Functional Characterization of mSWI/SNF Complexes using Perturb-seq

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Chemical Biology

have examined a dissertation entitled

Functional Characterization of mSWI/SNF Complexes using Perturb-seq

presented by Jordan Otto Jagielski

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature Cigall Kadoch

Typed name: Prof. Cigall Kadoch, Ph.D.

Signature

Typed name: Prof. David Liu, Ph.D.

Signature Fred Winston

Typed name: Prof. Fred Winston, Ph.D.

Signature

Typed name: Prof. Mario Suvà, M.D. Ph.D.

Signature

Typed name: Prof.

Date: 4/30/2021

Functional Characterization of mSWI/SNF Complexes using Perturb-seq

A dissertation presented

by

Jordan Otto Jagielski

to

The Committee on Higher Degrees in Chemical Biology

In partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

In the subject of

Chemical Biology

Harvard University

Cambridge, Massachusetts

April, 2021

Dissertation Advisor: Dr. Cigall Kadoch Jordan Otto Jagielski

Functional Characterization of mSWI/SNF Complexes Using Perturb-seq

Abstract

Mammalian SWI/SNF (mSWI/SNF or BAF) complexes are multi-subunit chromatin remodeling machines that use the power derived from ATP hydrolysis to mobilize nucleosomes and facilitate chromatin accessibility. mSWI/SNF complexes exist in three forms: canonical BAF (cBAF), polybromo-associated BAF (PBAF), or non-canonical BAF (ncBAF), each of which has a distinct subunit composition, genomic localization pattern, and activity on chromatin. mSWI/SNF subunits are mutated in greater than 20% of human cancers as well as in a variety of neurodevelopmental disorders, which has motivated the study of their activities across a wide range of disease settings.

While studies have begun to define the roles for individual mSWI/SNF subunits in various specific tissue or disease contexts, this heterogeneity in genetic backgrounds has hindered efforts to comparatively assess mSWI/SNF subunit functions. To this end, we performed a CRISPR-Cas9 knockout screen followed by single cell RNA sequencing (Perturb-seq), to probe both the individual functions of subunits as well as their combinatorial logic upon perturbation of multiple subunits. We combined these insights with single cell chromatin accessibility profiles and bulk chromatin binding profiles to define complex-, module-, and subunit-specific roles in the regulation of diverse gene sets and pathways. We further identify the contribution of each subunit (both subcomplex-specific and shared) to unique subcomplex activity using a logistic regression classifier. Finally, we mined RNA-seq profiles of TCGA-cataloged primary tumors and identified tumors with high correlation to mSWI/SNF perturbation signatures that lack mutations in any BAF subunit. We probed their mutational landscapes and identified transcription factors that mediate expression of the highly correlated gene sets to identify potential convergent mechanisms of gene expression signatures.

In addition to this large-scale effort, a focused study of the roles of the paralogous mSWI/SNF subunits ARID1A and ARID1B was performed to elucidate their similarities and differences in functional domains, gene targeting roles, and effects on mSWI/SNF complex composition. Using CRISPR-Cas9 domain scanning, the critical role for a previously uncharacterized N-terminal region of ARID1B was identified. Additionally, we found a preference for SMARCD paralog integration based upon ARID1 paralog integration, as well as identified different chromatin binding patterns of ARID1A- and ARID1B-containing complexes. Taken together, this body of work represents a multidisciplinary effort to comparatively and individually assess the functional roles of mSWI/SNF subunits.

Table of Contents

Appendix

Acknowledgements

The work described in this dissertation is the result not only of my own labor, but was supported heavily by the thoughts, ideas, inspiration, and feedback of so many throughout my time in graduate school.

First and foremost, I want to thank my advisor **Dr. Cigall Kadoch** for her mentorship and support throughout my Ph.D. experience. Through the many meetings, discussions, drafts, and presentations, she has shaped my critical thinking, scientific understanding, writing, and confidence in my approach to scientific research. Thank you for giving me the opportunity to learn, grow, and become an independent scientist under your tutelage.

I want to thank my collaborators **Dr. Oana Ursu** and **Alexander Wu** for their incredible scientific contributions to this body of work. As the computational experts for this project, their efforts and insights made these analyses possible. I would also like to thank their mentors **Dr. Aviv Regev** and **Dr. Bonnie Berger** for their support in this endeavor. I specifically want to thank Aviv for her invaluable insights and ideas as we built this project.

A big thank you goes to **Dr. David Liu**, **Dr. Fred Winston**, and **Dr. Mario Suvà** for their service on my dissertation advisory committee. Their feedback and insights heavily guided and shaped my projects into this thesis, and I am grateful for the many helpful and supportive conversations.

I want to thank all of the incredible past and present Kadoch Lab members for their scientific contributions and feedback as well as their friendship throughout the past five years. My labmates are some of the best people I know, and they made it a joy to come to lab every day. I especially want to thank **Kaylyn Williamson**, **Hayley Zullow**, and **Dawn Comstock** for their friendship and support. I am grateful to have spent (so much) time with you every day.

To my Chemical Biology Program friends, **Laura Doherty** and **Amanda Clark**, I could not have gotten through moving to Boston, classes, the qualifying exam, or the many tough spots in graduate school without you. I am so thankful we walked this journey together.

To my friends **Michele Fredlake**, **Kathleen VanGilder**, and **Samantha Fisher**, thank you from the bottom of my heart for all of your support from afar. Our decade of friendship is something I treasure dearly, and I can't state enough how much I look forward to every visit, call, and message from you.

To my sister **Taylor Otto**, I could not be more grateful to have a sister I can also call my best friend. You brighten every day and make my life better just by existing. I can always count on you for the best jokes, cheering up, support, and advice.

To my mom, **Amy Otto**, and my dad, **Kevin Otto**, I cannot overstate how much you have impacted my ability to get here. You are the most supportive and caring parents. You taught me the value of hard work and the importance of kindness. Thank you for always believing in me. With your support, I was able to follow my dreams and actually achieve them. For this, I am forever grateful.

To my husband, **Patrick Jagielski**, thank you for walking with me through all of the ups and downs both in graduate school and in life. You have been my anchor during this stressful time, and I cannot thank you enough for the support you've given me over the past few years. When you didn't know how to help, you'd always say, "I believe in you," and that was always exactly what I needed to hear.

Last and certainly not least, I need to thank **Selene Otto Jagielski**, my feline companion, for keeping me grounded with constant reminders for chin scratches and tuna dinners.

Chapter 1: Introduction

I. Epigenetic regulation in normal and disease states

Each cell in the human body holds a colossal amount of genetic information encoded by nearly two meters of DNA measured end-to-end[1]. The nucleus of each cell (where DNA resides) measures on average less than 10 micrometers in diameter, so DNA has to be highly compacted to fit in this small space[2]. The process by which DNA is compacted is highly regulated and occurs in an organized, precise manner. DNA is wrapped tightly around a set of proteins called histone octamers, much like thread is wrapped around a spool, to form the functional unit known as a nucleosome[3]. DNA is wrapped around one histone octamer approximately 1.7 times (or a 146 base pair length of DNA); the subsequent DNA is wrapped around new histone octamers in succession forming a structure called chromatin, which is described as having a “beads on a string” appearance[4]. Chromatin can be further compacted into denser formats by folding into secondary structures (interactions between nucleosomes) or tertiary structures (longer-range interactions between secondary structures), or it can remain more loosely coiled[5]. The tightly coiled form of chromatin is known as heterochromatin, while the more loosely coiled form is called euchromatin[6]. Heterochromatin is highly space efficient; however, it is very difficult for the cell to access the underlying DNA sequences in heterochromatin, since the DNA and proteins are so tightly packed[7]. In order for the cell to read the information encoded by DNA, segments of DNA must be exposed from the nucleosome-packed template; hence, it is necessary to uncoil heterochromatin in order for the cell to read the underlying genetic information[8]. When a gene or genetic element is not required to be read, it can be packed away into the dense heterochromatin structure[6].

The process by which chromatin is “opened” and “closed” is governed by a host of factors which work together to tightly regulate access to genes for factors necessary for transcribing DNA into mRNA, the messenger genetic molecule[9]. After mRNA is transcribed, it is then translated into protein to perform important functions in the cell. This workflow of DNA transcription to RNA, then RNA translation to protein is the central dogma of biology[10]. While the actual DNA sequence is very important for the correct protein to be made and mutations in DNA can lead to an incorrect protein sequence that has errant functions, the control of the timely transcription of DNA to RNA and subsequent translation to protein is also critical. While mutations in genes themselves give rise to a wide variety of human diseases, so too can improper regulation of the temporal activation or repression of DNA expression[11].

The study of the regulation of changes in gene expression and phenotypic outcome without changes in the actual DNA sequence is the main focus of the field of epigenetics. One of the first descriptions of the field of epigenetics was the study of how genotype is linked to the phenotype of a cell; this was first noted to be important specifically in development, as all cells in the human body have the same DNA yet have drastically different developmental outcomes[12]. Additional definitions have arisen over time as the concept of the field of epigenetics has evolved and as more was learned about the mechanisms by which gene expression is maintained and altered, and the field will continue to evolve in its understanding of all the factors that contribute to phenotypic cell states sans alterations in genetic sequence. While the qualification for epigenetic-level regulation has historically required changes to be heritable, the field of epigenetics has more recently been expanded to include non-heritable properties as well[13]. There are many important factors that govern the relationship between genotype and phenotype, some of the first identified being transcription factors, of which there are hundreds that coordinate expression by binding sites of DNA and either activate or repress their transcription[14].

In addition to transcription factors, three large classes of proteins and protein complexes that contribute to gene activation and repression include DNA modifying enzymes (e.g. DNMTs), histone modifying enzymes (e.g. HATs or HDACs), and chromatin remodeling complexes (e.g. mSWI/SNF)[15]. Both DNA and histone modifying enzymes work by chemically modifying either a DNA residue (in the case of DNA modifying enzymes) or a histone's amino acid side chain (in the case of histone modifying enzymes) with chemical modifications[16]. DNA methylation is catalyzed by enzymes known as DNA methyltransferases (DNMTs), which catalyze the conversion of cytosine to 5-methylcytosine; DNA methylation is associated with the repression of gene expression[17]. Histone modifications can take many forms, though the most common include acetylation and methylation. Lysine, arginine, and serine sidechains in the exposed histone tails have the potential to be chemically modified with a whole host of chemical groups, including the aforementioned acetylation and methylation, and additional modifications such as ubiquitination, phosphorylation, sumoylation, and crotonylation, amongst others[18,19]. These marks serve multiple purposes. Some evidence suggests that these marks may change the conformation or charge of the histone tail and as such begin to "loosen" the DNA wrapped around the histone to make it more permissible to transcriptional elements[20]. Additionally, these marks may serve as recruitment factors for chromatin remodelers and other transcriptional machinery which then can come in and perform their additional epigenetic regulatory functions[21].

One highly studied family of histone modifying proteins are the polycomb group proteins, which exist in two large complex forms: polycomb repressive complexes 1 and 2 (PRC1 and PRC2). These complexes each have an enzymatic subunit (RING1A/B, a ubiquitin ligase, or EZH2, a histone methyltransferase, respectively) which catalyze the transfer of chemical modifications onto histone tails[22]. Polycomb complexes are known for their activities in repression of gene expression, and are associated with silencing transcription and critical developmental repressive processes such as the maintenance of inactivated X-chromosome repression[23,24].

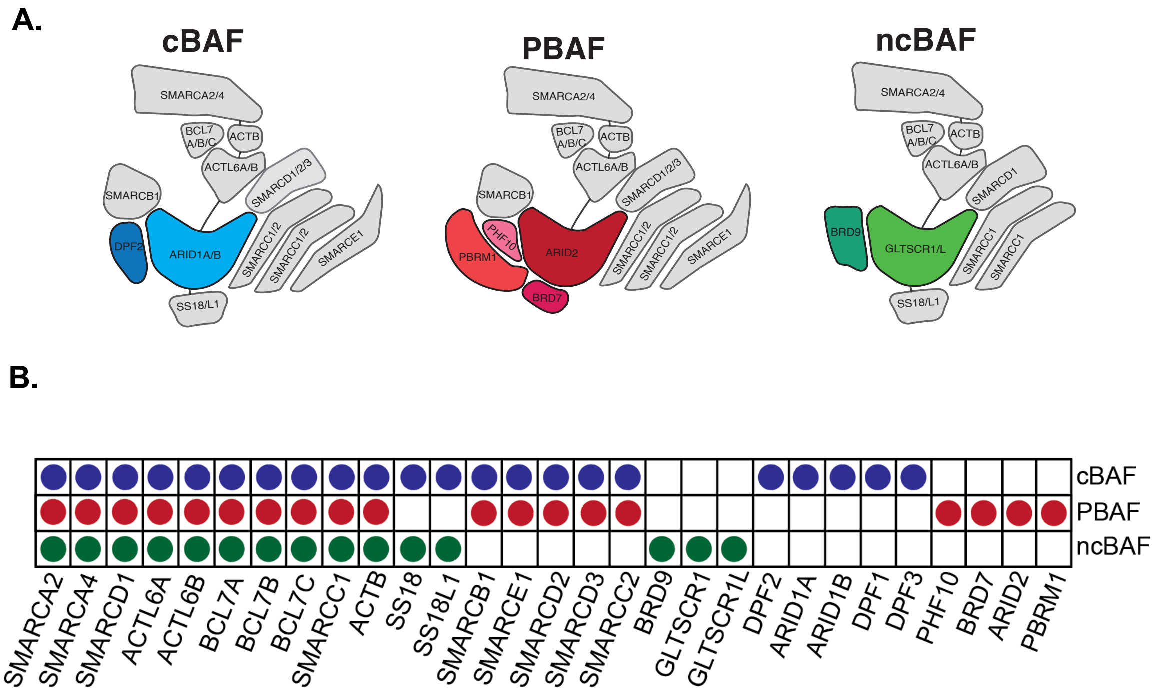
One additional class of epigenetic regulators are ATP-dependent chromatin remodeling complexes, which are multi-subunit molecular machines that use the power derived from ATP hydrolysis to mobilize, evict, or position nucleosomes[25]. This alteration of nucleosome positioning affects the accessibility of underlying DNA elements to transcriptional machinery[26]. The main families of chromatin remodelers are SWI/SNF, ISWI, NuRD, and INO80, each having a core ATPase subunit with peripheral subunits that confer different functional properties[27,28,29,30]. Each of these families of proteins work by increasing nucleosome mobility and altering nucleosome positioning, and many have been implicated in broader functional roles associated with chromatin architecture and the maintenance and progression of development-critical states[15].

The precise temporal coordination of gene activation and repression allows cells to achieve proper development states, maintain regular growth cycles, respond to external stimuli, maintain homeostasis, and undergo apoptosis when necessary[31]. Mutations in epigenetic regulatory elements lead to altered genetic regulation, which in certain cases can cause incorrect genes to be activated or repressed at inopportune times. When this dysregulation occurs in genes critical to differentiation, development, growth, or cell cycle, these mutations often lead to cancer[32]. Mutations in epigenetic modifiers are additionally found in a wide variety of human diseases including neurodevelopmental disorders as well as intellectual disability disorders[33]. Huge bodies of work have been conducted to elucidate disease-associated mechanisms in epigenetically dysregulated diseases. Oftentimes, repairing or inhibiting a mutated epigenetic regulatory protein will reverse disease-associated biology *in vivo*[34]. Additionally, epigenetic modifiers may be a susceptibility in certain cancers that are addicted to a specific oncogenic gene program[35]. Some of the best-studied epigenetic targets in human cancers include histone deacetylases and DNA-methyltransferases; these studies have led to development of targeted therapeutics, such as pan-DNMT inhibitors for the treatment of MDS and AML, HDAC inhibitors for a variety of hematological malignancies and solid tumors, and additional epigenetic targets in clinical trials include BET proteins, EZH2, and DOT1L[36]. Better understanding of the ways epigenetic regulatory elements contribute to disease states will allow for increased ability to therapeutically intervene in a variety of human cancers and diseases.

While many studies of the functionality of epigenetic regulatory elements have been inspired or motivated by observations in human disease, these elements are inherently interesting in their own right on a basic science level. The many layers of regulation, from the chemical modification of DNA itself, to the modifications of the histones around which DNA is wrapped, to the large molecular machines that alter nucleosome positioning, work together in symphony to allow cells to properly develop and differentiate as well as maintain homeostasis.

II. mSWI/SNF complexes and their functions in regulating gene expression

Mammalian SWI/SNF (mSWI/SNF or BAF) complexes are a class of ATP-dependent chromatin remodelers that utilize the energy derived from ATP hydrolysis to mobilize nucleosomes. BAF complexes are composed of 10-15 subunits arranged combinatorially from 29 different genes and can exist in three distinct forms: canonical BAF (cBAF), polybromo-associated BAF (PBAF), and non-canonical BAF (ncBAF)[37,38,39]. Each form of the BAF complex possesses unique subunits as well as subunits that are shared between multiple complex forms (**Figure 1.1**).



	SMARCA2	SMARCA4	SMARCD1	ACTL6A	ACTL6B	BCL7A	BCL7B	BCL7C	SMARCC1	ACTB	SS18	SS18L1	SMARCB1	SMARCE1	SMARCD2	SMARCD3	SMARCC2	BRD9	GLTSCR1	GLTSCR1L	DPF2	ARID1A	ARID1B	DPF1	DPF3	PHF10	BRD7	ARID2	PBRM1
cBAF	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●				●	●	●	●	●				
PBAF	●	●	●	●	●	●	●	●	●	●			●	●	●	●	●									●	●	●	●
ncBAF	●	●	●	●	●	●	●	●	●	●	●	●						●	●	●									

Figure 1.1 Composition of mSWI/SNF complexes and distribution of subunits across complex types.

A. Composition of the three mSWI/SNF subcomplex types: cBAF, PBAF, and ncBAF. Unique subunits that define complex identity are shown in blue (cBAF), red (PBAF), and green (ncBAF). Non-unique subunits that are shared across subcomplex types are shown in grey. **B.** Chart depicting which mSWI/SNF subunits can integrate into each subcomplex type.

mSWI/SNF complexes are members of the highly conserved SWI/SNF complex family, which was first discovered and characterized in *Saccharomyces Cerevisiae* as an important regulator of chromatin in eukaryotes. Genetic screens conducted in *S. cerevisiae* mutants identified a set of genes that led to defective mating type switching (hence their names including SWI for SWItch), as well a set of genes leading to defects in growth on sucrose media (hence their names including SNF for Sucrose Non-Fermenting)[40,41,42]. Mutations in these genes were subsequently found to have widespread effects on the expression of many genes, thus implicating them in transcriptional regulation[43]. The products of these genes were found to assemble into protein complexes, with the proteins being functionally interdependent in their roles for regulating transcription[44,45]. Later work characterized the mechanism behind these complexes' effects on transcription as the modulation of chromatin accessibility[46,47]. SWI/SNF complexes are highly evolutionarily conserved (from yeast to drosophila and mammals), and they have been studied in many model systems[48]. Foundational studies of the human SWI/SNF complex (mSWI/SNF) were later performed to characterize the functions of mSWI/SNF subunits within the context of mammalian gene expression and genome architecture, including the definition of core functional members of these complexes, the physical mechanisms that occur during remodeling, and the relationship between ATP hydrolysis and the maintenance of accessible chromatin states[49,50,51,52].

While mSWI/SNF subunits bind together as a complex to perform the concerted task of remodeling chromatin, there are many pieces of evidence that suggest that individual BAF complex subunits have unique contributing roles to the overall function of these complexes. These pieces of evidence include 1) tissue-specific expression of BAF subunits[53,54], 2) differentiation-dependent incorporation and switching of BAF subunits[55], and 3) the consistent and striking mutational pattern of individual subunits in a wide variety of human cancers and neurodevelopmental disorders[33,56]. One example of differentiation stage-linked, compositionally unique BAF complexes is esBAF, or embryonic stem cell BAF, which contains only SMARCC1 homodimers rather than the SMARCC1/SMARCC2 heterodimers which occur in later differentiation stages[57]. An example of subunit switching along discreet differentiation stages is the change from neural progenitor BAF(npBAF) to neuron BAF (nBAF) which is marked by the switch in incorporation of the SS18L1 subunit to SS18[58].

BAF complexes have largely been studied in the context of cancer, as they are mutated in more than 20% of human malignancies[59]. Many of these mutations occur consistently in an individual subunit tied to a specific tissue and cancer context (e.g. ARID1A in ovarian clear cell carcinoma[60], SMARCB1 in malignant rhabdoid tumors[61], PBRM1 in clear cell renal cell carcinoma[62], SS18 in synovial sarcoma[63], etc.) indicating there is likely some tissue-critical context and role for BAF complex subunits in the development of these malignancies. Additionally, there are mutations in many BAF complex subunits in a variety of neurodevelopmental disorders, such as Coffin-Siris syndrome and Nicolaides-Barraitser syndrome[64].

BAF complex subunits gain some of their distinct functionality through their variety of protein domains. Many functional protein domains are contained within BAF complex subunits, such as DNA binding domains (HMG, ARID, HSA) and epigenetic reader domains (Bromo, PHD finger)[37]. Each BAF complex contains either SMARCA2 or SMARCA4, which is the ATPase subunit and the workhorse of the complex[65]. These ATPase proteins contain a series of domains (including the SNF2, helicase, HSA, and SnAC domains) that function to catalyze the hydrolysis of ATP as well as grip and translocate the nucleosome[66]. The largest members of BAF complexes are ARID1A and ARID1B, which are 250 KD each. The presence of one of these mutually exclusive ARID1 proteins defines canonical BAF complexes, along with DPF2. Subunits defining PBAF complexes include ARID2, PBRM1, BRD7 and PHF10, while the subunits that define ncBAF complexes are GLTSCR1, GLTSCR1L, and BRD9[67].

One functional difference between cBAF, PBAF, and ncBAF complexes is their distinct localization patterns on chromatin. PBAF complexes largely associate with promoters, while cBAF complexes localize to more distal regulatory elements including enhancers[68]. ncBAF localizes across a variety of sites in the genome including promoter proximal and distal sites, as well as CTCF (a chromatin looping factor) binding sites[37,69]. With these distinct localization patterns, these complexes likely cooperate and individually regulate accessibility at varying distances between occupancy and the gene target: PBAF the immediate site near the promoter of the gene, cBAF the longer distance enhancer elements, and ncBAF the extremely long-range regulation involved in higher order chromatin loops (due to its association with CTCF sites). All forms of BAF complexes are typically thought to be activating complexes by increasing chromatin accessibility to transcription factors and transcriptional machinery, though recently roles in facilitating repressive states by allowing repressive transcription factors to bind or through positioning nucleosomes in a manner that is consistent with gene repression have been described[37,70].

Recently, the assembly order of BAF complexes was elucidated, giving insight into the process by which these large multi-subunit machines are assembled from the many constituent subunits. BAF complex formation is initiated by the dimerization (either homo- or hetero-) of the SMARCC subunits. Subsequent core subunits (SMARCB1, SMARCE1) join on to form the functional core of BAF complexes. After core formation, the ARID or GLTSCR subunit specific to the complex form (ARID1A/ARID1B for cBAF, ARID2 for PBAF, or GLTSCR1/GLTSCR1L for ncBAF) forms a bridge to attach to the ATPase module, which consists of SMARCA2/4, BCL7A/B/C, ACTL6A/B, and ACTB. The ATPase module assembles independently of the core formation process[71].

Many BAF subunits are members of paralog families, most of which are mutually exclusive within complexes except for the paralogs SMARCC1 and SMARCC2, of which two fit into a single BAF complex. The differential roles of paralogs within a family have remained largely elusive, though previously cited evidence of disease mutation frequency and subunit switching specifically within paralog families suggests that differential roles exist. Synthetic lethal relationships have been identified within two mSWI/SNF paralog families, ARID1A/ARID1B, and SMARCA2/SMARCA4, via large scale genome-wide fitness screens such as Project Achilles[72,73,74]. Understanding the roles of each complex type, the relationships between complex members, and specific subunit functions remain active areas of investigation in the mSWI/SNF field.

III. Critical unanswered questions in mSWI/SNF biology addressed in this thesis

While quite a bit is known about the functions of mSWI/SNF complexes, many questions remain about these complexes and their roles in epigenetic regulation, both in normal and disease states. One important open question is how mSWI/SNF subunits function in relation to one another; how are they functionally similar or functionally distinct? Individual mSWI/SNF subunits have been largely studied in singular tissue types or cell contexts (typically those where their roles are suspected to be critical based on insights from human disease or genetic screens), which has yielded insights into the tissue-specific context of individual subunits. For example, studies of ARID1A in colorectal carcinoma and in hepatocellular carcinoma led to the identification of functional mechanisms or direct gene targets that are critically important in the respective tissue types for maintaining normal cell functions[75,76], while SS18 and specifically the SS18-SSX fusion protein were shown to hijack mSWI/SNF complexes to new gene target sites in synovial sarcoma[63]. While understanding the roles of these subunits in individual, relevant tissue types and contexts important to understanding why these proteins are critical in various settings, it is difficult to comparatively assess the roles of the subunits to one another due to the heterogeneity of the background genetic and epigenetic environments of different cell contexts.

While studies aimed at elucidating the roles of individual BAF subunits have been performed in a variety of contexts, the comprehensive dissection of individual subunit functions in relationship to one another has remained elusive due to various technical limitations of current systems for profiling large numbers of subunits with readouts of sufficient resolution. While functional similarity networks have been mapped for mSWI/SNF complexes using large-scale fitness screening data, we do not yet understand the actual gene expression changes that underlie the similarities and differences in function between these subunits[67]. Some studies have recently emerged to address these questions by performing large-scale genomics experiments to understand the roles of these subunits in a single cell context; however, these studies have not yet profiled all mSWI/SNF subunits, nor have they assessed the transcriptomic effects of multiple subunit perturbations simultaneously[77,78]. Towards this goal, large scale studies to holistically dissect all BAF subunit roles have been limited by the labor-intensive process for arrayed assays of each subunit perturbation, long-term viability issues of critical BAF subunit losses in the creation of stable knockout lines, clonal variation in single cell clone-derived populations, and batch effects due to arrayed perturbations and large numbers of sequencing runs. To circumvent these limitations, in this work we use Perturb-seq to perturb each BAF complex subunit in a pooled CRISPR-Cas9 knockout screen, assaying transcriptomic changes to reveal the roles of BAF complex subunits with single cell resolution for the very first time. We also combinatorially perturb BAF complex subunits to investigate important biological questions about the roles of paralog families, subcomplex identities, and cooperation between subcomplexes.

Another area of mSWI/SNF research with outstanding questions is that of paralog family functions. There are many pieces of evidence that suggest that paralogs have different functions within a family, including aforementioned differentiation-stage switches and different mutational patterns in human diseases. It is interesting to note that many paralogous subunits are highly similar in sequence, suggesting paralogs also have shared functionality. Functional studies have begun to specify distinct roles for single paralogs within a family; for example the critical role of SMARCD2 in mediating granulopoeisis, or the role of ARID1A in regulation of the G2/M cell cycle phase transition[79,80]. However, there are still many questions as to how paralogs are functionally different from one another and which domains and/or functional sites of these protein contribute to their unique functions. For example, ARID1A and ARID1B are highly similar in sequence, but their functional and structural similarities and differences remain poorly described, even though they have very different mutational patterns in human disease. To better understand the differential roles of the ARID1A and ARID1B subunits, a variety of experiments to answer questions about differences in functional regions, composition, and gene targeting were performed in this body of work.

To summarize, the work described in this thesis addresses both of these unmet areas of research within the mSWI/SNF field, specifically by:

1. Dissecting the functional roles of mSWI/SNF subunits in a uniform genetic background, both through individual and combinatorial subunit perturbations
2. Elucidating the functional differences between the mSWI/SNF paralogs ARID1A and ARID1B in both their structures and activities

IV. References

1. Bloom, K. & Joglekar, A. Towards building a chromosome segregation machine. *Nature* **463**, 446–456 (2010).

2. Annuziato, A. DNA Packaging: Nucleosomes and Chromatin. *Nature Education* **1(1)**, (2008).

3. Kornberg, R. D. & Lorch, Y. Twenty-Five Years of the Nucleosome, Fundamental Particle of the Eukaryote Chromosome. *Cell* **98**, 285–294 (1999).

4. Olins, D. E. & Olins, A. L. Chromatin history: our view from the bridge. *Nature Reviews Molecular Cell Biology* **4**, 809–814 (2003).

5. Woodcock, C. L. & Dimitrov, S. Higher-order structure of chromatin and chromosomes. *Current Opinion in Genetics & Development* **11**, 130–135 (2001).

6. Penagos-Puig, A. & Furlan-Magaril, M. Heterochromatin as an Important Driver of Genome Organization. *Front Cell Dev Biol* **8**, 579137 (2020).

7. Lee, D. Y., Hayes, J. J., Pruss, D. & Wolffe, A. P. A positive role for histone acetylation in transcription factor access to nucleosomal DNA. *Cell* **72**, 73–84 (1993).

8. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics* **20**, 207–220 (2019).

9. Hahn, S. & Young, E. T. Transcriptional Regulation in Saccharomyces cerevisiae: Transcription Factor Regulation and Function, Mechanisms of Initiation, and Roles of Activators and Coactivators. *Genetics* **189**, 705–736 (2011).

10. Crick, F. Central Dogma of Molecular Biology. *Nature* **227**, 561–563 (1970).

11. Ladd-Acosta, C. & Fallin, M. D. The role of epigenetics in genetic and environmental epidemiology. *Epigenomics* **8**, 271–283 (2015).

12. Waddington, C. H. The Epigenotype. *International Journal of Epidemiology* **41**, 10–13 (2012).

13. Harvey, Z. H., Chen, Y. & Jarosz, D. F. Protein-Based Inheritance: Epigenetics beyond the Chromosome. *Mol Cell* **69**, 195–202 (2018).

14. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).

15. Ho, L. & Crabtree, G. R. Chromatin remodelling during development. *Nature* **463**, 474–484 (2010).

16. Allis, C. D. & Jenuwein, T. The molecular hallmarks of epigenetic control. *Nature Reviews Genetics* **17**, 487–500 (2016).

17. Suzuki, M. M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics* **9**, 465–476 (2008).

18. Wang, Y. *et al.* Linking Covalent Histone Modifications to Epigenetics: The Rigidity and Plasticity of the Marks. *Cold Spring Harb Symp Quant Biol* **69**, 161–170 (2004).

19. Kouzarides, T. Chromatin Modifications and Their Function. *Cell* **128**, 693–705 (2007).

20. Grunstein, M. Histone acetylation in chromatin structure and transcription. *Nature* **389**, 349–352 (1997).

21. Lawrence, M., Daujat, S. & Schneider, R. Lateral Thinking: How Histone Modifications Regulate Gene Expression. *Trends in Genetics* **32**, 42–56 (2016).

22. Margueron, R. & Reinberg, D. The Polycomb complex PRC2 and its mark in life. *Nature* **469**, 343–349 (2011).

23. Schuettengruber, B., Bourbon, H.-M., Di Croce, L. & Cavalli, G. Genome Regulation by Polycomb and Trithorax: 70 Years and Counting. *Cell* **171**, 34–57 (2017).

24. Plath, K. *et al.* Role of histone H3 lysine 27 methylation in X inactivation. *Science* **300**, 131–135 (2003).

25. Hota, S. K. & Bruneau, B. G. ATP-dependent chromatin remodeling during mammalian development. *Development* **143**, 2882–2897 (2016).

26. Wasylyk, B. & Chambon, P. Transcription by eukaryotic RNA polymerases A and B of chromatin assembled in vitro. *Eur J Biochem* **98**, 317–327 (1979).

27. Guyon, J. R., Narlikar, G. J., Sif, S. & Kingston, R. E. Stable remodeling of tailless nucleosomes by the human SWI-SNF complex. *Mol Cell Biol* **19**, 2088–2097 (1999).

28. Toto, M., D'Angelo, G. & Corona, D. F. V. Regulation of ISWI chromatin remodelling activity. *Chromosoma* **123**, 91–102 (2014).

29. Torchy, M. P., Hamiche, A. & Klaholz, B. P. Structure and function insights into the NuRD chromatin remodeling complex. *Cell Mol Life Sci* **72**, 2491–2507 (2015).

30. Conaway, R. C. & Conaway, J. W. The INO80 chromatin remodeling complex in transcription, replication and repair. *Trends in Biochemical Sciences* **34**, 71–77 (2009).

31. Toh, T. B., Lim, J. J. & Chow, E. K.-H. Epigenetics in cancer stem cells. *Mol Cancer* **16**, 29 (2017).

32. Nebbioso, A., Tambaro, F. P., Dell'Aversana, C. & Altucci, L. Cancer epigenetics: Moving forward. *PLoS Genet* **14**, e1007362 (2018).

33. López, A. J. & Wood, M. A. Role of nucleosome remodeling in neurodevelopmental and intellectual disability disorders. *Front. Behav. Neurosci.* **9**, (2015).

34. Dawson, M. A. & Kouzarides, T. Cancer Epigenetics: From Mechanism to Therapy. *Cell* **150**, 12–27 (2012).

35. Weinstein, I. B. Addiction to Oncogenes--the Achilles Heal of Cancer. *Science* **297**, 63–64 (2002).

36. Wimalasena, V. K., Wang, T., Sigua, L. H., Durbin, A. D. & Qi, J. Using Chemical Epigenetics to Target Cancer. *Mol Cell* **78**, 1086–1095 (2020).

37. Centore, R. C., Sandoval, G. J., Soares, L. M. M., Kadoch, C. & Chan, H. M. Mammalian SWI/SNF Chromatin Remodeling Complexes: Emerging Mechanisms and Therapeutic Strategies. *Trends in Genetics* **36**, 936–950 (2020).

38. Michel, B. C. *et al.* A non-canonical SWI/SNF complex is a synthetic lethal target in cancers driven by BAF complex perturbation. *Nat Cell Biol* **20**, 1410–1420 (2018).

39. Alpsoy, A. & Dykhuizen, E. C. Glioma tumor suppressor candidate region gene 1 (GLTSCR1) and its paralog GLTSCR1-like form SWI/SNF chromatin remodeling subcomplexes. *Journal of Biological Chemistry* **293**, 3892–3903 (2018).

40. Stern, M., Jensen, R. & Herskowitz, I. Five SWI genes are required for expression of the HO gene in yeast. *J Mol Biol* **178**, 853–868 (1984).

41. Neigeborn, L. & Carlson, M. Genes affecting the regulation of SUC2 gene expression by glucose repression in Saccharomyces cerevisiae. *Genetics* **108**, 845–858 (1984).

42. Peterson, C. L., Dingwall, A. & Scott, M. P. Five SWI/SNF gene products are components of a large multisubunit complex required for transcriptional enhancement. *Proc Natl Acad Sci U S A* **91**, 2905–2908 (1994).

43. Ryan, M. P., Jones, R. & Morse, R. H. SWI-SNF complex participation in transcriptional activation at a step subsequent to activator binding. *Mol Cell Biol* **18**, 1774–1782 (1998).

44. Laurent, B. C., Treitel, M. A. & Carlson, M. Functional interdependence of the yeast SNF2, SNF5, and SNF6 proteins in transcriptional activation. *Proc Natl Acad Sci U S A* **88**, 2687–2691 (1991).

45. Peterson, C. L. & Herskowitz, I. Characterization of the yeast SWI1, SWI2, and SWI3 genes, which encode a global activator of transcription. *Cell* **68**, 573–583 (1992).

46. Côté, J., Quinn, J., Workman, J. L. & Peterson, C. L. Stimulation of GAL4 derivative binding to nucleosomal DNA by the yeast SWI/SNF complex. *Science* **265**, 53–60 (1994).

47. Hirschhorn, J. N., Brown, S. A., Clark, C. D. & Winston, F. Evidence that SNF2/SWI2 and SNF5 activate transcription in yeast by altering chromatin structure. *Genes Dev* **6**, 2288–2298 (1992).

48. Martens, J. A. & Winston, F. Recent advances in understanding chromatin remodeling by Swi/Snf complexes. *Curr Opin Genet Dev* **13**, 136–142 (2003).

49. Phelan, M. L., Sif, S., Narlikar, G. J. & Kingston, R. E. Reconstitution of a core chromatin remodeling complex from SWI/SNF subunits. *Mol Cell* **3**, 247–253 (1999).

50. Aoyagi, S. *et al.* Nucleosome remodeling by the human SWI/SNF complex requires transient global disruption of histone-DNA interactions. *Mol Cell Biol* **22**, 3653–3662 (2002).

51. Imbalzano, A. N., Schnitzler, G. R. & Kingston, R. E. Nucleosome disruption by human SWI/SNF is maintained in the absence of continued ATP hydrolysis. *J Biol Chem* **271**, 20726–20733 (1996).

52. Kwon, H., Imbalzano, A. N., Khavari, P. A., Kingston, R. E. & Green, M. R. Nucleosome disruption and enhancement of activator binding by a human SW1/SNF complex. *Nature* **370**, 477–481 (1994).

53. Lickert, H. *et al.* Baf60c is essential for function of BAF chromatin remodelling complexes in heart development. *Nature* **432**, 107–112 (2004).

54. Wenderski, W. *et al.* Loss of the neural-specific BAF subunit ACTL6B relieves repression of early response genes and causes recessive autism. *PNAS* **117**, 10055–10066 (2020).

55. Staahl, B. T. *et al.* Kinetic analysis of npBAF to nBAF switching reveals exchange of SS18 with CREST and integration with neural developmental pathways. *J. Neurosci.* **33**, 10348–10361 (2013).

56. Kadoch, C., Copeland, R. A. & Keilhack, H. PRC2 and SWI/SNF Chromatin Remodeling Complexes in Health and Disease. *Biochemistry* **55**, 1600–1614 (2016).

57. Ho, L. *et al.* An embryonic stem cell chromatin remodeling complex, esBAF, is essential for embryonic stem cell self-renewal and pluripotency. *Proc Natl Acad Sci U S A* **106**, 5181–5186 (2009).

58. Lessard, J. *et al.* An Essential Switch in Subunit Composition of a Chromatin Remodeling Complex during Neural Development. *Neuron* **55**, 201–215 (2007).

59. Kadoch, C. *et al.* Proteomic and bioinformatic analysis of mammalian SWI/SNF complexes identifies extensive roles in human malignancy. *Nat Genet* **45**, 592–601 (2013).

60. Jones, S. *et al.* Frequent Mutations of Chromatin Remodeling Gene ARID1A in Ovarian Clear Cell Carcinoma. *Science* **330**, 228–231 (2010).

61. Chun, H.-J. E. *et al.* Genome-Wide Profiles of Extra-cranial Malignant Rhabdoid Tumors Reveal Heterogeneity and Dysregulated Developmental Pathways. *Cancer Cell* **29**, 394–406 (2016).

62. Liao, L. *et al.* Multiple tumor suppressors regulate a HIF-dependent negative feedback loop via ISGF3 in human clear cell renal cancer. *eLife* **7**, (2018).

63. McBride, M. J. *et al.* The SS18-SSX Fusion Oncoprotein Hijacks BAF Complex Targeting and Function to Drive Synovial Sarcoma. *Cancer Cell* **33**, 1128-1141.e7 (2018).

64. Aref-Eshghi, E. *et al.* BAFopathies' DNA methylation epi-signatures demonstrate diagnostic utility and functional continuum of Coffin-Siris and Nicolaides-Baraitser syndromes. *Nat Commun* **9**, 4885 (2018).

65. Pan, J. *et al.* The ATPase module of mammalian SWI/SNF family complexes mediates subcomplex identity and catalytic activity-independent genomic targeting. *Nat. Genet.* **51**, 618–626 (2019).

66. Fernando, T. M. *et al.* Functional characterization of SMARCA4 variants identified by targeted exome-sequencing of 131,668 cancer patients. *Nat Commun* **11**, (2020).

67. Pan, J. *et al.* Interrogation of Mammalian Protein Complex Structure, Function, and Membership Using Genome-Scale Fitness Screens. *Cell Systems* **6**, 555-568.e7 (2018).

68. Nakayama, R. T. *et al.* SMARCB1 is required for widespread BAF complex–mediated activation of enhancers and bivalent promoters. *Nature Genetics* **49**, 1613–1623 (2017).

69. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283 (2011).

70. Rafati, H. *et al.* Repressive LTR nucleosome positioning by the BAF complex is required for HIV latency. *PLoS Biol* **9**, e1001206 (2011).

71. Mashtalir, N. *et al.* Modular Organization and Assembly of SWI/SNF Family Chromatin Remodeling Complexes. *Cell* **175**, 1272-1288.e20 (2018).

72. Helming, K. C. *et al.* ARID1B is a specific vulnerability in ARID1A-mutant cancers. *Nat Med* **20**, 251–254 (2014).

73. Hoffman, G. R. *et al.* Functional epigenetics approach identifies BRM/SMARCA2 as a critical synthetic lethal target in BRG1-deficient cancers. *PNAS* **111**, 3128–3133 (2014).

74. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564-576.e16 (2017).

75. Mathur, R. *et al.* ARID1A loss impairs enhancer-mediated gene regulation and drives colon cancer in mice. *Nat. Genet.* **49**, 296–302 (2017).

76. Sun, X. *et al.* Arid1a Has Context-Dependent Oncogenic and Tumor Suppressor Functions in Liver Cancer. *Cancer Cell* **32**, 574-589.e6 (2017).

77. Schick, S. *et al.* Systematic characterization of BAF mutations provides insights into intracomplex synthetic lethalities in human cancers. *Nat Genet* **51**, 1399–1410 (2019).

78. Rubin, A. J. *et al.* Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks. *Cell* **176**, 361-376.e17 (2019).

79. Priam, P. *et al.* SMARCD2 subunit of SWI/SNF chromatin-remodeling complexes mediates granulopoiesis through a CEBPε dependent mechanism. *Nat Genet* **49**, 753–764 (2017).

80. Nagl, N. G. *et al.* The p270 (ARID1A/SMARCF1) Subunit of Mammalian SWI/SNF-Related Complexes Is Essential for Normal Cell Cycle Arrest. *Cancer Res* **65**, 9236–9244 (2005).

Chapter 2: Structural and functional dissection of mammalian SWI/SNF chromatin remodeling complexes using Perturb-seq

Jordan E. Otto[1,2,3,*], Oana Ursu[2,*], Alexander P. Wu[2,5*], Evan B. Winter[1], Michael Cuoco[2], Sai Ma[2], Kristin Qian[4], Brittany C. Michel[4], Jason D. Buenrostro[2], Bonnie Berger[2,5], Aviv Regev[2,5,+] and Cigall Kadoch[1,2,3,4+]

[1] Department of Pediatric Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA, USA.
[2] Broad Institute of MIT and Harvard, Cambridge, MA, USA.
[3] Chemical Biology Program, Harvard University, Cambridge, MA, USA.
[4] Biological and Biomedical Sciences Program, Harvard Medical School, Boston, MA, USA.
[5] Massachusetts Institute of Technology, Cambridge, MA, USA.
[*] Denotes Equal Contribution to Authorship
[+] Denotes Corresponding Authors

Author Contributions

JEO (Jordan Otto Jagielski)*:* Conceived of project, performed most experimental work unless otherwise noted, wrote chapter
OU: Performed computational analyses for single cell sequencing and dial-out alignment including cell qc metrics and filtering, linear model analysis for combinatorial perturbations, SHARE-seq analysis, and BAF activity classifier.
APW: Performed computational analyses for NMF-derived signatures and UMAP representations of single cells, performed TCGA mining and comparison to Perturb-seq signatures
EBW: Assisted in cloning Perturb-seq guide library
MC: Assisted with dial-out sequencing
SM: Performed SHARE-seq protocol and initial data processing
KQ: Created functional network maps from Project Achilles data and Perturb-seq data
BCM: Generated MOLM-13 ChIP-seq experimental data
JDB: Advised on project
BB: Advised on project
AR: Conceived of project, advised on project
CK: Conceived of project, supervised study

Acknowledgements

This work was supported by the National Science Foundation Graduate Research Fellowship (2016230210). We are grateful to A. Dixit and P. Thakore for helpful conversations about experimental design and protocols. We are thankful to the Broad Institute Genetics Perturbation Platform (GPP) for helpful discussions and for processing 10x Genomics samples as well as sequencing for single-cell transcriptomics and SHARE-seq samples.

Disclosure

CK is the Scientific Founder, fiduciary Board of Directors member, SAB member, shareholder, and consultant for Foghorn Therapeutics, Inc. (Cambridge, MA).

I. Abstract:

mSWI/SNF complexes are multi-subunit chromatin remodeling machines that perform critical functions in modulating chromatin accessibility and gene expression. Studies have begun to elucidate the roles of constituent subunits in the regulation of gene expression in individual tissue and disease contexts; however, the heterogeneity of genetic backgrounds in these studies has made it difficult to comparatively assess the roles of mSWI/SNF complex subunits. In this work, we performed CRISPR-Cas9 knockout screens followed by single cell transcriptome sequencing (Perturb-seq) to comprehensively dissect the contributions of individual and combinations of mSWI/SNF subunits to gene regulation. Combining these insights with data from single cell ATAC-sequencing and bulk chromatin profiling yielded complex-, module-, and subunit-specific roles for the regulation of gene expression and chromatin accessibility. Using combinatorial perturbations, we found both additive and synergistic gene sets across paralog families, as well as recapitulated insights from mSWI/SNF complex assembly. Additionally, we quantified the contributions of subcomplex-defining and shared subunits to subcomplex identity and activity using a logistic regression classifier, and found specific shared subunits with varying degrees of activity within their complex types. Finally, an exploratory comparison of mSWI/SNF subunit perturbation signatures to gene expression signatures of TCGA tumors identified a subset of tumors with high similarity to cBAF and ncBAF perturbation signatures that lack mSWI/SNF subunit perturbations. We probed their mutational landscapes and the transcription factors associated with their highly correlated gene expression profiles to identify routes through which the similarity may arise. Taken together, this comprehensive dissection of the contributions of specific mSWI/SNF subunits provides an important roadmap for future mechanistic studies and underlines the utility of large-scale single cell sequencing datasets to elucidate and uncover gene expression and chromatin accessibility signatures relevant to human disease states.

II. Introduction

Mammalian SWI/SNF (mSWI/SNF or BAF) complexes are large ATP-dependent chromatin remodeling machines that regulate chromatin accessibility and modulate gene expression. BAF complexes are combinatorial assemblies of 10-15 subunits arranged from the products of 29 genes and exist in three distinct forms: canonical BAF (cBAF), polybromo-associated BAF (PBAF) and non-canonical BAF (ncBAF)[1,2]. While BAF complex subunits work together in a concerted fashion for their overall function of gene regulation, they each likely have unique roles in controlling gene expression as evidenced by i) tissue- and cell-type specific expression of certain BAF complex subunits, ii) coordinated subunit switching at critical stages of differentiation, and iii) the striking mutational pattern of individual subunits in specific subtypes of human cancers and neurodevelopmental disorders[2,3,4,5].

The functional importance of individual BAF subunits in specific disease contexts has motivated a wide variety of studies implicating BAF subunit mutations or loss of expression with disease-associated gene regulation and pathology in relevant cell types. For example, ARID1A has been largely studied in ovarian clear cell carcinoma (OCCC), colorectal carcinoma, and hepatocellular carcinoma[6,7,8], SMARCB1 has been studied in malignant rhabdoid tumors and Coffin-Siris syndrome (CSS)[9,10], SS18 has been studied in the context of synovial sarcoma[11], and SMARCE1 has largely been studied in the context of clear cell meningioma[12]. As such, from these previous studies in varying contexts, it is difficult to disentangle the contribution of each subunit from the context specificity of BAF. While studies aimed to elucidate the roles of BAF complex subunits have been performed in a variety of contexts, the field has not yet elucidated the roles of all BAF subunits in a singular context to comparatively assess the roles of each subunit and determine subunit-, subcomplex-, and module-specific requirements of gene activation, though promise has been shown for comparatively assessing subsets of mSWI/SNF complex subunits in a singular context recently[13,14]. Efforts to perturb all subunits in a singular context have so far been limited by the critical nature of many BAF subunits to cell viability making it difficult to create stable knockout lines, the labor and resource-intensive process for arrayed assays of each subunit perturbation, and batch effects that occur when processing arrayed perturbations and sequencing runs.

To work around these limitations, we use Perturb-seq, a method entailing a pooled CRISPR-Cas9 knockout screen followed by single cell transcriptomic profiling, to assay the effects of each mSWI/SNF subunit on gene expression[15]. We also combinatorially perturb BAF complex subunits to investigate important biological questions about the roles of paralog families, subcomplex identities, and cooperation between subcomplexes. We find that while there are core sets of genes regulated by the three subcomplex types, subunits also have distinct roles in a wide variety of gene expression programs including differentiation, proliferation, cell cycle regulation, signaling, and tissue specific development such as neural, cardiac, and renal. Moreover, for combination perturbations, we identify unique synergistic gene sets for paralog family losses, revealing underlying roles of paralog families that could not be identified by singular perturbation alone. The integration of chromatin accessibility and chromatin binding reveals the overarching circuitry and logic of BAF subunits in the regulation of gene expression. Finally, exploratory comparisons of our BAF-perturbed signatures to primary tumor sequencing datasets identified a set of tumors without BAF mutations that have similar gene programs differentially expressed, suggesting a similar or potentially convergent mechanism in these cancer settings.

III. Results

IIIA. Development of Perturb-seq system for the study of BAF complex subunits

To use Perturb-seq to study the roles of BAF complex subunits in coordinating gene regulation, four guides were cloned targeting each of 28 BAF complex subunits (excluding B-actin due to its presence in multiple critical cell components), together with negative control guides including both non-targeting guides as well as guides targeting intergenic sites (**Figure 2.S1A** and **Table 2.T1**). We packaged the virus for each guide separately (as proposed recently as a strategy for maximizing guide and barcode coupling[16]), and then pooled all guide viruses into a library (for the single perturbation experiment) or used them in arrayed combinations for transduction into Cas9-expressing cells. The combinations were chosen based on their ability to answer biological questions involving the roles of paralog families, subcomplex identities, and subcomplex cooperation or interactions (**Figure 2.1A**). We chose the AML cell line, MOLM-13, for this study due to its lack of BAF mutations, its sensitivity to BAF subunit perturbations (based upon data from Project Achilles[17]), successful CRISPR-Cas9 editing tests, and amenability to standard genomic profiling methods (ChIP-seq, ATAC-seq). We sorted cells for BFP fluorescence two days after transduction to select for successfully infected cells (**Figure 2.S1B-C**). Finally, following 5 days of recovery (total of 7 days post-transduction) we used the 10x Chromium 3' Gene Expression platform to isolate single cells and prepare and sequence libraries. To assign cells to the guides they received, we performed dial-out PCR, which enriched for transcripts specifying the barcodes associated with each guide. We recovered in total ~105,000 cells of good quality (filtering out low quality cells distinguished by high mitochondrial read fractions and low total counts), 30,891 of which had a single BAF subunit perturbation, with a median of 260 cells per perturbation (or approximately 1040 cells per gene, when pooling across guides) (**Figure 2.1B-C**). We explored the role of fitness effects of each perturbation by comparing the titer of each virus in the guide pool to cell recovery, which typically followed expected patterns of fewer cells recovered for more detrimental perturbations and greater than expected recovery of control guides (**Figure 2.S1D**).

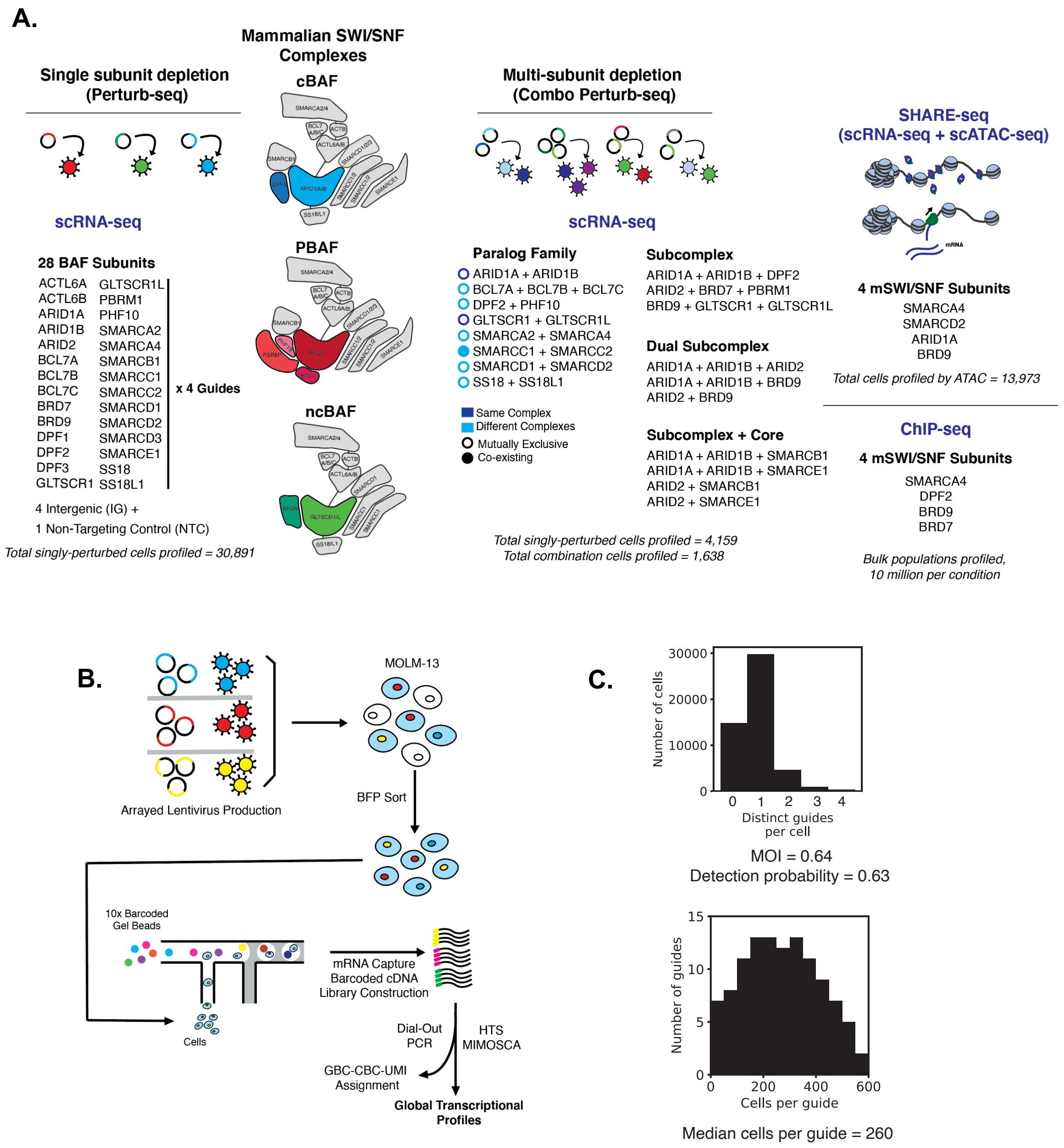


Figure 2.1 Using single cell technologies to investigate the differential roles of BAF subunits in gene regulation.

A. Schematic depicting the experimental strategy for uncovering the roles of individual and combinations of BAF subunits in regulating gene expression. 28 individual BAF subunits and 18 combinations of perturbations were profiled using scRNA-seq in MOLM-13 cells. 4 BAF subunits were profiled using SHARE-seq (scRNA+scATAC-seq) and 4 BAF subunits were profiled using bulk chromatin immunoprecipitation. **B.** Schematic for Perturb-seq experimental process. **C.** Histogram depicting the number of distinct guides captured per cell in the single perturbation Perturb-seq experiment (top) and number of cells recovered per guide condition (bottom).

We assessed the quality of our dataset in multiple ways. First, we determined that there were minimal batch effects across the 15 channels of single cell mRNA capture (**Figure 2.S1E**). We then asked whether the perturbations resulted in lower expression of the targeted genes, and found this to be the case for a large majority of the guides, while the rest included either guides that may be ineffective, guides that work as expected but did not result in NMD-mediated degradation of the targeted gene's transcripts, or guides targeting genes that are expressed at low levels in unperturbed cells, reducing our power to detect differential expression (these genes include DPF1, DPF3, ACTL6B, and SMARCD3 which are not expressed in this cell line, and GLTSCR1/1L which are expressed at very low levels.) (**Figure 2.2A-B and 2.S2**). Second, in general, guides targeting the same gene had similar effects on the transcriptome (**Figure 2.S3**), with exceptions including guides that do not result in lowered expression of the target gene. Third, reassuringly, we found that the set of guides targeting subunits that are not expressed in MOLM-13 cells showed the expected similarity to the negative control guides. Finally, we computed a set of "outlier" scores to quantify the degree to which a perturbation deviates from cells harboring negative control guides, and found good concordance in effect size between guides targeting the same gene, as well as low outlier scores for guides targeting the subunits not expressed. We thus concluded the high quality of the dataset.

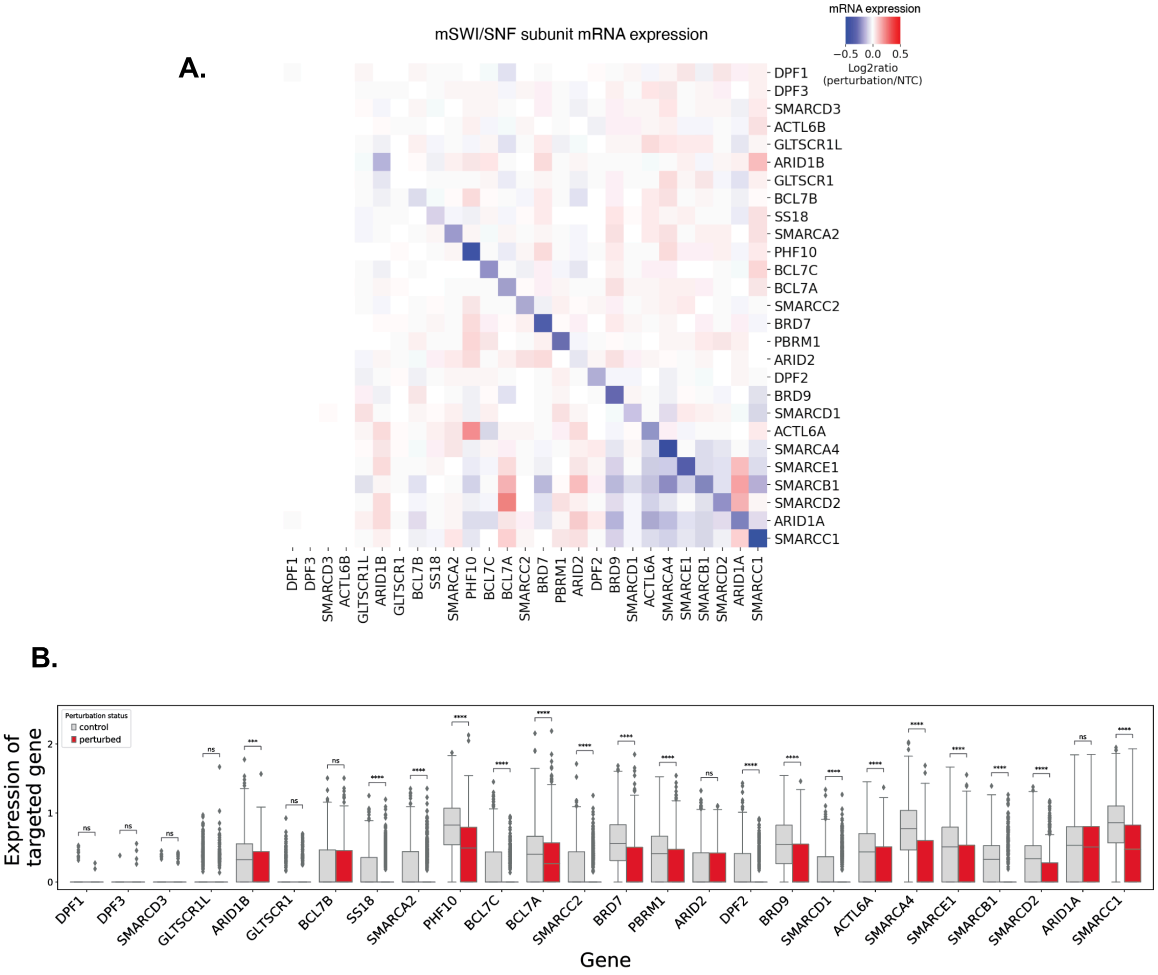


Figure 2.2 Depletion of mSWI/SNF subunits in single perturbation Perturb-seq experiment **A.** Heatmap of expression of mSWI/SNF subunits in single perturbation Perturb-seq experiment, cells are pooled by subunit perturbed. **B.** Boxplot of targeted subunit expression in control guide-receiving cells (left, gray bar) and in the subunit-perturbed cells (right, red bar). Significance of depletion is denoted with an asterisk (*) above each set of bars.

IIIB. Pooled Perturb-seq screen for single mSWI/SNF subunit perturbation in acute myeloid leukemia cells

Distinct modules of perturbations reveal subunits with similar functions and associated effects on gene expression control

After regressing out cell cycle terms and controlling for two distinct cell states within the cell line (as originally characterized in the isolation of this cell line from patient biopsy; we restricted analysis to the large population of cells that makes up 85%+ of all cells[18]), Louvain clustering of the transcriptomes of singly perturbed cells identified unique clusters enriched for various BAF perturbations (**Figures 2.3A** and **2.S4A-B**). Interestingly, clustering based upon global transcriptome similarity identified 2 highly similar groups of perturbations (**Figure 2.3B**). These groups include what we will define as the cBAF/functional core (ARID1A, SMARCA4, SMARCB1, SMARCC1, SMARCD2, SMARCE1), and the ncBAF group (BRD9, SMARCD1). The PBAF subunits BRD7, PBRM2, and ARID2 also form a less strongly correlated but still distinct cluster. Mapping functional similarity using global transcriptome signatures, there is strong agreement between the theoretical functional similarity derived from Project Achilles fitness screens and the transcriptomes of BAF-perturbed cells recovered in this study (**Figure 2.3C**). A notable exception includes the relationships between the ncBAF group of BRD9 and SMARCD1 to the GLTSCR1 subunit, which will be discussed later as a unique case. It is also important to note here that SMARCD1 correlates highly with the noncanonical BAF subunit BRD9 even though SMARCD1 can integrate into in all forms of BAF complexes. While SMARCD1 can exist in any BAF complex, it is the only SMARCD family member that can integrate into ncBAF. So, while SMARCD1 perturbation affects each BAF subcomplex type, it effectively perturbs 100% of ncBAF complexes while only perturbing a smaller fraction of PBAF and cBAF complexes, which can also incorporate SMARCD2 or SMARCD3. Additionally, in the Perturb-seq dataset, SMARCA4 clusters more closely with the cBAF/core subunits than it does with the noncanonical BAF group, while SMARCA4 appears to be more equidistant between both complex types based upon fitness data from Project Achilles.

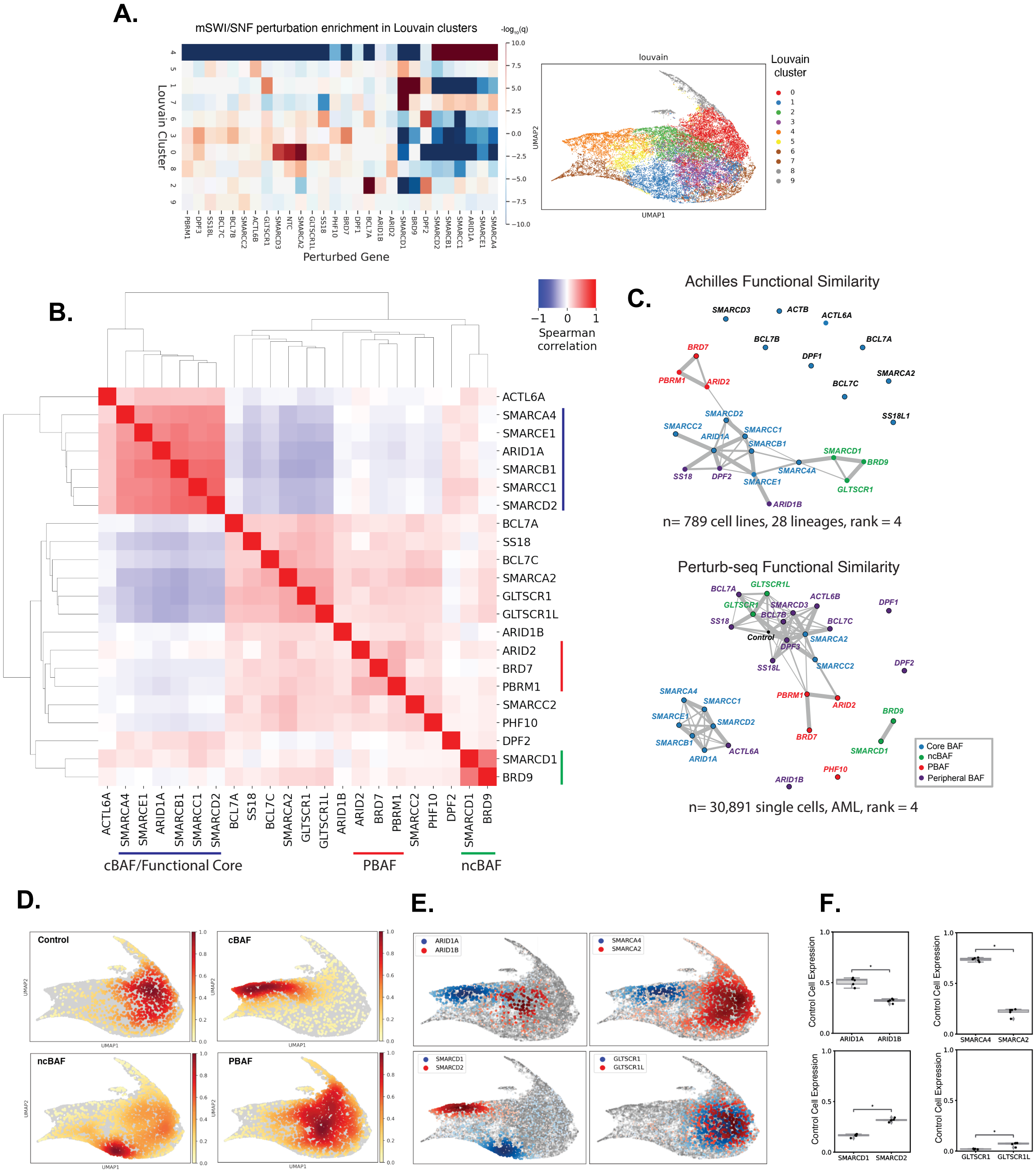


Figure 2.3 Distinct single cell gene expression outcomes associated with mSWI/SNF complex subtypes (cBAF, PBAF, ncBAF) and paralog subunit pairs.

A. (Left) Heatmap showing the enrichment and depletion of each of the perturbations in the clusters from (right), with values representing the significance of the enrichment/depletion via log10(q-values); (right) Low-dimensional representation of singly-perturbed cells in UMAP space, colored by unsupervised Louvain clustering. **B.** Spearman correlation matrix between perturbations by global transcriptome similarity. PBAF, ncBAF, and cBAF/functional core subunits are annotated. **C.** Functional similarity maps based on fitness screening data from Project Achilles (top) and the global transcriptome similarity obtained from the single perturbation Perturb-seq experiment (bottom). **D.** Low-dimensional representation of singly-perturbed cells in UMAP

Figure 2.3 (Continued) space, colored by the density of cells receiving perturbations specific to each of the 3 subcomplex/functional group types: cBAF/Core (ARID1A, SMARCA4, SMARCB1, SMARCC1, SMARCD2), PBAF (PBRM1, BRD7, ARID2), ncBAF (SMARCD1, BRD9, GLTSCR1), and control cells. **E.** Distributions of mSWI/SNF family paralog subunit pairs in UMAP space. **F.** Expression of paralog subunits in control cells for each pair of subunits from (E), with expression represented as log (1 + TP10k). Significant difference in expression is denoted by asterisk (*) above boxes, *p<0.01.

Complex defining groups (cBAF/core, ncBAF, and PBAF) all occupy distinct distributions in UMAP space, with the PBAF distribution being more similar to control cells than the other complex types (**Figure 2.3D**). Additionally, individual paralogs within a family have unique distributions in UMAP space (**Figures 2.3E** and **2.S4C**). While many paralogs within a family seem to have this distinct distribution in UMAP space, those subunits with a lesser effect localize more tightly with the control cells. Similarly, the paralogs where both of the family members have fewer effects are more tightly localized to this control cell region; thus, there is less difference between subunits within a paralog family when fewer effects happen for all family members, like GLTSCR1/1L, or DPF2/PHF10. The differences underlying the distributions of paralogous subunits in UMAP space is likely a combination of effects from actual differences in gene targeting or function as well as expression differences between the paralogs in unperturbed cells (**Figure 2.3F** and **2.S4D**). In general, the paralog that is expressed at a lower level tends to localize more closely to control cells than the more highly expressed paralog, with the exceptions of SMARCD1/SMARCD2 and GLTSCR1/GLTSCR1L as mentioned earlier. SMARCD1 and SMARCD2 are the most distinct paralogous subunits, due to their strong correlations to the ncBAF group and the cBAF/functional core groups respectively (**Figure 2.S4E**).

BAF subunit perturbations affect diverse gene pathways

Perturbations of mSWI/SNF subunits at the single cell level resulted in a wide range of affected gene ontologies, ranging from immune activation, signaling, proliferation, and development terms, to diverse tissue type gene sets including neural, renal, and cardiac terms (**Figure 2.S5A**). Tissue-specific gene sets were intriguing to observe given that this Perturb-seq screen was performed in one cell line (AML cell line MOLM-13), indicating the power of this method to detect changes pertinent to other cell contexts. As expected from functional similarity analyses, mSWI/SNF subunits with high concordance in our dataset similarly affected many gene pathways and ontologies (and in similar directions). Further, while we identified many pathways changing across all subunits within a given functional group, there are also groups of ontologies that are unique for certain subunits.

To better identify the core gene sets underlying the differences in response to mSWI/SNF subunit perturbation, we applied non-negative matrix factorization (NMF) to our single cell transcriptomic dataset (**Figure 2.4A**). To ensure the optimal number of factors describing the range of outcomes of mSWI/SNF subunit perturbations, we 1) assessed the model's ability to recapitulate our dataset and 2) defined the unique sets of genes most strongly defining each factor or program (**Figure 2.4B**). We observed that for cBAF and core perturbations, there were unique down and upregulated programs described by our NMF model not shared by ncBAF complex perturbations, while for ncBAF perturbations, downregulated programs overlapped a subset of the cBAF/core programs while maintaining a strikingly unique signature in upregulated programs (**Figure 2.5A-B**). The upregulated gene sets that most uniquely define ncBAF complex perturbation are those involved in actin filament-based polymerization and processes as well as cytoskeletal and supramolecular fiber organization, potentially demonstrating ncBAF's importance in cell state changes, as has been suggested by a recent screen in the immune cell context[19]. The effects of cBAF/core perturbations most uniquely affected a variety of developmental pathways and processes, ranging from stem cell differentiation to specific tissue-related development terms, suggesting cBAF is a critical complex in governing a variety of important developmental processes. In general, PBAF subunits had the fewest effects on the transcriptomes in this cell type, yet they still maintained a unique perturbation signature. Significant factors for the PBAF subunits BRD7 and PBRM1 overlap to some degree with cBAF/core and ncBAF subunits, especially in the upregulated terms, though a factor most unique to the PBAF subunits BRD7 and PBRM1 was identified (**Figure 2.S5B**).

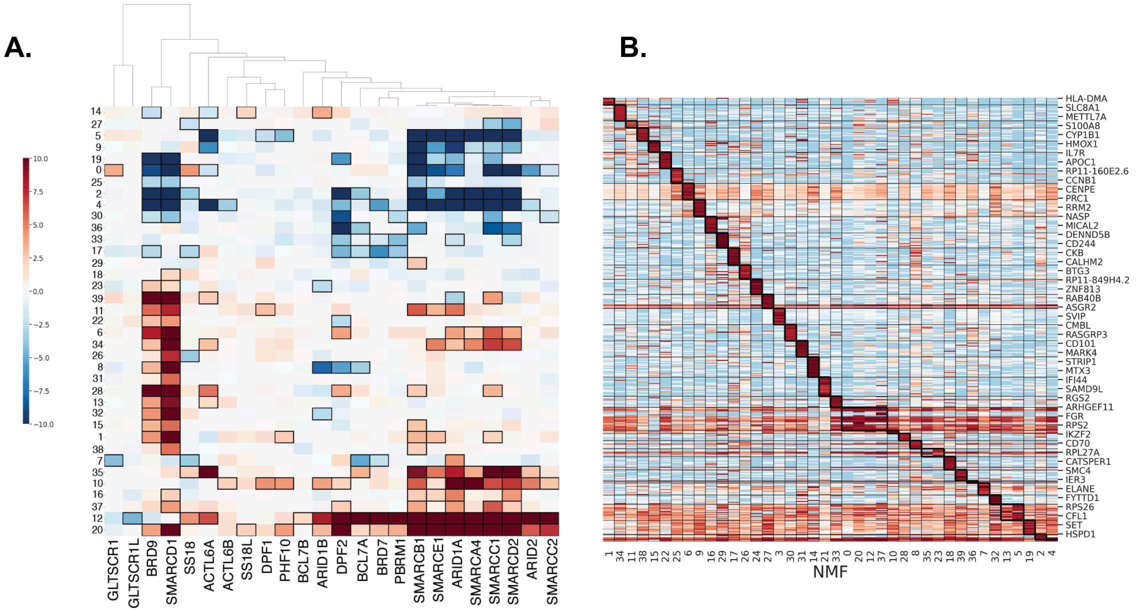


Figure 2.4 Non-negative Matrix Factorization defines programs that differentiate single cell transcriptomes

A. Heatmap showing the activity of gene programs as computed using non-negative matrix factorization (NMF) across a subset of mSWI/SNF subunit perturbations. **B.** Heatmap of NMF programs clustered by enriched genes showing the unique gene sets that define each factor.

To identify gene sets that mark subcomplex identity rather than unique subunit identity, we looked at the intersection between SMARCA4 sites and marker subunits for each subcomplex (ARID1A, ARID2, BRD9) (**Figure 2.5C**). In general, the cBAF subunit ARID1A had the strongest overlap with SMARCA4 gene sets, perhaps suggesting that cBAF has the strongest characteristic signature in this cell type and that the ATPase subunit of this family most dominantly supports the activity of the cBAF complex. There is a smaller yet still distinct set of genes similarly regulated between SMARCA4 and BRD9, and an even smaller overlap between SMARCA4 and ARID2. In general, PBAF subunits had the fewest effects on the transcriptome in this cell type, yet they still maintained a unique perturbation signature. It is important to note that there are many other gene sets changing for each subunit that are not co-regulated by SMARCA4. These sites likely come from unique contributions of mSWI/SNF subunits to complex targeting, activity, or association with other proteins, such as transcription factors.

To identify gene sets that are similarly regulated by all forms of mSWI/SNF complexes and hence mark the collective activity of all three subtypes within the mSWI/SNF family, we overlaid the differentially expressed genes (preserving directionality) of unique subunits from each complex type (ARID1A, ARID2, BRD9) and performed GSEA/GO analyses on the resulting gene sets (**Figure 2.5D**). There were a very small number of genes that were similarly regulated by all unique complex-defining subunits, which included genes from programs involving cell activation in immune response and IL-8 production. The greatest overlap was between ncBAF differentially expressed genes and cBAF differentially expressed genes, which is expected based upon their magnitudes of effects on the transcriptomes globally (compared to PBAF subunit perturbation), and the aforementioned patterning of NMF downregulated and upregulated gene programs for these complex types.

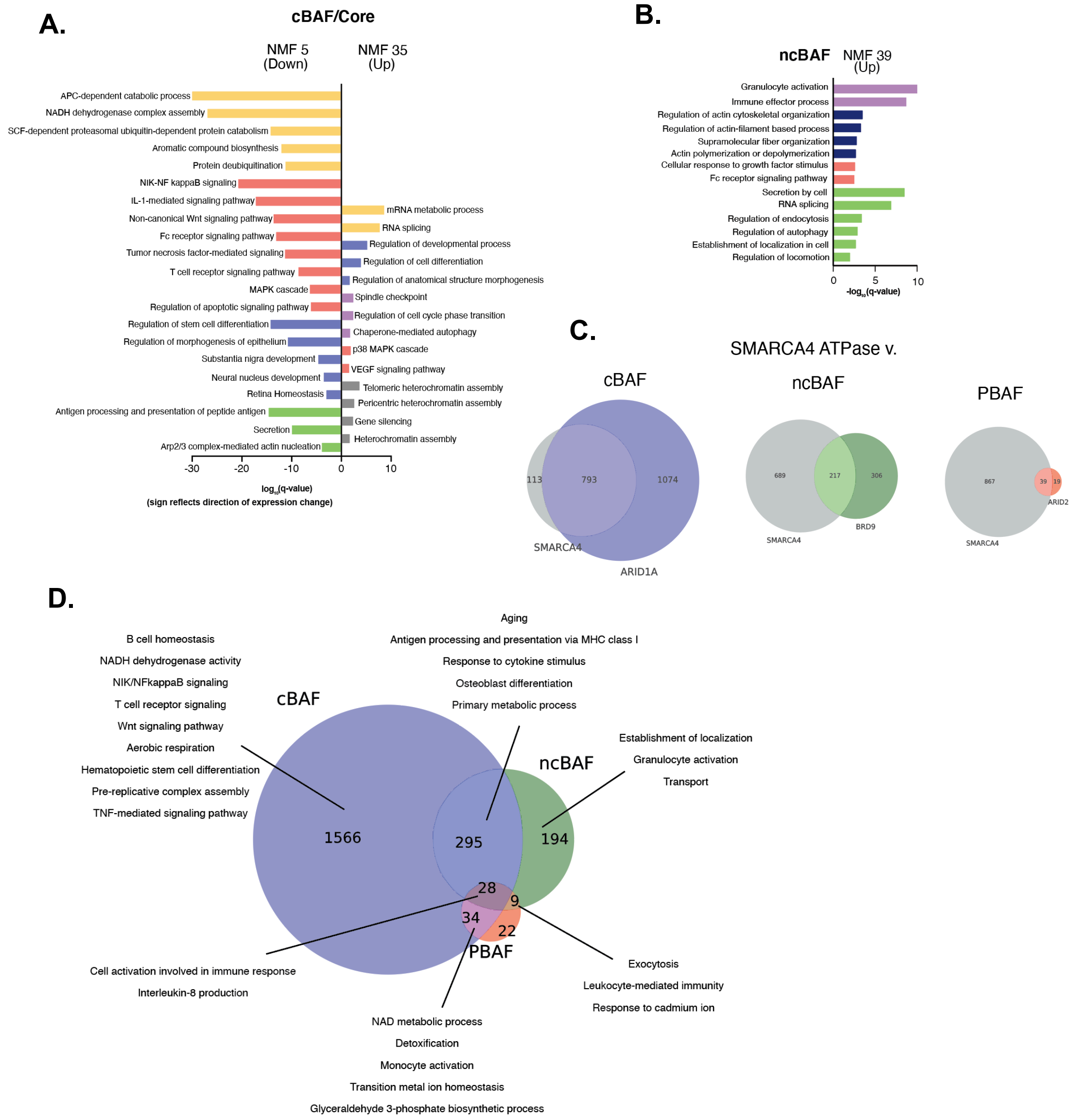


Figure 2.5 mSWI/SNF subcomplexes and subunits regulate distinct gene programs
A. Gene ontologies associated with up and downregulated gene sets (denoted by directionality of sign on x-axis) enriched in the NMF-defined factors uniquely aligned with cBAF/core subunit perturbations. **B.** Gene ontologies associated with the upregulated gene sets enriched in an NMF-defined factor uniquely aligned with ncBAF perturbation. **C.** Venn diagrams of SMARCA4-mediated differentially expressed genes compared to DE genes for complex-specific subunits (preserving directionality of DE genes). **D.** Genes differentially expressed as a function of perturbed subcomplex, compared via Venn diagram with highlighted gene ontologies for the different Venn groups.

Excitingly, using NMF, we identified gene signatures for subunits with very subtle effects overall in this experiment and in the literature to date, such as SS18, DPF2, and SS18L1. Specifically, the perturbation of SS18L1, while having very few overall effects, upregulated many genes in a pathway of dendrite extension (**Figure 2.S5C**). This corresponds nicely to its known role in neural development, where it plays a critical role in the development of dendrites in cortical neurons in a calcium-dependent manner[20]. ACTL6A is also a unique subunit to consider, as it is the only subunit included that is also present in other families of chromatin remodelers (i.e. INO80 and TIP60-p400)[21,22]. In general, it has similarities in NMF program effects with the unique cBAF/functional core NMFs as well as similarities to some of the unique ncBAF upregulated NMF-defined gene programs, making it a unique hybrid in terms of NMF enrichment patterns. Ontologies associated with this perturbation include upregulated terms involving cardiac muscle terms as well as response to glucocorticoid stimulus, while downregulated terms include mitochondrial metabolic processes involving electron transport (**Figure 2.S5D**).

Finally, while the gene sets described above are derived from analyses in which cell cycle terms have been regressed out (identifying factors orthogonal to cell cycle, which is a large source of the heterogeneity in single cell data), we wanted to also identify from these experiments the subunits most implicated in critical steps of cell cycle phase progression. For example, SMARCB1 has been shown to have critical roles in the G1-S phase transition via regulation of cyclin D1 expression[23,24]. Additionally, ARID1A was found to be critical for the transition past the G2/M checkpoint, while ARID1B was found to be unimportant to this transition[25,26]. To determine whether mSWI/SNF subunit perturbations affected cell cycle processes, each cell was scored for a set of marker genes for each cell cycle phase. The proportion of cells in each cell cycle phase was plotted and compared to the control cell distribution (**Figure 2.S6**). For many of the cBAF/core functional group perturbations, there was an increase in the population of non-cycling cells (lack of signature match to other cell cycle phases), with the largest effect from SMARCB1, which has approximately 40% of cells scoring in this phase (control cells were found to have 10% noncycling cells). These data may indicate these perturbations contribute to cell cycle stalling, senescence, or apoptosis. Additionally, we observed subtle changes in cell cycle distribution for a number of BAF perturbations, including ARID1B and SS18, which had subtle over-representations of cells in the G1/S phase. When we evaluated cell cycle signature changes to determine whether there was a subset of genes driving the change (potentially indicating primary targets of the perturbations), we found that signatures of each changed cell cycle phase were uniform over the marker genes. To better understand the order of activation of cell cycle marker genes upon BAF perturbation, an earlier time point or other experimental technique may need to be employed.

In summary, a list of the most important findings from the single perturbation Perturb-seq experiment include: 1) cBAF has the widest number and range of effects in this context, which are largely composed of downregulated genes upon perturbation of constituent subunits. Many of the cBAF-perturbed gene sets include a variety of developmental and differentiation-related terms. 2) ncBAF gene sets overlap with a subset of downregulated gene sets of cBAF perturbations, but are markedly unique in their upregulated gene sets. The most striking enrichment is that of actin fiber processes. 3) SMARCA4 seems to predominantly support the activity of cBAF complexes. 4) Though performed in a singular cell context, this dataset has applicability to other tissue-specific processes as highlighted by neural, renal, musculoskeletal, and cardiac-specific developmental gene sets.

BAF subcomplex binding profiles and chromatin accessibility changes highlight complex-specific roles in genomic architecture regulation

To link the differentially expressed genes identified using Perturb-seq to chromatin accessibility changes that occur upon BAF complex perturbation, we performed SHARE-seq on a set of BAF subunit perturbations spanning complex identities: SMARCA4 (pan-BAF), ARID1A (cBAF), SMARCD2 (cBAF and PBAF), and BRD9 (ncBAF)[27]. We combined this data with chromatin binding profiles for complex-specific subunits to mark localization of each BAF complex type on chromatin. These ChIP-seq profiles include BRD7 (PBAF), DPF2 (CBAF), BRD9 (ncBAF), and SMARCA4 (pan-BAF). RNA signatures identified using SHARE-seq correlated well with the respective single perturbation signatures defined in the Perturb-seq dataset, indicating the following ATAC signatures are tied to the gene expression changes observed in mSWI/SNF perturbed cells (**Figure 2.S7**).

mSWI/SNF perturbed single cells clustered by chromatin accessibility profiles highlighted two similar groups: the CBAF and core perturbations (ARID1A, SMARCA4, SMARCD2), and the BRD9 perturbation (**Figure 2.6A-B**). Upon analysis of accessibility over subcomplex-specific and co-subcomplex bound sites, accessibility is largely decreased over a host of these bound sites, especially for cBAF-specific sites upon perturbation of ARID1A and SMARCD2 (**Figure 2.6C-D**). There were fewer effects in general over ncBAF-specific sites, but BRD9 perturbation did decrease accessibility over more ncBAF sites than gained accessibility. It has been suggested that perturbation to BRD9 has only subtle effects on chromatin accessibility, which is consistent with these findings[19]. Interestingly, we found increases in accessibility over ncBAF-bound sites upon cBAF/core perturbation, suggesting the loss of functional cBAF complexes increases regulation over ncBAF sites, perhaps through stoichiometry or some other mechanism. There was little skew in the change in accessibility over PBAF-bound sites for any of the perturbations, including SMARCA4 and SMARCD2.

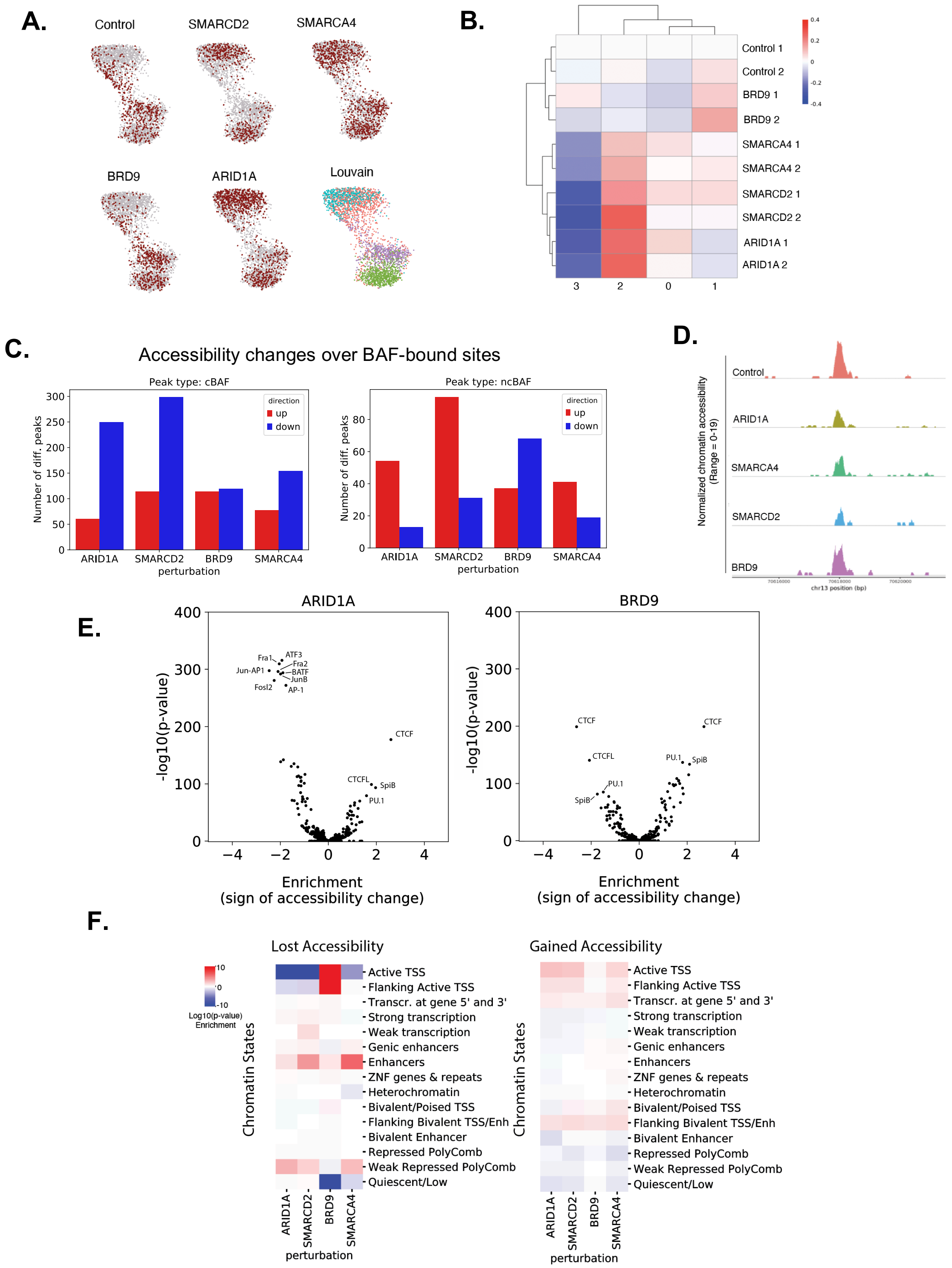


Figure 2.6 SHARE-seq reveals chromatin accessibility changes linked to gene expression signatures and identifies differential chromatin states and transcription factor motif accessibility associated with mSWI/SNF perturbation

A. Low-dimensional representation of localization of mSWI/SNF perturbed cells and control cells (left) and Louvain-identified clusters (right). **B.** Heatmap of proportions of single cells in Louvain-

Figure 2.6 continued defined clusters for chromatin accessibility. **C.** Bar graph of accessibility changes over unique complex-bound sites using chromatin binding profiles for DPF2 (cBAF), BRD7 (PBAF), BRD9 (ncBAF), and SMARCA4 (pan-BAF). **D.** Sample track for the accessibility changes over a cBAF/core lost accessibility site **E.** Motif enrichment over sites of changed accessibility for ARID1A and BRD9 perturbation conditions. **F.** Heatmap of chromatin states enriched in regions of increased and decreased accessibility upon BAF perturbation.

To better understand the hallmarks of sites with differential accessibility upon mSWI/SNF subunit perturbation, we analyzed the motifs within differentially accessible peaks and also probed the chromatin states associated with these peaks (**Figure 2.6E-F**)[28,29]. Notably for the cBAF-specific sites that lost accessibility upon subunit perturbation, the topmost motifs in were for AP-1 factors including FOS and JUN transcription factors, which corresponds nicely with identified relationships between BAF complexes and AP-1 factors[30]. For the ncBAF-specific sites that lost accessibility upon BRD9 perturbation, the topmost motifs were for CTCF and CTCFL (BORIS) which are core architectural proteins that establish 3D chromatin architecture through long range interactions[31]. Interestingly, the widespread loss of accessibility over ncBAF-specific sites was not seen in SMARCA4 perturbation, suggesting that perhaps the complex-defining subunits of ncBAF are critical in this association or relationship with CTCF/CTCFL factors, or that SMARCA4 is less critical to ncBAF identity or functions than it is to other complex types. Chromatin states associated with decreased accessibility and cBAF perturbations included those marked by enhancers, while BRD9 perturbation affected accessibility at active transition start sites (TSS) and flanking active TSSs.

Taken together, insights from SHARE-seq and chromatin binding analysis strengthen the overall picture of mSWI/SNF gene regulation, as chromatin accessibility changes were directly tied to gene expression changes, and binding profiles aligned nicely with expected trends in the accessibility profiles. Further, we show that ncBAF perturbation largely affects accessibility over CTCF and CTCFL sites, and perturbation of cBAF increases accessibility over these sites, potentially implicating stoichiometry of these subcomplex forms in generation of accessibility. Finally, SMARCA4 perturbation did not largely affect occupancy over ncBAF-bound sites, suggesting its activity is less critical in this subcomplex type.

IIIC. Arrayed Perturb-seq screen for combinatorial mSWI/SNF subunit knockouts in acute myeloid leukemia cells

Combinatorial perturbations identify paralog family traits and synergistic gene sets

In order to answer a series of biological questions best addressed by simultaneously perturbing multiple BAF subunits, we performed an arrayed combinatorial perturbation experiment to answer questions about the four following categories of subunits i) paralog families ii) all unique subunits of a complex type iii) unique subunits of two complexes and iv) one unique complex plus core subunits. The main goals for these categories were to perturb either a whole paralog family, all unique subunits for a complex type, perturb unique subunits for 2 subcomplex types, or perturb a unique complex subunit plus a core subunit to assess the roles of core subunits in the various complex types. There are of course many more combinations that would be interesting and insightful, however these combinations begin to answer many interesting questions in BAF biology.

In order to perform combination perturbations with the Perturb-seq platform, virus was made in an arrayed fashion and cells were transduced in an arrayed manner with each combination of BAF guide viruses. Instead of increasing MOI as in the original Perturb-seq methodology, we used the best guide for each subunit in each combination condition. This allowed us to profile far fewer cells; increasing the MOI would have necessitated sequencing a vast number of cells to recover sufficient cells for biologically insightful combinations. Using similar metrics to the method for filtering out poorly performing guides, we identified the best performing guides for the combination experiment. Our criteria for the best guides were that they 1) had high correlation to other guides in the set (based on global transcriptome similarity) and 2) effectively knocked out their target subunit as assessed by depletion of the target's mRNA transcript (see **Figure 2.S2**). After infection, cells were pooled and sorted for BFP fluorescence, then collected at 7 days post-transduction for single cell transcriptome profiling. While most combinations were recovered in sufficient quantities for analysis, some were sparser within the pool (**Figure 2.S8A**). We performed down sampling analysis to determine whether the smaller quantities of cells (~30) could still robustly represent the transcriptomic changes occurring in cells upon BAF perturbation (**Figure 2.S8B-C**). Using the sample set of genes that mark cell cycle and also the adjusted Rand index (ARI) for varying degrees of down sampling, we concluded that we can confidently and robustly detect transcriptomic changes for the quantities of cells present in this dataset.

A few different patterns emerged when visualizing single and combination perturbed cells in UMAP space (**Figure 2.7A-B**). Some combination perturbations, like ARID1A/ARID1B localized in a similar distribution as one of the constituent individual perturbations (ARID1A) whereas other combinations like SMARCC1/SMARCC2 occupied a new distribution in UMAP space. For many of the sets of paralogs, there were genes newly regulated upon combination perturbation beyond the expected additive effects of these perturbations, as shown by residual expression heatmaps (**Figure 2.7A-B** and **2.S8F**). The only paralog family set that appeared to be largely additive in nature is ARID1A and ARID1B. The remaining sets of paralogs had a large number of both upregulated and downregulated synergistic gene sets. For most dual paralog loss conditions, immune system activation terms were upregulated upon loss while markers of cell cycle phases were downregulated. These sets of genes make sense, as multiple perturbations are likely more deleterious to fitness (and thus decrease cell cycling) and a greater number of perturbations likely incite a greater immune response to the DNA damage involved in CRISPR-Cas9 editing. However, other synergistic gene sets excitingly include heme stem cell differentiation programs and signaling pathways such as Wnt and NF-KB (for SMARCC1/SMARCC2 combination), Heme/lymphoid development and motility (for ARID1A/ARID1B combination), and secretion and metabolism (for SMARCD1/SMARCD2).

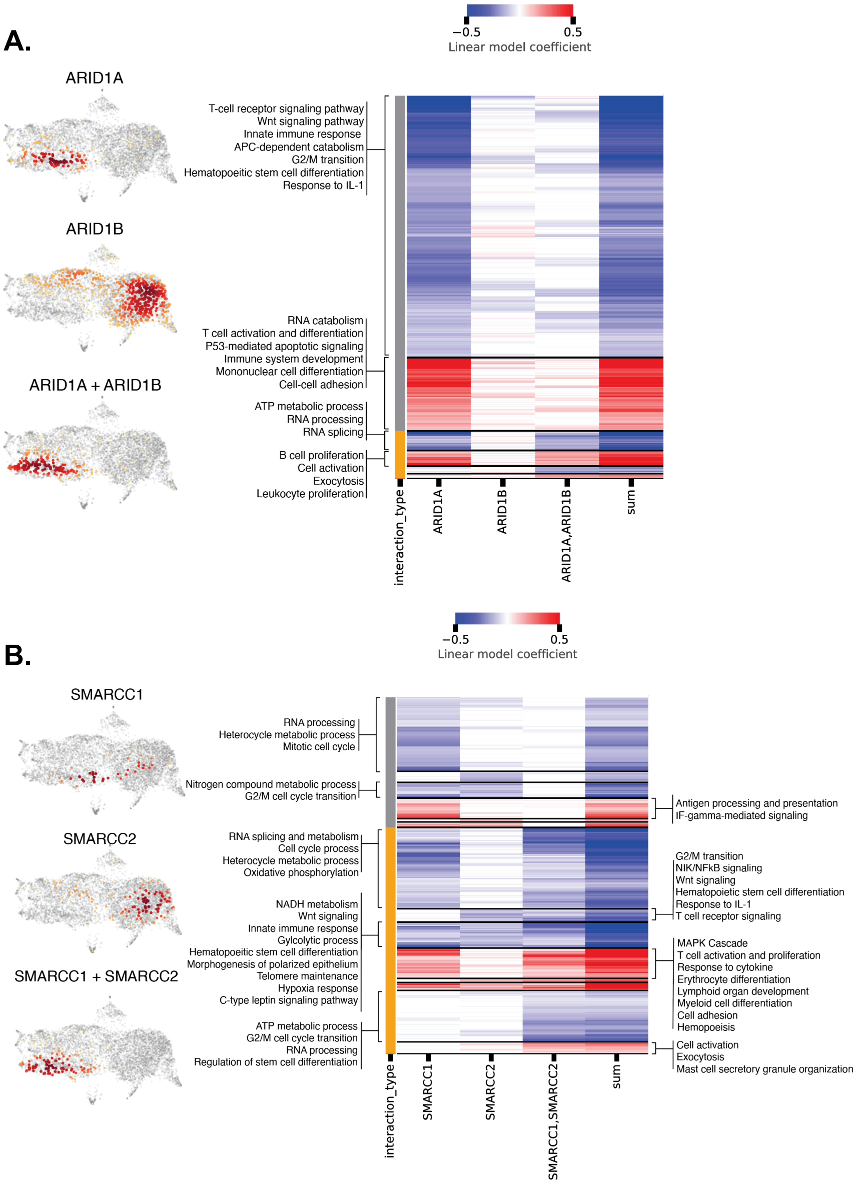
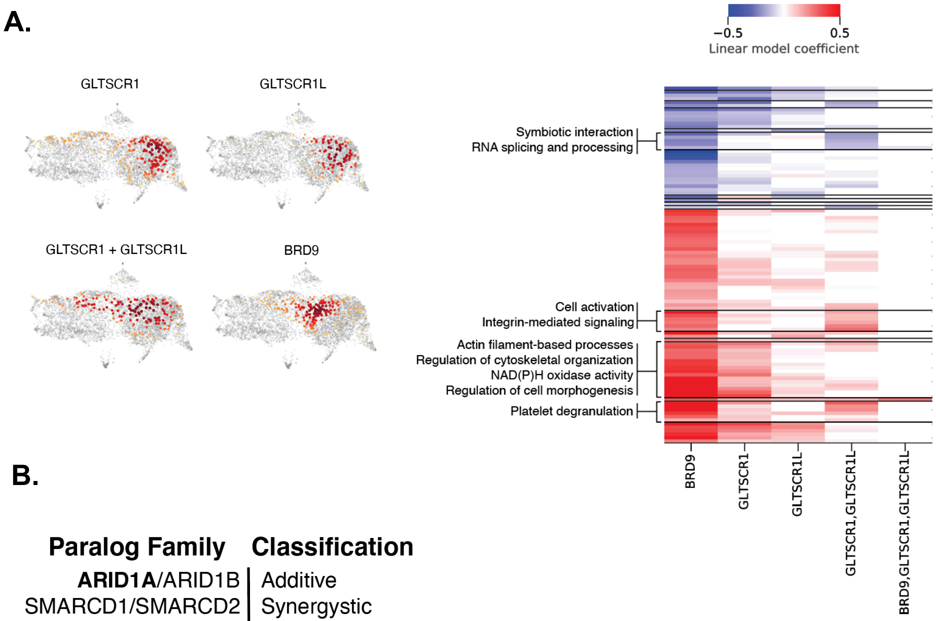


Figure 2.7 Combinatorial perturbations reveal additive and synergistic roles for mSWI/SNF paralogs

A. Low-dimensional representation in UMAP space of singly and doubly ARID1A/ARID1B perturbed cells recovered in the combination perturbation study, colored by the density of cells with specific perturbations (left). Heatmap of linear model coefficients for expression changes attributed to single perturbations or their associated combination perturbation (right). The coefficients associated with the combinatorial perturbations represent additional variation beyond additive effects of the individual perturbations. Interaction type is colored orange for synergistic gene sets or gray for additive gene sets. Gene clusters are annotated with gene ontologies. **B.** Same as (A), for SMARCC1/SMARCC2 paralog family.

One of the most interesting paralog family cases is the dual perturbation of GLTSCR1 and GLTSCR1L, which neither individually have many effects on the transcriptome. However, upon dual perturbation, there are a variety of new up and downregulated genes (**Figure 2.8A**). Interestingly, many of these genes are similarly regulated by BRD9. This suggests interesting characteristics specifically for the GLTSCR1 family, as they likely are able to compensate well for one another in the case of singular loss, but when both are lost, we see gene sets congruent with the ncBAF perturbation phenotype. According to these paralog family behaviors, we classified each family as behaving in an additive vs synergistic manner (based on the frequency of these combinatorial gene set behaviors) (**Figure 2.8B**).

The interpretation of these synergistic gene sets could come down to a few different theories: either the paralogs are able to compensate functionally in some way for each other (having some degree of functional redundancy) in which case we are best able to determine redundant gene targets in the context of whole paralog family loss. Alternatively (or additionally), these paralogs may be filling an important structural role within the assembly pathway, thus maintaining a degree of structural redundancy. In the case for structural redundancy, the additional differentially expressed gene targets may be due to the loss of structural integrity of BAF complexes upon paralog family loss.



B.

Paralog Family	Classification
ARID1A/ARID1B	Additive
SMARCD1/SMARCD2	Synergystic
SMARCA2/**SMARCA4**	Synergystic
SMARCC1/SMARCC2	Synergystic
GLTSCR1/GLTSCR1L	Synergystic

Figure 2.8 Combinatorial perturbation of GLTSCR1/1L captures ncBAF perturbation phenotype with synergistic gene sets

A. Distribution of singly and doubly perturbed GLTSCR1/GLTSCR1L cells in UMAP space, shown compared to the distribution of BRD9 cells (left). Linear model of gene expression attributed to the GLTSCR1/L singular and double perturbations as compared to BRD9 perturbation (right). **B.** Table of classifications of paralog families based upon behaviors observed in combination perturbation study.

IIID. Model of activation of the 3 mSWI/SNF subcomplex types in all perturbation conditions

Additional combination perturbations were performed to identify how unique subunits contribute to complex character, how core subunits relate to each complex type, and to determine how complexes may interact or cooperate. To this end, we targeted: all unique subunits of a complex type, unique complex subunits plus a core subunit, or unique subunits belonging to two complex types. To assess the contribution of each unique subunit to a complex type, vector diagrams were created to represent the direction and localization of the mean of each perturbed population in UMAP space (**Figure 2.9A**). Here, we see an excellent comparison to the assembly pathway in the ARID1 subunits plus DPF2 perturbation. DPF2 is the final subunit to assemble onto cBAF complexes, and it cannot bind without an ARID1 member on the complex[32]. Thus, we see that both the dual ARID1 loss and dual ARID1 loss plus DPF2 perturbation result in the highly similar localization and no additional changes upon perturbation of DPF2 on top of the ARID1A/1B combination perturbation (**Figure 2.9B**). This indicates that we can potentially resolve some structural information from the transcriptomic outcomes. For unique ncBAF subunits, we see the GLTSCR1 + GLTSCR1L combination perturbation occupy distinctly different space from the individual perturbations of these subunits, consistent with earlier findings about these paralogs only achieving ncBAF perturbation-like transcriptomes upon dual perturbation.

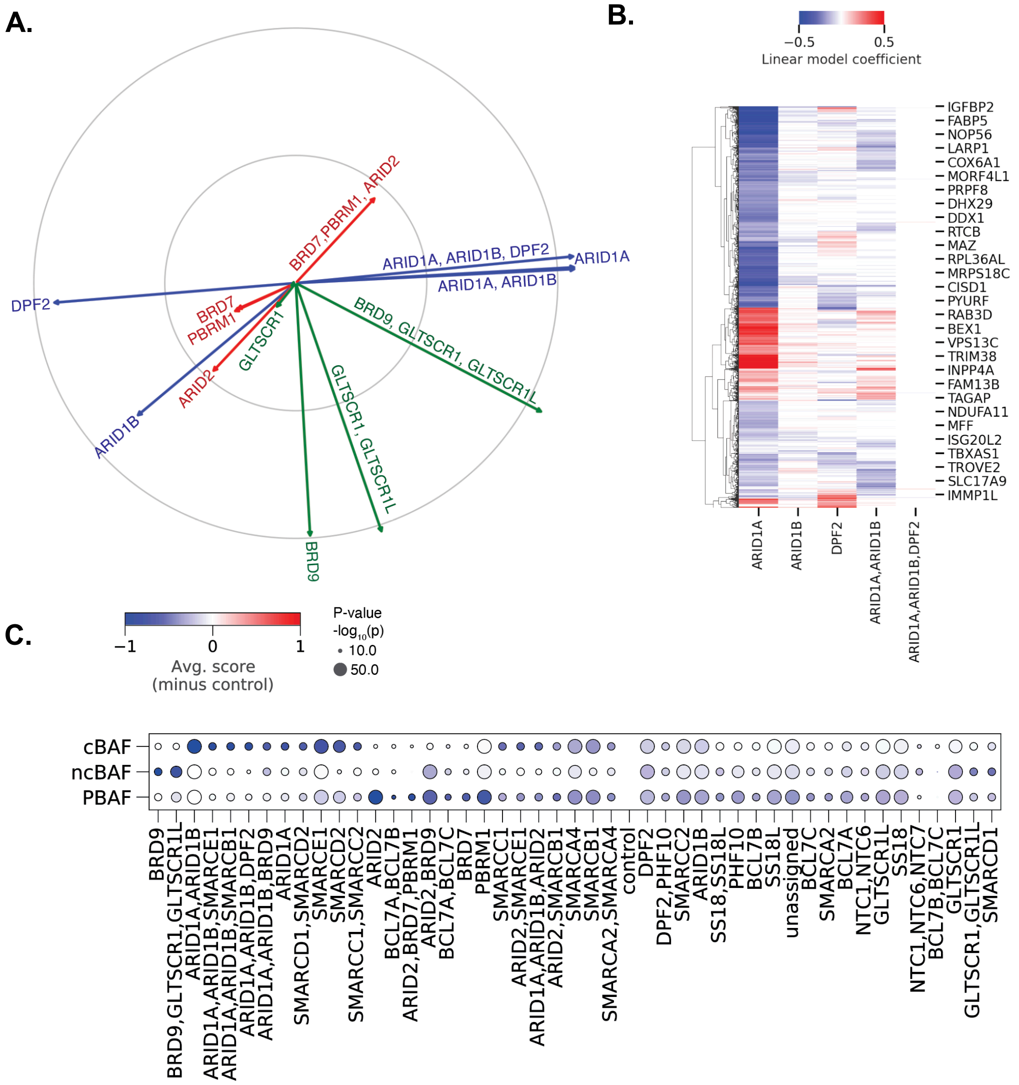


Figure 2.9 Combinatorial perturbations within and across mSWI/SNF complex types yield assembly-related insights and reveal altered subcomplex activity in response to mSWI/SNF subunit perturbations

A. Radar plot with velocity vectors defining the direction and magnitude of impact for each of the combination perturbations targeting all unique subunits within a single complex type. **B.** Linear model heatmap defining the gene sets differentially regulated upon depletion of all three cBAF-specific complex subunits, ARID1A, ARID1B, and DPF2, illustrating the lack of effect beyond ARID1A/B combination perturbation, consistent with insights from assembly of mSWI/SNF complexes. **C.** mSWI/SNF subcomplex activity scores by subunit perturbation calculated using a logistic regression classifier to distinguish complex activity levels. Classifier was trained for each complex based on the following perturbations: ARID1A + ARID1B combination perturbation (cBAF), ARID2 perturbation (PBAF), and BRD9 perturbation (ncBAF).

Finally, we created a model to compute the activity of each of the three complex types for each perturbation condition (**Figure 2.9C**). This was accomplished by training a logistic regression classifier to identify whether cells were likely to be control cells or BAF perturbed cells, and this classifier was trained on unique complex-defining subunits including ARID2, ARID1A + ARID1B, or BRD9 (**Figure 2.S9**). This model accounted for unique subunits and their effects on their various complexes very well, even when the signature was of a smaller magnitude as is the case for PBAF perturbations. Dual subcomplex perturbations followed expected trends in this model, where both subcomplex types were affected in activity upon their dual perturbation. Excitingly, we were able to resolve differences in the activities of different complex types upon perturbation of shared subunits. For example, SMARCB1 depletion equally affects both PBAF and cBAF, while SMARCE1 more harshly impacts the activity of cBAF than PBAF. SMARCD2 similarly has stronger effects on cBAF activity as compared to PBAF. SMARCD1 has the most dramatic impact on ncBAF but still impacts both cBAF and PBAF to a lesser extent, consistent with the knowledge that SMARCD1 is the only paralog that can incorporate into ncBAF. However, since it can still be incorporated in cBAF and PBAF, it makes sense that it would also impact the function of these complexes. SMARCA4 appears to affect the activity of both cBAF and PBAF, while hardly affecting the activity of ncBAF complexes. This corresponds nicely to the earlier findings in the single cell ATAC analysis that accessibility over ncBAF sites was largely unchanged upon SMARCA4 perturbation. Finally, the single perturbations of GLTSCR1 and GLTSCR1L again are mostly unremarkable, while we see a stronger ncBAF-perturbed characteristic upon dual loss. These insights suggest that shared subunits may have different roles or different magnitudes of effects in subcomplex types, and further investigation into these findings could yield insights as to why some subunits may be more important to one complex type over another.

IIIE. Disease relevance of Perturb-seq-defined mSWI/SNF perturbation signatures

Due to the extensive roles of mSWI/SNF mutations in a wide variety of human cancers, we sought to determine the relevance of our Perturb-seq-derived BAF perturbation signatures to RNA expression profiles of tumors in The Cancer Genome Atlas (TCGA)[33]. We first compared our NMF-defined mSWI/SNF subunit perturbation signatures to RNA expression profiles of rare, BAF-driven cancers including malignant rhabdoid tumors (MRT), small cell carcinoma of the ovary hypercalcemic type (SCCOHT), and epithelioid sarcomas (EpS) to determine if our dataset would recapitulate their gene expression signatures as a proof of concept[11]. We found good concordance between the gene expression signatures from our experiment and the rare cancer sequencing dataset, especially for MRT, with all tumors having high cosine similarity to SMARCB1 perturbation (which is the driver mutation in this cancer type) (**Figure 2.10A**). There was also high cosine similarity between SCCOHT tumors and the SMARCA4 perturbation signature, which was encouraging as these tumors are marked by loss of SMARCA4 expression (in addition to loss of SMARCA2). We then queried the RNA-sequencing database for all tumors in TCGA with signatures similar to our NMF-defined gene programs (**Figure 2.10B**). We found sets of tumors with high cosine similarity (>0.95) to cBAF perturbations as well as ncBAF perturbations, and excitingly, they spanned different tissue types and subsets of these tumors lacked mutations in any mSWI/SNF component (**Figure 2.11A**). We then sought to identify mutations that occur in these groups of tumors that could explain similarities to our mSWI-SNF perturbation signatures. We compared the mutational landscapes of each of the tumors to the background mutational rate and landscape in matched tumor types to identify enriched mutations in these specific samples. We identified a host of mutations overrepresented in these tumors, including many transcriptional coactivators among other proteins with varying enzymatic functions (**Figure 2.11B**).

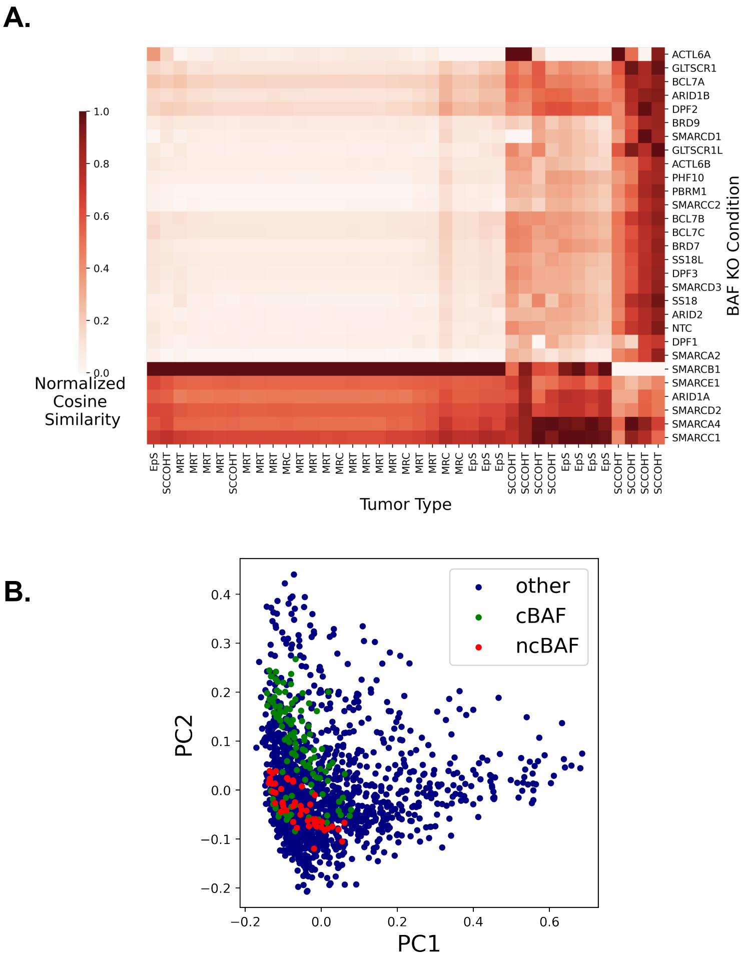


Figure 2.10 Perturb-seq-defined mSWI/SNF subunit perturbations signatures recapitulate BAF-driven rare cancer signatures and identify a set of TCGA tumors with high similarity that lack mutations in mSWI/SNF subunits

A. Heatmap of cosine similarity scores between gene program signatures identified by non-negative matrix factorization for mSWI/SNF perturbations in our perturb-seq dataset and RNA expression profiles of rare, mSWI/SNF perturbation-driven cancers. **B.** PCA plot of gene expression profiles of tumors in TCGA without mSWI/SNF mutations that have >0.95 cosine similarity to any mSWI/SNF signature, normalized to RNA expression in their matched normal tissue types.

For the most abundant class of TCGA tumors with high similarity to each of the cBAF and ncBAF perturbation signatures, we used LISA to identify the transcription factors associated with the regulation of genes that are differentially expressed in these tumors relative to the general set of tumors in the same class[29]. We found many factors that are known or highly suspected to interact with mSWI/SNF complexes or co-opt their activity, along with other transcription factors previously unexplored in relation to mSWI/SNF complex activity (**Figure 2.11C**)[34]. For example, the transcription factors associated with cBAF perturbation-similar gene sets include ERG and FLI1, both factors known to hijack BAF complex activities to achieve cancer-associated gene expression in other tumor types[35,36]. While this analysis is exploratory in nature, these mutations and transcription factor-associated gene expression programs could in theory point to a convergent disease mechanism hinging on mSWI/SNF complex activities. We are continuing work to refine these analyses and ensure robust signatures and similarities. It is encouraging to note that even though the screen was performed in a singular disease context, we found highly similar signatures across a wide variety of tissue types, showing that screens of this nature have the potential to be more broadly applicable to other contexts. This dataset could be used to mine other gene expression datasets where BAF is suspected to control critical processes, both for disease relevant contexts as well as for more general biological insights.

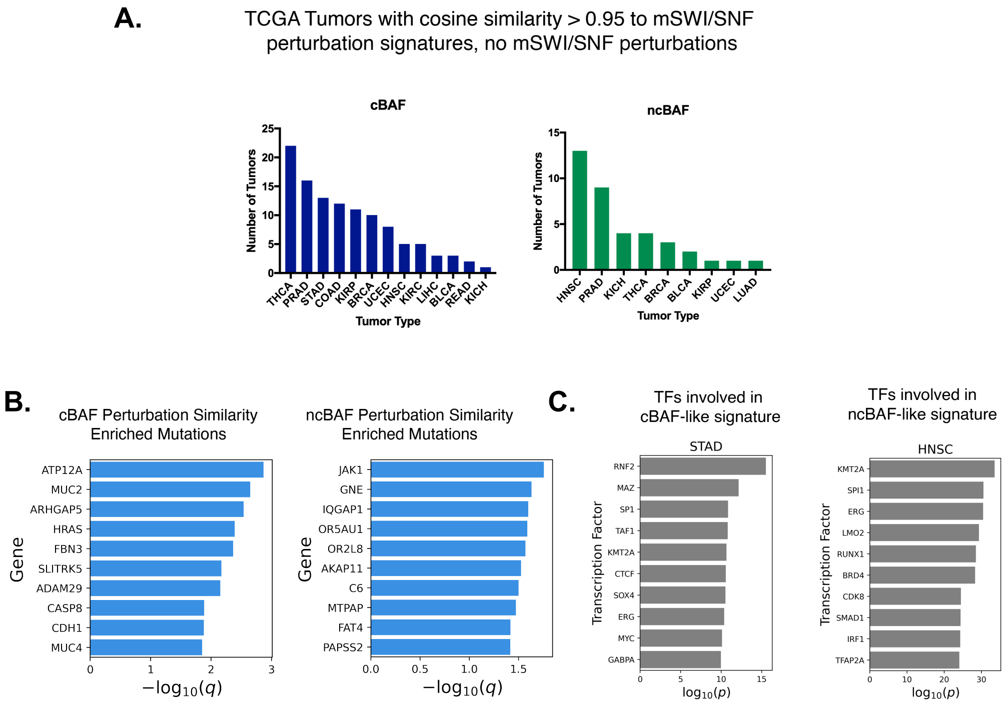


Figure 2.11 Analysis of TCGA tumors lacking mSWI/SNF perturbations with high cosine similarity to Perturb-seq signatures

A. Bar graph of the number of each tumor type without any mSWI/SNF perturbations that have high cosine similarity (> 0.95) to mSWI/SNF perturbation signatures for cBAF (left) and ncBAF (right) complex types. **B.** Bar graph showing the significantly enriched mutational landscape of tumors with high similarity scores (cosine similarity > 0.95) to cBAF (left) and ncBAF (right) perturbations. Mutation enrichment was calculated with respect to the mutational background within each tumor class using a hypergeometric test with the Benjamini-Hochberg correction. **C.** LISA analysis for transcription factors that regulate the gene signatures for tumors highly similar (cosine similarity >0.95) to our dataset for tumors without BAF mutations.

IV. Discussion

Taken together, these studies represent an effort to holistically dissect the contributions of each subunit, functional module, and paralog family within mSWI/SNF complexes in a singular context to enable this kind of systematic study. In summary, our dataset recapitulated predicted functional similarity relationships based on fitness screening and additionally resolved the gene sets driving these predicted similarities. We functionally resolve the three subcomplex types, and show that cBAF has the majority of effects in this context, while ncBAF exhibits a subset of cBAF character in its regulation of the same downregulated sites, but maintains its unique signature through upregulation of many processes, especially those including supramolecular fiber organization and actin polymerization. Analysis from this study also suggests that SMARCA4 is less important for ncBAF activity than it is for cBAF or PBAF activity. We arrive at this conclusion from multiple pieces of evidence including SMARCA4 association with cBAF/core signatures and the greatest overlap with ARID1A-regulated genes, the lack of chromatin accessibility changes over ncBAF occupied sites upon SMARCA4 knockout, and the lack of depletion in ncBAF activity score in the logistic regression classifier for subcomplex activity while SMARCA4 loss equally affected the activities of cBAF and PBAF.

We characterize additive and synergistic gene sets for mSWI/SNF paralog family perturbations, with ARID1A/B as the only paralog family with a majority of gene sets being additive upon perturbation. The interpretation of synergistic gene sets could include that paralogs may be functionally redundant to some degree, so the differential regulation of these gene sets may only be seen upon complete family loss. Alternatively (or additionally), the paralogs may maintain a level of structural redundancy, and synergistic gene sets may arise due to the newly destabilized or disassociated mSWI/SNF complexes upon completely losing a critical structural component in the form of whole paralog family loss, especially in subunits that bind early on in the assembly pathway, like SMARCC1/C2 or SMARCD1/D2. A unique behavior was found upon dual loss of the GLTSCR1/1L subunits, where neither perturbation alone had much of an effect on the transcriptome, but upon dual perturbation, cells became more like ncBAF-perturbed cells, as marked by BRD9 loss. This indicates that GLTSCR1 and GLTSCR1L may be able to functionally compensate for one another well in the case of singular loss, and there is minimal defect conferred to ncBAF activity until both GLTSCR1 paralogs are lost.

Targeting all unique subunits of a subcomplex type showed how much each conferred unique properties to their subcomplex. Specifically for cBAF, the loss of DPF2 on top of ARID1A and ARID1B was negligible, consistent with assembly insights that DPF2 cannot bind in the absence of ARID1 integration into the complex. Other combinatorial perturbations were useful in training and assessment of the logistic regression classifier for subcomplex activity. Through this classifier, we learned that while some shared subunits have similar effects on their constituent complex types, like SMARCB1 on both PBAF and cBAF, other subunits, like SMARCE1, affected one of their subcomplex types much more (cBAF, in this case). This opens up hypotheses to be tested about the functional roles of shared subunits within different subcomplex types.

Combining these insights with chromatin accessibility and binding profiles allowed for the identification of transcription factor motifs in regions of changing accessibility as well as the linked behaviors of binding, changes in chromatin accessibility, and finally the resultant effects on gene expression when mSWI/SNF subunits are perturbed. We found that BRD9 perturbation most strongly affects accessibility over CTCF and CTCFL sites, while ARID1A perturbation most strongly affects accessibility over FOS/JUN/AP-1 sites. Interestingly, upon cBAF depletion (ARID1A), there was an increase in accessibility over ncBAF-bound sites, suggesting that the stoichiometry of complexes could be impacted upon depletion of a specific subcomplex type.

Finally, the use of this dataset to identify similar signatures across TCGA-cataloged tumors highlighted a myriad of mutations and transcription factors that mimic the loss of BAF function. While this analysis is exploratory in nature, it was exciting to see top transcription factors associated with the similar gene expression profiles that are already known to hijack BAF complexes in cancer settings (such as FLI1, which forms a fusion oncoprotein in Ewing sarcoma, or ERG, which forms a fusion oncoprotein in prostate cancers) or ones that were highlighted under sites of changed chromatin accessibility in the SHARE-seq experiment, such as CTCF, SPI1, and CEBPB. These highlighted mutations and transcription factors will need to be further studied to determine whether they interact with or somehow converge with mSWI/SNF complex functions.

The datasets generated in this study still hold lots of potential for future mining of import functional roles and relationships between mSWI/SNF complexes, the logic of mSWI/SNF subunit and complex perturbations, and specific gene targets or gene sets critical for insight into normal or disease-associated biology. Integration of similar datasets in different cell contexts and at different time points would allow for identification of even wider sets of gene targets and pathways, give context for timing of the effects of each BAF subunit perturbation, and allow for greater discovery relating to BAF-mediated transcriptional circuitry. Finally, it would also be exciting to determine BAF complex subunit functions in dynamic processes such as differentiation or immune activation to determine how these complexes govern processes marked by rapid changes in chromatin accessibility and gene expression. Taken together, these insights are a strong foundation for the understanding of comparative roles of mSWI/SNF complex subunits, and additionally provide new hypotheses for both functional roles of subunits on a basic science level as well as potential new mechanisms for disease-associated biology.

V. Materials and Methods

Data and code

Code reproducing all analysis and figures in this paper is at https://github.com/broadinstitute/Perturbseq_BAF_complex.

Cell Lines and Culture

HEK293T and HEK293T LentiX cells were grown in DMEM medium (high glucose, no glutamine) supplemented with 10% FBS, 1% Penicilin-Streptomycin, 1% GlutaMax, 1% Sodium Pyruvate, 1% HEPES, and 1% Non-essential Amino Acids (NEAA) (All TC reagents from Gibco). MOLM13 cells were grown in RMPI 1640 medium (no glutamine) supplemented with 10% FBS (Omega), 1% Penicillin-Streptomycin (Gibco), and 1% GlutaMax (Gibco). Cells were maintained in an incubator at 37 degrees Celsius with 5% Carbon Dioxide. HEK293T were obtained from ATCC, HEK293T LentiX cells were obtained from Clontech, and MOLM-13 cells were a generous gift from Dr. Jay Bradner (Broad Institute).

Plasmid Construction and Cloning

The Perturb-seq vector (pBA439, Addgene #85967) with 18 nucleotide barcode (pBA571, Addgene #85968) was used to generate the BAF subunit-targeting and control guides for the Perturb-seq experiments. The barcoded vectors were digested with BstXI (NEB) and BlpI (NEB) to create an insertion site for the 20 nucleotide guide RNA sequences. The guides were designed using the Broad Institute sgRNA Designer for CRISPR knockout. The sense and antisense guide oligos were annealed, phosphorylated (T4 PNK, NEB), diluted, and ligated into the digested perturb-seq vector (T4 DNA ligase, NEB).

The lentiCas9-Blast plasmid (Addgene #52962) was used to achieve Cas9 expression.

Lentiviral Generation

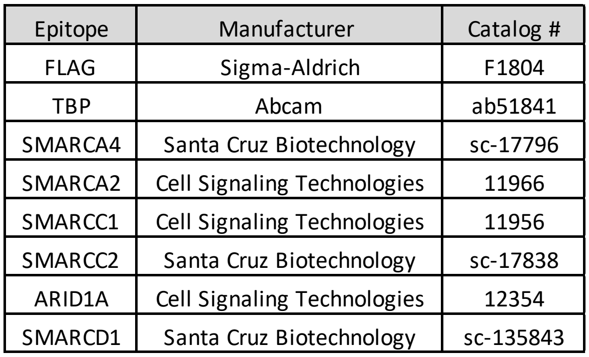
Lentivirus was produced using the second generation virus packaging constructs psPAX2 and pMD2.G. PEI was used to transfect HEK293T LentiX cells with the packaging constructs and the expression constructs (Cas9 or barcoded perturb-seq vector). Perturb-seq guide vector viruses were produced in an arrayed fashion to avoid guide and barcode swapping between vectors during the packaging process. Viral supernatant was harvested 48 hours after transfection and was concentrated by ultracentrifugation at 20,000 rpm for 2.5 hours at 4 degrees C. Viral pellets were resuspended in PBS and frozen at -80°C.

Lentiviral Infection

MOLM-13 cells were spinfected with virus for 1.5 hours at 2000 RPM at 25 degrees Celsius using the concentrated lentivirus and 5 mg/mL polybrene. Cas9 spinfection was performed first with blasticidin selection beginning 2 days after spinfection and western blot confirmation of expression at one week post-infection. Subsequent spinfection with the Perturb-seq guides was performed. HEK293T cells were infected by adding concentrated virus dropwise to the media, adding 10 ug/uL polybrene, and allowing cells to incubate for 48 hours before removing the viral media and replacing with fresh DMEM.

Immunoblotting

Nuclear extraction was performed by resuspending cells in hypotonic solution (no salt) to rupture the cell membrane, and nuclei were spun down to separate from the cytosolic content. Nuclei were lysed in 300mM NaCl and chromatin was spun out of the nuclear extract. The supernatant was combined with 1:4 LDS and 10mM DTT and heated at 95 degrees C for 5 minutes before loading into a 4-12% Bis-Tris protein gel (NuPage). Proteins were transferred to PVDF membrane overnight via wet transfer. 5% milk was used to block membranes for 1 hour, then primary antibody was added and incubated on a shaker for 4 hours at room temperature. 3 x 5 minute washes were performed in PBST, then secondary antibody (Li-Cor Imaging Products) for one hour before 3 more washes. Blots were imaged on a LiCor Odyssey imager.



Epitope	Manufacturer	Catalog #
FLAG	Sigma-Aldrich	F1804
TBP	Abcam	ab51841
SMARCA4	Santa Cruz Biotechnology	sc-17796
SMARCA2	Cell Signaling Technologies	11966
SMARCC1	Cell Signaling Technologies	11956
SMARCC2	Santa Cruz Biotechnology	sc-17838
ARID1A	Cell Signaling Technologies	12354
SMARCD1	Santa Cruz Biotechnology	sc-135843

Fluorescence-Activated Cell Sorting

MOLM-13 or HEK293T cells were resuspended in FACS buffer containing 10mM EDTA and 2% fetal bovine serum in PBS. Cells were sorted for alive populations and for BFP expression using either the Violet laser of the BD FACSAria II or BD FACSAria II UV. Cells were sorted into media containing 20% fetal bovine serum and allowed to recover from sorting for 2 days before media was replaced with fresh media containing 10% fetal bovine serum.

10x Single Cell Gene Expression Library Generation and high-throughput sequencing

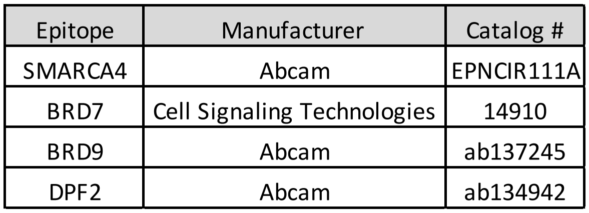
MOLM-13 cells were prepared by diluting in RPMI with 10% FBS to a concentration of approximately 500 cells/uL, and 6,000 cells were loaded onto each channel of the 10x chip (10x Genomics). For the pilot experiment, Chromium Single Cell 3′ Gene Expression version 2 chemistry was used, while for the full-scale perturb-seq and combination perturbations version 3 chemistry was used (10x Genomics). 2 10x channels were run for the pilot experiment, 15 channels were run for the full-scale single perturbation experiment, and 3 channels were run for the combination perturbation experiment. After generation of cDNA libraries from the harvested mRNA, libraries were amplified on a standard thermocycler. Perturb-seq libraries were sequenced on an Illumina Hi-Seq at a ratio of 3 channels of 10x to 1 lane of sequencing (1 for the pilot, 5 for the single perturbation experiment, and 1 for the combination perturbation experiment), using paired end reads as follows. For the pilot experiment, read 1 is 26 bases long (16 for cell barcode, 10 for UMI), read 2 is 98 bases long and sample index read is 8 bases long. For the single guide experiment, read 1 is 28 bases long (16 for cell barcode, 12 for UMI), read 2 is 91 bases long and sample index read is 8 bases long.

Dial-Out PCR and sequencing for assigning perturbations to cells

An aliquot of the cDNA library from each 10x channel was used for Dial-Out PCR using primers designed to amplify the segment between the sgRNA constant region and the BFP expression cassette in the Perturb-seq vector in order to obtain amplified GBC-CBC-UMI (guide barcode-cell barcode-UMI) combinations. The sequences of the primers used are given in table 2.T2. Dial-out PCR products were sequenced on an Illumina Mi-seq, paired-end (for the pilot experiment, read 1 is 26 bases and read 2 is 60 bases long, for the single guide experiment, read 1 is 28 bases and read 2 is 60 bases long), with read 1 having the same structure as in the high-throughput sequencing of the libraries. Each experiment required only one Mi-Seq run to process the dial-out libraries.

Chromatin Immunoprecipitation and sequencing (ChIP-seq)

MOLM-13 cells were harvested and crosslinked with 1% formaldehyde for 10 minutes at 37 degrees Celsius. The reaction was quenched with 125 nM glycine and was incubated 5 more minutes at 37 degrees C. Fixed cells were sheared using a Covaris E220. 10 million cells were used per IP condition, and were incubated overnight with antibody (see table). Antibody and bound protein/DNA were retrieved using Protein G Dynabeads (Thermo Scientific) via incubation for 3 hours at 4 degrees C. Beads were washed, and protein/DNA was eluted. Samples underwent reverse crosslinking and were treated with RNAse A (Roche) and Proteinase K (Thermo Scientific). AMPure beads were used to recover DNA fragments.



Epitope	Manufacturer	Catalog #
SMARCA4	Abcam	EPNCIR111A
BRD7	Cell Signaling Technologies	14910
BRD9	Abcam	ab137245
DPF2	Abcam	ab134942

Perturb-seq data processing: 10x gene expression and dialout PCR

We used cellranger software (version 2.1.1 for the pilot experiment, and 3.02 for the single guide experiment) to align reads to the GRCh38 human transcriptome (transcriptome version GRCh38-1.2.0), obtaining a matrix of counts for each gene in each cell. We used the feature barcoding option from cellranger to process the dialout PCR data together with the expression data, resulting in assignments of which guides were present in which cell. We then used the scanpy package to perform quality-control filtering, normalization and scaling for downstream analyses. We filtered out cells with <1000 genes or if their percent mitochondrial reads were above 20% of the total reads. We normalized the counts per cell such that the counts in each cell sum to 10000. We then transformed these normalized counts to log(normalized count +1), resulting in what we refer to as the raw expression values. We selected a subset of variable genes using standardized dispersion (keeping genes with standardized dispersion greater than 0.5). To account for sources of technical variation, we regressed out batch effect, total counts and percent mitochondrial reads. Finally, we converted the resulting values to z-scores, such that each gene had a mean of 0 and variance of 1 across the cells. We performed PCA for dimensionality reduction, keeping the first 50 principal components. Finally, we represented the cells in a low dimension using UMAP.

Non-negative Matrix Factorization and gene ontology analysis of perturbation pathways

Non-negative Matrix Factorization was used to define differential gene expression programs by transforming all negative relative expression values to zero. Gene expression profiles for all cells were decomposed into two matrices; one describing the activation of gene programs and the other defining the contribution of each gene to each gene program. Number of NMFs was optimized by defining the number of factors that best reconstructed the single cell transcriptomic profiles in performance for a test set of cells compared to a set of cells held back from the analysis. Gene signatures for each gene program were determined by the contributions of genes to each gene program and running gene ontology analysis for each program with respect to the ranking of genes for their respective gene programs.

Overlap enrichment comparison with ENCODE data

To study the relationships between the perturbed BAF subunits and the chromatin binding landscape in the control cells, we asked whether the genes affected by the perturbation of a given subunit fall preferentially near specific transcription factor binding sites of histone marks. Specifically, for each gene program, we retrieved the top genes associated with it (see above), and then defined a window around these genes to look for enrichment. We quantified this enrichment of binding in two separate analyses: a promoter-centric one and a regulatory region-based one.

For the promoter-centric approach, for a given set of genes we identified their respective promoters, defined as 10 kb upstream of the gene transcription start site (TSS), as retrieved from GENCODE[37]. We then used LOLA[28] to compute the enrichment of various ENCODE datasets in the promoters of interest, and defined significant enrichment using an FDR of 0.05. We performed this analysis across all cell types present in the dataset, but found as expected that most enriched binding was specific to leukemia or related blood cell types.

For the regulatory region-based approach, we linked each gene of interest to putative regulatory elements within 100 kb of the gene's TSS. We defined regulatory elements as peaks called from ATAC-seq data in the same cell line. Our universe was all regulatory elements present in the cell line. Starting from a set of genes of interest, we obtained the union of regulatory elements linked to those genes, and used LOLA as before to quantify the enrichment of these regulatory elements of interest in ENCODE binding datasets.

SHARE-seq Experimental

Perturbations were performed by infecting Cas9-expressing MOLM-13 cells with individual guides in replicate. Populations were selected with puromycin to enrich for guide populations 48 hours after transduction. At the 7 day time point, cells were crosslinked and subjected to SHARE-seq experimental protocol as described by Ma et al. *Cell* 2020[27]. Libraries were pooled 3:2 ATAC:RNA before sequencing using a 150 cycle high output Illumina NextSeq sequencing kit.

SHARE-seq data analysis

ATAC Processing

Reads were trimmed and aligned to hg19 using bowtie2[38]. Data was demultiplexed according to their 4 sets of barcodes, tolerating one base mismatch in each of the barcodes. Reads that were discarded include: quality <Q30, mitochondrial reads, improperly paired reads, reads mapped to unmapped contigs, or to the Y chromosome. Duplicates were discarded. MACS2 was used to call peaks[39]. Peaks were merged and filtered out if they overlapped ENCODE blacklisted regions. HOMER (Heinz et all 2010) was used annotate peaks with their genes. ChromVAR was used to calculate fragment counts in peaks and TF scores (Schep et al 2017).

RNA Processing

Base calls were converted into fastq files using bcl2fastq and reads were trimmed. Reads without proper poly-T at the beginning of Read 2 were filtered out, and the remailing reads were aligned to hg19. Demultiplmexing allowed for one mismatch in each of the barcodes. FeatureCounts was used to annotate exonic and intronic reads[40]. UMI tools was used to create matrices of gene counts for each cell[41]. Cells were filtered out if they had <300 genes or more than 1% mitochondrial reads. Seurat V3 was used to to scale gene expression matrices by total UMIs[42].

Combinatorial perturbation linear model derivation

For each pair of perturbations, a linear regression model was trained using the expression of each gene (z-score) across all cells. The model was defined as: Expression of gene ~ perturbation 1 + perturbation 2 + (perturbation 1)*(perturbation 2) + intercept. The first two terms capture any additional variation in gene expression not explained by additive effects, as described by Dixit et al. *Cell* 2016[15]. We trained one model across all groups of perturbations, and assessed the significance of coefficients by permuting the identities of cells to determine p-values corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure. Only genes with a significant coefficient in at least one of the conditions are shown in this chapter's heat maps. Genes are categorized as additive if they do not have a significant interaction term, or they are categorized as synergistic if they do.

BAF activity model via logistic regression classifier

To determine the extent to which each cell has perturbed activity of cBAF, ncBAF, and PBAF subcomplexes, we used a classification framework to score the magnitude of perturbation of each subcomplex. We trained a multi-class logistic regression classifier to distinguish between 4 classes of cells including control cells plus the 3 subcomplex types marked by the following perturbations: ARID1A+ARID1B for cBAF, ARID2 for PBAF, and BRD9 for ncBAF. The model was evaluated using auPRC and found the best performance for the cBAF and ncBAF classes compared to PBAF, which had less-defined effects. We used the model to predict the class in which each cell belonged, and used the resulting outputs as the measure of the perturbation of specific subcomplex activity. In the figure given, we show the average subcomplex scores for each condition, subtracted from the average scores of the control cells. This model accounts only for a subcomplex being perturbed in one direction, and other frameworks are needed to capture multi-directional effects.

TCGA mining and comparison to Perturb-seq signatures

For each tumor, gene expression signatures were normalized to gene expression in normal tissue samples from matched tissue types. BAF knockout signatures were also normalized against control cell population to standardize single cell RNA-seq profiles for comparison across bulk RNA-seq data in the tumor datasets. Gene programs defined by differentially activated NMFs were used to identify the top 10 genes that correspond to each program, and these genes were used for calculation of a gene program activity score for each tumor. Pseudo-bulk RNA-seq samples were created by averaging expression profiles for each knockout condition, and the same approach was applied for calculating gene program activity scores. Softmax transformation was applied to gene activity scores for each knockout condition or tumor (transformed scores sum to 1 for each condition or tumor) to normalize for robust comparisons. BAF knockout condition gene program activity scores were then compared to tumor gene program activity scores using cosine similarity.

VI. References

1. Michel, B. C. *et al.* A non-canonical SWI/SNF complex is a synthetic lethal target in cancers driven by BAF complex perturbation. *Nat Cell Biol* **20**, 1410–1420 (2018).

2. Kadoch, C., Copeland, R. A. & Keilhack, H. PRC2 and SWI/SNF Chromatin Remodeling Complexes in Health and Disease. *Biochemistry* **55**, 1600–1614 (2016).

3. Lessard, J. *et al.* An Essential Switch in Subunit Composition of a Chromatin Remodeling Complex during Neural Development. *Neuron* **55**, 201–215 (2007).

4. Lickert, H. *et al.* Baf60c is essential for function of BAF chromatin remodelling complexes in heart development. *Nature* **432**, 107–112 (2004).

5. López, A. J. & Wood, M. A. Role of nucleosome remodeling in neurodevelopmental and intellectual disability disorders. *Front. Behav. Neurosci.* **9**, (2015).

6. Jones, S. *et al.* Frequent Mutations of Chromatin Remodeling Gene ARID1A in Ovarian Clear Cell Carcinoma. *Science* **330**, 228–231 (2010).

7. Mathur, R. *et al.* ARID1A loss impairs enhancer-mediated gene regulation and drives colon cancer in mice. *Nat. Genet.* **49**, 296–302 (2017).

8. Sun, X. *et al.* Arid1a Has Context-Dependent Oncogenic and Tumor Suppressor Functions in Liver Cancer. *Cancer Cell* **32**, 574-589.e6 (2017).

9. Nakayama, R. T. *et al.* SMARCB1 is required for widespread BAF complex–mediated activation of enhancers and bivalent promoters. *Nature Genetics* **49**, 1613–1623 (2017).

10. Filatova, A. *et al.* Mutations in SMARCB1 and in other Coffin-Siris syndrome genes lead to various brain midline defects. *Nat Commun* **10**, 2966 (2019).

11. McBride, M. J. *et al.* The SS18-SSX Fusion Oncoprotein Hijacks BAF Complex Targeting and Function to Drive Synovial Sarcoma. *Cancer Cell* **33**, 1128-1141.e7 (2018).

12. Tauziede-Espariat, A. *et al.* Loss of SMARCE1 expression is a specific diagnostic marker of clear cell meningioma: a comprehensive immunophenotypical and molecular analysis. *Brain Pathol.* **28**, 466–474 (2018).

13. Schick, S. *et al.* Systematic characterization of BAF mutations provides insights into intracomplex synthetic lethalities in human cancers. *Nat Genet* **51**, 1399–1410 (2019).

14. Rubin, A. J. *et al.* Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks. *Cell* **176**, 361-376.e17 (2019).

15. Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853-1866.e17 (2016).

16. Adamson, B., Norman, T. M., Jost, M. & Weissman, J. S. Approaches to maximize sgRNA-barcode coupling in Perturb-seq screens. *bioRxiv* 298349 (2018) doi:10.1101/298349.

17. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564-576.e16 (2017).

18. Matsuo, Y. *et al.* Two acute monocytic leukemia (AML-M5a) cell lines (MOLM-13 and MOLM-14) with interclonal phenotypic heterogeneity showing MLL-AF9 fusion resulting from an occult chromosome insertion, ins(11;9)(q23;p22p23). *Leukemia* **11**, 1469–1477 (1997).

19. Loo, C.-S. *et al.* A Genome-wide CRISPR Screen Reveals a Role for the Non-canonical Nucleosome-Remodeling BAF Complex in Foxp3 Expression and Regulatory T Cell Function. *Immunity* **53**, 143-157.e8 (2020).

20. Aizawa, H. *et al.* Dendrite development regulated by CREST, a calcium-regulated transcriptional activator. *Science* **303**, 197–202 (2004).

21. Conaway, R. C. & Conaway, J. W. The INO80 chromatin remodeling complex in transcription, replication and repair. *Trends in Biochemical Sciences* **34**, 71–77 (2009).

22. Lu, W. *et al.* Actl6a protects embryonic stem cells from differentiating into primitive endoderm. *Stem Cells* **33**, 1782–1793 (2015).

23. Versteege, I., Medjkane, S., Rouillard, D. & Delattre, O. A key role of the hSNF5/INI1 tumour suppressor in the control of the G1-S transition of the cell cycle. *Oncogene* **21**, 6403–6412 (2002).

24. Zhang, Z.-K. *et al.* Cell cycle arrest and repression of cyclin D1 transcription by INI1/hSNF5. *Mol Cell Biol* **22**, 5975–5988 (2002).

25. Nagl, N. G. *et al.* The p270 (ARID1A/SMARCF1) Subunit of Mammalian SWI/SNF-Related Complexes Is Essential for Normal Cell Cycle Arrest. *Cancer Res* **65**, 9236–9244 (2005).

26. Flores-Alcantar, A., Gonzalez-Sandoval, A., Escalante-Alcalde, D. & Lomelí, H. Dynamics of expression of ARID1A and ARID1B subunits in mouse embryos and in cells during the cell cycle. *Cell Tissue Res* **345**, 137–148 (2011).

27. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103-1116.e20 (2020).

28. Sheffield, N. C. & Bock, C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* **32**, 587–589 (2016).

29. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

30. Vierbuchen, T. *et al.* AP-1 Transcription Factors and the BAF Complex Mediate Signal-Dependent Enhancer Selection. *Molecular Cell* **68**, 1067-1082.e12 (2017).

31. Ghirlando, R. & Felsenfeld, G. CTCF: making the right connections. *Genes Dev* **30**, 881–891 (2016).

32. Mashtalir, N. *et al.* Modular Organization and Assembly of SWI/SNF Family Chromatin Remodeling Complexes. *Cell* **175**, 1272-1288.e20 (2018).

33. Hutter, C. & Zenklusen, J. C. The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. *Cell* **173**, 283–285 (2018).

34. Qin, Q. *et al.* Lisa: inferring transcriptional regulators through integrative modeling of public chromatin accessibility and ChIP-seq data. *Genome Biology* **21**, 32 (2020).

35. Sandoval, G. J. *et al.* Binding of TMPRSS2-ERG to BAF Chromatin Remodeling Complexes Mediates Prostate Oncogenesis. *Molecular Cell* **71**, 554-566.e7 (2018).

36. Boulay, G. *et al.* Cancer-Specific Retargeting of BAF Complexes by a Prion-like Domain. *Cell* **171**, 163-178.e19 (2017).

37. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**, D766–D773 (2019).

38. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).

39. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).

40. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

41. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**, 491–499 (2017).

42. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* **20**, 296 (2019).

Chapter 3: Structural and functional dissection of the mSWI/SNF complex subunits ARID1A and ARID1B

Contributions

Jordan Otto Jagielski: Conceived of project; performed CRISPR-Cas9 domain scanning editing efficiency tests and pilot experiments, KO cell line generation, IP-MS experiments, cloning and reintroduction of mutants, and ChIP-seq experiments.
Yiliang Wei, PhD: Performed pooled CRISPR-Cas9 screening at CSHL
Osama El Demerdash: Designed and cloned ARID1B domain-targeting guide library at CSHL
Drew D'Avino: Identified differential ARID1A/B crosslinking from CXMS studies
Clayton Collings: Performed computational analysis for ChIP-seq studies
Chris Vakoc, PhD: Advised on domain scanning study
Cigall Kadoch, PhD: Conceived of project, supervised study

Acknowledgements

This work was supported by the National Science Foundation Graduate Research Fellowship (2016230210). We thank the Vakoc Lab at Cold Spring Harbor for advising on CRISPR-Cas9 domain scanning experiment and for sharing experimental resources. We are grateful to the Taplin Mass Spectrometry Facility for processing and analyzing mass spectrometry samples. We are also grateful to the Knoechel lab for assistance in processing samples for NGS.

Disclosure

CK is the Scientific Founder, fiduciary Board of Directors member, SAB member, shareholder, and consultant for Foghorn Therapeutics, Inc. (Cambridge, MA).

I. Abstract

ARID1A and ARID1B are the largest constituent members of mSWI/SNF complexes. As mutually exclusive paralogs of the cBAF complex, these subunits have remained enigmatic in their functional similarities and differences. Mutational patterns in human disease suggest that there are distinct differences in function, as ARID1A is often mutated in a variety of human cancers while ARID1B is often mutated in neurodevelopmental disorders. To comprehensively dissect the roles of these paralogous subunits in mSWI/SNF functions, we assessed the functional domains, influence on mSWI/SNF complex composition, and gene targeting roles of ARID1A and ARID1B. In this work, we identify a novel functional domain of ARID1B and describe differential complex assemblies as well as gene targeting functions dependent on ARID1A or ARID1B incorporation.

II. Introduction to ARID1 proteins in mSWI/SNF biology

ARID1A and ARID1B are paralogous mSWI/SNF subunits that incorporate into cBAF complexes in a mutually exclusive manner. They are the largest BAF complex subunits at 250 kD each, and their protein sequences are very similar to one another, especially within their functional domains. Both proteins have a highly conserved AT-rich interaction domain (ARID) which binds DNA, though its DNA binding activity is not specific to AT-rich sequences as its name would suggest[1,2]. The C-terminal end of the proteins consist of two conserved domains recently described as the Core Binding Region A (CBRA) and the Core Binding Region B (CBRB) which together allow ARID1 proteins to bind to BAF complexes (**Figure 3.1**)[3].

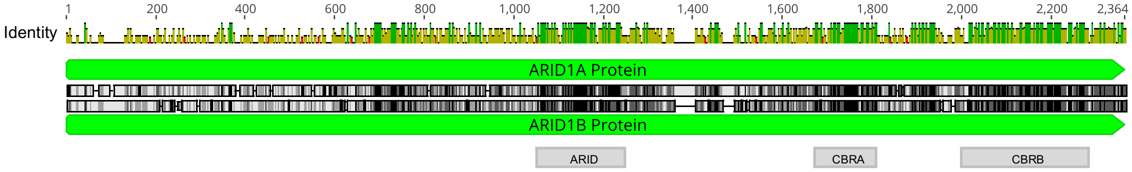


Figure 3.1 ARID1A and ARID1B protein alignment
Alignment of the two human paralogs ARID1A and ARID1B. Identity score is shown above the alignment for each amino acid. Between each annotated protein, identical amino acids are shown in black while similar amino acids are shown in grey. Domains for both ARID1A and ARID1B are annotated below. Alignment and image generated using Genious version 10.2 created by Biomatters.

Though ARID1A and ARID1B are highly similar in sequence and both contain ARID, CBRA, and CBRB domains, there is quite a bit of evidence to suggest they have some degree of difference in function. One observation leading to this conclusion is the differential mutation patterns in disease. ARID1A is notably the most frequently mutated BAF subunit in human cancers; disease-associated mutations in ARID1A are often early truncations or frameshifts that render the whole protein non-expressed or largely non-functional, rather than point mutations[4]. ARID1A mutations occur in 55% of ovarian clear cell carcinoma and 35% of endometrioid carcinomas, and also occur in a host of other cancers including hepatocellular carcinoma and colorectal carcinoma[5,6,7,8]. ARID1B mutations are rarely seen in cancers; however, they are frequently implicated in neurodevelopmental disorders such as Coffin-Siris syndrome, intellectual disability disorders, and autism spectrum disorders[9,10,11,12]. One additional piece of evidence for difference in function is that these paralogs have been differentially implicated in cell cycle control; ARID1A has been shown to be critical in governing the cell cycle stage transition from G2 to M phase, while ARID1B is not necessary for this transition[13,14]. While there may be differences in function between ARID1A and ARID1B, it is critical for a cell to have at least one of these paralogs. Through data from Project Achilles, a large scale RNAi screen to discover weaknesses in genomically-defined cancer cell lines, it was discovered that ARID1B is a synthetic lethality in ARID1A-loss settings[15]. This finding was validated and expanded upon in subsequent experimental work[16,17]. Since ARID1A loss is a fairly common event in a variety of human cancers, ARID1B has become an attractive drug target for these specific indications[18]. It is interesting to note that in a few rare cases of dedifferentiated endometroid carcinoma, loss of both ARID1A and ARID1B has been observed[19]. While the mechanism of the loss of both of these proteins has not yet been elucidated, it is possible that the synthetic lethal relationship may not hold in all tissue contexts or disease settings.

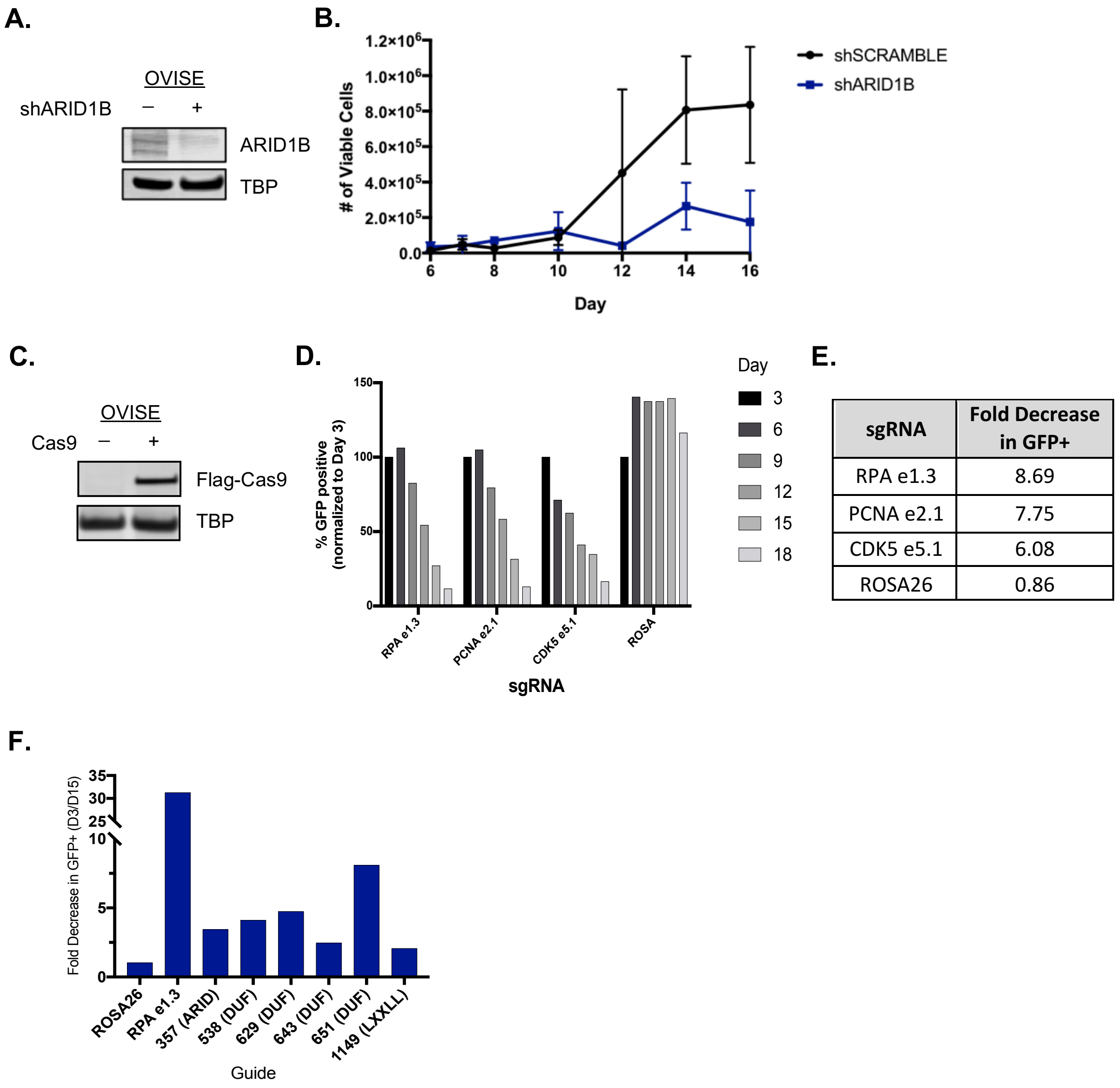
It is important to note that these paralogs tend to be expressed at different levels in different tissue contexts. In general, ARID1A is expressed at a higher level in almost all tissue contexts, according to expression data from the Genotype Tissue Expression Portal (GTEx)[20]. While their difference in expression might partially explain their magnitudes of effects in different tissue contexts, their expression likely cannot explain differences in function, as both paralogs are expressed relatively comparably in all tissues assayed in the GTEx portal, and this doesn't explain the differential patterns of mutation in diseases. Another potential explanation for their difference in function is that there may be some difference between the proteins in the regions of low similarity. One such region is the N-terminus of the proteins. The N-terminus of both ARID1A and ARID1B is very GC-rich and largely predicted to be disordered in structure. The high GC content and large size of these proteins has made them difficult to study via traditional methods reliant on cloning and reintroduction. However, there are a few key questions we sought to answer about the functional roles of these proteins as well as their functional redundancies or differences. The areas we sought to explore include: 1) the identification and characterization of novel functional regions of the proteins ARID1A and ARID1B, 2) the compositional biases of ARID1A- or ARID1B-containing BAF complexes, and 3) the differences in genomic targeting and activities of ARID1A and ARID1B. To this end, we performed a series of experiments to identify and elucidate the functional roles of this paralog family in mSWI/SNF activity.

IIIA. Identification of a novel ARID1B protein domain using CRISPR-Cas9 Domain Scanning

To identify novel functional domains of the ARID1 proteins, we employed a method developed by the Vakoc lab at Cold Spring Harbor known as CRISPR-Cas9 domain scanning[21]. For the CRISPR-Cas9 domain screen to be effective, the protein that is scanned must confer a fitness defect to the target cell when lost. Accordingly, we leveraged the synthetic lethal relationship between ARID1A and ARID1B to perform this assay by targeting ARID1B in an ARID1A-deficient ovarian clear cell carcinoma cell line, OVISE. To ensure that ARID1B loss indeed proved deleterious to the fitness of OVISE cells, shRNAs were used to target ARID1B and viability was measured over time after viral introduction of the shRNA. Indeed, upon knockdown of ARID1B, OVISE cells had reduced viability and proliferation compared to OVISE cells receiving a scrambled control shRNA (**Figure 3.2A-B**).

To perform CRISPR-Cas9 domain scanning, guides are designed tiling the whole length of the target protein. The guide plasmid has a GFP reporter for a readout of guide-infected cell populations. The main principle of domain scanning is that cells receiving guides targeting functionally important sites of a protein have decreased fitness compared to cells receiving guides targeted to a non-functional region of the protein. This occurs because the vast majority of mutations induced by the guide cutting, whether frameshift, truncation, or even missense, are deleterious to the function of a critical domain and thus the overall function of the protein, which results in reduced cell fitness. Alternatively, missense mutations in a non-functional region would not impact protein functions, and thus not impact fitness. In a population of cells with various regions targeted, we would expect that guides targeting functionally important regions would drop out of the population at a faster rate than those targeting non-functional regions.

To test the efficacy of Cas9 editing and the readout system, Cas9-expressing OVISE cells were infected in an arrayed format with guides targeted toward essential genes or a control locus with a target ratio of approximately 50% transduction to allow for sufficient quantities of infected and non-infected cells to complete. GFP (which is co-expressed with the guide plasmid) positive cell quantities were assessed at 3 day intervals to observe whether the guides targeting essential genes were outcompeted by the non-transduced cells. Indeed, there was strong depletion in GFP+ cells in the positive control conditions while there was no GFP+ cell depletion in the control guide population (**Figure 3.2C-E**). As a second control experiment, we wanted to ensure we would see dynamic changes in fitness responses to guides targeting different regions of ARID1B.



sgRNA	Fold Decrease in GFP+
RPA e1.3	8.69
PCNA e2.1	7.75
CDK5 e5.1	6.08
ROSA26	0.86

Figure 3.2 Validation of experimental system for ARID1B CRISPR-Cas9 domain scan
A. Immunoblot of ARID1B expression in control and ARID1B knockdown conditions post-infection with shARID1B or shSCRAMBLE constructs. **B.** Viability time course for OVISE cells infected with shSCRAMBLE construct or shARID1B construct. **C.** Immunoblot of Flag-tagged Cas9 expression in OVISE cells. **D.** GFP+ cell proportions over time for each guide condition targeting 3 essential genes and one control locus in OVISE cells. **E.** Fold-reduction calculations for time course given in (D). **F.** Fold-depletion bar graph for negative and positive control guides along with guides targeting different regions of ARID1B in OVISE cells.

Along with the control negative and control positive guides, 6 guides targeting different sites on ARID1B were tested in the same GFP+ cell depletion assay. A range of depletion in guide abundance was observed ranging from near negative control-level to 8-fold in the ARID1B-targeted conditions (**Figure 3.2F**). For the full ARID1B scan, 958 guides were designed targeted along the length of ARID1B (**Table 3.T1**). These guides were pooled and transduced into the Cas9-expressing OVISE cell population. After 5 cell population doublings, guide abundance in the population was determined by sequencing amplified DNA libraries derived from the cell population. Dropout for each guide was determined and plotted along the length of ARID1B (**Figure 3.3**).

Encouragingly, the previously described known functional domains were identified as functionally important in the Cas9 domain scan. The highest dropout was observed in the ARID domain, which is the protein's critical domain for DNA binding. Additionally, the CBRA and CBRB domains (which bind ARID1 proteins to BAF complexes) both experienced high dropout as well, while the small flexible loop between them was unimportant for fitness in this screen, consistent with expectations. The most exciting finding was a novel region of approximately 200 amino acids near the N-terminus that dropped out to approximately the same extent as CBRA and CBRB. Since this region is predicted to be highly disordered, it may be important for interactions with other proteins or factors (such as transcription factors) or it could serve other functions in the cell. The high GC content and large size of ARID1B have so far proven prohibitive to obtaining a full-length clone of this gene. In order to fully study this novel region of ARID1B, alternative tactics to traditional cloning and mutagenesis may need to be employed in future work.

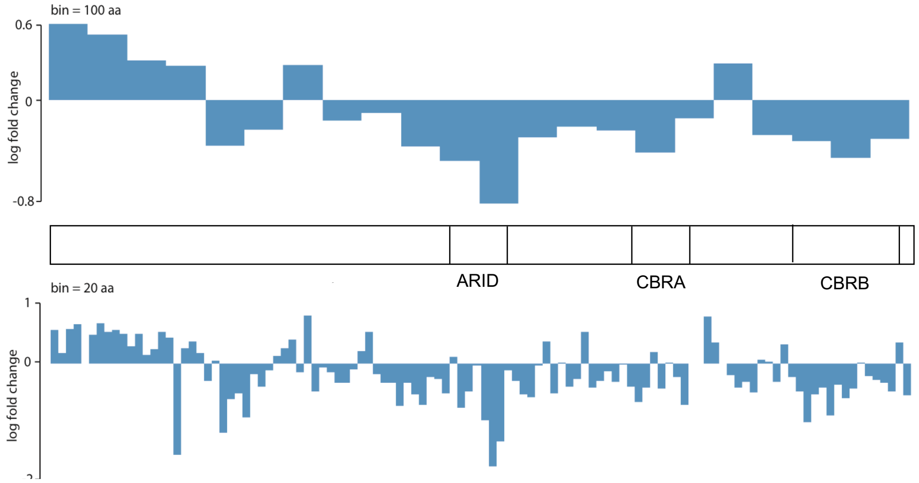


Figure 3.3 ARID1B CRISPR-Cas9 domain scan results
Dropout scores for cells receiving guides targeting the indicated regions tiled along the length of the ARID1B protein sequence. Dropout scores are binned in 100 amino acid regions (top) and 20 amino acid regions (bottom) to show broader and more granular changes in fitness by region. Currently annotated functional regions are depicted.

IIIB. Characterization of compositional preferences of ARID1A- and ARID1B-containing complexes

Insights from cross-linking mass spectrometry analysis of purified BAF complexes suggest bias in the composition of BAF complexes based upon their incorporation of ARID1A or ARID1B, which are mutually exclusive members of the complex[3]. Specifically, there was a skew in the number of crosslinks between the ARID1 and the SMARCD paralogs (which are also mutually exclusive in the complex), with more crosslinks between ARID1A and SMARCD1, and more crosslinks between ARID1B and SMARCD2 (**Figure 3.4**). If there was no preference in binding or composition, we would have expected there to be an equal proportion of SMARCD1/2 binding between complexes containing either ARID1A or ARID1B. Instead, ARID1B had a ratio of SMARCD2:SMARCD1 crosslinks 4-fold higher than that of ARID1A. These increased crosslinks could be explained by other phenomena such as lysine crosslinking efficiency in these regions, so we sought to determine whether we could recapitulate this skew in binding between the ARID1 paralogs and the SMARCD paralogs in non-crosslinked settings.

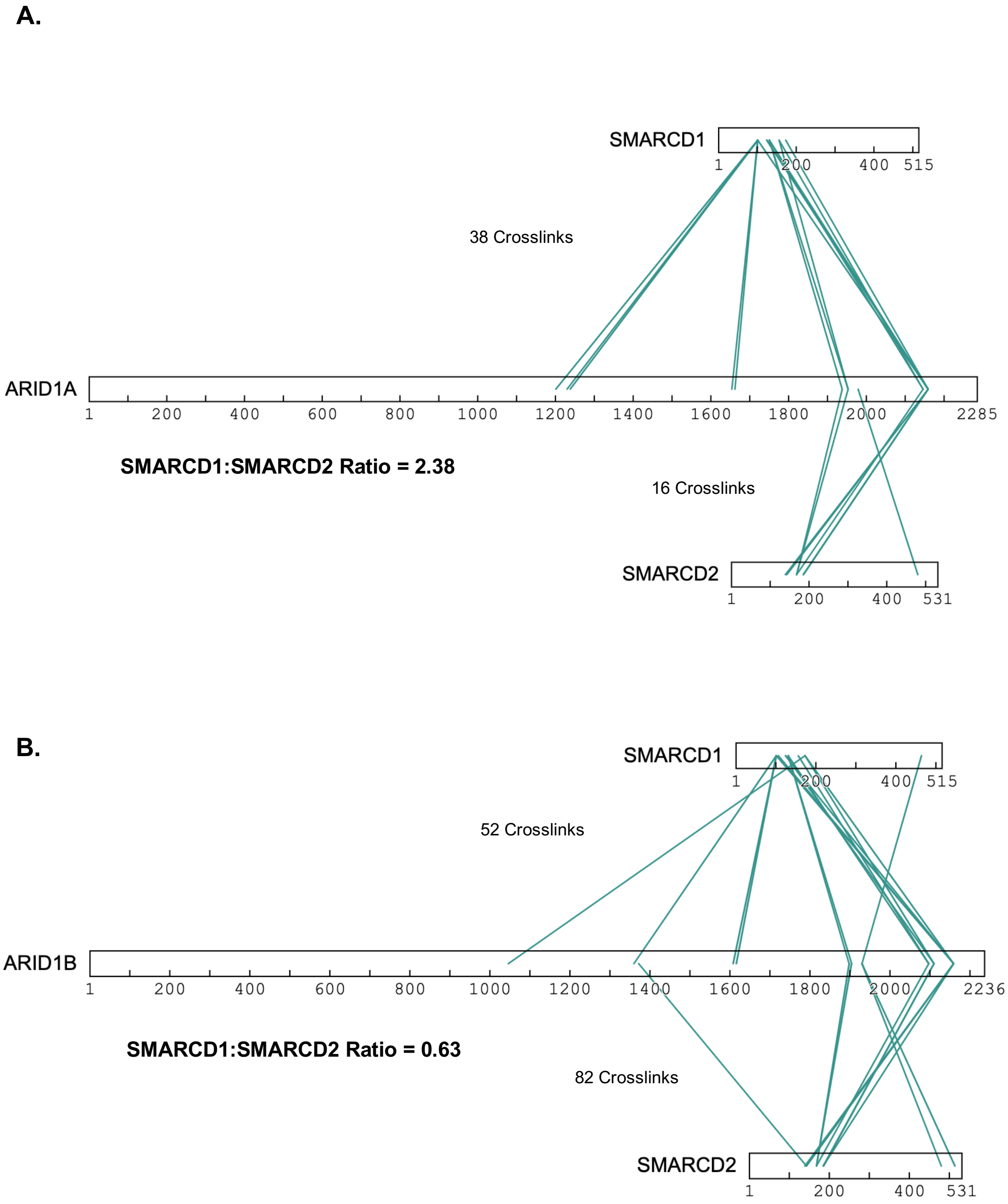


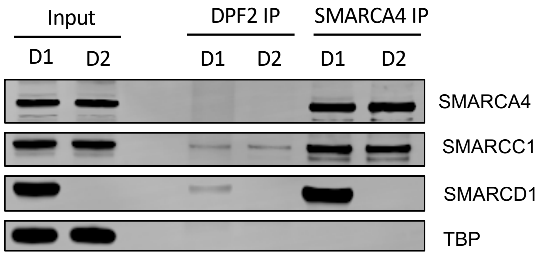
Figure 3.4 Crosslinking mass spectrometry results depicting associations between ARID1 and SMARCD paralogs

A. Crosslinks identified between ARID1A and SMARCD1 or SMARCD2 annotated along protein sequences for each protein. Total number of crosslinks and ratio of SMARCD1:SMARCD2 crosslinks is given. **B.** Same as (A) for ARID1B.

To determine whether this crosslinking mass-spectrometry observation replicates in non-crosslinked BAF complexes, immunoprecipitation for BAF complexes was performed on native proteins in nuclear extract. To eliminate the issue of varying capture efficiencies of different antibodies (for example comparing ARID1A capture to ARID1B capture), cell lines lacking all but one paralog in the SMARCD family were created via CRISPR-Cas9 knockout so the abundance of ARID1A/B could be analyzed with mSWI/SNF complex immunoprecipitation (pulling down the same subunit in both knockout conditions) and subsequent mass spectrometry. Knockout was confirmed on the protein level by western blot (**Figure 3.S1A-B**).

In each knockout context, immunoprecipitation was performed using antibodies against SMARCA4 (pan-BAF complex subunit) and DPF2 (cBAF specific subunit), and the samples were denatured and submitted for mass spectrometry (**Figure 3.5A**). The mass spectrometric analysis identified a similar bias of SMARCD1-containing complexes having more ARID1A integrated while SMARCD2-containing complexes had a higher proportion of ARID1B integration based on unique peptide count (**Figure 3.5B and 3.S1C**). This difference was apparent in the SMARCA4 IP, but was even more magnified in the DPF2 IP, which specifically immunoprecipitates cBAF complexes. It was also encouraging to note that there were no peptides identified for the other SMARCD paralogs in the respective knockout conditions. While it appears there might be some preference for incorporation of ARID1 and SMARCD paralogs, it does not appear to be to be a completely exclusive pairing. There are some shortcomings with this method as well; since these paralogs are so similar to one another, there are few unique peptides. Methods for assessing differential composition have also been hindered by the lack of quality specific antibodies to these paralogs, as again, they are very similar in sequence. Further work will need to be done to fully understand the preferential binding of the SMARCD and ARID1 paralogs.

A.



B.

293T Type	ARID1B:ARID1A Unique Peptide Ratio	
	BRG1 IP	**BAF45D IP**
SMARCD1 Only	0.66	0.42
SMARCD2 Only	0.77	0.81
ARID1B Fold-Increase	**1.16**	**1.93**

Figure 3.5 IP-MS analysis of SMARCD1-containing and SMARCD2-containing BAF complexes
A. Immunoblot of immunoprecipitation products of DPF2 and SMARCA4 in the SMARCD1 and SMARCD2-only cell populations. These IP products were submitted for mass spectrometry analysis. **B.** Unique peptide ratio of ARID1B:ARID1A in SMARCD1- and SMARCD2-only cells. For DPF2 immunoprecipitation, there was an approximate 2-fold increase in ARID1B abundance over ARID1A for the SMARCD2-only containing cells.

IIIC. Gene targeting differences between ARID1A- and ARID1B-containing complexes

To help define differences in functionality of ARID1A and ARID1B, we sought to understand whether these proteins have different gene targeting functions. While previous work in the lab has focused on using different antibodies toward ARID1A and ARID1B, we wanted to expand upon these datasets by performing chromatin immunoprecipitation that does not depend on the affinity and specificity of different antibodies in comparison to one another. In order to accomplish this, clonal populations of ARID1A and ARID1B knockouts in ES2 cells (an ovarian cancer cell line) were generated. Since ARID1A specifically is so important in ovarian contexts, we hypothesized that this would be a good settings in which to discover the differences in gene targeting for ARID1A and ARID1B. The knockout of ARID1A and ARID1B in clonal cell populations was confirmed via protein immunoblot (**Figure 3.6A**).

Chromatin immunoprecipitation was performed in naïve, ARID1A knockout, and ARID1B knockout ES2 cells using antibodies against SMARCA4 and SMARCC1. The peaks identified by SMARCA4 and SMARCC1 chromatin immunoprecipitation were overlaid to have the highest confidence sites for BAF binding on chromatin. There were a number of lost and gained BAF binding sites for both ARID1A and ARID1B knockout conditions (**Figure 3.6B**). Sites were identified where BAF binding was lost in both ARID1A and ARID1B knockout cells and also sites that were only lost in ARID1A knockout conditions or ARID1B knockout conditions (**Figure 3.6C-D**). In order to characterize the sites that are selectively lost upon ARID1A or ARID1B knockout, LOLA was used to define which transcription factor binding profiles are most highly correlated with these lost sites[22]. For ARID1A-lost sites, ESR1, AR, and FOXA1 ChIPs had the highest similarity (**Figure 3.6E**). This pairs nicely with the suspected role of ARID1A as critical in hormone-responsive tissues, such as ovarian and breast[23]. Additionally, the LXXLL motifs present in the ARID1 proteins are known hormone receptor binding motifs (specifically for AR

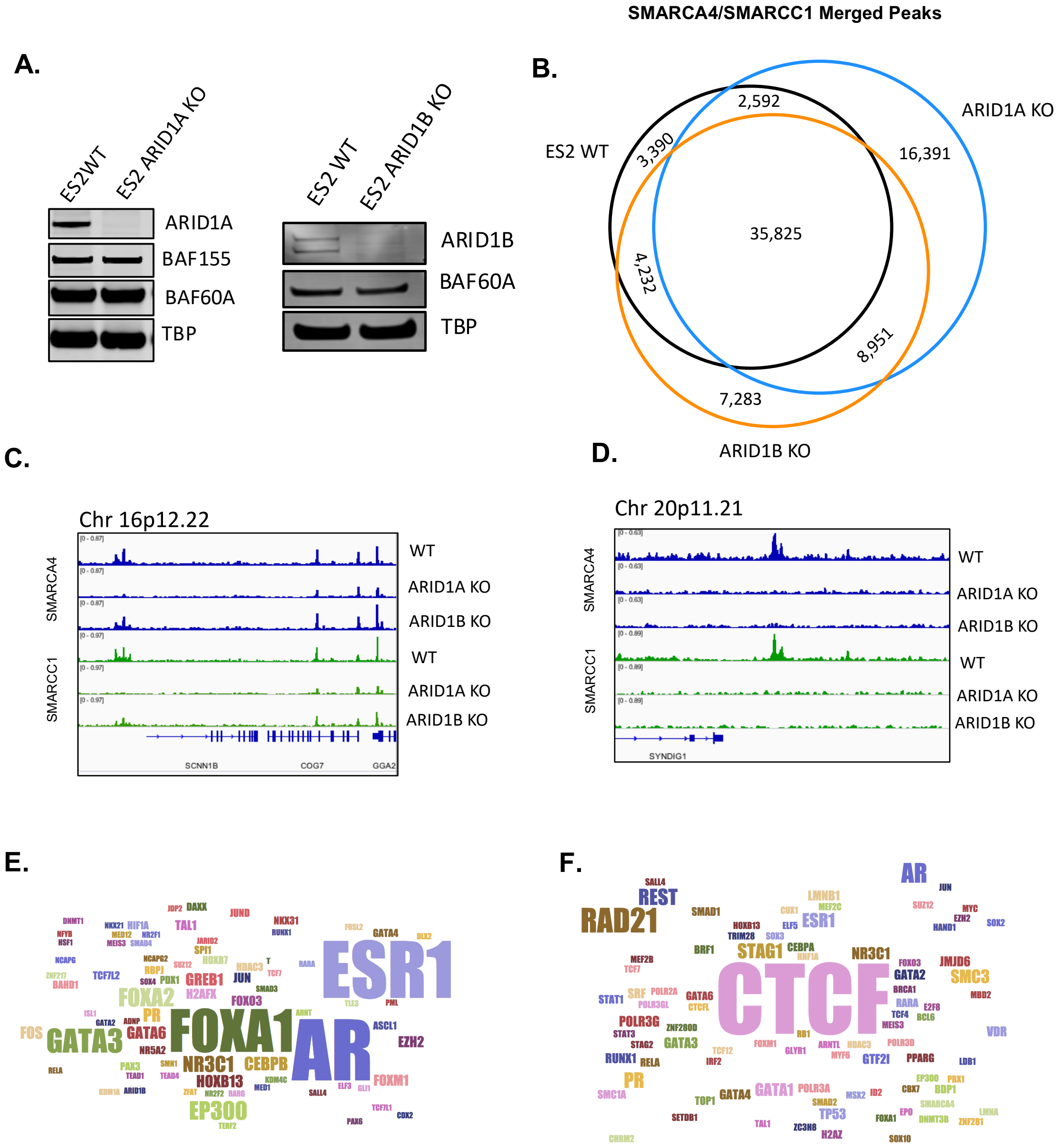


Figure 3.6. ChIP-seq studies identify differential localization patterns for ARID1A- and ARID1B-containing BAF complexes

A. Immunoblots of ARID1A and ARID1B expression in knockout ES2 cell nuclear extract. **B.** Venn diagram of BAF-occupied peaks in ARID1A and ARID1B knockout ES2 cells. **C.** ChIP-seq track showing a lost site of BAF binding in ARID1A knockout condition. **D.** ChIP-seq track showing a lost site of BAF binding in both ARID1A and ARID1B knockout conditions. **E.** LOLA enrichment for factors with binding profiles similar to lost sites in ARID1A knockout ES2 cells. **F.** Same as (E) for ARID1B knockout ES2 cells.

ESR1, and GR), so it makes sense that these kinds of transcription factors may be associated with ARID1A binding[24]. For unique ARID1B lost sites, the highest enrichment included ChIP profiles for CTCF, RAD21, and REST (**Figure 3.6F**). Some nuclear hormone receptors are similarly enriched on this list (such as AR) but are further down the ranking in significance.

The difference in binding sites and associations with transcription factor binding profiles confirms that in this context, there are functional differences in the genomic targeting of ARID1A and ARID1B. Some caveats to these studies include that they were performed in clonal populations, which may not accurately resemble a bulk population finding, depending on how the clones adapt to survive target gene knockout. Further studies need to be performed to characterize the nature of the differences that were defined, and to perhaps identify whether these factors with similar binding profiles interact with or otherwise co-opt BAF functions.

IIID. ARID1A N-terminus functional studies

Since binding profiles and localization on the genome were found to differ between ARID1A- and ARID1B-containing complexes, we sought to identify whether the differences in genome targeting could be attributed to the N-terminus of these proteins, which is the region of least similarity between the two paralogs. Additionally, a few point mutations have been described in the N-terminal region of ARID1A in breast cancer and endometrial cancers[25,26]. Finally, the identified N-terminal region in the ARID1B CRISPR-Cas9 domain scan motivated the study into the N-terminal region of these proteins. To determine whether this region has functional importance to the gene targeting abilities of ARID1A, HA-tagged full-length ARID1A and a truncated version with the first 600 amino acids removed were cloned. These constructs were expressed using lentivirus in OVISE cells, which lack ARID1A (**Figure 3.S2**). Chromatin immunoprecipitation was performed targeting the HA-tag of the constructs as well as SMARCA4 which marks all BAF complexes. Upon analysis of localization across the genome, there were very few differences in localization between the ARID1A full length and N-terminal deletion constructs. Further investigation will be needed to understand whether this region is critical for ARID1A function, or if there are other assays or methods to better characterize potential differences in function.

IV. Discussion

Taken together, the work described in these studies has begun to define differential roles for the mSWI/SNF paralogous subunits ARID1A and ARID1B, specifically in their functional domains, gene targeting functions, and their influence on the composition of mSWI/SNF complexes. For gene targeting functions, there were differences in binding patterns observed between ARID1A- and ARID1B-containing BAF complexes as evidenced by differential peaks and their peak associations with different transcription factor binding profiles. Future work needs to be done to understand which domain(s) confer the specificity between the paralogs in their gene targeting functions. Preliminary work on studies of the N-terminus of ARID1A suggests it may not affect the localization properties of ARID1A, though perhaps it is important for other functions that need to be assessed through different assays.

The relationship between ARID1 paralog integration and SMARCD paralog integration was also investigated, and a preferential skew between the pairs ARID1A and SMARCD1 and ARID1B and SMARCD2 was identified, as was predicted based on prior insights based on crosslinking mass spectrometry. The composition of the paralogs was not mutually exclusive, and further work should be done to validate these findings. It would be especially important to verify these findings with a method that does not rely on mass spectrometry for detection, as both the method used that initially identified this compositional bias and was subsequently used to validate the trend relied on mass spectrometry. Since these paralogs are highly similar to one another, reagents to definitively resolve them to quantify relative amounts has so far been limiting.

Finally, the most exciting aspect of this body of work was the identification of a novel functionally important region in the N-terminus of ARID1B. Further work needs to be done to validate these findings, which so far has been hindered by technical limitations of cloning full-length ARID1B. However, the implications of this functional region could be highly important, as ARID1B is a synthetic lethality in ARID1A-deficient settings, which occurs in many human cancers. Hence, targeting this region (or other functional regions) of ARID1B could be a viable therapeutic strategy in treating ARID1A-deficient cancers.

V. Materials and Methods

Cell Lines and Cell Culture

OVISE cells and ES2 cells (ATCC) were grown in RPMI 1640 medium (no glutamine, Gibco) supplemented with 10% fetal bovine serum (Omega), 1% GlutaMax (Gibco), and 1% penicillin-streptomycin (Gibco). HEK293T LentiX cells (ATCC) were cultured in modified MEF media containing DMEM (no glutamine, high glucose), 10% fetal bovine serum, 1% penicillin-streptomycin, 1% HEPES, 1% sodium pyruvate, 1% GlutaMax, and 1% non-essential amino acids (all Gibco). Cells were washed with PBS and trypsinized using 0.25% trypsin (Gibco) during passaging and collection for flow cytometry, lysate collection, and ChIP-seq experiments.

Constructs and cloning

Constructs for the CRISPR-Cas9 domain scanning include Cas9 Puro (addgene #108100), positive and negative control guides, and the empty guide vector, which were generous gifts of Chris Vakoc at Cold Spring Harbor. Constructs for the ARID1A truncation reintroduction were cloned into a modified vector (see backbone for addgene #31780, phND2-N106), with the insert under the EF1a promoter. Not1 enzyme (NEB) was used to digest the plasmid and the PCR amplified inserts were dropped in using Infusion enzyme (NEB).

Lentiviral production

Lentivirus was created by transfecting HEK293T LentiX cells with the insert, psPAX2, and pMD2.g. PEI (Polysciences) and OptiMem (Gibco) were used as the transfection medium, and the viral supernatant was harvested 48 hours after transfection. Lentivirus was concentrated by spinning at 20,000 RPM for 2.5 hours at 4 degrees Celsius. Virus was resuspended in PBS and frozen at -80 degrees C.

Lentiviral infection

OVISE (ATCC) cells were infected by adding concentrated virus dropwise along with 10ug/mL polybrene. Cells were incubated for 48 hours in viral media before a media change and before adding selection.

Immunoblotting

For whole cell lysis, cells were harvested and lysed by resuspending cell pellets in 10% SDS in PBS. Lysate was cleared by sonication to shear genomic DNA, then loading dye (Invitrogen) was added 1:4 and 1 M DTT was added 1:10 before samples were incubated at 95 degrees C for 5 minutes. For nuclear extraction, the cell membrane was ruptured by incubating cell pellets in hypotonic solution (no salt), after which the nuclei were spun into a pellet. These nuclei were resuspended in 300 mM salt solution to lyse and release the nuclear contents. Chromatin was spun down at high speed, and supernatant (nuclear protein content) was kept. LDS was added 1:4 and 1 M DTT was added 1:10, and samples were incubated at 95 degrees C for 5 minutes. 4-12% bis-tris gel (NuPage) wells were loaded with 20 mg protein each and run for sufficient band separation. Proteins were transferred to PVDF membrane overnight using wet transfer method with high methanol buffer. Membranes were blocked with 5% milk in PBS for 1 hour, then incubated with primary antibody (various, see table) for 4 hours at room temperature. 3 x 5min washes were performed with PBST before a 1 hour incubation with secondary antibodies (Li-Cor Odyssey imaging products). Blots were imaged on a Licor Odyssey imager.

Epitope	Manufacturer	Catalog #
ARID1A	Cell Signaling Technologies	12354
ARID1B	NOVUS Biologicals	57485
TBP	Abcam	ab51841
SMARCA4	Santa Cruz Biotechnology	sc-17796
SMARCC1	Cell Signaling Technologies	11956
SMARCD1	Santa Cruz Biotechnology	sc-135843

Fluorescence-Activated Cell Sorting

OVISE cells were trypsinized, washed with PBS, and resuspended in FACS buffer containing 10mM EDTA and 2% fetal bovine serum in PBS. Cells were sorted for alive populations and GFP expression on a tabletop Accuri flow cytometer.

CRISPR Cas9 knockout clonal population generation

ARID1A and ARID1B clonal knockout populations were created by transfecting ES2 cells (ATCC) with Santa Cruz knockout constructs (product numbers 400469 and 402365) using Lipofectamine 3000 reagents and protocol (Thermo Fisher Scientific). Media was changed at 24 hours and cells were allowed to grow another 24 hours before single cell plating. Single cell plating was performed using dilution, and clones were grown up and tested for protein expression via immunoblot.

Immunoprecipitation and mass spectrometry (IP-MS)

Cells were subjected to the nuclear extract protocol (see immunoblot methods) and nuclear protein content was quantified by BCA. 1 mg of protein was incubated overnight with 2 ug of antibody (Santa Cruz BRG1 G-7, or Abcam DPF2 REQ) and antibody-bound complexes were retrieved using Protein G Dynabeads (Thermo Scientifiec) via incubation for 2 hours. After several washes, protein was eluted from beads using LDS (nuPAGE) and 100 mM DTT, then submitted for mass spectrometry.

Chromatin immunoprecipitation

Cells were harvested following their transduction with indicated constructs and 7 days of 10ug/mL blasticidin selection. Chromatin precipitation was performed using a standard protocol with some modifications. Cells were crosslinked with 1% formaldehyde for 10 minutes at 37 degrees C. 125 mM glycine was added to quench the reaction and was incubated for another 5 minutes. Fixed cells had chromatin fragmented using a Covaris E220 and 10 million cells were used for each IP condition. Lysate was incubated with the appropriate antibody overnight (see table). Protein G Dynabeads (Thermo Scientific) were added and incubated for 3 hours at 4 C to retrieve antibody/protein/chromatin conjugates. Beads were washed and eluted. Samples were reverse crosslinked and treated with RNAse A (Roshe) and Proteinase K (Thermo Scientific). DNA was recovered using AMPure beads (Beckman Coulter).

Epitope	Manufacturer	Catalog #
ARID1A	Cell Signaling Technologies	12354
ARID1B	NOVUS Biologicals	57485
SMARCA4	Abcam	EPNCIR111A
SMARCC1	Cell Signaling Technologies	11956

Library Preparation and Sequencing

Library preparation was performed using NEBNext Ultra II DNA library preparation kit (New England Biolabs) and protocol. Library quality was assessed on an Agilent TapeStation 4200 instrument with D1000 high sensitivity tape and reagents (Agilent). Sequencing was performed on a NextSeq500 instrument with a 75 cycle high output sequencing kit (Illumina) using standard denaturing and dilution protocol.

ChIP-seq data analysis

ChIP-seq reads were mapped using Bowtie2 to the human genome hg19[27]. Peaks were called using MACS2 against input reads, and duplicate reads were discarded[28]. BAF complex sites were determined by standalone and merged peak sites for BRG1 occupancy and BAF155 occupancy. Venn diagrams were made using the R statistical package. LOLA was used to determine overlap with ChIP-seq datasets of transcription factors[22].

VI. References

1. Wang, T., Zhang, J., Zhang, X. & Tu, X. Solution structure of SWI1 AT-rich interaction domain from Saccharomyces cerevisiae and its nonspecific binding to DNA. *Proteins* **80**, 1911–1917 (2012).

2. Kim, S., Zhang, Z., Upchurch, S., Isern, N. & Chen, Y. Structure and DNA-binding sites of the SWI1 AT-rich interaction domain (ARID) suggest determinants for sequence-specific DNA recognition. *J Biol Chem* **279**, 16670–16676 (2004).

3. Mashtalir, N. *et al.* Modular Organization and Assembly of SWI/SNF Family Chromatin Remodeling Complexes. *Cell* **175**, 1272-1288.e20 (2018).

4. Wu, J. N. & Roberts, C. W. M. ARID1A mutations in cancer: another epigenetic tumor suppressor? *Cancer Discov* **3**, 35–43 (2013).

5. Jones, S. *et al.* Frequent Mutations of Chromatin Remodeling Gene ARID1A in Ovarian Clear Cell Carcinoma. *Science* **330**, 228–231 (2010).

6. Wiegand, K. C. *et al.* ARID1A mutations in endometriosis-associated ovarian carcinomas. *N Engl J Med* **363**, 1532–1543 (2010).

7. Sun, X. *et al.* Arid1a Has Context-Dependent Oncogenic and Tumor Suppressor Functions in Liver Cancer. *Cancer Cell* **32**, 574-589.e6 (2017).

8. Mathur, R. *et al.* ARID1A loss impairs enhancer-mediated gene regulation and drives colon cancer in mice. *Nat. Genet.* **49**, 296–302 (2017).

9. Tsurusaki, Y. *et al.* Mutations affecting components of the SWI/SNF complex cause Coffin-Siris syndrome. *Nat Genet* **44**, 376–378 (2012).

10. Filatova, A. *et al.* Mutations in SMARCB1 and in other Coffin-Siris syndrome genes lead to various brain midline defects. *Nat Commun* **10**, 2966 (2019).

11. Sim, J. C. H., White, S. M. & Lockhart, P. J. ARID1B-mediated disorders: Mutations and possible mechanisms. *Intractable Rare Dis Res* **4**, 17–23 (2015).

12. Halgren, C. *et al.* Corpus callosum abnormalities, intellectual disability, speech impairment, and autism in patients with haploinsufficiency of ARID1B. *Clin Genet* **82**, 248–255 (2012).

13. Flores-Alcantar, A., Gonzalez-Sandoval, A., Escalante-Alcalde, D. & Lomelí, H. Dynamics of expression of ARID1A and ARID1B subunits in mouse embryos and in cells during the cell cycle. *Cell Tissue Res* **345**, 137–148 (2011).

14. Nagl, N. G. *et al.* The p270 (ARID1A/SMARCF1) Subunit of Mammalian SWI/SNF-Related Complexes Is Essential for Normal Cell Cycle Arrest. *Cancer Res* **65**, 9236–9244 (2005).

15. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564-576.e16 (2017).

16. Helming, K. C. *et al.* ARID1B is a specific vulnerability in ARID1A-mutant cancers. *Nat Med* **20**, 251–254 (2014).

17. Kelso, T. W. R. *et al.* Chromatin accessibility underlies synthetic lethality of SWI/SNF subunits in ARID1A-mutant cancers. *Elife* **6**, (2017).

18. Sato, E. *et al.* ARID1B as a Potential Therapeutic Target for ARID1A-Mutant Ovarian Clear Cell Carcinoma. *Int J Mol Sci* **19**, (2018).

19. Coatham, M. *et al.* Concurrent ARID1A and ARID1B inactivation in endometrial and ovarian dedifferentiated carcinomas. *Modern Pathology* **29**, 1586–1593 (2016).

20. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580–585 (2013).

21. Shi, J. *et al.* Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. *Nat Biotechnol* **33**, 661–667 (2015).

22. Sheffield, N. C. & Bock, C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* **32**, 587–589 (2016).

23. Xu, G. *et al.* ARID1A determines luminal identity and therapeutic response in estrogen-receptor-positive breast cancer. *Nat Genet* **52**, 198–207 (2020).

24. Savkur, R. S. & Burris, T. P. The coactivator LXXLL nuclear receptor recognition motif. *J Pept Res* **63**, 207–212 (2004).

25. Cornen, S. *et al.* Mutations and deletions of ARID1A in breast tumors. *Oncogene* **31**, 4255–4256 (2012).

26. Gibson, W. J. *et al.* The genomic landscape and evolution of endometrial carcinoma progression and abdominopelvic metastasis. *Nat Genet* **48**, 848–855 (2016).

27. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).

28. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).

Chapter 4: Conclusions and Future Directions

The work described in this thesis spans two distinct types of experimental work: 1. Large scale screening to characterize whole systems and 2. Targeted experimental work towards mechanistic understanding of a specific function or property of a gene or protein. Both of these routes of study are highly important, and they often are intertwined, as the large scale screening discoveries tend to motivate focused mechanistic studies after the identification of interesting phenomena. The work in this thesis for both the Perturb-seq system as well as the focused ARID1A/ARID1B study answers questions about the mSWI/SNF system and also proposes new hypotheses and interesting avenues for further exploration.

I. Perturb-seq Study

The single cell genomics characterization of mSWI/SNF complexes faithfully recapitulated previously proposed behaviors, such as the functional similarity of its constituent subunits as well as assembly-related insights for complex formation[1,2]. However, these earlier defined properties were based on fitness, not the actual gene expression changes underlying these similarities in behavior. We identify the gene programs driving the similarities and differences between subunits and subcomplex types, showing that there while there are shared behaviors between subcomplex forms, they are each marked by their own unique patterning of transcriptomic effects upon perturbation. ncBAF was particularly interesting in this case, because it similarly regulated a subset of the downregulated gene programs as cBAF/core perturbations, however it was highly unique in its upregulated gene programs, notably those regarding actin/filament based processes. cBAF had the strongest character in this cell system as marked by the magnitude of effects on gene expression, largely marked by differentially regulated expression of processes critical in development and differentiation.

Through combinatorial perturbations, we identify additive and synergistic gene sets for paralog families and find different behaviors within each family. While ARID1A and ARID1B were largely additive in nature, the other paralog families all had a significant number of synergistic gene sets upon dual perturbation. Specifically for the GLTSCR1/1L subunits, which have subtle effects on the transcriptome alone, they achieve ncBAF-perturbed phenotype upon dual knockout, suggesting they may be able to structurally or functionally compensate for one another. Further investigation into the interdependence or functional similarities and differences of the paralog families will reveal underlying biology to explain these observations.

Other interesting findings include the differential magnitudes of roles of shared complex subunits. While the cBAF and PBAF subunit SMARCB1 had equal effects in perturbing both these complexes' activities, SMARCE1 perturbed cBAF activity more strongly than PBAF activity. Similarly, SMARCA4 perturbed both cBAF and PBAF to a similar extent while less significantly perturbing the activity of ncBAF. Together with other insights such as the lack of chromatin accessibility changes over ncBAF-bound sites upon SMARCA4 knockout, it may be that SMARCA4 activity is less critical to the activities of ncBAF complexes. Further biochemical characterization of the roles of these shared subunits could help elucidate these mechanisms, perhaps through purification of the various complex types with and without the shared subunit perturbation to assay in vitro behaviors such as nucleosome remodeling efficiency. Further to this point, studies regarding the relationships of inter-complex stoichiometry would also be useful to understand how the perturbation of one complex may affect the abundance of the other subcomplex forms. Experimental results from SHARE-seq suggests that the decrease in cBAF-associated regulation leads to an increase of activity over ncBAF sites (as marked by the CTCF motif accessibility gain and increased accessibility over ncBAF-bound sites in these conditions).

Finally, and potentially most impactfully for human disease, further work to validate the mining data from TCGA-cataloged tumors could yield fruitful insights into related mechanisms of epigenetic perturbations in human malignancies. Encouragingly, the mining efforts in this project identified transcription factors already known to co-opt BAF complexes in cancer-associated settings, including FLI1 and ERG[3,4]. Other factors including CTCF and SPI1 were factors identified in the SHARE-seq analysis as motifs under regions of altered chromatin accessibility upon mSWI/SNF subunit perturbation. The remaining mutation enrichments and transcription factor-associated gene expression represent a set of new testable hypotheses for routes of perturbation that lead to similar signatures as BAF loss, potentially through a convergent mechanism.

Finally, while this study was performed in a singular context, it would be interesting to see how these studies compare to similar profiling in a multitude of different tissue or disease context, which would ideally expand the gene sets known to be under regulation of mSWI/SNF complexes. It would also be interesting to perturb mSWI/SNF subunits over a dynamic process, such as differentiation or immune cell activation, to see how individual mSWI/SNF subunits regulate and govern rapid changes in chromatin states.

II. ARID1A/ARID1B Study

The work towards better understanding the structural and functional properties of ARID1A and ARID1B has yielded a variety of insights that importantly define new avenues of interrogation for these paralogous complex members. In terms of the structural considerations, the identification of a novel, functionally important region of ARID1B has potential implications in 1. Better understanding the functional roles of subunit composition and domains and 2. Therapeutic targeting for ARID1A-deficient cancers, of which there are many. While this region was identified as critical to functionality of this protein, its behaviors and reasons for being functionally important need to be elucidated. Advanced cloning methods may need to be employed, as the length and high GC-content has so far prevented the isolation of full-length ARID1B for performing mutagenesis studies. However, upon the creation of such a clone, these kinds of studies will be imperative to understand how this region functions within the protein.

The exploration into compositional preference or bias in ARID1A- or ARID1B-containing mSWI/SNF complexes identified a skew in incorporation of SMARCD paralogs, with SMARCD1 complexes containing more ARID1A, while SMARCD2 complexes contain more ARID1B. While these patterns were not mutually exclusive, the shift in incorporation could have potential effects on the differential activities of mSWI/SNF complexes that contain one paralog or the other. Future work needs to be done to confirm this compositional preference, ideally using methodology not reliant on mass spectrometry, as the identification of this potential trend was accomplished using CX-MS and the follow up studies were performed using IP-MS. The paralogs within both families are highly similar to one another, so another method that does not rely on unique peptide count from mass spectrometry would be an ideal system for validation.

Finally, assessments of the gene targeting functions of ARID1A and ARID1B proteins suggest differential binding profiles for the mSWI/SNF complexes containing these proteins, and potential different interactions with transcription factors based on comparison to ChIP-seq binding profiles for these factors. Taken together, this set of studies highlight both structural and functional differences between these paralogous subunits and motivates a more intensive study into their roles.

III. References

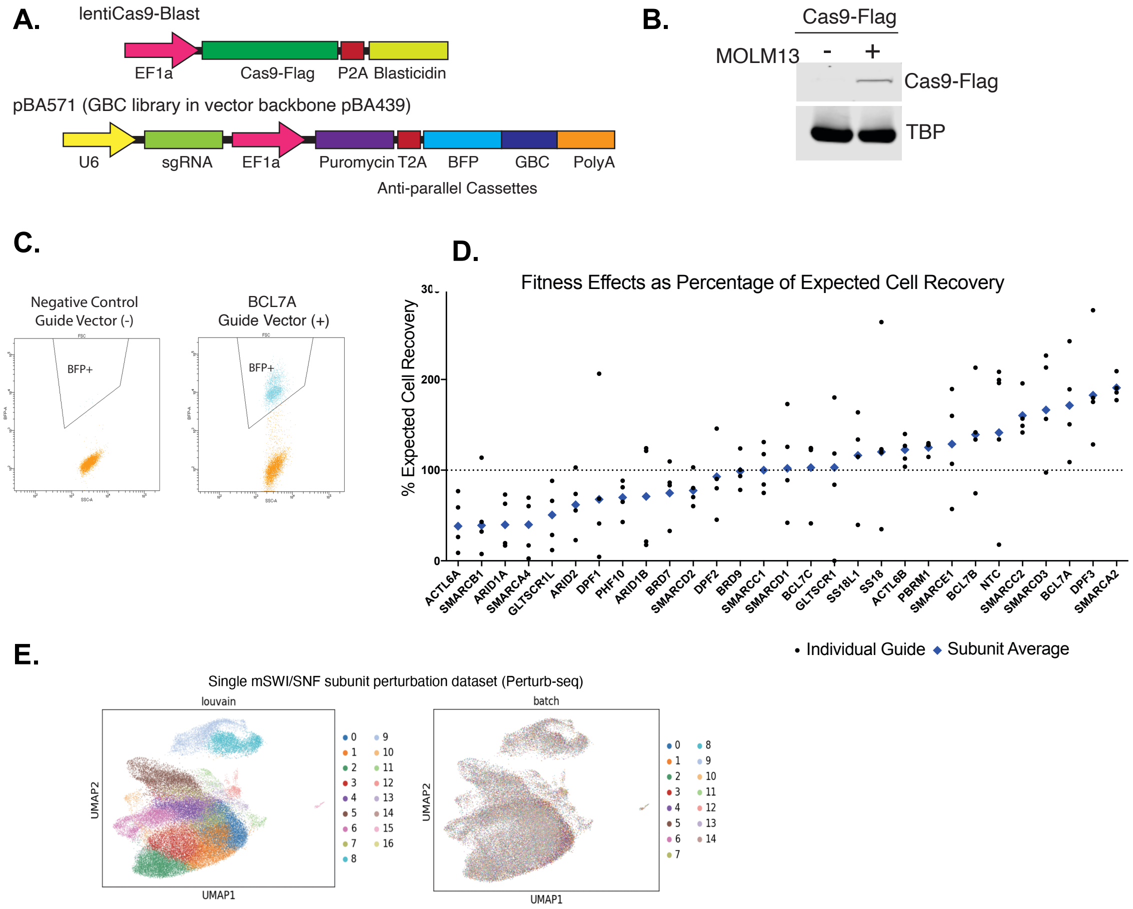
1. Pan, J. *et al.* Interrogation of Mammalian Protein Complex Structure, Function, and Membership Using Genome-Scale Fitness Screens. *Cell Systems* **6**, 555-568.e7 (2018).

2. Mashtalir, N. *et al.* Modular Organization and Assembly of SWI/SNF Family Chromatin Remodeling Complexes. *Cell* **175**, 1272-1288.e20 (2018).

3. Boulay, G. *et al.* Cancer-Specific Retargeting of BAF Complexes by a Prion-like Domain. *Cell* **171**, 163-178.e19 (2017).

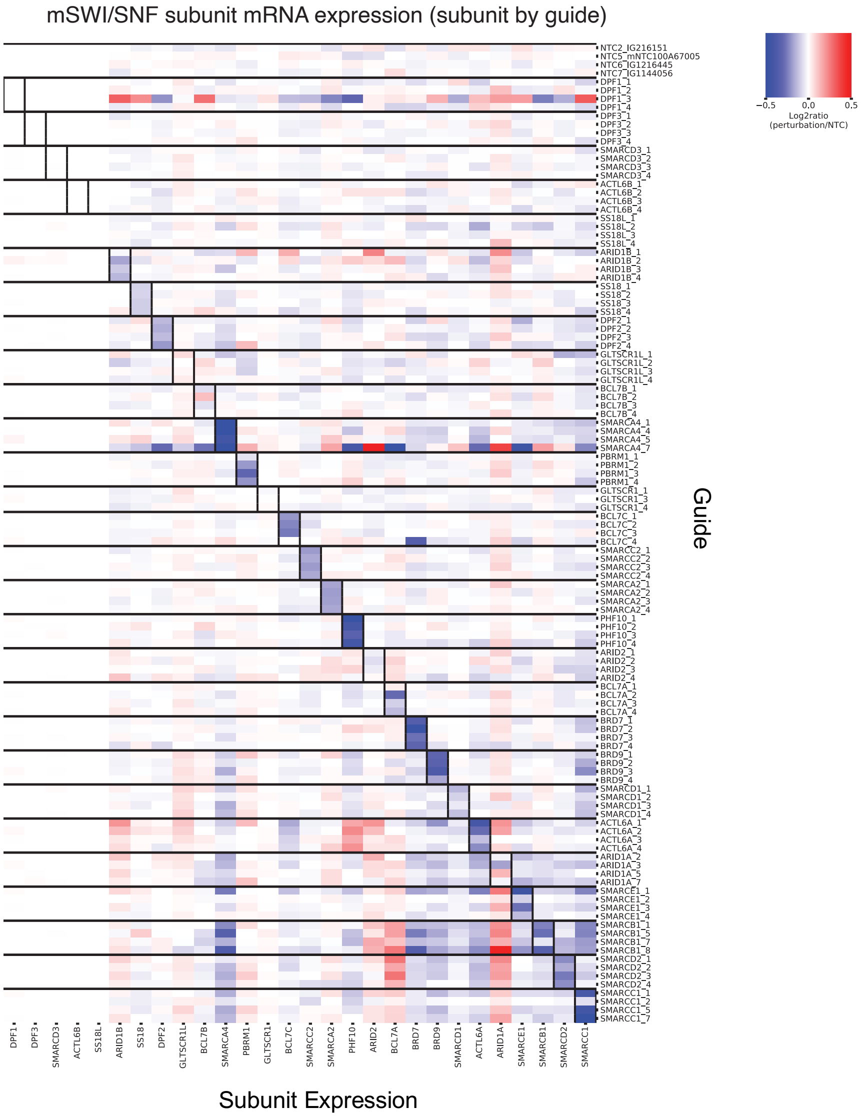
4. Sandoval, G. J. *et al.* Binding of TMPRSS2-ERG to BAF Chromatin Remodeling Complexes Mediates Prostate Oncogenesis. *Molecular Cell* **71**, 554-566.e7 (2018).

Chapter 2 Supplemental Figures

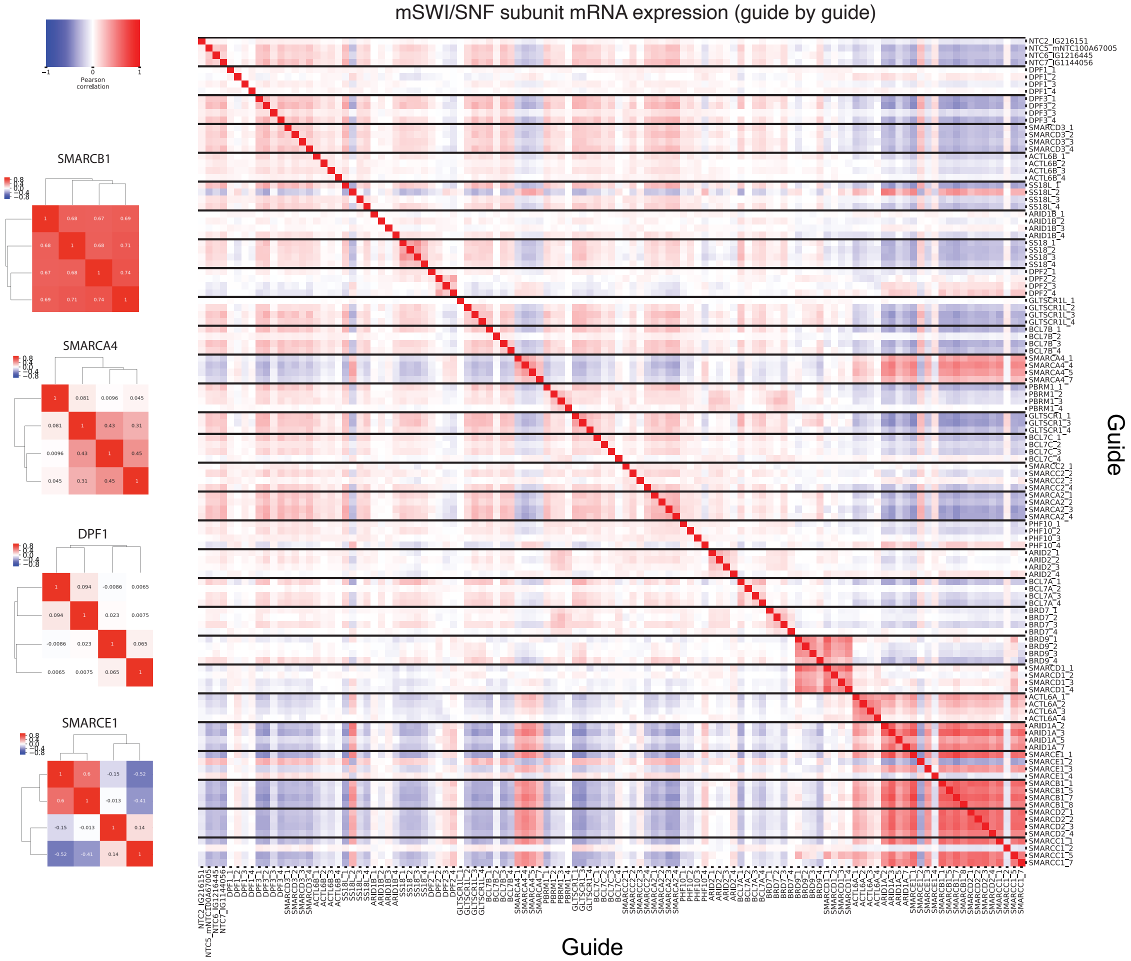


Supplemental Figure 2.S1 Perturb-seq experimental guide systems, validation, and cell recovery

A. Vector diagram of lentiCas9-Blast and pBA571 barcoded guide library constructs used in the Perturb-seq experimental system. **B.** Immunoblot of Flag-tagged Cas9 expression in MOLM-13 Cells. **C.** Fluoresence activated cell sorting (FACS) plots for cells without guide vector transduction (left) and with guide vector transduction (right) showing cell populations positive for BFP expression. **D.** Fitness effects calculated as percent expected recovery using the cell/guide recovery form 10x single cell profiling and the initial viral titer in the guide pool as calculated by BFP sort values given in (C). **E.** UMAP Louvain and batch plots demonstrating transcriptionally unique clusters obtained by single cell profiling and little to no batch effects across 15 channels of 10x single cell isolation and library construction, respectively.

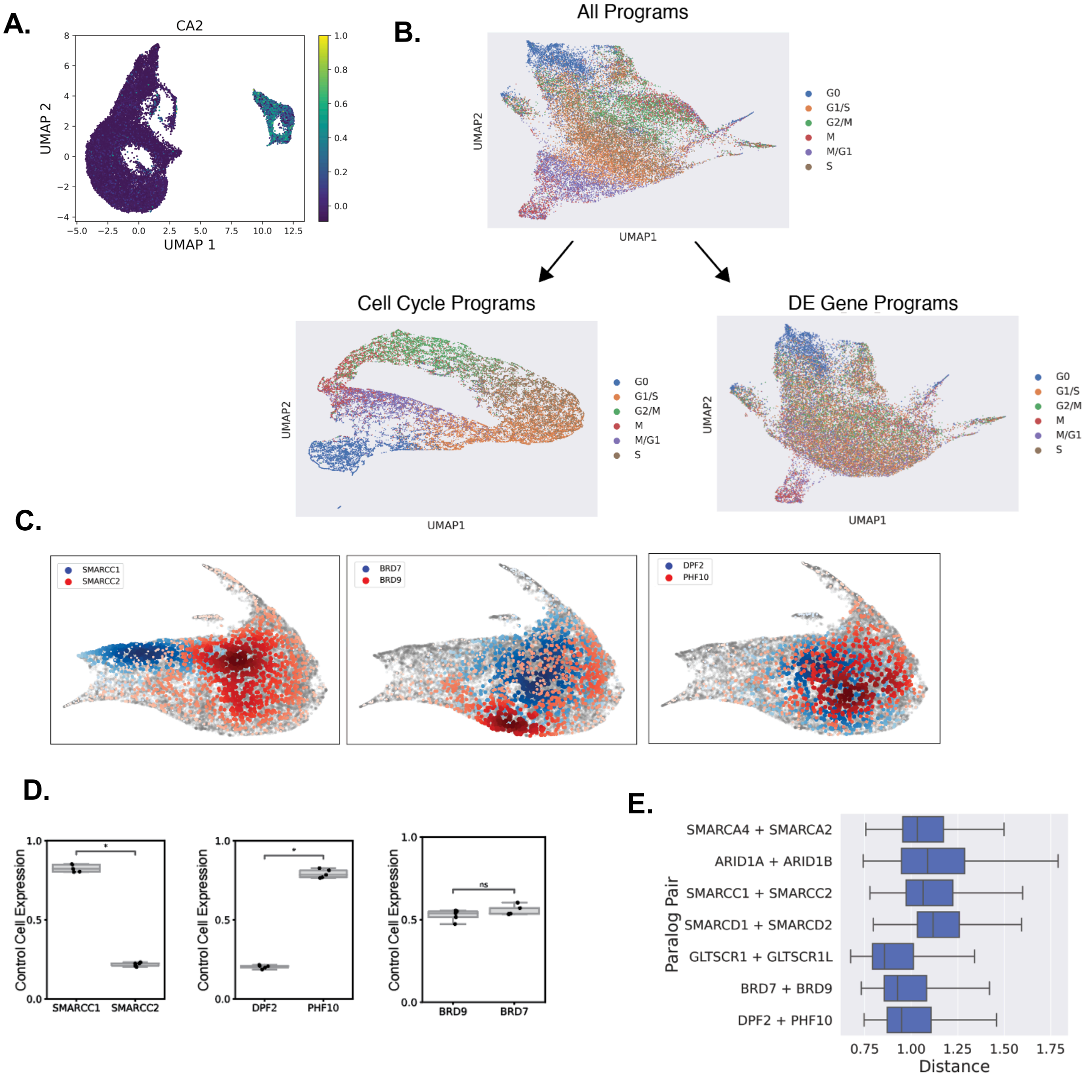


Supplemental Figure 2.S2 mSWI/SNF subunit expression by guide condition
Heatmap of mSWI/SNF subunit mRNA expression levels relative to control cell expression for each guide condition. Sets of guides are left/right outlined in black within the heatmap.



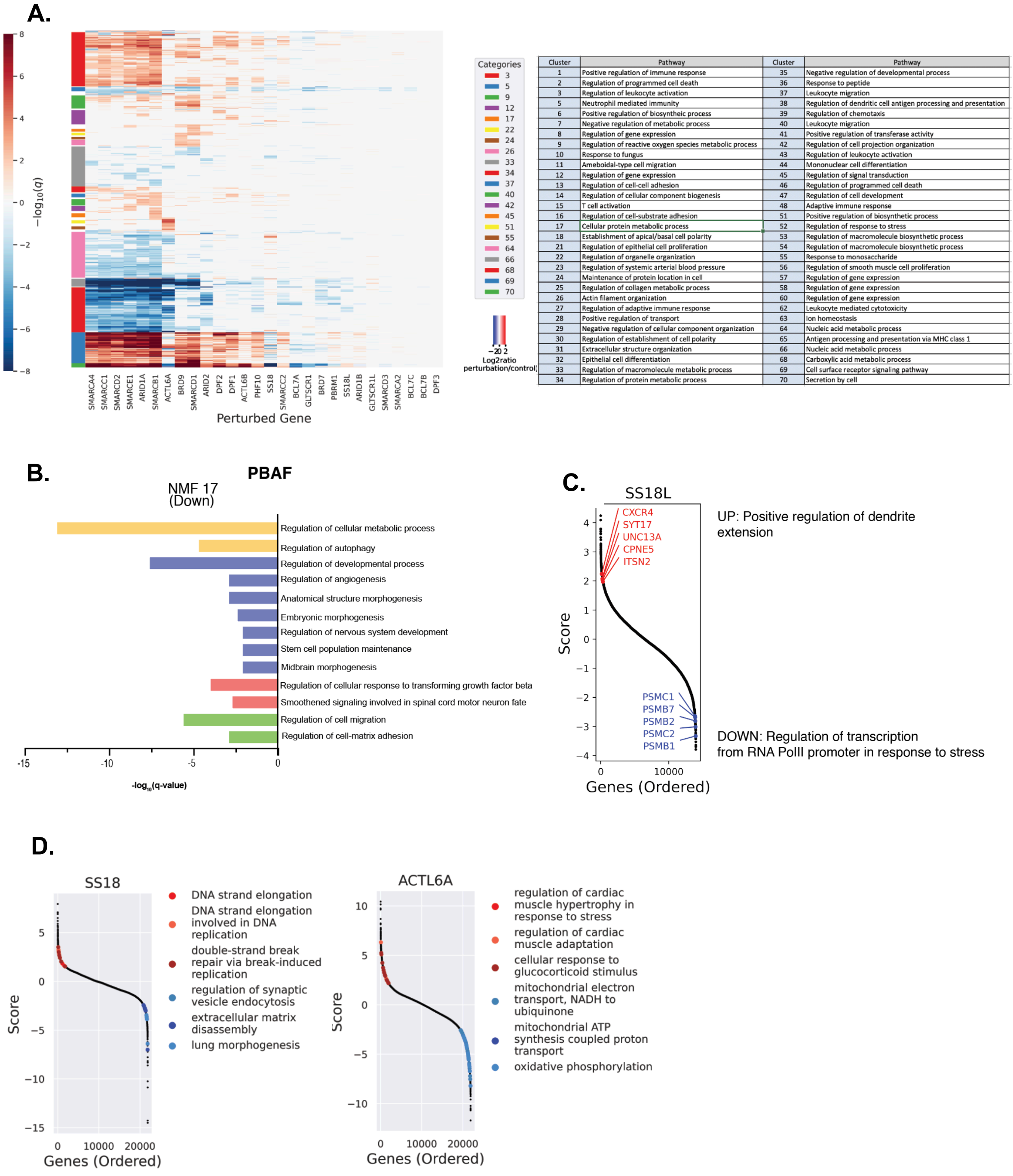
Supplemental Figure 2.S3 Guide correlation for mSWI/SNF subunit single perturbation experiment.

Left (Guide correlation heatmaps for 4 mSWI/SNF subunit perturbations depicting multiple case scenarios of similarity between transcriptomes in each guide condition; (right) global transcriptome similarity for all guides.



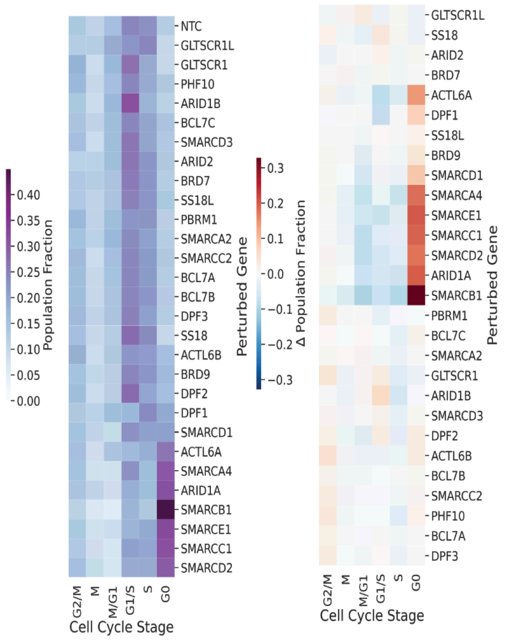
Supplemental Figure 2.S4 Cell program processing and localization changes upon mSWI/SNF paralog perturbation

A. UMAP plot reflecting all cells recovered from Perturb-seq experiment, showing two distinct populations of cells within the MOLM-13 cell line. The expression of CA2 is highly correlated with the differences in these unique subpopulations. Downstream analysis is restricted to the larger subpopulation of cells (>85%) **B.** Schematic depicting the decomposition of terms related to cell cycle or non-cell cycle related differentially expressed gene programs. **C.** Distribution of paralog pairs SMARCC1/C2, BRD7/9, and DPF2/PHF10 in UMAP space. **D.** Expression of paralog subunits in control cells for each pair of subunits from (C) with expression represented as log(1+TP10K). Significant difference in expression is denoted by asterisk (*) above boxes, p<0.01. **E.** Boxplot of distance between paralog pairs computed as the difference in the low-dimensional representations of cells of the 2 paralogs.

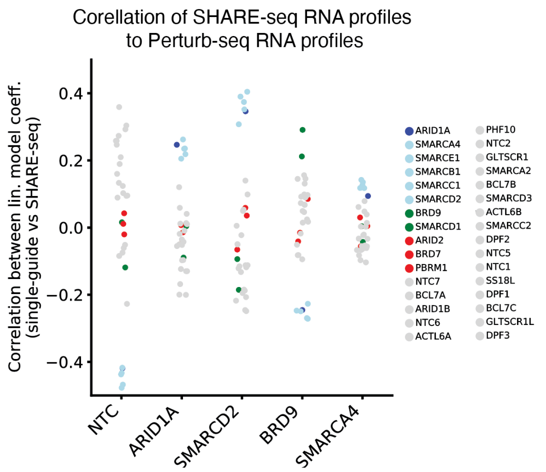


Cluster	Pathway	Cluster	Pathway
1	Positive regulation of immune response	35	Negative regulation of developmental process
2	Regulation of programmed cell death	36	Response to peptide
3	Regulation of leukocyte activation	37	Leukocyte migration
5	Neutrophil mediated immunity	38	Regulation of dendritic cell antigen processing and presentation
6	Positive regulation of biosyntheic process	39	Regulation of chemotaxis
7	Negative regulation of metabolic process	40	Leukocyte migration
8	Regulation of gene expression	41	Positive regulation of transferase activity
9	Regulation of reactive oxygen species metabolic process	42	Regulation of cell projection organization
10	Response to fungus	43	Regulation of leukocyte activation
11	Ameboidal-type cell migration	44	Mononuclear cell differentiation
12	Regulation of gene expression	45	Regulation of signal transduction
13	Regulation of cell-cell adhesion	46	Regulation of programmed cell death
14	Regulation of cellular component biogenesis	47	Regulation of cell development
15	T cell activation	48	Adaptive immune response
16	Regulation of cell-substrate adhesion	51	Positive regulation of biosynthetic process
17	Cellular protein metabolic process	52	Regulation of response to stress
18	Establishment of apical/basal cell polarity	53	Regulation of macromolecule biosynthetic process
21	Regulation of epithelial cell proliferation	54	Regulation of macromolecule biosynthetic process
22	Regulation of organelle organization	55	Response to monosaccharide
23	Regulation of systemic arterial blood pressure	56	Regulation of smooth muscle cell proliferation
24	Maintenance of protein location in cell	57	Regulation of gene expression
25	Regulation of collagen metabolic process	58	Regulation of gene expression
26	Actin filament organization	60	Regulation of gene expression
27	Regulation of adaptive immune response	62	Leukocyte mediated cytotoxicity
28	Positive regulation of transport	63	Ion homeostasis
29	Negative regulation of cellular component organization	64	Nucleic acid metabolic process
30	Regulation of establishment of cell polarity	65	Antigen processing and presentation via MHC class 1
31	Extracellular structure organization	66	Nucleic acid metabolic process
32	Epithelial cell differentiation	68	Carboxylic acid metabolic process
33	Regulation of macromolecule metabolic process	69	Cell surface receptor signaling pathway
34	Regulation of protein metabolic process	70	Secretion by cell

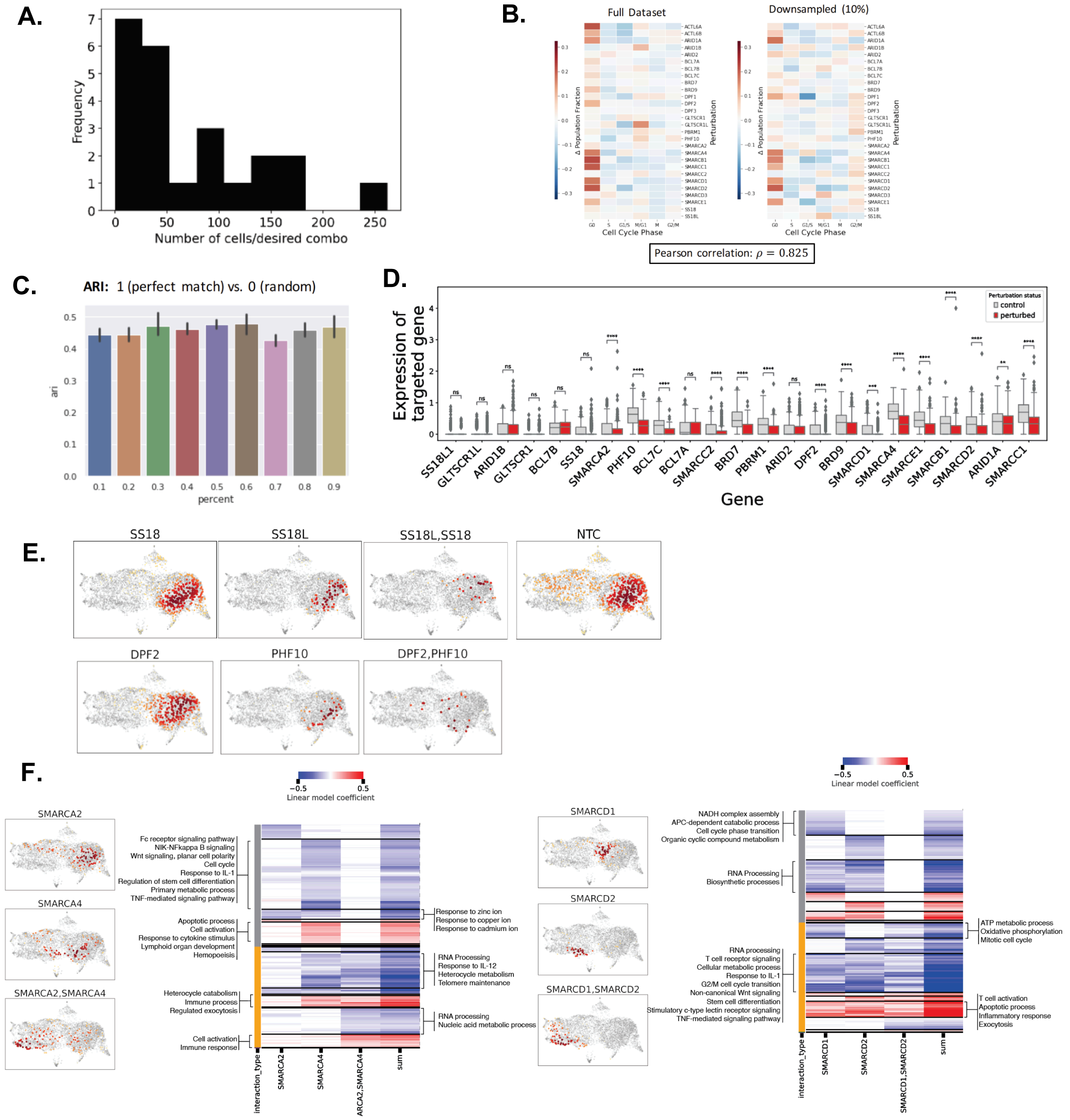
Supplemental Figure 2.S5 Differential gene sets affected by perturbation of specific complex subtypes and subunits. A. Heatmap of gene ontologies for all BAF subunit perturbations (left). Level 3 GO tree term is used to annotate each cluster in table (right). **B.** Gene ontologies associated with the selected NMF 17, which defines the PBAF subunit perturbations BRD7 and PBRM1. **C.** Differentially expressed genes composing the indicated pathways for the SS18L1 subunit perturbation. **D.** Gene ontologies for differentially expressed gene sets for select mSWI/SNF subunit perturbations.



Supplemental Figure 2.S6 Cell cycle effects of mSWI/SNF Perturbations
Cell Cell cycle phase distributions of cells receiving mSWI/SNF subunit and control perturbations (left), and the change in cell cycle phase distribution for mSWI/SNF perturbed cells normalized to the control cell cycle phase distribution (right).

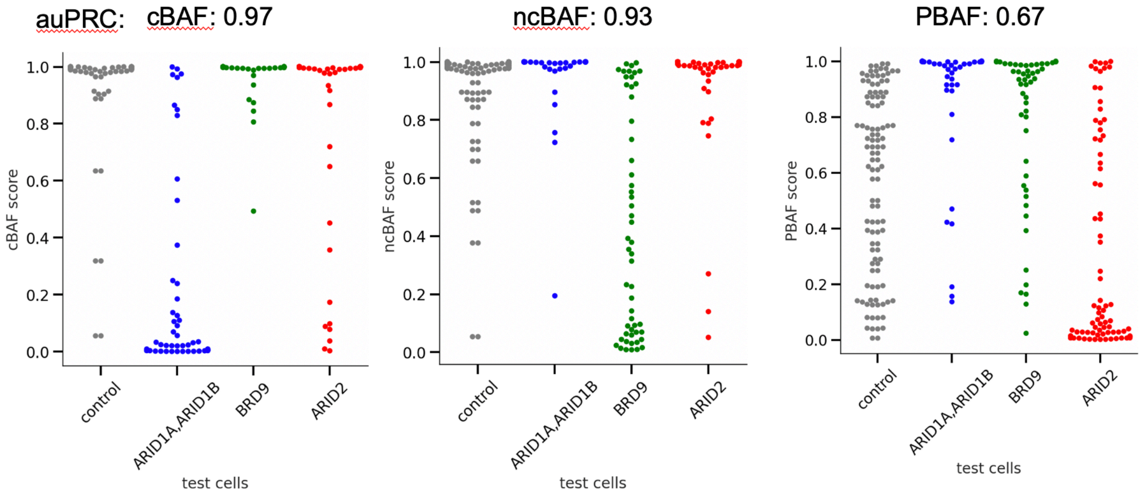


Supplemental Figure 2.S7 SHARE-seq RNA correlation to Perturb-seq RNA
Dotplot of correlation scores for each SHARE-seq perturbation RNA expression profile to Perturb-seq defined transcriptomic changes.



Supplemental Figure 2.S8 Combinatorial perturbation experiment yields high quality data and allows for identification of new cell states upon combination perturbation in comparison to single perturbation states

A. Histogram of recovery of cells for each desired combination of guides. **B.** Downsampling analysis based on cell cycle signatures in single cells, showing high correlation of expression signatures in full and downsampled dataset. **C.** ARI values for downsampled fractions of cells in the combination study. **D.** Expression of each mSWI/SNF subunit gene in control guide-receiving cells (left, gray) and expression in cells with the given mSWI/SNF subunit perturbation in singly infected cells (Right, red). **E.** UMAP distributions of singly and doubly perturbed cells recovered in the combination perturbation study. **F.** Low-dimensional representation in UMAP space of singly and multiply perturbed cells alongside linear interaction term heatmaps for the paralog families of SMARCA2/4 (left) and SMARCD1/2 (right).



Supplemental Figure 2.S9 QC for logistic regression classifier for BAF activity model auPRC values for the performance of each of the test populations for perturbations of each of the complex-defining subunits indicated.

Chapter 2 Supplemental Tables

Subunit	Guide#	Guide Sense Sequence	Barcode
SMARCA4	5	GGAGCGGCTGACCTGTGAGG	CCCATGTTCAACCAGTAG
SMARCA4	4	GATCATCAAGGACGACGCGG	CTACGCCCAAGTGCGTCT
SMARCA4	7	GACACCCCATCCCCACCCAG	TCTGGGTGCGTAGTCGTG
SMARCA4	1	GGCATGCTCAGAGCCACCCA	GCACTTCTACACAGACAC
ARID1A	5	GTTGCCCAGGCTGCTGGCGG	GAGATTACCCAGTGAAAT
ARID1A	7	TGGCGGCAGCAGCGATGGGG	TAACCAATTTGAAAAATG
ARID1A	3	GATGCCAGGCAGGTGAGGGG	AGACCTCTGACTCTAAAA
ARID1A	2	GAGGCGCTGGAGGAGGGAGG	GGAGCCCTTTTACACATC
SMARCB1	1	GAGAACCTCGGAACATACGG	ACACATCCGGCCGGGTGG
SMARCB1	7	GCGACCAGGACAGGAACACG	GTGTTTTGCGAAACTAAC
SMARCB1	5	GCAGATCGAGTCCTACCCCA	CCATTCAATTAGCTGTCG
SMARCB1	8	AGAGATACCCCTCACTCTGG	GCCGGGAACGTCACCTAG
SMARCC1	2	GAATGAGGAGGATTATGAGG	ACGCCTTTACAAGTGAAA
SMARCC1	5	GGAGACCCTTCTACTCCTGG	AAGGTGTATCACAGTGTC
SMARCC1	7	GGTGGGATCCACTTTCCCAG	GTACCCACTCTCCCGTAG
SMARCC1	1	CACGGGCTCGGGGATTGCGG	CCTAATTCATATTTGAGT
ARID2	1	GTAAGCCAGCCAGCTCAACA	CAACGTCTTCCTGTACTA
ARID2	2	GCAGTCTCCATTACACACAG	CTGTCTGGAGGCATGTGG
ARID2	3	TGTGGTAGGAGTAAAACGGA	AGGACGGGTCAGACGATG
ARID2	4	TTTACTACTTGCTAATGCCG	TGTGCGATGTCTACAGCG
PHF10	1	AGGTTATCCAGGTACCTCAA	GATACGCATTATTGCCAG
PHF10	2	CACCATCACTGTCTAGAGCA	GTAAAGTGACGCCATAGG
PHF10	3	CCCTTCAGATACATTGCCAG	TTCGAGGGACTGGTTCGG
PHF10	4	GGACCCAGCCATCCAAAAGG	AGTGGCATGTCTAAGCAG
SS18	1	ATGATGGGTCAAGTTAACCA	GCAGAGAGTAACGCCCGG
SS18	2	AATCAGATGACAATGAGTCA	CTCTTAATGCATGTACAG
SS18	3	GGCATGTTGTGAGAGCGTGG	CAGCCAAGAAGTACGTAT
SS18	4	CCTAACCATATGCCTATGCA	AGTCCGTCGACACACGCG
SMARCE1	1	TATGTAAGCAAGGTACGCGG	CACTACATCGTTTACGCG
SMARCE1	2	TCGACAGAGACAATCTCGCA	CACGTCCACCGTAGAGTG
SMARCE1	3	TGAAATTCTTAGTGAGAGTG	CATGGACTTCGGGAGTCG
SMARCE1	4	TTGATTCTCCTACCGTGACC	CCAACTTTGGGTTAAGAC
SMARCD1	1	GAAACGGCTAGATATCCAAG	GAATATCATCGAGAATCG
SMARCD1	2	TGACAAACTCCCGCTCGTGA	CCGACGTAAGCATCGTCG
SMARCD1	3	CCTGGTAATCCAGCATCAGT	GCCCTCCGATCGGTGACG
SMARCD1	4	GAGCGGTACAGCCCTTGACC	TATACACTGCGTTAGAGG
SMARCD2	1	GGGAGTGCACACGCAGACGA	AGCAGTCATGTGTGCAAG
SMARCD2	2	ACCAGACCATTGCTCGCAAG	GGACCGCTGTTATAATTG
SMARCD2	3	AAGCAGGATGCTCTTACCCC	GCACGTGTGACATAAATG
SMARCD2	4	AAGGCGGAAGGCGATAGTGC	TAAAAACGTATCGTCCTT
SMARCD3	1	CGGGGATCCAGTTTGAACTG	AACTACCTCGCTCGAGTG
SMARCD3	2	CTCTTGGTCTTACCTCAACG	CCGCTTGGACGGCGTTAC
SMARCD3	3	GGTCTTCACATACTGCCACA	CGCAGGCACACTTGGACT
SMARCD3	4	CCATGTAAGCCTGGGACTCG	AGACGCCAGATACCGGTA
SMARCA2	1	CCGTGGAACTAAAAGCACTT	GCCTTACAGACTGTACCA
SMARCA2	2	CTCCCAGTCCTACTACACCG	CTTAGCTGAAAGGACCTA
SMARCA2	3	GTCTCCAGCCCTATGTCTGG	TTAAATCGTTAGCCACCG
SMARCA2	4	CAGATTGGCTACATACTCAT	CTCTGCCCCACGACCCAG
ARID1B	1	TTGAGTGCAAGATCGAACGT	TAGGAGGTCCTAAACCGA
ARID1B	2	TGCCAATTGGATACCGCTGT	CGGTTGCGGCCAGACTGT
ARID1B	3	GTAATTATTAAACTCCGGGA	GAATGTTGTTGCACACGT
ARID1B	4	TGTGAACCGTAAGGCACAGG	TCAGCAGAGTGCAGTCCA
DPF1	1	GTTTCCGCATGACCTCGAGG	AATCCCTCCTGGTAGGAC
DPF1	2	CCTGCGAGTACAAGATCGGT	ATGATGTGGCGGAAGCTT
DPF1	3	GTACGTGTAAATCTGTCCCG	GATGACGCACTCTCGCAT
DPF1	4	CCTCGACTCGCAGACCGGCG	TTGTGGACTGGAGGCGAT
DPF2	1	GAAGATACTCCCAAGCGTCG	GCCGAACCCTCCAGAGCG
DPF2	2	TGGATGGAAAAGCGACACCG	TGCGAAGTAACCCGTAGG
DPF2	3	ATAGATGGGAAGGAAAGTCG	AGCGTCGTCCAGCTAGCA
DPF2	4	ATCATCAACTCGGGGATCCG	ACGAGCGAGCCAGAACGG
DPF3	1	CACCTGAAGATCCAAAACTG	TGGTCTAAGGGACGTGAG
DPF3	2	AGCGCTACAAGAACCGACCG	CAGGTTCGGAGACGTTCT
DPF3	3	CAAGTAGGCACTCACGCCTG	CAGTCAGACATCAGCAGT
DPF3	4	AAGCGAAAGAACAGGACTAG	GCTAACGCCTACTACACG
SMARCC2	1	GACTCGGGGATCGACGACAG	TAGTGTGCGCTGCAAGGA
SMARCC2	2	TATGCCTATTACCTGCACCA	AACCTATCCGGACCCACG
SMARCC2	3	GGTACGACCAGTCATGAAGA	TTGAGCGGGTGGAGCCAG
SMARCC2	4	ATGAGGATGAGAACAGTACG	ACAGACAATCAATTCTGA
BCL7A	1	TCGTAGGGATGTGTCACCAA	CCCATTCGGCACACGCCG
BCL7A	2	TGGCTCAGAGGTGACCACTC	CTCACCGGACTAGTTGTT
BCL7A	3	TGATATCAAGAGGGTCATGG	ACGTATACGGCGTAGAGA
BCL7A	4	ATACTCACGTCATCAACCTT	CTTTGCAGAGGCACAGGG
BCL7B	1	AAGCTGATAGACGTCAGACA	ACGGACTCCACATACTAA
BCL7B	2	AACCCACTTAAATATCCTCA	GCTCGCAGGCGCCGAGAG
BCL7B	3	TGGCGGCCATCGAGAAAGTG	AACTGACTCCATGTCAGG
BCL7B	4	GTCATCCGTGCGGAAGTCGG	CACAGCAATAGCTGCACG
BCL7C	1	GCACCCACTTGAAGATACGA	GGTGAACTCTGGACTACA
BCL7C	2	TCCTGCTGGATCTTAATGGT	ACCTTGACAGACCCAATG
BCL7C	3	CCATCGAGAAGGTCCGGAGA	CGTCGTATGCGGCAACTG
BCL7C	4	ATGAGAGGGCCACCCCCTCG	ATACGATGCTCCAGTTCG
BRD7	1	GAAGTCACCGAACTCTCCAC	CATCTGAACCGAGCGCCG
BRD7	2	CCTGTGGATCCCATTGTAGG	CCGGAGGATTTACTTACG
BRD7	3	AACTGATGAGACAATTGCAG	TGGAGTAAGGCGCTACCG
BRD7	4	AGGCAAGTCTAATCTCACAG	AATGCACAACTGGGCTTG
BRD9	1	AGATACCGTGTACTACAAGT	GTCGTACCTTCCCATGTG
BRD9	2	AGGGAGCACTGTGACACGGA	GGCTGATAGGGAGTCTGA
BRD9	3	CTTGACGGACAGTACCGCAG	AAGTCGGTGAAACTATAG
BRD9	4	ACTCCAGTTACTATGATGAC	CTACACGAATTGGCCTCG
GLTSCR1	1	AGAGTTCCCATTGAGCGTGG	AGGAGGCTAGCGATTTTG
GLTSCR1	2	ACCATCTGGAAGGTCCGAGG	ACGCCTTTACAAGTGAAA
GLTSCR1	3	CATGAACGTCAGGTTCTGCG	GACTAAGCTGACATATTC
GLTSCR1	4	GCCAGCTCCTTTGGCGACGG	CCAAACATTAGAACTCAG
GLTSCR1L	1	ATGGCTTTATGCAACATGTG	CCGCACGTATTTATACTC
GLTSCR1L	2	TGACAGGTGTAATACAGTTG	ATTGTGACACAGAGAGCG
GLTSCR1L	3	AAGAGATCCCTTGTTGTACG	AAAGGTCATATCCTGTCG
GLTSCR1L	4	CAATATCACTGAACAGACAT	TCCACCACCGTCCGCAGT
PBRM1	1	TTGAAAATAATCGCTACCGT	AGACCGGTCGGTAACCCA
PBRM1	2	ATGTGCGATGTAAGCCTGAG	CCAAGGCACTCTTAGCGG
PBRM1	3	AGTAATAAAAGAGCAGTACA	CCACCGTCAAAGAATGCG
PBRM1	4	CAAATCCCAGAGTTTGCAAG	AATGCAACACGAATCATG
ACTL6A	1	ACTGCAATTCCAGTCCACGA	AGTGGGTGTACTGTCCGA
ACTL6A	2	AGAAGTTGCCTCAGGTTACG	TTTTGCGCTGAGGCAGGG
ACTL6A	3	ACCACCATACCAATAGCTGT	GCCGTACTTATTTAGCAG
ACTL6A	4	CTAATGCTCTGCGTGTTCCG	AGACAAGAGAGACGCAGG
ACTL6B	1	ATGCCAGCTTTGCAAACGGG	GATTCCCTCGCCCTGCAG
ACTL6B	2	CACTCACTGTTCATCGTAGG	AAACCGACGCCTTTTTGG
ACTL6B	3	ACCTTACATGATCGCAGCCA	GTGGTAGCATCGAGAACG
ACTL6B	4	TTCTCCACAGTGGAACACAC	GAACCTGTTACCCGGACA
SS18L	1	CGCCGAGCTGTAGTGCGACG	GCTAGCTCCCACTAAGCG
SS18L	2	GGTACTCACGGCAGGAAGCA	TCTTGTTGTCGGCGACAG
SS18L	3	CCATCGGCAACTACGTGTCT	ACCCGCAGACGCCCCCCG
SS18L	4	ATCCTGGAGTACCAGAGCAA	GATTTGTACCGAGGAAACA
Control	1	TGATGGGCTTCAAGCAGTAG	CACACGACGACAGTATCG
Control	2	CAAGTGTCACAATGGACCAT	AAGTTAGACGCCGTCCTC
Control	5	TATGATCGTATGGCCCTTCC	CTGCGCTCAGGTGACCTG
Control	6	CCAGCTGCCCAGGAATTCAG	CCGAGGACAACAATTTCG
Control	7	AAACTCACAGATCTAGAGGA	CGGCCAAAGACCACAATG

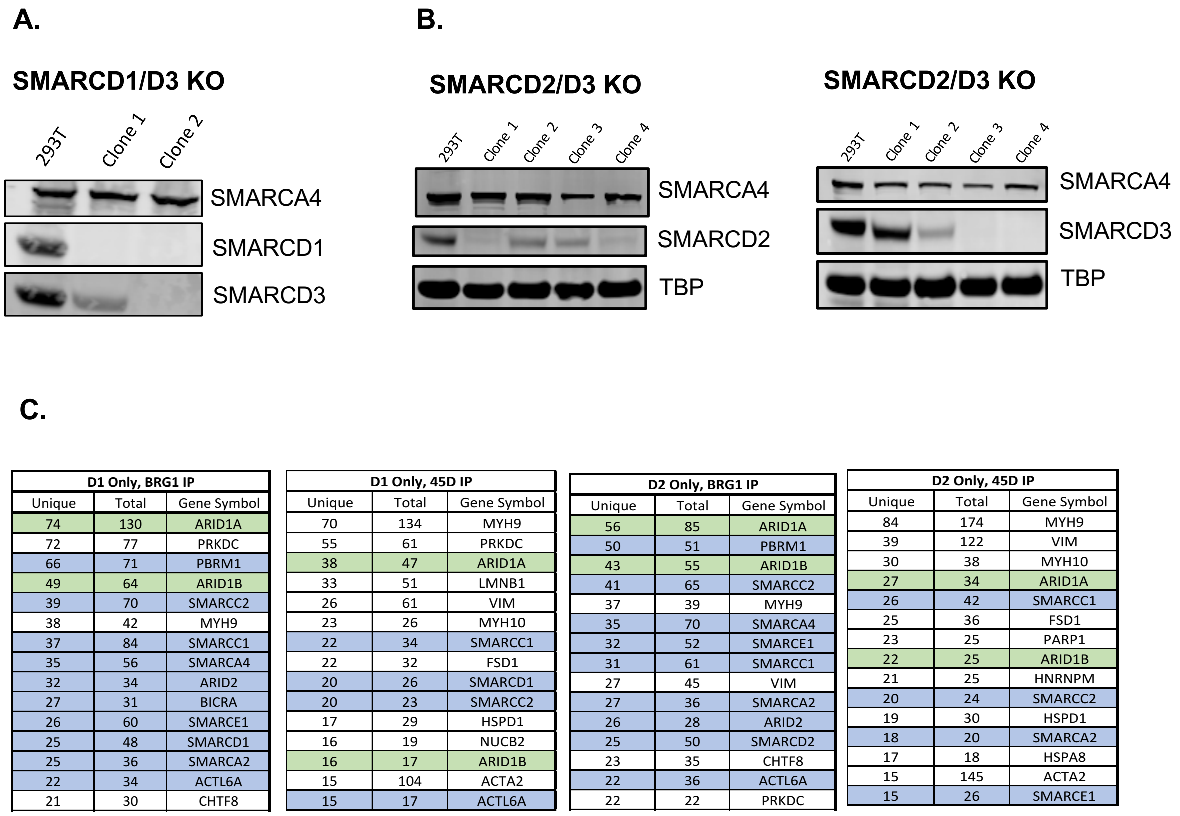
Table S2.T1 CRISPR-Cas9 knockout guides used in Perturb-seq study

Dial Out Primers	
P7 sequence:	CAAGCAGAAGACGGCATACGAGAT
Read 2:	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC
BFP binding sequence:	TAGCAAACTGGGGCACAAGCTTAAT
TruSeq Universal Primer:	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

Primer Number	Index
1	GTTACCTG
2	ACAGTTCC
3	TACGTCAT
4	GGAATACA
5	TATTCCGT
6	AGTGTACG
7	GATTCGAT
8	TCGTACGT
9	TACCTTTG
10	GCTCGCTT
11	TAGGCTGA
12	GTAGTAAT
13	TAAGCCTA
14	AATGGCAT
15	CCTAGATG

Tables S2.T2-3 Dial Out primer design
Constant region sequences for dial out primers used in Perturb-seq experiment (top) and variable index sequences (bottom).

Chapter 3 Supplemental Figures



D1 Only, BRG1 IP		
Unique	Total	Gene Symbol
74	130	ARID1A
72	77	PRKDC
66	71	PBRM1
49	64	ARID1B
39	70	SMARCC2
38	42	MYH9
37	84	SMARCC1
35	56	SMARCA4
32	34	ARID2
27	31	BICRA
26	60	SMARCE1
25	48	SMARCD1
25	36	SMARCA2
22	34	ACTL6A
21	30	CHTF8

D1 Only, 45D IP		
Unique	Total	Gene Symbol
70	134	MYH9
55	61	PRKDC
38	47	ARID1A
33	51	LMNB1
26	61	VIM
23	26	MYH10
22	34	SMARCC1
22	32	FSD1
20	26	SMARCD1
20	23	SMARCC2
17	29	HSPD1
16	19	NUCB2
16	17	ARID1B
15	104	ACTA2
15	17	ACTL6A

D2 Only, BRG1 IP		
Unique	Total	Gene Symbol
56	85	ARID1A
50	51	PBRM1
43	55	ARID1B
41	65	SMARCC2
37	39	MYH9
35	70	SMARCA4
32	52	SMARCE1
31	61	SMARCC1
27	45	VIM
27	36	SMARCA2
26	28	ARID2
25	50	SMARCD2
23	35	CHTF8
22	36	ACTL6A
22	22	PRKDC

D2 Only, 45D IP		
Unique	Total	Gene Symbol
84	174	MYH9
39	122	VIM
30	38	MYH10
27	34	ARID1A
26	42	SMARCC1
25	36	FSD1
23	25	PARP1
22	25	ARID1B
21	25	HNRNPM
20	24	SMARCC2
19	30	HSPD1
18	20	SMARCA2
17	18	HSPA8
15	145	ACTA2
15	26	SMARCE1

Figure 3.S1 SMARCD1 and SMARCD2-only cell line creation and mass spectrometry analysis

A. Immunoblot of SMARCD1 and SMARCD3 expression in HEK293T clonal SMARCD knockout populations. Clone 2 is the population used for the IP-MS studies of SMARCD2-only cells. **B.** Immunoblots of SMARCD2 and SMARCD3 expression in HEK293T clonal SMARCD knockout populations. Clone 4 is the population used for the IP-MS studies of SMARCD2 only-cells. **C.** Mass spectrometry peptide analysis for SMARCD1-only or SMARCD2-only-containing cells in both the SMARCA4 and DPF2 IP conditions.

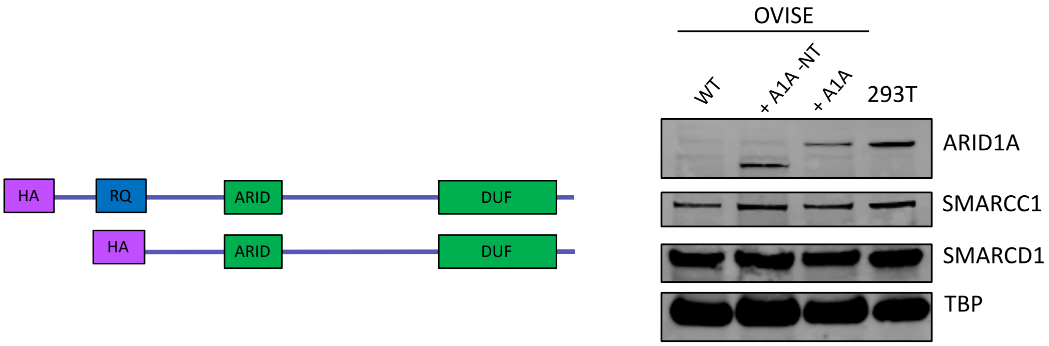


Figure 3.S2 ARID1A expression constructs for investigation of N-terminus
(Left) Full length and truncated ARID1A expression constructs and (Right) immunoblot of expression of these constructs in OVISE cells.

Chapter 3 Supplemental Tables

Table 3.T1 Domain-targeting guides for CRISPR-Cas9 domain scan

internal_id	strand	cut_codon	cut_site	sequence	Domain
ARID1B_6750.1151	-	419	1257	GTGAGCAGCTGATTGAGGGT	LXXLL /region_name: "propagated from UniProtKB/Swiss-Prot (Q8NFD5.2)"
ARID1B_6750.1150	-	420	1261	CGAGGTGAGCAGCTGATTGA	LXXLL /region_name: "propagated from UniProtKB/Swiss-Prot (Q8NFD5.2)"
ARID1B_6750.1149	-	420	1262	GCGAGGTGAGCAGCTGATTG	LXXLL /region_name: "propagated from UniProtKB/Swiss-Prot (Q8NFD5.2)"
ARID1B_6750.262	+	712	2137	CCAAGGGGATCAGAGCAACC	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.1036	-	716	2150	GGGAGAAAGGCGACTGCGCC	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.1035	-	717	2151	GGGGAGAAAGGCGACTGCGC	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.1034	-	721	2163	GGGGACGCATGTGGGGAGAA	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.1033	-	723	2170	GAGATGAGGGGACGCATGTG	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.1032	-	723	2171	AGAGATGAGGGGACGCATGT	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.1031	-	724	2172	GAGAGATGAGGGGACGCATG	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.1030	-	727	2182	CGGGATGCTGGAGAGATGAG	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.1029	-	727	2183	CCGGGATGCTGGAGAGATGA	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.1028	-	728	2184	CCCGGGATGCTGGAGAGATG	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.263	+	729	2188	CCCTCATCTCTCCAGCATCC	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.264	+	729	2189	CCTCATCTCTCCAGCATCCC	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.265	+	730	2190	CTCATCTCTCCAGCATCCCG	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.266	+	730	2191	TCATCTCTCCAGCATCCCGG	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.267	+	730	2192	CATCTCTCCAGCATCCCGGG	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.1027	-	731	2194	AGATGGGCCCCCCGGGATGC	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.1026	-	733	2201	GAGAGGGAGATGGGCCCCCC	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.1025	-	734	2202	GGAGAGGGAGATGGGCCCCC	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.1024	-	736	2210	AGCCAACAGGAGAGGGAGAT	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.1023	-	737	2211	GAGCCAACAGGAGAGGGAGA	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.268	+	737	2213	GGCCCATCTCCCTCTCCTGT	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.1022	-	739	2217	ACAGGAGAGCCAACAGGAGA	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.1021	-	739	2218	TACAGGAGAGCCAACAGGAG	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.1020	-	741	2223	CTTCCTACAGGAGAGCCAAC	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.269	+	741	2225	TCTCCTGTTGGCTCTCCTGT	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.1019	-	745	2235	CGAGACTGGTTGCTTCCTAC	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.270	+	748	2246	GGAAGCAACCAGTCTCGATC	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.1018	-	749	2249	AGATTGGGCCAGATCGAGAC	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.1017	-	754	2264	GGATACTTGCAGGAGAGATT	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.1016	-	755	2265	GGGATACTTGCAGGAGAGAT	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.272	+	766	2300	CAGATGCCTCCGCAGCCACC	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.1012	-	767	2301	TGGCTCCCGGGTGGCTGCGG	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.273	+	767	2301	AGATGCCTCCGCAGCCACCC	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.1011	-	768	2304	GACTGGCTCCCGGGTGGCTG	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.1010	-	770	2310	GATTCTGACTGGCTCCCGGG	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.1009	-	771	2313	CTGGATTCTGACTGGCTCCC	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.1008	-	771	2314	ACTGGATTCTGACTGGCTCC	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.1007	-	773	2321	GATGGGAACTGGATTCTGAC	Glyco_hydro_81 /region_name: "Glycosyl hydrolase family 81; cl02310"
ARID1B_6750.922	-	1069	3207	TCGACCCAGAGCTTTCTCTC	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.343	+	1069	3207	ATGAGCCAGAGAGAAAGCTC	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.344	+	1069	3208	TGAGCCAGAGAGAAAGCTCT	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.921	-	1076	3230	CTTCCATGAAGGTGAGGTAT	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.345	+	1077	3232	CGACCGATACCTCACCTTCA	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.920	-	1078	3236	CTCTCTCTTCCATGAAGGTG	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.919	-	1080	3241	AGAGCCTCTCTCTTCCATGA	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.346	+	1080	3242	CTCACCTTCATGGAAGAGAG	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.918	-	1088	3264	ACGGCAGGCAGACTTGAGAC	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.347	+	1090	3271	TGTCTCAAGTCTGCCTGCCG	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.348	+	1090	3272	GTCTCAAGTCTGCCTGCCGT	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.917	-	1093	3279	AGGGGCTTCTTGCCCACGGC	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.916	-	1094	3283	GTCCAGGGGCTTCTTGCCCA	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.349	+	1095	3286	TGCCGTGGGCAAGAAGCCCC	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.915	-	1099	3297	TAGAGTCGGAACAGGTCCAG	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.914	-	1099	3298	GTAGAGTCGGAACAGGTCCA	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.913	-	1099	3299	CGTAGAGTCGGAACAGGTCC	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.912	-	1101	3305	CGCAGACGTAGAGTCGGAAC	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.911	-	1103	3311	CTTTGACGCAGACGTAGAGT	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.350	+	1107	3323	TACGTCTGCGTCAAAGAGAT	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.351	+	1108	3324	ACGTCTGCGTCAAAGAGATC	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.352	+	1108	3325	CGTCTGCGTCAAAGAGATCG	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.353	+	1108	3326	GTCTGCGTCAAAGAGATCGG	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.354	+	1110	3331	CGTCAAAGAGATCGGGGGTT	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.357	+	1122	3367	AAACAAGAAGTGGCGTGAGC	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.358	+	1128	3386	CTGGCAACCAACCTAAACGT	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.908	-	1129	3388	TGAGGTGCCAACGTTTAGGT	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.907	-	1130	3392	TGCTTGAGGTGCCAACGTTT	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.906	-	1135	3406	GGAGCTCGCTGCACTGCTTG	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.905	-	1142	3427	CTGAATATACTGCTTTTTCA	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.904	-	1142	3428	ACTGAATATACTGCTTTTTC	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.903	-	1150	3452	TCTTGCACTCAAAGGCAAAC	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.902	-	1153	3460	ACGTTCGATCTTGCACTCAA	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.359	+	1155	3467	TTTGAGTGCAAGATCGAACG	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.360	+	1156	3468	TTGAGTGCAAGATCGAACGT	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.361	+	1156	3469	TGAGTGCAAGATCGAACGTG	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;
ARID1B_6750.362	+	1157	3472	GTGCAAGATCGAACGTGGGG	BRIGHT /region_name: "BRIGHT, ARID (A/T-rich interaction domain) domain;

ARID1B_6750.651	-	1951	5854	TGACAAGCTACGGACAATAT	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.650	-	1954	5864	CAGGCACGAATGACAAGCTA	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.519	+	1956	5870	CGTAGCTTGTCATTCGTGCC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.649	-	1961	5883	GACATTTCGGCATCATTGCC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.648	-	1965	5896	GCCTGGATGTTTGGACATTT	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.520	+	1966	5900	GCCGAAATGTCCAAACATCC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.521	+	1968	5905	AATGTCCAAACATCCAGGCC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.647	-	1968	5905	CAGCACCAGGCCTGGATGTT	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.646	-	1971	5913	CCCAGGATCAGCACCAGGCC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.522	+	1972	5917	TCCAGGCCTGGTGCTGATCC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.645	-	1972	5918	GCTTCCCCAGGATCAGCACC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.523	+	1972	5918	CCAGGCCTGGTGCTGATCCT	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.524	+	1973	5919	CAGGCCTGGTGCTGATCCTG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.644	-	1976	5930	GAAGAAGAATCAGCTTCCCC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.643	-	1984	5954	GCTTTCTCTCTGGATGCTCG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.642	-	1988	5964	TGCGGTGCTCGCTTTCTCTC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.641	-	1994	5982	TCCTCTTTCTCATAGGTCTG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.525	+	1995	5986	ACCGCAGACCTATGAGAAAG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.526	+	1996	5989	GCAGACCTATGAGAAAGAGG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.640	-	1996	5989	CTCATCCTCCTCTTTCTCAT	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.527	+	1998	5995	CTATGAGAAAGAGGAGGATG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.528	+	2000	6001	GAAAGAGGAGGATGAGGACA	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.529	+	2000	6002	AAAGAGGAGGATGAGGACAA	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.530	+	2001	6003	AAGAGGAGGATGAGGACAAG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.531	+	2001	6004	AGAGGAGGATGAGGACAAGG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.532	+	2002	6007	GGAGGATGAGGACAAGGGGG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.639	-	2008	6025	CCACCACTCATCTTTGCTGC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.533	+	2009	6027	TGGCCTGCAGCAAAGATGAG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.534	+	2010	6030	CCTGCAGCAAAGATGAGTGG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.535	+	2011	6033	GCAGCAAAGATGAGTGGTGG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.536	+	2011	6034	CAGCAAAGATGAGTGGTGGT	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.537	+	2015	6046	GTGGTGGTGGGACTGCCTCG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.538	+	2018	6054	GGGACTGCCTCGAGGTCTTG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.539	+	2018	6055	GGACTGCCTCGAGGTCTTGA	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.638	-	2018	6056	TGTTATCCCTCAAGACCTCG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.540	+	2022	6067	GGTCTTGAGGGATAACACGT	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.541	+	2025	6076	GGATAACACGTTGGTCACGT	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.542	+	2029	6089	GTCACGTTGGCCAACATTTC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.543	+	2030	6090	TCACGTTGGCCAACATTTCC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.637	-	2031	6094	GTCTAGCTGCCCGGAAATGT	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.636	-	2034	6103	AGCAGACAAGTCTAGCTGCC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.544	+	2038	6115	GCTAGACTTGTCTGCTTACA	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.545	+	2046	6139	AAGCATCTGCTTGCCAATTT	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.546	+	2047	6143	ATCTGCTTGCCAATTTTGGA	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.635	-	2049	6147	TGCAGCAAGCCATCCAAAAT	LXXLL /region_name: "propagated from UniProtKB/Swiss-Prot (Q8NFD5.2)
ARID1B_6750.547	+	2052	6156	TTTTGGATGGCTTGCTGCAC	DUF3518 /region_name: "Domain of unknown function (DUF3518)
ARID1B_6750.548	+	2053	6160	GGATGGCTTGCTGCACTGGA	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.549	+	2059	6178	GATGGTGTGCCCGTCTGCAG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.634	-	2060	6182	GATCTTGTGCCTCTGCAGAC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.633	-	2061	6183	GGATCTTGTGCCTCTGCAGA	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.550	+	2067	6202	ACAAGATCCCTTTCCAACTG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.551	+	2067	6203	CAAGATCCCTTTCCAACTGT	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.632	-	2068	6204	TTGGGTCCCACAGTTGGAAA	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.631	-	2068	6205	GTTGGGTCCCACAGTTGGAA	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.630	-	2070	6210	ACCGAGTTGGGTCCCACAGT	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.552	+	2071	6214	TCCAACTGTGGGACCCAACT	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.629	-	2074	6222	TGAGGCGACAGGACCGAGTT	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.628	-	2074	6223	CTGAGGCGACAGGACCGAGT	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.627	-	2077	6233	GCACAAGTCTCTGAGGCGAC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.626	-	2080	6240	GTCTCCAGCACAAGTCTCTG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.553	+	2080	6241	GTCGCCTCAGAGACTTGTGC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.625	-	2087	6262	CTGGATACTGAGTTTACAGA	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.624	-	2087	6263	CCTGGATACTGAGTTTACAG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.554	+	2089	6268	CCTCTGTAAACTCAGTATCC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.555	+	2093	6280	CAGTATCCAGGACAATAATG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.623	-	2093	6281	TCAGGTCCACATTATTGTCC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.556	+	2097	6292	CAATAATGTGGACCTGATCT	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.622	-	2099	6299	ATGGAGGAGTGGCCAAGATC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.621	-	2103	6310	CTGACGACTAAATGGAGGAG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.620	-	2105	6315	TTCTCCTGACGACTAAATGG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.557	+	2105	6316	CACTCCTCCATTTAGTCGTC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.619	-	2106	6318	AATTTCTCCTGACGACTAAA	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.558	+	2114	6342	AATTCTATGCTACATTAGTT	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.559	+	2116	6350	GCTACATTAGTTAGGTACGT	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.560	+	2117	6351	CTACATTAGTTAGGTACGTT	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.561	+	2117	6352	TACATTAGTTAGGTACGTTG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.618	-	2126	6380	TGGACATTTCTCGACAGACT	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.617	-	2127	6381	ATGGACATTTCTCGACAGAC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.562	+	2129	6388	AGTCTGTCGAGAAATGTCCA	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.616	-	2133	6400	AAGGTTCGATAAAAGCGCCA	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"

Table 3.T1 (continued) Domain-targeting guides for CRISPR-Cas9 domain scan

ARID1B_6750.565	+	2138	6415	TTTATCGAACCTTGCCCAAG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.615	-	2139	6419	CTAGTGCGTCCCCTTGGGCA	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.614	-	2141	6424	TGCTGCTAGTGCGTCCCCTT	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.613	-	2141	6425	TTGCTGCTAGTGCGTCCCCT	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.566	+	2144	6432	AAGGGGACGCACTAGCAGCA	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.567	+	2144	6433	AGGGGACGCACTAGCAGCAA	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.612	-	2150	6451	GCTTCCTTTCTGCACAGCTA	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.568	+	2150	6452	AGGGCCATAGCTGTGCAGAA	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.569	+	2153	6461	GCTGTGCAGAAAGGAAGCAT	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.570	+	2161	6484	AAACTTGATAAGCTTCCTAG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.571	+	2162	6488	TTGATAAGCTTCCTAGAGGA	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.572	+	2163	6489	TGATAAGCTTCCTAGAGGAT	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.573	+	2163	6490	GATAAGCTTCCTAGAGGATG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.611	-	2164	6494	CCATCGTGACCCCATCCTCT	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.574	+	2166	6499	CCTAGAGGATGGGGTCACGA	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.610	-	2172	6517	GTGCTGGCTCTGCTGGTACT	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.609	-	2172	6518	TGTGCTGGCTCTGCTGGTAC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.608	-	2174	6524	TGAGGTTGTGCTGGCTCTGC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.607	-	2177	6533	GCATGTGCATGAGGTTGTGC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.606	-	2180	6542	GCGGGGGCTGCATGTGCATG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.575	+	2184	6553	GCACATGCAGCCCCCGCCCC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.605	-	2186	6558	CTAGGTGGTTCCAGGGGCGG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.604	-	2186	6559	GCTAGGTGGTTCCAGGGGCG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.603	-	2186	6560	CGCTAGGTGGTTCCAGGGGC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.602	-	2187	6561	ACGCTAGGTGGTTCCAGGGG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.601	-	2188	6564	TCTACGCTAGGTGGTTCCAG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.600	-	2188	6565	GTCTACGCTAGGTGGTTCCA	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.599	-	2188	6566	TGTCTACGCTAGGTGGTTCC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.598	-	2191	6573	CACATCATGTCTACGCTAGG	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.597	-	2192	6576	CTGCACATCATGTCTACGCT	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.576	+	2194	6582	CTAGCGTAGACATGATGTGC	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"
ARID1B_6750.577	+	2194	6583	TAGCGTAGACATGATGTGCA	DUF3518 /region_name: "Domain of unknown function (DUF3518); pfam12031"

Table 3.T1 (continued) Domain-targeting guides for CRISPR-Cas9 domain scan