘parasite’ sequences to functional elements. *Journal of Applied Genetics* 56, 133–145.

Picelli, S., Björklund, A., Aasa K., Reinius, B., Sagasser, S., Winberg, G., and Sandberg, R. (2014). Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Research* 24, 2033–2040.

Pritham, E.J., and Thomas, J. (2015). Helitrons, the Eukaryotic Rolling-circle Transposable Elements. *Microbiology Spectrum* 3.

Qin, J.Y., Zhang, L., Clift, K.L., Hular, I., Xiang, A.P., Ren, B.-Z., and Lahn, B.T. (2010). Systematic Comparison of Constitutive Promoters and the Doxycycline-Inducible Promoter. *PLoS ONE* 5, e10611.

- Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* 35, 833–844.
- Routh, A., Head, S.R., Ordoukhanian, P., and Johnson, J.E. (2015). ClickSeq: Fragmentation-Free Next-Generation Sequencing via Click Ligation of Adaptors to Stochastically Terminated 3'-Azido cDNAs. *Journal of Molecular Biology* 427, 2610–2616.
- Routh, A., Ji, P., Jaworski, E., Xia, Z., Li, W., and Wagner, E.J. (2017). Poly(A)-ClickSeq: click-chemistry for next-generation 3'-end sequencing without RNA enrichment or fragmentation. *Nucleic Acids Research* 45, e112–e112.
- Sabari, B.R., Dall'Agnesse, A., Boija, A., Klein, I.A., Coffey, E.L., Shrinivas, K., Abraham, B.J., Hannett, N.M., Zamudio, A.V., Manteiga, J.C., et al. (2018). Coactivator condensation at super-enhancers links phase separation and gene control. *Science* 361, eaar3958.
- Salvador-Martínez, I., Grillo, M., Averof, M., and Telford, M.J. (2019). Is it possible to reconstruct an accurate cell lineage using CRISPR recorders? *ELife* 8, e40292.
- Sarkar, A., Sim, C., Hong, Y.S., Hogan, J.R., Fraser, M.J., Robertson, H.M., and Collins, F.H. (2003). Molecular evolutionary analysis of the widespread piggyBac transposon family and related “domesticated” sequences. *Mol Genet Genomics* 270, 173–180.
- Saunders, F., Sweeney, B., Antoniou, M.N., Stephens, P., and Cain, K. (2015). Chromatin Function Modifying Elements in an Industrial Antibody Production Platform - Comparison of UCOE, MAR, STAR and cHS4 Elements. *PLoS ONE* 10, e0120096.

Shively, C.A., Liu, J., Chen, X., Loell, K., and Mitra, R.D. (2019). Homotypic cooperativity and collective binding are determinants of bHLH specificity and function. *Proc Natl Acad Sci USA* 116, 16143–16152.

Singer, T., McConnell, M.J., Marchetto, M.C.N., Coufal, N.G., and Gage, F.H. (2010). LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes? *Trends in Neurosciences* 33, 345–354.

Staller, M.V., Holehouse, A.S., Swain-Lenz, D., Das, R.K., Pappu, R.V., and Cohen, B.A. (2018). A High-Throughput Mutational Scan of an Intrinsically Disordered Acidic Transcriptional Activation Domain. *Cell Systems* 6, 444-455.e6.

Sultana, T., van Essen, D., Siol, O., Bailly-Bechet, M., Philippe, C., Zine El Aabidine, A., Pioger, L., Nigumann, P., Sacconi, S., Andrau, J.-C., et al. (2019). The Landscape of L1 Retrotransposons in the Human Genome Is Shaped by Pre-insertion Sequence Biases and Post-insertion Selection. *Molecular Cell* 74, 555-570.e7.

Tipanee, J., Chai, Y.C., VandenDriessche, T., and Chuah, M.K. (2017). Preclinical and clinical advances in transposon-based gene therapy. *Bioscience Reports* 37, BSR20160614.

Trubitsyna, M., Michlewski, G., Finnegan, D.J., Elfick, A., Rosser, S.J., Richardson, J.M., and French, C.E. (2017). Use of mariner transposases for one-step delivery and integration of DNA in prokaryotes and eukaryotes by transfection. *Nucleic Acids Research* 45, e89–e89.

Vargas, J.E., Chicaybam, L., Stein, R.T., Tanuri, A., Delgado-Cañedo, A., and Bonamino, M.H. (2016). Retroviral vectors and transposons for stable gene therapy: advances, current challenges and perspectives. *J Transl Med* 14, 288.

Vriend, L.E.M., Prakash, R., Chen, C.-C., Vanoli, F., Cavallo, F., Zhang, Y., Jasin, M., and Krawczyk, P.M. (2016). Distinct genetic control of homologous recombination repair of Cas9-induced double-strand breaks, nicks and paired nicks. *Nucleic Acids Research* 44, gkw179-5217.

Walker, R., and Pretorius, I. (2018). Applications of Yeast Synthetic Biology Geared towards the Production of Biopharmaceuticals. *Genes* 9, 340.

Yoshida, J., Akagi, K., Misawa, R., Kokubu, C., Takeda, J., and Horie, K. (2017). Chromatin states shape insertion profiles of the piggyBac, Tol2 and Sleeping Beauty transposons and murine leukemia virus. *Scientific Reports* 7, 43613.

Yusa, K., Zhou, L., Li, M.A., Bradley, A., and Craig, N.L. (2011). A hyperactive piggyBac transposase for mammalian applications. *Proceedings of the National Academy of Sciences of the United States of America* 108, 1531–1536.

Zhang, F., Thornhill, S.I., Howe, S.J., Ulaganathan, M., Schambach, A., Sinclair, J., Kinnon, C., Gaspar, H.B., Antoniou, M., and Thrasher, A.J. (2007). Lentiviral vectors containing an enhancer-less ubiquitously acting chromatin opening element (UCOE) provide highly reproducible and stable transgene expression in hematopoietic cells. *Blood* 110, 1448–1457.

Zhang, F., Frost, A.R., Blundell, M.P., Bales, O., Antoniou, M.N., and Thrasher, A.J. (2010). A Ubiquitous Chromatin Opening Element (UCOE) Confers Resistance to DNA Methylation-mediated Silencing of Lentiviral Vectors. *Molecular Therapy* 18, 1640–1649.

(2006). Oligonucleotide primers for cDNA synthesis (1 mg/ml). *Cold Spring Harb Protoc* 2006, pdb.rec8886.

Appendix 1: Mammalian Calling Cards

Quick Start Guide

(A version of this appendix has been published on protocols.io:

<https://doi.org/10.17504/protocols.io.xurfnv6>)

A1.1 Abstract

Transposon calling cards can identify transcription factor (TF) binding sites. This involves fusing your favorite TF (YFTF) to the hyperactive *piggyBac* transposase (HyPBase). This is delivered to cells in conjunction with a *piggyBac* transposon. The TF will visit sites in the genome and YFTF-HyPBase will deposit transposons near binding sites. We then generate sequencing libraries to map the genome-wide localization of transposons. Finally, we identify significant clusters of insertions to identify TF binding sites.

A1.2 Guidelines

These experiments are intended to introduce you to the calling cards assay. We recommend following them to establish baseline confidence in your transcription factor constructs before proceeding to your favorite model system. Alternatively, if you just wish to use the directed *piggyBac* transposase, you can use this protocol to familiarize yourself with our molecular workflow.

A1.3 Materials

Name	Catalog #	Vendor
Lipofectamine 3000	L3000015	Thermo Fisher Scientific

- **pRM1258/pENTR_myc-hyPBase:** This Gateway entry vector contains the HyPBase gene. It is convenient to make the YFTF-HyPBase fusions in this vector because you can

easily port the fusion into a variety of other plasmids for AAV packaging, homologous recombination into the *Rosa26* locus, etc.

- **pRM1114/CMV_HyPBase:** This is our standard positive control transposase plasmid. For some experiments it may be more convenient to make your YFTF-HyPB fusion in this construct.
- **pRM1304/PB_SRT_Rz_Puro:** This is a *piggyBac* self-reporting transposon (SRT) encoding a puromycin resistance gene. Cells transfected with this plasmid and HyPBase survive puromycin selection.
- **pRM1535/SRT_tdTomato.** This is a *piggyBac* self-reporting transposon (SRT) donor encoding a tdTomato fluorescent reporter. Cells transfected with this plasmid and HyPBase can be sorted for based on high fluorescence signal.
- **pRM1294/BrokenHeart:** This plasmid is a reporter of *piggyBac* transposase activity. It encodes the DsRed fluorescent protein gene interrupted by a *piggyBac* transposon. When cells are co-transfected with BrokenHeart and *piggyBac*, the transposon is removed and the DsRed reading frame is restored. These cells fluoresce brightly, while cells transfected with BrokenHeart alone will not fluoresce.
- A non-fluorescent control, or "empty," plasmid. This can be something you use regularly in-house or a commercially available plasmid, such as NEB's pUC19 vector (#N3041S).
- Optional: a GFP expression plasmid as a transfection control. This can be something used routinely in your lab or a commercially available vector, such as Addgene #54767.
- HCT-116 cells. Stocks may be obtained from ATCC (#CCL-247) if necessary.

A1.4Steps

A.1.4.1 Cloning and Sequence Validation of YFTF-HyPBase Fusions

1. *Make C- and N- terminal fusions of your favorite transcription factor (YFTF) with HyPBase.*

It is important to include a linker sequence between these genes. We strongly recommend the following amino acid linker sequence: KLGGGAPAVGGGPKAADK. We have tested many and have found this sequence works best. It is often convenient to make these fusion constructs by using In-Fusion (Clontech/Takara) or Gibson (NEB) cloning to drop YFTF into pRM1258/pENTR_myc-hypPBase. We do not have good antibodies to the *piggyBac* transposase, so we recommend designing your construct so that the chimeric protein is tagged with *myc*.

2. *Validate the constructs.* Perform restriction digest analysis on the plasmid with at least 3 restriction enzymes to make sure there were no gross rearrangements. Next, Sanger sequence the full chimeric gene, or alternatively perform Illumina sequencing on the whole plasmid. It is important to do the restriction digest and EITHER of the sequencing strategies.

3. *(Only required for Gateway strategy).* Move the chimeric gene from the pENTR vector to an expression vector.

A1.4.2 Functional Validation of YFTF-HyPBase Fusions

4. After creating the YFTF-HyPBase and HyPBase-YFTF fusions, the next steps are to validate them. First we will assess whether the fusions retain *piggyBac* transposase activity. We recommend transforming HCT-116 cells with the YFTF fusions and the BrokenHeart transposon along with appropriate controls. BrokenHeart plasmid is available from Addgene (#86950). Empty plasmid can be any vector that does not have transposase or transposon sequence, e.g. pUC19, pBluescript, etc. We recommend using Lipofectamine 3000 (following manufacturer's

instructions) to deliver 1 µg total DNA to approximately 200,000 cells in each well of a 6-well plate. The following table summarizes each condition and the expected results.

Condition	Empty plasmid	pRM1294 BrokenHeart	pRM1114 CMV_HyPBBase	YFTF-HyPBBase	HyPBBase-YFTF	Red cells?	Notes
Negative control	0.5 µg	0.5 µg	NA	NA	NA	None	
Positive control	NA	0.5 µg	0.5 µg	NA	NA	Many	
YFTF-HyPBBase	NA	0.5 µg	NA	0.5 µg	NA	Some	Perform in duplicate
HyPBBase-YFTF	NA	0.5 µg	NA	NA	0.5 µg	Some	Perform in duplicate

Optional control #1 -- a "lipofection only" negative control. This is not a bad idea, particularly if you are new to lipofections or are testing a new cell line and are concerned about toxicity. These cells should show high viability and no fluorescence signal. If these cells are viable but cells transfected with DNA are not, it may indicate issues with plasmid isolation (e.g. endotoxin contamination).

Optional control #2 -- a GFP expression plasmid could be transfected in parallel to estimate overall transfection efficiencies.

5. The second validation will test whether the YFTF-fusions successfully redirect *piggyBac* insertions near YFTF binding sites. Since *piggyBac* inserts into TTAAAs, we would like to be able to distinguish unique insertions into the same TTAA. For this reason, we recommend 6 replicates per condition, at least for your two “test” samples, and the unfused *piggyBac* (1 6-well plate each, 3 total). In addition, we recommend running one well as a transposon-only negative control and one well as a mock lipofection negative control. Here we will use puromycin selection to obtain cells with transpositions. Once again, we will work with HCT-116 cells.

Condition	Empty plasmid	pRM1304 PB_SRT_Rz_Puro	pRM1114 CMV_HyPBBase	YFTF-HyPBBase	HyPBBase-YFTF	Alive cells?	Notes
-----------	---------------	------------------------	----------------------	---------------	---------------	--------------	-------

No transfection control	1 μ g	NA	NA	NA	NA	None	No colonies after selection
SRT only control	0.5 μ g	0.5 μ g	NA	NA	NA	None	No colonies after a few days
Positive control	NA	0.5 μ g	0.5 μ g	NA	NA	Many	Perform 6 replicates
YFTF-HyPBase	NA	0.5 μ g	NA	0.5 μ g	NA	Some	Perform 6 replicates
HyPBase-YFTF	NA	0.5 μ g	NA	NA	0.5 μ g	Some	Perform 6 replicates

We typically split each well 1:1 and transfer to a 10 cm dish after 24 hours. We add puromycin to a final concentration of 2 μ g/ml after cells have seeded, typically 6-8 hours after transferring. Media is replenished every two days. All replicates are cultured separately. Cells are harvested after the SRT-only control transfectants are dead.

A1.4.3 Calling Card Library Preparation

6. Calling card libraries can now be made from successfully selected cells. For first-time users, we recommend following the bulk calling card protocol for making libraries (Appendix 2).
7. (*Advanced*) Depending on your application, you may also be interested in making single cell calling card libraries (Appendix 3).

A1.4.4 Sequencing, Analyzing, and Visualizing Calling Card Data

8. If your library preparation has been successful, you are ready to sequence your calling card libraries. We have successfully sequenced libraries on the Illumina MiSeq, MiniSeq, HiSeq, and NextSeq platforms. Due to relatively low sequence complexity, we typically run our libraries with 50% PhiX genome spiked in.

A1.4.5 Next Steps

9. If you've made it this far and your data look great, congratulations! One or both of the YFTF fusions have worked and successfully enriched for insertions near YFTF binding sites. You may now wish to repeat the above experimental workflow with your model systems, or try different transgenesis techniques (e.g. electroporation, nucleofection, viral transduction). Otherwise, you are now ready to move into your model system and study YFTF binding. Please let us know if you have any difficulties and we will do our best to provide assistance.

Appendix 2: Bulk Calling Cards Library

Preparation

(A version of this appendix has been published on protocols.io:

<https://doi.org/10.17504/protocols.io.xwhfpb6>)

A2.1 Abstract

This protocol describes how to create calling card libraries from bulk RNA. This protocol assumes you have successfully transformed cells with *piggyBac* self-reporting transposons and either undirected *piggyBac* transposase or your favorite transcription factor (YTF) fused to *piggyBac*. Your cells are now ready for RNA extraction, SRT amplification, and library preparation.

A2.2 Guidelines

Please read this protocol in its entirety before starting. For several steps, it may help to pre-program your thermocycler with the listed settings.

Please read and familiarize yourself with the manuals for the QIAGEN RNEasy Plus Mini Kit and the Nextera XT Tagmentation Kit. The instructions are meant to summarize those workflows; however, when in doubt, please refer to the manufacturer's instructions for guidance.

Ensure that you have performed multiple (i.e. 8-12), independent replicates of your experiment before proceeding. The calling card assay relies on the clustering of multiple nearby insertions to identify TF binding sites. Some regions of the genome may have relatively few insertion sites for the transposase. Therefore, doing multiple independent replicates increases the statistical power to discriminate between a true binding site and background noise.

This protocol is meant to describe how we prepare calling card libraries. While it is possible that another kit or component could equally suffice, we have not tested any substitutions and do not officially support deviations from this protocol. This document enumerates what we have had success with and is a starting point from which we can best help troubleshoot.

A2.3 Materials

Name	Catalog #	Vendor
Agencourt Ampure XP	A63880	Beckman Coulter
dNTP	639125	Takara
2x Kapa HiFi Hotstart Readymix	KK2602	Kapa Biosystems
Maxima H Minus Reverse Transcriptase (200 U/uL)	EP0752	Thermo Fisher Scientific
Qubit dsDNA HS Assay Kit	Q32851	Thermo Fisher Scientific
Capillary electrophoresis instrument (e.g. Agilent TapeStation 4200)		
RNEasy Plus Mini Kit	74134	Qiagen
2-mercaptoethanol	21985023	Gibco - Thermo Fisher
Qubit RNA HS Assay Kit	Q32852	Thermo Fisher Scientific
RNaseOUT™ Recombinant Ribonuclease Inhibitor	10777019	Thermo Fisher Scientific
RNase H	M0297S	New England Biolabs
Nextera XT DNA Library Preparation Kit	FC-131-1024	Illumina, Inc.
High Sensitivity D1000 Reagents	5067-5585	Agilent Technologies
High Sensitivity D1000 ScreenTape	5067-5584	Agilent Technologies

Primers

>SMART_dT18VN

AAGCAGTGGTATCAACGCAGAGTACGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT

VN

>SRT_PAC_F1

CAACCTCCCCTTCTACGAGC

>SRT_tdTomato_F1

TCCTGTACGGCATGGACGAG

>SMART

AAGCAGTGGTATCAACGCAGAGT

>Raff_ACTB_F

CCTCGCCTTTGCCGATCCG

>Raff_ACTB_R

GGATCTTCATGAGGTAGTCAGTCAGGTCC

Barcoded *piggyBac* primers, for example:

>OM-PB-ACG (barcode sequence is underlined)

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGA
TCTACGTTTACGCAGACTATCTTTCTAG

You should have multiple barcoded primers. See Table 2.4 for more examples.

Indexed Nextera N7 primers, for example:

>Nextera_N701 (index sequence is underlined)

CAAGCAGAAGACGGCATAACGAGATTCGCCTTAGTCTCGTGGGCTCGG

You should have multiple indexed primers. These can be either official Nextera indexes or custom, lab-specific indexes. For a comprehensive list of official Nextera indexes, consult the Illumina Adapter Sequences Document.

These primers can be ordered purified by standard desalting.

Other reagents

- Ethanol (96-100%)
- Ethanol (70%)
- Molecular biology grade water (ddH₂O)

A2.4Steps

A2.4.1 RNA Extraction with QIAGEN's RNEasy Plus Mini Kit

1. Harvest cells. Process each replicate independently. Do not overload gDNA Eliminator

columns. If you have more than 10⁷ cells, split cells in half and process on two columns, then merge the RNA pools. Adherent cells may have to be dissociated using trypsin or a cell scraper.

Pellet cells by centrifuging at 300g for 5 minutes. Aspirate all of the supernatant.

2. Add Buffer RLT Plus (with added 2-mercaptoethanol) to the pellet. Use the following table as a guide.

# cells	Buffer RLT Plus
< 5e6	350 µl
5e6 to 1e7	600 µl

Note the volume used here

Mix by vortexing or pipetting.

- 3.** Homogenize the lysate by vortexing briefly, then letting rest on bench for 1 minute.
Alternatively, cells can be homogenized using QIAshredder spin columns or by repeatedly passing through a 20-gauge needle.
- 4.** Transfer lysate to a gDNA Eliminator spin column placed in a 2 ml collection tube. Centrifuge for 30 seconds at $\geq 8,000g$. Ensure no liquid remains on the column membrane. Repeat centrifugation if necessary. Keep the flow-through and discard the column.
- 5.** Add 1 volume (i.e. 350 or 600 μ l) 70% ethanol to the flow-through and mix by pipetting.
- 6.** Transfer up to 700 μ l of the sample to an RNEasy spin column placed in a 2 ml collection tube. Spin for 15 seconds at $\geq 8,000g$. Discard the flow-through. If sample volume was greater than 700 μ l, centrifuge sample in successive batches on the same column, discarding the flow-through at every step.
- 7.** Add 700 μ l of Buffer RW1 to the column and spin for 15 seconds at $\geq 8,000g$ to wash.
Discard flow-through.
- 8.** Prepare DNase solution by adding 10 μ l of DNaseI to 70 μ l of Buffer RDD for each sample.
Add 80ul of DNase solution to each column. Incubate at room temperature for 15 minutes.
- 9.** Add 350 μ l of Buffer RW1 to the column and spin for 15 seconds at $\geq 8,000g$ to wash.
Discard flow-through.
- 10.** Add 500 μ l of Buffer RPE to the column. Spin for 15 seconds at $\geq 8,000g$. Discard flow-through.
- 11.** Repeat Step 9 but spin for 2 minutes. Discard the flow-through and the collection tube.
- 12.** Place the spin column in a new collection tube and centrifuge for 1 minute at $\geq 8000g$.

13. Place the spin column in a new 1.5 ml collection tube. Add 40 μ l RNase-free water to the column and spin for 1 minute at $\geq 8,000g$ to elute RNA. RNA can be stored at -80°C .

14. Dilute 1 μ l of RNA in 9 μ l of ddH₂O and quantitate using the Qubit RNA HS Assay Kit.

A2.4.2 cDNA Synthesis

15. For cDNA synthesis, continue processing each replicate separately. Prepare the reverse transcription (RT) reaction mix:

- 2 μ g total RNA
- 1 μ l of 50 μ M SMART_dT18VN primer
- 1 μ l of 10 mM dNTPs
- Raise to 14 μ l with ddH₂O

Incubate RT mix at 65 $^{\circ}\text{C}$ for 00:05:00

Place on ice for 1 minute

16. Create 1x Maxima RT buffer:

- For 5 or fewer samples, combine 1 uL of 5X Maxima RT buffer with 4 uL of ddH₂O.
- Mix by pipetting and store on ice.

17. Create a 0.5x Maxima RT H Minus enzyme dilution:

- Mix an equal volume of Maxima RT H Minus Enzyme with the 1x Maxima RT buffer made in step 16 (e.g. 2 uL of Enzyme + 2 uL of 1x buffer).

You will need 1 uL of the 0.5x enzyme dilution for every sample being processed. Avoid pipetting volumes < 1 uL.

18. Add the following to the RT mix:

- 4 µl 5X Maxima RT Buffer
- 1 µl RNaseOUT
- 1 µl of 0.5X Maxima RT H Minus enzyme (1:1 mixture of 1X Maxima RT Buffer and Maxima RT H Minus enzyme = 100 U)

Mix by pipetteing and incubate at 50 °C for 01:00:00

19. Heat inactivate the reaction by incubating at 85 °C for 00:10:00

20. Clean up reaction using 1 µl RNase H and incubating at 37 °C for 00:30:00

Digestion with RNase H removes the complementary RNA strand from the DNA-RNA first strand duplex. This is thought to aid amplification of longer cDNA molecules (> 1 kb)

21. cDNA can be stored at -20°C

A2.4.3 Amplification of Self-Reporting Transcripts

22. This PCR will specifically amplify self-reporting transcripts from cDNA libraries. Prepare the following solution:

- 25 µl 2X Kapa HiFi HotStart ReadyMix
- 1 µl of 25 µM Reverse Primer (SMART)
- 2 µl of cDNA

- 21 μ l of ddH₂O
- 1 μ l of Forward Primer, either:
 - 25 μ M SRT_PAC_F1 primer, if using PB-SRT-Puro
 - 25 μ M SRT_tdTomato_F1, if using PB-SRT-tdTomato

This PCR can be run as half-size reactions by halving each of the listed volumes. If you find yourself doing this PCR repeatedly, this can be a way to decrease costs.

If you have multiple replicates, amplify them separately.

23. Perform PCR using the following thermocycling parameters:

- 95°C for 3 minutes
- 20 cycles of:
 - 98°C for 20 seconds
 - 65°C for 30 seconds
 - 72°C for 5 minutes
- 72°C for 10 minutes
- 4°C forever

24. At this point, gel electrophoresis can be performed to check the quality of amplification. We recommend running 5 μ L of the PCR product from step 23 on 1% TAE agarose gel. The expected product is a smear extending from ~1 kb up to 5 kb.

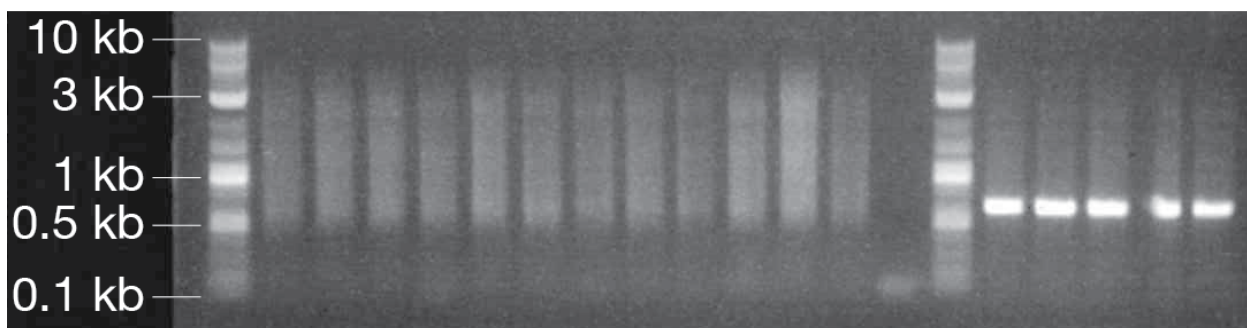


Figure A2.1: Representative products of SRT amplification. 1% TAE gel showing expected products of SRT amplification. Left: the first 12 lanes are biological replicates of a calling card experiment, while the thirteenth is a no template control. The calling card libraries appear as smears extending up to 5 kb. Right: amplification of β -actin with Raff_ACTB_F and Raff_ACTB_R from the same RT product as the SRT samples produces the expected 626-bp product. The ladder is NEB's 1 kb Plus (previously, 2-Log) DNA Ladder (#N3200S).

As a control, we recommend that the constitutive β -actin gene be amplified in parallel to the calling card libraries in steps 22 and 23. The control amplification uses the same PCR mix and thermocycler settings as step 22 and 23, but replaces the calling card forward and reverse primers with human β -actin primers (sequence provided in Materials as Raff_ACTB_F and Raff_ACTB_R). The expected product of the B-actin amplification is 626 bp (see Figure 1 in <https://doi.org/10.2144/97233st02>).

A2.4.4 Purification of PCR Products

25. Vortex AMPure XP beads to resuspend them. Beads should be brought to room temperature for at least 30 minutes prior to use.

26. Add 30 μ l beads to each 50 μ l PCR mixture (0.6x ratio; if you did a half-size PCR, add 15 μ l beads). Mix by pipetting 10 times until evenly dispersed.

27. Incubate at room temperature for 00:05:00

28. Place on a magnetic rack for 2 minutes. Aspirate supernatant and discard.

29. Add 200 μ l of freshly-prepared 70% ethanol and incubate \geq 30 seconds. Aspirate supernatant and discard.

30. Repeat Step #29.

31. Air dry the pellet at room temperature for 2 minutes.

32. Remove the tube from the magnetic rack. Add 20 μ l ddH₂O to elute PCR products. Mix by pipetting until evenly dispersed. Incubate off the rack for 2 minutes.

33. Place on magnetic rack for 1 minute, or until supernatant is clear.

34. Transfer supernatant to new tube. Create a 1:10 dilution and quantitate using the Qubit dsDNA HS Assay Kit.

Expected concentration of product should be 10-20 ng/ μ l.

A2.4.5 Generation of Bulk Calling Card Libraries

35. The tagmentation protocol fragments the long PCR products into libraries suitable for sequencing. This protocol is based on the standard Drop-seq library preparation workflow. Continue processing each replicate independently.

36. Preheat thermocycler to 55 °C

37. Take 1 ng of PCR product and resuspend in a total of 5 μ l ddH₂O in a PCR strip tube.

38. Add 10 μ l of Nextera Tagment DNA (TD) Buffer and 5 μ l of Amplicon Tagment Mix (ATM). Pipette to mix and briefly spin down; bubbles are normal. Incubate at 55 °C for 00:05:00

39. Add 5 μ l of Neutralization Tagment (NT) Buffer. Pipette to mix and briefly spin down; bubbles are normal. Incubate at room temperature for 00:05:00

40. Add the following to each PCR tube in order:

- 15 μ l Nextera PCR Mix (NPM)
- 8 μ l ddH₂O
- 1 μ l of 10 μ M barcoded *piggyBac* primer (e.g. OM-PB-ACG)
- 1 μ l of 10 μ M indexed Nextera N7 primer (e.g. Nextera_N701)

Each replicate should be identifiable by its barcode-index combination. It would be ideal if each replicate had a unique barcode *and* a unique index assigned to it. For some experimental setups, that may not be feasible. One option might be to assign a different index for different conditions/treatments, and within a condition/treatment, assign different barcodes to each replicate.

41. Perform PCR using the following thermocycling parameters:

- 95°C for 3 minutes
- 13 cycles of:
 - 95°C for 10 seconds
 - 50°C for 30 seconds
 - 72°C for 30 seconds
 - 72°C for 5 minutes

- 4°C forever

42. Purify PCR libraries using AMPure XP beads. Vortex AMPure XP beads to resuspend them.

Beads should be brought to room temperature for at least 30 minutes prior to use.

43. Add 35 μ l beads to each 50 μ l PCR mixture (0.7x ratio). Mix by pipetting 10 times until evenly dispersed.

44. Incubate at room temperature for 00:05:00

45. Place on a magnetic rack for 2 minutes. Aspirate supernatant and discard.

46. Add 200 μ l of freshly-prepared 70% ethanol and incubate \geq 30 seconds. Aspirate supernatant and discard.

47. Repeat Step #46.

48. Air dry the pellet at room temperature for 2 minutes.

49. Remove the tube from the magnetic rack. Add 11 μ l ddH₂O to elute PCR products. Mix by pipetting until evenly dispersed. Incubate off the rack for 2 minutes.

50. Place on magnetic rack for 1 minute, or until supernatant is clear. Transfer supernatant to new tube.

A2.4.6 Final Quantitation and Sequencing

51. Create a 1:10 dilution of each final library. Measure concentrations using the Qubit dsDNA HS Assay Kit or on a TapeStation device with a High Sensitivity D1000 ScreenTape. Libraries should be smoothly distributed between 300-60 bp.

Expected concentration of product should be 2-4 ng/ μ l.

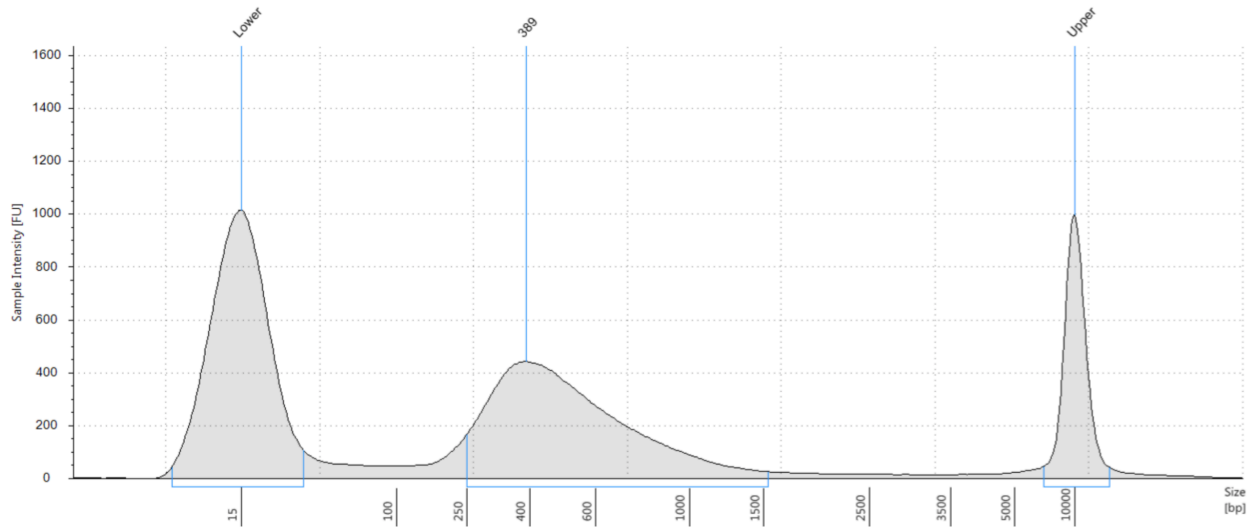


Figure A2.2: Representative TapeStation trace of bulk calling card libraries. This sample was run on a High Sensitivity D5000 ScreenTape.

52. Libraries can be sequenced on any Illumina sequencing platform. Due to the low complexity nature of calling card libraries, we recommend adding PhiX at a final concentration of 50%.

Bulk calling card libraries only use the information from read 1 for mapping insertions. Therefore, single-end sequencing should be sufficient, with at least 75 bp for read 1. An index 1 read will also be necessary for demultiplexing samples.

Appendix 3: Single Cell Calling Cards

Library Preparation

(A version of this appendix has been published on protocols.io:

<https://doi.org/10.17504/protocols.io.xwifpce>)

A3.1 Abstract

This protocol describes how to create calling card libraries from single cell RNA. We assume you have successfully transformed cells with *piggyBac* self-reporting transposons and either undirected *piggyBac* transposase or your favorite transcription factor (YFTF) fused to *piggyBac*. We also assume you have optimized the dissociation protocol for your specific cells or tissues and can generate single cell suspensions.

A3.2 Guidelines

Please read this protocol in its entirety before starting. For several steps, it may help to pre-program your thermocycler or heat block with the listed settings. While single cell calling card (scCC) libraries can, in principle, be generated from any poly(A)-based scRNA-seq method, this protocol specifically describes how to proceed from 10x Chromium 3' scRNA-seq libraries. Please obtain all additional kits, reagents, and equipment as specified in the 10x Chromium Single Cell 3' User Guide.

The components in this protocol are sensitive to repeated freeze-thaw cycles, specifically the modified primers for amplifying self-reporting transcripts and the components of the Nextera Mate Pair Library Prep kit. We recommend pipetting the primers (at 100 μ M) and kit buffers (CB: Circularization Buffer 10X; ERP3: End Repair Mix; ATL2: A-Tailing Mix; LIG2: Ligation

Mix; STL: Stop Ligation Buffer; EPM: Enhanced PCR Mix) into five-use aliquots and storing at –20°C until needed. Sterilize the Axygen 1.7 ml tubes in an autoclave.

In bulk calling cards, we recommend collecting 8-12 independent biological replicates to ensure sufficient statistical power for identifying true binding sites. This is not necessary in single cell calling cards, where each cell is barcoded and is thus considered an independent replicate.

This protocol is meant to describe how we prepare calling card libraries. While it is possible that another kit or component could equally suffice, we have not tested any substitutions and do not officially support deviations from this protocol. This document enumerates what we have had success with and is a starting point from which we can best help troubleshoot.

A3.3 Materials

Name	Catalog #	Vendor
dNTP	639125	Takara
2x Kapa HiFi Hotstart Readymix	KK2602	Kapa Biosystems
Dynabeads M-280 Streptavidin	11205D	Thermo Fisher Scientific
Maxima RT 5X Buffer	Provided with EP0752	Thermo Fisher Scientific
Maxima H Minus Reverse Transcriptase (200 U/uL)	EP0752	Thermo Fisher Scientific
High Sensitivity D5000 ScreenTape	5067-5592	Agilent Technologies
High Sensitivity D5000 Reagents	5067-5593	Agilent Technologies
RNaseOUT™ Recombinant Ribonuclease Inhibitor	10777019	Thermo Fisher Scientific
High Sensitivity D1000 Reagents	5067-5585	Agilent Technologies
High Sensitivity D1000 ScreenTape	5067-5584	Agilent Technologies
Chromium Single Cell 3' Library & Gel Bead Kit v2	120267	10x Genomics
Chromium Single Cell A Chip Kit	1000009	10x Genomics
Chromium i7 Multiplex Kit	120262	10x Genomics
Nextera Mate Pair Library Prep Kit	FC-132-1001	Illumina, Inc.
1.7 ml Axygen Maxymum Recovery Microcentrifuge Tubes	MCT-175-L-C	Axygen
Covaris T6 (6 x 32 mm) glass tubes	520031	Covaris
Covaris Snap-Cap - Teflon Silicone Septa 8 mm	520042	
Ficoll PM-400	17030010	Ge Healthcare

Primers

All primers should be resuspended in Low TE Buffer (10 mM Tris-HCl pH 8.0, 0.1 mM EDTA) at a final concentration of 100 μ M and stored at -20° C.

The following primers should be purified by HPLC and stored as 5 μ l aliquots:

>Bio_Illumina_Seq1_scCC_10X_3xPT

/5Phos/ACACTCTTTCCC/iBiodT/ACACGACGCTCTTCCGA*T*C*T

>Bio_Long_PB_LTR_3xPT

/5Phos/GCGTCAATTTTACGCAGAC/iBiodT/ATCTTTC*T*A*G

These primers should be purified by PAGE:

>scCC_PB_CustomRead2

CGTGTAGGGAAAGAGTGTGCGTCAATTTTACGCAGACTATCTTTCTAG

>scCC_CustomIndex1

GAGACTGGCAAGTACACGTCGCACTCACCATGA

These primers can be purified by standard desalting:

>scCC_P5_adapter

AATGATACGGCGACCACCGAGATCTTCACTCATTCCACACGACTCCTTGCCAGTC

TC*T

>scCC_P7_adapter

/5Phos/GAGACTGGCAAGTACACGTCGCACTCACCATGA[index]ATCTCGTATGCCG
TCTTCTGCTTG

Note that you can replace [index] with an 8-10 bp sequence for multiplexing samples. For more guidance, consult the Illumina Adapter Sequences Document.

>scCC_P5_primer

AATGATACGGCGACCACCGAGATC

>scCC_P7_primer

CAAGCAGAAGACGGCATAACGAGAT

>10x_TSO

AAGCAGTGGTATCAACGCAGAGTACATrGrGrG

Equipment

- 10x Chromium Controller
- Thermocycler for PCR
- Heat blocks or programmable thermoshaker
- Covaris AFA Ultrasonicator, model S2, S220, or E220.

Other reagents

- Consult the 10x Chromium Single Cell 3' User Guide for scRNA-specific consumables

- Ethanol (96-100%)
- ddH₂O

A3.4Steps

A3.4.1 Single Cell Barcoding and Reverse Transcription

1. Prepare cells for isolation and encapsulation in gel bead emulsions (GEMs). If your experiment involves a *piggyBac* transposase with PB-SRT-Puro transposons, cells that have survived selection should be dissociated and resuspended in solution. If you are using *piggyBac* with PB-SRT-tdTomato, we recommend using FACS to isolate tdTomato-positive cells, running cells transfected with PB-SRT-tdTomato alone as a gating control.

2. Follow [10x's instructions](#) for GEM Generation & Barcoding, with this modification:

- Step 1.1: Replace the RT Primer with an equivalent volume of Low TE Buffer. In v3 chemistry, The RT primer has been renamed Template Switch Oligo.

Proceed with Steps 1.2–1.5 as instructed: loading the Single Cell 3' chip, running the controller, transferring GEMs, and reverse transcription.

Incubate the RT reaction under standard conditions.

- Set lid temperature to 53 °C
- 00:45:00 53 °C
- 00:05:00 85 °C
- Hold at 4 °C

3. Step 2.1: clean the GEM-RT mixture using the Recovery Agent and DynaBeads MyOne Silane per 10x's instructions. At the final elution stage, add 36.5 μ l Elution Solution I to the tube, mix by pipetting, and incubate at room temperature for 1 minute. Place the tube in a 10x Magnetic Separator in the Low position until the solution turns clear. Transfer 36 μ l of the eluted sample to a new tube.

4. Divide the eluate into two 18- μ l aliquots. These can be stored at -20°C until needed. One aliquot will be used for scRNA-seq library preparation, while the other will be used to generate scCC libraries.

A3.4.2 Single Cell RNA-seq Library Preparation and Sequencing

5. To continue preparing scRNA-seq libraries, we need to add the template switch oligonucleotide to first strand synthesis products from the RT reaction. Take one of the 18 μ l aliquots and thaw on ice.

6. Prepare the following 1X master mix:

- 20 μ l Maxima 5X RT buffer
- 20 μ l 20% w/v Ficoll PM-400
- 10 μ l 10 mM dNTPs
- 2.5 μ l RNaseOUT
- 2.5 μ l 100 μ M 10x_TSO

7. To the mix, add 18 μ l of first strand RT product and 22 μ l H_2O . Add 5 μ l Maxima H- RTase to the reaction, flick to the mix, and centrifuge briefly.

8. Incubate:

- 00:30:00 25 °C
- 01:30:00 50 °C
- 00:05:00 85 °C

9. Clean up following 10x's post GEM-RT Cleanup protocol, starting with the addition of DynaBeads MyOne Silane (Step 2.1, part D). Clean samples per manufacturer's instructions.

10. Complete cDNA amplification and library construction according to the 10x's instructions (Steps 2.2–3.7). For each sample, record which index sample index was used for the final PCR. Quantiate each library by running a 1:10 dilution on an Agilent TapeStation High Sensitivity D1000 ScreenTape.

11. Finished scRNA-seq libraries can be pooled and sequenced on Illumina MiSeq, NextSeq, HiSeq, and NovaSeq platforms.

A3.4.3 Amplification of Self-Reporting Transcripts

12. To prepare single cell calling cards libraries, we start by amplifying self-reporting transcripts from the other aliquot of first-strand synthesis product. As before, thaw the remaining 18 μ l aliquot on ice.

13. Prepare a PCR primer cocktail in a PCR tube:

- 5 μ l of 100 μ M Bio_Illumina_Seq1_scCC_10X_3xPT primer
- 5 μ l of 100 μ M Bio_Long_PB_LTR_3xPT primer
- 10 μ l of Low TE Buffer

Mix by vortexing and spin down briefly. This cocktail can be stored at -20°C .

14. Prepare the following PCR mix in PCR tube:

- 25 μl of 2X Kapa HiFi Hotstart Readymix
- 18 μl of first-strand synthesis product
- 6 μl of ddH₂O
- 1 μl of PCR primer cocktail

Keep on ice until ready for PCR.

15. Perform PCR using the following thermocycling parameters:

- 98°C for 3 minutes
- 20 cycles of:
 - 98°C for 20 seconds
 - 67°C for 30 seconds
 - 72°C for 5 minutes
 - 72°C for 10 minutes
- 4°C forever.

The number of cycles may need to be adjusted depending on the cell type and number of cells represented in the library. If uncertain, you can use 9 μl of first-strand synthesis product as template, reserving the other 9 μl for another round of PCR with more cycles as needed.

A3.4.4 Purification of PCR Products

16. Vortex AMPure XP beads to resuspend them. Beads should be brought to room temperature for at least 30 minutes prior to use.

17. Add 30 μ l beads to the 50 μ l PCR mixture (0.6x ratio). Mix by pipetting 10 times until evenly dispersed.

18. Incubate at room temperature for 00:05:00

19. Place on a magnetic rack for 00:05:00

Aspirate supernatant and discard.

20. While the tube is still on the rack, add 200 μ l of 70% ethanol and incubate \geq 30 seconds.

Aspirate supernatant and discard.

21. Repeat Step #20

22. Air dry the pellet at room temperature for 00:02:00

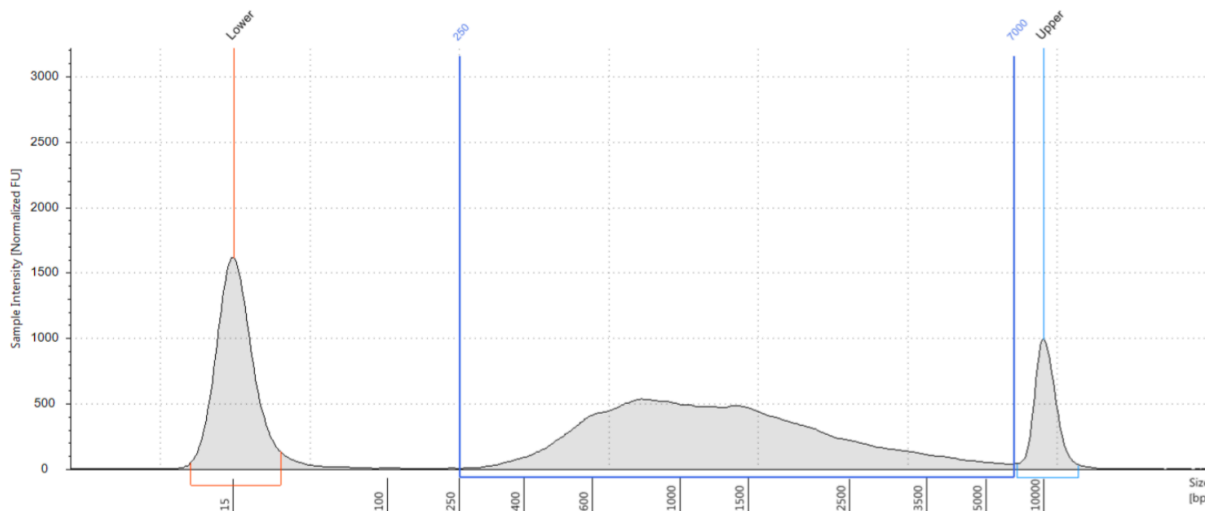
23. Remove the tube from the magnetic rack. Add 40 μ l QIAGEN Elution Buffer to elute PCR products. Mix by pipetting until evenly dispersed. Incubate off the rack for 00:05:00

24. Place on magnetic rack for 00:05:00 or until supernatant is clear. Transfer supernatant to new 1.7 ml tube.

25. Take 1 μ l of the eluate and dilute in 9 μ l of ddH₂O to make a 1:10 dilution. Quantitate on TapeStation using a High Sensitivity D5000 ScreenTape. Measure the molar concentration of the sample, taking everything from 250 bp to 7000 bp. Ideally, the diluted sample will be at least 750 pM, corresponding to 7.5 nM for the original eluate. If you kept half of the template aside, you

can increase the number of PCR extension cycles until you get to a minimum of 7.5 nM of product.

This is what a representative TapeStation trace looks like. The library should be smooth and unimodal.



Region Table

From [bp]	To [bp]	Average Size [bp]	Conc. [pg/ul]	Region Molarity [pmol/l]	% of Total	Region Comment	Color
250	7000	1445	1170	1790	96.59	Everything	■

Figure A3.1: Representative TapeStation trace of SRT amplification from 10x 3' scRNA-seq library.

A3.4.5 Single Cell Calling Cards – Circularization

26. Thaw an aliquot of CB – Circularization Buffer 10X on ice.

27. Add the following components to a new 1.7 ml tube in this order:

1. 300 fmol self-reporting transcripts from Step 25
2. ddH₂O up to a total of 268 µl
3. 30 µl CB
4. 2 µl Circularization Ligase

To calculate what volume of eluate corresponds to 300 fmol, divide 300 by the molar concentration (in nM) of the eluate. For example, if the concentration is 10 nM, $300 \text{ fmol}/10 \text{ nM} = 30 \text{ }\mu\text{l}$, and consequently you would need 238 μl ddH₂O. If you have less than 300 fmol total, you may proceed but might need to make adjustments at the final PCR step. If you do not have a way to quantitate the molarity of your solution, we have observed that 300 fmol of self-reporting transcripts is (very) approximately 200 ng.

We strongly encourage you to calculate the volume of SRT solution based on molarity, not by mass. The circularization reaction is sensitive to starting concentration. If overloaded, it can lead to excess intermolecular ligations and, subsequently, increased noise with respect to the assignment of insertions to cell types.

Mix by flicking the tube and spin down briefly. Incubate at 30 °C overnight (12-16 hours).

A3.4.6 Single Cell Calling Cards – Exonuclease and Setup

28. Add 9 μl of PS1 – Exonuclease directly to the overnight circularization mixture. Flick to mix, spin down, and incubate as following:

- 00:30:00 37 °C
- 00:30:00 70 °C

29. While the exonuclease digestion proceeds, prepare for the rest of the library preparation. Fill a large ice bucket with ice. Thaw, on ice, aliquots of:

- STL – Stop Ligation Buffer
- ERP3 – End Repair Mix

- ATL2 – A-tailing Mix
- LIG2 – Ligation Mix
- EPM – Enhanced PCR Mix

Also thaw the following oligonucleotides:

- scCC_P5_adapter (100 μ M)
- scCC_P7_adapter (100 μ M)
- scCC_P5_primer (25 μ M)
- scCC_P7_primer (25 μ M)

Finally, thaw NEBuffer 2

30. While the exonuclease incubates, anneal the scCC adapters. Prepare the following mixture in a PCR tube, using a different indexed scCC_P7_adapter for each sample:

- 4.5 μ l scCC_P5_adapter
- 4.5 μ l scCC_P7_adapter
- 1 μ l NEBuffer 2

31. Anneal scCC adapters in a thermocycler using the following settings:

- 95°C for 5 minutes
- 70°C for 15 minutes

- Ramp down to 25°C as slowly as possible
- 25°C for 5 minutes
- 4°C forever

scCC adapters can be kept on ice until needed.

Adapters should be prepared fresh. NEBuffer 2 contains magnesium salts which can promote DNase activity, leading to degradation of adapters.

32. Prepare the streptavidin-coated magnetic beads. These instructions are for 1 sample; up to 5 can be prepared in a single 1.7 ml tube. Resuspend Dynabeads M-280 by vortexing briefly.

33. Transfer 20 μ l of beads to a clean 1.7 ml tube.

34. Place on a magnetic rack for 1 minute. Once clear, aspirate and discard supernatant.

35. Add 40 μ l BBB – Bead Bind Buffer. Incubate for 1 minute, then aspirate and discard supernatant.

36. Repeat Step #35.

37. Remove from rack and add 300 μ l BBB. Beads can be stored at room temperature until needed.

38. The exonuclease digestion should be complete by now. Add 12 μ l STL – Stop Ligation Buffer. Flick to mix and centrifuge gently.

A3.4.7 Single Cell Calling Cards – Shearing and Capture

39. Transfer the entire sample (now approximately 320 μ l) to a Covaris T6 tube. Add ddH₂O as necessary to fill to the top, then cap the tube. Check to make sure there are no air bubbles.

40. Shear DNA on a Covaris ultrasonicator. Here are recommended settings for various models (we have tested this protocol on the E220):

Model	S2	S220	E220
Peak Power Intensity	N/A	240	200
Intensity	8	N/A	N/A
Duty Cycle/Factor	20%	20%	20%
Cycles Per Burst	200	200	200
Time	40	40	40
Temperature	6	6	6

Recommended shearing settings for preparing scCC libraries

41. Transfer the sample to a new 1.7 ml tube. Add 300 μ l of bead solution to the sheared DNA.

42. Incubate 20 °C 00:15:00

If incubating on a thermoshaker, shake at 1000 RPM. Otherwise, flick to mix every 2 minutes.

43. Centrifuge briefly (5-10 seconds), then place on a magnetic rack for 1 minute. Discard the supernatant.

44. Wash 4 times with BWB – Bead Wash Buffer:

- Add 200 μ l BWB
- Remove from rack, flick to mix, and spin down briefly (1-2 seconds)
- Place on rack for 30 seconds
- Discard supernatant

45. Wash 2 times with RSB – Resuspension Buffer:

- Add 200 μ l RSB
- Remove from rack, flick to mix, and spin down briefly
- Place on rack for 30 seconds
- Discard supernatant

For the second wash, do not discard supernatant until ready to add the master mix in the next step.

Repeat for a total for 4 washes

A3.4.8 Single Cell Calling Cards – End Repair, A-Tailing, and Adapter Ligation

46. Prepare master mixes for End Repair and A-Tailing as follows.

1X End Repair Master Mix:

- 40 μ l ERP3 – End Repair Mix
- 60 μ l ddH₂O

1X A-Tailing Master Mix:

- 12.5 μ l ATL2 – A-Tailing Mix
- 17.5 μ l ddH₂O

47. Discard all supernatant from the DNA sample. Centrifuge briefly, then place on a magnetic rack.

48. Use a 10 μ l pipette to aspirate any residual supernatant.

49. Add 100 μ l End Repair reaction mix, remove from the rack, flick to mix, and centrifuge briefly (do not allow beads to pellet).

50. Incubate 30 °C 00:30:00

If incubating on a thermoshaker, shake at 1000 RPM, to prevent beads from settling.

51. Centrifuge briefly (5-10 seconds), then place on a magnetic rack for 1 minute. Discard the supernatant.

52. Wash 4 times with BWB – Bead Wash Buffer:

- Add 200 μ l BWB
- Remove from rack, flick to mix, and spin down briefly (1-2 seconds)
- Place on rack for 30 seconds
- Discard supernatant

Repeat for a total for 4 washes

53. Wash 2 times with RSB – Resuspension Buffer:

- Add 200 μ l RSB
- Remove from rack, flick to mix, and spin down briefly
- Place on rack for 30 seconds
- Discard supernatant

For the second wash, do not discard supernatant until ready to add the master mix in the next step.

54. Discard all supernatant from the DNA sample. Centrifuge briefly, then place on a magnetic rack. Use a 10 μ l pipette to aspirate any residual supernatant.

55. Add 30 μ l A-Tailing reaction mix, remove from the rack, flick to mix, and centrifuge briefly (do not allow beads to pellet).

56. Incubate 37 °C 00:30:00

If incubating on a thermoshaker, shake at 1000 RPM, to prevent beads from settling.

57. Add the following components in order to the A-tailing mix:

- (30 μ l A-tailing reaction)
- 2.5 μ l LIG2 – Ligation Mix
- 4 μ l ddH₂O
- 1 μ l annealed scCC adapter

Flick to mix and centrifuge briefly (do not allow beads to pellet).

58. Incubate 30 °C 00:10:00

59. Add 5 μ l STL – Stop Ligation Buffer. Flick to mix.

60. Centrifuge briefly (5-10 seconds), then place on a magnetic rack for 1 minute. Discard the supernatant.

61. Wash 4 times with BWB – Bead Wash Buffer:

- Add 200 μ l BWB
- Remove from rack, flick to mix, and spin down briefly (1-2 seconds)
- Place on rack for 30 seconds
- Discard supernatant

Repeat for a total for 4 washes

62. Wash 2 times with RSB – Resuspension Buffer:

- Add 200 μ l RSB
- Remove from rack, flick to mix, and spin down briefly
- Place on rack for 30 seconds
- Discard supernatant

For the second wash, do not discard supernatant until ready to add the master mix in the next step.

A3.4.9 Single Cell Calling Cards – Final PCR and Purification

63. Prepare a 1X PCR master mix in a new 1.7 ml tube:

- 20 μ l EPM – Enhanced PCR Mix
- 28 μ l ddH₂O
- 1 μ l scCC_P5_primer (25 μ M)

- 1 μ l scCC_P7_primer (25 μ M)

64. Discard all supernatant from the DNA sample. Centrifuge briefly, then place on a magnetic rack. Use a 10 μ l pipette to aspirate any residual supernatant.

65. Add 50 μ l PCR reaction mix to the sample and pipette to mix. Transfer to PCR tubes.

66. Incubate in a thermocycler with the following settings:

- 98°C for 30 seconds
- 15 cycles of:
 - 98°C for 10 seconds
 - 60°C for 30 seconds
 - 72°C for 2 minutes
- 72°C for 5 minutes
- 4°C forever.

If you started with less than 300 fmol of self-reporting transcripts, you can increase the number of extension cycles here. More cycles will increase the risk of artifacts, however, so we recommend increasing by the minimum necessary to obtain reasonable sequencing libraries. The most we have pushed this PCR is to 17 extension cycles.

67. Vortex AMPure XP beads to resuspend them. Beads should be brought to room temperature for at least 30 minutes prior to use.

68. Place PCR tubes on a magnetic rack for 1 minute. Transfer 50 μ l of supernatant to new tubes.

69. Add 35 μ l beads to the 50 μ l PCR mixture (0.7x ratio). Flick to mix and centrifuge briefly.

70. Incubate at room temperature for 00:05:00

71. Place on a magnetic rack for 00:05:00

Aspirate supernatant and discard.

72. Add 200 μ l of 70% ethanol and incubate \geq 30 seconds. Aspirate supernatant and discard.

73. Repeat Step #72

74. Air dry the pellet at room temperature for 00:02:00

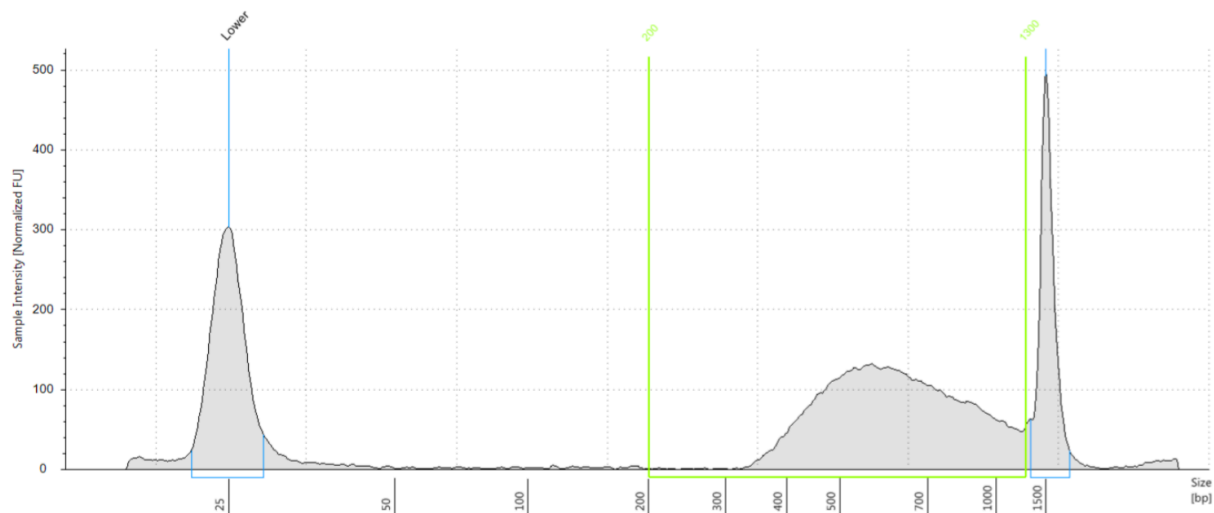
75. Remove the tube from the magnetic rack. Add 25 μ l RSB – Resuspension Buffer to elute PCR products. Mix by pipetting until evenly dispersed. Incubate off the rack for 00:05:00

76. Place on magnetic rack for 00:05:00 or until supernatant is clear. Transfer supernatant to new 1.7 ml tube.

77. Make a 1:10 dilution of the eluate and quantitate on TapeStation using a High Sensitivity D1000 ScreenTape. Measure the molar concentration of the sample, taking everything from 200 bp to 1300 bp.

This is what a representative TapeStation trace looks like. The library should be smooth and unimodal, peaking between 500-700 bp. Occasionally, you may see a primer-dimer peak.

However, as scCC libraries are sequenced from the middle and not the ends, the primer-dimer product will not sequence on the Illumina flow cell.



Region Table

From [bp]	To [bp]	Average Size [bp]	Conc. [pg/ul]	Region Molarity [pmol/l]	% of Total	Region Comment	Color
200	1300	663	731	1870	88.88	scCC	■

Figure A3.2: Representative TapeStation trace of scCC libraries.

A3.4.10 Single Cell Calling Cards – Sequencing

78. Single cell calling cards libraries use a mix of standard and custom primers for sequencing and rely on dual-indexing for proper demultiplexing. We have sequenced scCC libraries on Illumina NextSeq 500 machines, using v2 Reagent Cartridges. These libraries use the standard Illumina primers BP10 and BP14 for read 1 and index 2, respectively. Read 1 sequences the cell barcode and unique molecular index (UMI), while index 2 reads into the terminal repeat of the *piggyBac* transposon, confirming that molecules successfully circularized.

In addition, we use the custom sequencing primers scCC_PB_CustomRead2 and scCC_CustomIndex1 for read 2 and index 1, respectively. Read 2 anneals at the end of the transposon and sequences into the genome. The first six base pairs typically begin "GGTTAA", which are the terminal two base pairs of the *piggyBac* repeat followed by the insertion site tetramer. The remainder of the read is genomic DNA sequence. Index 1 sequences the sample-specific sequence on the scCC adapter and is used to demultiplex libraries.

Due to the low complexity nature of calling card libraries, we recommended adding PhiX at a final concentration of 50%.

79. While index 1 should be sufficient to demultiplex libraries, we have observed a biphasic response when sequencing scCC libraries at low and high concentrations. At low library concentrations, (e.g., 1-2%) the index 1 read generates high-quality reads and can demultiplex libraries; however at higher concentrations (i.e., 50%) the index 1 read can fail, yielding all N's. If this happens, libraries can be demultiplexed by the index 2 read alone: scCC reads that have successfully circularized will have "GCGTCAAT" as the index 2 sequence.

After this, scCC reads can be assigned to specific samples using the cell barcodes obtained from the corresponding scRNA-seq libraries. Different libraries may, by chance, have cells that share the same cell barcode. Typically, these represent a very small fraction of cells (< 1% per library) and we discard these reads and cells from downstream calling cards and scRNA-seq analyses, respectively.

Appendix 4: Processing Bulk Calling Card Sequencing Data

(A version of this appendix has been published on protocols.io:

<https://doi.org/10.17504/protocols.io.xwjfpcn>)

A4.1 Abstract

Here we present a computational pipeline for processing bulk RNA calling card data. These data will have been generated from transfection/-duction of either undirected *piggyBac* transposase or your favorite transcription factor (YFTF) fused to *piggyBac*. Multiple biological replicates should have been generated, each with a unique combination of primer barcode and index sequences. This workflow demonstrates how to analyze a single replicate; the workflow can be parallelized on distributed computing architectures (e.g. slurm).

A4.2 Guidelines

Please make sure you have installed the required software and packages (see Materials section).

This protocol describes how to analyze a **SINGLE** biological replicate from a bulk RNA calling cards* experiment. Multiple replicates (e.g. 10-12) should be analyzed in each experiment to distinguish independent transposition events into the same insertion site. This is essential for adequate statistical power to detect transcription factor binding sites. Each replicate can be processed following this protocol, making appropriate changes to the primer barcode sequence and/or the index sequence(s). Data from multiple calling card replicates can be pooled at the end into a single file.

*If you are unfamiliar with calling card libraries, we recommend reading our [quick start guide](#) (Appendix 1) and our [library preparation protocol](#) (Appendix 2).

A4.3 Materials

The following external programs are required:

- [cutadapt](#) (≥ 1.16)
- [samtools](#) (≥ 1.9)

You will need a genomic aligner. Here we will use novoalign, but in theory any aligner should be sufficient (e.g. bowtie2, GATK, STAR, etc.). Also, you will need a .2bit version of the genome sequence you are aligning to; these are readily available from the [UCSC Genome Browser](#). (They can also be generated from a FASTA file using the [faToTwoBit](#) utility)

The following programs are optional, but highly recommended:

- [bedtools](#) (≥ 2.27)
- [bedops](#) (≥ 2.4)

In addition, this workflow calls some calling card-specific scripts, which use Python 3. It is recommended that your Python installation be relatively up-to-date (i.e. ≥ 3.4). To check your python version, type

```
python -V
```

You will need to install the following Python modules:

- [numpy](#)

- [pandas](#)
- [pysam](#)
- [twobitreader](#)

All of these packages are available on PyPI and can be installed via pip:

```
pip install numpy pandas pysam twobitreader
```

(If Python3 is not the default on your system, replace pip with pip3)

Finally, these are the calling card-specific scripts you will need, all of which are available on [GitHub](#):

- TagBam.py
- AnnotateInsertionSites.py
- BamToCallingCard.py

A4.4Steps

A4.4.1 Preamble

1. The objective of this protocol is to take sequencing reads from a calling cards library and process them into a .ccf file. A CCF file is a modified BED file (BED3+3) that concisely enumerates every transposition event in the sequenced library.

CCF files typically have six columns:

- chrom: chromosome
- start: beginning coordinate of the insertion site

- end: ending coordinate of the insertion site; since *piggyBac* inserts into TTAA's, this typically spans the motif itself.
- count: the number of reads supporting this insertion
- strand: + or -, indicating which strand was targeted (optional but highly recommended)
- barcode: a string identifying the library from which this insertion originated (optional but highly recommended)

This workflow will walk through how to perform quality control, alignment, filtering, and processing of calling card sequencing libraries to generate a CCF file. This file can then be used in downstream applications, such as visualization on the (legacy) [WashU Epigenome Browser](#) (instructions [here](#)), and as input for peak calling.

2. To illustrate the workflow, let's say that we have performed bulk RNA calling cards on our favorite transcription factor (YFTF) in a human cell line. We have prepared libraries from 10 biological replicates of cells transfected with wild-type *piggyBac* transposase, and 10 replicates of cells with YFTF-*piggyBac*. We have sequenced these libraries and now need to map these insertions across the genome.

We will consider a single replicate; the workflow can then be repeated for all remaining replicates. At the end we can combine the data from the 10 *piggyBac* replicates, and the 10 YFTF-*piggyBac* replicates, respectively, into a single CCF file each.

3. In this example, we will be analyzing a single replicate from the wild-type *piggyBac* libraries: PBase_rep1. The read 1 sequencing file is PBase_rep1_L001_R1_001.fastq.gz

For bulk RNA calling card libraries, only read 1 is analyzed, as it contains the junction between the transposon and genome.

This biological replicate had GAT as its primer barcode and CTCACGGTGA as its index sequence. It was prepared by PCR ligation with the following primers:

>OM-PB-GAT (barcode in bolded)

```
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC
TGATTTTACGCAGACTATCTTTCTAG
```

>Nextera_N7_CTCACGGTGA (index in bold; note the reverse complement orientation)

```
CAAGCAGAAGACGGCATAACGAGATTCACCGTGAGGTCTCGTGGGCTCGG
```

Thus, each read 1 *should* begin with GATTTTACGCAGACTATCTTTCTAG.

A4.4.2 Adapter Trimming

4. The purpose of this step is to check that reads have (1) the appropriate primer barcode sequence and (2) the transposon sequence is correct and ends in TTAA, *piggyBac*'s insertion motif. If these conditions are true, those bases are trimmed (hard clipped), to facilitate genomic alignment. Only reads with perfect matches to the barcode and transposon sequence are carried forward.

```
cutadapt \  
-g ^GATTTTACGCAGACTATCTTTCTAGGGTTAA \  
--minimum-length 1 \  
--discard-untrimmed \  

```

```
-e 0 \  
  
--no-indels \  
  
-o PBase_repl_trimBC.fastq.gz \  
  
PBase_repl_L001_R1_001.fastq.gz
```

Typically 70-90% of reads will pass this filter, although there may be sample-dependent variation.

5. Next, we re-examine the passing reads and trim any reads that end in the Nextera adapter that was added during tagmentation. This step reduces the amount of non-genomic bases, which should accelerate alignment. Only a small fraction (5-10%) typically have any adapter sequence at all, so virtually every read will pass this filter.

```
cutadapt \  
  
-a  
  
CTGTCTCTTATACACATCTCCGAGCCCACGAGACTCTCACGGTGATCTCGTATGCCGTCTTCTG  
CTTG \  
  
--minimum-length 1 \  
  
-o PBase_repl_trimmed.fastq.gz \  
  
Base_repl_trimBC.fastq.gz
```

The index sequence has been emphasized in bold, but if you are processing libraries with many index different indexes, you can replace the bolded sequence with N's (keeping the length same). cutadapt can handle degenerate bases in adapters.

A4.4.3 Alignment

6. Now that our reads are trimmed, we are ready to align them to the genome. This step can be done with any aligner; we typically use novoalign, so that is what we will demonstrate here.

```
novoalign \  
  
-d hg38.nvx \  
  
-f PBase_repl_trimmed.fastq.gz \  
  
-n 40 \  
  
-o SAM \  
  
-o SoftClip > PBase_repl_trimmed.sam
```

The "-n 40" flag tells novoalign to align only the first 40 bases of the read. We have found that this reduction can increase the speed of alignment with minimal impact on total number of insertions recovered. Faster aligners (e.g. bowtie2, GATK, STAR) may not need this setting.

7. After alignment, we filter out reads that mapped to multiple locations in the genome (e.g. in a repetitive element) and convert to the more space-efficient BAM format.

```
samtools view \  
  
-bS -h -F 260 \  
  
PBase_repl_trimmed.sam | \  
  
samtools sort - -o PBase_repl_mapped.bam
```


A4.4.4 Annotation

8. The BAM format provides a flexible way to annotate reads through the use of short tags.

These tags remain with the reads in the BAM file, which makes for a simple and portable archive of a calling card experiment. We use the following custom tags:

- XP: primer barcode
- XJ: index 1 sequence
- XK: index 2 sequence (optional; reserved for future use)
- XI: insertion site annotation
- XZ: adjacent sequence (to verify transposase motif)

9. First, we will annotate reads with the XP tag for the primer barcode GAT.

```
python TagBam.py \  
  
--tag XP:Z:GAT \  
  
PBase_rep1_mapped.bam \  
  
PBase_rep1_tagged.bam
```

10. Next, we will annotate reads with the XJ tag for the index sequence CTCACGGTGA.

```
python TagBam.py \  
  
--tag XJ:Z:CTCACGGTGA \  
  
PBase_rep1_tagged.bam \  
  
PBase_rep1_tagged2.bam
```

11. Lastly, we will annotate reads with respect to the insertion site. This script checks each read to make sure that it maps next to the *piggyBac* insertion site motif TTAA. Remember, this part of read 1 was trimmed in step 4. By double checking that the read maps next to a genomic TTAA, we add an extra layer of specificity to the alignment. The sequence of the adjacent bases will also be annotated with the XZ tag. Reads that pass will be annotated with the insertion site coordinates in the XI tag and written to the output file.

```
python AnnotateInsertionSites.py \  
  
--transposase PB \  
  
-f \  
  
PBase_repl_tagged2.bam \  
  
hg38.2bit \  
  
PBase_repl_final.bam
```

You can provide a path to the .2bit file if your genome references are in another directory.

A4.4.5 Finishing Up

12. To finish, we first index the BAM file.

```
samtools index PBase_repl_final.bam
```

13. Next, clean up intermediate files.

```
rm PBase_repl_trimBC.fastq.gz
```

```
rm PBase_repl_trimmed.fastq.gz
```

```
rm PBase_repl_trimmed.sam
```

```
rm PBase_repl_mapped.bam
```

```
rm PBase_repl_tagged.bam
```

```
rm PBase_repl_tagged2.bam
```

14. Lastly, convert the BAM file to a CCF file.

```
python BamToCallingCard.py \
```

```
-b XP XJ \
```

```
-i PBase_repl_final.bam \
```

```
-o PBase_repl_final.ccf
```

This will use the combination of primer barcode and index sequence (XP and XJ, respectively) to identify insertions derived from different biological replicates.

Here is an example of a CCF file:

chr1	28575	28579	2	+	GCA/TCGCCACCC
chr1	28575	28579	10	+	TAG/GAGGTACAG
chr1	28575	28579	1	+	GAT/GAGGTACAG
chr1	31191	31195	1	+	GCA/TCGCCACCC
chr1	31191	31195	49	+	TAG/TCGCCACCC
chr1	46620	46624	5	+	CTA/GAGGTACAG
chr1	54136	54140	42	-	GCA/TCGCCACCC
chr1	54818	54822	16	-	CTA/TCGCCACCC
chr1	57829	57833	6	-	CGT/GAGGTACAG
chr1	58414	58418	40	+	CTA/TCGCCACCC

A4.4.6 Notes

15. This workflow described how to process a **SINGLE** biological replicate. After each replicate has been processed, CCF files can be combined to consolidate all insertions from a given

experiment. For example, to combine data from all replicates from our wild-type *piggyBac* libraries, we can use `cat` and `bedops` (preferred):

```
cat \  
  
PBase_rep1_final.ccf \  
  
PBase_rep2_final.ccf \  
  
PBase_rep3_final.ccf \  
  
PBase_rep4_final.ccf \  
  
PBase_rep5_final.ccf \  
  
PBase_rep6_final.ccf \  
  
PBase_rep7_final.ccf \  
  
PBase_rep8_final.ccf \  
  
PBase_rep9_final.ccf \  
  
PBase_rep10_final.ccf | sort-bed - > PBase.ccf
```

Similarly, CCFs from the YFTF replicates can be combined:

```
cat YFTF-PBase_rep*_final.ccf | sort-bed - > YFTF-PBase.ccf
```

The concatenated CCF files can also be sorted using `bedtools`:

```
cat PBase_rep*_final.ccf | bedtools sort -i - > PBase.ccf
```

Or, using the standard shell `sort` command:

```
cat PBase_rep*_final.ccf | sort -k1V -k2n -k3n > PBase.ccf
```

16. Analogously, we can combine BAM files from biological replicates into a single archival-quality BAM file for an entire experiment:

```
samtools merge PBase.bam PBase_rep*_final.bam
```

17. Ideally, each biological replicate will have a unique primer barcode AND unique index sequence. However, sometimes this is not possible. If so, each replicate should be identifiable from a unique combination of primer barcode and index sequence. If multiple replicates share an index, their reads will be found in the same FASTQ file. This is okay as step 3 can separate each replicate based on an exact match to the primer barcode sequence. In that case, you will have to provide the same input file to step 4 multiple times, each with a different primer barcode at the start of the adapter.

Appendix 5: Processing Single Cell Calling Card Sequencing Data

(A version of this appendix has been published on protocols.io:

<https://doi.org/10.17504/protocols.io.4phgvj6>)

A5.1 Abstract

Here we present a computational pipeline for processing single cell calling card (scCC) data.

These data will have been generated from single cell RNA-seq libraries following transfection/-duction of either undirected *piggyBac* transposase or your favorite transcription factor (YFTF) fused to *piggyBac*. This workflow demonstrates how to process scCC sequencing data derived from a 10x Chromium-based scCC library; the workflow can be parallelized on distributed computing architectures (e.g. slurm).

A5.2 Guidelines

Please make sure you have installed the required software and packages (see Materials section).

This protocol describes how to analyze a single cell calling cards* (scCC) experiment. These libraries are derived from single cell RNA-seq libraries (scRNA-seq). Currently, we only support 10x Chromium 3'-based libraries. Unlike bulk calling card libraries, in which we require multiple replicates (e.g. 10-12), each cell in a scCC library is considered a replicate. Sensitivity is driven by the total number of insertions recovered, which is directly proportional to the number of transformed cells in the scRNA-seq library. Therefore, we advise processing as many cells as feasible to maximize discovery of transcription factor binding sites.

*If you are unfamiliar with calling card libraries, we recommend reading our [quick start guide](#) (Appendix 1) and our [scCC library preparation protocol](#) (Appendix 3).

A5.3 Materials

The following external programs are required:

- [cutadapt](#) (≥ 1.16)
- [samtools](#) (≥ 1.9)
- [cellranger](#) ($\geq 2.1.0$)

You will need a .2bit version of the genome sequence you are aligning to; these are readily available from the [UCSC Genome Browser](#). (They can also be generated from a FASTA file using the [faToTwoBit](#) utility)

The following programs are optional, but highly recommended:

- [bedtools](#) (≥ 2.27)
- [bedops](#) (≥ 2.4)

In addition, this workflow calls some calling card-specific scripts, which use Python 3. It is recommended that your Python installation be relatively up-to-date (i.e. ≥ 3.4). To check your python version, type

```
python -V
```

You will need to install the following Python modules:

- [numpy](#)

- [pandas](#)
- [pysam](#)
- [twobitreader](#)

All of these packages are available on PyPI and can be installed via pip:

```
pip install numpy pandas pysam twobitreader
```

(If Python3 is not the default on your system, replace pip with pip3)

Finally, these are the calling card-specific scripts you will need, all of which are available on [GitHub](#):

- TagBam.py
- AnnotateInsertionSites.py
- BamToCallingCard.py
- UMIFilter.py
- FilterUniqueBarcodes.py
- FilterBAMByBarcodes.py
- FilterCCFByBarcodes.py

A5.4Steps

A5.4.1 Preamble

1. The objective of this protocol is to take sequencing reads from single cell calling cards (scCC)

library and process them into a CCF (calling card format; .ccf) file. A CCF file is a modified

BED file (BED3+3) that concisely enumerates every transposition event in the sequenced library.

CCF files typically have six columns:

- chrom: chromosome
- start: beginning coordinate of the insertion site
- end: ending coordinate of the insertion site; since *piggyBac* inserts into TTAA's, this typically spans the motif itself.
- count: the number of reads supporting this insertion
- strand: + or -, indicating which strand was targeted (optional but highly recommended)
- barcode: a string identifying the library or cell from which this insertion originated. For scCC, this is the sequence of the cell barcode.

Each line of the final .ccf file represents an independent calling card insertion, and the value in the cell barcode column specifies in which cell this insertion was observed. If your sample is heterogeneous, you may find that your cells can be grouped into biologically meaningful clusters (e.g. different cell types) based on their scRNA-seq expression profiles. In this case, you will have assigned cell barcodes to each cluster, and can use this information to split the .ccf file to generate insertion profiles for each cluster (see Step 19). These split .ccf files can then be used to identify differentially bound loci, or for visualization of TF binding in different clusters.

This workflow walks through how to perform quality control, alignment, filtering, and processing of single cell calling card sequencing libraries to generate a .ccf file. This file can

then be used in downstream applications, such as visualization on the (legacy) [WashU Epigenome Browser](#) (instructions [here](#)), and as input for peak calling.

2. To illustrate the workflow, let's say that we have performed scCC on our favorite transcription factor (YFTF) in a human cell line. We have prepared scRNA-seq and scCC libraries from 10,000 cells transfected with wild-type *piggyBac* transposase and 10,000 cells transfected with YFTF-*piggyBac*. Both the wild-type and YFTF transfectants were loaded across two wells each of a Chromium chip.

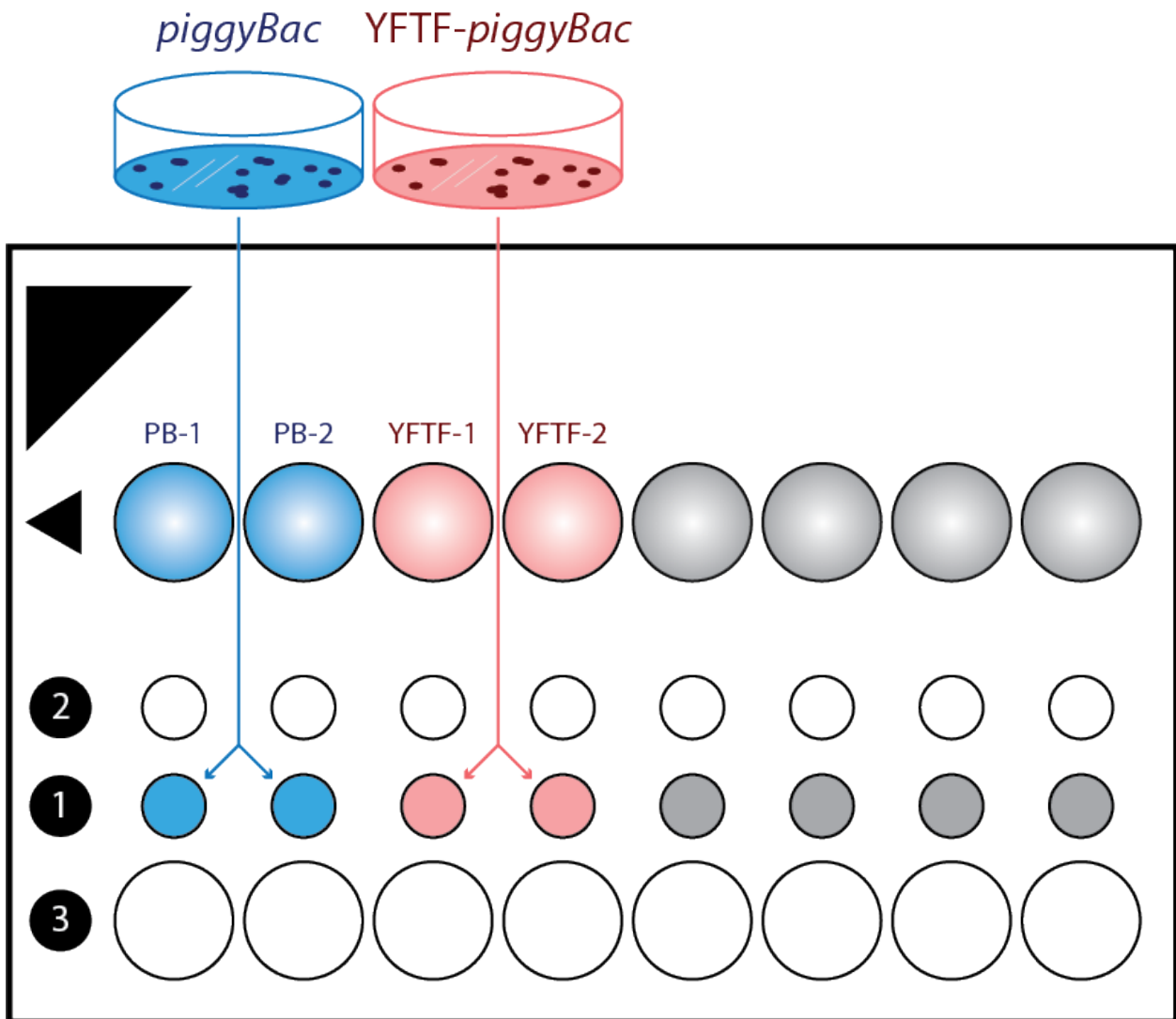


Figure A5.1: Example of a 10x Chromium chip loaded for a single cell calling cards experiment.

This image illustrates our example experiment. Two wells have been loaded with *piggyBac*-transfected cells in Row 1 (blue) and two wells with YFTF-*piggyBac* transfectants (pink). The remaining wells, in grey, were empty (loaded with 50% glycerol). After emulsion generation, the resulting libraries—in the row marked by the triangle—were kept separate and processed according to the [scCC molecular protocol](#).

3. scCC analysis requires two separate sequencing runs: one for the scRNA-seq library (which will be used for dimensionality reduction and cell type identification) and one for the scCC library (to assign insertions to specific cells). The scRNA-seq library can be sequenced using 10x's standard recommendations and processed using cellranger. **We will assume this step has already been completed.** This step establishes a set of high-quality cell barcodes from each library; we will cross-reference these with the scCC library to assign insertions to specific cells.

4. The scCC library should have been sequenced as recommended in our [scCC molecular workflow](#). Specifically, on a dual indexed-compatible Illumina sequencer; we prefer to sequence these libraries on an Illumina NextSeq 500 with 50% phiX, allocating 26 bases to read 1, 50 bases to read 2, and 8 bases each to index 1 and index 2. Although scCC libraries should be demultiplexable with unique index sequences, this does not always work and the index 1 read can fail, reporting all N's. If this happens, reads from all libraries will be mixed together. We can identify scCC reads from phiX and other artifacts by demultiplexing with the index 2 read (should be GCGTCAAT). To further identify reads from the constituent libraries, we will demultiplex using the cell barcode.

5. In this example, we will be analyzing scCC sequencing run demultiplexed using index 2 only.

The read 1 file will contain the cell barcode and UMI, while read 2 will contain the junction

between the transposon and the genome. The read 1 and read 2 files are, respectively:

- PB_YFTF-PB_combined_R1_001.fastq.gz
- PB_YFTF-PB_combined_R2_001.fastq.gz

Read 1 is 26 bases long: the first 16 bases comprise the cell barcode, while the final 10 bases are the UMI.

Read 2 is 50 bases long: the first 2 bp will contain transposon sequence (GG) followed by the 4bp TTAA insertion site; the rest of the read is genomic sequence (and maybe some P7 adapter).

A5.4.2 Adapter Trimming

6. We first ensure that read 2 begins with GGTTAA. If it does, those bases are trimmed (hard clipped) to facilitate genomic alignment. Only reads with perfect matches are carried forward.

```
cutadapt \  
  
-g ^GGTTAA \  
  
-o PB_YFTF-PB_combined-trim1_R2_001.fastq.gz \  
  
-p PB_YFTF-PB_combined-trim1_R1_001.fastq.gz \  
  
--minimum-length 1 \  
  
--discard-untrimmed \  
  
-e 0 \  

```

```
--no-indels \
```

```
PB_YFTF-PB_combined_R2_001.fastq.gz \
```

```
PB_YFTF-PB_combined_R1_001.fastq.gz
```

Typically 70-90% of reads will pass this filter, although there may be sample-dependent variation.

7. Next, we re-examine the passing reads and trim any reads that end in the P7 adapter that was added during scCC library preparation. This step reduces the amount of non-genomic bases, which should accelerate alignment. Only a small fraction (5-10%) typically have any adapter sequence at all, so the majority of reads pass this filter.

```
cutadapt \
```

```
-a
```

```
AGAGACTGGCAAGTACACGTCGCACTCACCATGANNNNNNNNNATCTCGTATGCCGTCTTCTGCTTG \
```

```
-o PB_YFTF-PB_S1_L001_R2_001.fastq.gz \
```

```
-p PB_YFTF-PB_S1_L001_R1_001.fastq.gz \
```

```
--minimum-length 1 \
```

```
PB_YFTF-PB_combined-trim1_R2_001.fastq.gz \
```

```
PB_YFTF-PB_combined-trim1_R1_001.fastq.gz
```

Here, NNNNNNNNNN indicates where the index 1 sequence would be. The N's do not have to be replaced as cutadapt can tolerate ambiguous bases.

For running cellranger in next step, the input FASTQ filenames MUST conform to the following pattern:

```
[Sample Name]_S1_L00[Lane Number]_[Read Type]_001.fastq.gz
```

[Read Type] is R1 or R2. [Lane Number] can be varied; for simplicity, we use L001 here.

A5.4.3 Alignment

8. Now that our reads are trimmed, we are ready to align them to the genome. At the same time, we need to perform error-correction on our cell barcode and UMI sequences. We will use cellranger, as it can perform both whole-genome alignment and barcode curation at once. We need to specify the directory where the trimmed FASTQ files can be found. Note that this directory should be "flat", i.e have no subdirectories.

```
cellranger count \  
  
--id=PB_YFTF-PB_map_sccc \  
  
--fastqs=fastq_dir/ \  
  
--transcriptome=/opt/refdata-cellranger-GRCh38-3.0.0 \  
  
--sample=PB_YFTF-PB \  
  
--expect-cells=5000 \  
  
--nosecondary \  
  
--chemistry=SC3Pv2 \  

```

```
--localcores=16 \
```

```
--localmem=30
```

A few notes on this command:

- `--id=` will specify the output directory that will be created. This will be familiar to anyone who has worked with cellranger before.
- `--expect-cells=` is an estimate for how many cells are present in the library. This is much more important for scRNA-seq libraries than scCC libraries, and so can probably be safely excluded.
- `--nosecondary` skips the dimensionality reduction step of the cellranger pipeline. We are only concerned with mapping insertions to the genome.
- `--chemistry=` specifies the chemistry of the 10x kit. We prefer to explicitly specify this. Here, we used version 2 of the single cell 3' kit. The scCC workflow should also be immediately compatible with v3 chemistry.
- `--localcores=` and `--localmem=30` specify machine settings. Here, we used 16 cores and 30 GB of memory. These can be adjusted to fit your setup.

cellranger automatically performs barcode whitelisting and error-correction of UMIs, which are encoded in the program's output .bam file. The **CB** tag contains the read's verified cell barcode, and the **UB** tag denotes the corrected UMI. A full description of cellranger BAM tags can be found [here](#).

9. We then filter mapped reads for primary alignments, to eliminate multi-mapped reads:

```
samtools view \
```

```
-b -h -F 260 \
```

```
-o PB_YFTF-PB_map_scCC_uniq.bam \
```

```
PB_YFTF-PB_map_scCC/outs/possorted_genome_bam.bam
```

A5.4.4 Annotation

10. We then filter mapped reads for primary alignments, to eliminate multi-mapped reads:

```
samtools view \
```

```
-b -h -F 260 \
```

```
-o PB_YFTF-PB_map_scCC_uniq.bam \
```

```
PB_YFTF-PB_map_scCC/outs/possorted_genome_bam.bam
```

11. Now we will annotate reads with respect to the insertion site. This script checks each read to make sure that it maps next to the *piggyBac* insertion site motif TTAA. Remember, this part of read 1 was trimmed in step 5. By double checking that the read maps next to a genomic TTAA, we add an extra layer of specificity to the alignment. The sequence of the adjacent bases will also be annotated with the XZ tag. Reads that pass will be annotated with the insertion site coordinates in the XI tag and written to the output file.

```
python AnnotateInsertionSites.py \
```

```
--transposase PB \
```

```
-f \
```



```
PB_YFTF-PB_map_scCC_uniq.bam \
```

```
hg38.2bit \
```

```
PB_YFTF-PB_map_scCC_tagged.bam
```

You can provide a path to the .2bit file if your genome references are in another directory.

12. We perform one last quality control check on the processed scCC reads. Since scCC libraries involve intramolecular circularization, there is a small chance that concatamerization can occur. These would appear as singleton events where a cell barcode and UMI are linked to an insertion in a different cell. To guard against this, we require all insertions in a given cell (i.e. sharing the same cell barcode) to have at least two different UMIs each. This yields libraries with excellent specificity (see e.g. Figure 2.13B)

```
python UMIFilter.py \
```

```
-p 10x \
```

```
-i PB_YFTF-PB_map_scCC_tagged.bam \
```

```
--verbose \
```

```
-o PB_YFTF-PB_map_scCC_final.bam
```

13. Finally, we convert this BAM file to a (sorted) CCF file. The sorting step relies on bedops; see [here](#) for alternatives.

```
python BamToCallingCard.py \
```

```
-b CB \
```

```
-i PB_YFTF-PB_map_scCC_final.bam \  
  
-o PB_YFTF-PB_map_scCC_unsorted.ccf  
  
sort-bed PB_YFTF-PB_map_scCC_unsorted.ccf > PB_YFTF-  
PB_map_scCC_final.ccf
```

A5.4.5 Demultiplexing

14. At this point, PB_YFTF-PB_map_scCC_final.bam contains all insertions from all cells in our single cell library, both from the wild-type *piggyBac* transfectants and the YFTF-*piggyBac* transfectants. How can we determine which insertions came from which library? We will use the cell barcodes to further demultiplex the CCF file.

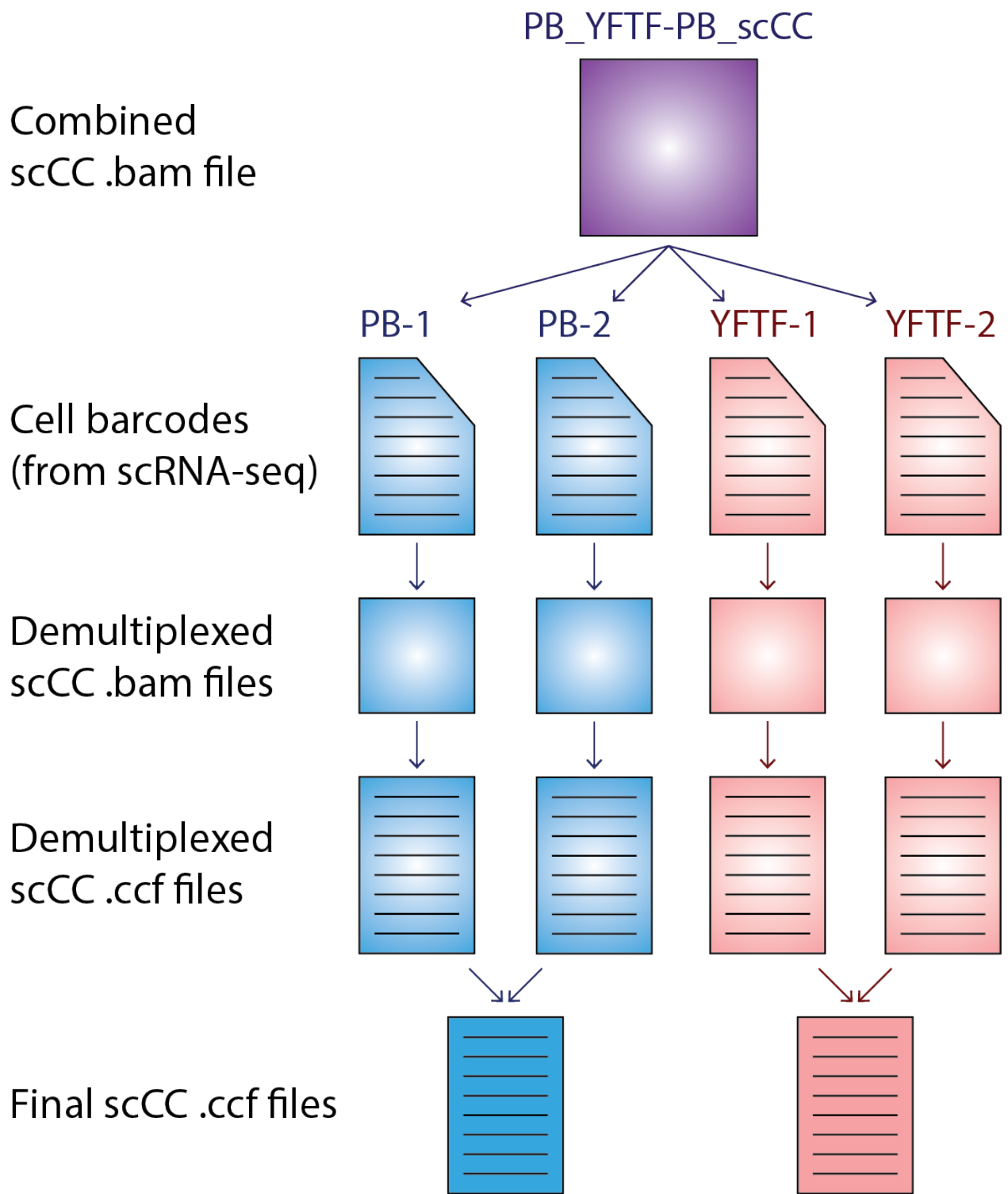


Figure A5.2: Demultiplexing scCC libraries.

This figure summarizes Steps 15-18, wherein we use the cell barcodes (obtained from the respective scRNA-seq libraries) to demultiplex the combined scCC .bam file, generate .ccf files from each library, and finally create master .ccf files for each condition (i.e. *piggyBac* and YFTF-*piggyBac* treatments).

15. In step 2, we prepared our 10x libraries by loading four wells of the Chromium chip: two for wild-type PB, and two for YFTF-PB. Let us call these libraries PB-1, PB-2, YFTF-1, and YFTF-2; further, assume that we have completed the scRNA-seq portions of the scCC workflow, including analysis with cellranger. For each of these four libraries, we can get a list of high-quality barcodes.

After cellranger has finished, each library's cell barcodes can be found in the following locations, respectively:

```
PB-1/outs/filtered_gene_bc_matrices/hg38/barcodes.tsv
```

```
PB-2/outs/filtered_gene_bc_matrices/hg38/barcodes.tsv
```

```
YFTF-1/outs/filtered_gene_bc_matrices/hg38/barcodes.tsv
```

```
YFTF-2/outs/filtered_gene_bc_matrices/hg38/barcodes.tsv
```

Starting with cellranger v3, the barcodes.tsv file may be gzipped (barcodes.tsv.gz). If that is the case, you will need to unzip before proceeding (`gunzip barcodes.tsv.gz`)

16. Since GEM generation was performed independently for each sample, there is a small chance (see note below) that the same cell barcode was captured more than once across libraries. This could, in theory, confound interpretation of TF binding, as a shared cell barcode may belong to cells of different types or states. While the effect of these is likely small, we recommend

discarding shared barcodes between libraries. The following command takes in a set of barcode files; for each, an output file is created containing the subset of unique barcodes found only in the respective input file.

```
python UMIFilter.py \  
  
-i PB-1/outs/filtered_gene_bc_matrices/hg38/barcodes.tsv \  
  
PB-2/outs/filtered_gene_bc_matrices/hg38/barcodes.tsv \  
  
YFTF-1/outs/filtered_gene_bc_matrices/hg38/barcodes.tsv \  
  
YFTF-2/outs/filtered_gene_bc_matrices/hg38/barcodes.tsv \  
  
-o PB-1_unique_barcodes.txt \  
  
PB-2_unique_barcodes.txt \  
  
YFTF-1_unique_barcodes.txt \  
  
YFTF-2_unique_barcodes.txt
```

Note that this command can take in any number of input files (2, 3, 4, 5, etc.). The only requirements are: (1) a matched list of output files is provided; and (2) the input barcode files contain one barcode per line.

For the curious: the probability of a shared barcode (i.e. barcode collision) between two 10x scRNA-seq libraries is quite small, but is dependent on library size. For two libraries of 5,000 cells each, the probability is < 1%. As the number of libraries increases, the probability of collision increases approximately. We have filtered unique cell barcodes across as many as six libraries and have discarded no more than 5% of total cell barcodes.

17. Now that we have a list of cell barcodes unique to each library, we can demultiplex our calling card BAM file. The example show is for PB_1 but can be generalized to all samples.

```
python FilterBAMByBarcodes.py \  
  
-i PB_YFTF-PB_map_scCC_final.bam  
  
-b PB-1_unique_barcodes.txt \  
  
-o PB-1_scCC_final.bam
```

We can now convert this BAM file to CCF output.

```
python BamToCallingCard.py \  
  
-b CB \  
  
-i PB-1_scCC_final.bam \  
  
-o PB-1_scCC_unsorted.ccf
```

We can also combine the two wild-type *piggyBac* libraries into a single, sorted CCF file. (The second step requires bedops; see [here](#) for alternative sorting commands).

```
cat PB-1_scCC_unsorted.ccf PB-2_scCC_unsorted.ccf | sort-bed - >  
PB_scCC_final.ccf  
  
cat YFTF-PB-1_scCC_unsorted.ccf YFTF-PB-2_scCC_unsorted.ccf |  
sort-bed > YFTF-PB_scCC_final.ccf
```

18. At last, we have a CCF file containing all insertions across all (unique) cells in a scCC experiment. This file can be further visualized on the WashU Epigenome Browser and used as input for peak calling. Here is an example of scCC CCF output.

chr1	29884	29888	3	+	GCATGATCAGACGTAG-1
chr1	30355	30359	4	-	CAGCTGGTCGCAAAC-1
chr1	32116	32120	11	-	GTGTGCGAGCTTCGCG-1
chr1	32303	32307	674	+	GTCGTAAAGGTAGCTG-1
chr1	33031	33035	2	-	TTAGTTCTCAACACTG-1
chr1	33031	33035	21	+	GCAATCAGTGGTTTCA-1
chr1	33031	33035	25	+	GCACTCTAGTAGCCGA-1
chr1	33031	33035	98	+	GTTTCTACAGACGCAA-1
chr1	33169	33173	26	-	CAAGAAAGTACAGCAG-1
chr1	34572	34576	4	-	CGTTCTGCAAATTGCC-1

A5.4.6 Notes

19. In the course of analyzing your scRNA-seq data, you may find biologically meaningful clusters and may wish to identify differentially bound loci. Let us suppose that in the your analysis of the YFTF-PB transfectants, you find two clusters of cells (Alfa and Bravo) and wish to stratify insertions specific to each cluster. If the cell barcodes in each cluster are in Barcodes_Alfa.txt and Barcodes_Bravo.txt, we can directly filter insertions from the YFTF-PB CCF file, instead of going back to the BAM file.

```
python FilterCCFByBarcodes.py \

-i YFTF-PB_scCC_final.ccf \

-b Barcode_Alfa.txt \

-o YFTF-PB_scCC_Alfa.ccf

python FilterCCFByBarcodes.py \

-i YFTF-PB_scCC_final.ccf \
```

```
-b Barcode_Bravo.txt \
```

```
-o YFTF-PB_scCC_Bravo.ccf
```

Note that if the input CCF file is sorted, the output file should automatically be sorted as well.

20. Currently, the scCC pipeline **does not** support 10x scRNA-seq libraries merged using *cellranger aggr*. Guidance on this will be provided in the future.

Appendix 6: Calling Peaks on *piggyBac* **Calling Card Data**

(A version of this appendix has been published on protocols.io:

<https://doi.org/10.17504/protocols.io.bb9xir7n>)

A6.1 Abstract

This protocol describes how to call peaks on mammalian calling card data using either undirected, or transcription factor fusions, to the *piggyBac* transposase. It is applicable for both bulk as well as single cell calling card data.

A6.2 Guidelines

Please make sure you have installed the required software and packages (see Materials section).

This protocol describes how to go from a CCF file, the processed output of a calling cards experiment, to a set of peaks enriched for transcription factor (TF) binding. If you are unfamiliar CCF files, we recommend reading our [bulk](#) or [single cell](#) calling card data processing protocols first. We assume that you have performed either bulk calling cards (with sufficient replicates) or single cell calling cards with undirected *piggyBac* (to map BRD4 binding) and, optionally, with a TF-*piggyBac* fusion for your favorite TF (YFTF). To identify peaks in BRD4 binding, you should prepare a single CCF containing insertions from all replicates of undirected *piggyBac* calling cards. To call peaks on YFTF peaks, you will need two CCF files: one from all replicates of undirected *piggyBac* experiments, and one from all replicates of YFTF-*piggyBac* experiments.

A6.3 Materials

This protocol requires a FASTA file for your genome of interest (e.g. hg38.fa) and the following script:

- kmer.cc

In addition, this workflow relies on some calling card-specific scripts, which use Python 3. It is recommended that your Python installation be relatively up-to-date (i.e. ≥ 3.4). To check your python version, type:

```
python -V
```

You will need the following Python modules:

- [numpy](https://numpy.org/)
- [pandas](https://pandas.pydata.org/)
- [scipy](https://www.scipy.org/)
- [statsmodels](https://statsmodels.org/)
- [pybedtools](https://pybedtools.readthedocs.io/)
- [astropy](https://astropy.org/)

All of these packages are available on PyPI and can be installed via pip:

```
pip install numpy pandas scipy statsmodels pybedtools astropy
```

(If Python3 is not the default on your system, replace pip with pip3)

These are the calling card-specific scripts you will need, all of which are available on [GitHub](https://github.com):

- SegmentCCF.py
- CCFIdeogram.py
- BBPeakCaller_TF.py
- BBPeakCaller_BRD4.py

The following programs are optional, but highly recommended:

- [bedtools](#) (≥ 2.27)
- [bedops](#) (≥ 2.4)

A6.4Steps

A6.4.1 Preprocessing

1. Before calling peaks on calling card data, it is useful to create a file listing the location of every TTAA tetramer in the genome. The *piggyBac* transposase inserts almost exclusively into this motif. Moreover, we use the presence of a TTAA adjacent to a mapped read as an internal quality check when creating CCF files. This section will walk you through how to quickly find all TTAAAs in a genome.

2. Compile the kmer.cc program as follows:

```
g++ kmer.cc -o kmer
```

The result should be a C++ executable in your directory called **kmer**.

3. This program takes as input a FASTA file and k-mer and outputs a BED file of all exact matches to that k-mer. Here we use it to find all exact matches to the 4-mer TTAA.

Download or copy to your working directory a FASTA file of your genome of interest. Using the latest human genome build as an example:

```
./kmer hg38.fa TTAA > hg38_TTAA.bed
```

The file hg38_TTAA.bed now lists all TTAA's in hg38.fa.

It is important that the FASTA file that you use in this step is the same FASTA sequence used for aligning calling card reads. For example, if you mapped to a repeat-masked genome earlier, you should supply a repeat-masked FASTA file here.

4. (Optional) More recent builds of the human and mouse genomes contain unplaced contigs and alternate haplotypes. You may be interested in restricting your analysis to the "canonical" chromosomes (e.g. 1-22, X & Y for humans). Here is a simple way to filter only "canonical" TTAA's:

```
grep -v '_' hg38_TTAA.bed > hg38_TTAA_canon.bed
```

(If you are being nitpicky, this file will include TTAA's on the mitochondrial chromosome, but we have not found this to be a problem for peak calling).

If you filter only "canonical" TTAA's, it is important that you also filter your CCF file so it contains only insertions mapping to "canonical" chromosomes. The above command can be used to do so. If you do not do this, you may get "divide by zero" errors in subsequent steps.

5. The TTAA file only needs to be generated once per genome. Afterwards, all experiments using the same reference genome can use the same TTAA file.

A6.4.2 Bayesian Blocks

6. The core of calling peaks in calling card data is Bayesian blocks. This algorithm, originally developed in astrophysics, segments one-dimensional datasets into regions of piecewise-constant density. We use it to initially partition the genome into intervals, where each interval contains a constant rate of *piggyBac* insertions. These intervals are referred to as **blocks**; two adjacent blocks are characterized by different insertion rates and, accordingly, different insertion densities. One attractive reason for using Bayesian blocks is that it can find a mathematically optimal partition of the data into blocks. Peak calling then proceeds by testing each block to see if it contains more insertions than expected by some background model.

This much overview of Bayesian blocks is sufficient to understand peak calling. For more details, we recommend reading the original paper ([Scargle et al. 2013](#)) or this [blog post by Jake VanderPlas](#).

7. We generate a list of blocks from CCF files, but these files must first be sorted. Here are three ways to sort CCF files, in order of preference:

Using bedops:

```
sort-bed sample.ccf > sample_sorted.ccf
```

Using bedtools:

```
bedtools sort -i sample.ccf > sample_sorted.ccf
```

Using the standard shell sort command:

```
sort -k1V -k2n -k3n sample.ccf > sample_sorted.ccf
```

For the remainder of this protocol, we assume your CCF files are already sorted.

A6.4.3 Calling BRD4 Peaks

8. Here we will describe how to call BRD4 peaks from undirected *piggyBac* insertions. Our sample file is HCT-116_PBase.ccf, which contains insertions from 10 replicates of bulk RNA calling cards in the HCT-116 cell line. This file contains 1.5 million insertions:

```
wc -l HCT-116_PBase.ccf  
  
1521048 HCT-116_PBase.ccf
```

9. We start by creating creating blocks from the CCF file. To do this, we use the SegmentCCF.py script:

```
python SegmentCCF.py HCT-116_PBase.ccf | sed -e '/^\s*$/d' >  
HCT-116_PBase.blocks
```

The output file is a BED-formatted list of blocks inferred by Bayesian blocks. The piped sed command simply removes blank lines.

You may see a warning about false positive rates for event data, as well as possibly a dividing by zero warning. These are automatically generated by astropy, the library which contains the Bayesian blocks algorithm and can be safely ignored. We have successfully called peaks with the blocks generated despite these warnings.

Segmenting the CCF file is often the most time-consuming step. The Bayesian blocks algorithm has quadratic runtime complexity. If one CCF file has twice as many insertions as another, the former is expected to take roughly four times longer to segment as the latter.

10. We then provide the CCF and blocks files, as well as the TTAA file, to `BBPeakCaller_BRD4.py`, which tests each block for statistical significance. This script performs the following steps:

1. Divide the number of insertions by the number of TTAAAs in the TTAA file. This defines a global rate parameter (r) under a null model assuming a uniform distribution of insertions.
2. For each block b , count the number of TTAAAs in b and multiply it by r . This value specifies the expected number of insertions in b if insertions were uniformly distribution (denoted λ_b).
3. For each block b , let x_b be the number of observed insertions in the block. The script then performs a one-tailed Poisson significance test on the block. This is calculated as the probability of observing x_b insertions or more in the block parametrized by a Poisson distribution with expected value λ_b .
4. Multiple hypothesis correction is performed (based on user preferences).
5. Finally, blocks that pass multiple hypothesis correction are polished and written to file.

The output file is in BED format.

11. `BBPeakCaller_BRD4.py` takes four required positional arguments: CCF file, blocks file, TTAA file, and output filename. An example command would look like this:

```
python BBPeakCaller_BRD4.py HCT-116_PBase.ccf HCT-116_PBase.blocks hg38_TTAA_canon.bed HCT-116_PBase_peaks.bed
```

This command **WILL NOT RUN** as written because it does not specify how peaks should be thresholded. See Step 12 for details.

12. Required flag:

BBPeakCaller_BRD4.py has one required flag which specifies the statistical threshold for filtering blocks. There are two options for this: a straight p-value cutoff or with a multiple hypothesis correction method. The former is specified by the `-p` flag; the value supplied to it must be the $-\log_{10}$ transformation of the desired p-value.

Example: select only those blocks with $p < 10^{-9}$

```
-p 9
```

Alternatively, you can control for multiple hypotheses at a desired alpha level. This is specified by the `-a` flag and the value supplied is **not** transformed. If this option is used, you must supply a method (`-m`) of multiple hypothesis correction. Valid methods are listed [here](#).

Example: Bonferroni correction at an alpha of 0.05

```
-a 0.05 -m bonferroni
```

Example: Benjamini-Hochberg correction at false discovery rate of 10%

```
-a 0.1 -m fdr_bh
```

Remember: You must use **EITHER** `-p` **OR** `-a -m` for the program to run.

BRD4 peaks may require choosing a p-value cutoff that is more stringent than, for example, Bonferroni correction. This appears to scale with size of the dataset: with more insertions, a more

stringent cutoff is needed. To guide settling on an optimal p-value, we recommend calling peaks at a variety of cutoffs, visualizing CCF data and peak files (eg. on the WashU Epigenome Browser), and choosing a value whose peak boundaries reasonably accord with insertion densities.

Optional flags:

Multiple significant blocks may occur in close proximity to one another. If you want to merge these into larger peaks, you can specify a **distance (-d)**. Significant blocks within this distance will be merged together.

Example: merge blocks within 12.5 kb

```
-d 12500
```

Finally, while the primary output of BBPeakCaller_BRD4.py is a list of peaks in BED format, an **intermediate filename (-i)** can be supplied to write information about each block and its p-value. This file will be written in CSV format.

Example: write an intermediate file for the HCT-116 PBase dataset

```
-i HCT-116_PBase_intermediate.csv
```

13. The blocks file, in addition to being used to call peaks, can also be used to calculate normalized insertion densities across the genome. This is done using the CCFIdeogram.py script, which takes as input a CCF file and a blocks file and outputs a bedgraph file. Each entry in the bedgraph file is a block and the numerical value for each block is the number of insertions in that block per million mapped insertions per kilobase (IPKM).

```
python CCFIdeogram.py HCT-116_PBase.ccf HCT-116_PBase.blocks  
HCT-116_PBase.bedgraph
```

This script is named after ideograms because the resulting bedgraph files, when visualized as densities, create banding patterns that resemble karyotyped chromosomes.

14. Let's put this all together. Here is the command to recreate our analysis from our recent preprint:

```
python BBPeakCaller_BRD4.py -p 30 -d 12500 HCT-116_PBase.ccf  
HCT-116_PBase.blocks hg38_TTAA_canon.bed HCT-116_PBase_peaks.bed
```

This generates a BED file containing nearly 2000 peaks.

```
wc -l HCT-116_PBase_peaks.bed
```

```
1939 HCT-116_PBase_peaks.bed
```

15. Here is the output of our sample analysis as visualized on the WashU Epigenome Browser.

The top track is the raw CCF data. The next track is the per-block insertion densities as calculated by CCFIdeogram.py. The third track is the same as the second but with in-browser smoothing (15 px). We then show peak boundaries at a variety of p-value thresholds, in order of increasing stringency. Notice how peaks grow, merge, shrink, and vanish at different cutoffs.

The dark blue peaks track corresponds to the threshold used in the previous step. BRD4 and H3K27ac data, marks of enhancers and super-enhancers, are shown for reference.

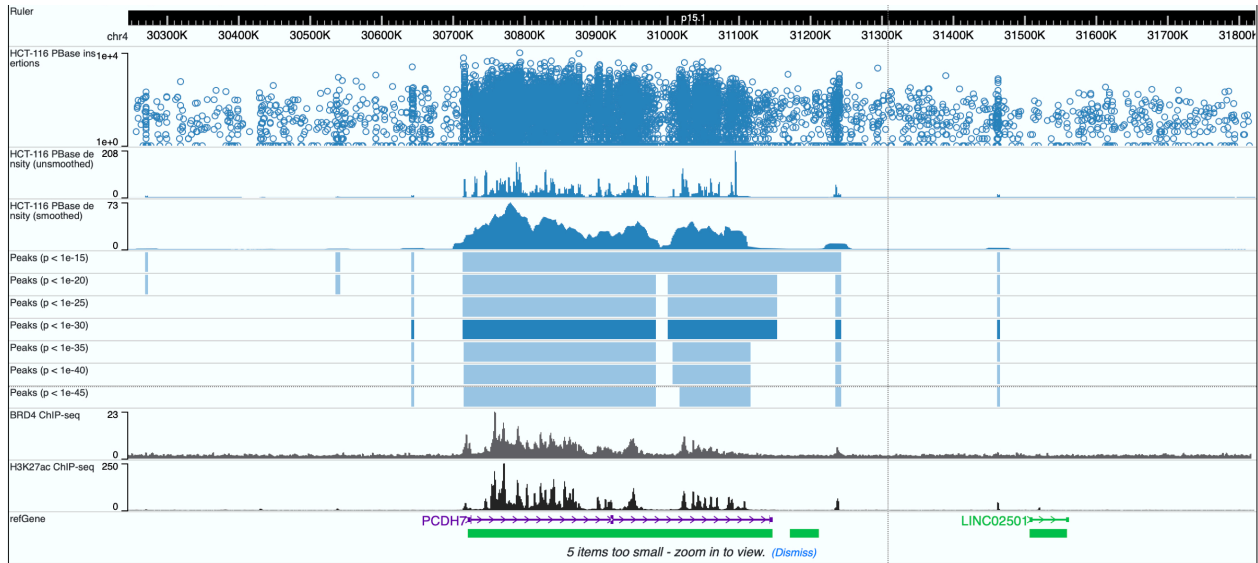


Figure A6.1: Example of a BRD4-directed calling cards peak.

A6.4.4 Calling TF Peaks

16. Peak calling on TF-directed *piggyBac* insertions is similar to calling BRD4 peaks. The major distinction is in the choice of background model. With undirected *piggyBac*, the background is modeled as a uniform distribution of the observed number of insertions. For TF-directed *piggyBac*, however, the background is the undirected *piggyBac* dataset from the same cell line or system.

For this example, we will use HCT-116_SP1-PBase.ccf. This file was generated from 10 replicates of bulk RNA calling cards with SP1-*piggyBac* in HCT-116 cells. The control file is the undirected *piggyBac* data from the same system, i.e. HCT-116_PBase.ccf.

```
wc -l HCT-116_SP1-PBase.ccf
```

```
410588 HCT-116_SP1-PBase.ccf
```

17. As before, we start by creating creating blocks from the CCF file:

```
python SegmentCCF.py HCT-116_SP1-PBase.ccf | sed -e '/^\s*$/d' >
HCT-116_SP1-PBase.blocks
```

The same notes from Step 9 apply here as well.

18. We then provide the CCF and blocks files, as well as the background CCF file, to `BBPeakCaller_TF.py`, which tests each block for statistical significance. This script performs the following steps:

1. Divide the number of insertions in the TF CCF file by the number of insertions in the background CCF file. This defines a global scaling parameter (s). This enables us to account for library size differences between the TF-directed and undirected control libraries.
2. For each block b , count the number of insertions from the background CCF file in b and multiply it by s , then add a pseudocount c . This value specifies the normalized expected number of insertions in b from the undirected control experiment (denoted λ_b).
3. For each block b , let x_b be the number of insertions from the TF-directed CCF file. The script then performs a one-tailed Poisson significance test on the block. This is calculated as the probability of observing x_b insertions or more in the block parametrized by a Poisson distribution with expected value λ_b .
4. Multiple hypothesis correction is performed (based on user preferences).
5. Finally, blocks that pass multiple hypothesis correction are polished and written to file.

The output file is in BED format.

19. BBPeakCaller_TF.py takes four required positional arguments: TF-directed CCF file, TF-directed blocks file, undirected CCF file, and output filename. An example command would look like this:

```
python BBPeakCaller_TF.py HCT-116_SP1-PBase.ccf HCT-116_SP1-  
PBase.blocks HCT-116_PBase.ccf HCT-116_SP1-PBase_peaks.bed
```

This command **WILL NOT RUN** as written because it does not specify how peaks should be thresholded. See Step 20 for details.

20. Required flag:

BBPeakCaller_TF.py has one required flag which specifies the statistical threshold for filtering blocks. There are two options for this: a straight p-value cutoff or with a multiple hypothesis correction method. The former is specified by the `-p` flag; the value supplied to it must be the `-log10` transformation of the desired p-value.

Example: select only those blocks with $p < 10^{-9}$

```
-p 9
```

Alternatively, you can control for multiple hypotheses at a desired alpha level. This is specified by the `-a` flag and the value supplied is **not** transformed. If this option is used, you must supply a method (`-m`) of multiple hypothesis correction. Valid methods are listed [here](#).

Example: Bonferroni correction at an alpha of 0.05

```
-a 0.05 -m bonferroni
```

Example: Benjamini-Hochberg correction at false discovery rate of 10%

```
-a 0.1 -m fdr_bh
```

Remember: You must use **EITHER** `-p` **OR** `-a -m` for the program to run.

Optional flags:

Multiple significant blocks may occur in close proximity to one another. If you want to merge these into larger peaks, you can specify a **distance** (`-d`). Significant blocks within this distance will be merged together.

Example: merge blocks within 250 bp

```
-d 250
```

Peaks are composed of one or more blocks. Bayesian blocks draws block boundaries halfway between adjacent insertions. This can, in some cases, lead to unnecessarily wide peaks.

The **refine** (`-r`) flag constrains the block edges so that they start and end at insertions. This, in turn, helps increase the resolution of peak calls.

Peaks can be further filtered based on a size threshold. You can specify a **minimum** (`-n`) and **maximum** (`-x`) size bound on reported peaks.

Example: report all peaks less than 5 kb in length

```
-x 5000
```

Example: report only peaks between 100 and 500 bp in length

```
-n 100 -x 500
```

TF-*piggyBac* fusions redirect, but do not abolish, *piggyBac*'s natural affinity for BRD4. This is why TF-directed experiments must use an undirected calling card experiment as a control. This can also pose a challenge for peak calling: whereas most TF's have narrow, sharp peaks, BRD4 can bind much broader stretches of the genome. Peak calling may not completely eliminate this signal, which is typically reflected in large, but statistically significant, peaks. Broad peaks can also occur in the shoulder regions flanking a TF binding site, likely from "spillover" of insertions by the increased local concentration of transposase.

A simple way to increase peak specificity is to threshold on peak size. In our experience, in a number of cell lines with a variety of TFs, 5 kb is a reasonable upper bound for filtering peaks. This threshold is greater than the median peak sizes we have observed, which lets us preserve the majority of called peaks.

By default, the pseudocount added to all peaks is 1. This **value** (**-c**) can be changed if desired.

Example: use a pseudocount of 0.1

```
-c 0.1
```

Finally, while the primary output of `BBPeakCaller_TF.py` is a list of peaks in BED format, an **intermediate filename** (**-i**) can be supplied to write information about each block and its p-value. This file will be written in CSV format.

Example: write an intermediate file for the HCT-116 SP1-PBase dataset

```
-i HCT-116_SP1-PBase_intermediate.csv
```

21. As before, TF-directed CCF files can also be used to create insertion density tracks, following the instructions in Step 13.

22. Let's put this all together. This command calls peaks from SP1-PBase at a false discovery rate of 5%, merging significant blocks within 250 bp, refining block edges, and outputting all peaks less than 5 kb in length:

```
python BBPeakCaller_TF.py -a 0.05 -m fdr_bh -d 250 -r -x 5000
HCT-116_SP1-PBase.ccf HCT-116_SP1-PBase.blocks HCT-116_PBase.ccf
HCT-116_SP1-PBase_peaks.bed
```

This generates a BED file containing around 5600 peaks.

```
wc -l HCT-116_SP1-PBase_peaks.bed
```

```
5615 HCT-116_SP1-PBase_peaks.bed
```

23. Here is the output of our sample analysis as visualized on the WashU Epigenome Browser. The top track are the undirected insertions, followed by the SP1-directed insertions. The next track is the per-block insertion densities for the SP1-PBase data with in-browser smoothing (3 px). Finally, we plot peak boundaries at a variety of p-value thresholds, in order of decreasing stringency. Notice how peaks grow, merge, shrink, and vanish at different cutoffs. The dark blue peaks track corresponds to the threshold used in the previous step. The dark blue track shows all significant peaks at 5% FDR less than 5 kb in length. The light blue track shows all peaks without size restriction. Notice how imposing a maximum peak size filters out potentially artifactual peaks.

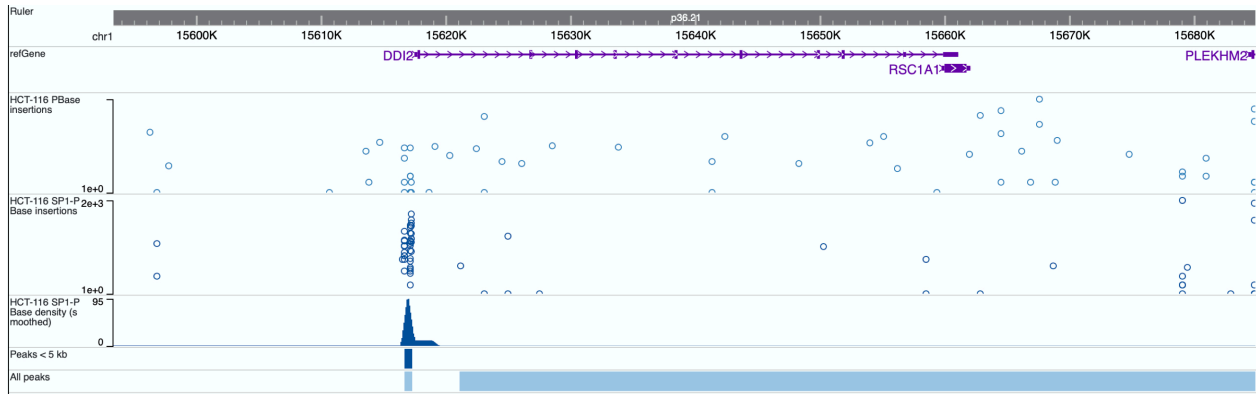


Figure A6.2: Example of an SP1-directed calling cards peak.

A6.4.5 Final Thoughts

24. BBPeakCaller_TF.py can also be used to call differential peaks between two different undirected or TF-directed calling card datasets, such as between cell types or treatments. In this use case, one sample serves as the "background" for the other. For example, let's imagine that we have done one undirected (a.k.a. BRD4) calling cards experiment on cells treated with DMSO and another treated with dexamethasone. To identify peaks that are enriched in the dexamethasone condition, we could call:

```
python BBPeakCaller_TF.py -p 9 -d 12500 DEX.ccf DEX.blocks
DMSO.ccf DEX_peaks.bed
```

Since these tests are one-tailed, to find peaks that are enriched in the other direction, i.e. in the DMSO condition, we simply swap the datasets:

```
python BBPeakCaller_TF.py -p 9 -d 12500 DMSO.ccf DMSO.blocks
DEX.ccf DMSO_peaks.bed
```

25. The output of these peak calling scripts are BED files listing peak coordinates only. They do not annotate the peaks themselves. There are a number of programs for secondary analyses:

- Peaks can be annotated with overlapping and nearest genes using [HOMER](#).
- Peaks can be connected to putative genetic targets using [GREAT](#).
- *De novo* motif analysis on peaks can be done with either [HOMER](#) or [MEME](#).

26. For guidance on how to visualize calling card data, see our Appendix 7. Documentation is also available from the [WashU Epigenome Browser](#).

Appendix 7: Visualizing Calling Card Data on the WashU Epigenome Browser

(A version of this appendix has been published on protocols.io:

<https://doi.org/10.17504/protocols.io.bca8ishw>)

A7.1 Abstract

This document explains how to visualize calling card insertions as well as density and peak tracks on the WashU Epigenome Browser.

A7.2 Guidelines

Please make sure you have installed the required software and packages (see Materials section).

This protocol describes how to visualize calling card data, such as raw insertions (in a CCF file), insertion densities (bedgraph file), and peaks (BED file). Ideally, you will be able to store files on a publicly-accessible webserver. This requires [enabling Cross-Origin Resource Sharing \(CORS\)](#). If this is not possible, data files can also be directly uploaded from your computer to the browser. This approach is described in the second half of the protocol.

A7.3 Materials

If you are hosting files on a server, you will require the following software package:

- [htslib](#)

A7.4 Steps

A7.4.1 Introduction

1. During the course of a calling card experiment, or after all the analysis has been said and done, you may want to visually inspect the data. These can be the raw insertions (stored as a CCF file),

insertion densities (as a bedgraph file), or the peak boundaries (as a BED file). Instructions for generating these kinds of files can be found in the following protocols:

- [Processing Bulk RNA Calling Card Sequencing Data](#) (Appendix 4)
- [Processing Single Cell Calling Card Sequencing Data](#) (Appendix 5)
- [Calling Peaks on piggyBac Calling Card Data](#) (Appendix 6)

2. The WashU Epigenome Browser natively supports visualizing CCF files as a calling card track. In addition, it supports standard genomic file formats such as BED, bedgraph, bigWig, and HiC. This allows us to compare and contrast calling card data to popular genomic assays like ChIP-seq, DNase-seq, ATAC-seq, and Hi-C.

There are two ways load data onto the browser:

1. Hosting your data on a publicly-accessible server and letting the browser fetch the data.
2. Directly uploading text files to the browser.

The first method is preferred: track information and preferences can be stored in a simple text file, which can then be shared with collaborators. This saves the effort and cost of moving around potentially large datasets. Instead, the data remains in one place and can be accessed by anyone from anywhere.

The second option can be used if a public server is not available to you, or if you want to rapidly look at files that are already on your computer.

3. For this protocol, we will use the following files:

- HCT-116_PBase.ccf, a file of raw calling card insertions
- HCT-116_PBase.bedgraph, a file of calling card insertion densities across the genome
- HCT-116_PBase_peaks.bed, a file of peaks inferred from calling card data

A7.4.2 Uploading Data from an External Server

4. BED, bedgraph, and CCF files are all text-based formats. Before being hosted on an external server, these files must be compressed and indexed. This allows for fast, random-access to the data. For this step, make sure you have installed htlib (see Materials).

To compress the file:

```
bgzip HCT-116_PBase.ccf
```

This will compress the original file, creating HCT-116_PBase.ccf.gz

To index the compressed file:

```
tabix -p bed HCT-116_PBase.ccf.gz
```

This will create an index file, HCT-116_PBase.ccf.gz.tbi

Both HCT-116_PBase.ccf.gz and HCT-116_PBase.ccf.gz.tbi should be copied to a publicly-accessible webserver. These steps also apply to bedgraph and BED files.

5. Next, we create a JSON file. JSON is a generic standard describing objects and their properties. Here, we use it to describe each track, its data, and any options we may want to customize. The JSON file is plain text; it is recommended that you use a text editor (e.g. Sublime, Atom, Visual Studio Code, vim/emacs, etc.) instead of a word processor (e.g. Microsoft Word) to create this file.

We will create HCT-116_PBase.json for our data. Here is the structure of a JSON file for a single track depicting calling card data:

```
[ { "type": "callingcard",  
  "url": "https://htcf.wustl.edu/files/xX18ZAXy/HCT-  
116_PBase.ccf.gz", "name": "piggyBac insertions",  
  "showOnHubLoad": "true", "options": { "color": "#3182bd",  
  "height": 100, "logScale": "log10",  
  "markerSize": 3,  
  "opacity": [100],  
  "show": "all", "sampleSize": 1000,  
  },  
  },  
  ],
```

All tracks in the JSON file must be between the square brackets [...]. Curly braces {...} and commas separate individual tracks, and within a track, another set of curly braces may be used to specify grouped options. Let's consider each of these entries in turn:

```
"type": "callingcard",
```

This specifies the track type, which in this case is a calling card track.

```
"url": "https://htcf.wustl.edu/files/xX18ZAXy/HCT-  
116_PBase.ccf.gz",
```

This points to the web address where your data are stored. Note that this must point to the compressed file. The browser will automatically locate the index file. (If the index file is not present, it will throw an error).

```
"name": "piggyBac insertions",
```

This is the track label.

```
"showOnHubLoad": "true",
```

This specifies whether the track should be displayed immediately after the JSON file has been uploaded. The default value is "false."

There are also a number of options that can be specified to customize the track appearance.

While all of them can be changed after the tracks have been loaded, specifying them in the JSON file helps to record and reproduce settings. Note that these only need to be set if you wish to override defaults.

```
"color": "#f4916c",
```

This sets the color of the individual calling card markers. The default is blue.

```
"height": 100,
```

This sets the height of the track in pixels. The default is 40.

```
"logScale": "log10",
```

This transforms the y-axis from a linear scale to a logarithmic scale. By default, the track uses a linear scale, but for calling card experiments we recommend using a logarithmic scale.

```
"markerSize":3,
```

This is the radius of the marker in pixels. The default value is 3.

```
"opacity":[100],
```

This specifies the opacity of the markers, which can help emphasize data-dense regions. This takes an integer value between 0 and 100. The value supplied must be in square brackets. The default value is 100.

```
"show":"all", "sampleSize":1000,
```

This pair of options specify whether the browser should display all data points, or simply a random subsample. If there are many (e.g. thousands) of data points in view, it can slow down your web browser. Subsampling can help preserve a global representation of your data while still remaining responsive and memory-friendly. The "show" option can take two values: "all" (default) or "sample." The former draws all data points. If the latter is specified, then the "sampleSize" option is applied. This number determines how many points to subsample and is set to 1000 by default.

6. We can also add the bedgraph and BED files to the JSON:

```
[ { "type":"callingcard",  
  "url":"https://htcf.wustl.edu/files/xXl8ZAXy/HCT-  
116_PBase.ccf.gz", "name":"piggyBac insertions",  
  "showOnHubLoad":"true", "options":{ "color":"#3182bd",  
  "height":100, "logScale":"log10",  
  
  "markerSize":3,
```



```

"opacity":[100],

"show":"all", "sampleSize":1000,

},

},

{ "type":"bedgraph",
"url":"https://htcf.wustl.edu/files/xX18ZAXy/HCT-
116_PBase.bedgraph", "name":"HCT-116 PBase density",
"showOnHubLoad":"true", "options":{ "color":"#3182bd",
"height":50,

},

},

{ "type":"bed",
"url":"https://htcf.wustl.edu/files/xX18ZAXy/HCT-116_PBase.bed",
"name":"HCT-116 PBase peaks", "showOnHubLoad":"true", "options":
{ "color":"#3182bd",

"height":20, "displayMode":"density",

}


},


]


```

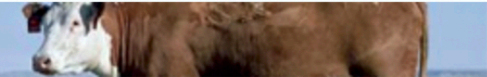
Many options have the same names and settings as the calling card track. A complete list of options for bedgraph and BED tracks can be found [here](#).

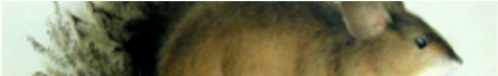
7. Now we can upload this file to the browser and visualize it. Go to the [homepage](#) and select the appropriate genome:

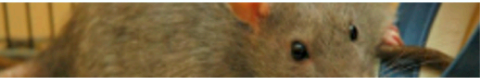
Human 

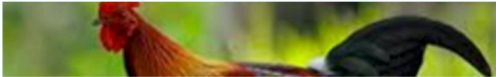
Chimp 


Rhesus 


Cow 


Mouse 


Rat 

Chicken 

Zebrafish 

Fruit fly 

C.elegans 

Arabidopsis 

hg19
 hg38

Go ⇒

Figure A7.1: WashU Epigenome Browser splash page.

8. Once the genome has loaded, click on the blue "Tracks" button and select "Custom tracks."

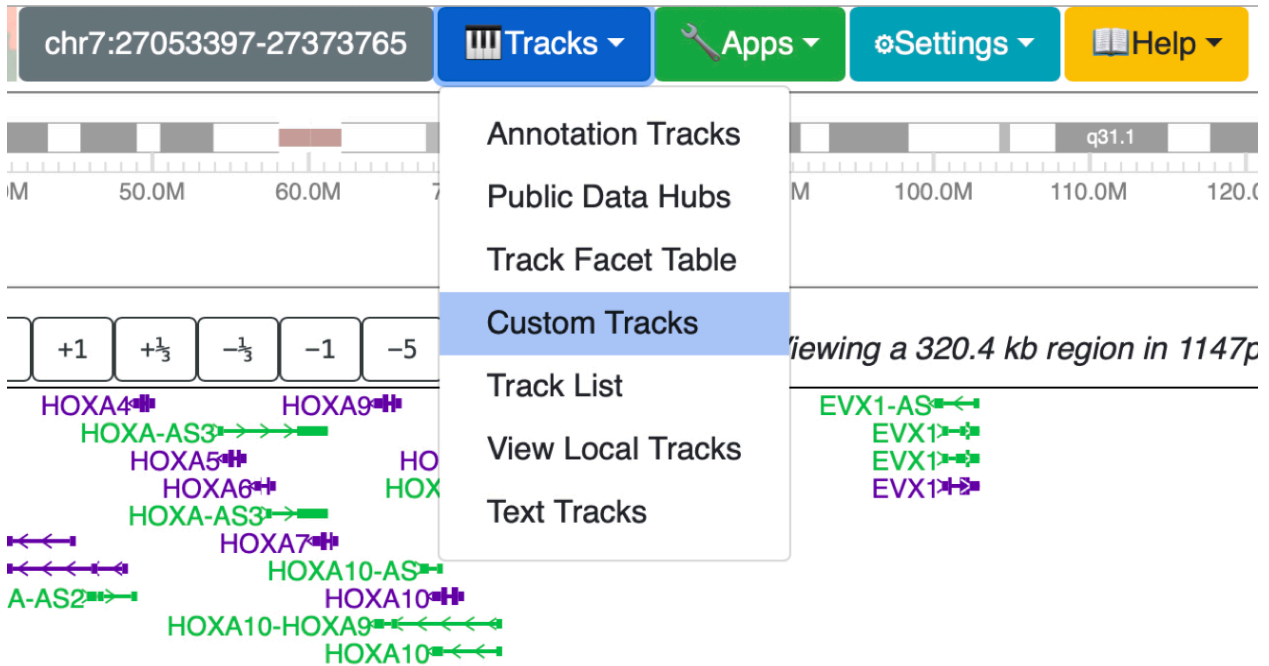


Figure A7.2: Custom tracks pane.

9. Select the second tab, "Add Custom Data Hub."



Figure A7.3: Load custom data hub.

Clicking on the second dialog box ("Choose datahub file") will open a filesystem navigator window. Find the JSON file on your system and upload it.

Upon successful upload, you should see a table listing the uploaded tracks:

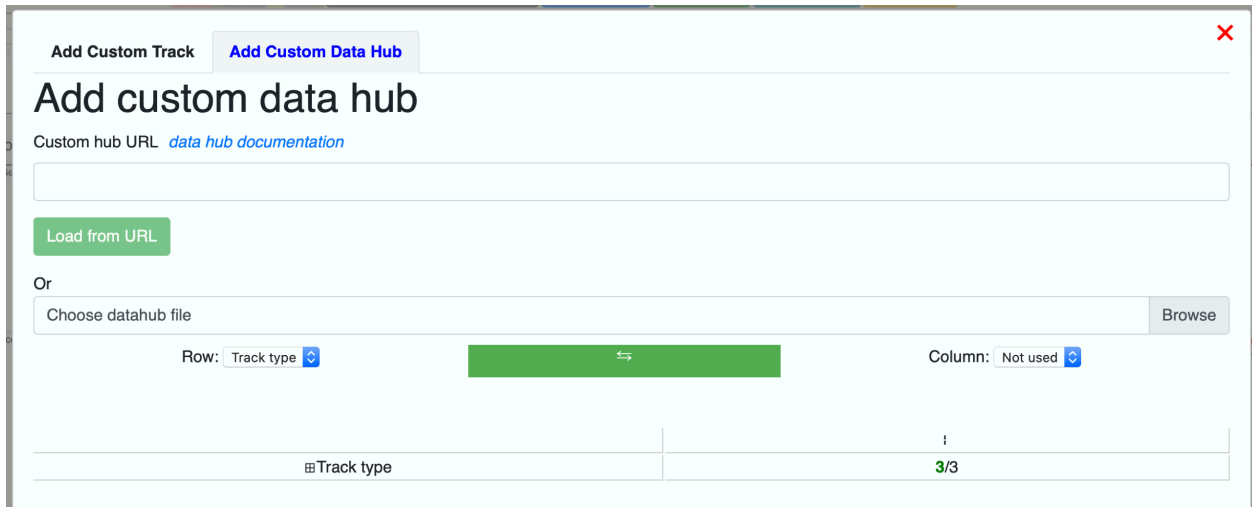


Figure A7.4: Successfully uploaded data hub.

10. After closing the custom track pane (with the red X in the upper right corner), all three tracks should be visible:

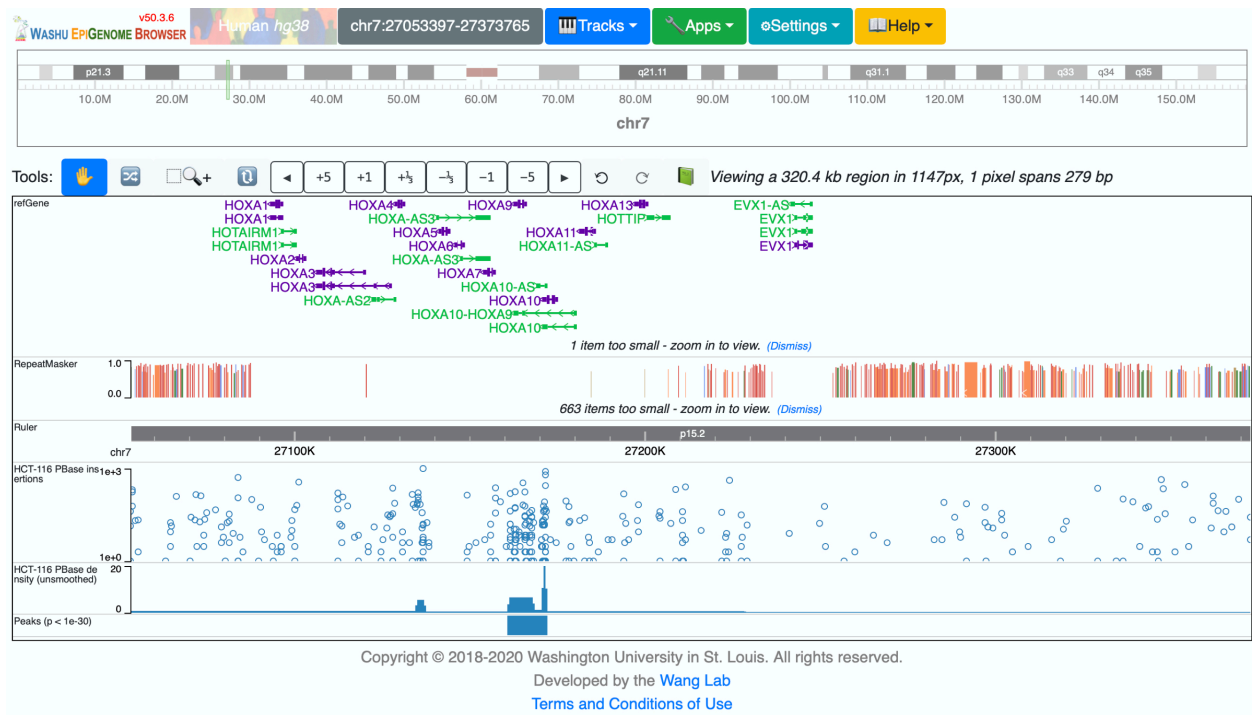


Figure A7.5: Default view after uploading data hub.

Note that the gencodeV29 track has been removed for this screenshot.

A7.4.3 Uploading Local Data Files

11. As an alternative to hosting your data externally, you can directly upload text-based data files, including CCF, bedgraph, and BED. If you choose this option, your files **should not** be compressed. Also, for very large files, this approach may slow down your browser.

12. Click on the blue "Tracks" button and select "Text track."

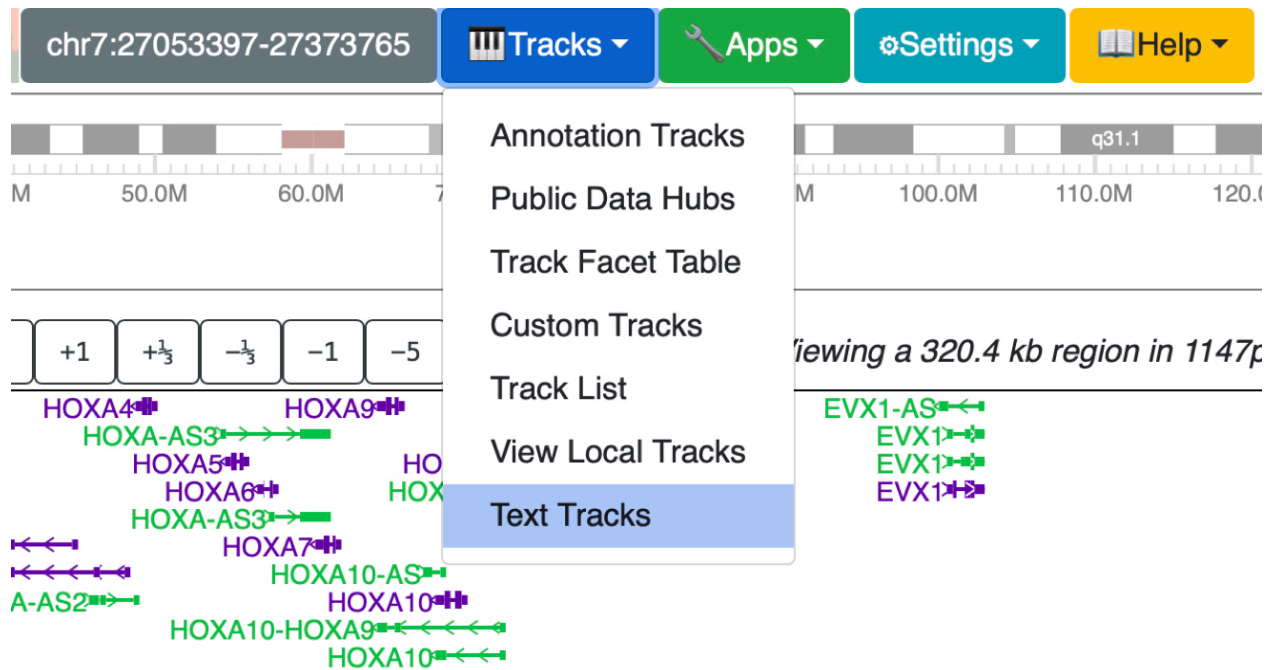


Figure A7.6: Text track pane.

13. Select "callingcard" from the dropdown menu, then click on the "Choose Files" button.

Select one or more CCF files to upload. The browser will then load the tracks; this may take a few moments, depending on the size of the file(s).

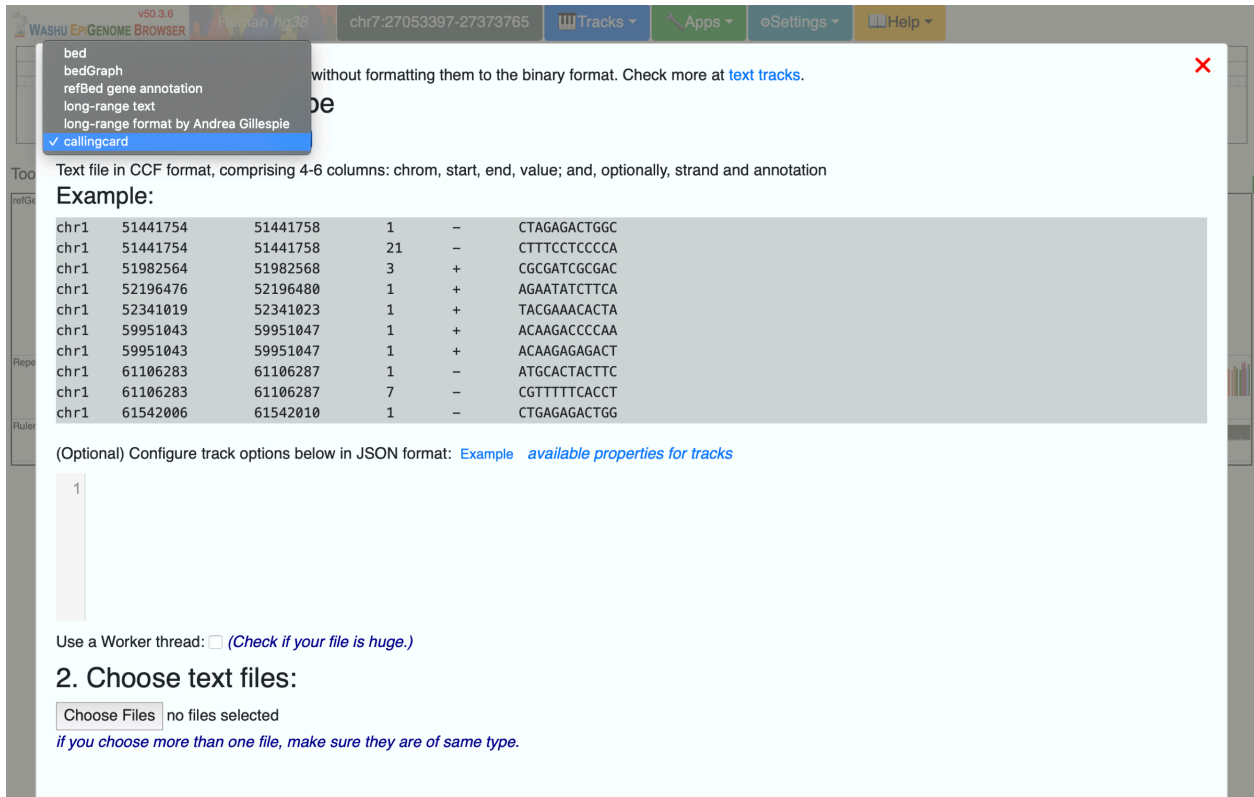


Figure A7.7: Uploading a calling cards text file.

14. After the tracks have been loaded, close the pane by clicking on the red X in the top right corner. When directly uploading text files, tracks will be rendered with default options.



Figure A7.8: Default view after uploading text file.

These steps can be repeated for BED and bedgraph files.

A7.4.4 Interacting with the Calling Card Track

15. The calling card track was designed for interactive exploration. This is chiefly accomplished by a rollover box that appears as the mouse cursor nears a data point. An approximate location is at the top of the pane, while the bottom lists the read count, strand, and barcode for each insertion:

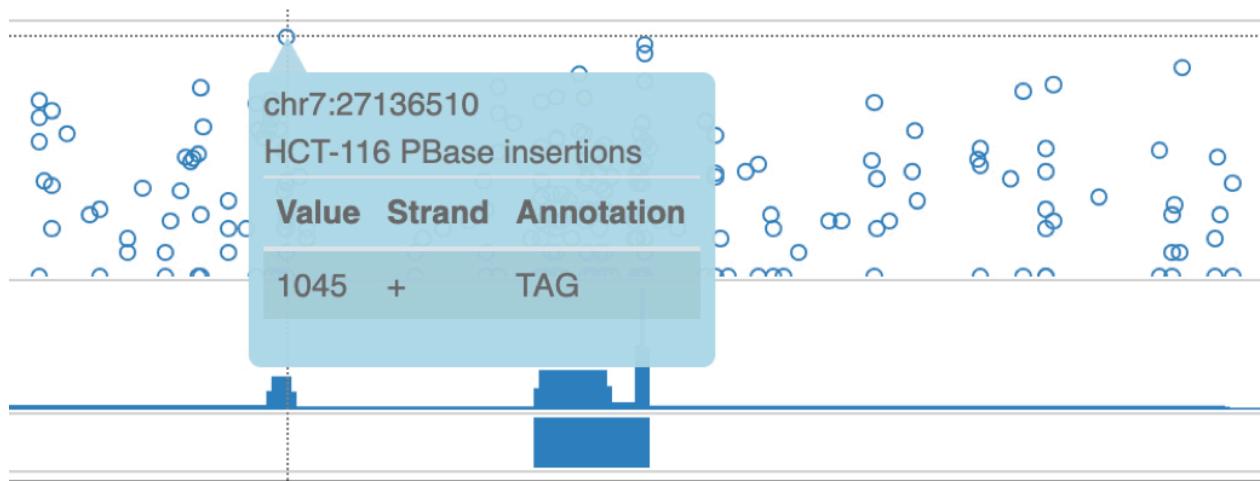


Figure A7.9: Hovering over a calling cards insertion.

16. Right clicking on the calling card track will bring up a preference pane, where tracks can be dynamically customized:

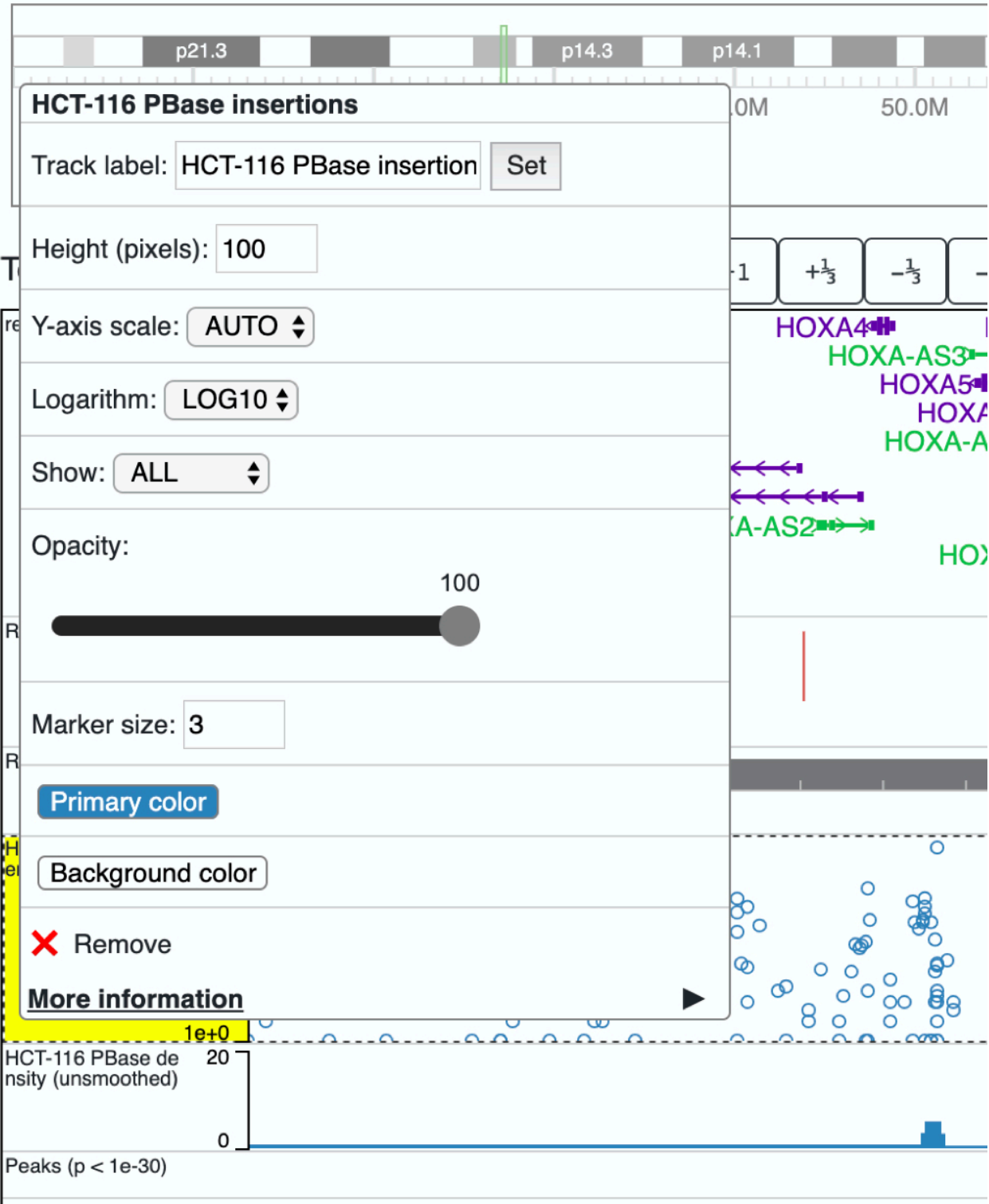


Figure A7.10: Calling cards customization pane.

Similar preference panes exist for the BED and bedgraph tracks.

17. This example showcases all the different options available for calling card tracks, such as color, marker size, opacity, and subsampling:

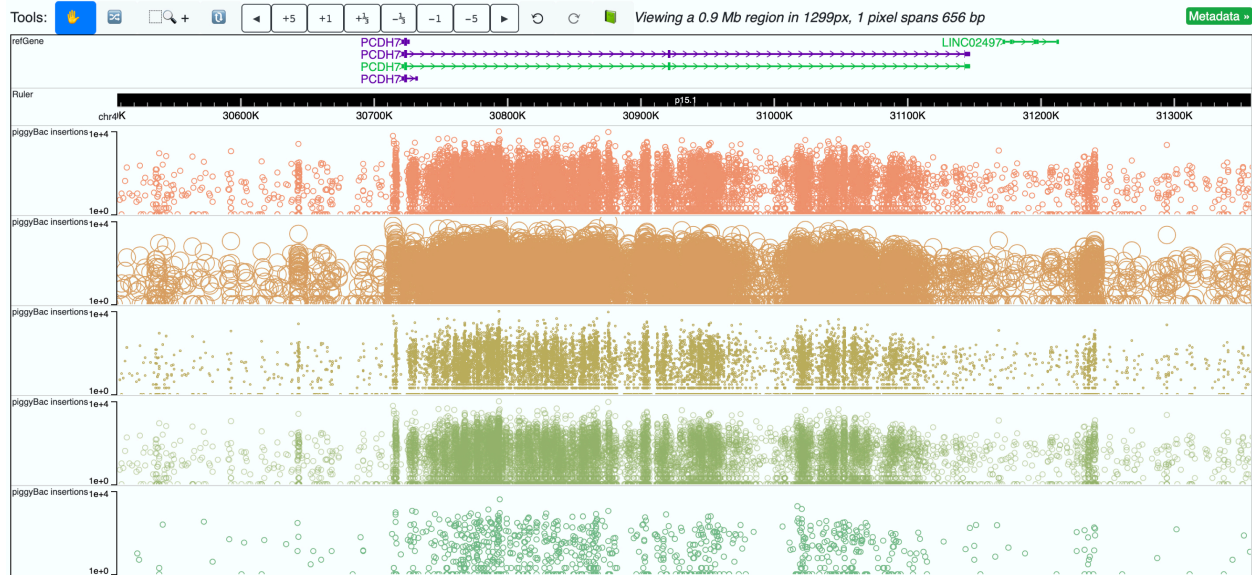


Figure 11: Various customizations applied to calling card tracks.

Appendix 8: Online Resources

A8.1 Links to Online Resources

Computational analyses are a necessary aspect of modern biology and this thesis would not have been possible without custom, task-specific software. While a number of software repositories and other online resources are mentioned throughout the text, we have collected them here for easy reference:

- Main calling cards repository: https://github.com/arnavm/calling_cards
 - This houses scripts referenced in Chapter 2 and Appendices 4-6. Additional scripts are provided specifically for use on the High-Throughput Computing Facility (HTCF) cluster at Washington University in St. Louis. Finally, this repository also has several notebooks that detail the single cell analyses in Chapter 2.
- WashU Epigenome Browser with Calling Card support: <https://github.com/arnavm/eg-react>
 - This fork of the new WashU Epigenome Browser incorporates code for visualizing the calling card track (Appendix 7). These edits have since been merged back into the main branch of the WashU Epigenome Browser.
- Legacy WashU Epigenome Browser with Calling Card support:
<https://github.com/arnavm/eg>

- This was the original fork of the WashU Epigenome Browser that originated the calling card track. While it has been superseded by the previous repository, references to this version can be found in Chapter 2.
- Blockify: <https://github.com/arnavm/blockify>
 - This repository contains the source code for blockify, a genomic segmentation algorithm based on Bayesian blocks (Chapter 4).
 - A packaged version of this code is available from the Python Package Index: <https://pypi.org/project/blockify/>.
 - Accompanying documentation can be found at <https://blockify.readthedocs.io/>
- Mirror of the calling cards repository: https://gitlab.com/arnavm/calling_cards
 - This is a fork of the main calling cards codebase on GitHub. One key difference is that this repository has a Ref folder containing data files that are fetched in blockify's unit tests.
- Transposon calling cards protocols: <https://www.protocols.io/workspaces/calling-cards/>
 - This folder contains online versions of transposon calling cards protocols (Appendices 1-7).
- kmer.cc: <https://gist.github.com/arnavm/039e76a34a386a4f29b82682bc8e6c72>
 - This is a fast program for finding exact sequence matches in a genome and is referenced in Appendix 6.

- Multi-modal scRNA-seq figure: <https://github.com/arnavm/multimodal-scRNA-seq>
 - This repository contains the multi-modal scRNA-seq figure graphic referenced in Chapter 1 as well as links to relevant references.